Essays in Econometrics

by

David W. Hughes

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Economics and Statistics
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June 2022

Author .....
Department of Economics
May 13, 2022

Certified by .....
Whitney K. Newey
Ford Professor of Economics
Thesis Supervisor

Certified by .....
Anna Mikusheva
Associate Professor of Economics
Thesis Supervisor

Certified by .....
Alberto Abadie
Professor of Economics
Thesis Supervisor

Accepted by .....
Abhijit V. Banerjee
Ford International Professor of Economics
Chairman, Department Committee on Graduate Theses

Essays in Econometrics

by

David W. Hughes

Submitted to the Department of Economics
on May 13, 2022, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Economics and Statistics

Abstract

This thesis consists of three chapters on the development and analysis of methods in econometrics. In the first two chapters I consider the use of jackknife bias correction techniques to deal with the incidental parameters bias that arises from including fixed effect parameters in nonlinear models. The final chapter deals with the properties of common linear instrumental variables methods in the presence of many endogenous regressors.

Chapter 1 considers estimation of a directed network model in which outcomes are driven by dyad-specific variables (such as measures of homophily) as well as unobserved agent-specific parameters that capture degree heterogeneity. I develop a jackknife bias correction to deal with the incidental parameters problem that arises from fixed effect estimation of the model. In contrast to previous proposals, the jackknife approach is easily adaptable to different models and allows for non-binary outcome variables. Additionally, since the jackknife estimates all parameters in the model, including fixed effects, it allows researchers to construct estimates of average effects and counterfactual outcomes. I also show how the jackknife can be used to bias-correct fixed effect averages over functions that depend on multiple nodes, e.g. triads or tetrads in the network. As an example, I implement specification tests for dependence across dyads, such as reciprocity or transitivity. Finally, I demonstrate the usefulness of the estimator in an application to a gravity model for import/export relationships across countries.

In Chapter 2, joint with Jinyong Hahn, I compare the properties of two bias correction methods, the leave-one-out jackknife and the split-sample jackknife, in a nonlinear panel data model with individual fixed effects. Since both estimators are asymptotically unbiased with equal asymptotic variances, we derive higher-order bias and variance expressions for both bias corrections, and show that the split-sample jackknife has larger higher-order variance. This difference in higher-order variances can be important in practice, particularly in settings where the time-series dimension T is not large. In addition, the remaining bias (after bias correction) is larger for the split-sample estimator. Simulations confirm these findings, and show significant distortions in coverage when the asymptotic distribution is used for inference on the split-sample jackknife estimator.

Chapter 3 considers the properties of linear IV estimators when used to estimate models in

which there are many potentially endogenous regressors. One common setting in which many endogenous regressors naturally arise, is the interaction of endogenous treatment variables with exogenous covariates in models that aim to capture heterogeneity in treatment effects. I extend existing results on linear IV estimation by considering asymptotics under which the number of endogenous regressors is allowed to grow with the sample size, and derive consistency and asymptotic normality results for the jackknife IV estimator (JIVE), as well as the heteroskedasticity robust k-class style estimators (including the HLIM and HFUL). In simulations, the HFUL estimator is shown to outperform others in models with both many endogenous regressors and many instruments.

JEL classification: C13, C23, C26

Thesis Supervisor: Whitney K. Newey
Title: Ford Professor of Economics

Thesis Supervisor: Anna Mikusheva
Title: Associate Professor of Economics

Thesis Supervisor: Alberto Abadie
Title: Professor of Economics

Acknowledgments

I would like to express my sincere gratitude to the many people who have contributed to both my academic and personal lives over the course of my Ph.D.

Whitney Newey has spent innumerable hours patiently listening to my ideas and answering my many questions, and our discussions have always left me with a renewed passion and interest for econometrics. Much of the research in this thesis builds upon his own work, and his insights have been extremely valuable. His constant support and enthusiasm for my research has enabled me to continue through many difficult periods in my studies, and his unerring confidence in my abilities has often sustained me when my own was lacking.

Anna Mikusheva has provided advice and insightful feedback over the years, and I have learnt a great deal from her. I am particularly grateful for the care and concern she has shown, and her support throughout the often stressful job market period. She has demonstrated to me the importance of not only being a great researcher, but a wonderful teacher and person as well.

I would also like to thank Alberto Abadie, who I have enjoyed many interesting and helpful discussions with, and who has always found time to offer advice, feedback and support. He has been instrumental in helping me work through the requirements of the joint degree in Economics and Statistics. Victor Chernozhukov and Isaiah Andrews have provided helpful feedback for my work, and Jerry Hausman, whose generous fellowship supported my research.

I have had the pleasure of spending time with many wonderful classmates who have dramatically improved my Ph.D experience, both academically and, more importantly, on a personal level. In particular, I have shared many drinks and some insightful (and not so insightful) conversations with Ben Deaner, Sylvia Klosin, Claire Lazar, Jeremy Majerovitz, Matthew Ridley, Parinitha Sastry, Sammy Young, and Sean Wang.

I would also like to thank Maya Bidanda, who has been a never ending source of support and encouragement. She has been there to celebrate each of my successes, and provided love and care at every setback. Finally, I thank my parents. They instilled in me a passion for curiosity and learning and have supported all of my endeavours. Despite being on the other side of the world, their unconditional love and support has sustained me throughout my studies. I could not have achieved any of this without them.

Contents

1 Estimating Nonlinear Network Data Models with Fixed Effects **15**

1.1 Introduction 15

1.2 Dyadic linking model and jackknife correction 19

1.2.1 Model 19

1.2.2 Empirical setting - gravity equation 22

1.2.3 Incidental parameters problem 23

1.2.4 Jackknife bias correction 25

1.2.5 A weighted jackknife 29

1.3 Estimating average effects 30

1.3.1 Averages over single observations 31

1.3.2 Specification testing 33

1.4 Asymptotic theory 37

1.4.1 Asymptotic analysis for the common parameters 37

1.4.2 Asymptotic analysis for fixed effect averages 40

1.4.3 Examples 45

1.5 Empirical example 46

1.5.1 Zero-inflated binomial model 47

1.5.2 Testing for strategic interactions 49

1.5.3 Comparison with conditional logit estimator 50

1.6 Simulations 52

1.7 Conclusion 57

2 The Higher-order Variance of Bias-corrected Panel Estimators **59**

2.1 Introduction 59

2.2 Nonlinear fixed effects models and jackknife bias correction 61

2.2.1 The fixed effects model 61

2.2.2 Bias corrections . . . 62
2.2.3 Assumptions . . . 64
2.3 Higher-order variances . . . 65
2.3.1 Understanding the effect of bias correction . . . 67
2.3.2 Accuracy in estimating the bias . . . 69
2.4 Higher-order bias . . . 70
2.5 Fixed effect averages . . . 71
2.6 Examples . . . 73
2.6.1 An analytical example . . . 73
2.6.2 Monte Carlo analysis . . . 75
2.7 Conclusion . . . 79

3 Estimation of Linear IV Models with Many Endogenous Regressors 81
3.1 Introduction . . . 81
3.2 The model . . . 84
3.3 Linear IV estimators . . . 86
3.4 Consistency . . . 88
3.5 Asymptotic normality and inference . . . 91
3.6 Simulations . . . 95
3.7 Conclusion . . . 97

Appendices 107
A Appendix for Chapter 1 . . . 107
A.1 Notation and norms . . . 107
A.2 Asymptotic expansions . . . 108
A.3 Jackknife results for β . . . 111
A.4 Jackknife results for average effects . . . 123
B Appendix for Chapter 2 . . . 137
B.1 Expansions and V-statistic forms . . . 137
B.2 Proof of Theorem 1 . . . 141
B.3 Proof of Theorem 2 . . . 145
B.4 Proof of Theorem 3 . . . 147
C Appendix for Chapter 3 . . . 150
C.1 Two useful lemmas . . . 150
C.2 Consistency of estimators . . . 151

C.3 Asymptotic normality 159
C.4 HLIM and HFUL parameter 165
C.5 Consistency of variance of jackknife estimator 167
C.6 Proof of $c'\hat{\Xi}_n c \to c'\Xi_n c$ 179
C.7 Consistency of variance of jackknife k-class estimator 187

List of Figures

1.1 Diagram of leave-out sets $\mathcal{I}_k$ for $k = 1, 2, 3$ 26
1.2 Transitive triangles 35
1.3 Weights $(\widehat{W}_{(k)})$ versus estimates $(\widehat{\beta}_{(k)})$ in leave-out samples 55

List of Tables

1.1 Estimated model coefficients 48
1.2 Estimated average effects 49
1.3 Strategic interaction tests 50
1.4 Coefficient estimates for logit model 51
1.5 Fixed effect distributions 52
1.6 Simulation results 53
1.7 Simulation results - median and range 54
1.8 Simulation results $N = 101$ 55
1.9 Comparison of jackknife corrections 57

2.1 Estimation of θ_0 76
2.2 Estimation of μ_0 78

3.1 Simulation results: $\lambda_{min} = 100$ 98
3.2 Simulation results: $\lambda_{min} = 30$ 99

Chapter 1

Estimating Nonlinear Network Data Models with Fixed Effects

1.1 Introduction

Networks are common in both economic and social contexts, and it is important to understand the factors that play a role in both the formation and strength of the links between agents. The econometric analysis of networks faces a number of challenges that have received much attention in recent literature (see de Paula (2020) and Graham (2020) for reviews of this literature). One common modeling approach is to assume a dyadic network structure (one in which decisions are made bilaterally between agents), but allow for linking decisions to depend on unobserved agent-specific heterogeneity. These models are common in practice since they are straightforward to implement while still being able to capture important aspects of observed networks. Controlling for agent-specific heterogeneity is important since in many real world networks agents vary significantly in the number and strength of connections made. Ignoring this heterogeneity can lead to large biases in estimated effects.

In this paper, we consider the estimation of dyadic models, where the presence of unobserved heterogeneity is accounted for by two sets of agent-specific fixed effects – a sender and a receiver effect. The fixed effects approach is appealing as it does not require strong assumptions about the unobserved component as in random effects models. In addition, the network setting does not suffer from the 'fixed T' issues of panel data, since we observe every agent interacting with $N-1$ other agents in the network, so that fixed

effect estimates are consistent. However, the large number of fixed effects (proportional to the square root of the sample size) does create an incidental parameter problem (Neyman and Scott, 1948). This paper proposes a jackknife approach to bias correction, which has a number of benefits over existing methods. Importantly, the jackknife is easily adaptable to a range of settings, including models for non-binary outcome variables. Additionally, since the jackknife estimates all of the parameters in the model, including the fixed effects, we are able to construct estimates of average effects and counterfactual outcomes. We also show how the jackknife can be used to bias correct averages over functions of multiple observations (e.g. dyads or triads in the network), which we show is useful for constructing specification test statistics, such as tests for the presence of certain strategic interactions like reciprocity or transitivity.

We demonstrate the consistency and asymptotic normality of the jackknife estimator under asymptotic sequences in which a single network grows in size, while the network remains 'dense'. The network model we consider is one in which agents make bilateral decisions about link-specific outcomes, independently of other relationships. This type of dyadic model (a *dyad* is a pair of agents) has received much attention in the literature, because of its tractability and its ability to replicate some key features of observed networks. In particular, it allows for: *homophily*, the tendency of agents to form stronger ties with other agents that are similar to them; and *degree heterogeneity*, where the number/strength of links in the networks can vary substantially across nodes.

In the case of a binary outcome variable, the model we consider is one of link formation, and is an extension of the model by Holland and Leinhardt (1981). There are several alternative approaches to address the incidental parameters problem in this setting. Graham (2017), Charbonneau (2017), and Jochmans (2018) all consider versions of this model in which the latent disturbances follow a logistic distribution, and use conditioning arguments to remove dependence on the fixed effects. The conditioning approach has the advantage of being applicable under certain *sparse network* asymptotic sequences, but is limited to models in which sufficient statistics for the fixed effects exist, and is not able to recover counterfactuals or average effects. Yan et al. (2019) also studies the logistic model and provides asymptotic results for the incidental parameters. Graham (2017) considers an analytical correction for the logistic model, while Dzemski (2019) derives the analytical correction for a probit model. The analytical bias correction approach is limited to *dense network* sequences, as in this paper, and similarly to this paper can recover average effects. The advantage of the jackknife correction relative to an analytical approach is that it provides

an off-the-shelf approach that researchers may apply to new settings, without the need to first derive bias expressions. Candelaria (2020), Toth (2017), and Gao (2020) study identification of the common parameters without a known parametric form for the disturbance term, while Zeleneev (2020) allows for nonparametric structure in the unobserved heterogeneity term.

Although the focus of the literature on dyadic network models has been on the binary outcome case, researchers often have access to outcome variables that are non-binary. Examples of these settings include the value of exports between countries, the value of loans between banks, or the number of workers migrating between states. The results in this paper are derived for a general M-estimator satisfying basic regularity conditions and so cover a range of models for both binary and non-binary outcome variables, as well as a range of estimation approaches, including MLE, quasi-MLE and nonlinear least squares estimation.

As a demonstration of the technique in an empirical setting, we estimate a model of international trade relationships. Gravity models have been a workhorse model in the trade literature for many years, and the importance of including country-specific fixed effects is well known (Anderson and Van Wincoop, 2003). We estimate the zero-inflated negative binomial model of Burger et al. (2009), which combines both a model for the decision of countries to engage in trade, as well as a model for the *value* of exports conditional on some trade occurring. The model addresses to key issues in the gravity model literature: it allows for a large proportion of zero trade flows in the network, and it captures the observed high dispersion of export values across countries. We obtain bias-corrected estimates of both the model parameters, as well as average effects.

The jackknife bias correction also allows for the construction of various specification tests. Many models of network formation include strategic aspects in which agents' decisions are influenced by the state of the network. For instance, agent i may find it more beneficial to link with j if they already share many other links in common. Graham and Pelican (2020) derives the locally best similar test for a class of alternatives in a logit model, using conditioning arguments. Dzemski (2019) tests for the presence of transitive links with triads (groups of three agents) in a probit model, and derives an analytical bias correction for the statistic. We demonstrate that a range of test statistics, including that of Dzemski (2019), can be bias-corrected using the jackknife. This extends the set of tests available to researchers, as well as the range of models they can be applied to. As an example, we test for reciprocity and transitivity in trade links between countries and find evidence that the decisions of countries to engage in trade are reciprocal (if country i exports to country j

then it is likely that j also exports to i), but do not find evidence of transitive relationships.

The network jackknife extends previous results on jackknife bias correction in panel data. Hahn and Newey (2004) introduced a jackknife correction for panel estimators with individual fixed effects, based on re-estimating the parameters on data sets that exclude a single time period. Dhaene and Jochmans (2015) present a split-sample version of this idea based on estimating the model in the first and second halves of time periods separately. Fernández-Val and Weidner (2016) develop a general framework that allows for both time and individual fixed effects. The analysis in this paper builds heavily off of the asymptotic expansions in Fernández-Val and Weidner (2016).

Analogously to the panel data setting, the network jackknife is constructed by forming 'leave-out' estimates that exclude certain subsets of the data. Cruz-Gonzalez et al. (2017) and Chen et al. (2021) have suggested jackknife approaches for network data, although without formal proof, based on either a split-sample approach or a leave-one-out approach that drops a single agent at a time. We propose a different approach to jackknifing network data that is based on a novel partitioning of the data set that constructs leave-out estimates that remove bias from both sender and receiver of fixed effects in one step. We extend the asymptotic expansions of Fernández-Val and Weidner (2016) to allow for formal analysis of the jackknife estimator. The jackknife proposed here drops a single observation per fixed effect at each step, so that our approach is likely to have better finite sample variance properties than the split-sample approach (see Hughes and Hahn (2020) for a formal argument in the panel setting). In contrast to a jackknife that drops all observations from a single agent in the network, our jackknife retains all agents in each leave-out estimation, so that the distribution of unobserved effects is held constant. We demonstrate the small-sample effectiveness of our approach in comparison to previous suggestions in simulations that show that our jackknife is more robust to settings with meaningful levels of unobserved heterogeneity and in networks with lower density.

In addition, we introduce a weighted jackknife, that differs from standard implementations of the jackknife approach by taking a weighted-average of the leave-out estimates. This version puts less weight on noisier leave-out estimates, which improves the finite-sample properties of the jackknife in sparser settings. The weighted jackknife idea may be useful elsewhere, for example in binary-outcome panel data models with few successes for some individuals (so called 'rare events'). Finally, we also introduce a 'leave-l-out' version of the jackknife. This version requires only $(N-1)/l$ additional estimations of the model, and may allow researchers to reduce the computational burden in settings where model estimation is

difficult.

The rest of the paper is organized as follows. Section 2 introduces the network model and discusses implementation of the jackknife procedure for the estimation of model parameters, while Section 3 discusses estimation of average effects, and the construction of specification tests. Section 4 provides asymptotic results for the estimators, and discusses the main assumptions under which they hold. In Section 5 we demonstrate the method by estimating a model of international trade flows, while Section 6 reports simulation results that are consistent with the jackknife theory.

1.2 Dyadic linking model and jackknife correction

1.2.1 Model

The researcher observes a network of N agents; these agents may, for example, be individuals, firms, or countries. For each potential *directed* connection, $i \to j$, we observe an associated link-specific outcome variable Y_{ij}. The variable Y_{ij} may capture the presence (or absence) of a link between two agents, in which case Y_{ij} is binary, or may represent a measure of the strength of the link between agents. For example, Y_{ij} may be the value of exports from country i to country j in a particular year, or the number of times agents i and j interacted in some period. Links are directed, meaning that $Y_{ij} \neq Y_{ji}$ in general, and so, following the literature, we term i the 'sender' and j the 'receiver' in link Y_{ij}.

The researcher also observes a set of link-specific covariates X_{ij}. The covariates capture characteristics of the relationship between agents that may impact the linking outcome. Often these will be interpreted as measures of *homophily*, that is, the tendency for agents to link with other agents that are similar to themselves. For example, countries may engage in greater levels of trade if they share a common language, or are geographically close.

Agents are endowed with two fixed effects, α_i and γ_i, which capture unobserved *degree heterogeneity*, that is, the tendency of some agents to form more (or stronger) links than others. The 'sender' fixed effect α_i accounts for heterogeneity in out-degree (the number or strength of links from agent i to other agents), while the 'receiver' fixed effect accounts for in-degree heterogeneity. Degree heterogeneity is an important feature of many networks, for example, we would expect countries with larger GDPs to engage in more trade than smaller countries *ceteris paribus* (see Anderson and Van Wincoop (2003) for an example of such a model). Since the network considered here is a directed one, we allow for the

sender and receiver fixed effects to differ; some countries may have structural tastes for importing goods over exporting, that is, they run trade deficits (or vice versa), so that $\alpha_i < \gamma_i$. Failure to account for degree heterogeneity in a network can lead to incorrect conclusions about the strength of homophily in a network. For example, observing that the United States imports more from China than from Canada may lead to the conclusion that distance between countries is unimportant for trade if we do not account for a China export effect. Graham (2017) provides some further intuition for why failing to account for degree heterogeneity can bias conclusions about homophily in a network.

We make the assumption that linking decisions are bilateral in nature, so that

$$Y_{ij} \perp\!\!\!\perp Y_{kl} | X, \beta, \alpha, \gamma \qquad \forall (k,l) \notin \{(i,j),(j,i)\}, \tag{1.1}$$

where $\perp\!\!\!\perp$ denotes independence of the outcomes conditional on observed covariates and fixed effects. This assumption does allow for dependence between the two links within a pair of agents (a dyad), but implies the decision between i and j is independent of that between i and k for instance. Importantly, this independence is conditional on the covariates and agent-specific fixed effects. Unconditionally, country i's exports to country j are correlated with their exports to country k, since both are determined by the exporter effect α_i. In many settings, the inclusion of fixed effects will be important in establishing the plausibility of (1.1). Assumption (1.1) may not be appropriate in situations where linking decisions are strategic. Estimation of models with strategic interactions is substantially more challenging, and is likely to require multiple observations of the network over time. Nonetheless, the dyadic model presented here still represents an important baseline model, and can be used to construct tests for the presence of strategic interactions against the null hypothesis of (1.1). We discuss examples of such tests in Section 1.3.2.

We leave the specific model for the network outcomes unspecified, and assume only that the model parameters $(\beta_0, \alpha_0, \gamma_0)$ are solutions to the population maximization problem

$$\max_{(\beta,\alpha,\gamma)\in\mathbb{R}^{\dim\beta+2N}} \bar{E}\big[\boldsymbol{\mathcal{L}}_N(\beta,\alpha,\gamma)\big],$$

$$\boldsymbol{\mathcal{L}}_N(\beta,\alpha,\gamma) = \frac{1}{N-1}\sum_i \sum_{j\neq i} \ell(Y_{ij}, X_{ij}, \beta, \alpha_i + \gamma_j) - \frac{b}{2N}\big(\sum_i \alpha_i - \sum_i \gamma_i\big)^2, \tag{1.2}$$

where $\bar{E}$ represents expectation conditional on the exogenous covariates and fixed effects, and ℓ is a known function that is maximized in expectation at the true parameters. Many

models can be estimated by maximizing objective functions of the form in (1.2), including MLE, quasi-MLE, and nonlinear least squares estimators. The researcher need only specify the objective function $\mathcal{L}$ and does not need to specify the distribution of the fixed effects, or how they relate to the covariates.

We assume that the unobserved effects enter in an additively separable manner, i.e. as $\alpha_i + \gamma_j$, identification of the two sets of fixed effects parameters requires a normalization, for which we choose

$$\sum_i \alpha_i = \sum_i \gamma_i.$$

The term $\frac{b}{2N}\big(\sum_i \alpha_i - \sum_i \gamma_i\big)^2$ is a penalty term intended to impose this normalization on the fixed effect parameters, where $b > 0$ as an arbitrary constant.[1] Note that the vectors of fixed effects α and γ are dependent on the network size N, although we leave this dependence implicit in the notation. When functions are evaluated at the true parameters $(\beta_0, \alpha_0, \gamma_0)$ we will typically drop them from notation. We will also use the shorthand $\ell_{ij} = \ell(Y_{ij}, X_{ij}, \beta_0, \alpha_{0,i} + \gamma_{0,j})$ when convenient.

Example. ***Maximum likelihood***

If we specify the full conditional distribution of the outcome variable as

$$Y_{ij}|X, \beta, \alpha, \gamma \sim f(Y_{ij}|X_{ij}, \beta, \alpha_i + \gamma_j),$$

then ℓ will be the log-likelihood function

$$\ell(Y_{ij}, X_{ij}, \beta, \alpha_i + \gamma_j) = \log f(Y_{ij}|X_{ij}, \beta, \alpha_i + \gamma_j).$$

Note that $\mathcal{L}$ need not be a true log-likelihood, since it may be that the observations Y_{ij} and Y_{ji} are conditionally dependent, in which case $\mathcal{L}$ is a pseudo log-likelihood.

As an example, consider a model of directed link formation according to

$$Y_{ij} = \mathbb{1}\{X_{ij}'\beta + \alpha_i + \gamma_j - \varepsilon_{ij} \geq 0\},$$

where ε_{ij} follows a known distribution F. This linking rule is compatible with a model in which utility is transferable across linked agents as in Bloch and Jackson (2007). Given the distributional assumption for ε_{ij}, the probability of a link forming, conditional on covariates

[1] In practice the constraint could simply be eliminated by substituting it into the objective. We follow Fernández-Val and Weidner (2016) in choosing this normalization as it is convenient in the proofs.

and fixed effects is $p_{ij} = F(X_{ij}'\beta + \alpha_i + \gamma_j)$, and the log-likelihood is $\ell_{ij} = Y_{ij} \log p_{ij} + (1 - Y_{ij}) \log(1 - p_{ij})$. This is an extension of the linking model of Holland and Leinhardt (1981) and has been used extensively in empirical literature.

Example. ***Nonlinear least squares***

The researcher may choose to specify only the conditional mean function for the outcome, rather than its full distribution, e.g $E\big[Y_{ij}|X, \beta, \alpha, \gamma\big] = h(X_{ij}, \beta, \alpha_i + \gamma_j)$. In this case, we may estimate the parameters of the model by setting

$$\ell(Y_{ij}, X_{ij}, \beta, \alpha_i + \gamma_j) = -\big(Y_{ij} - h(X_{ij}, \beta, \alpha_i + \gamma_j)\big)^2.$$

1.2.2 Empirical setting - gravity equation

Although much of the focus of the incidental parameters bias-correction literature has been on the binary outcome case, researchers often have access to non-binary outcome variables and will be interested in modeling networks in which the links are weighted. As a working example throughout the paper, we will consider a model of country-level trade relationships using a data set consisting of a directed network of export volumes between 136 countries (136×135 country pair observations) in 1990. The data are taken from Santos Silva and Tenreyro (2006), and additional details on their construction can be found in their paper. The outcome variable is the value of exports from country i to country j. We also use several covariates to capture homophily in trade relationships, which include: *log distance*, the log of the distance between the capitals of the countries*;* *border*, an indicator of whether the countries share a common border; *language*, an indicator for whether the countries share a language; *colonial*, and indicator for whether either country had colonized the other at some point in history; and *trade agreement*, an indicator for the presence of a joint preferential trade agreement between the two countries.

The Anderson and Van Wincoop (2003) gravity equation expresses the trade volume from country i to country j as

$$Y_{ij} = \alpha_0 G_i G_j d_{ij}^{\beta} e^{\alpha_i + \gamma_j} \tag{1.3}$$

where Y_{ij} is the trade flow from country i to country j, G_i is GDP of country i, and d_{ij} is inversely proportional to the distance between the two countries (which is generally taken to include all factors that create resistance to trade). The inclusion of exporter and importer fixed effects (α_i and γ_i) is intended to control for multilateral resistance terms, which may bias results if excluded.

A simple method for estimating the parameters in (1.3) is to first log-linearize the model. Unfortunately, this raises the issue of how to deal with the presence of zero outcomes that are common in trade data. In the country-level trade data introduced above, just under half of all country pairs engage in no trade. A number of solutions to this problem have been suggested. Several authors use Tobit models or two-step Heckman style models, which combine a binary selection equation (predicting whether or not any trade occurs) and a separate equation for the value of trade (conditional on selection); see for example Helpman et al. (2008), Rose (2004), Linders and de Groot (2006).

Another popular approach, suggested by Santos Silva and Tenreyro (2006), is to use a Poisson pseudo-maximum-likelihood estimator, which provides a natural way to incorporate zero-valued outcomes, as well as being robust to heteroskedasticity issues that can arise when log-linearizing multiplicative equations. However, there are two key drawbacks to modeling trade flows with a Poisson distribution. Firstly, the proportion of zeroes observed in typical trade data is much larger than that predicted by a Poisson model. Secondly, since the variance of a Poisson is restricted to be equal to its mean, outcomes are typically much more dispersed than would be expected under the Poisson model. In order to address these two issues, Burger et al. (2009) propose a zero-inflated negative binomial model, in which the value of trade between i and j is given by the product of two variables, $Y_{ij} = z_{ij}Y^*_{ij}$, where $z_{ij} \in \{0, 1\}$ is a binary decision to enter into a trading relationship, while Y^*_{ij} is the value of exports that will be realized, conditional on $z_{ij} = 1$. The binary decision is modeled using as a probit function, while the latent outcome Y^*_{ij} is modeled as a negative binomial variable, which allows for overdispersion in the model for Y^*_{ij}, that is, it allows the variance to differ from the mean. Since the distribution of Y_{ij} is parametrically specified, we may estimate the model using maximum likelihood, so that this model is captured by the framework in (1.2).

1.2.3 Incidental parameters problem

In total, the model contains $\dim(\beta) + 2N$ parameters to be estimated, from the $N(N-1)$ observations (Y_{ij}, X'_{ij}), which we will typically refer to using the shorthand (i, j). As is well known, nonlinear estimators with fixed effects suffer from an incidental parameter problem (Neyman and Scott, 1948). To describe the problem, consider the maximization problem (1.2) after first concentrating out the fixed effect parameters

$$\widehat{\alpha}(\beta), \widehat{\gamma}(\beta) = \arg\max_{\alpha,\gamma} \mathcal{L}_N\big(\beta, \alpha, \gamma\big),$$

$$\widehat{\beta} = \arg\max_{\beta} \mathcal{L}_N\big(\beta, \widehat{\alpha}(\beta), \widehat{\gamma}(\beta)\big). \tag{1.4}$$

Replacing the population functions $\alpha(\beta), \gamma(\beta) = \arg\max_{\alpha,\gamma} \bar{E}\big[\mathcal{L}_N(\beta, \alpha, \gamma)\big]$ with their sample values, results in an objective function that is biased, in the sense that

$$\beta_0 \neq \beta_N = \arg\max_{\beta} \bar{E}\big[\mathcal{L}_N\big(\beta, \widehat{\alpha}(\beta), \widehat{\gamma}(\beta)\big)\big]. \tag{1.5}$$

To see why, observe that the first-order condition for $\widehat{\alpha}_i(\beta)$ depends only on the $N-1$ observations (i,j) for $j \neq i$. Similarly, the first-order condition for $\widehat{\gamma}_i(\beta)$ depends on the $N-1$ observations (j,i) for $j \neq i$. Expanding $\widehat{\alpha}_i(\beta)$ around $\alpha_i(\beta)$ (and similarly for γ) will therefore result in a bias of order $O(N^{-1})$. Under regularity conditions discussed in Section 1.4, we show that the bias of the maximizer of the profile objective function (1.5) is approximately given by

$$\beta_N - \beta_0 \approx \frac{B_N}{N-1} \tag{1.6}$$

for some bias term B_N. Analogously to the panel literature, the bias is inversely proportional to the number of observations used to estimate each of the fixed effect parameters; the exporter effect for country i is estimated using the data on the $N-1$ other countries in the network that i may export to. As the size of the network grows, $N \to \infty$, we will have that $\beta_N \to \beta_0$ so that parameter estimates are consistent. Considering $\widehat{\beta}$ as an estimator for β_N, we can show that $N(\widehat{\beta} - \beta_N) \Rightarrow \mathcal{N}(0, V)$. However, since the bias $\beta_N - \beta_0$ is of the same order as the estimation error, $O(N^{-1})$, $\widehat{\beta}$ will be *asymptotically biased*, that is

$$N(\widehat{\beta} - \beta_0) \Rightarrow \mathcal{N}\big(B, V\big).$$

The incidental parameters generate an *asymptotic bias* in the network model analogous to the panel setting with both N and T growing to infinity at the same rate. Similar asymptotic expansion arguments have been used in the panel data literature on nonlinear models with fixed effects. Hahn and Newey (2004) derive expansions for models with individual fixed effects, while Fernández-Val and Weidner (2016) derive expansions that apply to general models with additively separable unobserved effects, and Chen et al. (2021) consider the setting with interactive effects. The expansions used in this paper rely heavily on these prior results. Dzemski (2019) also applies the Fernández-Val and Weidner (2016) expansions to the network model structure to derive bias expressions for a probit model. In this paper, we extend the asymptotic expansions to higher order so that they may be used to justify

jackknife bias correction procedures. We demonstrate that, under mild additional regularity conditions, the jackknife estimator introduced below is asymptotically normal and mean zero, so that valid inference can be performed on model parameters.

1.2.4 Jackknife bias correction

The jackknife bias-corrected estimator is constructed as a linear combination of the full-sample parameter estimates and an average of 'leave-out' estimators that exclude certain observations in the data set. The particular linear combination chosen can be motivated by asymptotic expansions of the estimator, in particular the form of the bias in (1.6).

Suppose we were to drop observations from our data set in such a way that for every country i we exclude one observation in which i is the exporter, and one observation in which i is the importer (recall that a single observation is one export-import relationship, of which we observe $N(N-1)$ in total). We show that the new estimator using this 'leave-out' sample, $\tilde{\beta}$, has a bias that is approximately

$$\bar{E}[\tilde{\beta}] - \beta \approx \frac{B_N}{N-2}.$$

The form of the bias for $\tilde{\beta}$ can be explained by two important factors: (i) $\tilde{\beta}$ is estimated using only $N-2$ observations per fixed effect, since we excluded one observation related to each fixed effect parameter, and (ii) $\tilde{\beta}$ is estimated on a random sample generated from the same set of fixed effects, so that the bias expression B_N is the same as that in (1.6).

Taking advantage of the fact that the estimate $\tilde{\beta}$ has a larger bias than the full-sample estimate $\widehat{\beta}$ by the factor $\frac{N-1}{N-2}$, we can construct a new estimator $\widehat{\beta}_{jack} = (N-1)\widehat{\beta} - (N-2)\tilde{\beta}$ which has no asymptotic bias

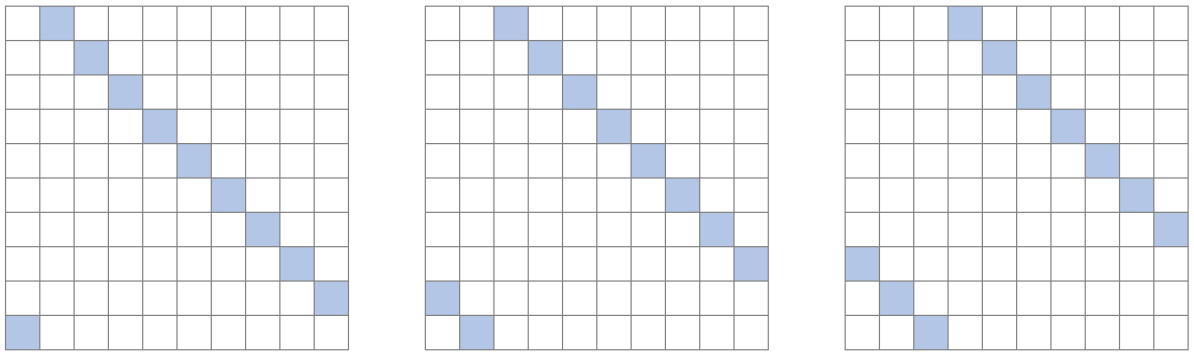
$$\bar{E}\big[(N-1)\widehat{\beta} - (N-2)\tilde{\beta}\big] - \beta \approx 0.$$

To describe the construction of the leave-out estimators, we first define a partition of the $N(N-1)$ observations of directed pairs (i,j) into $N-1$ sets of the form[2]

$$\mathcal{I}_k = \big\{(i,j) : j = (i+k) \pmod N\big\}, \quad \text{for } k = 1, \ldots, N-1$$

[2]In the modulo notation we consider agent N equivalent to agent 0.

Figure 1.1: Diagram of leave-out sets $\mathcal{I}_k$ for $k = 1, 2, 3$



Observation (i, j) in the network is represented by the corresponding position in each matrix. The blue squares are the observations contained in the leave-out sets $\mathcal{I}_k$.

that is, the set of directed pairs $\{(1, 1+k), \ldots, (N-k, N), (N-k+1, 1), \ldots, (N, k)\}$.[3]

Figure 1.1 represents the structure of the first three sets $\mathcal{I}_k$ for a network of $N = 10$ agents. Observations are ordered in an $N \times N$ matrix so that the (i, j) cell represents the corresponding observation in the network (the diagonal elements are empty since there are no (i, i) observations). The leave-out sets take diagonal sections from the data matrix. Importantly, constructing the sets this way ensures that each contains exactly one observation related to each sender and receiver fixed effect; i.e. there is one observation taken from every row and every column.

Let $1_{ij}^k = \mathbb{1}\{(i, j) \notin \mathcal{I}_k\}$, be an indicator variable that is equal to one whenever the observation (i, j) is *not* in the k-th leave-out set. The k-th leave-out estimates are

$$(\widehat{\beta}_{(k)}, \widehat{\phi}_{(k)}) = \arg \max_{(\beta,\phi)\in\Theta} \frac{1}{N-2} \sum_i \sum_{j \neq i} \ell_{ij}(\beta, \phi) \times \mathbb{1}_{ij}^k, \tag{1.7}$$
$$\text{subject to } \sum_i \alpha_i = \sum_i \gamma_i,$$

that is, the estimates obtained by excluding the observations in $\mathcal{I}_k$ from the data. We can

[3]The construction of the leave-out sets assumes that the labelling of the nodes is arbitrary. This will generally be true, but the researcher may ensure this by randomizing the ordering of nodes prior to estimation.

then construct the jackknife bias-corrected estimator

$$\widehat{\beta}_J = (N-1)\widehat{\beta}_N - (N-2)\frac{1}{N-1}\sum_{k=1}^{N-1}\widehat{\beta}_{(k)}. \tag{1.8}$$

The construction of the leave-out estimators is analogous to jackknife bias correction in the panel data setting; however, the structure of the jackknife proposed here is new. The procedure relies on dropping sets of observations that contain a single observation related to every sender fixed effect α_i as well as every receiver fixed effect γ_i. In this way, the bias from both types of fixed effects can be addressed simultaneously, while holding the distribution of fixed effects constant across the leave-out samples. This is in contrast to an approach which drops all observations from a single agent, which removes that agents' fixed effects from the leave-out sample and alters the distribution of unobserved heterogeneity. We show in simulations that the method proposed here is more robust to networks that have more unobserved heterogeneity or are less dense.

We prove in Section 1.4 that the jackknife bias correction is consistent, and asymptotically normal, with mean zero and variance equal to that of the full-sample estimate

$$N(\widehat{\beta}_J - \beta_0) \Rightarrow \mathcal{N}\big(0, V\big).$$

The fact that the jackknife is able to remove bias without affecting the asymptotic variance of the estimator may seem surprising. This important feature is achieved by averaging across the $N-1$ different leave-out estimators $\widehat{\beta}_{(k)}$. Since the sets $\mathcal{I}_1, \ldots, \mathcal{I}_{N-1}$ form a partition of the $N(N-1)$ observations in the network, each observation is excluded from exactly one of the leave-out estimates. This balanced treatment of observations ensures that the jackknife procedure does not affect the first-order asymptotic approximation of the estimator.[4]

Remark 1. The construction of the leave-out sets depends on the labelling of nodes, and so the final estimator will be dependent on this labelling (since the make-up of the leave-out sets will change). While the researcher could re-randomize the node labels, construct $\widehat{\beta}_J$ for each randomization, and then average to remove some of the arbitrariness of the node labels, this is not necessary. The estimators with different labelings should be very similar so that

[4]In the panel data setting, Dhaene and Jochmans (2015) note that forming a jackknife using overlapping subpanels (across time) results in an inflation of the asymptotic variance, since some time periods are used more that others. That the $\mathcal{I}_k$ form a partition ensures that each observation is used an equal number of times in the average $\frac{1}{N-1}\sum_{k=1}^{N-1}\widehat{\beta}_{(k)}$.

the additional computations will have little effect.[5]

The jackknife estimator requires N estimations of the model, and so may be computationally intensive for large networks, although speed may be improved by computing the leave-out estimates in parallel, and using good starting values such as the full sample estimates. As an alternative, we present a 'leave-l-out' version of the jackknife, which reduces the number of additional estimations of the model by dropping l observations per fixed effect, as opposed to just one. To describe the estimator, let $N_l = \frac{N-1}{l}$ (we assume here that $N-1$ is divisible by l for simplicity). We can construct the k-th leave-l-out set by combining l of the leave-one-out sets as follows $\mathcal{I}_k^l = \cup_{j=0}^{l-1} \mathcal{I}_{k+jN_l}$, for $k = 1, \ldots, N_l$. This results in $N_l = \frac{N-1}{l}$ non-overlapping leave-l-out sets, with corresponding estimates $\hat{\beta}_{l,(k)}$, which are the estimates from using all observations except those in the k-th leave-l-out set $\mathcal{I}_k^l$. A jackknife bias-corrected estimate can then be constructed as

$$\hat{\beta}_{J,l} = \frac{N-1}{l}\hat{\beta}_N - \frac{N-1-l}{l}\frac{1}{N_l}\sum_{k=1}^{N_l}\hat{\beta}_{l,(k)}, \tag{1.9}$$

where $N_l = \frac{N-1}{l}$.[6]

Remark 2. The leave-l-out jackknife bias correction has the same asymptotic variance as the standard leave-one-out jackknife and the full-sample estimator. However, there may be some finite-sample efficiency loss, particularly when l is large or when the network is not sufficiently dense. Hughes and Hahn (2020) show in the panel case that the leave-one-out jackknife is higher-order more efficient than the split-sample jackknife (i.e. its variance to $O(N^{-1})$ is smaller), and it is likely that the same result applies here, although this is beyond the scope of the present paper.

[5] In simulations (see Section 1.6 for the design) the estimates based on different labelings were close to identical and so the additional calculations had almost no effect on the estimation.

[6] In the case that $N-1$ is not divisible by l, let $N_l = \lfloor \frac{N-1}{l} \rfloor$. Then we may partition the data into $r = N-1-lN_l$ leave-$(l+1)$-out sets and $N_l - r$ leave-l-out sets. Denote $\hat{\beta}_{(k)}$ as the leave-$(l+1)$-out estimates for $k = 1, \ldots, r$ and the leave-l-out estimates for $k = r+1, \ldots, N_l$. Then, let

$$\bar{\beta} = \frac{1}{(N-1)(N_l-1)}\Big(\sum_{k=1}^{r}(N-l-2)\hat{\beta}_{(k)} + \sum_{k=r+1}^{N_l}(N-l-1)\hat{\beta}_{(k)}\Big)$$

The jackknife bias-corrected estimator is then given by

$$\hat{\beta}_J = N_l\hat{\beta}_N - (N_l-1)\bar{\beta}$$

1.2.5 A weighted jackknife

The jackknife relies on large dense network asymptotics, but in finite samples it is possible for some leave-out estimates to drop a number of important observations all at once. This is more likely to occur when N is small, there are few links for some nodes, or when we are using the leave-l-out jackknife with large l.

The performance of the jackknife can be improved in these settings by taking a weighted average of the estimates $\widehat{\beta}_{(k)}$. Define the weights

$$\widehat{W}_{(k)} = -\frac{1}{N}\Big(\partial_{\beta\beta'}\widehat{\mathcal{L}}_{(k)} - \frac{1}{N}(\partial_{\phi\beta'}\widehat{\mathcal{L}}_{(k)})(\partial_{\phi\phi'}\widehat{\mathcal{L}}_{(k)})^{-1}(\partial_{\beta\phi}\widehat{\mathcal{L}}_{(k)})\Big).$$

The weighted-jackknife estimator is given by

$$\widehat{\beta}_{wJ} = (N-1)\widehat{\beta}_N - (N-2)\bar{W}_J^{-1}\Big(\frac{1}{N-1}\sum_{k=1}^{N-1}\widehat{W}_{(k)}\widehat{\beta}_{(k)}\Big), \tag{1.10}$$

where $\bar{W}_J = \frac{1}{N-1}\sum_{k=1}^{N-1}\widehat{W}_{(k)}$.

The weights $\widehat{W}_{(k)}$ are the Hessian for β, after concentrating out the fixed effects. In the special case that $\mathcal{L}$ is a log-likelihood function, W_N is the Fisher information for β, and so is equal to the inverse of the asymptotic variance. In this case, we are using an inverse variance weighting scheme, which down-weights leave-out samples that produce particularly noisy estimates of the common parameters. The weighting scheme is equally applicable to non-likelihood settings, although it no longer carries the inverse variance interpretation. In simulations, this weighted version of the jackknife significantly improves the performance of the estimator in sparser networks (see Section 1.6 for more details).

Remark 3. Asymptotically the weights have no effect on the estimator, since all $\widehat{W}_{(k)}$ converge to the same quantity. This implies that the asymptotic variance of the weighted jackknife is the same as that of the standard jackknife. In finite samples, variation in the weights depends on the number of nodes N, as well as the density of the network (i.e. the variation in outcomes for each node). The weighting scheme is likely to have a large impact for small or less dense networks, but in denser (or larger) networks we will have $\bar{W}_J^{-1}\widehat{W}_{(k)} \approx I_{\dim\beta}$, so that the weighted and unweighted jackknife estimates are very similar.

Remark 4. The motivation behind the particular choice of weights comes from the first-order asymptotic expansion of the estimator. A Taylor expansion of the first-order conditions

of the objective function gives an expression of the form

$$W_N(\widehat{\beta} - \beta) \approx A + B,$$

where A is mean zero and asymptotically normal with variance $\bar{\Omega}$ (as in Theorem 1.1), while B is an additional term responsible for the asymptotic bias of the estimator.

As demonstrated in Lemmas A.3 and A.4 in the Appendix, the jackknife procedure applied to $A + B$ successfully demeans B and leaves A unchanged, i.e.

$$(N-1)A - (N-2)\frac{1}{N-1}\sum_{k=1}^{N-1} A_{(k)} = A,$$
$$\bar{E}\Big[(N-1)B - (N-2)\frac{1}{N-1}\sum_{k=1}^{N-1} B_{(k)}\Big] = 0.$$

This results in an estimator with no asymptotic bias and unchanged asymptotic variance. In practice however, the jackknife is applied to $\widehat{\beta}$, so we must consider the effect of the jackknife procedure on $W_N^{-1}(A + B)$.

The validity of the jackknife relies on the fact that $W_{(k)} \approx W_N$, which is guaranteed in large samples under our assumptions, in particular, dense network asymptotics. However, in finite samples, $W_{(k)}$ could vary substantially in some leave-out samples when N is small, there are few links for some nodes, or we are using the leave-l-out jackknife with large l. This motivates instead averaging over $\widehat{W}_{(k)}\widehat{\beta}_{(k)}$, so that variation in $W_{(k)}^{-1}$ has less impact on the quality of the asymptotic approximation. Intuitively, we are jackknifing the first-order condition for $\widehat{\beta}$, rather than $\widehat{\beta}$ itself.[7]

1.3 Estimating average effects

In addition to estimation of the common parameter β, researchers may also be interested in estimating certain averages over the distribution of exogenous regressors and fixed effects. An important advantage of the jackknife bias correction, over methods based on conditioning on sufficient statistics (e.g. Graham (2017), Jochmans (2018)), is that by estimating the fixed

[7]Of course, W^{-2}, W^{-3} and similar terms also appear in high-order parts of the expansion. The validity of the large-network approximation $W_{(k)}^{-1} \approx \bar{W}^{-1}$ is still necessary for the jackknife to be consistent and asymptotically normal. The weighting scheme simply aims to improve the finite-sample properties of the estimator.

effect parameters we are able to construct estimators for these averages. Common examples include average and marginal effects, as well as counterfactual outcomes. In the network setting, these are averages over functions of a single potential link (i, j) in the network. We will additionally show that averages over multiple links also provide interesting objects of interest; for example, averages over dyads $\{(i, j), (j, i)\}$, triads (groups of three nodes), or other network patterns. As an example, we focus on how these objects can be used to construct tests of the assumption of independent link formation stated in (1.1), but they may have wider relevance in empirical work. Estimation of the many fixed effect parameters means that these averages also suffer from an incidental parameter problem. We show that the jackknife can be used to bias-correct average effects estimates and obtain correct inference.

1.3.1 Averages over single observations

A simple fixed effect average may be expressed as

$$\begin{aligned} \delta &= E\big[\Delta_N(\beta_0, \phi_0)\big], \\ \Delta_N &= \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \neq i} m(X_{ij}, \beta, \alpha_i + \gamma_j), \end{aligned} \tag{1.11}$$

where the expectation is taken over the joint distribution of covariates X_{ij} and fixed effects (α_i, γ_j), and the function m represents the effect of interest. Here we specify two possible parameters of interest, δ the population average, and Δ_N, the sample average; this choice will affect the asymptotic distribution of the estimator, a point we return to in Section 1.4. As earlier, we will impose that the fixed effects enter the function m in an additively separable way, as $\pi_{ij} = \alpha_i + \gamma_j$; this will imply that the choice of fixed effect normalization will not affect the estimator.

Example. ***Marginal effect***

As an example, consider a binary outcome model with $P(Y_{ij} = 1|X, \beta, \alpha, \gamma) = F(\beta X_{ij} + \alpha_i + \gamma_j)$. We may be interested in estimating the average partial effect of the covariate, in which case we would have

$$m(X_{ij}, \beta, \alpha_i + \gamma_j) = \beta \frac{\partial}{\partial X} F(\beta X_{ij} + \alpha_i + \gamma_j).$$

Alternatively, we may be interested in the average partial effect at some fixed value of

$X_{ij} = x$, in which case, $m(X_{ij}, \beta, \alpha_i + \gamma_j) = \beta \frac{\partial}{\partial X} F(\beta x + \alpha_i + \gamma_j)$.

Example. ***Counterfactual change***

Alternatively, assume that we estimate the conditional mean function $E[Y_{ij}|X_{ij}, \beta, \phi] = h(\beta X_{ij} + \alpha_i + \gamma_j)$. We may be interested in the counterfactual change in predicted outcome from a change in the value of the covariate X_{ij} from x_0 to x_1, e.g. the effect of entering or exiting a trade agreement. In this case

$$m(X_{ij}, \beta, \alpha_i, \gamma_j) = h(\beta x_1 + \alpha_i + \gamma_j) - h(\beta x_0 + \alpha_i + \gamma_j).$$

The average effect in (1.11) can be estimated by plugging in estimates of the model parameters

$$\widehat{\Delta}_N = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \neq i} m(X_{ij}, \widehat{\beta}, \widehat{\alpha}_i, \widehat{\gamma}_j).$$

As with estimates of the parameters themselves, the average effect estimate $\widehat{\Delta}_N$ is asymptotically biased, that is, $N(\widehat{\Delta}_N - \Delta_N)$ converges to a normal distribution that is not centered at zero. The asymptotic bias in $\widehat{\Delta}_N$ stems from three sources: (i) bias in the common parameter estimates $\widehat{\beta}$, (ii) averaging over a nonlinear function of noisy fixed effect estimates (a Jensen inequality type bias), and (iii) correlation between the fixed effect errors and $m(X_{ij}, \beta, \alpha_i, \gamma_j)$.

The average effect estimator can be bias-corrected using the jackknife in an almost identical way to the bias correction of $\widehat{\beta}$. Let

$$\widehat{\Delta}_{(k)} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \neq i} m(X_{ij}, \widehat{\beta}_{(k)}, \widehat{\alpha}_{(k),i}, \widehat{\gamma}_{(k),j})$$

be the average effect estimate that uses the parameter estimates in the k-th leave-out sample (that is, the sample which we drop observations $(i, j) \in \mathcal{I}_k$). A jackknife bias-corrected estimator is

$$\widehat{\Delta}_J = (N-1)\widehat{\Delta}_N - (N-2)\frac{1}{N-1} \sum_{k=1}^{N-1} \widehat{\Delta}_{(k)}. \tag{1.12}$$

We show in Section 1.4 that the bias-corrected estimator is asymptotically normal with mean zero. Note that there is no need to use a bias-corrected estimate of $\widehat{\beta}$ in the construction of the average effects. The jackknife takes care of the bias generated by bias in $\widehat{\beta}$ as well as the other sources of bias in a single step.

1.3.2 Specification testing

The parameter in (1.11) is an average over a function that depends on a single observation in the network. In some cases, we may be interested in averages over functions that depend on multiple observations such as patterns depending on pairs of nodes (dyads), groups of three or four nodes (triads or tetrads), or other structures. Although these averages may be of interest in their own right, they prove particularly useful in developing specification tests, and we focus on this case. In this section, we show that like simple average effects, these averages also suffer from the incidental parameters bias problem, but can be bias corrected using the jackknife approach.

Let λ be a set of observations in the network; for example, $\lambda = \{(i,j),(j,i)\}$ collects the two observations within a dyad, and $\lambda = \{(i,j),(j,k),(k,i)\}$ collects a sequence of potential links between three nodes. Let Λ_N be the set of all possible λ formed by permuting the nodes for a network of size N. We consider averages of the form

$$\Delta_N = \frac{1}{|\Lambda_N|} \sum_{\lambda} m(Y_\lambda, X_\lambda, \beta, \pi_\lambda), \tag{1.13}$$

where $Y_\lambda = \{Y_{ij}\}_{(i,j)\in\lambda}$, $X_\lambda = \{X_{ij}\}_{(i,j)\in\lambda}$, and $\pi_\lambda = \{\alpha_i + \gamma_j\}_{(i,j)\in\lambda}$ collect the outcomes, covariates and fixed effects for the observations in λ. These generalize the averages in (1.11) in two ways: (i) they allow for averages over functions of multiple observations in the network, and (ii) they allow the function m to depend on the outcome variable Y_{ij}.

One important application of the type of averages in (1.13) is to specification testing. The model presented in this paper assumes that decisions are made bilaterally, that is, agents i and j decide on Y_{ij} independently of other outcomes in the network. In some settings, we may expect that decision making has some strategic aspect, in that an agent's utility from a link depends on the presence (or strength) of other links. One way to model such a phenomenon is to include network statistics in the utility function. For example, imagine that i sends a link to j according to

$$Y_{ij} = \mathbb{1}\big\{\delta S_{ij} + \beta' X_{ij} + \alpha_i + \gamma_j \geq \varepsilon_{ij}\big\}, \tag{1.14}$$

where S_{ij} is the value of some network statistic, and ε_{ij} are independent shocks. Models of this form generally result in multiple equilibria, which raises a number of difficulties for estimation. However, under the null hypothesis $H_0 : \delta = 0$, the model is the dyadic link formation model considered in this paper and can be consistently estimated. This suggests

that a test statistic based on S_{ij} may be useful for testing the null hypothesis of the dyadic model (i.e. assumption (1.1)) against an alternative of the form (1.14). One possible test statistic is

$$T_N = \frac{1}{N(N-1)} \sum_i \sum_{j \neq i} (Y_{ij} - p_{ij}) S_{ij}, \tag{1.15}$$

where $p_{ij} = E[Y_{ij}|X, \beta, \alpha, \gamma]$. The statistic tests the ability of the network statistic S_{ij} to predict the model errors $Y_{ij} - p_{ij}$. Under the null model, $E[(Y_{ij} - p_{ij})S_{ij}] = 0$ and so values of T_N far from zero suggest the presence of unexplained strategic interactions. Graham and Pelican (2020) consider a similar setup in the case of a logit model and derive the locally best conditional similar test for the null hypothesis $H_0 : \delta = 0$ in (1.14). The resulting statistic is similar to (1.15), although this certainly does not imply any optimality of the test proposed here. The motivation for the statistic suggested here is heuristic, but has the advantage of being applicable to a wide set of models (e.g. models that do not admit a sufficient statistic) and has the simplicity of using asymptotic critical values. Some examples of potential test statistics in this framework may be useful.

Example. Reciprocity in link formation

Consider a model in which links are reciprocal, that is the presence of a link from j to i increases the utility of the reverse link from i to j. In this case we could let $S_{ij} = Y_{ji}$ in (1.15), which gives the statistic

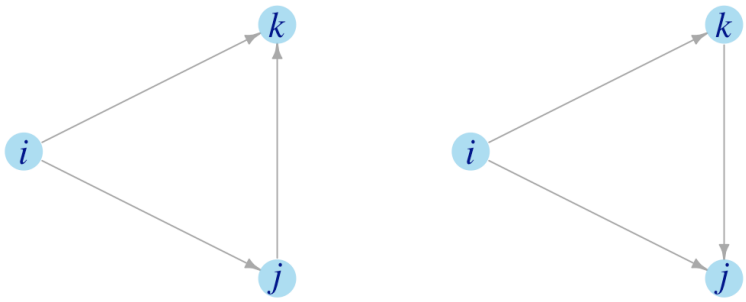
$$T_N = \frac{1}{N(N-1)} \sum_i \sum_{j \neq i} (Y_{ij} - p_{ij}) Y_{ji} \tag{1.16}$$

This statistic measures whether the average prediction error for reciprocal links differs from the average prediction error for non-reciprocal links. Note that reciprocity is allowed for under the assumptions of this paper, so that a rejection of the null hypothesis of no reciprocity does not affect the interpretation of model estimates.

Example. Transitivity

Consider a *triad,* three nodes in the network (i, j, k). Under the dyadic model, linking decisions are independent across the three pairs of nodes in the triad, but in many settings it may be reasonable to think that the existence of links between two of these pairs may affect the formation of the links in the third. For example, imagine that i has formed a directed link to j, and j has formed a link to k. We may expect the existence of the indirect path from i to k (passing through j) to increase the likelihood of observing the direct link from i

Figure 1.2: Transitive triangles



There are six potential transitive triangles that may exist within any triad; the figure above shows two of these (in which i is the sender of two links). Additional links may exist within the triad - these do not affect the existence of a transitive triangle.

to k. This linking structure, $i \to j, i \to k, k \to j$ (shown in the right diagram of Figure 1.2), is known as a *transitive triangle*.

Transitivity in linking is a feature of many models of strategic network formation and we may test the null hypothesis that linking decisions are dyadic against an alternative in which transitivity exists by choosing $S_{ij} = \frac{1}{N-2} \sum_{k \neq \{i,j\}} Y_{ik} Y_{kj}$ and using the statistic

$$T_N = \frac{1}{N(N-1)(N-2)} \sum_i \sum_{j \neq i} \sum_{k \neq \{i,j\}} (Y_{ij} - p_{ij}) Y_{ik} Y_{kj}. \tag{1.17}$$

Note that in both examples, and more generally, the outcome Y_{ij} need not be binary. The test applies equally to non-binary outcomes, for example, in a trade network the presence of large export flows of a particular good from i to k and from k to j may reduce the expected direct exports of that good from i to j.

The framework in (1.14) generates just one possible set of specification tests, and many others statistics of the form (1.13) are possible. One alternative method is to compare the observed frequency of some possible subgraph configuration with the expected frequency under the assumed dyadic model. Such a test, for the case of transitive and cyclic triangles, was proposed by Dzemski (2019), who also derived an analytical bias correction for the statistic in a binary outcome model with normal disturbances. The statistic suggested by

Dzemski (2019) is of the form

$$T_N = \frac{1}{N(N-1)(N-2)} \sum_{i=1}^{N} \sum_{j \neq i} \sum_{k \neq \{i,j\}} \Big(Y_{ij} Y_{ik} Y_{kj} - p_{ij} p_{ik} p_{kj} \Big),$$

where $Y_{ij}Y_{ik}Y_{kj}$ is an indicator for a transitive triangle and $p_{ij}p_{ik}p_{kj}$ is the probability of observing such a triangle when all three links are independent. Many test statistics of this form could be derived in the same way, by taking some function of network outcomes between multiple agents and comparing it to its expectation under the dyadic model. For example, an alternative statistic for reciprocity would be $\frac{1}{N(N-1)} \sum_i \sum_{j \neq i} (Y_{ij} Y_{ji} - p_{ij} p_{ji})$. A key advantage of the jackknife procedure proposed in this paper is that it applies to such a wide variety of statistics. Given the plethora of potential test statistics, an analysis of their power properties would certainly be useful, although is beyond the scope of this paper.

Like estimates of the common parameter β, the test statistics discussed above suffer from an incidental parameter bias. Although the infeasible test statistics are mean zero under a correctly specified model, the feasible versions, which replace parameters $\beta_0, \alpha_0, \gamma_0$ with their estimated values, have an asymptotic bias that leads to incorrect inference. However, we show in Section 1.4 that statistics of the form in (1.13), may be jackknife bias corrected.

To describe the jackknife bias correction for these statistics, denote the number of observations contained in λ as r. Let $\mathbb{1}_\lambda^k = \prod_{(i,j) \in \lambda} \mathbb{1}_{ij}^k$ be an indicator that is *zero* whenever any of the observations in λ are included in the k-th leave-out set $\mathcal{I}_k$. Define the leave-out estimate

$$\widehat{\Delta}_{(k)} = \frac{N-1}{N-r-1} \frac{1}{|\Lambda_N|} \sum_{\lambda} m(Y_\lambda, X_\lambda, \widehat{\beta}_{(k)}, \widehat{\pi}_{\lambda,(k)}) \times \mathbb{1}_\lambda^k,$$

where $\widehat{\beta}_{(k)}, \widehat{\pi}_{\lambda,(k)}$ are parameter estimates from the k-th leave-out estimation. The factor $\frac{N-1}{N-r-1}$ accounts for the fact that m_λ is dropped from the average whenever any of the r observations it is a function of are dropped.[8] A jackknife bias-corrected estimator is again given by

$$\widehat{\Delta}_J = (N-1)\widehat{\Delta}_N - (N-2)\frac{1}{N-1} \sum_k \widehat{\Delta}_{(k)}. \tag{1.18}$$

In Section 1.4, we show that the bias-corrected statistic is asymptotically normal and well

[8] This jackknife differs from (1.12) by jackknifing the average itself, not just the parameter estimates. This is because m here may depend on the outcome Y_{ij}. In settings where m does not depend on outcomes, or it is separable between outcomes and parameters, the simpler jackknife in (1.12) can be used. For example, when $m_\lambda = Y_{ij}Y_{ik}Y_{kj} - p_{ij}p_{ik}p_{kj}$, the average over the second term ($p_{ij}p_{ik}p_{kj}$) may be jackknifed separately and the average over the first term ($Y_{ij}Y_{ik}Y_{kj}$) left as is.

centered. In the case of the specification test statistics discussed above, we have $\bar{E}[\Delta_N] = 0$ and so $N\widehat{\Delta}_J \Rightarrow N(0, V_\Delta)$, where the form of the variance is shown in Theorem 1.2. This allows us to test hypotheses in the usual way, comparing $N\widehat{\Delta}_J/\sqrt{V_\Delta}$ to the quantiles of a standard normal distribution. Further details on the implementation of these tests are discussed below.

1.4 Asymptotic theory

We consider an asymptotic framework in which a single network of N agents grows in size, i.e. $N \to \infty$. Recall that the parameters of interest maximize the objective function in (1.2) for some function $\ell_{ij} = \ell(Y_{ij}, X_{ij}, \beta, \alpha_i + \gamma_j)$ of the observables $Z_{ij} = (Y_{ij}, X_{ij})$ and additive unobserved fixed effects $\alpha_i + \gamma_j$. The asymptotic theory relies on expansions of the objective function, for which we require certain differentiability and moment conditions. We let $\phi' = (\alpha', \gamma')$ denote the $2N \times 1$ vector of fixed effect parameters and $\pi_{ij} = \alpha_i + \gamma_j$ represent the additive index through which they enter the objective function. We denote derivatives of the function ℓ with respect to parameters by $\partial_\beta \ell_{ij}(\beta, \alpha, \gamma) = \partial \ell_{ij}(\beta, \alpha, \gamma)/\partial\beta$, $\partial_{\pi^q} \ell_{ij}(\beta, \alpha, \gamma) = \partial^q \ell_{ij}(\beta, \alpha, \gamma)/\partial\pi^q$ etc. When evaluating these objects at the true parameter values, we simply write $\partial_{\pi^q}\ell_{ij}$ and so on.

1.4.1 Asymptotic analysis for the common parameters

The results below are derived under the following set of assumptions. Proofs are provided in the Appendix.

Assumption 1.1. *Let $\varepsilon > 0$. For every (i, j) let $\mathcal{B}_{\varepsilon,ij}$ be a subset of $\mathbb{R}^{\dim\beta+1}$ that contains an ε-neighborhood of $(\beta_0, \pi_{0,ij})$ for all N.*

(i) Conditional on (X, α, γ), dyads are independent, that is,

$$Y_{ij} \perp\!\!\!\perp Y_{kl} | X, \beta, \alpha, \gamma \qquad \forall (k, l) \notin \{(i, j), (j, i)\}.$$

(ii) For all i, j and N we have that $\bar{E}[\partial_\beta \ell_{ij}] = \bar{E}[\partial_\pi \ell_{ij}] = 0$. For all N, the objective function $\mathcal{L}$ is strictly concave over $\mathbb{R}^{\dim\beta+2N}$, and the matrix $\bar{\mathcal{H}} = -\partial_{\phi\phi'}\bar{\mathcal{L}}$ is positive definite.

(iii) For all (i, j), the function $(\beta, \pi) \mapsto \ell_{ij}(\beta, \pi)$ is five times continuously differentiable over $\mathcal{B}_{\varepsilon,ij}$ almost surely. For all (i, j), the partial derivatives of ℓ_{ij} with respect to the elements

of (β, π) *up to fifth order are bounded in absolute value uniformly over* $(\beta, \pi) \in \mathcal{B}_{\varepsilon,ij}$ *by a function* $M(Z_{ij}) > 0$ *a.s., where* $\max_{i,j} \bar{E}[M(Z_{ij})^{16}]$ *is a.s. uniformly bounded over* N.

(iv) Let

$$\bar{W}_N = -\frac{1}{N}\big(\partial_{\beta\beta'}\bar{\mathcal{L}} - (\partial_{\phi\beta'}\bar{\mathcal{L}})(\partial_{\phi\phi'}\bar{\mathcal{L}})^{-1}(\partial_{\beta\phi}\bar{\mathcal{L}})\big)$$
$$\bar{\Omega}_N = \bar{E}\big[\big(\partial_{\beta}\mathcal{L} - (\partial_{\beta\phi'}\bar{\mathcal{L}})(\partial_{\phi\phi'}\bar{\mathcal{L}})^{-1}(\partial_{\phi}\mathcal{L})\big)^2\big].$$

The limits $plim_{N\to\infty}\bar{W}_N = W$ *and* $plim_{N\to\infty}\bar{\Omega}_N = \Omega$ *exist for* W *and* Ω *positive definite matrices.*

(v) There exist constants b_{min} *and* b_{max} *such that for all* $(\beta, \pi) \in \mathcal{B}_{\varepsilon,ij}$

$$0 < b_{min} \leq -\bar{E}[\partial_{\pi^2}\ell_{ij}(\beta, \pi)] \leq b_{max}$$

a.s. uniformly over i, j *and* N.

Assumption 1.1 (i) specifies that outcomes depend on dyad specific variables only, and not on other features of the network. Conditional on the observed covariates and fixed effects, the outcome Y_{ij} is independent of other outcomes in the network, with the exception of Y_{ji}. Note that unconditionally outcomes are allowed to be dependent, through dependence across covariates X_{ij} and the fixed effects.

Assumption 1.1 (ii) contains the parametric restriction of the model and requires that the true parameters $\beta_0, \alpha_0, \gamma_0$ are solutions to the first-order equations of the objective function. Concavity of the objective function ensures that the population problem has a unique solution. This is satisfied in many common nonlinear models, including the class of regression models with log-concave densities (as well as censored and truncated versions of these models), which includes probit, logit, ordered probit, Tobit, gamma and beta models among others (see Pratt (1981), Newey and McFadden (1994)).

Assumption 1.1 (iii) provides basic smoothness conditions for the objective function. The derivative and moment conditions are required to ensure the validity of the asymptotic expansions to high enough order to establish the properties of the jackknife procedure, and to ensure that remainder terms are well bounded. Analysis of the jackknife requires higher order expansions than are required for characterization of the analytical bias and first-order asymptotic properties of the estimator, and so (iii) is somewhat stronger than the equivalent assumption employed in Fernández-Val and Weidner (2016).

Assumption 1.1 (iv) ensures that the variance of $\widehat{\beta}$ is non-degenerate. The term $\bar{W}_N$ is

the Hessian matrix for the common parameters β, after concentrating out the fixed effect parameters, while $\bar{\Omega}_N$ is the variance of the score for β (again after concentrating out the fixed effect parameters). To describe estimators of these terms, let

$$\Xi_{ij} = -\frac{1}{N}\sum_{s}\sum_{t\neq s}\big(\bar{\mathcal{H}}^{-1}_{(\alpha\alpha)is} + \bar{\mathcal{H}}^{-1}_{(\gamma\alpha)jt} + \bar{\mathcal{H}}^{-1}_{(\alpha\gamma)it} + \bar{\mathcal{H}}^{-1}_{(\gamma\gamma)st}\big)\bar{E}[\partial_{\beta\pi}\ell_{st}],$$

and define $D_\beta\ell_{ij} = \partial_\beta\ell_{ij} - \Xi_{ij}(\partial_\pi\ell_{ij})$. This term is the score for β after partialling out the fixed effect parameters. Estimators of the variance terms can be created in the usual way by plugging in estimates of the model parameters, i.e.

$$\begin{aligned}\widehat{W}_N &= -\frac{1}{N}\big(\partial_{\beta\beta'}\widehat{\mathcal{L}} - (\partial_{\phi\beta'}\widehat{\mathcal{L}})(\partial_{\phi\phi'}\widehat{\mathcal{L}})^{-1}(\partial_{\beta\phi}\widehat{\mathcal{L}})\big),\\ \widehat{\Omega}_N &= \frac{1}{N(N-1)}\sum_{i}\sum_{j<i}(\widehat{D_\beta\ell_{ij}} + \widehat{D_\beta\ell_{ji}})^2,\end{aligned} \tag{1.19}$$

where the terms $\partial_{\beta\beta'}\widehat{\mathcal{L}}$, $\widehat{D_\beta\ell_{ij}}$ etc. are evaluated at the estimates $\widehat{\beta}$, $\widehat{\alpha}$, $\widehat{\gamma}$. Note that the estimator $\widehat{\Omega}_N$ allows for correlation between the Y_{ij} and Y_{ji} outcomes.

Finally, Assumption 1.1 (v) ensures that the Hessian for the fixed effect parameters is positive definite. This requires sufficient variation in the outcomes across both dimensions – i.e., variation in Y_{ij} over j (for fixed sender i) and over i (for fixed receiver j). In the binary outcome case, it implies that the model generates a *dense network*, one in which the number of links formed by each node tends to infinity as the size of the network grows. The assumption may not be reasonable in all empirical settings – in simulations we investigate the robustness of the estimator to sparsity in finite samples. The density assumption can be avoided in settings where sufficient statistics for the incidental parameters exists, such as the conditional logit framework, since estimation of the fixed effects is avoided. This comes at the expense of no longer being able to estimate average effects or counterfactual outcomes.

We now state the main theorem of the paper, on the asymptotic distribution of the jackknife bias-corrected estimator.

Theorem 1.1. *Let Assumption 1.1 hold and let $\widehat{\beta}_J$ be the jackknife bias-corrected estimator (1.8), the leave-l-out estimator (1.9) or the weighted jackknife (1.10). Let $V_N = \bar{W}_N^{-1}\bar{\Omega}_N\bar{W}_N^{-1}$ and assume that $V = plim_{N\to\infty}V_N$ exists and is positive definite. Then,*

$$N(\widehat{\beta}_J - \beta_0) \Rightarrow \mathcal{N}(0, V).$$

Let $\widehat{V}_N = \widehat{W}_N^{-1}\widehat{\Omega}_N\widehat{W}_N^{-1}$ be an estimator of the asymptotic variance, where $\widehat{W}_N$ and $\widehat{\Omega}_N$ are the plug-in estimators shown in (1.19). Then $\widehat{V}_N \to V$.

The jackknife estimator is asymptotically normally distributed and unbiased. It also has the same asymptotic variance as the non-bias-corrected estimator in (1.2). The variance is the usual sandwich form one, and is easily computed. In the case of maximum likelihood we will have that $\bar{W}_N = \bar{\Omega}_N$ so that the variance simplifies to $V_N = \bar{W}_N^{-1}$. In general this will not be true, for example the researcher may wish to allow for correlation between Y_{ij} and Y_{ji} by clustering $\bar{\Omega}_N$ at the dyad level as in (1.19).

1.4.2 Asymptotic analysis for fixed effect averages

Here we present asymptotic results for averages of functions that may take more than one observation as arguments. This structure will cover a number of interesting cases such as standard average effects, averages over dyads, triads or other structures in the network, and specification tests. Recall that λ is a set of observations (i, j), and Λ_N is the collection of all such sets formed by permuting the nodes in λ. We let $Y_\lambda = \{Y_{ij}\}_{(i,j)\in\lambda}$, $X_\lambda = \{X_{ij}\}_{(i,j)\in\lambda}$, and $\pi_\lambda = \{\alpha_i + \gamma_j\}_{(i,j)\in\lambda}$ collect the outcomes, covariates and fixed effects for the observations in λ. The function of interest is $m_\lambda = m(Y_\lambda, X_\lambda, \beta, \pi_\lambda)$, which is a function of $(Y_{ij}, X_{ij}, \alpha_i + \gamma_j)$ for each $(i, j) \in \lambda$.

It will be useful to define three separate quantities

$$\delta = E\big[\Delta_N\big], \quad \bar{\Delta}_N = \bar{E}\big[\Delta_N\big], \text{ and } \Delta_N = \frac{1}{|\Lambda_N|}\sum_{\lambda\in\Lambda_N} m_\lambda.$$

Here, Δ_N is the average effect computed in the observed sample (at the true parameter values), $\bar{\Delta}_N$ is the expectation of the average effects conditional on the distribution of covariates and fixed effects in the observed sample, while δ_N is the population expectation. We can decompose the estimation error $\widehat{\Delta}_N - \delta$ into three sources

$$\widehat{\Delta}_N - \delta = (\widehat{\Delta}_N - \Delta_N) + (\Delta_N - \bar{\Delta}_N) + (\bar{\Delta}_N - \delta). \tag{1.20}$$

The first term, $\widehat{\Delta}_N - \Delta_N$, represents variation caused by estimation of the parameters in the model, including fixed effects. The next term, $\Delta_N - \bar{\Delta}_N$, is variation of the sample outcomes m_λ around their conditional expectations $\bar{m}_\lambda = \bar{E}[m_\lambda]$. In the case that m is a function of the data only through X_λ, i.e. it does not depend on outcomes Y_λ, we will have $\bar{m}_\lambda = m_\lambda$

and this second term will vanish. Finally, $\bar{\Delta}_N - \delta$ captures differences in the distribution of covariates and fixed effects in the observed network, relative to the population. In the case that the full network is observed, or whenever $\bar{\Delta}_N = 0$ as is the case for specification tests, we will have that $\bar{\Delta}_N = \delta_N$.

The results below will rely on Assumption 1.1, as well as additional restrictions on the choices of λ and m.

Assumption 1.2. *Let λ be a set of r observations (i,j) containing p distinct agents. For every λ, let $\mathcal{B}_{\varepsilon,\lambda}$ be a subset of $\mathbb{R}^{\dim\beta+r}$ that contains an ε-neighborhood of $(\beta_0, \pi_{0,\lambda})$ for all N, with $\varepsilon > 0$.*

(i) The number of observations and unique agents in λ, r and p, are fixed over N. The set Λ_N contains all $\frac{N!}{(N-p)!}$ permutations of agents in the set of observations λ.

(ii) The function m depends on (α,γ) only through $\pi_\lambda = \{\alpha_i + \gamma_j\}_{(i,j)\in\lambda}$. For all λ, the function $(\beta,\pi_\lambda) \mapsto m(Z_\lambda,\beta,\pi_\lambda)$ is five times continuously differentiable over $\mathcal{B}_{\varepsilon,\lambda}$ a.s. For all λ, the partial derivatives of m with respect to the elements of (β,π_λ) up to fifth order are bounded in absolute value uniformly over $(\beta,\pi_\lambda) \in \mathcal{B}_{\varepsilon,\lambda}$ by a function $M(Z_\lambda) > 0$ a.s., and $\max_\lambda \bar{E}[M(Z_\lambda)^{16}]$ is a.s. uniformly bounded over N.

(iii) We have that $0 < \min_\lambda E[m_\lambda^2] - E[m_\lambda]^2 \leq \max_\lambda E[m_\lambda^2] - E[m_\lambda]^2 < \infty$ uniformly over N

Assumption 1.2 (i) restricts m to be a function of a fixed number of edges in the network, which allows us to construct leave-out sets to bias correct using the jackknife. It also assumes that Δ_N is an average of all possible arrangements of the nodes in λ, ensuring that averaging occurs over all dimensions. For example, an average of the form $\frac{1}{N}\sum_{j\neq i} m(X_{ij},\alpha_i,\gamma_j)$ is not allowed since we are only averaging over the receiver dimension, while holding i fixed.

Assumption 1.2 (ii) is analogous to Assumption 1.1 (iii), and imposes the same differentiability and moment conditions on m as are imposed on ℓ. This allows for asymptotic expansions of $\widehat{\Delta}_N$ to be derived, in the same way as for $\widehat{\beta}$. Finally, (iii) ensures that the unconditional second moments of m are well defined.

The asymptotic distribution depends on the choice of target parameter, either a conditional or population average. The following theorem states the asymptotic result for the jackknife bias-corrected estimator of the conditional fixed effect average $\bar{\Delta}_N$.

Theorem 1.2. *Let Assumptions 1.1 and 1.2 hold, and let $\widehat{\Delta}_J$ be the jackknife bias-corrected*

estimator (1.18). Then

$$N(\widehat{\Delta}_J - \bar{\Delta}_N) \Rightarrow \mathcal{N}(0, V_\Delta).$$

If we additionally assume that $E\big[(m_\lambda - \bar{m}_\lambda)(m_{\lambda'} - \bar{m}_{\lambda'})\big] \neq 0$ *for sets* λ *and* λ' *that share exactly one observation in common, then the asymptotic variance is*

$$V_\Delta = \lim_{N\to\infty} \frac{1}{N(N-1)} \sum_i \sum_{j<i} \bar{E}\big[(h_{ij} + s_{ij})^2\big]$$

where $h_{ij} = -N(\partial_\theta \bar{\Delta}_N)(\partial_{\theta\theta'} \bar{\mathcal{L}})^{-1}\big((\partial_\theta \ell_{ij}) + (\partial_\theta \ell_{ji})\big)$*, for* $\theta' = (\beta, \alpha', \gamma')$*, and* $s_{ij} = \frac{(N-p)!}{(N-2)!} \sum_{\lambda \in \tilde{\Lambda}_{ij}} (m_\lambda - \bar{m}_\lambda)$*, with* $\tilde{\Lambda}_{ij} = \{\lambda : (i,j) \in \lambda \text{ or } (j,i) \in \lambda\}$.

If we have either $m_\lambda = \bar{m}_\lambda$ *or* $E\big[(m_\lambda - \bar{m}_\lambda)(m_{\lambda'} - \bar{m}_{\lambda'})\big] = 0$ *for sets* λ *and* λ' *that share exactly one observation in common, then the asymptotic variance is*

$$V_\Delta = \lim_{N\to\infty} \frac{1}{N^2} \sum_i \sum_{j<i} \bar{E}\big[h_{ij}^2\big].$$

In either case, let $\widehat{V}_\Delta$ *be the plug-in estimator for* V_Δ *that replaces the unknown* θ *with estimates* $\widehat{\theta}$*. Then* $\widehat{V}_\Delta \to V_\Delta$.

Some explanation for the form of the variance may be useful. The two terms h_{ij} and s_{ij} relate to the first two components of (1.20). The first component $\widehat{\Delta}_J - \Delta_N$ contains variation from estimation of the common parameters β and fixed effects α, γ. In the Appendix it is shown that this term can be approximated using the delta-method

$$\begin{aligned} N(\widehat{\Delta}_J - \Delta_N) &= -N(\partial_\theta \bar{\Delta}_N)(\partial_{\theta\theta'} \bar{\mathcal{L}})^{-1}(\partial_\theta \ell_{ij}) + o_p(1) \\ &= \frac{1}{N-1} \sum_i \sum_{j \neq i} h_{ij} + o_p(1). \end{aligned}$$

Note that, replacing the jackknife estimate $\widehat{\Delta}_J$ with the standard estimator $\widehat{\Delta}_N$ would result in additional terms appearing in the above first-order approximation, related to the incidental parameter bias. The second component is

$$N(\Delta_N - \bar{\Delta}_N) = \frac{N}{|\Lambda_N|} \sum_\lambda (m_\lambda - \bar{m}_\lambda).$$

The variance of this term depends on the conditional covariance between m_λ and $m_{\lambda'}$ for

distinct sets λ and λ'. Note that if λ and λ' share no dyads in common then they are conditionally independent. The variance of this term depends on the condition $E\big[(m_\lambda - \bar{m}_\lambda)(m_{\lambda'} - \bar{m}_{\lambda'})\big] \neq 0$ for sets λ and λ' share exactly one observation in common. Under this condition, the variance of $\sum_\lambda (m_\lambda - \bar{m}_\lambda)$ is dominated by covariances between m_λ and $m_{\lambda'}$ for λ and λ' that share exactly one dyad in common; although λ with two or more common dyads also contribute to the variance, there are an order of magnitude fewer such combinations, and so these represent smaller order contributions that do not appear in the asymptotic variance. In settings where $E\big[(m_\lambda - \bar{m}_\lambda)(m_{\lambda'} - \bar{m}_{\lambda'})\big] = 0$ for λ and λ' that share exactly one observation in common, $\Delta_N - \bar{\Delta}_N$ is a degenerate U-statistic and its variance is asymptotically of smaller order than the variance from parameter estimation, i.e. $\widehat{\Delta}_J - \Delta_N$, and so may be ignored.

Theorem 1.2 shows how we may construct confidence sets for the parameter of interest $\bar{\Delta}_N$. When the object of interest is the unconditional average δ, the convergence of the estimator will be dominated by variation from the third component in (1.20), $\bar{\Delta}_N - \delta$. To describe the statistic in this setting, it is useful to use its U-statistic representation. We will additionally assume that $X_{ij} = h(X_i, X_j)$. This condition appears in other work on dyadic models, for example in Graham (2017). When X_{ij} measures the similarity (or difference) between i and j in some measure, we will commonly have $X_{ij} = d(X_i - X_j)$, for d some distance function. Alternatively, if X_{ij} captures common membership in some group, we may have $X_{ij} = X_i X_j$ where X_i is an indicator for i's membership.

To give $\bar{\Delta}_N$ a U-statistic representation, we first sum together all $\bar{m}_\lambda$ which share the same set of agents. Since there are p agents in each λ, this gives $p!$ different λ that can be created from a given set of agents. We denote these sets of unordered agents by τ. We have, for $\tilde{m} = \bar{m} - E[m]$

$$\begin{aligned}
\bar{\Delta}_N - \delta &= \frac{1}{N \cdots (N-p+1)} \sum_\lambda \tilde{m}_\lambda \\
&= \frac{p!}{N \cdots (N-p+1)} \sum_\tau \Big(\frac{1}{p!} \sum_{\lambda \in \tau} \tilde{m}_\lambda \Big) \\
&= \binom{N}{p}^{-1} \sum_\tau u_\tau,
\end{aligned}$$

where $u_\tau = \frac{1}{p!} \sum_{\lambda \in \tau} \tilde{m}_\lambda$. The term u_τ is a symmetric function of $\{\beta, X_i, \alpha_i, \gamma_i\}$ for p agents i. Assuming that the $\{X_i, \alpha_i, \gamma_i\}$ are i.i.d. over agents, $\bar{\Delta}_N - \delta$ is a U-statistic of order p and we may apply standard theory on such statistics to compute its asymptotic distribution.

Theorem 1.3. *Let Assumptions 1.1 and 1.2 hold. Additionally, assume that $X_{ij} = h(X_i, X_j)$ where X_i is an observed agent-specific characteristic, and also that $(\alpha_i, \gamma_i, X_i)$ is i.i.d. over i. Let*

$$\Sigma_1 = Cov\big(u_\tau, u_{\tau'}\big),$$

for τ, τ' such that $\tau \cap \tau' = \{i\}$. Then, for $V_\delta = p^2\Sigma_1$

$$\sqrt{N}(\widehat{\Delta}_J - \delta) \Rightarrow \mathcal{N}(0, V_\delta).$$

Additionally, the variance estimator $\widehat{V}_\delta$ in (1.21) is consistent, i.e. $\widehat{V}_\delta \to V_\delta$.

The convergence rate in Theorem 1.3 is slower than the rate in Theorem 1.2. While m_λ and $m_{\lambda'}$ are *conditionally* independent when λ and λ' share no dyads in common, the two are *unconditionally* independent only when they share no agents in common. Since there are many more sets λ that share a single agent i than share a dyad (i, j), the variance of $\bar{\Delta}_N - \delta$ is an order of magnitude larger than that of $\widehat{\Delta}_N - \Delta_N$, and so the convergence rate is slower.

Similarly to Theorem 1.2, the variance is dominated by covariances between sets λ that share exactly one node in common. The term Σ_1 is the covariance between u_τ and $u_{\tau'}$ when τ and τ' share exactly one agent in common. A consistent estimator for this quantity is given by

$$\begin{aligned}
\widehat{V}_\delta &= \frac{1}{N}\sum_i \tilde{\mu}_i^2, \\
\tilde{\mu}_i &= \frac{(N-p)!}{(N-1)!}\sum_{\lambda: i\in\lambda} (\widehat{\bar{m}}_\lambda - \widehat{\mu}), \\
\widehat{\mu} &= \frac{(N-p)!}{N!}\sum_\lambda \widehat{m}_\lambda,
\end{aligned} \tag{1.21}$$

where $\tilde{\mu}_i$ is the average over all sets λ containing agent i, $\widehat{\bar{m}}_\lambda$ is a plug-in estimator for $\bar{m}_\lambda$, and $\widehat{\mu} = \sum_\lambda \widehat{m}_\lambda$ is the overall mean.

Since the rate of convergence in Theorem 1.3 is $N^{-1/2}$, there is in fact no asymptotic bias generated by the incidental parameters. The bias from the estimation of the fixed effects is of order N^{-1}, which is smaller than the variation in the sampled distribution of fixed effects around is population distribution. Nonetheless, we would still recommend bias correcting estimators as it is likely to improve the finite sample properties of inference, in terms of correct centering of confidence sets, with little or no cost in terms of additional variance. In the panel data setting, Fernández-Val and Weidner (2016) report such improvements in

simulations.

1.4.3 Examples

Theorems 1.2 and 1.3 suggest that the construction of confidence sets for estimates and hypothesis testing may be performed in the usual way, by using the asymptotic normal approximations. Standard plug-in estimates for the variance expressions may be used. Here we provide some examples of how these results can be used.

Example 1.1. ***Average marginal effect in a probit model***

The average marginal effect in the probit model can be estimated using

$$\widehat{\Delta} = \frac{1}{N(N-1)} \sum_i \sum_{j \neq i} \beta \varphi(\beta' X_{ij} + \alpha_i + \gamma_j),$$

where φ is the standard normal density function. In this setting we have $\lambda = (i, j)$ and $m_{ij} = \bar{m}_{ij}$ so that $s_{ij} = 0$. The variance in Theorem 1.2 is then the standard delta-method variance of

$$\begin{aligned} V_\Delta &= (\partial_\theta \bar{\Delta}_N)(\partial_{\theta\theta'} \bar{\mathcal{L}})^{-1} \bar{\Omega} (\partial_{\theta\theta'} \bar{\mathcal{L}})^{-1} (\partial_\theta \bar{\Delta}_N), \\ \bar{\Omega} &= \frac{1}{N(N-1)} \sum_{i<j} \bar{E}\big[(\partial_\theta \ell_{ij} + \partial_\theta \ell_{ji})(\partial_\theta \ell_{ij} + \partial_\theta \ell_{ji})'\big], \end{aligned}$$

and standard plug-ing estimators may be used. A $(1-\alpha)$-per cent confidence set could be constructed for Δ_N using $\widehat{\Delta}_J \pm c_{1-\alpha/2} V_\Delta^{1/2}/N$, where $c_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution.

If instead we are interested in inference with respect to the population parameter δ, Theorem 1.3 states that we may compute the asymptotic variance as

$$\widehat{V}_\delta = \frac{1}{N} \sum_i \Big(\frac{1}{N-1} \sum_{j \neq i} (\beta \varphi_{ij} + \beta \varphi_{ji} - 2\widehat{\Delta}) \Big)^2.$$

A $(1-\alpha)$-per cent confidence set for δ is $\widehat{\Delta}_J \pm c_{1-\alpha/2} \widehat{V}_\delta^{1/2}/\sqrt{N}$.

Example 1.2. ***Testing transitivity in a probit model***

Recall that a statistic for testing the presence of transitivity is

$$\widehat{\Delta}_N = \frac{1}{N(N-1)(N-2)} \sum_i \sum_{j \neq i} \sum_{k \neq \{i,j\}} (Y_{ij} - p_{ij}) Y_{ik} Y_{kj},$$

where $p_{ij} = E[Y_{ij}|X, \beta, \alpha, \gamma]$. Fitting Y_{ij} with a probit regression, we have $p_{ij} = \Phi(\beta' X_{ij} + \alpha_i + \gamma_j)$. Since $\bar{m}_\lambda = \bar{E}\big[(Y_{ij} - p_{ij}) Y_{ik} Y_{kj}\big] = 0$, we have $\bar{\Delta}_N = \delta = 0$ so we may determine the asymptotic distribution of the test statistic using Theorem 1.2. We have

$$\begin{aligned} s_{ij} = \frac{1}{N-2} \sum_{k \neq \{i,j\}} \Big(& (Y_{ij} - p_{ij}) Y_{ik} Y_{kj} + (Y_{ji} - p_{ji}) Y_{jk} Y_{ki} \\ & + (Y_{ik} - p_{ik}) Y_{ij} Y_{jk} + (Y_{jk} - p_{jk}) Y_{ji} Y_{ik} \\ & + (Y_{kj} - p_{kj}) Y_{ki} Y_{ij} + (Y_{ki} - p_{ki}) Y_{kj} Y_{ji} \Big). \end{aligned}$$

From the likelihood for a probit model, we have $\ell_{ij} = Y_{ij} \log(p_{ij}) + (1 - Y_{ij}) \log(1 - p_{ij})$, which gives $\partial_\pi \ell_{ij} = H_{ij}(Y_{ij} - p_{ij})$ and $\partial_\beta \ell_{ij} = H_{ij}(Y_{ij} - p_{ij}) X_{ij}$ for $H_{ij} = \varphi_{ij} / p_{ij}(1 - p_{ij})$. Also,

$$\begin{aligned} \partial_\beta \bar{\Delta}_N &= -\frac{1}{N(N-1)(N-2)} \sum_i \sum_{j \neq i} \sum_{k \neq \{i,j\}} \varphi_{ij} p_{ik} p_{kj} X_{ij}, \\ \partial_{\alpha_i} \bar{\Delta}_N &= -\frac{1}{N(N-1)(N-2)} \sum_{j \neq i} \sum_{k \neq \{i,j\}} \varphi_{ij} p_{ik} p_{kj}, \\ \partial_{\gamma_i} \bar{\Delta}_N &= -\frac{1}{N(N-1)(N-2)} \sum_{j \neq i} \sum_{k \neq \{i,j\}} \varphi_{ji} p_{jk} p_{ki}, \end{aligned}$$

from which we can construct h_{ij} using the estimated Hessian matrix and the formula given in Theorem 1.2.

1.5 Empirical example

We illustrate the jackknife procedure on a data set consisting of a directed network of export volumes between 136 countries (136×135 country pair observations) in 1990. The data are taken from Santos Silva and Tenreyro (2006), and additional details on their construction can be found in their paper. The outcome variable Y_{ij} is the value of exports from country i to country j. We also use several covariates to capture homophily in trade relationships,

which include: *log distance*, the log of the distance between the capitals of the countries; *border*, an indicator of whether the countries share a common border; *language*, an indicator for whether the countries share a language; *colonial*, and indicator for whether either country had colonized the other at some point in history; and *trade agreement*, an indicator for the presence of a joint preferential trade agreement between the two countries.

1.5.1 Zero-inflated binomial model

Burger et al. (2009) propose a zero-inflated negative binomial model. The value of trade between i and j is given by the product of two variables $Y_{ij} = z_{ij}Y_{ij}^*$, where $z_{ij} \in \{0, 1\}$ is a binary decision to enter into a trading relationship, while Y_{ij}^* is the value of exports that will be realized, conditional on $z_{ij} = 1$. The binary decision is modeled using as a probit function, while the latent outcome Y_{ij}^* is modeled as a negative binomial variable.

In this example, the objective function is given by

$$f(Y_{ij}|X_{ij}, \theta) = \mathbb{1}\{Y_{ij} = 0\}p_{ij} + (1 - p_{ij})g(Y_{ij}|X_{ij}, \theta)$$

where $\theta = (\beta, \alpha, \gamma, \nu)$, and

$$\begin{aligned} p_{ij} &= \Phi(X_{ij}'\beta^z + \alpha_i^z + \gamma_j^z) \\ g(Y_{ij}|X_{ij}, \theta) &= \frac{\Gamma(Y_{ij} + \nu)}{\Gamma(\nu)Y_{ij}!}\Big(\frac{\nu}{\nu + \mu_{ij}}\Big)^{\nu}\Big(\frac{\mu_{ij}}{\nu + \mu_{ij}}\Big)^{Y_{ij}} \\ \mu_{ij} &= \exp(X_{ij}'\beta^y + \alpha_i^y + \gamma_j^y) \end{aligned}$$

The parameter ν captures the degree of overdispersion in the model for Y_{ij}^*, with $\nu \to \infty$ resulting in a model with equal mean and variance (as in the Poisson), while smaller ν lead to greater degrees of dispersion.

Estimates of the parameters in the model are presented in Table 1.1. Most variables change by only small amounts after bias correction. However, the effect of sharing a common border on the probability of engaging in zero trade changes significantly after bias correction; while the maximum likelihood estimate suggests that common borders are important for link formation, the bias corrected estimate is no longer significant. This suggests that the sharing a common border has little effect on the likelihood of engaging in trade, but does affect the volume of trade. The results also suggests a substantial impact of trade agreements, both on the probability of engaging in trade and on the volume of trade, a result that is robust to

bias correction. The overdispersion parameter ν is less than a half, suggesting a significant amount of overdispersion, i.e. export volumes have far greater variation across country pairs than suggested by a Poisson model.

The rightmost column in the table reports the difference between the MLE and jackknife bias-corrected estimators in terms of their standard errors. For a number of variables in the model of export volumes, the change in estimate is around three-quarters of the standard deviation or more, which has an important impact on inference. To give some idea of the scale of these biases, a bias of three-quarters of a standard error is enough for a five per cent test two reject around 12 per cent of the time (more than twice nominal size), while bias of 1.5 standard errors leads to a rejection rate of more than 30 per cent.

Table 1.1: Estimated model coefficients

	MLE	Jackknife	SE	(Bias/SE)
Zero model				
log distance	0.721	0.721	0.029	0.00
border	0.628	0.157	0.120	3.93
language	−0.330	−0.306	0.053	0.45
colonial	−0.305	−0.282	0.056	0.41
trade agreement	−1.168	−1.126	0.180	0.24
Volume model				
log distance	−1.243	−1.218	0.033	0.77
border	0.437	0.483	0.129	0.36
language	0.405	0.418	0.068	0.18
colonial	0.399	0.335	0.073	0.88
trade agreement	1.055	0.960	0.131	0.73
ν	0.492	0.459	0.008	4.38

Table 1.2 contains estimates of the average effect of a regressor on expected export volume, conditional on non-zero trade, over the distribution of regressors and fixed effects for trading country pairs. That is, we calculate (for $n_1 = \sum_i \sum_{j \neq i} \mathbb{1}\{Y_{ij} > 0\}$)

$$\Delta_N = \frac{1}{n_1} \sum_i \sum_{j \neq i} \mathbb{1}\{Y_{ij} > 0\} \beta_{dist} \exp(X_{ij}' \beta^y + \alpha_i^y + \gamma_j^y)$$

Table 1.2: Estimated average effects

	MLE	Jackknife	SE	(Bias/SE)
log distance	−116.2	−113.8	9.08	0.26
border	47.5	50.2	16.95	0.16
language	43.4	41.0	8.64	0.28
colonial	44.4	31.0	10.00	1.34
trade agreement	140.2	107.1	28.10	1.18

for the continuous regressor *log distance* and

$$\Delta_N = \frac{1}{n_1}\sum_i \sum_{j\neq i} \mathbb{1}\{Y_{ij}>0\}\big(\exp(\beta^{y\prime}X_{ij}^{(1)}+\alpha_i^y+\gamma_j^y)-\exp(\beta^{y\prime}X_{ij}^{(0)}+\alpha_i^y+\gamma_j^y)\big)$$

for binary regressors, where $X_{ij}^{(1)}$ sets the binary regressor of interest to one for all (i,j) and $X_{ij}^{(0)}$ sets it to zero (leaving other regressors unchanged). Again, the jackknife bias correction has an important impact on two of the effects; for example, the effect of a trade agreement on expected export volumes decreases by about a quarter (more than a one standard error change in magnitude). Note that, as is the case here, a small bias in the coefficient on some variable does not necessarily imply low bias in the corresponding marginal effect.

1.5.2 Testing for strategic interactions

To demonstrate the use of the jackknife for specification testing, we implement the test in (1.15) in a binary model for the probability of country i exporting to country j. We test for two types of strategic interaction: reciprocity, using $S_{ij} = Y_{ji}$; and transitivity, using $S_{ij} = \frac{1}{N-2}\sum_{k\neq\{i,j\}} Y_{ik}Y_{kj}$. In both cases we construct the statistics by estimating p_{ij} using a probit model and jackknife the statistic. Standard errors are computed using the expressions in Theorem 1.2. Table 1.3 presents the values of the statistics as well as the standardized values $t_N = NT_N/\sqrt{V_T}$ and $t_J = NT_J/\sqrt{V_T}$ for both tests.

For the reciprocity statistic, the jackknife bias correction appears to have little effect. The statistic rejects the null of no reciprocity strongly suggesting that the existence of an export relationship from i to j increases the likelihood of i also importing from j. This is perhaps not surprising. It is important to note that the model considered in this paper allows for reciprocity so that this conclusion has no impact on the validity of model estimates. The presence of reciprocity does suggest that standard errors should be clustered at the dyad

level to account for correlation between the outcomes Y_{ij} and Y_{ji}.

Table 1.3: Strategic interaction tests

	NT_N	NT_J	t_N	t_J
Reciprocity	3.160	3.249	16.96	17.43
Transitivity	−0.094	0.012	−6.75	0.88

In contrast, the jackknife bias correction appears to have an important effect on the transitivity statistic. The uncorrected statistic leads to a rejection of the null, and the conclusion that indirect paths of trade (exports paths from i to j through a third-party country) are associated with a lower probability of a direct export relationship than is expected given the model. However, the jackknifed statistic is close to zero so that we do not reject the null hypothesis that trade decisions are bilateral in nature.

1.5.3 Comparison with conditional logit estimator

As a comparison with an existing approach to the incidental parameters problem in networks, we consider estimating a logit model for the probability that country i exports to country j. Under the logit specification, it is possible to estimate the common parameters in the model by first removing the fixed effects. This approach has been suggested by Graham (2017) for an undirected network, and Jochmans (2018) for a directed network. The conditional logit estimator works by forming difference-in-differences style contrasts among sets of four nodes (tetrads). Let

$$z_{ij,kl} = \frac{(Y_{ik} - Y_{il}) - (Y_{jk} - Y_{jl})}{2}$$
$$r_{ij,kl} = (X_{ik} - X_{il}) - (X_{jk} - X_{jl})$$

Given the logistic specification

$$P(Y_{ij} = 1|X, \beta, \alpha, \gamma) = \frac{\exp(\beta' X_{ij} + \alpha_i + \gamma_j)}{1 + \exp(\beta' X_{ij} + \alpha_i + \gamma_j)}$$

and conditional independence of outcomes across dyads, we may write

$$P(z_{ij,kl} = 1|z_{ij,kl} \in \{-1, 1\}, X, \beta, \alpha, \gamma) = \frac{\exp(\beta' r_{ij})}{1 + \exp(\beta' r_{ij})} \tag{1.22}$$

Table 1.4: Coefficient estimates for logit model

	MLE	Jackknife	Conditional logit
log distance	-1.341	-1.305	-1.125
	(0.062)	(0.062)	(0.059)
border	-1.192	-1.157	0.866
	(0.265)	(0.265)	(0.268)
language	0.590	0.573	0.488
	(0.105)	(0.105)	(0.104)
colonial	0.509	0.493	0.579
	(0.110)	(0.110)	(0.106)
trade agreement	2.057	1.998	1.653
	(0.407)	(0.407)	(0.349)

Standard errors are shown in parentheses. For the MLE and jackknife, the standard errors are those shown in Theorem 1.1, while for the conditional logit standard errors are computed using the asymptotic distribution in Jochmans (2018).

That is, conditional on the event $z_{ij,kl} \in \{-1, 1\}$, the outcome $z_{ij,kl}$ follows a logit model without any fixed effect parameters. This allows us to estimate the common parameter β by a standard logit regression of $z_{ij,kl}$ on $r_{ij,kl}$ in the subsample of $z_{ij,kl} \in \{-1, 1\}$. Estimates from this model, as well as the jackknife bias-corrected estimates, are shown in Table 1.4.

Interestingly, although the jackknife bias correction suggests little bias in the original logit parameter estimates, the conditional logit estimates are significantly different. For example, while the coefficient on distance is similar across the MLE and jackknife estimates (-1.34 and -1.31 respectively), the conditional logit estimate differs by more than three standard errors.

There may be a number of reasons for such a discrepancy between the estimates. One possibility is that the model is misspecified in the sense that the correct link function is not the logistic function. In this case, maximum likelihood estimates a set of pseudo parameters that represent the parameters that minimize the Kullback-Leibler distance between the logit model and the true model (White, 1982). Since the jackknife theory does not rely on any information equalities, the jackknife bias correction is asymptotically unbiased for these pseudo parameters. In contrast, the conditional logit estimator estimates a different set of pseudo parameters, those that minimize the KL distance conditional on $z_{ijkl} \in \{-1, 1\}$, for a logit regression of z_{ijkl} on r_{ijkl}. To the best of our knowledge, there is no reason to suspect that these two sets of pseudo parameters would coincide.

Table 1.5: Fixed effect distributions

Name	C_N^l	C_N^u	Density[(a)]
bal	$-\log\log N$	$\log\log N$	0.50
llog	$-\log\log N$	0	0.19
slog	$-\log^{1/2} N$	0	0.12
log	$-\log N$	0	0.03

[a] Values are the average density over 100 simulations in a network of $N = 50$ nodes.

An alternative explanation is that the assumption of independence between dyads is violated in the data. In this case, the likelihood function for the conditional logit will be incorrect, since the identity (1.22) will no longer hold. In this setting the bias correction given by the jackknife estimator is also likely to be incorrect since it will not account for bias terms generated by the dependence across dyads.

1.6 Simulations

Here I demonstrate the effectiveness of the jackknife in simulations. I repeat the simulation design of Dzemski (2019), which has also been used in a number of other network papers. The binary outcome Y_{ij} is determined by

$$Y_{ij} = 1\{\theta X_{ij} + \alpha_i + \gamma_j > \varepsilon_{ij}\}$$

where $\theta = 1$ and $\varepsilon_{ij} \sim N(0,1)$. Individual i is characterized by the binary scalar $X_i = 1 - 2 \times 1\{i \text{ is odd}\}$, and the homophily variable is given by $X_{ij} = X_i X_j$, i.e. it is one for pairs with the same sign and minus one for pairs with opposing signs. The fixed effects are given by

$$\alpha_i = \gamma_i = C_N^l - \frac{N-i}{N-1}(C_N^u - C_N^l)$$

which is a sequence from C_N^l to C_N^u. The value of (C_N^l, C_N^u) is intended to control the sparsity of the network, and we consider four choices, shown in Table 1.5.

In the balanced setting ('bal'), fixed effects range between $\pm \log\log N$, generating a dense network in which around half of all links are formed. Subsequent settings feature increasingly sparse networks in which some nodes remain well connected, while others make few links. In the sparsest setting ('log') only around 3 per cent of all links are formed.

Table 1.6: Simulation results

	Bias (mean)				SD				Rejection (5%)			
	$\hat{\theta}_{MLE}$	$\hat{\theta}_{BC}$	$\hat{\theta}_{J}$	$\hat{\theta}_{wJ}$	$\hat{\theta}_{MLE}$	$\hat{\theta}_{BC}$	$\hat{\theta}_{J}$	$\hat{\theta}_{wJ}$	$\hat{\theta}_{MLE}$	$\hat{\theta}_{BC}$	$\hat{\theta}_{J}$	$\hat{\theta}_{wJ}$
bal	0.06	0.00	-0.01	0.00	0.04	0.04	0.04	0.04	0.29	0.03	0.03	0.03
llog	0.07	0.01	-0.01	0.00	0.06	0.05	0.06	0.05	0.27	0.04	0.04	0.04
slog	0.11	0.02	-0.03	0.01	0.10	0.09	0.13	0.12	0.26	0.05	0.05	0.05
log	0.51	-	-1.23	0.04	0.60	-	1.37	0.42	0.27	0.23	0.67	0.29

[a] $N = 50$, although in the sparsest setting, there were on average 29 agents in the connected network
[b] In sparse settings the analytical correction generated some extreme outliers. Median and 95-5 percentile range are reported in Table 1.7.

I simulate the model 1000 times and compute the MLE, analytical bias-corrected estimate, and both the standard and weighted jackknife bias-corrected estimates for each fixed effect distribution.

Table 1.6 presents the bias, standard deviation, and rejection rates for a 5 per cent test of the null hypothesis $\beta_0 = 1$ for each estimator. As expected, the MLE is biased, with the size of the bias increasing in the sparsity of the network. In each case, the bias is around one standard deviation in size, resulting in substantial over-rejection. In contrast, the jackknife estimator is approximately unbiased in the first two designs. In the sparsest design the jackknife estimator does not appear to perform well, and shows similar bias to the MLE. The weighted jackknife performs well, even in the sparse designs; it is unbiased in the first three designs, and removes the majority of the bias even in the sparsest design. In each case, the jackknife estimators have smaller standard deviation than the MLE (with the exception of the standard jackknife in the sparsest design). Rejection rates are at or below the nominal level in three of the four settings, although in the sparsest setting all estimators over reject.

Weighted jackknife

To help explain the properties of the weighted jackknife, we investigate how the weight given to each leave-out estimate $\widehat{\beta}_{(k)}$ correlates with the contribution of that estimate to the total error of the jackknife estimate. Define

$$e_{(k)} = (N-1)\widehat{\beta}_{MLE} - (N-2)\widehat{\beta}_{(k)} - \beta_0$$

as the contribution to the error $\widehat{\beta}_J - \beta_0$ from a single leave-out estimate $\widehat{\beta}_{(k)}$. The total error of the standard jackknife is simply the average of these errors $\widehat{\beta}_J - \beta_0 = \frac{1}{N-1}\sum_k e_{(k)}$,

Table 1.7: Simulation results - median and range

	Bias (median)				95-5 range			
	$\hat{\theta}_{MLE}$	$\hat{\theta}_{BC}$	$\hat{\theta}_{J}$	$\hat{\theta}_{wJ}$	$\hat{\theta}_{MLE}$	$\hat{\theta}_{BC}$	$\hat{\theta}_{J}$	$\hat{\theta}_{wJ}$
bal	0.06	0.00	-0.01	0.00	0.14	0.13	0.13	0.13
llog	0.07	0.00	-0.01	0.00	0.19	0.18	0.17	0.18
slog	0.10	0.01	-0.03	0.00	0.27	0.25	0.23	0.25
log	0.23	0.11	-0.95	-0.06	1.55	1.38	4.24	1.13

[a] top panel is $N = 50$, lower panel is $N = 70$
[b] $\bar{N}$ is the average number of nodes in the connected network

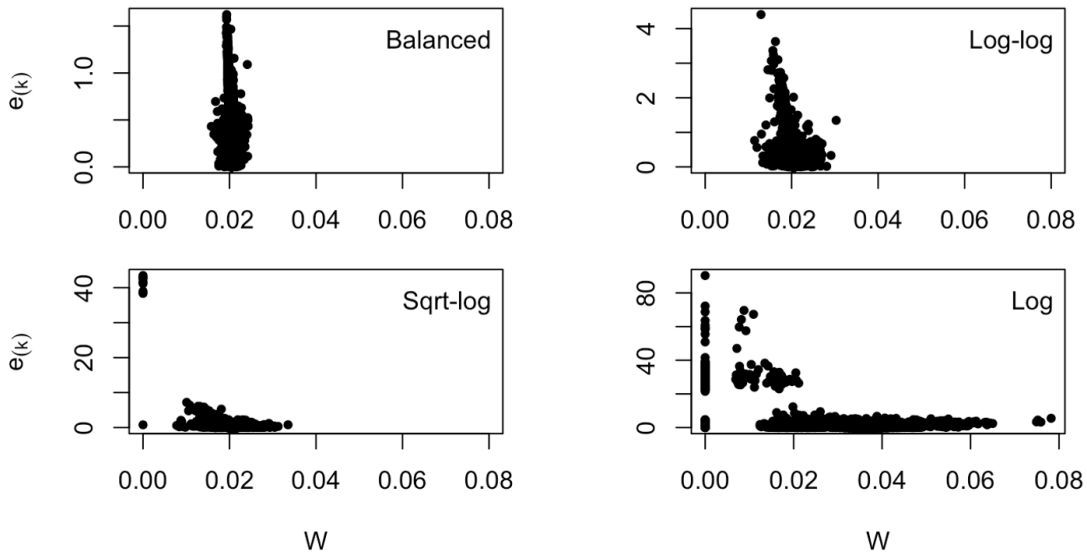
while the weighted jackknife gives differing weights to different leave-out estimates. Figure 1.3 plots the absolute value of these errors against their weights in the weighted jackknife across the four simulation designs.

For the balanced design (upper left panel), which is the densest network, the weights are concentrated around $\frac{1}{N-1} \approx 0.02$ with very little variation. In this case, the weighted jackknife is almost identical to the standard jackknife, as is clear in Tables 1.6 and 1.7. As the level of sparsity in the network increases, so does the dispersion in weights and estimates across the leave-out samples. In the sparsest design, $C_N^l = -\log N$ (lower right panel), the weights vary considerably, with many close to zero. Note that the sparser designs also exhibit large outliers, with extremely large errors, but that these outliers typically receive weights close to zero. It is this feature that appears to drive the success of the weighted jackknife over the standard jackknife in the sparser designs ($C_N^l = -\{\log^{1/2} N, \log N\}$) in Tables 1.6 and 1.7.

Simulations for leave-l-out style jackknife

Table 1.8 reports the results of simulations of the leave-l-out style jackknife in a network of $N = 101$ nodes. The model is the same as that in the previous section, with fixed effect distributions given in Table 1.5. Results are shown for $l = \{5, 10, 20, 50\}$, which corresponds to $20, 10, 5$, and 2 leave-out estimates in each case. The leave-l-out jackknife performs well, even for l as large as 50 in the dense network scenarios. In the sparsest scenario, the jackknife performs reasonably well for l as large as 20, but appears to break down for $l = 50$. For large $l = 50$, half of all observations in the network are dropped for each leave-out estimation, which appears to create some issues when the network is sparse.

Figure 1.3: Weights ($\widehat{W}_{(k)}$) versus estimates ($\widehat{\beta}_{(k)}$) in leave-out samples



The y-axis is which measures the absolute error of the jackknife estimator using the single leave-out estimate. The x-axis is the weight given to that leave-out estimate in the weighted jackknife.

Table 1.8: Simulation results $N = 101$

	Bias (mean)				SD				Rejection (5%)			
$l =$	5	10	20	50	5	10	20	50	5	10	20	50
Standard jackknife												
bal	0.00	0.00	0.00	0.00	0.02	0.02	0.02	0.02	0.05	0.06	0.04	0.05
llog	0.00	0.00	0.00	-0.01	0.03	0.03	0.03	0.03	0.05	0.05	0.03	0.06
slog	-0.01	-0.01	-0.01	-0.01	0.04	0.04	0.04	0.04	0.05	0.04	0.05	0.09
log	-0.09	-0.08	-0.13	-0.53	0.21	0.24	0.32	0.65	0.13	0.15	0.27	0.75
Weighted jackknife												
bal	0.00	0.00	0.00	0.00	0.02	0.02	0.02	0.02	0.05	0.05	0.04	0.05
llog	0.00	0.00	0.00	0.00	0.03	0.03	0.03	0.03	0.05	0.05	0.03	0.06
slog	0.00	0.00	0.00	-0.01	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.10
log	0.01	0.02	0.02	-0.53	0.13	0.16	0.20	0.65	0.10	0.10	0.20	0.75

Comparison with other jackknife bias corrections

Finally, we compare the jackknife suggested in this paper to previous suggestions. Specifically, Cruz-Gonzalez et al. (2017) suggest two jackknife bias corrections for network models of the type considered in this paper.[9] The first bias correction is based on a split-sample approach. Divide the agents into two halves, A_1 and A_2. Define $\widehat{\beta}_{\alpha,\gamma/2} = \frac{1}{2}(\widehat{\beta}_{\alpha,\gamma\in A_1} + \widehat{\beta}_{\alpha,\gamma\in A_2})$, where $\widehat{\beta}_{\alpha,\gamma\in A_1}$ is the estimator that uses only observations in which the receiver is in the first set of agents A_1, and $\widehat{\beta}_{\alpha,\gamma\in A_2}$ uses only observations in which the receiver is in A_2. Similarly, define $\widehat{\beta}_{\alpha/2,\gamma} = \frac{1}{2}(\widehat{\beta}_{\alpha\in A_1,\gamma} + \widehat{\beta}_{\alpha\in A_2,\gamma})$ as the average of the two estimators that split the sample based on sending agents. A split-sample jackknife is given by

$$\widehat{\beta}_{ss} = 3\widehat{\beta}_N - \widehat{\beta}_{\alpha/2,\gamma} - \widehat{\beta}_{\alpha,\gamma/2}.$$

The second bias correction is based on dropping all observations associated with a particular agent. Let $\widehat{\beta}_{(i)}$ be the estimate using only observations in the sub-network that excludes agent i. Cruz-Gonzalez et al. (2017) define the 'double' correction as

$$\widehat{\beta}_d = N\widehat{\beta}_N - (N-1)\frac{1}{N}\sum_i \widehat{\beta}_{(i)}.$$

Both jackknife corrections differ from $\widehat{\beta}_J$ in (1.8), only $\widehat{\beta}_J$ preserves the distribution of fixed effects in each leave-out estimate. This appears to be an important property for the jackknife to perform well in settings with substantial unobserved heterogeneity. Table 1.9 reports results from simulations of the same model discussed above for the standard and weighted jackknife estimators as well as the split-sample and double corrections suggested by Cruz-Gonzalez et al. (2017). As is clear from the results, although the split-sample and double corrections appear to work well in the densest network DGP, they do not perform well in settings with more heterogeneity and sparser networks. In particular, the split-sample correction has much larger variance, and removes less bias than the leave-one-out style corrections (Hughes and Hahn (2020) derive higher-order bias and variance expressions that explain this phenomenon in the panel setting).

[9] Cruz-Gonzalez et al. (2017) suggest these jackknife corrections for the network model although do not prove their validity. Establishing the validity of the jackknife corrections requires the higher-order asymptotic expansions that are derived in this paper.

Table 1.9: Comparison of jackknife corrections

	Bias (mean)				SD				Rejection (5%)			
	$\hat{\theta}_J$	$\hat{\theta}_{wJ}$	$\widehat{\theta}_d$	$\widehat{\theta}_{ss}$	$\hat{\theta}_J$	$\hat{\theta}_{wJ}$	$\widehat{\theta}_d$	$\widehat{\theta}_{ss}$	$\hat{\theta}_J$	$\hat{\theta}_{wJ}$	$\widehat{\theta}_d$	$\widehat{\theta}_{ss}$
bal	-0.01	0.00	0.00	-0.02	0.04	0.04	0.04	0.04	0.03	0.03	0.05	0.06
llog	-0.01	0.00	-0.02	-0.04	0.06	0.05	0.05	0.10	0.04	0.04	0.03	0.11
slog	-0.03	0.00	-0.09	-0.26	0.11	0.09	0.25	0.31	0.06	0.05	0.07	0.43
log	-1.23	0.05	-1.18	-1.51	1.38	0.47	1.55	1.42	0.25	0.28	0.64	0.61

[a] $\widehat{\theta}_J$ and $\widehat{\theta}_{wJ}$ are the jackknife and weighted jackknife proposed in this paper. $\widehat{\theta}_d$ and $\widehat{\theta}_{ss}$ are the 'double' and 'ss2' jackknife estimators proposed in Cruz-Gonzalez et al. (2017).

1.7 Conclusion

This paper presents a new method for bias correcting nonlinear dyadic network models with fixed effects. We provide a novel formulation of the jackknife method that applies to networks with both sender and receiver fixed effects. The jackknife method provides an 'off-the-shelf' procedure for bias correction that is easy to apply, and applicable to a wide set of models. It allows for discrete multivalued and continuous outcome variables, and is able to obtain estimates of average effects and counterfactual outcomes.

In addition, we show how the jackknife can be used to bias correct averages of functions that depend on multiple observations, including dyads, triads, and tetrads in the network. These averages can be used to produce a wide array of test statistics for the presence of strategic interactions in the network, such as reciprocity or transitivity. In simulations, we show that the jackknife performs well, even in relatively low density networks, and outperforms previous suggestions for jackknife procedures.

There are a number of interesting areas in which the work in this paper might be usefully extended. The jackknife procedure developed in this paper applies to data on a single observation of a network. It would be interesting to extend these results to networks observed over multiple time periods, perhaps with the addition of time fixed effects. It is expected that a similar jackknife procedure could also be used in this setting, with appropriate splitting across the time dimension to account for dynamics in the network. The jackknife procedures proposed in this paper may also be useful in the interactive fixed effect model of Chen et al. (2021), and establishing their validity in this setting would also be useful.

Chapter 2

The Higher-order Variance of Bias-corrected Panel Estimators

2.1 Introduction

Panel data allow researchers to control for unobserved but time-invariant individual heterogeneity, which may otherwise confound inference on other explanatory variables. Methods to control for such unobserved heterogeneity are well established (see for example Chamberlain (1984), Arellano and Honoré (2001) for reviews in the linear case, and Arellano and Hahn (2010) for a review on nonlinear panel methods). One approach is to allow for individual fixed effects in the model; unfortunately, these estimators are typically subject to the incidental parameters problem described by Neyman and Scott (1948). In some cases, such as static linear or logit models, fixed-T consistent estimators are available (Andersen, 1970). However, this type of consistency is often impossible, notably in settings with dynamic effects or in nonlinear models (see Chamberlain, 2010). In general, the best that can be achieved in a fixed-T setting is partial identification; this is especially true for policy relevant parameters such as average marginal effects (see Chernozhukov et al., 2013).

Even under sequences in which T grows at the same rate as n, parameters may be asymptotically biased, as shown by Hahn and Kuersteiner (2002) for the dynamic linear model, and Hahn and Newey (2004) for the static nonlinear model. Given the typical size of panel datasets, in which n is much larger than T, it is desirable to find estimators that have biases of order $O(T^{-2})$ or smaller, rather than the typical $O(T^{-1})$.

One method of obtaining such estimators is to bias correct using the jackknife formula of

Quenouille (1956) and Tukey (1958). For a static model, Hahn and Newey (2004) show that a 'leave-one-out' jackknife estimator is asymptotically normal and centered at the truth when n and T grow at the same rate. Other styles of jackknife bias correction are also possible, for example, Dhaene and Jochmans (2015) suggest a split-sample bias correction that, in its simplest form, is constructed by splitting the sample into two half-panels of length $T/2$. The split-sample bias corrections are successful in both i.i.d. and dynamic panel settings. While other methods of bias correction exist, jackknife type corrections are attractive since they are typically easy to construct and do not rely on the derivation of analytical expressions for the bias. Importantly, both the 'leave-one-out' jackknife and split-sample bias corrections do not affect the asymptotic variance of the estimator, and so both have the same asymptotic distributions. Despite this, the suitability of these asymptotic approximations for inference is far from guaranteed. It is the goal of this paper to compare the two bias corrections in terms of their higher-order efficiency.

In the cross-sectional setting, Hahn et al. (2002) derive higher-order variance expressions for a range of bias-corrected MLEs. The authors argue that any asymptotically unbiased and efficient estimators with equal first- and second-order expansion terms (i.e. up to terms of order $O(n^{-1})$) must have equal higher order variances. Several bias corrections, including the jackknife, bootstrap, and analytic corrections, are shown to affect only the third-order $O(n^{-3/2})$ expansion term and hence be higher-order equivalent. This includes the estimator of Pfanzagl and Wefelmeyer (1978), implying that these bias-corrected estimators are also higher-order efficient.

In this paper, we derive similar higher-order variance expressions for jackknife and split-sample bias corrections of a panel maximum likelihood estimator with individual fixed effects. While both bias corrections share the same asymptotic variance, the jackknife correction is shown to have strictly smaller higher-order variance. Intuition for this result is given by showing that while the jackknife correction affects the third-order $O(T^{-3/2})$ expansion term, the split-sample correction has a second-order $O(T^{-1})$ effect.

In addition, although both estimators remove the $O(T^{-1})$ bias term, and so are asymptotically unbiased, the remaining $O(T^{-2})$ bias is shown to be larger for the split-sample correction. This suggest that the jackknife correction is likely to have smaller finite sample bias, particularly when T is small. Although we focus on the maximum likelihood setting, the results are applicable to a broader set of moment condition estimators under suitable assumptions on the moment functions. Simulations of a probit model with one common parameter and individual fixed effects support the theory. The split-sample

corrected estimator for the common parameter is shown to have much larger variance than the jackknife, and more bias. This comparison is also true of estimates of a marginal effect parameter.

In Section 2 we introduce the panel model, some assumptions, and discuss the jackknife and split-sample bias correction methods. Section 3 presents the main results of the paper, comparing the higher-order variance expressions for the two methods. In Section 4, we discuss the remaining bias of the estimators, while Section 5 extends the results to bias-corrected estimation of average fixed effects. Finally, Section 6 contains Monte Carlo evidence for the usefulness of the higher-order variance comparisons.

2.2 Nonlinear fixed effects models and jackknife bias correction

2.2.1 The fixed effects model

We begin with a description of the fixed effects maximum likelihood estimator. Let z_{it}, for $i = 1, \ldots, n$ and $t = 1, \ldots, T$, be a vector of observed data. Denote θ a $p \times 1$ parameter vector and α_i a scalar unobserved individual effect.[1] The data have density function $f(z|\theta, \alpha)$ with respect to some measure, and so (treating the α_i as parameters to be estimated) we may estimate θ via maximum likelihood. Assuming that the z_{it} are independent across both i and t, the MLE solves

$$\hat{\theta}_T \equiv \arg\max_{\theta} \sum_{i=1}^{n} \sum_{t=1}^{T} \ln f(z_{it}|\theta, \hat{\alpha}_i(\theta))$$
$$\hat{\alpha}_i(\theta) \equiv \arg\max_{\alpha} \sum_{t=1}^{T} \ln f(z_{it}|\theta, \alpha)$$

The incidental parameters problem is well known in this setting ((Neyman and Scott, 1948)). For fixed T, the probability limit of the estimator $\theta_T = p\lim_{n\to\infty} \hat{\theta}_T$ is given by

$$\arg\max_{\theta} \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} E\big[\sum_{t=1}^{T} \ln f(z_{it}|\theta, \hat{\alpha}_i(\theta))\big]$$

[1] The results in the paper and their proofs are presented for $p = 1$ for notational simplicity, but should be expected to hold for any finite $p > 1$.

which differs from θ_0 since each $\hat{\alpha}_i(\theta)$ depends only on T observations and so will differ from α_i. The estimation of the fixed effects $\hat{\alpha}_i(\theta)$ leads to three sources of bias in the common parameter: (i) the $\hat{\alpha}_i(\theta)$ are themselves biased, (ii) there exists correlation between $\hat{\theta}$ and $\hat{\alpha}_i(\theta)$ due to the fact that both are estimated using the same data, and (iii) the estimator is a nonlinear function of the $\hat{\alpha}_i(\theta)$, so that the variance of $\hat{\alpha}_i(\theta)$ also becomes bias.

As shown by Hahn and Newey (2004), even for $n/T \to \rho$ the estimator $\hat{\theta}$ will remain asymptotically biased, i.e. $\sqrt{nT}(\hat{\theta} - \theta_0) \Rightarrow N(\sqrt{\rho}\mathbf{B}, \Omega)$ for some bias term $\mathbf{B}$. Although the bias is of order $O(T^{-1})$, so that the estimator is still consistent, the bias may be substantial, particularly in applications where T is small.

For a broader interpretation of the results, it is useful to think about θ_0 and α_i as solutions to a set of moment equations given by the score functions

$$0 = \sum_{i=1}^{n} E\big[\frac{\partial}{\partial \theta} \ln f(z_{it}|\theta_0, \alpha_i)\big]$$
$$0 = E\big[\frac{\partial}{\partial \alpha_i} \ln f(z_{it}|\theta_0, \alpha_i)\big]$$

The asymptotic expansions presented below are based on expansions of these first order conditions, and hence the results are likely to apply to other moment estimators, under regularity conditions on the smoothness of the moment functions. We may also consider other quantities of interest that can be defined via some moment condition, for example an average effect parameter μ_0, defined as the solution to

$$0 = \mu_0 - \frac{1}{n} \sum_{i=1}^{n} E\big[m(z_{it}, \theta_0, \alpha_i)\big]$$

Stacking these moment conditions, we can define the common parameter to be (θ, μ), so that the results presented below will also apply to these types of parameters.

2.2.2 Bias corrections

An expansion of the fixed-T limit of $\hat{\theta}$ may be used to explain the success of the jackknife and split-sample bias corrections. Let the fixed-T limit of the estimator $\hat{\theta}$ have the expansion

$$\theta_T = \theta_0 + \frac{B}{T} + \frac{D}{T^2} + O(T^{-3})$$

This expansion suggests that a linear combination of two estimators using different numbers of time periods may be formed that removes the $O(T^{-1})$ bias. The 'leave-one-out' jackknife estimator is

$$\tilde{\theta}_J = T\hat{\theta} - (T-1)\frac{1}{T}\sum_{t=1}^{T}\hat{\theta}_{(t)}$$

where $\hat{\theta}_{(t)}$ is the estimator formed from the subsample that excludes time period t. More generally we may form a 'leave-k-out' jackknife as

$$\tilde{\theta}_{J,k} = \frac{T}{k}\hat{\theta} - \frac{T-k}{k}\bar{\theta}_k$$

where $\bar{\theta}_k = \binom{T}{k}^{-1}\sum_{\tau\in\mathcal{T}_k}\hat{\theta}_{(\tau)}$ for $\mathcal{T}_k$ the set of all k-period combinations from the possible T time periods, and $\hat{\theta}_{(\tau)}$ the estimator that excludes the k time periods in τ. Then, the fixed-T limit of the jackknife is

$$\frac{T}{k}\theta_T - \frac{T-k}{k}\theta_{T-k} = \theta_0 - \frac{1}{T(T-k)}D + O(T^{-3})$$

The bias of this estimator is $o(T^{-1})$ so long as $(T-k)\to\infty$, and in the case of the 'leave-one-out' jackknife of Hahn and Newey (2004), with $k=1$, the estimator has bias of order $O(T^{-2})$. As a consequence, the estimator will be asymptotically well centered so long as $n/T^3\to 0$. In the analysis below we will focus on the 'leave-one-out' jackknife, and refer to it as the 'full-sample jackknife', or simply 'the jackknife'.

The averaging over all possible leave-k-out estimates is not necessary for the bias correction properties to hold. We may instead use only a subset of these estimators in the construction of the jackknife correction. Dhaene and Jochmans (2015) propose the use of split-panel jackknifes that form weighted averages of estimates from subpanels of consecutive time periods. For example, take T to be even and divide the panel into two halves along the time dimension. Let $\hat{\theta}_1$ be the estimate using observations from the first half of time periods, $\hat{\theta}_2$ the estimate that uses the second half of time periods, and $\bar{\theta}_{1/2} = \frac{1}{2}(\hat{\theta}_1 + \hat{\theta}_2)$. The split-sample jackknife estimator is

$$\tilde{\theta}_{1/2} = 2\hat{\theta} - \bar{\theta}_{1/2}$$

and its fixed-T limit is

$$2\theta_T - \theta_{T/2} = \theta_0 - \frac{2}{T^2}D + O(T^{-3})$$

This bias-correction also reduces the bias to be of order $O(T^{-2})$. Other choices of split-sample jackknife are available; however, the results in Dhaene and Jochmans (2015) show that non-overlapping sub-panels in general have lower asymptotic variance, and that among the non-overlapping options, splitting in two leads to the smallest inflation of higher-order bias. Hence, in this paper we focus on this half sample version, and simply refer to it as the split-sample jackknife.

Note that this estimator is a form of 'leave-k-out' jackknife, with $k = T/2$, but one that does not average over all possible subsets of size $T/2$. Two observations are apparent here: (i) the dependence of the order of bias on k in the 'leave-k-out' jackknife suggests that smaller choices of k (e.g. $k = 1$) may be preferable in terms finite-sample bias reduction, and the choice $k = T/2$ may be less successful, particularly when T is quite small; and, (ii) we may expect that the averaging over all possible subsets in the 'leave-k-out' jackknife is beneficial in terms of variance, as it makes better use of the available data, suggesting that the half-sample jackknife may have larger finite sample variance. We demonstrate the validity of these statements, with particular emphasis on the relative efficiency of the two estimators.

It is important to note that Dhaene and Jochmans (2015) present the split-sample method in the context of dynamic panel models, where the restriction to using only subsets of consecutive time periods is important. In the non-i.i.d. context the bias terms in the probability limits used above depend on the time series structure of the data and so will differ for estimators that use different numbers of time periods as well as, crucially, estimators that use non-consecutive time periods. This will cause a 'leave-k-out' style jackknife to generally be unsuccessful in the dynamic case. The focus of this paper is on higher-order variance comparisons in the i.i.d. case, with the dynamic case left for future work.

2.2.3 Assumptions

Firstly, we provide some additional notation. Let

$$u_{it}(\theta, \alpha) = \frac{\partial}{\partial \theta} \ln f(z_{it}|\theta, \alpha)$$
$$V_{it}(\theta, \alpha) = \frac{\partial}{\partial \alpha_i} \ln f(z_{it}|\theta, \alpha)$$

be the score functions. When evaluating functions at the true value of parameters, arguments will be dropped, e.g. $u_{it} = u_{it}(\theta_0, \alpha_i)$. Further, let $U_{it}(\theta, \alpha) = u_{it}(\theta, \alpha) - \delta V_{it}(\theta, \alpha)$,

for $\delta = E[u_{it}V_{it}]/E[V_{it}^2]$, be the score for θ with the fixed effects concentrated out. All expectations are taken with respect to the distribution for an individual i, that is $E[h(z_{it})] = \int h(z)f(z|\theta,\alpha_i)dz$. Denote partial derivatives of these functions with superscripts, e.g. $\partial U_{it}/\partial\theta = U_{it}^{\theta}$.

The following assumptions are imposed throughout the paper and follow those used in Hahn and Newey (2004).

Assumption 2.1. *$n, T \to \infty$, with $n/T \to \rho$ for $0 < \rho < \infty$*

Assumption 2.2. *(i) The data z_{it} are independent over i and t and identically distributed over t according to the density $f(z|\theta,\alpha)$; (ii) the log density $\ln f(z|\theta,\alpha)$ is continuous in both θ and α; (iii) there exists a function $M(z_{it})$ such that $|\ln f(z_{it}|\theta,\alpha_i)| \leq M(z_{it})$,*

$$\Big|\frac{\partial \ln f(z_{it}|\theta,\alpha_i)}{\partial(\theta,\alpha_i)}\Big| \leq M(z_{it})$$

and $\sup_i E[M(z_{it})^{33}] < \infty$.

Assumption 2.3. *For each $\eta > 0$,*

$$\inf_i \Big[G_i(\theta_0,\alpha_i) - \sup_{\{(\theta,\alpha):|(\theta,\alpha)-(\theta_0,\alpha_i)|>\eta\}} G_i(\theta,\alpha)\Big] > 0$$

where $G_i(\theta,\alpha) \equiv E[\ln f(z_{it}|\theta,\alpha)]$.

Assumption 2.4. *(i) There exists some $M(z_{it})$ such that*

$$\Big|\frac{\partial^{m_1+m_2} \ln f(z_{it}|\theta,\alpha)}{\partial\theta^{m_1}\partial\alpha^{m_2}}\Big| \leq M(z_{it})$$

for $0 \leq m_1 + m_2 \leq 7$, and $\sup_i E[M(z_{it})^Q] < \infty$ for some $Q > 64$;
(ii) $\lim_{n\to\infty} \mathcal{I}_n > 0$, where $\mathcal{I}_n \equiv \frac{1}{n}\sum_i E[U_{it}^2]$;
(iii) $\min_i E[V_{it}^2] > 0$

2.3 Higher-order variances

Before we can describe the higher-order variances for each of our estimators, we must discuss the asymptotic expansions on which they are based. It is shown in the Appendix that the

MLE $\hat{\theta}$ has an expansion of the form

$$\sqrt{nT}(\hat{\theta} - \theta_0) = \sqrt{n}A_n + \frac{\sqrt{n}}{\sqrt{T}}B_n + \frac{\sqrt{n}}{T}C_n + O_p(T^{-1}) \tag{2.1}$$

where the expansions terms satisfy $Var(\sqrt{n}A_n) = O(1)$, $Var(\sqrt{n}B_n) = O(1)$, and $Var(\sqrt{n}C_n) = O(1)$. The first-order term in the expansion is given by

$$\sqrt{n}A_n = \mathcal{I}_n^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}U_{it}$$

This is simply the influence function for the MLE $\hat{\theta}$, and is mean zero with variance $\mathcal{I}_n^{-1}$, the inverse of the information for θ. The second-order term is in general not mean zero, and under asymptotic sequences in which $n/T \to \rho$, the $O(T^{-1})$ bias of the MLE is given by $\sqrt{\rho}\mathbf{B}$, where $\mathbf{B} = \lim_{n\to\infty} E[B_n]$.

We may write similar expansions for the bias-corrected estimators

$$\sqrt{nT}(\tilde{\theta}_J - \theta_0) = \sqrt{n}\tilde{A}_{n,J} + \frac{\sqrt{n}}{\sqrt{T}}\tilde{B}_{n,J} + \frac{\sqrt{n}}{T}\tilde{C}_{n,J} + O_p(T^{-1})$$
$$\sqrt{nT}(\tilde{\theta}_{1/2} - \theta_0) = \sqrt{n}\tilde{A}_{n,1/2} + \frac{\sqrt{n}}{\sqrt{T}}\tilde{B}_{n,1/2} + \frac{\sqrt{n}}{T}\tilde{C}_{n,1/2} + O_p(T^{-1})$$

Both the jackknife and split-sample corrections have no impact on the first-order term in the expansion, so that $\tilde{A}_{n,J} = \tilde{A}_{n,1/2} = A_n$. This implies that both bias-corrected estimators have the same asymptotic variance, equal to $\lim_{n\to\infty} Var(\sqrt{n}A_n) = \lim_{n\to\infty} \mathcal{I}_n^{-1}$. The remaining terms in the expansion differ for the different estimators, although for both we have $E[\tilde{B}_{n,J}] = E[\tilde{B}_{n,1/2}] = 0$, so that both estimators are asymptotically unbiased.

To compute the higher-order variance for the estimators, we take the variance of the first three expansions terms, retaining terms up to $O(T^{-1})$. Let the variance of these terms, for some estimator $\tilde{\theta}$ be

$$Var\Big(\sqrt{n}\tilde{A}_n + \frac{\sqrt{n}}{\sqrt{T}}\tilde{B}_n + \frac{\sqrt{n}}{T}\tilde{C}_n\Big) \equiv V_{1,n} + \frac{1}{T}V_{2,n} + R_{n,T}$$

where

$$V_{1,n} \equiv Var(\sqrt{n}\tilde{A}_n)$$
$$V_{2,n} \equiv Var(\sqrt{n}\tilde{B}_n) + 2Cov\big(\sqrt{n}\tilde{A}_n, \sqrt{nT}\tilde{B}_n + \sqrt{n}\tilde{C}_n\big)$$

$$R_{n,T} \equiv \frac{1}{T^2} Var(\sqrt{n}\tilde{C}_n) = o(T^{-1})$$

The higher-order variance of $\tilde{\theta}$ is then defined as $V_n^* \equiv V_{1,n} + \frac{1}{T}V_{2,n}$. This higher-order variance concept was introduced by Nagar (1959) and is common in comparisons of the properties of estimators. As discussed in Rothenberg (1984), under some regularity conditions, rankings based on this higher-order variance definition will correspond to rankings based on the variance of an Edgeworth approximation. Given that the first-order term for both bias-corrected estimators is the same, comparisons between the jackknife and split-sample corrections will depend on their respective higher-order terms. The following theorem shows that this term is around twice as large for the split-sample bias-correction compared with the jackknife correction.

Theorem 2.1. *Let V_J^* and $V_{1/2}^*$ be the higher-order variances of the jackknife and the split-sample bias-corrected estimators. Then these variances may be written as*

$$V_J^* = V_{1,n} + \frac{1}{T-1}V_{2,n}$$
$$V_{1/2}^* = V_{1,n} + \frac{2}{T}V_{2,n}$$

where

$$V_{1,n} = \mathcal{I}_n^{-1}$$
$$V_{2,n} = \mathcal{I}_n^{-2}\frac{1}{n}\sum_i \frac{1}{E[V_{it}^2]^2}\Big(\frac{1}{2}E[U_{it}^{\alpha\alpha}]^2 + 2E[U_{it}^{\alpha\alpha}]E[V_{it}U_{it}^{\alpha}] + E[V_{it}^2]E[(U_{it}^{\alpha})^2] + E[V_{it}U_{it}^{\alpha}]^2\Big) + o(1)$$

2.3.1 Understanding the effect of bias correction

We will attempt to provide some intuition for the relative performance of the two bias corrections; the full proof of Theorem 2.1 is provided in the Appendix. Intuitively, the result follows from the fact that, while the jackknife averages over all possible leave-one-out estimates, the split-sample correction uses only two of the $\binom{T}{2}$ possible $T/2$-period estimates. The impact of this on the higher-order variances of the two bias corrections is explained by

how they impact the second-order term in the expansion

$$\hat{\theta} \approx \theta_0 + \frac{1}{\sqrt{T}} A_n + \frac{1}{T} B_n + \frac{1}{T\sqrt{T}} C_n$$

As shown in the Appendix, the second-order term $\frac{1}{T} B_n$ is made up of V-statistics of the form

$$W_n = \frac{1}{n} \sum_i \frac{1}{T^2} \sum_{s,t} k_1(z_{is}) k_2(z_{it})$$

where $k_1(z_{it})$ and $k_2(z_{it})$ are both mean zero functions of single observations. It is these terms that contribute to the bias of the MLE, since for $s = t$ we have $E[k_1(z_{it})k_2(z_{it})] \neq 0$ in general. The jackknife removes the terms for which $s = t$, so that $\frac{1}{T}\tilde{B}_{n,J}$ contains U-statistics of the form

$$\begin{aligned} W_{n,J} &= TW_n - (T-1)\frac{1}{T}\sum_t \frac{1}{n}\sum_i \frac{1}{(T-1)^2} \sum_{(s_1,s_2)\neq t} k_1(z_{is_1})k_2(z_{is_2}) \\ &= \frac{1}{n}\sum_i \frac{1}{T(T-1)} \sum_{s \neq t} k_1(z_{it})k_2(z_{it}) \end{aligned}$$

In fact, this is true for the general 'leave-k-out' formulation of the jackknife as well so that the higher-order variance is the same for any choice of k with $T - k = O(T)$. In contrast, the split-sample jackknife removes all terms where s and t are in the same half of time periods so that $\frac{1}{T}\tilde{B}_{n,1/2}$ contains elements of the form

$$\begin{aligned} W_{n,1/2} &= 2W_n - \frac{1}{2}\Big(\frac{1}{n}\sum_i \frac{1}{(T/2)^2} \sum_{(s,t)\leq \frac{1}{2}T} k_1(z_{is})k_2(z_{it}) \\ &\qquad + \frac{1}{n}\sum_i \frac{1}{(T/2)^2} \sum_{(s,t)> \frac{1}{2}T} k_1(z_{is})k_2(z_{it})\Big) \\ &= 2 \times \frac{1}{n}\sum_i \frac{1}{T^2} \sum_{s\leq \frac{1}{2}T} \sum_{t> \frac{1}{2}T} \big(k_1(z_{is})k_2(z_{it}) + k_1(z_{it})k_2(z_{is})\big) \end{aligned}$$

Considering the variances of these two terms, it is then straightforward to see that $Var(\sqrt{nT}W_{n,1/2}) = 2\frac{T-1}{T} Var(\sqrt{nT}W_{n,J})$, which explains the result in Theorem 2.1.

As well as impacting the variance of the second-order terms, bias-correction also has an impact on the covariances between terms. As shown above, the first-order term $\frac{1}{\sqrt{T}} A_n$ is a sample average of score functions $\mathcal{I}_n^{-1} U_{it}$. Since $E[U_{ir}k_1(z_{is})k_2(z_{it})] = 0$ whenever $s \neq t$,

bias correction also results in the second-order expansion terms being uncorrelated with A_n. So, while the higher-order variance of the MLE $\hat{\theta}$ contains terms from $Cov(A_n, B_n)$, the higher-order variances for the bias-corrected estimators do not, since $Cov(A_n, \tilde{B}_{J,n}) = Cov(A_n, \tilde{B}_{1/2,n}) = 0$.

We can similarly examine the impact of bias correction on the third order term $T^{-3/2}C_n$, which can be shown to contain statistics of the form $W_n = \frac{1}{n}\sum_i \frac{1}{T^3}\sum_{r,s,t} k_1(z_{ir})k_2(z_{is})k_3(z_{it})$. Bias-correction does not completely remove the covariance of these terms with A_n; however, it does result in $Cov(\sqrt{n}A_n, \sqrt{n}\tilde{C}_n) = o(1)$ for both types of correction, so that these covariances no longer impact the higher order variance.

2.3.2 Accuracy in estimating the bias

One alternative way of understanding the differences in the higher-order variances of the estimators is by considering the accuracy with which each estimates the bias term. The form of the bias for nonlinear panel estimators was derived in Hahn and Newey (2004) for the i.i.d. setting, and is given by the expectation of the second-order term in (2.1). Each of the bias-corrected estimators implicitly subtracts an estimate of this bias from the MLE $\hat{\theta}$, resulting in an estimator of the form $\tilde{\theta} = \hat{\theta} - \frac{1}{T}\hat{\beta}$.

For the 'leave-one-out' jackknife,

$$\frac{1}{T}\hat{\beta}_J = (T-1)\big(\frac{1}{T}\sum_{t=1}^{T}\hat{\theta}_{(t)} - \hat{\theta}\big)$$

while the split-sample jackknife uses the bias estimate

$$\frac{1}{T}\hat{\beta}_{1/2} = (\bar{\theta}_{1/2} - \hat{\theta})$$

Recalling that $\mathbf{B} = \lim_{n\to\infty} E[B_n]$, the following theorem establishes the accuracy of $\hat{\beta}_J$ and $\hat{\beta}_{1/2}$ as estimators for $\mathbf{B}$.

Theorem 2.2. *Let $\hat{\beta}_J = (T-1)(\frac{1}{T}\sum_{t=1}^{T}\hat{\theta}_{(t)} - \hat{\theta})$ and $\hat{\beta}_{1/2} = (\bar{\theta}_{1/2} - \hat{\theta})$ be the jackknife and split-sample estimators for the bias term $\mathbf{B}$. Then,*

$$\sqrt{nT}\frac{1}{T}(\hat{\beta}_J - \mathbf{B}) = O_p(T^{-1})$$

and

$$\sqrt{nT}\frac{1}{T}(\hat{\beta}_{1/2} - \mathbf{B}) = O_p(T^{-1/2})$$

From Theorem 2.2, we could alternatively write the expansion for the jackknife bias corrected estimator as

$$\begin{aligned}\sqrt{nT}(\tilde{\theta}_J - \theta_0) = \sqrt{n}A_n + \frac{1}{\sqrt{T}}\sqrt{n}(B_n - \mathbf{B}) \\ + \frac{1}{T}\big(\sqrt{n}C_n - \sqrt{nT}(\hat{\beta}_J - \mathbf{B})\big) + O_p(T^{-1})\end{aligned}$$

and for the split-sample bias corrected estimator as

$$\begin{aligned}\sqrt{nT}(\tilde{\theta}_{1/2} - \theta_0) = \sqrt{n}A_n + \frac{1}{\sqrt{T}}\big(\sqrt{n}(B_n - \mathbf{B}) - \sqrt{n}(\hat{\beta}_{1/2} - \mathbf{B})\big) \\ + \frac{1}{T}\sqrt{n}C_n + O_p(T^{-1})\end{aligned}$$

The key difference between the two estimators is that the jackknife bias correction affects the third-order, $O_p(T^{-1})$, part of the expansion, while the split-sample bias correction appears as a second-order, $O_p(T^{-1/2})$, term. This implies that the jackknife bias estimate will only impact the higher-order variance through its covariance with the first-order term A_n, i.e. through the term $Cov\big(\sqrt{n}A_{n,J}, \sqrt{nT}(\hat{\beta}_J - \mathbf{B})\big)$. In contrast, the split-sample bias estimate appears in the higher-order variance both through its covariance with A_n, as well as through its own variance, $Var\big(\sqrt{n}(\hat{\beta}_{1/2} - \mathbf{B})\big)$. In fact, it is straightforward to show that $Var\big(\sqrt{n}(\hat{\beta}_{1/2} - \mathbf{B})\big) = Var\big(\sqrt{n}(B_n - \mathbf{B})\big) + o(1)$, while the covariance between these two terms is also $o(1)$, which explains the form of the higher-order variance in Theorem 2.1. It is this extra variance term that accounts for the loss in efficiency of the split-sample correction relative to the jackknife.

2.4 Higher-order bias

It is also worth highlighting the effect of the two bias corrections on the $O(T^{-2})$ bias of the estimators, which remains even after bias correction. To clarify this, it is useful to extend the expansion in (2.1) to include a further term

$$\sqrt{nT}(\hat{\theta} - \theta_0) = \sqrt{n}A_n + \frac{\sqrt{n}}{\sqrt{T}}B_n + \frac{\sqrt{n}}{T}C_n + \frac{\sqrt{n}}{T\sqrt{T}}D_n + o_p(T^{-1})$$

where again, $Var(\sqrt{n}D_n) = O(1)$. Using this expansion, we can define $\frac{1}{T}\sqrt{\rho}\mathbf{D} = \frac{1}{T}\sqrt{\rho}\lim_{n\to\infty}\{\sqrt{T}E[C_n] + E[D_n]\}$ as the $O(T^{-2})$ bias. While both bias-corrections remove the asymptotic bias of the MLE, they do not remove the $O(T^{-2})$ bias, and in fact will inflate this term. The following theorem establishes the higher-order biases of the two bias-corrected estimators, relative to that of the MLE $\hat{\theta}$.

Theorem 2.3. *Let $\tilde{\theta}_J$ and $\tilde{\theta}_{1/2}$ be the jackknife and split-sample bias-corrected estimates. The $O(T^{-2})$ biases of these estimators are $\frac{1}{T}\sqrt{\rho}\mathbf{D}_J$ and $\frac{1}{T}\sqrt{\rho}\mathbf{D}_{1/2}$ respectively, where*

$$\mathbf{D}_J = \lim_{n\to\infty}\{\sqrt{T}E[\tilde{C}_{n,J}] + E[\tilde{D}_{n,J}]\} = -\frac{T}{T-1}\mathbf{D} + o(1)$$
$$\mathbf{D}_{1/2} = \lim_{n\to\infty}\{\sqrt{T}E[\tilde{C}_{n,1/2}] + E[\tilde{D}_{n,1/2}]\} = -2\mathbf{D} + o(1)$$

The jackknife correction inflates the $O(T^{-2})$ bias by a factor of $\frac{T}{T-1}$, while in the split-sample case the bias is inflated by a factor of 2, leading to strictly larger bias for $T > 2$. As shown by Dhaene and Jochmans (2015), this factor of 2 is the smallest inflation of higher-order bias among a variety of forms of split-sample correction, for example, splits into more than two subsets and/or splits into subsets of unequal length. It can also be shown that the general leave-k-out jackknife inflates the $O(T^{-2})$ bias by a factor of $\frac{T}{T-k}$, so that the choice $k = 1$ has the smallest higher-order bias among this class.[2]

2.5 Fixed effect averages

Researchers are often interested in averages over some function of the parameters rather than the parameters themselves. We can easily extend the results above to such objects. Assume that the researcher is interested in estimating

$$\mu = \frac{1}{n}\sum_{i=1}^{n} E[m(z_{it}, \theta, \alpha_i)]$$

[2]It is of course possible to remove this $O(T^{-2})$ bias also. For example, the average of all leave-k-out and all leave-j-out estimates (with $j < k$) can be combined with the MLE to give $\tilde{\theta}_{2J} = \frac{T^2}{jk}\hat{\theta} - \frac{(T-j)^2}{j(k-j)}\bar{\theta}_j + \frac{(T-k)^2}{k(k-j)}\bar{\theta}_k$, which has bias of order $O(T^{-3})$. Similarly, Dhaene and Jochmans (2015) suggest $\tilde{\theta}_{1/(2,3)} = 3\hat{\theta} - 3\bar{\theta}_{1/2} + \bar{\theta}_{1/3}$, where $\bar{\theta}_{1/2}$ and $\bar{\theta}_{1/3}$ are the averages of the two half-sample and three third-sample estimates. The properties of these higher-order bias corrections, in particular their impact on the higher-order variance is a subject of future research.

for a known function m that satisfies the same conditions on smoothness and existence of moments that we have previously imposed on the score function. A plug-in estimator for this quantity is

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{T}\sum_{t=1}^{T} m(z_{it}, \hat{\theta}, \hat{\alpha}_i(\hat{\theta}))$$

As with estimation of θ, the estimator $\hat{\mu}$ is asymptotically biased, but may be bias corrected using either the jackknife or split-sample correction. It can be shown that the first-order bias of the plug-in MLE has the form[3]

$$\begin{aligned}\mathbf{B}_{n,\mu} = \frac{1}{2}\sqrt{\frac{n}{T}}\frac{1}{n}\sum_i \Big\{ &\frac{E[m_{it}^{\alpha_i\alpha_i}]}{E[V_{it}^2]} + E[m_{it}^{\alpha_i}]\frac{E[V_{it}^{\alpha_i\alpha_i}] + 2E[V_{it}V_{it}^{\alpha_i}]}{E[V_{it}^2]^2} \\ &+ \frac{2E[m_{it}^{\alpha_i}V_{it}]}{E[V_{it}^2]} + \big(E[m_{it}^{\theta}] + E[m_{it}^{\alpha_i}]\alpha_i^{\theta}\big)\mathbf{B}_n \Big\}\end{aligned}$$

The first three terms in the bias are related to the estimation of the fixed effects, while the final term comes from the biased estimation of the common parameter θ. The first term results from averaging over a nonlinear function of the fixed effects, and will be zero whenever m is linear in the fixed effects so that $m^{\alpha\alpha}$ is zero. The second term relates to the first-order bias of the estimated fixed effects, while the third term comes from the fact that the fixed effects are estimated using the same data that is used to estimate the time series average $\frac{1}{T}\sum_{t=1}^{T} m(z_{it}, \hat{\theta}, \hat{\alpha}_i(\hat{\theta}))$; this term will be zero whenever the function m does not depend on the data (i.e. has no argument z_{it}).

In the same way as above, bias-corrected estimators can be formed using the jackknife and split-sample techniques, i.e.

$$\tilde{\mu}_J = T\hat{\mu} - (T-1)\frac{1}{T}\sum_t \hat{\mu}_{(t)}$$

$$\tilde{\mu}_{1/2} = 2\hat{\mu} - \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2)$$

where $\hat{\mu}_{(t)}$ are the estimators that exclude period t, while $\hat{\mu}_1$ and $\hat{\mu}_2$ are the estimators that use only the first or second halves of time periods. By expressing μ as the solution to a moment equation, as was done with the first-order condition for θ, we may derive an

[3]This formula for the bias of averages over fixed effects was shown in Hahn and Newey (2004), but evaluated at the bias-corrected estimates $\tilde{\theta}$ and $\tilde{\alpha}_i$ rather than $\hat{\theta}$ and $\hat{\alpha}_i$, so that the final term was not present.

expansion for $\hat{\mu}$ (as well as its bias-corrected versions) of the same form as (2.1). Given this, the higher-order variance and bias comparisons presented for estimates of θ apply identically to this setting.

Theorem 2.4. *Let Σ_J^* and $\Sigma_{1/2}^*$ be the higher-order variances (as defined in Section 2.3) of the jackknife and the split-sample bias-corrected estimators for μ. Then these variances may be written as*

$$\Sigma_J^* = \Sigma_{1,n} + \frac{1}{T-1}\Sigma_{2,n}$$
$$\Sigma_{1/2}^* = \Sigma_{1,n} + \frac{2}{T}\Sigma_{2,n}$$

for some $O(1)$ terms $\Sigma_{1,n}$ and $\Sigma_{2,n}$. Furthermore, let the $O(T^{-2})$ bias (as defined in Section 2.4) of $\hat{\mu}$ be $\frac{1}{T}\mathbf{D}_\mu$. Then, the $O(T^{-2})$ biases of $\tilde{\mu}_J$ and $\tilde{\mu}_{1/2}$ may be written as $\frac{1}{T}\mathbf{D}_{\mu,J}$ and $\frac{1}{T}\mathbf{D}_{\mu,1/2}$ respectively, where

$$\mathbf{D}_{\mu,J} = -\frac{T}{T-1}\mathbf{D}_\mu + o(1) \quad \textit{and,}$$
$$\mathbf{D}_{\mu,1/2} = -2\mathbf{D}_\mu + o(1)$$

2.6 Examples

2.6.1 An analytical example

A simple example may help to highlight the results. Consider the estimation of a common variance parameter using observations made from normally distributed variables with differing means.[4]

$$\hat{\theta} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}(z_{it} - \bar{z}_i)^2, \quad z_{it} \sim N(\alpha_i, \theta)$$

Standard calculation gives that $E[\hat{\theta}] = \frac{T-1}{T}\theta$ so that the MLE has bias of $\frac{1}{T}\mathbf{B} = -\frac{1}{T}\theta$ (in this case there is only bias of order $O(T^{-1})$, and no higher-order biases). The jackknife

[4]This is a standard example of the incidental parameters problem and appeared as Example 2 in Neyman and Scott (1948).

bias-corrected estimator has the form

$$\tilde{\theta}_J = \frac{1}{n(T-1)} \sum_{i=1}^{n} \sum_{t=1}^{T} (z_{it} - \bar{z}_i)^2$$

while the split-sample estimator is

$$\begin{aligned}
\tilde{\theta}_{1/2} &= 2\frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} (z_{it} - \bar{z}_i)^2 \\
&\quad - \frac{1}{2}\Big(\frac{1}{nM} \sum_{i=1}^{n} \sum_{t=1}^{M} (z_{it} - \bar{z}_{i,1})^2 + \frac{1}{nM} \sum_{i=1}^{n} \sum_{t=M+1}^{T} (z_{it} - \bar{z}_{i,2})^2 \Big) \\
&= \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} (z_{it} - \bar{z}_i)^2 + \frac{1}{2n} \sum_{i=1}^{n} \Big((\bar{z}_{i,1} - \bar{z}_i)^2 + (\bar{z}_{i,2} - \bar{z}_i)^2 \Big)
\end{aligned}$$

where $T = 2M$, and $\bar{z}_{i,1}$ and $\bar{z}_{i,2}$ are the sample means in the first and second halves of time periods. In this simple model we have

$$\begin{aligned}
U_{it} &= -\frac{1}{2\theta} + \frac{1}{2\theta^2}(z_{it} - \alpha_i)^2 \\
V_{it} &= \frac{1}{2\theta}(z_{it} - \alpha_i)
\end{aligned}$$

and taking further derivatives and applying them to the formula given in Theorem 2.1 gives

$$\begin{aligned}
V_J^* &= \frac{2T}{T-1}\theta^2 \\
V_{1/2}^* &= \frac{2T+4}{T}\theta^2
\end{aligned}$$

We can easily confirm that these are also the exact finite sample variances of the two estimators in this case. Considering the estimation of the bias itself, we can see that the jackknife bias-correction estimates the bias as

$$\frac{1}{T}\hat{\beta}_J = -\frac{1}{T}\frac{1}{n(T-1)} \sum_{i=1}^{n} \sum_{t=1}^{T} (z_{it} - \bar{z}_i)^2$$

while the split-sample correction estimates the bias using

$$\frac{1}{T}\hat{\beta}_{1/2} = -\frac{1}{2n}\sum_{i=1}^{n}\Big((\bar{z}_{i,1}-\bar{z}_i)^2+(\bar{z}_{i,2}-\bar{z}_i)^2\Big)$$

Both are unbiased estimators of the bias term in this case, i.e. $E[\hat{\beta}_J] = E[\hat{\beta}_{1/2}] = -\theta$. Let us now examine the variances of the two bias estimates. Firstly, using the fact that $\sum_{t=1}^{T}(z_{it}-\bar{z}_i)^2/\theta \sim \chi^2_{T-1}$

$$\begin{aligned} Var\big(\sqrt{\frac{n}{T}}(\hat{\beta}_J - \mathbf{B})\big) &= Var\Big(\frac{1}{\sqrt{nT}(T-1)}\sum_{i=1}^{n}\sum_{t=1}^{T}(z_{it}-\bar{z}_i)^2\Big) \\ &= \frac{\theta^2}{T(T-1)^2}Var(\chi^2_{T-1}) \\ &= \frac{2\theta^2}{T(T-1)} \end{aligned}$$

since $Var(\chi^2_k) = 2k$. Next, since $\bar{z}_{i,1}$ and $\bar{z}_{i,2}$ are independent and distributed $N(\alpha_i, \theta/M)$, we have $\{(\bar{z}_{i,1}-\bar{z}_i)^2+(\bar{z}_{i,2}-\bar{z}_i)^2\}\frac{M}{\theta} \sim \chi^2_1$, so

$$\begin{aligned} Var\big(\sqrt{\frac{n}{T}}(\hat{\beta}_{1/2} - \mathbf{B})\big) &= Var\Big(\frac{\sqrt{nT}}{2n}\sum_{i=1}^{n}\big((\bar{z}_{i,1}-\bar{z}_i)^2+(\bar{z}_{i,2}-\bar{z}_i)^2\big)\Big) \\ &= \frac{T}{4}\big(\frac{\theta}{M}\big)^2 Var(\chi^2_1) \\ &= \frac{2\theta^2}{T} \end{aligned}$$

The variance of the split-sample bias estimate is larger by a T factor, as predicted by Theorem 2.2.

2.6.2 Monte Carlo analysis

Next, to highlight the relevance of the results in a more practical setting, we conduct a Monte Carlo exercise of a probit model with fixed effects for sample sizes $n = \{100, 200\}$

Table 2.1: Estimation of θ_0

Estimator	Bias	SD	10%	5%	Bias	SD	10%	5%
		$N = 100$, $T = 8$				$N = 200$, $T = 8$		
$\hat{\theta}$	0.193	0.152	0.414	0.288	0.180	0.104	0.589	0.473
$\tilde{\theta}_J$	−0.038	0.119	0.087	0.042	−0.041	0.082	0.102	0.054
$\tilde{\theta}_{1/2}$	−0.067	0.261	0.406	0.330	−0.070	0.175	0.409	0.326
		$N = 100$, $T = 12$				$N = 200$, $T = 12$		
$\hat{\theta}$	0.129	0.097	0.400	0.293	0.126	0.066	0.641	0.506
$\tilde{\theta}_J$	−0.022	0.081	0.080	0.036	−0.021	0.055	0.087	0.050
$\tilde{\theta}_{1/2}$	−0.043	0.171	0.402	0.318	−0.032	0.116	0.395	0.311

and $T = \{8, 12\}$. The design follows that used in Hahn and Newey (2004)[5]

$$
\begin{aligned}
y_{it} &= 1\{\theta_0 x_{it} + \alpha_i + \varepsilon_{it} > 0\}, \quad \alpha_i \sim N(0,1), \quad \varepsilon_{it} \sim N(0,1) \\
x_{it} &= t/10 + x_{i,t-1}/2 + u_{it}, \quad x_{i0} = u_{i0}, \quad u_{it} \sim U(-1/2, 1/2) \\
\theta_0 &= 1
\end{aligned}
$$

As noted in Hahn and Newey (2004), although y_{it} is independent over time conditional on x_{it} and α_i, the serial correlation and non-stationarity of x_{it} means that this DGP does not completely align with the i.i.d. framework presented in this paper. We nevertheless maintain it here as the design is used elsewhere in the literature (see Heckman (1981)), and the departure from the i.i.d. assumption on the regressor may be relevant in practice. Results from a similar design with i.i.d. x_{it} are reported in the Supplementary Appendix.

Table 2.1 reports the bias, standard deviation, and rejection rates for estimation of the common parameter θ_0. Rejection rates are computed from confidence intervals constructed using a Hessian-based estimator of the asymptotic variance. The estimators presented are: $\hat{\theta}$, the (biased) maximum likelihood estimate; $\hat{\theta}_J$, the jackknife bias-corrected estimate; and $\hat{\theta}_{1/2}$, the split-sample bias-corrected estimate.

From the first row in each section of Table 2.1, it is clear that the MLE for θ has a significant positive bias of about 18-19 per cent for $T = 8$, and 13 per cent for $T = 12$. Both the jackknife and split-sample estimators remove most, but not all of this bias. The split-sample estimator has a negative bias of around 7 and 3-4 per cent, for $T = \{8, 12\}$; in

[5]Hahn and Newey (2004) report results for $T = \{4, 8\}$. We chose to use larger T since the split-sample estimator did not perform well for $T = 4$; this was due to the (half sample) 2-period estimates resulting in large outliers in many cases.

comparison, the jackknife does slightly better with negative biases of just 4 and 2 per cent. These results are consistent with the ratio between the $O(T^{-2})$ biases of the estimators of $2\frac{T-1}{T}$ which was derived above – this is approximately 1.8 for the samples sizes used here.

As is evident from the theory, the cross-sectional sample size n does not affect the bias of the estimator, but does affect its variance. This results in worsening coverage as n increases for a fixed T, as the bias of the estimators becomes larger relative to its dispersion. The jackknife bias correction substantially improves coverage of the estimators, as was noted by Hahn and Newey (2004). This is largely the result of proper centering of the confidence intervals, but coverage is also improved by a small reduction in the variance of estimates. This could be explained by the fact that the bias-correction removes the covariance between the first and second-order terms in the expansion, and reduces the order of the covariance between the first and third-order terms (see Section 2.3.1).[6] In comparison, the half-sample bias correction increases the standard deviation of the estimator by around 75 per cent, resulting in substantially worse coverage. In fact, for $n = 100$, the inflation in the variance overrides the benefits of bias reduction and leads to worse coverage than for the biased MLE estimate. This suggests that bias correction need not result in a reduction in the MSE of an estimator, and highlights the fact that inference based on the asymptotic distribution of the split-sample estimator can be highly misleading.

Dhaene and Jochmans (2015) report results of the same Monte Carlo experiment using an alternative estimator for the asymptotic variance to form confidence intervals for the split-sample jackknife. Their variance estimate is formed by estimating the variance in each of the two half samples, and then averaging. This method results in estimates of the asymptotic variance that are significantly larger than the single full-sample estimate, and consequently leads to much better coverage for the split-sample jackknife – rejection rates of 7.1 per cent and 6.3 per cent are reported for a 5 per cent test, with $n = 100$ and T equal to 8 and 12 respectively. We choose to maintain a consistent estimate of the asymptotic variance across estimators, for the purpose of comparison between them.[7]

Table 2.2 reports the same simulations for an average marginal effect parameter. The

[6] It is not clear that these covariances need be positive or negative in general, and so this may not hold for other models.

[7] The high variance of the split-sample estimator appears to be partly due to the fact that the proportion of 'stayers' (individuals for whom we observe either zero or one in all time periods, and hence do not impact estimation of θ) is much larger in the half samples than in the full sample. This issue also inflates the asymptotic variance estimates using the half samples, which may explain the better coverage of this form of the variance estimate. It is not clear that this variance estimator will perform as well in general; for example, it does not help in the example of Section 2.6.1.

Table 2.2: Estimation of μ_0

Estimator	Bias	SD	10%	5%	Bias	SD	10%	5%
		$N = 100$, $T = 8$				$N = 200$, $T = 8$		
$\hat{\mu}$	0.006	0.023	0.141	0.072	0.004	0.016	0.128	0.078
$\hat{\mu}_J$	0.003	0.024	0.140	0.074	0.001	0.016	0.127	0.072
$\hat{\mu}_{1/2}$	0.030	0.041	0.511	0.424	0.026	0.028	0.556	0.481
		$N = 100$, $T = 12$				$N = 200$, $T = 12$		
$\hat{\mu}$	0.004	0.018	0.152	0.088	0.004	0.012	0.119	0.066
$\hat{\mu}_J$	0.000	0.018	0.130	0.071	0.000	0.012	0.105	0.057
$\hat{\mu}_{1/2}$	0.033	0.030	0.616	0.526	0.034	0.020	0.779	0.711

parameter of interest is

$$\mu_0 = \theta_0 \frac{1}{n} \sum_{i=1}^{n} \phi(\theta_0 + \alpha_i)$$

that is, the average derivative evaluated at $x = 1$. Since the object of interest is sample dependent, it varies across simulations, but is on average around 0.22.

As has been noted elsewhere, marginal effects estimates in probit models can have quite small bias (see Fernández-Val (2009)). The bias of the maximum likelihood estimates are only around 2-3 per cent, about one-quarter of the standard deviation of the estimates. The jackknife removes about half of this small bias for $T = 8$ and is unbiased for $T = 12$; however, the half-sample jackknife appears to significantly increase the bias in each case.[8] The standard deviation of the jackknife is the same as that of the MLE, while the half-sample estimator has much larger standard deviation, by around the same proportion as for estimates of θ_0. It is interesting to note here (as was also noted by Hahn and Newey (2004)) that the asymptotic variance appears to underestimate the true variance in the case of the marginal effect for $T = 8$, so that coverage is distorted even though bias is minimal; however, this is no longer the case for $T = 12$.

[8]The estimator of the marginal effect is subject to yet another source of bias from the proportion of 'stayers' in the sample, which cannot be used for estimation. As noted above, we speculate that this may affect the split sample estimator more than the regular jackknife estimator, since the split-sample makes use of estimates with fewer time periods, explaining the increased bias.

2.7 Conclusion

We derived expressions for the higher-order variances of the jackknife and split-sample bias-corrected MLEs for a nonlinear panel model. The split-sample bias-corrected estimator has larger higher-order variance, suggesting that its finite sample efficiency may be less than that of the jackknife, particularly in settings where the time series dimension is not large. Moreover, the remaining bias in the split-sample estimator is larger than that of the jackknife estimator. The analysis suggests that in an i.i.d. panel setting, the jackknife bias correction should be preferred. However, in non-i.i.d. settings the standard jackknife cannot be used, while the split-sample jackknife remains consistent. In this setting, our results suggest that inference based on the asymptotic distribution is likely to underestimate the finite sample variance (particularly for small T). As suggested by Dhaene and Jochmans (2015), establishing the validity of other forms of inference, such as the bootstrap, would be useful in these cases. It would be valuable to have similar higher-order variance comparisons available for the non-i.i.d. setting, to compare the split-sample correction with alternatives, such as the analytic correction given in Hahn and Kuersteiner (2002). We leave this to future research.

Chapter 3

Estimation of Linear IV Models with Many Endogenous Regressors

3.1 Introduction

Empirical researchers may wish to estimate models in which a number of variables are potentially endogenous. This occurs, for example, in settings where there are multiple treatments available, or where a single treatment may take a number of levels. Multiple endogenous regressors can also arise when a single endogenous regressor is interacted with exogenous covariates in order to account for, or learn about, heterogeneity in the treatment effect.

Although much is known about the properties of various linear IV estimators with many instruments, covariates, and many weak instruments, much less is known about how well they function when used to estimate models with multiple endogenous regressors. This paper extends existing results on linear IV estimation by considering asymptotics under which the number of endogenous regressors is allowed to grow with the sample size. I derive consistency and asymptotic normality results for the jackknife IV (JIVE) estimator of Angrist et al. (1999), as well as the heteroskedasticity robust k-class style estimators (including the HLIM and HFUL estimators of Hausman et al. (2009)).

Under a framework that allows for the number of endogenous variables to grow with the sample size, as well as the presence of many weak instruments, I derive consistency for the parameter vector, as well as asymptotic normality for a linear combination of the parameters. The normalizing rate depends on the particular linear combination considered,

suggesting that the quality of the asymptotic approximation can vary depending on the object of interest.

The paper also provides simulation evidence on the properties of these estimators under moderate numbers of endogenous regressors. It is shown that the JIVE can be unstable in the presence of multiple endogenous regressors and many instruments, exhibiting wide dispersion and severe over-rejection in the simulations. In contrast, the HFUL estimator performs well and appears relatively robust to both multiple endogenous regressors and many weak instruments.

A considerable amount of research has investigated the properties of instrumental variables (IV) estimators in situations where there are many potential instruments or the instruments may only be weakly relevant for the endogenous regressors. Two-stage least squares (TSLS) is known to be inconsistent in these settings, and a number of more robust estimators have been studied. This paper analyzes the properties of some of these estimators when the number of endogenous regressors may also be large.

The literature on weak instruments uses asymptotic approximations that consider sequences of models in which the coefficients in the first stage regression are modeled as 'local-to-zero', i.e. they lie in a shrinking $n^{-1/2}$-neighborhood of zero (where the number of instruments is held fixed). For example, Staiger and Stock (1997) show that in this setting two-stage least squares (2SLS), limited information maximum likelihood (LIML), and other standard estimators are inconsistent. In linear IV models, the concentration parameter turns out to be a useful measure of the strength of instruments, and the inconsistency of estimators under local-to-zero asymptotics is reflected in the fact that the concentration parameter does not grow to infinity as the sample size grows, as it does under standard asymptotics.

A related strand of the literature has considered the problem of IV regression in the presence of 'many instruments'. These situations, in which the number of available instruments, K, may be large relative to sample size, are typically modeled by allowing K to grow to infinity (Morimune (1983), Bekker (1994)) with the sample size n. It can be shown that, under asymptotics in which K grows at the same rate as n, the 2SLS estimator is biased, but that LIML (as well as some other Fuller style k-class estimators) and jackknife IV estimators remain consistent. One may also consider asymptotics under which many weak instruments are available, as has been done for example by Chao and Swanson (2004), Chao and Swanson (2005), and Stock and Yogo (2005). When there are many instruments, it is possible for the concentration parameter to diverge, even when the individual instruments

are weak in the Staiger and Stock (1997) sense. Under homoskedasticity, LIML remains consistent under many weak instrument asymptotics, so long as the concentration parameter divided by $\sqrt{K}$ diverges (Chao and Swanson, 2005). Chao and Swanson (2004) and (Chao et al., 2012) show consistency of the JIVE estimator, under the same condition, but allowing for heteroskedasticity, while Hausman et al. (2009) introduce jackknife versions of the LIML and Fuller estimators that are also robust to heteroskedasticity. Mikusheva and Sun (2020) show that the concentration parameter divided by $\sqrt{K}$ growing to infinity is in fact necessary for a consistent test to exist, and develop an inference procedure that is valid regardless of identification strength (and consistent when concentration parameter divided by $\sqrt{K}$ diverges). Estimators based on moment conditions have also been considered in this setting by, for example, Han and Phillips (2006) and Newey and Windmeijer (2009).

This paper extends the many weak instrument results by considering asymptotics under which the number of endogenous regressors, J, is also allowed to grow to infinity. There are numerous examples of empirical applications in which a researcher may be faced with a large number of endogenous regressors. These include linear demand systems, in which numerous own-price as well as cross-price elasticities are estimated; models with treatment effect heterogeneity, in which the treatment may not be randomly assigned and is also interacted with a large number of covariates; and, nonparametric estimators such as the series NPIV regression, in which the endogenous regressors are a set of flexible functions of a vector endogenous explanatory variables.

I derive conditions under which JIVE and the HLIM and HFUL estimators of Hausman et al. (2009) are consistent and certain finite linear combinations of the parameter vector are asymptotically normal. Following the many weak instruments literature, the concentration parameter is allowed to grow at a rate r_n, which may be slower than n (in a way made precise below), allowing for the instruments to be weak for a given set of endogenous variables. When instruments are strong, consistency requires that the both the number of endogenous variables J_n as well as the number of instruments K_n does not grow too fast, but does not place direct restrictions on the rate at which these terms grow relative to each other.

As in previous work, the condition $\sqrt{K_n}/r_n \to 0$ will be requires, and when the number of instruments grows proportionally to the concentration parameter, the asymptotic variance will be made of two separate terms. In order to bound certain denominator terms in the jackknife estimators, I make use of a matrix concentration inequality for degenerate U-statistics from (Minsker and Wei, 2019).

This remainder of the paper is set out as follows. Section 2 outlines the linear IV setting,

briefly discusses some applications, and provides a review of the jackknife estimators. Section 3 introduces the asymptotic sequence and assumption used in the paper. Section 4 provides the main results on jackknife consistency, as well as the consistency (inconsistency) of the 2SLS estimator. Section 5 concludes with some brief simulation analysis. Proofs of results are provided in the appendices.

3.2 The model

This paper considers estimation of the following linear model, with potentially many endogenous explanatory variables

$$y_i = X_i'\beta + \varepsilon_i \tag{3.1}$$

$$X_i = \Upsilon_i + V_i \tag{3.2}$$

Here the dimension of the vector of endogenous explanatory variables X_i is J_n, which is allowed to grow to infinity with the sample size n. The error terms are assumed to be mean zero conditional on the reduced form and observed instruments, i.e. $E[\varepsilon_i|\mathcal{Z}] = E[V_i|\mathcal{Z}] = 0$, where $\mathcal{Z} = (\Upsilon, Z)$. The reduced form vector Υ_i may be a linear combination of the instruments, i.e. $\Upsilon_i = \Pi Z_i$, or we may consider the instruments as approximating the true reduced form. In each case, estimation is performed using a set of K_n observed instruments Z_i, where $K_n \geq J_n$. Since both the dimension of X_i and Z_i are growing under the asymptotics considered here, the elements of the model may depend on n; we suppress this notation throughout for notational convenience.

The elements of X_i need not all be endogenous and may also include controls; the asymptotic sequence will allow for the strength of identification to vary across elements of X_i. However, the framework does not treat controls separately to endogenous regressors, so that J_n is the total number of variables in the model. Models with a large number of covariates may benefit from a framework that allows exogenous variables to grow at a rate different to J_n, although this is beyond the scope of this paper. In settings where the researcher observes a vector of controls of fixed dimension, we may interpret the regressors and instruments as having these controls partialled out.

Empirical examples

There are a number of potential settings in which the many endogenous regressors framework is relevant in empirical applications.

Multiple treatments or discrete-valued treatments

In some cases, researchers may be interested in assessing the impact of multiple treatments. If the available instruments are relevant for multiple treatments, then exogeneity requires including all treatments in a single model. In other settings, the treatment may be discrete-valued and including each level of treatment as a separate binary regressor would allow the researcher to more flexibly estimate the effect of the treatment than simply assuming a linear effect.

Linear demand models with many prices

In demand estimation, researchers may wish to estimate both own-price and cross-price elasticities by including the prices of many products in the model. Since prices are generally endogenous, linear demand systems can result in instrumental variables problems with potentially many endogenous regressors.

Heterogeneous treatment effects

It is well known that in settings with heterogeneous treatment effects instrumental variables regressions estimate a weighted average treatment effect. For example, Evdokimov and Kolesar (2019) show that in a model with potentially many instruments and covariates, linear IV estimators estimate a weighted average of LATEs. They derive these weights for a number of estimators, including 2SLS and JIVE, and consider inference with respect to both the implied conditional and unconditional estimands. The particular weighted average uses weights that depend on the heterogeneity in first-stage coefficients, and its relationship with the heterogeneity in treatment effects, and is typically not the object of interest to the researcher. This motivates attempts to explain at least some part of this heterogeneity, and construct other weighted-average treatment effects.

One possibility is to allow treatment effects to vary over a set of observed covariates by interacting the treatment variable with these covariates.

$$y_i = \beta_0 + \beta_1' X_i + \beta_2' W_i + \beta_3' X_i \otimes W_i + \varepsilon_i \tag{3.3}$$

In this interacted specification, both the treatment X_i and the interactions $X_i \times W_i$ are endogenous, and given an initial instrument set Z_i, new instruments could be formed using interactions with the exogenous covariates, $Z_i \otimes W_i$. An object of interest may then be the

average of the treatment effect over some distribution of the covariates, $\theta = \beta_1 + \beta_3' E[W_i]$ for example.

Nonparametric IV

The endogenous variables may represent a series of approximating functions of some endogenous variable, as in the nonparametric IV literature, e.g. Newey and Powell (2003). In this setting, the estimation problem is known to be ill-posed. Blundell et al. (2007) derive a sieve minimum distance estimator whose convergence rate depends on a sieve-measure of ill-posedness, τ_n. In the setting of this paper, the concentration parameter sequence r_n would be related to the sieve measure of ill-posedness by $\tau_n = \sqrt{n/r_n}$ so that slower growth of the concentration parameter makes estimation more difficult. We do not pursue the nonparametric approach here, and instead focus on the estimation of linear models.

3.3 Linear IV estimators

There exists a large literature on the properties of linear IV estimators under many weak instruments. One such set of estimators are the so-called k-class estimators, which can be written, for some value of κ, as

$$\tilde{\beta}_\kappa = \Big(X'PX - \kappa X'X\Big)^{-1}\Big(X'Py - \kappa X'y\Big)$$

where $P = Z(Z'Z)^{-1}Z'$ as the projection matrix for the set of instruments Z. This class of estimators includes as special cases: two-stage least squares (2SLS), for $\kappa = 0$; bias-corrected two-stage least squares (BC2SLS), for $\kappa = (K_n - 2)/n$; and limited information maximum likelihood (LIML), for $\kappa = \min_{\|\gamma\|=1}(\gamma'\bar{X}'\bar{X}\gamma)^{-1}\gamma'\bar{X}'P\bar{X}\gamma$, where $\bar{X} = [y, X]$. Estimators of this form were considered in Chao and Swanson (2005), who showed that LIML and BC2SLS are consistent under many weak instrument asymptotics, with homoskedasticity, as long as $\sqrt{K_n}/r_n \to 0$, while the stronger conditions $K_n/n \to 0$ is required for consistency of 2SLS, which is known to suffer from many instrument bias (see Bekker (1994)).

Unfortunately, these estimators are inconsistent in the presence of heteroskedasticity (Bekker and Van Der Ploeg (2005), Chao and Swanson (2004)) and LIML may not have finite moments, which can result in large dispersion in some settings (Hahn and Newey (2004)). In contrast, the jackknife IV estimator is consistent with heteroskedasticity as well as under many weak instruments (see Phillips and Hale (2006), Blomquist and Dahlberg (1999), Angrist et al. (1999), Ackerberg and Devereux (2009), Chao and Swanson (2004)).

However, it has been noted in Monte Carlo evidence that JIVE can be significantly more dispersed that other IV estimators (Davidson and Mackinnon, 2006).

Motivated by these facts, Hausman et al. (2009) develop jackknife k-class estimators, which they show to be consistent and asymptotically normal under many weak instrument asymptotics with heteroskedasticity, while remaining as efficient as LIML under homoskedasticity. The class of jackknife k-class estimators have the form

$$\hat{\beta}_\kappa = \Big(\sum_{i \neq j} P_{ij} X_i X_j' - \kappa X'X \Big)^{-1} \Big(\sum_{i \neq j} P_{ij} X_i y_j - \kappa X'y \Big) \tag{3.4}$$

This class contains as special cases: the JIVE estimator, when $\kappa = 0$;[1] the HLIM estimator, when $\kappa = \tilde{\kappa}$ is the smallest eigenvalue of the matrix $(\bar{X}'\bar{X})^{-1}(\bar{X}'\tilde{P}\bar{X})$, for $\bar{X} = [y, X]$ and $\tilde{P}$ the projection matrix $Z(Z'Z)^{-1}Z'$ with diagonal elements set to zero; and, the HFUL estimator when

$$\kappa = \hat{\kappa} = \big(\tilde{\kappa} - (1 - \tilde{\kappa})C/n\big) / \big(1 - (1 - \tilde{\kappa})C/n\big) \tag{3.5}$$

for some constant C (Hausman et al. (2009) recommend using $C = 1$). This paper considers the properties of these estimators when the number of endogenous regressors may also be large.

Numerous other procedures have also been proposed in the literature, in an effort to improve the properties of jackknife estimators. Ackerberg and Devereux (2009) propose an adjustment to the JIVE estimator that removes a finite sample bias that occurs in the presence of many *included* exogenous regressors. Carrasco (2012) proposes regularizing the covariance matrix of the instruments as a way to improve the finite sample properties of IV estimators, and this technique is extended to LIML and settings with very high-dimensional, or even a continuum of, instruments in Carrasco and Tchuente (2015) and Carrasco and Tchuente (2016). Hansen and Kozbur (2014) present a ridge-regularized JIVE estimator.

In the context of moment estimators, Newey and Windmeijer (2009) demonstrate consistency and asymptotic normality of generalized empirical likelihood (GEL) estimators, under many weak instruments with heteroskedasticity. These estimators, which include the continuously updated estimator (CUE), are also asymptotically more efficient than

[1] Angrist et al. (1999) describe two jackknife IV estimators: the JIVE2 estimator is presented above, while the JIVE1 estimator instead uses the weights $P_{ij}/(1-P_{ii})$. Under the assumptions used here, the estimators have similar theoretical properties.

jackknifed GMM estimators. However, the estimators rely on estimation of an optimal weighting matrix that is robust to heteroskedasticity which can lead to poor finite sample properties (Hausman et al., 2009). The CUE estimator can also be computationally difficult, a problem that is likely to be even more acute in a setting with a large number of endogenous variables, and hence many parameters to estimate. For this reason we focus on linear estimators in this paper, although the extension to moment estimation may be interesting.

3.4 Consistency

This section provides conditions under which the estimator $\hat{\beta}_\kappa$ is consistent, and linear combinations of the parameter vector are asymptotically normal. In the proofs below I focus on the Euclidean norm for vectors, $\|a\|^2 = \sum_i a_i^2$, and the operator norm for matrices, $\|A\| = \lambda_{max}(A)$, where λ_{max} denotes the largest eigenvalue of a matrix. We begin with some assumptions on the model, and the many weak instrument asymptotic sequence.

Assumption 3.1. *Z_n includes a column of ones, and for $P = Z_n(Z_n'Z_n)^{-1}Z_n'$ there exists a $C < 1$ such that for large enough n we have $rank(Z_n) = rank(P) = K_n$, $P_{ii} \leq C < 1$ for all i, and $\max_i P_{ii} = O(K_n/n)$.*

Assumption 3.1 contains standard conditions for instrumental variables models, apart from the addition of the assumption $\max_i P_{ii} = O(K_n/n)$. This condition is used to prove bounds on the operator norms of high-dimensional matrix-valued U-statistics that appear in the jackknife estimators. It ensures that there is balance in the instrument values across individuals in the data set, and is satisfied in many common settings. For example, if the instruments are a set of G group dummies, then $P_{ii} = 1/n_{g_i}$, where n_{g_i} is the number of individuals in group g_i, the group to which individual i belongs. The condition then requires that $\frac{Gn_g}{n} > 0$ for all groups, which is implied by the balancing condition $\min_g n_g / \max_g n_g > 0$. This condition could be relaxed, at the expense of a stronger rate condition in Assumption 3.5.

Assumption 3.2. *$\Upsilon_i = S_n z_i/\sqrt{n}$ with $S_n = \bar{S}_n diag(\mu_{n1}, \ldots, \mu_{nJ})$ for $\bar{S}_n$ a nonsingular matrix with largest eigenvalue bounded above for all n. For each j, either $\mu_{nj} = \sqrt{n}$ or $\mu_{nj} = o(\sqrt{n})$; define $r_n = \min_{1\leq j\leq J_n} \mu_{nj}^2$. There exists $0 < c \leq C < \infty$ such that with probability one, for sufficiently large n, $\lambda_{min}(\frac{1}{n}\sum_i z_i z_i') > c$ and $\lambda_{max}(\frac{1}{n}\sum_i z_i z_i') \leq C$.*

This is the many weak instrument asymptotic sequence introduced in Chao et al. (2012) and Hausman et al. (2009). The structure allows for the reduced form to be fixed for some elements of X_i, while the first-stage coefficients may decline to zero for other elements. The matrix S_n controls whether the reduced form approaches zero for some X_i and at what rate, and consequently plays an important role in the convergence rate of the estimators. It is not necessary to know, or specify, S_n when performing inference.

As an example, consider a simple model in which $X_i = (X_{1i}, W_i)$, with W_i an exogenous covariate. In this case, we have $z_i = (z_{1i}, W_i)$ and

$$S_n = \begin{bmatrix} \pi_1 & \pi_2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \sqrt{r_n} & 0 \\ 0 & \sqrt{n} \end{bmatrix}$$

so that the first-stage equation for the endogenous covariate is

$$X_{1i} = \pi_1 z_i \sqrt{r_n/n} + \pi_2 W_i + V_{1i}$$

while the first-stage equation for the exogenous covariate is simply the identity.

Assumption 3.3. *There exists a sequence of $K_n \times J_n$ matrices Π_n such that*

$$\frac{1}{n} \sum_i \|z_i - \Pi_n' Z_i\|^2 \to 0$$

This assumption allows for the z_i to be unknown functions of the observed instruments, by assuming that they can be well approximated by a linear combination of them. Here the Π_n can be interpreted as the first stage coefficients from linear regression of X_i on Z_i so that the assumption allows us to replace the unknown z_i with observed Z_i. The next assumption imposes some standard moment conditions on the error terms in the model.

Assumption 3.4. *For $\mathcal{Z} = (\Upsilon, Z_n)$, the observations $(\varepsilon_{n1}, V_{n1}'), \ldots, (\varepsilon_{nn}, V_{nn}')$ are independent, with $E[\varepsilon_{n,i}|\mathcal{Z}] = 0$ and $E[V_{n,i}|\mathcal{Z}] = 0$ for all n. There exists some $C < \infty$ such that $\sup_i E[\varepsilon_{ni}^2|\mathcal{Z}] \leq C$ and $\sup_i \lambda_{max}\big(E[V_{ni}V_{ni}'|\mathcal{Z}]\big) \leq C$ almost surely.*

Consistency and asymptotic normality results rely on the following condition, which describes the rate at which the number of endogenous regressors, and the number of instruments, may grow, relative to the strength of identification as measured by the sequence r_n.

Assumption 3.5. *We assume that* $\sup_i \|\Pi_n Z_i\|^2 \leq CJ_n$*, and* $\sup_i \|V_i\|^2 \leq CJ_n$*. We also have that* $\frac{J_n(\log^3 J_n)K_n}{r_n^2} \to 0$ *and* $\frac{J_n \log J_n}{r_n} \to 0$

The sequence r_n is important for the rate conditions below, and is related to the concentration parameter in the weak identification literature. The conditions require $r_n \to \infty$, so that the model is asymptotically identified. This rules out the weak instruments setting of Staiger and Stock (1997) in which r_n is bounded. Chao and Swanson (2005) show consistency of the JIVE estimator under the condition $\sqrt{K_n}/r_n \to 0$ in the many weak instruments setting, while Mikusheva and Sun (2020) show that this condition is in fact necessary for an asymptotically consistent test to exist. The condition in Assumption 3.5 implies these conditions, and is equivalent when the number of endogenous regressors is bounded. When J_n is allowed to grow to infinity, the condition requires even stronger identification in the model in order for consistent estimation of all coefficients. The rates in Assumption 3.5 are needed to ensure that the denominator term in (3.4) converges, in the operator norm, to well-defined positive definite matrix. To demonstrate this we rely on a matrix concentration inequality for sample averages (as in Belloni et al., 2015), as well as a matrix concentration inequality for degenerate U-statistics from (Minsker and Wei, 2019).

Note that $r_n = \min_{1 \leq j \leq J_n} \mu_{nj}^2$ is the rate at which information accumulates in the worst case, or least well identified, direction. This ensures that the conditions are sufficient to ensure consistency of the entire parameter vector, and allows for inference on any linear combination of the structural coefficients. The presence of poorly identified endogenous regressors can affect the ability to perform inference on all regressors in the model, so it is important to have conditions that ensure consistency for all endogenous coefficients.

We may now state the consistency result.

Theorem 3.1. *Let Assumptions 3.1 to 3.5 hold. Then*

$$\|\hat{\beta}_{\kappa,n} - \beta_n\|^2 = O_p\big(\frac{J_n K_n}{r_n^2} + \frac{J_n}{r_n}\big) \to 0$$

The theorem implies consistency of the entire parameter vector in the L_2-norm. As in previous work on many weak instrument estimation, the convergence rate contains two separate conditions that correspond to two parts of the error term (see for example Chao and Swanson (2005)). When $K_n/r_n \to 0$ the part of the error term that corresponds to uncertainty from the many instruments will be asymptotically negligible compared to the standard error term. In contrast, when $K_n/r_n \to \infty$ (but slowly enough that Assumption

3.5 still holds), the many instruments term will dominate in the asymptotic variance. When $K_n/r_n \to c$, the two terms are of equivalent order and both parts of the error appear in the asymptotic variance. We derive asymptotic normality under the condition that the number of instruments grows proportionally to the concentration parameter, so that the standard errors presented below will be consistent in either regime, and the researcher need not specify which case she is in.

3.5 Asymptotic normality and inference

Since the parameter vector β is growing in dimension, in establishing asymptotic normality it is important to specify the object of interest for inference. Here we assume that the researcher is interested in performing inference on a scalar θ that is equal to a linear combination of the coefficients.

Assumption 3.6. *The parameter of interest is linear in the structural coefficients, $\theta = \alpha'\beta$, with $\|\alpha\| \leq C$.*

Although we consider a scalar θ, the results should extend easily to any finite vector, with α now a conformable matrix. Assumption 3.6 restricts the vector of weights to have bounded Euclidean norm. This covers the case of inference on a linear combination of some finite subvector of β, including inference on a single parameter, but covers other interesting cases as well. For example, we may be interested in performing inference on the effect of a single treatment, or in some weighted-average treatment effect for a multi-valued treatment. In the case where there is treatment effect heterogeneity, and the endogenous regressors are interactions between treatment and a set of covariates, we may be interested in the average treatment effect under some distribution of the covariates. The bound on the norm of α_n seems a necessary condition for consistency at the parametric rate (which here depends on strength of identification and ranges from $\sqrt{r_n}$ to $\sqrt{n}$), but will not be satisfied in all cases of interest. Asymptotic normality is likely possible under weaker conditions on α_n, but with slower convergence.

There are of course interesting cases which are not covered by Assumption 3.6. Researchers may be interested in performing inference on the largest or smallest treatment effect across some set of groups, e.g. $\theta = \max_j \beta_j$, or may wish to test the hypothesis that no group has a negative average treatment effect. These would require different asymptotic results, and are left for future research.

We next state the asymptotic variance of the estimators, and provide estimators for these quantities. The variances follow those derived in Hausman et al. (2009). Let $\sigma_i^2 = E\big[\varepsilon_i^2 \big| \mathcal{Z}\big]$, and define

$$\begin{aligned} H_n &= \frac{1}{n}\sum_i (1 - P_{ii}) z_i z_i' \\ D_n &= \frac{1}{n}\sum_i (1 - P_{ii})^2 z_i z_i' \sigma_i^2 \\ \Sigma_n &= \sum_{i \neq j} P_{ij}^2 S_n^{-1} \big(E[V_i V_i' | \mathcal{Z}] \sigma_j^2 \\ &\quad + E[V_i \varepsilon_i | \mathcal{Z}] E[V_i' \varepsilon_i | \mathcal{Z}] \big) (S_n^{-1})' \end{aligned}$$

The conditional (on $\mathcal{Z}$) asymptotic variance for the jackknife IV estimator is given by

$$V_{J,n} = b_n^2 \alpha_n' (S_n^{-1})' H_n^{-1} \big(D_n + \Sigma_n \big) H_n^{-1} S_n^{-1} \alpha_n$$

where $b_n = \|\alpha_n' (S_n^{-1})'\|^{-1}$. For the HLIM and HFUL estimators define

$$\begin{aligned} \tilde{\Sigma}_n &= \sum_{i \neq j} P_{ij}^2 S_n^{-1} \big(E[\tilde{V}_i \tilde{V}_i' | \mathcal{Z}] \sigma_j^2 \\ &\quad + E[\tilde{V}_i \varepsilon_i | \mathcal{Z}] E[\tilde{V}_i' \varepsilon_i | \mathcal{Z}] \big) (S_n^{-1})' \end{aligned}$$

where $\tilde{V}_i = V_i - \gamma_n \varepsilon_i$ and $\gamma_n = \sum_i E[V_i \varepsilon_i] / \sum_i \sigma_i^2$. Then, the conditional (on $\mathcal{Z}$) asymptotic variance of these estimators is

$$V_{\kappa,n} = b_n^2 \alpha_n' (S_n^{-1})' H_n^{-1} \big(D_n + \tilde{\Sigma}_n \big) H_n^{-1} S_n^{-1} \alpha_n$$

As noted in (Hausman et al., 2009), under homoskedasticity, the asymptotic variance of JIVE will be larger than that of HLIM and HFUL, since $E[V_i V_i'] \geq E[\tilde{V}_i \tilde{V}_i']$, although in general it is difficult to rank the two estimators.

Here b_n plays the role of a normalizing sequence that ensures that the asymptotic limits are well defined, and differs with the choice of linear combination α_n. Recall that $S_n/\sqrt{n} = \bar{S} \times diag(\mu_{n1}, \ldots, \mu_{nJ})/\sqrt{n}$ is the matrix of first-stage coefficients in the optimal first stage using z_i as instruments, so that $S_n S_n'$ plays a role similar to the concentration parameter

matrix. Then

$$b_n^2 = \frac{1}{\alpha_n'(S_n S_n')^{-1}\alpha_n}$$

so that b_n represents the strength of identification for the particular linear combination α_n. In the simplest case in which all directions of identification decline at the same rate, we have that $\mu_{jn} = \sqrt{r_n}$ for all j, and so $b_n = \sqrt{r_n}$. In general, Assumption 3.2 implies that $\min_j \mu_{jn} \leq b_n \leq \max_j \mu_{jn}$ so that the convergence rate is somewhere between the strongest and weakest directions of identification. This is at most at the $\sqrt{n}$-rate, and at the slowest $\sqrt{r_n}$.

As an example, consider the case in which we have a single treatment that is interacted with a set of mutually exclusive group dummies, with instruments formed by corresponding interactions of the group dummies with a single instrument. In this case, S_n is a diagonal matrix since the endogenous regressors are orthogonal to one another, as are the instruments.

$$S_n = \begin{bmatrix} \pi_1 & & 0 \\ & \ddots & \\ 0 & & \pi_J \end{bmatrix} \begin{bmatrix} \mu_1 & & 0 \\ & \ddots & \\ 0 & & \mu_J \end{bmatrix}$$

If we are interested in a weighted average of the J group treatment effects, with weights α_j, then

$$b_n^2 = \big(\sum_{j=1}^{J} \alpha_j^2 \pi_j^2 \mu_j^2\big)^{-1}$$

so that the rate of convergence is determined by a weighted average of the convergence rates of the group treatment effects. If we were instead only interested in inference on the j-th group, the rate of convergence would naturally be μ_j^{-2}. In this special case, the estimation of the treatment effect for group j is unaffected by the strength of identification in other groups. This is expected since fully interacting with group dummies is equivalent to running separate group-level regressions. In general however, whenever the first-stage fitted values are correlated across endogenous regressors (so that S_n is not diagonal) the strength of identification of all regressors will be important, even if we are only interested in a single regression coefficient.

With the addition of the following moment condition, we can now state the asymptotic normality result.

Assumption 3.7. *There exists a constant C such that $\frac{1}{n^2}\sum_i E\left[\|z_i\|^4\right] \to 0$, $\sup_{1\le i\le n} E[\varepsilon_i^4|\mathcal{Z}] \le C$, and $\sup_{1\le i\le n} E[V_{j,i}^4|\mathcal{Z}] \le C$ for all j.*

Theorem 3.2. *Let Assumptions 3.1 to 3.7 hold, and further assume that and $K_n/r_n \to c < \infty$. For the estimators $\hat{\theta}_{J,n} = \alpha_n'\hat{\beta}_J$ and $\hat{\theta}_\kappa = \alpha_n'\hat{\beta}_\kappa$, and for b_n, $V_{J,n}$ the JIVE estimator, and $V_{\kappa,n}$ either the HLIM or HFUL estimator, we have*

$$b_n V_{J,n}^{-1/2}\left(\hat{\theta}_{J,n} - \theta_n\right) \Rightarrow N(0,1)$$
$$b_n V_{\kappa,n}^{-1/2}\left(\hat{\theta}_{\kappa,n} - \theta_n\right) \Rightarrow N(0,1)$$

The theorem implies that inference on $\hat{\theta}_n$ might be performed using the approximation $\widehat{\theta} \sim N(\theta_n, b_n^{-2}V_n)$. We next prove that, for an appropriate estimator of the asymptotic variance $\hat{V}$, we have

$$b_n^2\hat{V} - V_n \to 0$$

so that $\widehat{V}$ may be used to construct confidence sets for the parameter of interest. For the JIVE, the variance estimator is given by

$$\begin{aligned}
\hat{\Xi}_J &= \sum_{i\neq j\neq k} P_{ik}P_{jk}X_iX_j'\hat{\varepsilon}_k^2 \\
&\quad + \sum_{i\neq j} P_{ij}^2\left(X_i\hat{\varepsilon}_iX_j'\hat{\varepsilon}_j + X_iX_i'\hat{\varepsilon}_j^2\right)' \\
\hat{H}_J &= \sum_{i\neq j} X_iP_{ij}X_j \\
\hat{V}_J &= \alpha_n'\hat{H}_J^{-1}\hat{\Xi}_J\hat{H}_J^{-1}\alpha_n
\end{aligned}$$

while the variance estimator for the HLIM and HFUL estimators is

$$\begin{aligned}
\hat{\Xi}_\kappa &= \sum_{i\neq j\neq k} P_{ik}P_{jk}\hat{X}_i\hat{X}_j'\hat{\varepsilon}_k^2 \\
&\quad + \sum_{i\neq j} P_{ij}^2\left(\hat{X}_i\hat{X}_i'\hat{\varepsilon}_j^2 + \hat{X}_i\hat{\varepsilon}_i\hat{X}_j'\hat{\varepsilon}_j\right) \\
\hat{H}_\kappa &= \sum_{i\neq j} X_iP_{ij}X_j - \kappa X'X \\
\hat{V}_\kappa &= \alpha_n'\hat{H}_\kappa^{-1}\hat{\Xi}_\kappa\hat{H}_\kappa^{-1}\alpha_n
\end{aligned}$$

Under a slightly stronger condition on the number of allowed endogenous regressors, we

may demonstrate consistency of the two asymptotic variance estimators.

Assumption 3.8. *The following rates hold:* $J_n^2 K_n / r_n^2 \to 0$ *and* $J_n^2 / r_n \to 0$

Theorem 3.3. *Let Assumptions 3.1 to 3.8 hold. Then*

$$b_n^2 \hat{V}_{J,n} - V_{J,n} \to 0$$
$$b_n^2 \hat{V}_{\kappa,n} - V_{\kappa,n} \to 0$$

and hence

$$\hat{V}_J^{-1/2}\big(\hat{\theta}_{J,n} - \theta_n\big) \Rightarrow N(0,1)$$
$$\hat{V}_\kappa^{-1/2}\big(\hat{\theta}_{\kappa,n} - \theta_n\big) \Rightarrow N(0,1)$$

Consistency of the asymptotic variance estimates, as shown in the theorem, implies that we may construct confidence sets for θ_n in the usual way, by using the asymptotic normal approximation.

3.6 Simulations

In this section we conduct some simple Monte Carlo experiments to provide further evidence on the finite sample behavior of IV estimators with multiple endogenous regressors. The simulations are based on those in Hausman et al. (2009), increasing the number of endogenous regressors. In each simulation we draw a sample of $n = 1000$ observations from the data generating process

$$y_i = \beta_0 + \sum_{j=1}^{J} \beta_j X_{j,i} + \varepsilon_i$$
$$X_{j,i} = \pi_j z_{j,i} + V_{j,i}$$

with $z_{j,i} \sim N(0,1)$ and $(V_{1,i}, \ldots, V_{J,i}) \sim N(0, I_J)$. For each choice of J, we set $\beta_j = (\sigma_\varepsilon^2 / \sum_j Var(X_{j,i}))^{1/2}$ so that the proportion of the variation in y_i explained by the endogenous regressors is held fixed.

The set of instruments used in estimation include $Z_i = (1, Z'_{1,i}, \ldots Z'_{J,i})$ with $Z'_{j,i} = (z_{j,i}, z_{j,i}^2, z_{j,i}^3, z_{j,i}^4, z_{j,i} D_{1,i}, \ldots, z_{j,i} D_{\tilde{K}-5,i})$, so that there are $K = J\tilde{K}$ instruments in total. The first-stage coefficients π_j are a sequence from π_u to π_l, where $\pi_u = 0.5$ and $\pi_l = \{\sqrt{0.1}, \sqrt{0.03}\}$,

so that the minimum eigenvalue of the concentration matrix $\Sigma_V^{-1/2}\Pi' Z' Z \Pi \Sigma_V^{-1/2}$ is given by $\lambda_{min} = \{100, 30\}$.

The structural error is heteroskedastic so that standard k-class estimators such as LIML are inconsistent under many instruments, but the heteroskedasticity-robust versions are consistent.

$$\varepsilon_i = \rho' V_i + \sqrt{\frac{1 - \rho' \Sigma_V \rho}{\phi^2 + (0.86)^4}} \left(\phi v_{1,i} + 0.86 v_{2,i} \right)$$

with $v_{1,i} \sim N(0, \|z_i\|^2 / J)$ and $v_{2,i} \sim N(0, (0.86)^2)$. To investigate how the choice direction α affects the properties of the estimator, we consider inference on both the first regressor, $\theta = \beta_1$, as well as the average of the coefficients $\theta = \frac{1}{J} \sum_j \beta_j$.

Table 3.1 shows the median bias, 5th-95th percentile range, and rejection rates for a 5% test when $\lambda_{min} = 100$. As expected, the TSLS estimator is approximately median unbiased in the just-identified settings, but becomes biased in overidentified models. The JIVE has a small bias that appears to be increasing in the number of endogenous regressors. Bias of the JIVE with many covariates has previously been noted by Ackerberg and Devereux (2009), who provide an improved version of the estimator that is robust to these settings. It would be interesting to investigate if such a correction is available in the many endogenous regressors setting. The HFUL estimator shows a small bias in the just-identified settings, but this bias remains small as both J and K increase. While the dispersion of the estimators is similar in the just-identified setting with small J, the estimators differ greatly in other settings. In particular, the JIVE becomes highly dispersed as J grows, while the HFUL estimator shows slightly lower dispersion than TSLS in just-identified settings, and much lower dispersion than JIVE in all settings. Rejection rates for HFUL are close to the nominal 5% level for $J \leq 10$, but become conservative for $J = 20$, while the high dispersion of JIVE leads to no rejection in the $J = 20$ case. It is interesting to note that there is more over-rejection for the $\theta = \frac{1}{J} \sum_j \beta_j$ parameter, which aligns with the fact that this is a less well identified direction in comparison to $\theta = \beta_1$, which captures the best identified parameter.

Table 3.2 shows the results of the simulation with $\lambda_{min} = 30$. Results for $\theta = \beta_1$ are similar to those in the previous simulation, while bias, dispersion and rejection rates are worse for all estimators of the average coefficient parameter. This again aligns with the fact that the strength of identification for β_1 is unchanged, while the identification of other coefficients is now weaker. Although this is a somewhat contrived example, in which all instruments and endogenous regressors are independent, it is instructive to see that in the many endogenous regressors case, strength of identification can differ for different objects of

interest. As was seen in the previous simulations, HFUL performs well in the presence of both overidentification and moderately large J.

3.7 Conclusion

This paper considers estimation of a linear IV model in which the number of endogenous regressors may increase with the sample size. We derive conditions under which the jackknife IV, HLIM and HFUL estimators are consistent, and show that certain linear combinations of the coefficients are asymptotically normal. Simulation evidence suggests that JIVE becomes unstable in settings with multiple endogenous regressors, and so we would suggest researchers do not use this estimator when estimating such models. In contrast, the HFUL estimators performs well in simulations with moderate numbers of endogenous regressors and many instruments.

This research could usefully be extended in a number of directions. Results for sequences of models in which the number of controls is increasing separately to the number of endogenous regressors may be of interest, see for example Cattaneo et al. (2018), which analyzes the partially linear model with a fixed number of coefficients of interest and and a number of controls that grows proportionally with sample size, and also Evdokimov and Kolesár (2019), which discusses IV estimators with many controls. Further results that allow for inference on objects that are not linear combinations of coefficients would also be important for allowing researchers to address practical policy questions. The conditions in this paper allow for moderate numbers of endogenous regressors, and so extensions to regularized models may also be of interest for settings in which the number of endogenous regressors is very large.

Simulation results suggest that the JIVE is particularly sensitive to settings with both moderate numbers of endogenous regressors and many instruments. The regularization role played by the term $\kappa_n X'X$ in the denominator of the HFUL estimator appears to be important for the superior performance of the HFUL estimator in these settings. Further analysis on the theoretical differences between these two estimators would be illuminating.

Table 3.1: Simulation results: $\lambda_{min} = 100$

		$\theta = \beta_1$			$\theta = \frac{1}{J}\sum_{j=1}^{J}\beta_j$		
J	K	TSLS	JIVE	HFUL	TSLS	JIVE	HFUL
				Median			
5	×1	-0.001	-0.021	0.007	0.001	-0.032	0.014
5	×5	0.060	-0.020	0.014	0.089	-0.039	0.012
5	×10	0.121	-0.024	0.012	0.172	-0.044	0.011
10	×1	-0.003	-0.043	0.006	0.001	-0.069	0.014
10	×5	0.113	-0.050	0.012	0.161	-0.087	0.014
10	×10	0.213	-0.069	0.011	0.283	-0.119	0.016
20	×1	0.000	-0.107	0.010	-0.001	-0.191	0.012
20	×5	0.192	-0.193	0.006	0.270	-0.296	0.017
20	×10	0.327	-0.113	0.015	0.413	-0.103	0.023
				5%-95% Range			
5	×1	0.622	0.648	0.612	0.357	0.389	0.348
5	×5	0.594	0.696	0.650	0.326	0.440	0.373
5	×10	0.539	0.740	0.676	0.285	0.492	0.401
10	×1	0.821	0.908	0.803	0.326	0.392	0.317
10	×5	0.684	1.055	0.861	0.261	0.532	0.349
10	×10	0.616	1.333	0.988	0.217	0.756	0.402
20	×1	1.148	1.669	1.116	0.325	0.633	0.315
20	×5	0.848	8.937	1.442	0.214	5.061	0.409
20	×10	0.687	17.366	2.011	0.169	8.019	0.556
				Rejection (5%)			
5	×1	0.043	0.043	0.048	0.048	0.043	0.058
5	×5	0.070	0.047	0.053	0.170	0.037	0.057
5	×10	0.120	0.041	0.051	0.505	0.030	0.057
10	×1	0.041	0.038	0.046	0.042	0.044	0.055
10	×5	0.078	0.026	0.040	0.509	0.019	0.049
10	×10	0.222	0.018	0.045	0.976	0.007	0.050
20	×1	0.037	0.018	0.044	0.039	0.033	0.054
20	×5	0.122	0.000	0.030	0.966	0.000	0.048
20	×10	0.371	0.000	0.019	1.000	0.000	0.038

Table 3.2: Simulation results: $\lambda_{min} = 30$

		$\theta = \beta_1$			$\theta = \frac{1}{J}\sum_{j=1}^{J}\beta_j$		
J	K	TSLS	JIVE	HFUL	TSLS	JIVE	HFUL
				Median			
5	×1	0.006	-0.017	0.014	0.001	-0.082	0.027
5	×5	0.066	-0.020	0.020	0.150	-0.090	0.030
5	×10	0.125	-0.027	0.020	0.252	-0.092	0.033
10	×1	0.005	-0.045	0.015	0.000	-0.175	0.025
10	×5	0.120	-0.053	0.025	0.235	-0.122	0.034
10	×10	0.218	-0.053	0.021	0.362	-0.026	0.048
20	×1	0.006	-0.063	0.016	-0.001	-0.090	0.028
20	×5	0.200	-0.070	0.020	0.351	0.077	0.069
20	×10	0.332	-0.019	0.058	0.479	0.243	0.105
				5%-95% Range			
5	×1	0.630	0.704	0.610	0.522	0.745	0.484
5	×5	0.578	0.877	0.667	0.418	1.338	0.638
5	×10	0.526	1.097	0.719	0.342	2.182	0.773
10	×1	0.834	1.435	0.808	0.465	1.533	0.431
10	×5	0.661	3.367	0.948	0.312	4.659	0.623
10	×10	0.595	4.908	1.138	0.242	6.048	0.832
20	×1	1.225	7.428	1.155	0.476	5.729	0.431
20	×5	0.815	11.906	1.809	0.240	7.270	0.820
20	×10	0.668	14.957	2.553	0.178	7.628	1.102
				Rejection (5%)			
5	×1	0.039	0.031	0.045	0.042	0.019	0.062
5	×5	0.070	0.024	0.045	0.242	0.013	0.062
5	×10	0.132	0.014	0.040	0.669	0.013	0.062
10	×1	0.033	0.012	0.039	0.034	0.004	0.056
10	×5	0.081	0.002	0.027	0.678	0.002	0.054
10	×10	0.241	0.001	0.025	0.994	0.001	0.055
20	×1	0.021	0.000	0.031	0.029	0.000	0.053
20	×5	0.129	0.000	0.012	0.992	0.000	0.042
20	×10	0.394	0.000	0.008	1.000	0.000	0.037

Bibliography

Ackerberg, D. A. and Devereux, P. J. 2009. Improved JIVE estimators for overidentified linear models with and without heteroskedasticity. *Review of Economics and Statistics*, 91(2):351–362.

Andersen, E. 1970. Asymptotic Properties of Conditional Maximum Likelihood Estimators. *Journal of the Royal Statistical Society, Series B*, 32:283–301.

Anderson, J. E. and Van Wincoop, E. 2003. Gravity with Gravitas: A Solution to the Border Puzzle. *American Economic Review*, 93(1):170–192.

Angrist, J. D., Imbens, G. W., and Krueger, A. B. 1999. Jacknife Instrumental Variables Estimation. *Journal of Applied Econometrics*, 14(1):57–67.

Arellano, M. and Hahn, J. 2010. Understanding Bias in Nonlinear Panel Models: Some Recent Developments. *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress, Volume III*, 3:381–409.

Arellano, M. and Honoré, B. *Panel Data Models: Some Recent Developments*. 2001.

Bekker, P. A. 1994. Alternative Approximations to the Distributions of Instrumental Variable Estimators. *Econometrica*, 62(3):657.

Bekker, P. A. and Van Der Ploeg, J. 2005. Instrumental variable estimation based on grouped data. *Statistica Neerlandica*, (59):506–508.

Belloni, A., Chernozhukov, V., Chetverikov, D., and Kato, K. 2015. Some New Asymptotic Theory for Least Squares: Pointwise and Uniform Results. *Journal of Econometrics*, 86 (2):345–366.

Bloch, F. and Jackson, M. O. 2007. The Formation of Networks with Transfers Among Players. *Journal of Econometric Theory*, 133(1):83–110.

Blomquist, S. and Dahlberg, M. 1999. Small Sample Properties of LIML and Jackknife IV Estimators : Experiments with Weak Instruments. *Journal of Applied Econometrics*, 14 (1):69–88.

Blundell, B. Y. R., Chen, X., and Kristensen, D. 2007. Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves. *Econometrica*, 75(6):1613–1669.

Burger, M., van Oort, F., and Linders, G. J. 2009. On the Specification of the Gravity Model of Trade: Zeros, Excess Zeros and Zero-inflated Estimation. *Spatial Economic Analysis*, 4 (2):167–190.

Candelaria, L. E. 2020. A Semiparametric Network Formation Model with Unobserved Linear Heterogeneity.

Carrasco, M. A regularization approach to the many instruments problem. In *Journal of Econometrics*, 2012.

Carrasco, M. and Tchuente, G. Regularized LIML for many instruments. In *Journal of Econometrics*, 2015.

Carrasco, M. and Tchuente, G. 2016. Efficient Estimation with Many Weak Instruments Using Regularization Techniques. *Econometric Reviews*, 35(8-10):1609–1637.

Cattaneo, M. D., Jansson, M., and Newey, W. K. 2018. Alternative asymptotics and the partially linear model with many regressors. *Econometric Theory*, 34(2):277–301.

Chamberlain, G. 1984. Panel data. *Handbook of Econometrics, Vol 2*.

Chamberlain, G. 2010. Binary Response Models for Panel Data: Identification and Information. *Econometrica*, 78:159–168.

Chao, J. C. and Swanson, N. R. 2004. Estimation and Testing Using Jackknife IV in Heteroskedastic Regressions with Many Weak Instruments.

Chao, J. C. and Swanson, N. R. Consistent estimation with a large number of weak instruments, 2005.

Chao, J. C., Swanson, N. R., Hausman, J. A., Newey, W. K., and Woutersen, T. 2012. Asymptotic distribution of JIVE in a heteroskedastic IV regression with many instruments. *Econometric Theory*, 28(1):42–86.

Charbonneau, K. B. 2017. Multiple Fixed Effects in Binary Response Panel Data Models. *The Econometrics Journal*, 20(3):1–13.

Chen, M., Fernández-Val, I., and Weidner, M. 2021. Nonlinear Factor Models for Network and Panel Data. *Journal of Econometrics*, 220:296–324.

Chernozhukov, V., Newey, W. K., Hahn, J., and Fernandez-Val, I. 2013. Average and Quantile Effects in Nonseparable Panel Models. *Econometrica*, 81:535–580.

Cruz-Gonzalez, M., Fernández-Val, I., and Weidner, M. 2017. Bias Corrections for Probit and Logit Models with Two-way Fixed Effects. *The Stata Journal*, 17(3):517–545.

Davidson, B. Y. R. and Mackinnon, J. G. 2006. The case against JIVE. *Journal of Applied Econometrics*, (21):827–833.

de Paula, Á. 2020. Econometrics of Network Models. *Annual Review of Economics*, 12: 775–799.

Dhaene, G. and Jochmans, K. 2015. Split-panel Jackknife Estimation of Fixed-effect Models. *Review of Economic Studies*, 82(3):991–1030.

Dzemski, A. 2019. An Empirical Model of Dyadic Link Formation in a Network with Unobserved Heterogeneity. *Review of Economics and Statistics*, 101(5):763–776.

Evdokimov, K. S. and Kolesár, M. 2019. Inference in Instrumental Variables Analysis with Heterogeneous Treatment Effects.

Fernández-Val, I. 2009. Fixed Effects Estimation of Structural Parameters and Marginal Effects in Panel Probit Models. *Journal of Econometrics*, 150:71–85.

Fernández-Val, I. and Weidner, M. 2016. Individual and Time Effects in Nonlinear Panel Models with Large N, T. *Journal of Econometrics*, 192(1):291–312.

Gao, W. Y. 2020. Nonparametric Identification in Index Models of Link Formation. *Journal of Econometrics*, 215(2):399–413.

Graham, B. S. 2017. An Econometric Model of Network Formation With Degree Heterogeneity. *Econometrica*, 85(4):1033–1063.

Graham, B. S. 2020. Network Data. *Handbook of Econometrics*, 7.

Graham, B. S. and Pelican, A. 2020. An Optimal Test for Strategic Interaction in Social and Economic Network Formation Between Heterogeneous Agents.

Hahn, J. and Kuersteiner, G. 2002. Asymptotically Unbiased Inference for a Dynamic Panel Model with Fixed Effects when Both n and T are Large. *Econometrica*, 70:1639–1657.

Hahn, J. and Newey, W. K. 2004. Jackknife and Analytical Bias Reduction for Nonlinear Panel Models. *Econometrica*, 72:1295–1319.

Hahn, J., Kuersteiner, G., and Newey, W. K. 2002. Higher Order Properties of Bootstrap and Jackknife Bias Corrected Maximum Likelihood Estimators.

Han, C. and Phillips, P. C. 2006. GMM with many moment conditions. *Econometrica*, (74): 147–192.

Hansen, C. and Kozbur, D. 2014. Instrumental variables estimation with many weak instruments using regularized JIVE. *Journal of Econometrics*.

Hansen, C. B., Hausman, J. A., and Newey, W. K. 2008. Estimation with Many Instrumental Variables. *Journal of Business & Economic Statistics*, 26(4):398–422.

Hausman, J. A., Newey, W. K., Woutersen, T., Chao, J. C., and Swanson, N. R. 2009. Instrumental Variable Estimation with Heteroskedasticity and Many Instruments. *Ssrn*, (0136869).

Heckman, J. J. 1981. The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time-Discrete Data Stochastic Process and Some Monte Carlo Evidence. *Structual Analysis of Discrete Data*.

Helpman, E., Melitz, M., and Rubinstein, Y. 2008. Estimating Trade Flows: Trading Partners and Trading Volumes. *Quarterly Journal of Economics*, 123(2):441–487.

Holland, P. W. and Leinhardt, S. 1981. An Exponential Family of Probability Distributions for Directed Graphs. *Journal of the American Statistical Association*, 76(373):33–50.

Hughes, D. W. and Hahn, J. 2020. The Higher-order Variance of Bias-corrected Panel Estimators.

Jochmans, K. 2018. Semiparametric Analysis of Network Formation. *Journal of Business and Economic Statistics*, 36(4):705–713.

Linders, G.-J. and de Groot, H. L. F. 2006. Estimation of the Gravity Equation in the Presence of Zero Flows.

Mikusheva, A. and Sun, L. 2020. Inference with Many Weak Instruments.

Minsker, S. and Wei, X. 2019. Moment inequalities for matrix-valued U-statistics of order 2. *Electronic Journal of Probability*, 24:1–32.

Morimune, K. 1983. Approximate Distributions of k-Class Estimators when the Degree of Overidentifiability is Large Compared with the Sample Size. *Econometric Theory*, 51(3): 821–841.

Nagar, A. L. 1959. The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations. *Econometrica*, 27:575–595.

Newey, W. K. and McFadden, D. 1994. Chapter 36: Large Sample Estimation and Hypothesis Testing. *Handbook of Econometrics*, 4:2111–2245.

Newey, W. K. and Powell, J. L. 2003. Instrumental Variables Estimation of Nonparametric Models. *Econometrica*, 71(5):1565–1578.

Newey, W. K. and Windmeijer, F. 2009. Generalized Method of Moments With Many Weak Moment Conditions. *Econometrica*, 77(3):687–719.

Neyman, J. and Scott, E. L. 1948. Consistent Estimates Based on Partially Consistent Observations. *Econometrica*, 16:1–32.

Pfanzagl, J. and Wefelmeyer, W. 1978. A Third-order Optimum Property of the Maximum Likelihood Estimator. *Journal of Multivariate Analysis*, 8:1–29.

Phillips, G. D. A. and Hale, C. 2006. The Bias of Instrumental Variable Estimators of Simultaneous Equation Systems. *International Economic Review*, 18(1):219.

Pratt, J. W. 1981. Concavity of the Log Likelihood. *Journal of the American Statistical Association*, 76:103–106.

Quenouille, M. H. 1956. Notes on Bias in Estimation. *Biometrika*, 43:353–360.

Rose, A. K. 2004. Do We Really Know That the WTO Increases Trade? *American Economic Review*, 94(1):98–114.

Rothenberg, T. J. 1984. Approximating the Distributions of Econometric Estimators and Test Statistics. *Handbook of Econometrics, Vol 2*.

Santos Silva, J. M. C. and Tenreyro, S. 2006. The Log of Gravity. *Review of Economic Studies*, 88(4):641–658.

Staiger, D. and Stock, J. H. 1997. Instrumental Variables Regression With Weak Instruments. *Econometrica*, 65(3):557–586.

Stock, J. H. and Yogo, M. Asymptotic Distributions of Instrumental Variables Statistics with Many Instruments. In *Identification and Inference in Honor of Thomas Rothenberg*, pages 109–120. Cambridge University Press, 2005.

Toth, P. *Semiparametric Estimation in Network Formation Models with Homophily and Degree Heterogeneity*. PhD thesis, 2017.

Tukey, J. W. 1958. Bias and Confidence in Not-Quite Large Samples (Abstract). *The Annals of Mathematical Statistics*, 29:614.

van der Vaart, A. W. *Asymptotic Statistics*. Cambridge University Press, 1998.

White, H. 1982. Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, 50 (1):1–25.

Yan, T., Jiang, B., Fienberg, S. E., and Leng, C. 2019. Statistical Inference in a Directed Network Model With Covariates. *Journal of the American Statistical Association*, 114 (526).

Zeleneev, A. 2020. Identification and Estimation of Network Models with Nonparameteric Unobserved Heterogeneity.

Appendices

A Appendix for Chapter 1

A.1 Notation and norms

The notation in the appendices follows that in the main paper. That is, denote partial derivatives of the objective function using subscripts, so that $\partial_\beta \mathcal{L}(\beta, \phi)$ denotes $\partial \mathcal{L}(\beta, \phi)/\partial \beta$ and so on, where $\phi' = (\alpha', \gamma')$. When functions are evaluated at β_0, ϕ_0 the dependence on these arguments is dropped. We also use ℓ_{ij} as shorthand for the function $\ell(Z_{ij}, \cdot)$, with $Z_{ij} = (Y_{ij}, X_{ij})$. Let

$$\mathcal{S}(\beta, \phi) = \partial_\phi \mathcal{L}(\beta, \phi), \qquad \mathcal{S}_\beta(\beta, \phi) = \partial_\beta \mathcal{L}(\beta, \phi)$$

be the first derivatives of the objective function. We also write

$$\mathcal{H}(\beta, \phi) = -\partial_{\phi\phi'} \mathcal{L}(\beta, \phi)$$

for the negative of the Hessian with respect to the fixed effects, a $2N \times 2N$ matrix.

We follow Fernández-Val and Weidner (2016) (FVW) in using the Euclidean norm $\|\cdot\|$ for $\dim \beta$ vectors, and the norm induced by the Euclidean norm for matrices and tensors, i.e.

$$\|\partial_{\beta\beta\beta} \mathcal{L}(\beta, \phi)\| = \max_{u,v \in \mathbb{R}^{\dim \beta} : \|u\|=1, \|v\|=1} \| \sum_{k,l=1}^{\dim \beta} u_k v_l \partial_{\beta\beta_k\beta_l} \mathcal{L}(\beta, \phi) \|$$

In the proofs we sometimes take β to be a scalar to simplify notation, although the results apply to any vector of fixed size. Since the number of fixed effect parameters in the model grows with N, the choice of norm for $\dim \phi$ vectors and matrices is important. Following FVW, we choose the ℓ_q-norm for $\dim \phi$ vectors and the corresponding induced norms for

matrices and tensors

$$\|\partial_{\phi\phi\phi}\mathcal{L}(\beta,\phi)\| = \max_{u,v\in\mathbb{R}^{\dim\phi}:\|u\|=1,\|v\|=1} \|\sum_{k,l=1}^{\dim\beta} u_k v_l \partial_{\phi\phi_k\phi_l}\mathcal{L}(\beta,\phi)\|_q$$

See FVW for more details on these norms.

We define the sets $\mathcal{B}(r,\beta_0) = \{\beta : \|\beta-\beta_0\| \leq r\}$, for $r > 0$, and $\mathcal{B}_q(r,\phi_0) = \{\phi : \|\phi-\phi_0\|_q \leq r\}$.

A.2 Asymptotic expansions

Verifying Assumption B.1 in Fernández-Val and Weidner (2016)

The results in the paper are based on asymptotic expansions of the objective function. Fernández-Val and Weidner (2016) (FVW) derive expansions for a general class of M-estimators with multiple incidental parameters, which includes the model studied here. In order to determine the properties of the jackknife estimator, I extend the expansions in FVW to higher-order, which requires additional conditions on the number of moments and derivatives of the objective function that exist. Derivations of higher-order terms and their bounds are quite long and largely similar to the derivations in FVW, and so are provided in the Supplementary Appendix. This appendix contains analysis based on the first-order expansions and focuses on results related to the jackknife, which are of most interest. To do so, I begin by verifying the conditions in Assumption B.1. of FVW, which will allow us to use some important results in that paper, including consistency of the common parameter and vector of fixed effects.

As is the paper, we use $\bar{E}$ to denote expectation conditional on exogenous covariates and fixed effects. Let $\bar{\mathcal{H}} = \bar{E}[\mathcal{H}]$, $\tilde{\mathcal{H}} = \mathcal{H} - \bar{\mathcal{H}}$, and similarly for other conditional expectations and their residuals. As in FVW, we may write

$$\bar{\mathcal{H}} = \begin{bmatrix} \bar{\mathcal{H}}^*_{\alpha\alpha} & \bar{\mathcal{H}}^*_{\alpha\gamma} \\ \bar{\mathcal{H}}^*_{\gamma\alpha} & \bar{\mathcal{H}}^*_{\gamma\gamma} \end{bmatrix} + \frac{1}{N} b v_N v_N'$$

where $\bar{\mathcal{H}}^*_{\alpha\alpha}$ and $\bar{\mathcal{H}}^*_{\gamma\gamma}$ are the diagonal matrices with elements

$$\left(\bar{\mathcal{H}}^*_{\alpha\alpha}\right)_{ii} = -\frac{1}{N-1}\sum_{j\neq i} \bar{E}[\partial_{\pi^2}\ell_{ij}]$$

$$\left(\bar{\mathcal{H}}^*_{\gamma\gamma}\right)_{ii} = -\frac{1}{N-1}\sum_{j\neq i}\bar{E}[\partial_{\pi^2}\ell_{ji}]$$

and $\bar{\mathcal{H}}^*_{\alpha\gamma} = (\bar{\mathcal{H}}^*_{\gamma\alpha})'$ has off-diagonal entries $\left(\bar{\mathcal{H}}^*_{\alpha\gamma}\right)_{ij} = -\bar{E}[\partial_{\pi^2}\ell_{ij}]/(N-1)$ and zeroes in diagonal entries. As in Lemma D.1 of FVW and Lemma A.1 in Dzemski (2019), we can show that $\bar{\mathcal{H}}^{-1}$ is dominated by its diagonal elements

$$\|\bar{\mathcal{H}}^{-1} - diag\big(\bar{\mathcal{H}}^*_{\alpha\alpha}, \bar{\mathcal{H}}^*_{\gamma\gamma}\big)^{-1}\|_{max} = O_p(N^{-1}) \tag{6}$$

We now demonstrate that the conditions in Assumption B.1 of FVW hold; a similar proof is provided by Dzemski (2019). Given the stronger moment conditions in Assumption 1.1, we can choose $q = 16$, $\epsilon = 1/(32+2\nu)$, $r_\beta = \log(N)N^{-1/4}$, and $r_\phi = N^{-3/16}$.

Assumption (i) holds trivially since $N = T$ in the network setting, while (ii) follows from Assumption 1.1.

For part (iv), by (6), the condition holds by Assumption 1.1 (v). Condition (v) follows similarly to the proofs in FVW for the panel case. For example, we have by Assumption 1.1 (iii) for $q \leq 16$

$$\begin{aligned}
&\bar{E}\Bigg[\sup_{\beta\in\mathcal{B}(r_\beta,\beta_0)}\sup_{\phi\in\mathcal{B}_q(r_\phi,\phi_0)}\frac{1}{N}\sum_i|\frac{1}{N-1}\sum_{j\neq i}\partial_{\beta_k\pi}\ell_{ij}|^q\Bigg] \\
&\leq \bar{E}\Bigg[\sup_{\beta\in\mathcal{B}(r_\beta,\beta_0)}\sup_{\phi\in\mathcal{B}_q(r_\phi,\phi_0)}\frac{1}{N}\sum_i\big(\frac{1}{N-1}\sum_{j\neq i}|\partial_{\beta_k\pi}\ell_{ij}|\big)^q\Bigg] \\
&\leq \bar{E}\Bigg[\frac{1}{N}\sum_i\frac{1}{N-1}\sum_{j\neq i}M(Z_{ij})^q\Bigg] \\
&= O_p(1)
\end{aligned}$$

and so

$$\begin{aligned}
\|\partial_{\beta_k\alpha}\mathcal{L}\|_q &= \Big(\sum_i|\frac{1}{N-1}\sum_{j\neq i}\partial_{\beta_k\pi}\ell_{ij}|^q\Big)^{1/q} \\
&= O_p(N^{1/q})
\end{aligned}$$

and similarly for $\|\partial_{\beta_k\gamma}\mathcal{L}\|_q$, and hence $\|\partial_{\beta_k\phi}\mathcal{L}\|_q$. For part (vi) the proofs in the panel case again carry over, for example see the proof in Lemma A.2 of Dzemski (2019). Inspection of that proof shows that we can get the bound $\|\tilde{\mathcal{H}}\| = O_p(N^{2\epsilon-\frac{1}{2}})$, for $\epsilon > \frac{1}{32}$, which will be

useful in the higher-order approximations.

Given Assumptions B.1 (i), (ii), (iv), (v), and (vi) we can apply Theorem B.3 in FVW to establish consistency of the estimates for common parameters and fixed effects, and to establish the bounds

$$\begin{aligned} \sup_{\beta \in \mathcal{B}(r_\beta, \beta_0)} \|\widehat{\phi}(\beta) - \phi_0\|_q &= o_p(r_\phi) \\ \|\widehat{\beta} - \beta_0\| &= O_p(N^{-1/2}) \end{aligned} \tag{7}$$

Asymptotic expansions

The next lemma gives the asymptotic expansion for the estimated fixed effects and common parameters. It is in part a restatement of Theorem B.1 in Fernández-Val and Weidner (2016), but the remainder terms are split into two parts. The expressions for the remainder terms are given in the Supplementary Appendix, and are used to derive the properties of the jackknife bias correction. The proof of Lemma A.1 is provided in the Supplementary Appendix.

Lemma A.1. *Let Assumption 1.1 hold. Then,*

$$\begin{aligned} \widehat{\phi} - \phi_0 = {} & \mathcal{H}^{-1}\mathcal{S} \\ & - W^{-1}\mathcal{H}^{-1}(\partial_{\beta\phi}\mathcal{L})\big(\mathcal{S}_\beta + (\partial_{\beta\phi}\mathcal{L})'\mathcal{H}^{-1}\mathcal{S}\big) \\ & + W^{-1}\mathcal{H}^{-1}(\partial_{\beta\phi}\mathcal{L})\frac{1}{2}\sum_g \big((\partial_{\beta\phi\phi_g}) + (\partial_{\phi\phi'\phi_g}\mathcal{L})\big)[\mathcal{H}^{-1}\mathcal{S}]_g\mathcal{H}^{-1}\mathcal{S} \\ & + \frac{1}{2}\mathcal{H}^{-1}\sum_g(\partial_{\phi\phi'\phi_g}\mathcal{L})[\mathcal{H}^{-1}\mathcal{S}]_g\mathcal{H}^{-1}\mathcal{S} \\ & + R_\phi + \tilde{R}_\phi \\ = {} & \mathcal{H}^{-1}\mathcal{S} - W^{-1}\mathcal{H}^{-1}(\partial_{\beta\phi}\mathcal{L})\big(U^{(0)} + U^{(1)}\big) \\ & + \frac{1}{2}\mathcal{H}^{-1}\sum_g(\partial_{\phi\phi'\phi_g}\mathcal{L})[\mathcal{H}^{-1}\mathcal{S}]_g\mathcal{H}^{-1}\mathcal{S} \\ & + R_\phi + \tilde{R}_\phi \end{aligned}$$

and

$$N\bar{W}_N(\widehat{\beta} - \beta_0) = U^{(0)} + U^{(1)} + R_\beta + \tilde{R}_\beta$$

where

$$
\begin{aligned}
U^{(0)} &= (\partial_{\beta}\mathcal{L}) + (\partial_{\beta\phi'}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}\mathcal{S} \\
U^{(1)} &= (\partial_{\beta\phi'}\tilde{\mathcal{L}})\bar{\mathcal{H}}^{-1}\mathcal{S} - (\partial_{\beta\phi'}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}\tilde{\mathcal{H}}\bar{\mathcal{H}}^{-1}\mathcal{S} \\
&\quad + \frac{1}{2}\sum_{g=1}^{\dim\phi}\left((\partial_{\beta\phi\phi_g}\bar{\mathcal{L}}) + (\partial_{\beta\phi'}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}(\partial_{\phi\phi'\phi_g}\bar{\mathcal{L}})\right)\left[\bar{\mathcal{H}}^{-1}\mathcal{S}\right]_g\bar{\mathcal{H}}^{-1}\mathcal{S}
\end{aligned}
$$

and the remainders satisfy $\|R_\phi\| = o_p(1)$, $\|R_\beta\| = o_p(1)$ *and* $\|\tilde{R}_\phi\| = o_p(N^{-1})$, $\|\tilde{R}_\beta\| = o_p(N^{-1})$.

A.3 Jackknife results for β

Here we allow for a more general construction of the leave-out sets, but impose two important conditions.

Condition A.1. Let $\mathcal{I}_k$ for $k = 1, \dots N_k$ be a partition of the observations in a network of size N. Define $\mathbb{1}_{ij}^k = \mathbb{1}\{(i,j) \notin \mathcal{I}_k\}$ as an indicator that the observation (i,j) is *not* included in the k-th set. We impose the following conditions on the sets:

(i) $\sum_{k=1}^{N_k}(1 - \mathbb{1}_{ij}^k) = 1$ for all (i,j)

(ii) $\sum_{j\neq i}(1 - \mathbb{1}_{ij}^k) = \sum_{i\neq j}(1 - \mathbb{1}_{ij}^k) = 1$ for all i, j, and k.

Condition A.1 imposes two important constrains on the sets $\mathcal{I}_k$: (i) that they are mutually exclusive, such that every edge (i,j) appears in exactly one of the sets, and (ii) that each set contains exactly one observation related to each of the N sender fixed effects α and each of the N receiver fixed effects γ. The first condition ensures that all observations are used equally in the jackknife, and is important for showing that the asymptotic variance of the estimator is not affected by the jackknife. The second condition ensures that each fixed effect parameter is affected equally in the leave-out sets, and that $\bar{\mathcal{H}}_{(k)}$ is well-defined and positive definite. We assume in the proofs below that $N_k = N - 1$ and that we are using the leave-one-out jackknife. All of the results still hold for the leave-l-out style jackknife for fixed l.

Some additional notation will also be useful for studying the jackknife estimates. Let A be some statistic and $A_{(k)}$ the same statistic in the k-th leave-out sample. We define the

jackknife operator $\mathbf{J}$ as

$$\mathbf{J}[A] = A_J = (N-1)A - (N-2)\frac{1}{N-1}\sum_{k=1}^{N-1} A_{(k)}$$

Additionally, we define a set of indicators that count the number of unique leave-out sets $\mathcal{I}_k$ that a group of edges $(i_1, j_1), \ldots, (i_t, j_t)$ are contained in. Let

$$I^r_{(i_1,j_1),\ldots,(i_t,j_t)} = \begin{cases} 1 & (i_1, j_1), \ldots, (i_t, j_t) \text{ span exactly } r \text{ of the sets } \mathcal{I}_k \\ 0 & \text{otherwise} \end{cases}$$

Assume that the researcher randomizes the ordering of the N node labels by choosing, with equal probability a labelling from the $N!$ possible orderings. Since the ordering of nodes is random, the indicator 1^k_{ij} is a random variable, with mean equal to the probability that $(i, j) \notin \mathcal{I}_k$ across each of the possible orderings. There are $N \times (N-2)!$ ways to order the N nodes keeping $(i, j) \in \mathcal{I}_k$ for any fixed k, so that $E[1^k_{ij}] = \frac{N-2}{N-1}$. The following lemma states that the expectation, over the randomness induced by the ordering of nodes, of sums in the leave-out sets is equal to that in the full sample.

Lemma A.2. *Let 1^k_{ij} satisfy Condition A.1, and define the sums over random variable A_{ij} in the full-sample and k-th leave-out sample*

$$A = \frac{1}{N-1}\sum_i \sum_{j \neq i} A_{ij}$$
$$A_{(k)} = \frac{1}{N-2}\sum_i \sum_{j \neq i} A_{ij} 1^k_{ij}$$

Then, we have that $\bar{E}[A] = \bar{E}[A_{(k)}]$ for any $k = 1, \ldots, N-1$.

Proof. The proof follows simply from the fact that $E[1^k_{ij}] = \frac{N-2}{N-1}$, where the expectation is taken over the randomness induced by the random ordering of nodes. Note that the ordering of nodes is independent of any other randomness in the sample, so that

$$\begin{aligned} \bar{E}[A_{ij} 1^k_{ij}] &= E[A_{ij} 1^k_{ij} | X_{ij}, \alpha_i, \gamma_j] \\ &= E[A_{ij} | X_{ij}, \alpha_i, \gamma_j] E[1^k_{ij}] \\ &= \frac{N-2}{N-1} E[A_{ij} | X_{ij}, \alpha_i, \gamma_j] \end{aligned}$$

Then we have

$$\begin{aligned}\bar{E}[A_{(k)}] &= \frac{1}{N-2}\sum_i \sum_{j\neq i} E[A_{ij}|X_{ij},\alpha_i,\gamma_j]\frac{N-2}{N-1} \\ &= \bar{E}[A]\end{aligned}$$

as required. □

An important corollary of A.2 is that $\bar{\mathcal{H}} = \bar{\mathcal{H}}_{(k)}$ and $\bar{W} = \bar{W}_{(k)}$.

Jackknifing first-order expansion terms

We next use this result to show the impact of the jackknife operator on the first-order expansion of $N(\widehat{\beta} - \beta)$. From the expansion in Lemma A.1, we have that a first-order approximation is given by

$$N\bar{W}_N(\widehat{\beta} - \beta) = \left(U^{(0)} + U^{(1)}\right) + o_p(1)$$

where

$$\begin{aligned}U^{(0)} &= (\partial_\beta \mathcal{L}) + (\partial_{\beta\phi'}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}\mathcal{S} \\ U^{(1)} &= (\partial_{\beta\phi'}\tilde{\mathcal{L}})\bar{\mathcal{H}}^{-1}\mathcal{S} - (\partial_{\beta\phi'}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}\tilde{\mathcal{H}}\bar{\mathcal{H}}^{-1}\mathcal{S} \\ &\quad + \frac{1}{2}\sum_{g=1}^{\dim\phi}\left((\partial_{\beta\phi\phi_g}\bar{\mathcal{L}}) + (\partial_{\beta\phi'}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}(\partial_{\phi\phi'\phi_g}\bar{\mathcal{L}})\right)\left[\bar{\mathcal{H}}^{-1}\mathcal{S}\right]_g\bar{\mathcal{H}}^{-1}\mathcal{S}\end{aligned}$$

The next lemmas demonstrate the effect of the jackknife on general sums of the forms in $U^{(0)}$ and $U^{(1)}$.

Lemma A.3. *Let 1_{ij}^k satisfy Condition A.1. For A_{ij} a mean-zero random variable, let*

$$\begin{aligned}A_i &= \frac{1}{N-1}\sum_{s\neq i} A_{is} \\ A_{(k),i} &= \frac{1}{N-2}\sum_{s\neq i} A_{is}1_{is}^k\end{aligned}$$

Define the jackknifed version of A_i as

$$A_{J,i} = \mathbf{J}[A_i] = (N-1)A_i - (N-2)\frac{1}{N-1}\sum_{k=1}^{N-1} A_i^k$$

Then, $A_{J,i} = A_i$.

Proof. First note that by Condition A.1, $\sum_k 1_{is}^k = N-2$. Then, $\sum_k A_{(k),i} = \sum_{s\neq i} A_{is}$ so that

$$\begin{aligned} A_{J,i} &= (N-1)A_i - (N-2)\frac{1}{N-1}\sum_{k=1}^{N-1} A_{(k),i} \\ &= (N-1)A_i - (N-2)\frac{1}{N-1}\sum_{s\neq i} A_{is} \\ &= \frac{1}{N-1}\sum_{s\neq i} A_{is} = A_i \end{aligned}$$

□

Lemma A.4. *Let 1_{ij}^k satisfy Condition A.1. For A_{ij} a mean-zero random variable with bounded fourth moment, let*

$$\begin{aligned} \mathbf{A} &= \frac{1}{N-1}\big(\{\sum_{s\neq i} A_{is}\}_{i=1,\ldots,N}, \{\sum_{s\neq j} A_{sj}\}_{j=1,\ldots,N}\big) \\ &= (\mathbf{A}_\alpha, \mathbf{A}_\gamma) \\ \mathbf{A}_k &= \frac{1}{N-2}\big(\{\sum_{s\neq i} A_{is}1_{is}^k\}_{i=1,\ldots,N}, \{\sum_{s\neq j} A_{sj}1_{sj}^k\}_{j=1,\ldots,N}\big) \\ &= (\mathbf{A}_{\alpha,k}, \mathbf{A}_{\gamma,k}) \end{aligned}$$

and let $\mathbf{B}$ and $\mathbf{B}_k$ be similarly defined vectors of sums involving mean-zero random variables B_{ij}. Assume that (A_{ij}, B_{ij}) are independent of (A_{st}, B_{st}) for $(i,j) \notin \{(s,t),(t,s)\}$. Define the jackknifed term

$$\mathcal{J}_0 = (N-1)\mathbf{A}'M\mathbf{B} - \frac{N-2}{N-1}\sum_k \mathbf{A}'_{(k)}M\mathbf{B}_{(k)}$$

where M is a non-random matrix that has $O_p(1)$ elements on its diagonal and $O_p(N^{-1})$ off-diagonal terms. Then we have: (i) $\bar{E}\big[\mathcal{J}_0\big] = o_p(1)$, and (ii) $\mathcal{J}_0 = o_p(1)$.

Proof. The most common choice of M will be $\bar{\mathcal{H}}^{-1}$, which satisfies the conditions for M by Assumption 1.1 and the approximation property in (6). We show the proof using $\bar{\mathcal{H}}^{-1}$, but note that it holds for any M satisfying the conditions stated above. We have

$$\begin{aligned}\mathbf{A}'\bar{\mathcal{H}}^{-1}\mathbf{B} = \mathbf{A}'_\alpha\bar{\mathcal{H}}^{-1}_{\alpha\alpha}\mathbf{B}_\alpha + \mathbf{A}'_\alpha\bar{\mathcal{H}}^{-1}_{\alpha\gamma}\mathbf{B}_\gamma \\ + \mathbf{A}'_\gamma\bar{\mathcal{H}}^{-1}_{\gamma\alpha}\mathbf{B}_\alpha + \mathbf{A}'_\gamma\bar{\mathcal{H}}^{-1}_{\gamma\gamma}\mathbf{B}_\gamma\end{aligned}$$

Recall that $I^1_{(ij)(st)}$ is one whenever (i,j) and (s,t) are contained in the same $\mathcal{I}_k$, and so $\sum_k 1^k_{ij}1^k_{st} = (N-2)I^1_{(ij)(st)} + (N-3)(1-I^1_{(ij)(st)})$.

$$\begin{aligned}\mathbf{A}'_\alpha\bar{\mathcal{H}}^{-1}_{\alpha\alpha}\mathbf{B}_\alpha &= \sum_{i,j}\mathbf{A}_{\alpha,i}(\bar{\mathcal{H}}^{-1}_{\alpha\alpha})_{ij}\mathbf{B}_{\alpha,j} \\ &= \frac{1}{(N-1)^2}\sum_{i,j}\sum_{s\neq i}\sum_{t\neq j}(\bar{\mathcal{H}}^{-1}_{\alpha\alpha})_{ij}A_{is}B_{jt} \\ \frac{1}{N-1}\sum_k\mathbf{A}'_{k,\alpha}\bar{\mathcal{H}}^{-1}_{\alpha\alpha}\mathbf{B}_{k,\alpha} &= \frac{1}{(N-1)(N-2)^2}\sum_k\sum_{i,j}\sum_{s\neq i}\sum_{t\neq j}(\bar{\mathcal{H}}^{-1}_{\alpha\alpha})_{ij}A_{is}B_{jt}1^k_{is}1^k_{jt} \\ &= \frac{1}{(N-1)(N-2)}\sum_{i,j}\sum_{s\neq i}\sum_{t\neq j}(\bar{\mathcal{H}}^{-1}_{\alpha\alpha})_{ij}A_{is}B_{jt}I^1_{(is)(jt)} \\ &+ \frac{N-3}{(N-1)(N-2)^2}\sum_{i,j}\sum_{s\neq i}\sum_{t\neq j}(\bar{\mathcal{H}}^{-1}_{\alpha\alpha})_{ij}A_{is}B_{jt}(1-I^1_{(is)(jt)})\end{aligned}$$

Then, we have

$$\begin{aligned}\mathcal{J}_{\alpha\alpha} =&(N-1)\mathbf{A}'_\alpha\bar{\mathcal{H}}^{-1}_{\alpha\alpha'}\mathbf{B}_\alpha - \frac{N-2}{N-1}\sum_k\mathbf{A}'_{\alpha,k}\bar{\mathcal{H}}^{-1}_{\alpha\alpha'}\mathbf{B}_{\alpha,k} \\ =&\frac{1}{(N-1)(N-2)}\sum_i\sum_j\sum_{s\neq i}\sum_{t\neq j}(\bar{\mathcal{H}}^{-1}_{\alpha\alpha'})_{ij}\big(A_{is}B_{jt}\big)(1-I^1_{(is)(jt)})\end{aligned}$$

Similar computations for the other three elements gives

$$\begin{aligned}\mathcal{J}_0 =&(N-1)\mathbf{A}'\bar{\mathcal{H}}^{-1}\mathbf{B} - \frac{N-2}{N-1}\sum_k\mathbf{A}'_k\bar{\mathcal{H}}^{-1}\mathbf{B}_k \\ =&\frac{1}{(N-1)(N-2)}\sum_i\sum_j\sum_{s\neq i}\sum_{t\neq j}\Big\{(\bar{\mathcal{H}}^{-1}_{\alpha\alpha'})_{ij} + (\bar{\mathcal{H}}^{-1}_{\alpha\gamma'})_{it} \\ &+ (\bar{\mathcal{H}}^{-1}_{\gamma\alpha'})_{sj} + (\bar{\mathcal{H}}^{-1}_{\gamma\gamma'})_{st}\Big\}\big(A_{is}B_{jt}\big)(1-I^1_{(is)(jt)})\end{aligned}$$

Now, since $\bar{E}[A_{is}B_{jt}(1-I^1_{(is)(jt)})] \neq 0$ only when $(j,t)=(s,i)$, we have

$$\begin{aligned}\bar{E}[\mathcal{J}_0] =& \frac{1}{(N-1)(N-2)}\sum_i\sum_{s\neq i}\Big\{(\bar{\mathcal{H}}^{-1}_{\alpha\alpha'})_{is}+(\bar{\mathcal{H}}^{-1}_{\alpha\gamma'})_{ii}\\ &+(\bar{\mathcal{H}}^{-1}_{\gamma\alpha'})_{ss}+(\bar{\mathcal{H}}^{-1}_{\gamma\gamma'})_{si}\Big\}\big(A_{is}B_{si}\big)\\ =& o_p(1)\end{aligned}$$

since, for for $i\neq s$, we have $(\bar{\mathcal{H}}^{-1}_{\alpha\alpha'})_{is}=O_p(N^{-1})$ and $(\bar{\mathcal{H}}^{-1}_{\gamma\gamma'})_{si}=O_p(N^{-1})$, while $(\bar{\mathcal{H}}^{-1}_{\alpha\gamma'})_{ii}$ and $(\bar{\mathcal{H}}^{-1}_{\gamma\alpha'})_{ss}$ are both $O_p(N^{-1})$ also.

Next, let $\Gamma_{ijst}=(\bar{\mathcal{H}}^{-1}_{\alpha\alpha'})_{ij}+(\bar{\mathcal{H}}^{-1}_{\alpha\gamma'})_{it}+(\bar{\mathcal{H}}^{-1}_{\gamma\alpha'})_{sj}+(\bar{\mathcal{H}}^{-1}_{\gamma\gamma'})_{st}$. We have

$$\Gamma_{ijst}=\begin{cases}O_p(1) & i=j \text{ or } s=t\\ O_p(N^{-1}) & \text{otherwise}\end{cases}$$

We can decompose $\mathcal{J}_0$ as

$$\begin{aligned}\mathcal{J}_0 =& \frac{1}{(N-1)(N-2)}\sum_i\sum_j\Big(\sum_{s<i}\sum_{t<j}\Gamma_{ijst}A_{is}B_{jt}+\sum_{s<i}\sum_{t>j}\Gamma_{ijst}A_{is}B_{jt}\\ &+\sum_{s>i}\sum_{t<j}\Gamma_{ijst}A_{is}B_{jt}+\sum_{s>i}\sum_{t>j}\Gamma_{ijst}A_{is}B_{jt}\Big)\\ =& \mathcal{J}_{0,11}+\mathcal{J}_{0,12}+\mathcal{J}_{0,21}+\mathcal{J}_{0,22}\end{aligned}$$

Then we have

$$\begin{aligned}\bar{E}[\mathcal{J}^2_{0,11}] =& \frac{1}{(N-1)^2(N-2)^2}\sum_i\sum_j\sum_k\sum_l\sum_{s<i}\sum_{t<j}\sum_{p<k}\sum_{q<l}\Gamma_{ijst}\Gamma_{klpq}\\ &\times E\big[A_{is}B_{jt}A_{kp}B_{lq}\big](1-I^1_{(is)(jt)})(1-I^1_{(kp)(lq)})\\ =& \frac{1}{(N-1)^2(N-2)^2}\sum_i\sum_j\sum_{s<i}\sum_{t<j}\Gamma^2_{ijst}\bar{E}[A^2_{is}B^2_{jt}](1-I^1_{(is)(jt)})\\ &+\frac{1}{(N-1)^2(N-2)^2}\sum_i\sum_j\sum_{s<i}\sum_{t<j}\Gamma_{ijst}\Gamma_{stij}\bar{E}[A_{is}B_{is}A_{jt}B_{jt}\big](1-I^1_{(is)(jt)})\\ =& \frac{1}{(N-1)^2(N-2)^2}\sum_i\sum_{s<i}\sum_{t<i}\Gamma^2_{iist}\bar{E}[A^2_{is}]\bar{E}[B^2_{it}](1-I^1_{(is)(it)})\\ &+\frac{1}{(N-1)^2(N-2)^2}\sum_i\sum_{j\neq i}\sum_{s<(i\wedge j)}\Gamma^2_{ijss}\bar{E}[A^2_{is}]\bar{E}[B^2_{js}](1-I^1_{(is)(js)})\end{aligned}$$

$$
\begin{aligned}
&+ \frac{1}{N^2(N-1)^2(N-2)^2} \sum_i \sum_{j\neq i} \sum_{s<i} \sum_{t<j, t\neq s} N^2 \Gamma_{ijst}^2 \bar{E}[A_{is}^2]\bar{E}[B_{jt}^2](1 - I^1_{(is)(jt)}) \\
&+ \frac{1}{(N-1)^2(N-2)^2} \sum_i \sum_{s<i} \sum_{t<i} \Gamma_{iist}\Gamma_{stii} \bar{E}[A_{is}B_{is}]\bar{E}[A_{it}B_{it}](1 - I^1_{(is)(it)}) \\
&+ \frac{1}{(N-1)^2(N-2)^2} \sum_i \sum_{j\neq i} \sum_{s\neq (i\wedge j)} \Gamma_{ijss}\Gamma_{ssij} \bar{E}[A_{is}B_{is}]\bar{E}[A_{js}B_{js}](1 - I^1_{(is)(js)}) \\
&+ \frac{1}{N^2(N-1)^2(N-2)^2} \sum_i \sum_{j\neq i} \sum_{s<i} \sum_{t<j, t\neq s} N^2 \Gamma_{ijst}\Gamma_{stij} \bar{E}[A_{is}B_{is}]\bar{E}[A_{jt}B_{jt}](1 - I^1_{(is)(jt)}) \\
&= O_p(N^{-1})
\end{aligned}
$$

Where the last line follows from the properties of Γ_{ijst}. The same result holds for $\bar{E}[\mathcal{J}_{0,12}^2]$, $\bar{E}[\mathcal{J}_{0,21}^2]$, and $\bar{E}[\mathcal{J}_{0,22}^2]$, hence $\mathcal{J}_0 = O_p(N^{-1/2}) = o_p(1)$. □

The following lemma derives the forms of $\mathbf{J}[U^{(0)}]$ and $\mathbf{J}[U^{(1)}]$.

Lemma A.5. *Let Assumption 1.1 hold. Then*

$$
\begin{aligned}
\mathbf{J}[U^{(0)}] &= U^{(0)} \\
\mathbf{J}[U^{(1)}] &= o_p(1)
\end{aligned}
$$

Proof. For $U^{(0)} = (\partial_\beta \mathcal{L}) + (\partial_{\beta\phi'}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}\mathcal{S}$ we can appeal to Lemma A.3 with $A_i = (\partial_\beta \mathcal{L})_i = \frac{1}{N-1}\sum_{j\neq i} \partial_\beta \ell_{ij}$ and with $A_i = \mathcal{S}_i = \frac{1}{N-1}\sum_{j\neq i} \partial_\pi \ell_{ij}$. Note that the jackknife operator is linear so that, since $(\partial_{\beta\phi'}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}$ are fixed across leave-out samples (by Lemma A.2),

$$
\begin{aligned}
\mathbf{J}[\partial_\beta \mathcal{L}] &= \sum_i \mathbf{J}[(\partial_\beta \mathcal{L})_i] \\
\mathbf{J}\left[(\partial_{\beta\phi'}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}\mathcal{S}\right] &= (\partial_{\beta\phi'}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}\mathbf{J}[\mathcal{S}]
\end{aligned}
$$

Now, for $U^{(1)}$ we can appeal to Lemma A.4. For the first term, $(\partial_{\beta\phi'}\tilde{\mathcal{L}})\bar{\mathcal{H}}^{-1}\mathcal{S}$ we set $A_{ij} = \partial_{\beta\pi}\tilde{\ell}_{ij}$ and $B_{ij} = \partial_\pi \ell_{ij}$. For the second term, $(\partial_{\beta\phi'}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}\tilde{\mathcal{H}}\bar{\mathcal{H}}^{-1}\mathcal{S}$ we note that

$$
(\partial_{\beta\phi'}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}\tilde{\mathcal{H}} = \left((\partial_{\beta\phi'}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}\tilde{\mathcal{H}}_{\cdot,\alpha}, (\partial_{\beta\phi'}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}\tilde{\mathcal{H}}_{\cdot,\gamma}\right)
$$

where the i-th element of $(\partial_{\beta\phi'}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}\tilde{\mathcal{H}}_{\cdot,\alpha}$ is

$$
(\partial_{\beta\phi'}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}\tilde{\mathcal{H}}_{\cdot,\alpha_i} = (\partial_{\beta\alpha'}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}_{\alpha\alpha}\tilde{\mathcal{H}}_{\alpha\alpha_i} + (\partial_{\beta\alpha'}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}_{\alpha\gamma}\tilde{\mathcal{H}}_{\gamma\alpha_i}
$$

$$
\begin{aligned}
&+ (\partial_{\beta\gamma'}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}_{\gamma\alpha}\tilde{\mathcal{H}}_{\alpha\alpha_i} + (\partial_{\beta\gamma'}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}_{\gamma\gamma}\tilde{\mathcal{H}}_{\gamma\alpha_i} \\
&= -(\partial_{\beta\alpha'}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}_{\alpha\alpha_i}\frac{1}{N-1}\sum_{j\neq i}\partial_{\pi^2}\tilde{\ell}_{ij} - \frac{1}{N-1}\sum_{j\neq i}(\partial_{\beta\alpha'}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}_{\alpha\gamma_j}\partial_{\pi^2}\tilde{\ell}_{ij} \\
&- (\partial_{\beta\gamma'}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}_{\gamma\alpha_i}\frac{1}{N-1}\sum_{j\neq i}\partial_{\pi^2}\tilde{\ell}_{ij} - \frac{1}{N-1}\sum_{j\neq i}(\partial_{\beta\gamma'}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}_{\gamma\gamma_j}\partial_{\pi^2}\tilde{\ell}_{ij}
\end{aligned}
$$

and similarly for elements of $(\partial_{\beta\phi'}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}\tilde{\mathcal{H}}_{\cdot,\gamma}$. So we can let

$$
A_{ij} = \big((\partial_{\beta\alpha'}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}_{\alpha\alpha_i} + (\partial_{\beta\alpha'}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}_{\alpha\gamma_j} + (\partial_{\beta\gamma'}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}_{\gamma\alpha_i} + (\partial_{\beta\gamma'}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}_{\gamma\gamma_j}\big)\partial_{\pi^2}\tilde{\ell}_{ij}
$$

and $B_{ij} = \partial_\pi \ell_{ij}$ in Lemma A.4. Note that we have A_{ij} and B_{ij} mean zero and the moment condition also holds by assumption. For the final term, we begin by demonstrating that both $\bar{\mathcal{H}}^{-1}\sum_g(\partial_{\beta\phi'\phi_g}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}$and $\bar{\mathcal{H}}^{-1}\sum_g(\partial_{\phi\phi'\phi_g}\bar{\mathcal{L}})\big[\bar{\mathcal{H}}^{-1}(\partial_{\beta\phi}\bar{\mathcal{L}})\big]_g\bar{\mathcal{H}}^{-1}$are matrices that satisfy the requirements for M in Lemma A.4. The proof for the second term is shown, with the result for the first term following nearly identically. Firstly,

$$
\begin{aligned}
&\sum_g(\partial_{\phi\phi'\phi_g}\bar{\mathcal{L}})\big[\bar{\mathcal{H}}^{-1}(\partial_{\beta\phi}\bar{\mathcal{L}})\big]_g \\
&\quad= \sum_{s,t}(\partial_{\phi\phi'\alpha_s}\bar{\mathcal{L}})\big(\bar{\mathcal{H}}^{-1}_{\alpha\alpha}\big)_{st}(\partial_{\beta\alpha_t}\bar{\mathcal{L}}) + \sum_{s,t}(\partial_{\phi\phi'\gamma_s}\bar{\mathcal{L}})\big(\bar{\mathcal{H}}^{-1}_{\gamma\alpha}\big)_{st}(\partial_{\beta\alpha_t}\bar{\mathcal{L}}) \\
&\quad+ \sum_{s,t}(\partial_{\phi\phi'\alpha_s}\bar{\mathcal{L}})\big(\bar{\mathcal{H}}^{-1}_{\alpha\gamma}\big)_{st}(\partial_{\beta\gamma_t}\bar{\mathcal{L}}) + \sum_{s,t}(\partial_{\phi\phi'\gamma_s}\bar{\mathcal{L}})\big(\bar{\mathcal{H}}^{-1}_{\gamma\gamma}\big)_{st}(\partial_{\beta\gamma_t}\bar{\mathcal{L}})
\end{aligned}
$$

Taking the first of these terms, the (i,j) element is given by

$$
\begin{aligned}
&\big[\sum_{s,t}(\partial_{\phi\phi'\alpha_s}\bar{\mathcal{L}})\big(\bar{\mathcal{H}}^{-1}_{\alpha\alpha}\big)_{st}(\partial_{\beta\alpha_t}\bar{\mathcal{L}})\big]_{ij} \\
&\quad= \sum_t(\partial_{\phi_i\phi'_j\alpha_t}\bar{\mathcal{L}})\big(\bar{\mathcal{H}}^{-1}_{\alpha\alpha}\big)_{tt}(\partial_{\beta\alpha_t}\bar{\mathcal{L}}) + \frac{1}{N}\sum_t\sum_{s\neq t}(\partial_{\phi_i\phi'_j\alpha_s}\bar{\mathcal{L}})\big(N\bar{\mathcal{H}}^{-1}_{\alpha\alpha}\big)_{st}(\partial_{\beta\alpha_t}\bar{\mathcal{L}})
\end{aligned}
$$

Now if $\phi_i = \phi_j = \alpha_i$ then

$$
\sum_t(\partial_{\phi_i\phi'_j\alpha_t}\bar{\mathcal{L}})\big(\bar{\mathcal{H}}^{-1}_{\alpha\alpha}\big)_{tt}(\partial_{\beta\alpha_t}\bar{\mathcal{L}}) = (\partial_{\alpha_i^3}\bar{\mathcal{L}})\big(\bar{\mathcal{H}}^{-1}_{\alpha\alpha}\big)_{ii}(\partial_{\beta\alpha_i}\bar{\mathcal{L}}) = O_p(1)
$$

and if $\phi_i = \phi_j = \gamma_i$ then

$$\sum_t (\partial_{\phi_i \phi'_j \alpha_t} \bar{\mathcal{L}}) \big(\bar{\mathcal{H}}^{-1}_{\alpha\alpha}\big)_{tt} (\partial_{\beta\alpha_t} \bar{\mathcal{L}}) = \frac{1}{N-1} \sum_{t \neq i} (\partial_{\alpha_t \gamma_i^2} \bar{\ell}_{ti}) \big(\bar{\mathcal{H}}^{-1}_{\alpha\alpha}\big)_{tt} (\partial_{\beta\alpha_t} \bar{\mathcal{L}}) = O_p(1)$$

Finally, if $i \neq j$ then we have either 0, or

$$\sum_t (\partial_{\phi_i \phi'_j \alpha_t} \bar{\mathcal{L}}) \big(\bar{\mathcal{H}}^{-1}_{\alpha\alpha}\big)_{tt} (\partial_{\beta\alpha_t} \bar{\mathcal{L}}) = \frac{1}{N-1} (\partial_{\pi\alpha_t \gamma_i} \bar{\ell}_{ti}) \big(\bar{\mathcal{H}}^{-1}_{\alpha\alpha}\big)_{tt} (\partial_{\beta\alpha_t} \bar{\mathcal{L}}) = O_p(N^{-1})$$

Identical results apply to the other elements in $\sum_g (\partial_{\phi\phi'\phi_g} \bar{\mathcal{L}}) \big[\bar{\mathcal{H}}^{-1} (\partial_{\beta\phi} \bar{\mathcal{L}})\big]_g$ and hence we can conclude that the matrix has $O_p(1)$ diagonal elements and $O_p(N^{-1})$ off-diagonal elements. It then follows that the same is true of $\bar{\mathcal{H}}^{-1} \sum_g (\partial_{\phi\phi'\phi_g} \bar{\mathcal{L}}) \big[\bar{\mathcal{H}}^{-1} (\partial_{\beta\phi} \bar{\mathcal{L}})\big]_g \bar{\mathcal{H}}^{-1}$. Then, we can apply Lemma A.4 with $\mathbf{A} = \mathbf{B} = \mathcal{S}$ to give the result. □

Lemma A.6. *Let Assumption 1.1 hold, and let $\widehat{\beta}_J$ be the either the jackknife, leave-l-out jackknife, or weighted jackknife estimator. Then, a first-order approximation to the estimator is given by*

$$\bar{W}_N N(\widehat{\beta}_J - \beta_0) = U^{(0)} + o_p(1)$$

where $U^{(0)} = (\partial_\beta \mathcal{L}) + (\partial_{\beta\phi'} \bar{\mathcal{L}}) \bar{\mathcal{H}}^{-1} \mathcal{S}$.

Proof. Recall from Lemma A.1 that

$$N \bar{W}_N (\widehat{\beta} - \beta_0) = U^{(0)} + U^{(1)} + R_\beta + \tilde{R}_\beta$$

Since $\bar{W}_N$ is fixed across leave-out samples (Lemma A.2), we can focus on the jackknife operator applied to the RHS. By Lemma A.5 we have that $\mathbf{J}[U^{(0)} + U^{(1)}] = U^{(0)} + o_p(1)$, while in the Supplementary Appendix (S.4) it is shown that $\mathbf{J}[R_\beta] = o_p(1)$. Finally,

$$\begin{aligned} \mathbf{J}[\tilde{R}_\beta] &= (N-1)\tilde{R}_\beta - (N-2)\frac{1}{N-1} \sum_k \tilde{R}_{\beta,(k)} \\ &= o_p(1) \end{aligned}$$

since each remainder term in the above is $o_p(N^{-1})$. □

Approximation of W_N

The next two results show that the sample version of the Hessian for the common parameters β is consistent, and that it is approximately the same across leave-out samples.

Lemma A.7. *Let Assumption 1.1 hold. Then, for $\epsilon \geq \frac{1}{32}$*

$$\|W_N - \bar{W}_N\| = O_p(N^{-\frac{1}{2}+2\epsilon})$$
$$\|W_{N,(k)} - \bar{W}_N\| = O_p(N^{-\frac{1}{2}+2\epsilon})$$

Let $\tilde{W}_N = W_N - \bar{W}_N$, then

$$\tilde{W}_N = \frac{1}{N}\partial_{\beta\beta}\tilde{\mathcal{L}} + \frac{1}{N}\big((\partial_{\beta\phi'}\mathcal{L})\mathcal{H}^{-1}(\partial_{\beta\phi}\mathcal{L}) - (\partial_{\beta\phi'}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}(\partial_{\beta\phi}\bar{\mathcal{L}})\big)$$

The first term is $\frac{1}{N}\partial_{\beta\beta}\tilde{\mathcal{L}} = O_p(N^{-1})$, since

$$\begin{aligned}
\frac{1}{N^2}\bar{E}\big[(\partial_{\beta\beta}\tilde{\mathcal{L}})^2\big] &= \frac{1}{N^2(N-1)^2}\sum_{i,s}\sum_{j\neq i}\sum_{t\neq s}\bar{E}\big[(\partial_{\beta\beta'}\tilde{\ell}_{ij})(\partial_{\beta\beta'}\tilde{\ell}_{st})\big] \\
&= \frac{1}{N^2(N-1)^2}\sum_{i}\sum_{j\neq i}\big(\bar{E}\big[(\partial_{\beta\beta'}\tilde{\ell}_{ij})^2\big] + \bar{E}\big[(\partial_{\beta\beta'}\tilde{\ell}_{ij})(\partial_{\beta\beta'}\tilde{\ell}_{ji})\big]\big) \\
&= O_p(N^{-2})
\end{aligned}$$

For the remaining term, we can decompose it as

$$\begin{aligned}
&\frac{1}{N}(\partial_{\beta\phi'}\mathcal{L})\mathcal{H}^{-1}(\partial_{\beta\phi}\mathcal{L}) - \frac{1}{N}(\partial_{\beta\phi'}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}(\partial_{\beta\phi}\bar{\mathcal{L}}) \\
&\qquad = \frac{1}{N}(\partial_{\beta\phi'}\tilde{\mathcal{L}})\bar{\mathcal{H}}^{-1}(\partial_{\beta\phi}\bar{\mathcal{L}}) + \frac{1}{N}(\partial_{\beta\phi'}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}(\partial_{\beta\phi}\tilde{\mathcal{L}}) \\
&\qquad + \frac{1}{N}(\partial_{\beta\phi'}\tilde{\mathcal{L}})\bar{\mathcal{H}}^{-1}(\partial_{\beta\phi}\tilde{\mathcal{L}}) + \frac{1}{N}(\partial_{\beta\phi'}\mathcal{L})\big(\mathcal{H}^{-1} - \bar{\mathcal{H}}^{-1}\big)(\partial_{\beta\phi}\mathcal{L})
\end{aligned}$$

By Assumption B.1 of FVW we have

$$\frac{1}{N}\|(\partial_{\beta\phi'}\tilde{\mathcal{L}})\bar{\mathcal{H}}^{-1}(\partial_{\beta\phi}\bar{\mathcal{L}})\| \leq O_p(N^{-1/2})$$
$$\frac{1}{N}\|(\partial_{\beta\phi'}\tilde{\mathcal{L}})\bar{\mathcal{H}}^{-1}(\partial_{\beta\phi}\tilde{\mathcal{L}})\| \leq O_p(N^{-1})$$

Also,

$$\begin{aligned}\frac{1}{N}\|(\partial_{\beta\phi'}\mathcal{L})\big(\mathcal{H}^{-1}-\bar{\mathcal{H}}^{-1}\big)(\partial_{\beta\phi}\mathcal{L})\| &\leq \frac{1}{N}\|\partial_{\beta\phi'}\mathcal{L}\|^2\|\mathcal{H}^{-1}-\bar{\mathcal{H}}^{-1}\| \\ &= O_p(N^{-\frac{1}{2}+2\epsilon})\end{aligned}$$

So we may write $\tilde{W}_N = O_p(N^{-\frac{1}{2}+2\epsilon})$. For the leave-out term note that the moment bounds in Assumption 1.1 (iii) imply identical bounds in the leave-out samples, simply by replacing $\partial_{\beta\pi}\ell_{ij}$ with $(\partial_{\beta\pi}\ell_{ij})1^k_{ij}\frac{N-1}{N-2}$ since we can condition on the node labels so that $1^k_{ij}\frac{N-1}{N-2}$ is simply an $O(1)$ constant. We can therefore apply the bounds in Assumption B.1 of FVW to the leave-out sample, as well as $\|\tilde{\mathcal{H}}_{(k)}\| = O_p(N^{-\frac{1}{2}+2\epsilon})$, and hence $\|\mathcal{H}^{-1}_{(k)} - \bar{\mathcal{H}}^{-1}\| = O_p(N^{-\frac{1}{2}+2\epsilon})$. Then, similar steps to the above proof also give $\|W_{N,(k)} - \bar{W}_N\| = O_p(N^{-\frac{1}{2}+2\epsilon})$.

Lemma A.8. *Let Assumption 1.1 hold, then for all k*

$$\begin{aligned}\|\widehat{W}_N - W_N\| &\to 0 \\ \|\widehat{W}_{N,(k)} - W_{N,(k)}\| &\to 0\end{aligned}$$

Proof. We prove the first statement, since the proof of the second is identical. A first-order Taylor expansion of $\widehat{W}_N$ gives

$$\begin{aligned}\widehat{W}_N = W_N(\widehat{\beta},\widehat{\phi}) = W_N(\beta_0,\phi_0) &+ \partial_\beta W_N(\bar{\beta},\bar{\phi})(\widehat{\beta}-\beta_0) \\ &+ \partial_{\phi'} W_N(\bar{\beta},\bar{\phi})(\widehat{\phi}-\phi_0)\end{aligned}$$

where $\bar{\beta}$ and $\bar{\phi}$ are intermediate values between (β_0,ϕ_0) and $(\widehat{\beta},\widehat{\phi})$. From Assumption 1.1 and the bounds in Assumption B.1 of FVW, we have that W_N is differentiable with derivatives that are $O_p(1)$, since

$$\begin{aligned}\partial_\beta W_N &= \frac{1}{N}\partial_{\beta\beta\beta}\mathcal{L} + \frac{2}{N}(\partial_{\beta\beta\phi'}\mathcal{L})\mathcal{H}^{-1}(\partial_{\beta\phi}\mathcal{L}) \\ &\quad - \frac{1}{N}(\partial_{\beta\phi'}\mathcal{L})\mathcal{H}^{-1}(\partial_{\beta\phi\phi'}\mathcal{L})\mathcal{H}^{-1}(\partial_{\beta\phi}\mathcal{L}) \\ &= O_p(1)\end{aligned}$$

where $\frac{1}{N}(\partial_{\beta\phi'}\mathcal{L})\mathcal{H}^{-1}(\partial_{\beta\phi\phi'}\mathcal{L})\mathcal{H}^{-1}(\partial_{\beta\phi}\mathcal{L}) = O_p(1)$ follows simply by expansion of the matrix product and the properties of $\mathcal{H}^{-1}$ in (6). Similarly,

$$
\begin{aligned}
(\partial_{\phi'} W_N)(\widehat{\phi} - \phi_0) &= \frac{1}{N}(\partial_{\beta\beta\phi'}\mathcal{L})(\widehat{\phi} - \phi_0) \\
&\quad + \frac{2}{N}(\partial_{\beta\phi'}\mathcal{L})\mathcal{H}^{-1}(\partial_{\beta\phi\phi'}\mathcal{L})(\widehat{\phi} - \phi_0) \\
&\quad - \frac{1}{N}\sum_{g=1}^{2N}(\partial_{\beta\phi'}\mathcal{L})\mathcal{H}^{-1}(\partial_{\phi\phi'\phi_g}\mathcal{L})\mathcal{H}^{-1}(\partial_{\beta\phi}\mathcal{L})(\widehat{\phi}_g - \phi_g) \\
&\leq 2\|\widehat{\phi} - \phi\|_\infty |\frac{1}{N(N-1)}\sum_i \sum_{j\neq i}(\partial_{\beta\beta\pi}\ell_{ij})| \\
&\quad + \|\widehat{\phi} - \phi\|_\infty |\frac{2}{N}\sum_{g=1}^{2N}(\partial_{\beta\phi'}\mathcal{L})\mathcal{H}^{-1}(\partial_{\beta\phi\phi_g}\mathcal{L})| \\
&\quad + \|\widehat{\phi} - \phi\|_\infty |\frac{1}{N}\sum_{g=1}^{2N}(\partial_{\beta\phi'}\mathcal{L})\mathcal{H}^{-1}(\partial_{\phi\phi'\phi_g}\mathcal{L})\mathcal{H}^{-1}(\partial_{\beta\phi}\mathcal{L})| \\
&= O_p(1) \times \|\widehat{\phi} - \phi\|_\infty
\end{aligned}
$$

Then, since $\|\widehat{\beta} - \beta_0\| \to 0$ and $\|\widehat{\phi} - \phi\|_\infty \to 0$ by (7), we get the result. □

Proof of Theorem 1.1

From Lemma A.6 we have that

$$
\begin{aligned}
\bar{W}_N N(\widehat{\beta}_J - \beta_0) &= (\partial_\beta \mathcal{L}) + (\partial_{\beta\phi'}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}\mathcal{S} + o_p(1) \\
&= \frac{1}{N-1}\sum_i \sum_{j<i}\left(D_\beta \ell_{ij} + D_\beta \ell_{ji}\right)
\end{aligned}
$$

where $D_\beta \ell_{ij} = \partial_\beta \ell_{ij} - \partial_\pi \ell_{ij}\Xi_{ij}$ for

$$
\begin{aligned}
\Xi_{ij} &= -\frac{1}{N-1}\sum_s \sum_{t\neq s}\Gamma_{ijst}\bar{E}[\partial_{\beta\pi}\ell_{st}] \\
\Gamma_{ijst} &= (\bar{\mathcal{H}}^{-1}_{\alpha\alpha})_{is} + (\bar{\mathcal{H}}^{-1}_{\gamma\alpha})_{js} + (\bar{\mathcal{H}}^{-1}_{\alpha\gamma})_{it} + (\bar{\mathcal{H}}^{-1}_{\gamma\gamma})_{jt}
\end{aligned}
$$

The result then follows from a standard CLT argument, noting that $\left(D_\beta \ell_{ij} + D_\beta \ell_{ji}\right)$ are independent over (i, j). For the weighted jackknife, we use the fact that Lemmas A.7 and

A.8, and the triangle inequality give $\|\bar{W}_J - \bar{W}_N\| = o_p(1)$ and hence

$$\|\bar{W}_J^{-1}\widehat{W}_{(k)} - I\| \leq \|\bar{W}_J^{-1}\|\|\widehat{W}_{N,(k)} - \bar{W}_J\| = o_p(1)$$

so that

$$\begin{aligned}\frac{1}{N-1}\sum_k \bar{W}_J^{-1}\widehat{W}_{(k)}\widehat{\beta}_{(k)} &= \frac{1}{N-1}\sum_k \widehat{\beta}_{(k)} + \frac{1}{N-1}\sum_k \Big(\bar{W}_J^{-1}\widehat{W}_{(k)} - I\Big)\widehat{\beta}_{(k)} \\ &= \frac{1}{N-1}\sum_k \widehat{\beta}_{(k)} + o_p(1)\end{aligned}$$

and hence the weighted jackknife is equal to the standard jackknife to first-order.

To show consistency of the plug-in estimator $\widehat{\Omega}_N$, we note that by Assumption 1.1 (iii), $D_\beta \ell_{ij}$ has a first-order Taylor approximation that is a continuously differentiable function of the parameters. Then by the continuous mapping theorem and the consistency results $\|\widehat{\beta} - \beta_0\| \to 0$ and $\|\widehat{\phi} - \phi\|_\infty \to 0$ in (7), $\widehat{\Omega}_N \to \Omega$ as required (see for example Lemma S.1 in FVW).

A.4 Jackknife results for average effects

We begin by stating a first-order asymptotic expansion for the average effect estimator that will be used in the proof of Theorem 1.2. The proof of this result is provided in the Supplementary Appendix.

Lemma A.9. *Let Assumptions 1.1 and 1.2 hold. Then*

$$\begin{aligned}N(\widehat{\Delta}_N - \Delta_N) &= (\partial_\beta \Delta_N)N(\widehat{\beta} - \beta) + N(\partial_{\phi'}\Delta_N)(\widehat{\phi} - \phi) \\ &\quad + \frac{1}{2}N(\widehat{\phi} - \phi)'(\partial_{\phi\phi'}\Delta_N)(\widehat{\phi} - \phi) + R^1_\Delta + \tilde{R}^1_\Delta \\ &= \Big[(\partial_\beta \bar{\Delta}_N) - (\partial_{\phi'}\bar{\Delta}_N)\bar{\mathcal{H}}^{-1}(\partial_{\beta\phi}\bar{\mathcal{L}})\Big]\bar{W}_N^{-1}\big(U^{(0)} + U^{(1)}\big) \\ &\quad + N(\partial_{\phi'}\bar{\Delta}_N)\bar{\mathcal{H}}^{-1}\mathcal{S} \\ &\quad + N(\partial_{\phi'}\tilde{\Delta}_N)\bar{\mathcal{H}}^{-1}\mathcal{S} - N(\partial_{\phi'}\bar{\Delta}_N)\bar{\mathcal{H}}^{-1}\tilde{\mathcal{H}}\bar{\mathcal{H}}^{-1}\mathcal{S} \\ &\quad + \frac{1}{2}N\mathcal{S}'\bar{\mathcal{H}}^{-1}\Big((\partial_{\phi\phi'}\bar{\Delta}_N) + \sum_g (\partial_{\phi\phi'\phi_g}\bar{\mathcal{L}})[(\partial_{\phi'}\bar{\Delta}_N)\bar{\mathcal{H}}^{-1}]_g\Big)\bar{\mathcal{H}}^{-1}\mathcal{S} \\ &\quad + R_\Delta + \tilde{R}_\Delta\end{aligned}$$

where $\|R_\Delta^1\| = o_p(1)$, $\|\tilde{R}_\Delta^1\| = o_p(N^{-1})$, $\|R_\Delta\| = o_p(1)$, *and* $\|\tilde{R}_\Delta\| = o_p(N^{-1})$.

In order to establish an equivalent asymptotic expansion for the leave-out estimators $\widehat{\Delta}_{(k)}$ we need to determine the value of expectations in the leave-out samples. The next lemma does this, and is analogous to Lemma A.2, which states the same result for averages over single observations.

Lemma A.10. *Let* 1_{ij}^k *satisfy Condition A.1, and define* $1_\lambda^k = \prod_{(i,j)\in\lambda} 1_{ij}^k$ *for* λ *a set of* r *observations* (i,j). *Then, for sums*

$$A = \frac{1}{|\Lambda_N|}\sum_\lambda A_\lambda$$
$$A_{(k)} = \frac{N-1}{N-r-1}\frac{1}{|\Lambda_N|}\sum_\lambda A_\lambda 1_\lambda^k$$

we have

$$\bar{E}[A_{(k)}] = \bar{E}[A]\big(1 + O(N^{-2})\big)$$

Proof. We will prove that $E[1_\lambda^k] = \frac{N-r-1}{N-1} + O(N^{-2})$ from which the statement in the lemma follows since

$$\begin{aligned}
\bar{E}[A_{(k)}] &= \frac{N-1}{N-r-1}\frac{1}{|\Lambda_N|}\sum_\lambda \bar{E}[A_\lambda]E[1_\lambda^k] \\
&= \frac{1}{|\Lambda_N|}\sum_\lambda \bar{E}[A_\lambda] + \frac{N-1}{N-r-1}\frac{1}{|\Lambda_N|}\sum_\lambda \bar{E}[A_\lambda]\times O(N^{-2}) \\
&= \bar{E}[A]\big(1 + O(N^{-2})\big)
\end{aligned}$$

Let I_λ be equal to the number of leave-out sets $\mathcal{I}_k$ spanned by the observations in λ. Since any observation is equally likely to appear in any leave-out set,

$$E[1_\lambda^k] = \sum_{s=1}^{r} E[1_\lambda^k | I_\lambda = s]P(I_\lambda = s)$$

We begin by arguing that $P(I_\lambda = r) = 1 - O(N^{-1})$, that is, the sets λ do not contain multiple observations from within the same leave-out set with probability approaching one. It is sufficient to consider sets λ that contain $p = 2r$ unique agents, since this case represents the most likely scenario for λ containing observations in the same leave-out set as within a leave-out set no two senders can be the same, and no two receivers may be the same.

There are $\frac{N!}{(N-2r)!}$ possible choices for the ordered set of agents in λ. Then, there are *at most* $N(N-1)(N-2)(N-4)\cdots(N-2r+2)(N-3r+2)$ orderings of agents in λ which span r different leave-out sets ($N(N-1)$ choices for the first observation (i,j), then $(N-2)(N-4)$ choices for the second observation (s,t) since (s,t) cannot belong in the same leave-out set as (i,j), and so on). This gives

$$P(I_\lambda = r) \geq \frac{N(N-1)(N-2)(N-4)\cdots(N-2r+2)(N-3r+2)}{N(N-1)\cdots(N-2r+1)}$$

which is the product of $2r$ ratios each of which is equal to $1 - O(N^{-1})$, which implies $P(I_\lambda^r = 1) = 1 - O(N^{-1})$.

Next, note that whenever $I_\lambda = s$, we have $1_\lambda^k = 1$ only if $\mathcal{I}_k$ is not one of the s leave-out sets spanned by λ. This happens with probability $\binom{N-2}{s}/\binom{N-1}{s} = \frac{N-s-1}{N-1}$.

$$\begin{aligned} E[1_\lambda^k] &= \sum_{s=1}^{r} \frac{N-s-1}{N-1} P(I_\lambda = s) \\ &= \frac{N-r-1}{N-1} \\ &\quad + \sum_{s=1}^{r-1} \frac{r-s}{N-1} P(I_\lambda = s) \end{aligned}$$

and so

$$\frac{1}{N-1}\big(1 - P(I_\lambda = r)\big) \leq E[1_\lambda^k] - \frac{N-r-1}{N-1} \leq \frac{r-1}{N-1}\big(1 - P(I_\lambda = r)\big)$$

Since we have $P(I_\lambda = r) = 1 - O(N^{-1})$, we can conclude $E[1_\lambda^k] = \frac{N-r-1}{N-1} + O(N^{-2})$. □

Lemma A.11. *Let λ be a set of r observations (i,j) involving p unique agents, and Λ_N be the collection of all such λ formed by permuting the agents in λ. Then, under Assumption 1.2,*

(i) $\partial_\phi \bar{\Delta}_N$ has $O_p(N^{-1})$ elements

(ii) $\partial_{\alpha\alpha'}\bar{\Delta}_N$, $\partial_{\alpha\gamma'}\bar{\Delta}_N$, $\partial_{\gamma\alpha'}\bar{\Delta}_N$, and $\partial_{\gamma\gamma'}\bar{\Delta}_N$ each have $O_p(N^{-1})$ diagonal elements and $O_p(N^{-2})$ off-diagonal elements

Proof. Let λ_α denote the set of p_α sender agents in the observations within λ, and λ_γ the set of p_γ receiving agents. There are $|\Lambda_N| = \frac{N!}{(N-p)!}$ ways of selecting the p agents in λ. Among these permutations, agent i is a sender $p_\alpha \frac{(N-1)!}{(N-p)!}$ times, while node j is the receiver $p_\gamma \frac{(N-1)!}{(N-p)!}$

times. Using this, the first derivatives of $\bar{\Delta}_N$ with respect to the fixed effects are

$$\partial_{\alpha_i}\bar{\Delta}_N = \frac{1}{|\Lambda_N|}\sum_{\lambda: i\in\lambda_\alpha}\partial_{\alpha_i}\bar{m}_\lambda = O_p(N^{-1})$$
$$\partial_{\gamma_i}\bar{\Delta}_N = \frac{1}{|\Lambda_N|}\sum_{\lambda: i\in\lambda_\gamma}\partial_{\gamma_i}\bar{m}_\lambda = O_p(N^{-1})$$

where the $O_p(N^{-1})$ statements come from the fact that $p_\alpha\frac{(N-1)!}{(N-p)!}/\frac{N!}{(N-p)!} = p_\alpha/N$. An identical result applies to the diagonal elements of $\partial_{\phi\phi'}\bar{\Delta}_N$, i.e. $\partial_{\alpha_i\alpha_i}\bar{\Delta}_N = O_p(N^{-1})$ and $\partial_{\gamma_j\gamma_j}\bar{\Delta}_N = O_p(N^{-1})$ since they are sums over the same sets of λ. Also, if the presence of i as a sender agent implies that i is also a receiver in λ (e.g. the cyclic triangle $\{(i,j),(j,k),(k,i)\}$) then it will be the case that $\partial_{\alpha_i\gamma_i}\bar{\Delta}_N = O_p(N^{-1})$ also (if this is not true it will be lower order).

Next, consider the off-diagonal components of $\partial_{\phi\phi'}\bar{\Delta}_N$. If $p_\alpha = 1$ then $\partial_{\alpha_i\alpha_j}\bar{\Delta}_N = 0$, otherwise, there are $\binom{p_\alpha}{2}\frac{(N-2)!}{(N-p)!}$ permutations that contain both i and j as senders. Similarly, for $p_\gamma \geq 2$, there are $\binom{p_\gamma}{2}\frac{(N-2)!}{(N-p)!}$ permutations that contain both i and j as receivers. Finally, there are *at most* $p_\alpha p_\gamma\frac{(N-2)!}{(N-p)!}$ permutations in which i is a sender and j a receiver (this is an upper bound since with i in a particular sender position, not all receiver positions may be valid for j). This, along with Assumption 1.2, gives the results

$$\partial_{\alpha_i\alpha_j}\bar{\Delta}_N = O_p(N^{-2})$$
$$\partial_{\alpha_i\gamma_j}\bar{\Delta}_N = O_p(N^{-2})$$
$$\partial_{\gamma_i\gamma_j}\bar{\Delta}_N = O_p(N^{-2})$$

which demonstrates the lemma. □

Lemma A.12. *Let 1^k_{ij} satisfy Condition A.1 and let $1^k_\lambda = \prod_{(i,j)\in\lambda} 1^k_{ij}$. Let A_{ij} be a mean-zero random variable with bounded fourth moment, and define*

$$\begin{aligned}\mathbf{A} &= \frac{1}{N-1}\big(\{\sum_{s\neq i}A_{is}\}_{i=1,\dots,N}, \{\sum_{s\neq j}A_{sj}\}_{j=1,\dots,N}\big)\\ &= (\mathbf{A}_\alpha, \mathbf{A}_\gamma)\\ \mathbf{A}_k &= \frac{1}{N-2}\big(\{\sum_{s\neq i}A_{is}1^k_{is}\}_{i=1,\dots,N}, \{\sum_{s\neq j}A_{sj}1^k_{sj}\}_{j=1,\dots,N}\big)\\ &= (\mathbf{A}_{\alpha,k}, \mathbf{A}_{\gamma,k})\end{aligned}$$

and let $\mathbf{B}$ *and* $\mathbf{B}_k$ *be defined as*

$$\begin{aligned}\mathbf{B} &= \frac{N}{|\Lambda_N|}\big(\{\sum_{s\neq i}\sum_{\lambda\in\Lambda_{is}} B_\lambda\}_{i=1,\dots,N}, \{\sum_{s\neq j}\sum_{\lambda\in\Lambda_{sj}} B_\lambda\}_{i=1,\dots,N}\big)\\ &= (\mathbf{B}_\alpha, \mathbf{B}_\gamma)\\ \mathbf{B}_k &= \frac{N-1}{N-r-1}\frac{N}{|\Lambda_N|}\big(\{\sum_{s\neq i}\sum_{\lambda\in\Lambda_{is}} B_\lambda 1_\lambda^k\}_{i=1,\dots,N}, \{\sum_{s\neq j}\sum_{\lambda\in\Lambda_{sj}} B_\lambda 1_\lambda^k\}_{i=1,\dots,N}\big)\\ &= (\mathbf{B}_{\alpha,k}, \mathbf{B}_{\gamma,k})\end{aligned}$$

for mean zero B_λ *with bounded fourth moment. Assume that* A_{ij} *is independent of* A_{st} *for* $(i,j) \notin \{(s,t),(t,s)\}$*, and independent of* B_λ *whenever* λ *does not contain either* (i,j) *or* (j,i)*. Define the jackknifed term*

$$\mathcal{J}_0 = (N-1)\mathbf{A}'M\mathbf{B} - \frac{N-2}{N-1}\sum_k \mathbf{A}'_{(k)}M\mathbf{B}_{(k)}$$

where M *is a non-random matrix that has* $O_p(1)$ *elements on its diagonal and* $O_p(N^{-1})$ *off-diagonal terms. Then we have:*

(i) $\bar{E}\big[\mathcal{J}_0\big] = o_p(1)$,

(ii) $\mathcal{J}_0 = o_p(1)$.

Proof. The most common choice of M will be $\bar{\mathcal{H}}^{-1}$, which satisfies the conditions for M by Assumption 1.1 and (6). We show the proof using $\bar{\mathcal{H}}^{-1}$, but note that it holds for any M satisfying the conditions stated above. We have

$$\begin{aligned}\mathbf{A}'\bar{\mathcal{H}}^{-1}\mathbf{B} = \mathbf{A}'_\alpha\bar{\mathcal{H}}^{-1}_{\alpha\alpha}\mathbf{B}_\alpha + \mathbf{A}'_\alpha\bar{\mathcal{H}}^{-1}_{\alpha\gamma}\mathbf{B}_\gamma\\ + \mathbf{A}'_\gamma\bar{\mathcal{H}}^{-1}_{\gamma\alpha}\mathbf{B}_\alpha + \mathbf{A}'_\gamma\bar{\mathcal{H}}^{-1}_{\gamma\gamma}\mathbf{B}_\gamma\end{aligned}$$

Let $\Lambda_{is} = \{\lambda : (i,s) \in \lambda\}$ be the set of λ containing observation (i,s). The full sample and leave-out versions of the first term are

$$\begin{aligned}\mathbf{A}'_\alpha\bar{\mathcal{H}}^{-1}_{\alpha\alpha}\mathbf{B}_\alpha &= \sum_{i,j}\mathbf{A}_{\alpha,i}(\bar{\mathcal{H}}^{-1}_{\alpha\alpha})_{ij}\mathbf{B}_{\alpha,j}\\ &= \frac{N}{(N-1)|\Lambda_N|}\sum_{i,j}\sum_{s\neq i}\sum_{t\neq j}\sum_{\lambda\in\Lambda_{jt}}(\bar{\mathcal{H}}^{-1}_{\alpha\alpha})_{ij}A_{is}B_\lambda\end{aligned}$$

$$
\begin{aligned}
&\frac{1}{N-1}\sum_k \mathbf{A}'_{k,\alpha}\bar{\mathcal{H}}^{-1}_{\alpha\alpha}\mathbf{B}_{k,\alpha} \\
&\quad= \frac{N}{(N-r-1)(N-2)|\Lambda_N|}\sum_k\sum_{i,j}\sum_{s\neq i}\sum_{t\neq j}\sum_{\lambda\in\Lambda_{jt}}(\bar{\mathcal{H}}^{-1}_{\alpha\alpha})_{ij}A_{is}B_\lambda 1^k_{is}1^k_\lambda \\
&\quad= \frac{N}{(N-r-1)(N-2)|\Lambda_N|}\sum_k\sum_{i,j}\sum_{s\neq i}\sum_{t\neq j}\sum_{\lambda\in(\Lambda_{jt}\cap\Lambda_{is})}(\bar{\mathcal{H}}^{-1}_{\alpha\alpha})_{ij}A_{is}B_\lambda 1^k_\lambda \\
&\quad+ \frac{N}{(N-r-1)(N-2)|\Lambda_N|}\sum_k\sum_{i,j}\sum_{s\neq i}\sum_{t\neq j}\sum_{\lambda\in(\Lambda_{jt}\setminus\Lambda_{is})}(\bar{\mathcal{H}}^{-1}_{\alpha\alpha})_{ij}A_{is}B_\lambda 1^k_{is}1^k_\lambda
\end{aligned}
$$

Let I_λ be equal to the number of leave-out sets $\mathcal{I}_k$ spanned by the r observations in λ so that $\sum_k 1^k_\lambda = N - I_\lambda - 1$. As shown in the proof of Lemma A.10, $|\{\lambda : I_\lambda < r\}|/|\Lambda_N| \to 0$, that is, the fraction of sets λ that contain two or more observations in the same $\mathcal{I}_k$ is a vanishingly small. Using this, we have

$$
\begin{aligned}
\mathcal{J}_{\alpha\alpha} =&(N-1)\mathbf{A}'_\alpha\bar{\mathcal{H}}^{-1}_{\alpha\alpha'}\mathbf{B}_\alpha - \frac{N-2}{N-1}\sum_k \mathbf{A}'_{\alpha,k}\bar{\mathcal{H}}^{-1}_{\alpha\alpha'}\mathbf{B}_{\alpha,k} \\
=&\frac{N}{|\Lambda_N|}\sum_{i,j}\sum_{s\neq i}\sum_{t\neq j}\sum_{\lambda\in(\Lambda_{jt}\cap\Lambda_{is})}(\bar{\mathcal{H}}^{-1}_{\alpha\alpha})_{ij}A_{is}B_\lambda\Big(1-\frac{\sum_k 1^k_\lambda}{N-r-1}\Big) \\
&+\frac{N}{|\Lambda_N|}\sum_k\sum_{i,j}\sum_{s\neq i}\sum_{t\neq j}\sum_{\lambda\in(\Lambda_{jt}\setminus\Lambda_{is})}(\bar{\mathcal{H}}^{-1}_{\alpha\alpha})_{ij}A_{is}B_\lambda\Big(1-\frac{\sum_k 1^k_{is}1^k_\lambda}{N-r-1}\Big) \\
=&\frac{N}{|\Lambda_N|}\sum_{i,j}\sum_{s\neq i}\sum_{t\neq j}\sum_{\substack{\lambda\in(\Lambda_{jt}\cap\Lambda_{is})\\ I_\lambda<r}}(\bar{\mathcal{H}}^{-1}_{\alpha\alpha})_{ij}A_{is}B_\lambda\Big(1-\frac{N-I_\lambda-1}{N-r-1}\Big) \\
&+\frac{N}{|\Lambda_N|}\sum_k\sum_{i,j}\sum_{s\neq i}\sum_{t\neq j}\sum_{\lambda\in(\Lambda_{jt}\setminus\Lambda_{is})}(\bar{\mathcal{H}}^{-1}_{\alpha\alpha})_{ij}A_{is}B_\lambda\Big(1-\frac{N-I_\lambda-2}{N-r-1}\Big)
\end{aligned}
$$

Letting $\Gamma_{ijst} = (\bar{\mathcal{H}}^{-1}_{\alpha\alpha'})_{ij} + (\bar{\mathcal{H}}^{-1}_{\alpha\gamma'})_{it} + (\bar{\mathcal{H}}^{-1}_{\gamma\alpha'})_{sj} + (\bar{\mathcal{H}}^{-1}_{\gamma\gamma'})_{st}$, similar computations for the other three elements gives

$$
\begin{aligned}
\mathcal{J}_0 =&(N-1)\mathbf{A}'\bar{\mathcal{H}}^{-1}\mathbf{B} - \frac{N-2}{N-1}\sum_k \mathbf{A}'_k\bar{\mathcal{H}}^{-1}\mathbf{B}_k \\
=&\frac{N}{|\Lambda_N|}\sum_{i,j}\sum_{s\neq i}\sum_{t\neq j}\sum_{\substack{\lambda\in(\Lambda_{jt}\cap\Lambda_{is})\\ I_\lambda<r}}\Gamma_{ijst}A_{is}B_\lambda\Big(1-\frac{N-I_\lambda-1}{N-r-1}\Big) \\
&\frac{N}{|\Lambda_N|}\sum_{i,j}\sum_{s\neq i}\sum_{t\neq j}\sum_{\lambda\in(\Lambda_{jt}\cap\Lambda_{si}\setminus\Lambda_{is})}\Gamma_{ijst}A_{is}B_\lambda\Big(1-\frac{N-I_\lambda-2}{N-r-1}\Big)
\end{aligned}
$$

$$+\frac{N}{|\Lambda_N|}\sum_{i,j}\sum_{s\neq i}\sum_{t\neq j}\sum_{\lambda\in(\Lambda_{jt}\setminus(\Lambda_{is}\cup\Lambda_{si}))}\Gamma_{ijst}A_{is}B_\lambda\Big(1-\frac{N-I_\lambda-2}{N-r-1}\Big)$$
$$=\mathcal{J}_{0,1}+\mathcal{J}_{0,2}+\mathcal{J}_{0,3}$$

Next, recall that $\Gamma_{ijst}=O_p(1)$ whenever $i=j$ or $s=t$, and is $O_p(N^{-1})$ otherwise. Note that $\bar{E}[\mathcal{J}_{0,3}]=0$ since λ does not contain (i,s) or (s,i). Taking expectations we get

$$\begin{aligned}\bar{E}[\mathcal{J}_0]=&\frac{N}{|\Lambda_N|}\sum_{i}\sum_{s\neq i}\sum_{\lambda\in\Lambda_{is}:I_\lambda<r}\Gamma_{iiss}\bar{E}[A_{is}B_\lambda]\Big(\frac{I_\lambda-r}{N-r-1}\Big)\\
&+\frac{N}{|\Lambda_N|}\sum_{i}\sum_{s\neq i}\sum_{t\neq\{i,s\}}\sum_{\lambda\in(\Lambda_{it}\cap\Lambda_{is}):I_\lambda<r}\Gamma_{iist}\bar{E}[A_{is}B_\lambda]\Big(\frac{I_\lambda-r}{N-r-1}\Big)\\
&+\frac{N}{|\Lambda_N|}\sum_{i}\sum_{j\neq i}\sum_{s\neq\{i,j\}}\sum_{\lambda\in(\Lambda_{js}\cap\Lambda_{is}):I_\lambda<r}\Gamma_{ijss}\bar{E}[A_{is}B_\lambda]\Big(\frac{I_\lambda-r}{N-r-1}\Big)\\
&+\frac{N}{|\Lambda_N|}\sum_{i}\sum_{j\neq i}\sum_{s\neq i}\sum_{t\neq\{s,j\}}\sum_{\lambda\in(\Lambda_{jt}\cap\Lambda_{is}):I_\lambda<r}\Gamma_{ijst}\bar{E}[A_{is}B_\lambda]\Big(\frac{I_\lambda-r}{N-r-1}\Big)\\
&+\frac{N}{|\Lambda_N|}\sum_{i}\sum_{s\neq i}\sum_{t\neq\{i,s\}}\sum_{\lambda\in((\Lambda_{it}\cap\Lambda_{si})\setminus\Lambda_{is})}\Gamma_{iist}\bar{E}[A_{is}B_\lambda]\Big(\frac{I_\lambda-r+1}{N-r-1}\Big)\\
&+\frac{N}{|\Lambda_N|}\sum_{i}\sum_{j\neq i}\sum_{s\neq\{i,j\}}\sum_{\lambda\in((\Lambda_{js}\cap\Lambda_{si})\setminus\Lambda_{is})}\Gamma_{ijss}\bar{E}[A_{is}B_\lambda]\Big(\frac{I_\lambda-r+1}{N-r-1}\Big)\\
&+\frac{N}{|\Lambda_N|}\sum_{i}\sum_{j\neq i}\sum_{s\neq i}\sum_{t\neq\{j,s\}}\sum_{\lambda\in((\Lambda_{jt}\cap\Lambda_{si})\setminus\Lambda_{is})}\Gamma_{ijst}\bar{E}[A_{is}B_\lambda]\Big(\frac{I_\lambda-r+1}{N-r-1}\Big)\\
=&o_p(1)\end{aligned}$$

since, $\frac{I_\lambda-r}{N-r-1}=O(N^{-1})$, $\frac{I_\lambda-r+1}{N-r-1}=O(N^{-1})$, $\frac{N}{|\Lambda_N|}|\lambda\in\Lambda_{is}:I_\lambda<r|=O(N^{-2})$, $\frac{N}{|\Lambda_N|}|\lambda\in(\Lambda_{it}\cap\Lambda_{is}):I_\lambda<r|=O(N^{-3})$, and so on applying the results on the size of sets Λ_{ij}, $I_\lambda<r$, and Γ_{ijst}.

Then,

$$\begin{aligned}\bar{E}[\mathcal{J}_{0,3}^2]=&\frac{N^2}{|\Lambda_N|^2}\sum_{i}\sum_{j}\sum_{k}\sum_{l}\sum_{s\neq i}\sum_{t\neq j}\sum_{p\neq k}\sum_{q\neq l}\sum_{\lambda\in(\Lambda_{jt}\setminus(\Lambda_{is}\cup\Lambda_{si}))}\sum_{\lambda\in(\Lambda_{ql}\setminus(\Lambda_{pk}\cup\Lambda_{kp}))}\\
&\Gamma_{ijst}\Gamma_{klpq}\bar{E}[A_{is}B_\lambda A_{pk}B_{\lambda'}]\Big(\frac{I_\lambda-r+1}{N-r-1}\Big)\Big(\frac{I_{\lambda'}-r+1}{N-r-1}\Big)\\
=&\frac{N^2}{|\Lambda_N|^2}\sum_{i}\sum_{j}\sum_{l}\sum_{s\neq i}\sum_{t\neq j}\sum_{q\neq l}\sum_{\lambda\in(\Lambda_{jt}\setminus(\Lambda_{is}\cup\Lambda_{si}))}\sum_{\lambda'\in((\Lambda_{ql}\cap(\Lambda_{jt}\cup\Lambda_{tj}))\setminus(\Lambda_{is}\cup\Lambda_{si}))}\end{aligned}$$

$$\Gamma_{ijst}\Gamma_{ilsq}\bar{E}[A_{is}(A_{is}+A_{si})]\bar{E}[B_\lambda B_{\lambda'}]\Big(\frac{I_\lambda - r + 1}{N-r-1}\Big)\Big(\frac{I_{\lambda'} - r + 1}{N-r-1}\Big)$$
$$+ \frac{N^2}{|\Lambda_N|^2}\sum_i\sum_j\sum_k\sum_l\sum_{s\neq i}\sum_{t\neq j}\sum_{p\neq k}\sum_{q\neq l}\sum_{\lambda\in(\Lambda_{jt}\cap(\Lambda_{pk}\cup\Lambda_{kp})\setminus(\Lambda_{is}\cup\Lambda_{si}))}\sum_{\lambda\in(\Lambda_{ql}(\Lambda_{is}\cup\Lambda_{si})\setminus(\Lambda_{pk}\cup\Lambda_{kp}))}$$
$$\Gamma_{ijst}\Gamma_{klpq}\bar{E}[A_{is}B_{\lambda'}]\bar{E}[A_{pk}B_\lambda]\Big(\frac{I_\lambda - r + 1}{N-r-1}\Big)\Big(\frac{I_{\lambda'} - r + 1}{N-r-1}\Big)$$

Note that $\frac{N}{|\Lambda_N|}|\lambda \in (\Lambda_{jt}\setminus(\Lambda_{is}\cup\Lambda_{si}))| = O(N^{-1})$, while $\frac{N}{|\Lambda_N|}|\lambda' \in ((\Lambda_{ql}\cap(\Lambda_{jt}\cup\Lambda_{tj}))\setminus(\Lambda_{is}\cup\Lambda_{si}))|$ is $O(N^{-1})$ if (q,l) equals (t,j) or (j,t), $O(N^{-2})$ if either $q\in\{t,j\}$ or $l\in\{t,j\}$ and $O(N^{-3})$ otherwise. Also, $\Big(\frac{I_\lambda-r+1}{N-r-1}\Big)\Big(\frac{I_{\lambda'}-r+1}{N-r-1}\Big) = O(N^{-2})$. Combining these facts with, $\Gamma_{ijst} = O_p(1)$ whenever $i=j$ or $s=t$, and $O_p(N^{-1})$ otherwise gives $\bar{E}[\mathcal{J}_{0,3}^2] = o_p(1)$.

An almost identical analysis applies to $\mathcal{J}_{0,1}$ and $\mathcal{J}_{0,2}$, giving the result $\mathcal{J}_0 = o_p(1)$. □

The next lemma states the first-order approximation for the jackknife bias-corrected average effect estimator.

Lemma A.13. *Let Assumptions 1.1 and 1.2 hold and let $\widehat{\Delta}_J$ be the jackknife bias-corrected estimator in (1.12). Then,*

$$N(\widehat{\Delta}_J - \Delta_N) = \Big[(\partial_\beta\bar{\Delta}_N) - N(\partial_{\phi'}\bar{\Delta}_N)\bar{\mathcal{H}}^{-1}(\partial_{\beta\phi}\bar{\mathcal{L}})\Big]\bar{W}_N^{-1}U^{(0)}$$
$$+ N(\partial_{\phi'}\bar{\Delta}_N)\bar{\mathcal{H}}^{-1}\mathcal{S} + o_p(1)$$

Proof. We can write an expansion for the leave-out estimate

$$N(\widehat{\Delta}_{(k)} - \Delta_N) = \Big[(\partial_\beta\bar{\Delta}_N) - (\partial_{\phi'}\bar{\Delta}_N)\bar{\mathcal{H}}^{-1}(\partial_{\beta\phi}\bar{\mathcal{L}})\Big]\bar{W}_N^{-1}\big(U_{(k)}^{(0)} + U_{(k)}^{(1)}\big)$$
$$+ N(\partial_{\phi'}\bar{\Delta}_N)\bar{\mathcal{H}}^{-1}\mathcal{S}_{(k)}$$
$$+ N(\partial_{\phi'}\tilde{\Delta}_{(k)})\bar{\mathcal{H}}^{-1}\mathcal{S}_{(k)} - N(\partial_{\phi'}\bar{\Delta}_N)\bar{\mathcal{H}}^{-1}\tilde{\mathcal{H}}_{(k)}\bar{\mathcal{H}}^{-1}\mathcal{S}_{(k)}$$
$$+ \frac{1}{2}N\mathcal{S}'_{(k)}\bar{\mathcal{H}}^{-1}\Big((\partial_{\phi\phi'}\bar{\Delta}_N) + \sum_g(\partial_{\phi\phi'\phi_g}\bar{\mathcal{L}})[(\partial_{\phi'}\bar{\Delta}_N)\bar{\mathcal{H}}^{-1}]_g\Big)\bar{\mathcal{H}}^{-1}\mathcal{S}_{(k)}$$
$$+ R_{(k),\Delta} + \tilde{R}_{(k),\Delta}$$

where $R_{(k),\Delta} = o_p(1)$ is the version of R_Δ in the leave-out sample and $\tilde{R}_{(k),\Delta} = o_p(N^{-1})$ is the leave-out version of $\tilde{R}_\Delta$ combined with the error from replacing terms like $\partial_\beta\bar{\Delta}_{(k)}$ with $\partial_\beta\bar{\Delta}_N$ (i.e. applying the result in Lemma A.10). Using the expansion for the leave-out estimate, we can apply the jackknife operator to each line above.

For the first term, we apply the result in Lemma A.5 to give

$$\begin{aligned}&\mathbf{J}\Big[\big((\partial_\beta\bar{\Delta}_N) - N(\partial_{\phi'}\bar{\Delta}_N)\bar{\mathcal{H}}^{-1}(\partial_{\beta\phi}\bar{\mathcal{L}})\big)\bar{W}_N^{-1}\big(U^{(0)}+U^{(1)}\big)\Big]\\&\quad=\big((\partial_\beta\bar{\Delta}_N) - N(\partial_{\phi'}\bar{\Delta}_N)\bar{\mathcal{H}}^{-1}(\partial_{\beta\phi}\bar{\mathcal{L}})\big)\bar{W}_N^{-1}U^{(0)}+o_p(1)\end{aligned}$$

Similarly, Lemma A.3 implies that jackknifing the second term gives $N(\partial_{\phi'}\bar{\Delta}_N)\bar{\mathcal{H}}^{-1}\mathcal{S}$.

For the third term, we note that by Lemma A.11 we can apply Lemma A.4 with $M=\bar{\mathcal{H}}^{-1}$, $A=\mathcal{S}$, and $B=N(\partial_{\phi'}\bar{\Delta}_N)\bar{\mathcal{H}}^{-1}\tilde{\mathcal{H}}$, and apply Lemma A.12 with $M=\bar{\mathcal{H}}^{-1}$, $A=\mathcal{S}$, and either $B=N(\partial_{\phi'}\tilde{\Delta}_N)$.

For the fourth term, we first show that

$$N\bar{\mathcal{H}}^{-1}\Big((\partial_{\phi\phi'}\bar{\Delta}_N)+\sum_g(\partial_{\phi\phi'\phi_g}\bar{\mathcal{L}})[(\partial_{\phi'}\bar{\Delta}_N)\bar{\mathcal{H}}^{-1}]_g\Big)\bar{\mathcal{H}}^{-1}$$

satisfies the conditions for M in Lemma A.4, from which we will be able to conclude that the jackknifed term will be $o_p(1)$. The above expression is non-random (conditional on exogenous regressors and fixed effects) and so we must demonstrate that it is a $2N\times 2N$ matrix with $O_p(1)$ diagonal elements, and $O_p(N^{-1})$ off-diagonal elements. Note that if two $2N\times 2N$ matrices both have $O_p(1)$ diagonal elements and $O_p(N^{-1})$ off-diagonal elements, then their product also shares this property. Since this is true of $\bar{\mathcal{H}}^{-1}$(see Lemma D.1 in Fernández-Val and Weidner (2016)), it remains to demonstrate this fact for the terms $N(\partial_{\phi\phi'}\bar{\Delta}_N)$ and $N\sum_g(\partial_{\phi\phi'\phi_g}\bar{\mathcal{L}})[(\partial_{\phi'}\bar{\Delta}_N)\bar{\mathcal{H}}^{-1}]_g$.

By Lemma A.11 we have that $N(\partial_{\phi\phi'}\bar{\Delta}_N)$ has $O_p(1)$ diagonal elements and $O_p(N^{-1})$ off-diagonal, with the possible exception of the elements $\partial_{\alpha_i\gamma_i'}\bar{\Delta}_N$. However, this still implies that $N\bar{\mathcal{H}}^{-1}(\partial_{\phi\phi'}\bar{\Delta}_N)\bar{\mathcal{H}}^{-1}$ satisfies the condition. For $N\sum_g(\partial_{\phi\phi'\phi_g}\bar{\mathcal{L}})[(\partial_{\phi'}\bar{\Delta}_N)\bar{\mathcal{H}}^{-1}]_g$, diagonal elements are given by (for $i\leq N$)

$$\begin{aligned}&\Big[N\sum_g(\partial_{\phi\phi'\phi_g}\bar{\mathcal{L}})[(\partial_{\phi'}\bar{\Delta}_N)\bar{\mathcal{H}}^{-1}]_g\Big]_{ii}\\&=N(\partial_{\alpha_i\alpha_i'\phi}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}(\partial_\phi\bar{\Delta}_N)\\&=N\sum_{s,t}(\bar{\mathcal{H}}_{\alpha\alpha}^{-1})_{st}(\partial_{\alpha_i\alpha_i\alpha_s}\bar{\mathcal{L}})(\partial_{\alpha_t}\bar{\Delta}_N)+N\sum_{s,t}(\bar{\mathcal{H}}_{\alpha\gamma}^{-1})_{st}(\partial_{\alpha_i\alpha_i\alpha_s}\bar{\mathcal{L}})(\partial_{\gamma_t}\bar{\Delta}_N)\\&+N\sum_{s,t}(\bar{\mathcal{H}}_{\gamma\alpha}^{-1})_{st}(\partial_{\alpha_i\alpha_i\gamma_s}\bar{\mathcal{L}})(\partial_{\alpha_t}\bar{\Delta}_N)+N\sum_{s,t}(\bar{\mathcal{H}}_{\gamma\gamma}^{-1})_{st}(\partial_{\alpha_i\alpha_i\gamma_s}\bar{\mathcal{L}})(\partial_{\gamma_t}\bar{\Delta}_N)\end{aligned}$$

$$= \frac{N}{N-1}\sum_t\sum_{j\neq i}(\bar{\mathcal{H}}^{-1}_{\alpha\alpha})_{it}(\partial_{\pi^3}\bar{\ell}_{ij})(\partial_{\alpha_t}\bar{\Delta}_N) + \frac{N}{N-1}\sum_t\sum_{j\neq i}(\bar{\mathcal{H}}^{-1}_{\alpha\gamma})_{st}(\partial_{\pi^3}\bar{\ell}_{ij})(\partial_{\gamma_t}\bar{\Delta}_N)$$
$$+ \frac{N}{N-1}\sum_t\sum_{s\neq i}(\bar{\mathcal{H}}^{-1}_{\gamma\alpha})_{st}(\partial_{\pi^3}\bar{\ell}_{is})(\partial_{\alpha_t}\bar{\Delta}_N) + \frac{N}{N-1}\sum_t\sum_{s\neq i}(\bar{\mathcal{H}}^{-1}_{\gamma\gamma})_{st}(\partial_{\pi^3}\bar{\ell}_{is})(\partial_{\gamma_t}\bar{\Delta}_N)$$

which is $O_p(1)$ since by Lemma A.11 and Lemma D.1 in Fernández-Val and Weidner (2016), and similarly for $i > N$. Off-diagonal components can be shown similarly, e.g. for $i < N$ and $j > N$ we have

$$\Big[N\sum_g(\partial_{\phi\phi'\phi_g}\bar{\mathcal{L}})[(\partial_{\phi'}\bar{\Delta}_N)\bar{\mathcal{H}}^{-1}]_g\Big]_{ij}$$
$$= N(\partial_{\alpha_i\gamma'_j\phi}\bar{\mathcal{L}})\bar{\mathcal{H}}^{-1}(\partial_\phi\bar{\Delta}_N)$$
$$= N\sum_{s,t}(\bar{\mathcal{H}}^{-1}_{\alpha\alpha})_{st}(\partial_{\alpha_i\gamma_j\alpha_s}\bar{\mathcal{L}})(\partial_{\alpha_t}\bar{\Delta}_N) + N\sum_{s,t}(\bar{\mathcal{H}}^{-1}_{\alpha\gamma})_{st}(\partial_{\alpha_i\gamma_j\alpha_s}\bar{\mathcal{L}})(\partial_{\gamma_t}\bar{\Delta}_N)$$
$$+ N\sum_{s,t}(\bar{\mathcal{H}}^{-1}_{\gamma\alpha})_{st}(\partial_{\alpha_i\gamma_j\gamma_s}\bar{\mathcal{L}})(\partial_{\alpha_t}\bar{\Delta}_N) + N\sum_{s,t}(\bar{\mathcal{H}}^{-1}_{\gamma\gamma})_{st}(\partial_{\alpha_i\gamma_j\gamma_s}\bar{\mathcal{L}})(\partial_{\gamma_t}\bar{\Delta}_N)$$
$$= \frac{N}{N-1}\sum_t(\bar{\mathcal{H}}^{-1}_{\alpha\alpha})_{it}(\partial_{\pi^3}\bar{\ell}_{ij})(\partial_{\alpha_t}\bar{\Delta}_N) + \frac{N}{N-1}\sum_t(\bar{\mathcal{H}}^{-1}_{\alpha\gamma})_{st}(\partial_{\pi^3}\bar{\ell}_{ij})(\partial_{\gamma_t}\bar{\Delta}_N)$$
$$+ \frac{N}{N-1}\sum_t(\bar{\mathcal{H}}^{-1}_{\gamma\alpha})_{jt}(\partial_{\pi^3}\bar{\ell}_{ij})(\partial_{\alpha_t}\bar{\Delta}_N) + \frac{N}{N-1}\sum_t(\bar{\mathcal{H}}^{-1}_{\gamma\gamma})_{jt}(\partial_{\pi^3}\bar{\ell}_{ij})(\partial_{\gamma_t}\bar{\Delta}_N)$$

which is $O_p(N^{-1})$. Finally, in the Supplementary Appendix (S.4) it is shown that $\mathbf{J}[R_\Delta] = o_p(1)$, and by $\tilde{R}_{\Delta,(k)} = o_p(N^{-1})$ for each k (and in the full sample) we have that $(N-2)\tilde{R}_{\Delta,(k)} = o_p(1)$ and $(N-1)\tilde{R}_\Delta = o_p(1)$ so that the jackknifed version of this term is also $o_p(1)$. □

Proof of Theorem 1.2

We can decompose $\widehat{\Delta}_J - \bar{\Delta}_N$ into

$$\widehat{\Delta}_J - \bar{\Delta}_N = (\widehat{\Delta}_J - \Delta_N) + (\Delta_N - \bar{\Delta}_N)$$

From Lemma A.13 we have

$$N(\widehat{\Delta}_J - \Delta_N) = \Big[(\partial_\beta\bar{\Delta}_N) - (\partial_{\phi'}\bar{\Delta}_N)\bar{\mathcal{H}}^{-1}(\partial_{\beta\phi}\bar{\mathcal{L}})\Big]\bar{W}_N^{-1}U^{(0)}$$
$$+ N(\partial_{\phi'}\bar{\Delta}_N)\bar{\mathcal{H}}^{-1}\mathcal{S} + o_p(1)$$

Some tedious matrix algebra shows that this expression is equivalent to

$$\begin{aligned} N(\widehat{\Delta}_J - \Delta_N) &= -N(\partial_\theta \bar{\Delta}_N)(\partial_{\theta\theta'}\bar{\mathcal{L}})^{-1}(\partial_\theta \mathcal{L}) \\ &= \big(- N(\partial_\theta \bar{\Delta}_N)(\partial_{\theta\theta'}\bar{\mathcal{L}})^{-1}\big) \frac{1}{N-1} \sum_i \sum_{j \neq i} \partial_\theta \ell_{ij} \\ &= \frac{1}{N-1} \sum_i \sum_{j \neq i} \tilde{h}_{ij} \end{aligned}$$

where $\theta' = (\beta, \phi')$, and $\tilde{h}_{ij} = -N(\partial_\theta \bar{\Delta}_N)(\partial_{\theta\theta'}\bar{\mathcal{L}})^{-1}\partial_\theta \ell_{ij}$. Next, let $\tilde{s}_{ij} = \frac{1}{|\Lambda_{ij}|}\sum_{\lambda \in \Lambda_{ij}}(m_\lambda - \bar{m}_\lambda)$, where $\Lambda_{ij} = \{\lambda : (i,j) \in \lambda\}$. Then

$$\begin{aligned} N(\Delta_N - \bar{\Delta}_N) &= \frac{N}{|\Lambda_N|} \sum_\lambda (m_\lambda - \bar{m}_\lambda) \\ &= \frac{1}{r} \frac{N}{|\Lambda_N|} \sum_i \sum_{j \neq i} \sum_{\lambda \in \Lambda_{ij}} (m_\lambda - \bar{m}_\lambda) \\ &= \frac{1}{N-1} \sum_i \sum_{j \neq i} \tilde{s}_{ij} \end{aligned}$$

since $|\Lambda_{ij}| = r\frac{(N-2)!}{(N-p)!}$. We then have

$$N(\widehat{\Delta}_J - \bar{\Delta}_N) = \frac{1}{N-1} \sum_i \sum_{j \neq i} (\tilde{h}_{ij} + \tilde{s}_{ij}) + o_p(1)$$

which is asymptotically normal by a standard CLT. We now determine the variance of this term.

$$Var\big(\frac{1}{N-1} \sum_i \sum_{j \neq i} (\tilde{h}_{ij} + \tilde{s}_{ij})\big) = \frac{1}{(N-1)^2} \sum_i \sum_{j \neq i} \sum_s \sum_{t \neq s} \bar{E}[(\tilde{h}_{ij} + \tilde{s}_{ij})(\tilde{h}_{st} + \tilde{s}_{st})]$$

To compute this, first note that $\sum_s \sum_{t \neq s} \bar{E}[\tilde{h}_{ij}\tilde{h}_{st}] = \bar{E}[\tilde{h}_{ij}(\tilde{h}_{ij} + \tilde{h}_{ji})]$. Also,

$$\begin{aligned} \sum_s \sum_{t \neq s} \bar{E}[\tilde{h}_{ij}\tilde{s}_{st}] &= \frac{1}{r} \frac{(N-p)!}{(N-2)!} \sum_s \sum_{t \neq s} \sum_{\lambda \in \Lambda_{st}} \bar{E}\big[\tilde{h}_{ij}(m_\lambda - \bar{m}_\lambda)\big] \\ &= \frac{(N-p)!}{(N-2)!} \sum_\lambda \bar{E}\big[\tilde{h}_{ij}(m_\lambda - \bar{m}_\lambda)\big] \end{aligned}$$

$$= \frac{(N-p)!}{(N-2)!} \sum_{\lambda \in (\Lambda_{ij} \cup \Lambda_{ji})} \bar{E}\big[\tilde{h}_{ij}(m_\lambda - \bar{m}_\lambda)\big]$$
$$= \bar{E}[\tilde{h}_{ij} s_{ij}]$$

where $s_{ij} = \frac{(N-p)!}{(N-2)!} \sum_{\lambda \in (\Lambda_{ij} \cup \Lambda_{ji})} (m_\lambda - \bar{m}_\lambda)$.

Let $D(\lambda)$ be the set of dyads formed from the observations in λ, i.e. if $(i,j) \in \lambda$ then $(i,j),(j,i) \in D(\lambda)$. Assume that λ and λ' both contain the observations (i,j),(i,k). There are $O(N^{p-3})$ sets λ containing the corresponding dyads, so that there are $O(N^{2p-3})$ such λ, λ' pairs ($O(N^3)$ choices of (i,j,k) and $O(N^{p-3})$ choices for each of λ and λ'). Similarly, there are $O(N^{2p-4})$ λ, λ' pairs that share two dyads made up for four agents, (i,j),(k,l). Using this,

$$\frac{1}{(N-1)^2}\Big(\frac{1}{r}\frac{(N-p)!}{(N-2)!}\Big)^2 \sum_i \sum_{j \neq i} \sum_{\lambda \in \Lambda_{ij}} \sum_s \sum_{t \neq s} \sum_{\lambda' \in \Lambda_{st}} \bar{E}\big[(m_\lambda - \bar{m}_\lambda)(m_{\lambda'} - \bar{m}_{\lambda'})\big]$$
$$= \frac{N^2}{|\Lambda_N|^2} \sum_\lambda \sum_{\lambda'} \bar{E}\big[(m_\lambda - \bar{m}_\lambda)(m_{\lambda'} - \bar{m}_{\lambda'})\big]$$
$$= \frac{N^2}{|\Lambda_N|^2} \sum_\lambda \sum_{\lambda' : |D(\lambda) \cap D(\lambda')| = 2} \bar{E}\big[(m_\lambda - \bar{m}_\lambda)(m_{\lambda'} - \bar{m}_{\lambda'})\big]$$
$$+ \frac{N^2}{|\Lambda_N|^2} \sum_\lambda \sum_{\lambda' : |D(\lambda) \cap D(\lambda')| > 2} \bar{E}\big[(m_\lambda - \bar{m}_\lambda)(m_{\lambda'} - \bar{m}_{\lambda'})\big]$$
$$= \frac{N^2}{|\Lambda_N|^2} \sum_i \sum_{j<i} \sum_{\lambda \in (\Lambda_{ij} \cup \Lambda_{ji})} \sum_{\lambda' : |D(\lambda) \cap D(\lambda')| = \{(i,j),(j,i)\}} \bar{E}\big[(m_\lambda - \bar{m}_\lambda)(m_{\lambda'} - \bar{m}_{\lambda'})\big]$$
$$+ \frac{N^2}{|\Lambda_N|^2} \sum_\lambda \sum_{\lambda' : |D(\lambda) \cap D(\lambda')| > 2} \bar{E}\big[(m_\lambda - \bar{m}_\lambda)(m_{\lambda'} - \bar{m}_{\lambda'})\big]$$
$$= \frac{N^2}{|\Lambda_N|^2} \sum_i \sum_{j<i} \sum_{\lambda \in (\Lambda_{ij} \cup \Lambda_{ji})} \sum_{\lambda' \in (\Lambda_{ij} \cup \Lambda_{ji})} \bar{E}\big[(m_\lambda - \bar{m}_\lambda)(m_{\lambda'} - \bar{m}_{\lambda'})\big]$$
$$- \frac{N^2}{|\Lambda_N|^2} \sum_i \sum_{j<i} \sum_{\lambda \in (\Lambda_{ij} \cup \Lambda_{ji})} \sum_{\substack{\lambda' \in (\Lambda_{ij} \cup \Lambda_{ji}): \\ |D(\lambda) \cap D(\lambda')| > 2}} \bar{E}\big[(m_\lambda - \bar{m}_\lambda)(m_{\lambda'} - \bar{m}_{\lambda'})\big]$$
$$+ O_p(N^{-1})$$
$$= \frac{1}{(N-1)^2} \sum_i \sum_{j<i} \bar{E}[s_{ij}^2] + o_p(1)$$

This implies that

$$\begin{aligned} Var\big(\frac{1}{N-1}\sum_i\sum_{j\neq i}(h_{ij}+\tilde{s}_{ij})\big) &= \frac{1}{(N-1)^2}\sum_i\sum_{j<i}\Big(\bar{E}[(\tilde{h}_{ij}+\tilde{h}_{ji})^2] \\ &\quad + 2\bar{E}[(\tilde{h}_{ij}+\tilde{h}_{ji})s_{ij}] + \bar{E}[s_{ij}^2]\Big) + o_p(1) \\ &= \frac{1}{(N-1)^2}\sum_i\sum_{j<i}\bar{E}\big[(h_{ij}+s_{ij})^2\big] + o_p(1) \end{aligned}$$

for $h_{ij} = \tilde{h}_{ij} + \tilde{h}_{ji}$. The asymptotic variance of $N(\widehat{\Delta}_J - \bar{\Delta}_N)$ is given by the limit of this expression. Assumptions 1.1 (iii) and 1.2 (ii) guarantee that both h_{ij} and s_{ij} have first-order approximations that are continuously differentiable in the parameters, so that the continuous mapping theorem and the consistency results $\|\widehat{\beta} - \beta_0\| \to 0$ and $\|\widehat{\phi} - \phi\|_\infty \to 0$ in (7), imply consistency of the plug-in estimator for V_Δ (see for example Lemma S.1 in FVW).

Proof of Theorem 1.3

We begin with a U-statistic representation of $\bar{\Delta}_N$, which will allow us to apply standard asymptotic results on U-statistics. We have defined m to be a function of the sets λ, which depend on an *ordered* set of p agents. For example the transitive triangle $\lambda = \{(i,j),(i,k),(k,j)\}$ depends on the agents $\{i,j,k\}$ in a non-symmetric manner. We first rewrite $\bar{\Delta}_N$ to be a sum over functions that are symmetric in agents. Denote the set of agents in λ by $N(\lambda)$, and let $\eta = \{i_1,\ldots,i_p\}$ be some set of p agents. Then we may define $\tau_\eta = \{\lambda : N(\lambda) = \eta\}$ as the collection of all λ that contain the same set of agents. Note that $|\tau_\eta| = p!$. We have, for $\tilde{m} = \bar{m} - E[m]$

$$\begin{aligned} \bar{\Delta}_N - \delta &= \frac{1}{N\cdots(N-p+1)}\sum_\lambda \tilde{m}_\lambda \\ &= \frac{p!}{N\cdots(N-p+1)}\sum_\tau\Big(\frac{1}{p!}\sum_{\lambda\in\tau}\tilde{m}_\lambda\Big) \\ &= \binom{N}{p}^{-1}\sum_\tau u_\tau \end{aligned}$$

where $u_\tau = \frac{1}{p!}\sum_{\lambda\in\tau}\tilde{m}_\lambda$. The variable u_τ is symmetric function of $\{\beta, X_i, \alpha_i, \gamma_i\}$ for p agents i. For example, there are $3! = 6$ possible transitive triangles using agents $\{i,j,k\}$ so that u is the average of the function m evaluated at these 6 different triangles. Assuming that the

$\{X_i, \alpha_i, \gamma_i\}$ are i.i.d. over agents, $\bar{\Delta}_N - \delta$ is a U-statistic of order p and we apply standard theory on such statistics to compute its asymptotic distribution. As in Theorem 12.3 in van der Vaart (1998) we have

$$\sqrt{N}(\bar{\Delta}_N - \delta) \Rightarrow N(0, p^2\zeta_1)$$

where, for τ and τ' sharing exactly one agent in common,

$$\zeta_1 = Cov(u_\tau, u_{\tau'})$$

An estimator of ζ_1 is

$$\begin{aligned}\sqrt{N}\binom{N}{p}^{-1}\sum_{\tau} u_\tau &= \frac{1}{\sqrt{N}}\sum_i \binom{N-1}{p-1}^{-1}\sum_{\tau: i\in\tau} u_\tau \\ &= \frac{1}{\sqrt{N}}\sum_i t_i\end{aligned}$$

The variance of t_i is given by

$$\begin{aligned}Var(t_i) &= \binom{N-1}{p-1}^{-2}\sum_{\tau:i\in\tau}\sum_{\tau':i\in\tau'} E[u_\tau u_{\tau'}] \\ &= \binom{N-1}{p-1}^{-2}\sum_{\tau:i\in\tau}\sum_{\tau':\tau\cap\tau'=\{i\}} E[u_\tau u_{\tau'}] \\ &+ \binom{N-1}{p-1}^{-2}\sum_{\tau:i\in\tau}\sum_{\substack{\tau':i\in\tau' \\ |\tau\cap\tau'|>1}} E[u_\tau u_{\tau'}] \\ &= \binom{N-1}{p-1}^{-2}\sum_{\tau:i\in\tau}\sum_{\tau':\tau\cap\tau'=\{i\}} E[u_\tau u_{\tau'}] + o(1) \\ &= \frac{(N-p)!(p-1)!}{(N-1)!}\frac{(N-p)!}{(N-2p+1)!(p-1)!}\zeta_1 + o(1) \\ &= \zeta_1 + o(1)\end{aligned}$$

To explain the final line, there are $\binom{N-1}{p-1}$ ways to choose τ containing i, and $\binom{N-p}{p-1}$ ways to choose the remaining $p-1$ agents in τ' so that τ and τ' share only agent i in common. The first term is therefore $O(1)$. Now assume τ and τ' share two agents in common (one of which is i). There are $N-1$ choices for the second common agent, $\binom{N-2}{p-2}$ ways to choose τ

containing i and the second common agent, and $\binom{N-p}{p-2}$ ways to choose the remaining agents in τ'. This implies that the sum for τ and τ' with two agents in common is $O(N^{-1})$. Similarly, the sums for three agents in common are $O(N^{-2})$ and so on.

This implies that the variance of t_i converges to ζ_1. We can alternatively express this variance as

$$\begin{aligned}
Var(t_i) &= \binom{N-1}{p-1}^{-2} \sum_{\tau: i\in\tau} \sum_{\tau': i\in\tau'} E[u_\tau u_{\tau'}] \\
&= \binom{N-1}{p-1}^{-2} \sum_{\tau: i\in\tau} \sum_{\tau': i\in\tau'} \frac{1}{p!} \sum_{\lambda\in\tau} \frac{1}{p!} \sum_{\lambda'\in\tau} E[\tilde{m}_\lambda \tilde{m}_{\lambda'}] \\
&= \frac{1}{p!^2} \binom{N-1}{p-1}^{-2} \sum_{\lambda: i\in\lambda} \sum_{\lambda': i\in\lambda'} E[\tilde{m}_\lambda \tilde{m}_{\lambda'}] \\
&= \frac{1}{p^2} E\Big[\Big(\frac{(N-p)!}{(N-1)!} \sum_{\lambda: i\in\lambda} \tilde{m}_\lambda\Big)^2\Big]
\end{aligned}$$

An estimator of $p^2\zeta_1^2$ is therefore given by

$$\begin{aligned}
\widehat{V}_\delta &= \frac{1}{N} \sum_i \tilde{\mu}_i^2 \\
\tilde{\mu}_i &= \frac{(N-p)!}{(N-1)!} \sum_{\lambda: i\in\lambda} (\widehat{\bar{m}}_\lambda - \widehat{\mu}) \\
\widehat{\mu} &= \frac{(N-p)!}{N!} \sum_\lambda \widehat{\bar{m}}_\lambda
\end{aligned}$$

and $\widehat{\bar{m}}_\lambda$ is a plug-in estimator for $\bar{m}_\lambda$. Assumption 1.2 and consistency of parameters $\|\widehat{\phi} - \phi_0\|_\infty \to 0$, $\|\widehat{\beta} - \beta_0\| \to 0$ ensures consistency of $\widehat{m}_\lambda$ and hence consistency of $\widehat{V}_\delta$.

B Appendix for Chapter 2

B.1 Expansions and V-statistic forms

The method of deriving the expansions in this paper follows that used in Hahn and Newey (2004). For notational simplicity the proofs are done for scalar θ, with the vector case expected to follow similarly. To describe the expansions, first note that the MLE solves the moment conditions

$$0 = \sum_{i=1}^{n}\sum_{t=1}^{T} U(z_{it};\hat{\theta},\hat{\alpha}_i(\hat{\theta}))$$
$$0 = \sum_{t=1}^{T} V(z_{it};\hat{\theta},\hat{\alpha}_i(\hat{\theta})) \quad \text{for all } i$$

for $U(z_{it};\theta,\alpha) = u(z_{it};\theta,\alpha) - \delta V(z_{it};\theta,\alpha)$, where $\delta = E[u_{it}V_{it}]/E[V_{it}^2]$. Let $F \equiv (F_1,\ldots,F_n)$ be the collection of distribution functions and $\hat{F}$ the corresponding collection of empirical distributions. Next, define $F(\epsilon) = F + \epsilon\sqrt{T}(\hat{F} - F)$ for $\epsilon \in [0, T^{-1/2}]$, and let $\alpha_i(\theta, F_i(\epsilon))$ and $\theta(F(\epsilon))$ be solutions to the moment equations

$$0 = \int V\big(z_{it};\theta,\alpha_i(\theta,F_i(\epsilon))\big)dF_i(\epsilon), \quad \text{for all } i$$
$$0 = \sum_{i=1}^{n}\int U\Big(z_{it};\theta(F(\epsilon)),\alpha_i\big(\theta(F(\epsilon)),F_i(\epsilon)\big)\Big)dF_i(\epsilon)$$

The expansion is generated via Taylor series expansion of $\theta(F(\epsilon))$

$$\hat{\theta} - \theta_0 = \frac{1}{\sqrt{T}}\theta^{\epsilon}(0) + \frac{1}{2}\frac{1}{T}\theta^{\epsilon\epsilon}(0) + \frac{1}{6}\frac{1}{T^{3/2}}\theta^{\epsilon\epsilon\epsilon}(0) + \frac{1}{24}\frac{1}{T^2}\theta^{\epsilon\epsilon\epsilon\epsilon}(0) + R_{n,T}$$

where $\theta^{\epsilon}(0) = d\theta(F(\epsilon))/d\epsilon$, $\theta^{\epsilon\epsilon}(0) = d^2\theta(F(\epsilon))/d\epsilon^2$ etc. evaluated at $\epsilon = 0$. In the main paper we use the notation $A_n = \theta^{\epsilon}(0)$, $B_n = \frac{1}{2}\theta^{\epsilon\epsilon}(0)$ and so on to avoid introducing complicated notation. Lemmas 1-5 of the Supplementary Appendix show that $E[A_n] = O(1)$, $E[B_n] = O(1)$, and so on, while $\sqrt{n}(A_n - E[A_n]) = O_p(1)$, $\sqrt{n}(B_n - E[B_n]) = O_p(1)$ etc.[2] Further, in Section D of the Supplementary Appendix we show that $R_{n,T} = o_p(T^{-1})$. These results

[2]The results are established up to sixth-order terms

allow us to write the expansion

$$\sqrt{nT}(\hat{\theta}-\theta_0)=\sqrt{n}A_n+\frac{\sqrt{n}}{\sqrt{T}}B_n+\frac{\sqrt{n}}{T}C_n+\frac{\sqrt{n}}{T\sqrt{T}}D_n+o_p(T^{-1}) \tag{8}$$

Normalized V-statistics

The proofs of most of the results in the paper make use of a particular structure for the terms in the above expansion. Consider a statistic of the form

$$\begin{aligned}W_{i,T,m}&\equiv\frac{1}{T^{m/2}}\sum_{t_1=1}^{T}\sum_{t_2=1}^{T}\cdots\sum_{t_m=1}^{T}k_1\left(x_{i,t_1}\right)k_2\left(x_{i,t_2}\right)\cdots k_m\left(x_{i,t_m}\right)\\&\equiv T^{m/2}\overline{k}_{i,1}\overline{k}_{i,2}\cdots\overline{k}_{i,m}\end{aligned}$$

where $E\left[k_j\left(x_{i,t}\right)\right]=0$. We will call the average

$$W_{T,(m)}=\frac{1}{n}\sum_{i=1}^{n}W_{i,T,m}$$

a normalized V-statistic of order m. The elements of the expansion can all be shown to be products of such V-statistics

$$W_{T,(m_1,\ldots,m_L)}=W_{T,m_1}\cdots W_{T,m_L}$$

Inspection of the expansion terms (see appendix to Hahn and Newey (2004)) shows the first-order term A_n is a V-statistic with $L=1$ and $m_1=1$, while the second-order term B_n contains V-statistics with $L\leq 2$ and $\sum_l m_l=2$, the third-order term C_n contains V-statistics with $L\leq 3$ and $\sum_l m_l=3$, and so on. Below we prove results for these general statistics, which subsequently imply the results for the expansion terms.

Jackknifing V-statistics

Here we show the impact of the jackknife and split-sample bias corrections on V-statistics up to third-order. We focus on $W_{i,T,m}$ rather than averages over i, since the bias corrections only act on the time series dimension. Recall that an m-th order expansion term is an average of

terms $T^{-m/2}W_{i,T,m}$. We can write the leave-one-out statistics as

$$\begin{aligned} W_{i,T,m}^{(-t)} &= \frac{1}{(T-1)^{m/2}} \sum_{t_1 \neq t}^{T} \sum_{t_2 \neq t}^{T} \cdots \sum_{t_m \neq t}^{T} k_1\left(x_{i,t_1}\right) k_2\left(x_{i,t_2}\right) \cdots k_m\left(x_{i,t_m}\right) \\ &= (T-1)^{m/2} \frac{T\bar{k}_{i,1} - k_1(x_{i,t})}{T-1} \cdots \frac{T\bar{k}_{i,m} - k_m(x_{i,t})}{T-1} \\ &= (T-1)^{-m/2} \Big(T\bar{k}_{i,1} - k_1(x_{i,t}) \Big) \cdots \Big(T\bar{k}_{i,m} - k_m(x_{i,t}) \Big) \end{aligned}$$

The corresponding jackknifed statistic is then

$$T^{-\frac{m}{2}} \tilde{W}_{i,T,m} = T \cdot T^{-\frac{m}{2}} W_{i,T,m} - (T-1) \cdot (T-1)^{-\frac{m}{2}} \frac{1}{T} \sum_{t=1}^{T} W_{i,T,m}^{(-t)}$$

The following Lemma outlines the impact of the jackknife for $m = 1, 2, 3$.

Lemma B.1. *Let* $\tilde{W}_{i,T,m} = TW_{i,T,m} - (T-1)(\frac{T}{T-1})^{m/2} \frac{1}{T} \sum_t W_{i,T,m}^{(-t)}$. *Then we have:*

$$\begin{aligned} \tilde{W}_{i,T,1} &= W_{i,T,1} \\ \tilde{W}_{i,T,2} &= \frac{1}{T-1} \sum_{t_1 \neq t_2} k_1(x_{i,t_1}) k_2(x_{i,t_2}) \\ \tilde{W}_{i,T,3} &= \frac{-1}{T^{1/2}(T-1)} \frac{1}{T} \sum_{t} k_1(x_{i,t}) k_2(x_{i,t}) k_3(x_{i,t}) \\ &\quad + \frac{1}{T^{1/2}(T-1)^2} \sum_{t_1 \neq t_2} \big(k_1(x_{i,t_1}) k_2(x_{i,t_2}) k_3(x_{i,t_2}) \\ &\quad\quad + k_1(x_{i,t_2}) k_2(x_{i,t_1}) k_3(x_{i,t_2}) + k_1(x_{i,t_2}) k_2(x_{i,t_2}) k_3(x_{i,t_1}) \big) \\ &\quad + \frac{T^2 - T - 2}{T^{1/2}(T-1)^2(T-2)} \sum_{t_1 \neq t_2 \neq t_3} k_1(x_{i,t_1}) k_2(x_{i,t_2}) k_3(x_{i,t_3}) \end{aligned}$$

The proof of this Lemma is simple algebra and is left for the Supplementary Appendix, which also details results for higher-order terms.

For the split-sample correction we write $W_{i,T,m}^{(1)}, W_{i,T,m}^{(2)}$ for the V-statistics using the first and second halves of time periods, so that the expansion terms take the form

$$T^{-m/2} \check{W}_{i,T,m} = 2 \times T^{-m/2} W_{i,T,m} - \frac{1}{2} \times (\frac{T}{2})^{-m/2} \big(W_{i,T,m}^{(1)} + W_{i,T,m}^{(2)} \big)$$

Lemma B.2. *Let* $\check{W}_{i,T,m} = 2W_{i,T,m} - 2^{m/2} \frac{1}{2} (W_{i,T,m}^{(1)} + W_{i,T,m}^{(2)})$, *and let* $\tau_1 = \{1, 2, \ldots, M\}$

and $\tau_2 = \{M+1, \ldots, T\}$ *for* $T = 2M$*. Then:*

$$\begin{aligned}
\check{W}_{i,T,1} &= W_{i,T,1} \\
\check{W}_{i,T,2} &= 2\Big(\frac{1}{T}\sum_{t_1\in\tau_1}\sum_{t_2\in\tau_2} k_1(x_{it_1})k_2(x_{it_2}) + \frac{1}{T}\sum_{t_1\in\tau_2}\sum_{t_2\in\tau_1} k_1(x_{it_1})k_2(x_{it_2})\Big) \\
\check{W}_{i,T,3} &= 2W_{i,T,3} - 4\Big(\frac{1}{T^{3/2}}\sum_{t_1,t_2,t_3\in\tau_1} k_1(x_{it_1})k_2(x_{it_2})k_3(x_{it_3}) \\
&\quad + \frac{1}{T^{3/2}}\sum_{t_1,t_2,t_3\in\tau_2} k_1(x_{it_1})k_2(x_{it_2})k_3(x_{it_3})\Big)
\end{aligned}$$

Again, the proof is left for the Supplementary Appendix.

B.2 Proof of Theorem 1

We first note that from Lemmas B.1 and B.2, neither the jackknife nor the split-sample correction impact the first-order expansion term so that $A_n = \tilde{A}_{n,J} = \tilde{A}_{n,1/2}$ and the first-order variance $V_{1,n} = Var(\sqrt{n}A_n)$ is the same for both estimators.

Next, we consider the variance of the second order term. There are two forms of V-statistic contained in B_n, $W_{T,(2)}$ and $W_{T,(1,1)}$ and we examine each in turn.

For the first case, we have

$$\begin{aligned}
Var(\sqrt{n}\tilde{W}_{T,(2)}) &= Var\Big(\frac{1}{(T-1)}\frac{1}{\sqrt{n}}\sum_i\sum_{t_1\neq t_2} k_1\left(x_{i,t_1}\right)k_2\left(x_{i,t_2}\right)\Big) \\
&= \frac{1}{n}\sum_i \frac{1}{(T-1)^2}\sum_{t_1\neq t_2}\Big(E\big[k_1(x_{i,t_1})^2\big]E\big[k_2(x_{i,t_2})^2\big] \\
&\quad + E\big[k_1(x_{i,t_1})k_2(x_{i,t_1})\big]E\big[k_1(x_{i,t_2})k_2(x_{i,t_2})\big]\Big) \\
&= \frac{T}{T-1}\frac{1}{n}\sum_i\Big(E\big[k_1(x_{i,t})^2\big]E\big[k_2(x_{i,t})^2\big] \\
&\quad + E\big[k_1(x_{i,t})k_2(x_{i,t})\big]^2\Big)
\end{aligned}$$

$$\begin{aligned}
&Var(\sqrt{n}\check{W}_{T,(2)}) \\
&= Var\Big(\frac{1}{\sqrt{n}}\sum_i\frac{2}{T}\sum_{t_1\in\tau_1}\sum_{t_2\in\tau_2}\big(k_1(x_{it_1})k_2(x_{it_2}) + k_1(x_{it_2})k_2(x_{it_1})\big)\Big)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_i \frac{4}{T^2}\sum_{t_1\in\tau_1}\sum_{t_2\in\tau_2} 2\Big(E[k_1(x_{it_1})^2]E[k_2(x_{it_2})^2] \\
&\quad + E[k_1(x_{it_1})k_2(x_{it_1})]E[k_1(x_{it_2})k_2(x_{it_2})]\Big) \\
&= \frac{2}{n}\sum_i \Big(E[k_1(x_{i,t})^2]E[k_2(x_{i,t})^2] + E[k_1(x_{i,t})k_2(x_{i,t})]^2\Big) \\
&= 2\frac{T-1}{T} Var(\sqrt{n}\tilde{W}_{T,(2)})
\end{aligned}
$$

For the second case we have

$$
\begin{aligned}
Var(\sqrt{n}\tilde{W}_{T,(1,1)}) &= Var\Big(\frac{1}{n^{3/2}}\sum_{i_1,i_2}\frac{1}{T-1}\sum_{t_1,t_2} k_1(x_{i_1,t_1})k_2(x_{i_2,t_2})\Big) \\
&= \frac{T-1}{T}\frac{1}{n^3}\sum_{i_1,i_2}\Big(E\big[k_1(x_{i_1,t})^2\big]E\big[k_2(x_{i_2,t})^2\big] \\
&\quad + E\big[k_1(x_{i_1,t})k_2(x_{i_1,t})\big]E\big[k_1(x_{i_2,t})k_2(x_{i_2,t})\big]\Big) \\
&= O(n^{-1}) \\
Var(\sqrt{n}\check{W}_{T,(1,1)}) &= 2\frac{T-1}{T} Var(\sqrt{n}\tilde{W}_{T,(1,1)}) = O(n^{-1})
\end{aligned}
$$

Finally, similar calculations give $Cov(\sqrt{n}\tilde{W}_{T,(2)}, \sqrt{n}\tilde{W}_{T,(1,1)}) = O(n^{-1})$ and $Cov(\sqrt{n}\check{W}_{T,(2)}, \sqrt{n}\check{W}_{T,(1,1)}) = O(n^{-1})$.

Next, we establish that $Cov(A_n, \tilde{B}_{n,J}) = Cov(A_n, \tilde{B}_{n,1/2}) = 0$. This result is immediate since the bias-corrected second order terms are sums over $k_1(x_{i,t_1})k_2(x_{i,t_2})$, for which $t_1 \neq t_2$. The product of these second-order terms with the first order terms in A_n will then contain triplets $k_1(x_{i,t_1})k_2(x_{i,t_2})k_3(x_{i,t_3})$ which must have mean zero by $t_1 \neq t_2$.

The final part of the higher-order variance comes from $Cov(\sqrt{n}A_n, \sqrt{n}\tilde{C}_{n,J})$ and $Cov(\sqrt{n}A_n, \sqrt{n}\tilde{C}_{n,1/2})$. Here we show that these terms are both $o(1)$. Recall from Lemma B.1 the third-order jackknife term has the form

$$
\begin{aligned}
\tilde{W}_{i,T,3} &= \frac{-1}{T^{1/2}(T-1)}\frac{1}{T}\sum_t k_1(x_{i,t})k_2(x_{i,t})k_3(x_{i,t}) \\
&\quad + \frac{1}{T^{1/2}(T-1)^2}\sum_{t_1\neq t_2}\big(k_1(x_{i,t_1})k_2(x_{i,t_2})k_3(x_{i,t_2}) \\
&\quad + k_1(x_{i,t_2})k_2(x_{i,t_1})k_3(x_{i,t_2}) + k_1(x_{i,t_2})k_2(x_{i,t_2})k_3(x_{i,t_!})\big)
\end{aligned}
$$

$$+\frac{T^2-T-2}{T^{1/2}(T-1)^2(T-2)}\sum_{t_1\neq t_2\neq t_3}k_1(x_{i,t_1})k_2(x_{i,t_2})k_3(x_{i,t_3})$$
$$=J_{i,1}+J_{i,2}+J_{i,3}$$

First consider $\tilde{W}_{T,(3)}=\frac{1}{n}\sum_i(J_{i,1}+J_{i,2}+J_{i,3})$ and its covariance with a first order statistic $\tilde{W}_{T,(1)}$.

$$\begin{aligned}
&Cov(\sqrt{n}\tilde{W}_{T,(1)},\sqrt{n}J_1)\\
&=\frac{-1}{T^{1/2}(T-1)}\frac{1}{n}\sum_{i_1,i_2}\frac{1}{T^{3/2}}\sum_{t_1,t_2}E\big[k_1(x_{i_1,t})k_2(x_{i_2,t})k_3(x_{i_3,t})k_4(x_{i_4,t})\big]\\
&=-\frac{1}{T-1}\frac{1}{n}\sum_i E\big[k_1(x_{i,t})k_2(x_{i,t})k_3(x_{i,t})k_4(x_{i,t})\big]\\
&=O(T^{-1})
\end{aligned}$$

$$\begin{aligned}
&Cov(\sqrt{n}\tilde{W}_{T,(1)},\sqrt{n}J_2)\\
&=\frac{1}{T(T-1)^2}\frac{1}{n}\sum_{i_1,i_2}\sum_t\sum_{t_2\neq t_3}E\Big[k_1(x_{i_1,t_1})\big(k_2(x_{i_2,t_3})k_3(x_{i_2,t_2})k_4(x_{i_2,t_2})\\
&\quad+k_1(x_{i_2,t_2})k_2(x_{i_2,t_3})k_3(x_{i_2,t_2})+k_1(x_{i_2,t_2})k_2(x_{i_2,t_2})k_3(x_{i_2,t_3})\big)\Big]\\
&=\frac{1}{n}\sum_i\frac{1}{T(T-1)^2}\sum_{t_1\neq t_2}\Big(E\big[k_1(x_{i,t_1})k_2(x_{i,t_1})\big]E\big[k_3(x_{i,t_2})k_4(x_{i,t_2})\big]\\
&\quad+E\big[k_1(x_{i,t_1})k_3(x_{i,t_1})\big]E\big[k_2(x_{i,t_2})k_4(x_{i,t_2})\big]\\
&\quad+E\big[k_1(x_{i,t_1})k_4(x_{i,t_1})\big]E\big[k_2(x_{i,t_2})k_3(x_{i,t_2})\big]\Big)\\
&=O(T^{-1})
\end{aligned}$$

Since J_3 contains terms with $t_1\neq t_2\neq t_3$, the covariance involving these terms is zero. Identical steps applied to $\sqrt{n}\tilde{W}_{T,(2,1)}$ and $\sqrt{n}\tilde{W}_{T,(1,1,1)}$ will confirm that the covariances of these terms with $\sqrt{n}\tilde{W}_{T,(1)}$ are also $O(T^{-1})$ or lower order, so that $Cov(\sqrt{n}A_n,\sqrt{n}\tilde{C}_{n,J})=o(1)$, as required.

For the split-sample case, we have

$$\check{W}_{i,T,3}=2W_{i,T,3}-4\Big(\frac{1}{T^{3/2}}\sum_{t_1,t_2,t_3\in\tau_1}k_1(x_{it_1})k_2(x_{it_2})k_3(x_{it_3})$$

$$
\begin{aligned}
&+ \frac{1}{T^{3/2}} \sum_{t_1,t_2,t_3 \in \tau_2} k_1(x_{it_1}) k_2(x_{it_2}) k_3(x_{it_3}) \Big) \\
&= 2W_{i,T,3} - 4S_1 - 4S_2
\end{aligned}
$$

The covariance with the first term is

$$
\begin{aligned}
&Cov(\sqrt{n}\check{W}_{T,(1)}, \sqrt{n}W_{T,(3)}) \\
&= \frac{1}{n^2} \sum_{i_1,i_2} \frac{1}{T^2} \sum_{t_1,t_2,t_3,t_4} E\big[k_1(x_{i_1,t_1}) k_2(x_{i_2,t_2}) k_3(x_{i_2,t_3}) k_4(x_{i_2,t_4})\big] \\
&= \frac{1}{n} \sum_{i} \frac{1}{T^2} \Big\{ \sum_{t} E\big[k_1(x_{i,t}) k_2(x_{i,t}) k_3(x_{i,t}) k_4(x_{i,t})\big] \\
&\quad + \sum_{t_1 \neq t_2} \Big(E\big[k_1(x_{i,t_1}) k_2(x_{i,t_1})\big] E\big[k_3(x_{i,t_2}) k_4(x_{i,t_2})\big] \\
&\quad + E\big[k_1(x_{i,t_1}) k_3(x_{i,t_1})\big] E\big[k_2(x_{i,t_2}) k_4(x_{i,t_2})\big] \\
&\quad + E\big[k_1(x_{i,t_1}) k_4(x_{i,t_1})\big] E\big[k_2(x_{i,t_2}) k_3(x_{i,t_2})\big] \Big) \Big\} \\
&= \frac{1}{T} \frac{1}{n} \sum_{i} E\big[k_1(x_{i,t}) k_2(x_{i,t}) k_3(x_{i,t}) k_4(x_{i,t})\big] \\
&\quad + \frac{T-1}{T} \frac{1}{n} \sum_{i} \Big(E\big[k_1(x_{i,t}) k_2(x_{i,t})\big] E\big[k_3(x_{i,t}) k_4(x_{i,t})\big] \\
&+ E\big[k_1(x_{i,t}) k_3(x_{i,t})\big] E\big[k_2(x_{i,t}) k_4(x_{i,t})\big] \\
&+ E\big[k_1(x_{i,t}) k_4(x_{i,t})\big] E\big[k_2(x_{i,t}) k_3(x_{i,t})\big] \Big)
\end{aligned}
$$

And for S_1

$$
\begin{aligned}
&Cov(\sqrt{n}\check{W}_{T,(1)}, \sqrt{n}S_1) \\
&= \frac{1}{n} \sum_{i} \frac{1}{T^2} \sum_{t_1,t_2,t_3,t_4 \in \tau_1} E\big[k_1(x_{it_1}) k_2(x_{it_2}) k_3(x_{it_3}) k_4(x_{it_4})\big] \\
&= \frac{1}{2} \frac{1}{T} \frac{1}{n} \sum_{i} E\big[k_1(x_{i,t}) k_2(x_{i,t}) k_3(x_{i,t}) k_4(x_{i,t})\big] \\
&\quad + \frac{1}{4} \frac{T-2}{T} \frac{1}{n} \sum_{i} \Big(E\big[k_1(x_{i,t}) k_2(x_{i,t})\big] E\big[k_3(x_{i,t}) k_4(x_{i,t})\big] \\
&+ E\big[k_1(x_{i,t}) k_3(x_{i,t})\big] E\big[k_2(x_{i,t}) k_4(x_{i,t})\big] \\
&+ E\big[k_1(x_{i,t}) k_4(x_{i,t})\big] E\big[k_2(x_{i,t}) k_3(x_{i,t})\big] \Big)
\end{aligned}
$$

and similarly for S_2. Combining, gives

$$\begin{aligned}
&Cov(\sqrt{n}\check{W}_{T,(1)}, \sqrt{n}\check{W}_{T,(3)}) \\
&= -2\frac{1}{T}\frac{1}{n}\sum_i E\big[k_1(x_{i,t})k_2(x_{i,t})k_3(x_{i,t})k_4(x_{i,t})\big] \\
&\quad + 2\frac{1}{T}\frac{1}{n}\sum_i \Big(E\big[k_1(x_{i,t})k_2(x_{i,t})\big]E\big[k_3(x_{i,t})k_4(x_{i,t})\big] \\
&\quad + E\big[k_1(x_{i,t})k_3(x_{i,t})\big]E\big[k_2(x_{i,t})k_4(x_{i,t})\big] \\
&+ E\big[k_1(x_{i,t})k_4(x_{i,t})\big]E\big[k_2(x_{i,t})k_3(x_{i,t})\big]\Big) \\
&= O(T^{-1})
\end{aligned}$$

Similar statements can be made for the terms $\check{W}_{T,(2,1)}$ and $\check{W}_{T,(1,1,1)}$, giving $Cov(\sqrt{n}A_n, \sqrt{n}\tilde{C}_{n,1/2}) = o(1)$ as before. We can now derive the expression for $V_{2,n}$ in the theorem. From Hahn and Newey (2004), we have

$$\begin{aligned}
B_n &= \frac{1}{2}\theta^{\epsilon\epsilon}(0) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{T}\sum_{s,t} k_1(x_{i,s})k_2(x_{i,t}) + o_p(1) \\
k_1(x_{i,t}) &= \mathcal{I}_n^{-1}\Big(\frac{E[U_{it}^{\alpha\alpha}]}{2E[V_{it}^2]}V_{it} + U_{it}^{\alpha}\Big) \\
k_2(x_{i,t}) &= \frac{1}{E[V_{it}^2]}V_{it}
\end{aligned}$$

We can then apply the form of $Var(\tilde{W}_{T,(2)})$ and $Var(\check{W}_{T,(2)})$ above to give the result.

B.3 Proof of Theorem 2

We begin by establishing an expansion for the bias estimators of the form

$$\begin{aligned}
\sqrt{nT}\frac{1}{T}(\hat{\beta}_J - \mathbf{B}) &= \frac{\sqrt{n}}{\sqrt{T}}(B_n - \tilde{B}_{n,J} - \mathbf{B}) + O_p(T^{-1}) \\
\sqrt{nT}\frac{1}{T}(\hat{\beta}_{1/2} - \mathbf{B}) &= \frac{\sqrt{n}}{\sqrt{T}}(B_n - \tilde{B}_{n,1/2} - \mathbf{B}) + O_p(T^{-1/2})
\end{aligned}$$

Since the implicit bias estimates are given by $\frac{1}{T}\hat{\beta} = \hat{\theta} - \tilde{\theta}$, the expansion in (8), and the equivalent expansions for the bias-corrected estimators, imply the result as long as $\frac{\sqrt{n}}{T}(C_n - \tilde{C}_{n,J}) = O_p(T^{-1})$ and $\frac{\sqrt{n}}{T}(C_n - \tilde{C}_{n,1/2}) = O_p(T^{-1/2})$. This follows straightforwardly from

Lemmas B.1 and B.2, and is shown in the Supplementary Appendix.

Given this expansion for the bias estimates, we analyze the V-statistic forms that are present in $\frac{\sqrt{n}}{\sqrt{T}}(B_n - \tilde{B}_{n,J} - \mathbf{B})$

$$\begin{aligned}
&\frac{\sqrt{n}}{\sqrt{T}}\big(W_{T,(2)} - \tilde{W}_{T,(2)} - E[W_{T,(2)}]\big) \\
&= \frac{1}{\sqrt{n}}\sum_i \frac{1}{T\sqrt{T}}\sum_{t_1,t_2}\big(k_1(x_{i,t_1})k_2(x_{i,t_2}) - E[k_1(x_{i,t_1})k_2(x_{i,t_2})]\big) \\
&\quad - \frac{1}{\sqrt{n}}\sum_i \frac{1}{(T-1)\sqrt{T}}\sum_{t_1\neq t_2} k_1(x_{i,t_1})k_2(x_{i,t_2}) \\
&= \frac{1}{T}\frac{1}{\sqrt{nT}}\sum_i\sum_t\big(k_1(x_{i,t})k_2(x_{i,t}) - E[k_1(x_{i,t})k_2(x_{i,t})]\big) \\
&\quad - \frac{1}{\sqrt{T}(T-1)}\frac{1}{\sqrt{n}}\sum_i \frac{1}{T}\sum_{t_1\neq t_2} k_1(x_{i,t_1})k_2(x_{i,t_2}) \\
&= O_p(T^{-1})
\end{aligned}$$

And similarly, we can show $\frac{\sqrt{n}}{\sqrt{T}}\big(W_{T,(1,1)} - \tilde{W}_{T,(1,1)} - E[W_{T,(1,1)}]\big) = O_p(T^{-1})$. Consequently, $\frac{\sqrt{n}}{\sqrt{T}}(B_n - \tilde{B}_{n,J} - \mathbf{B}) = O_p(T^{-1})$.

For the split-sample version of the statistics, let $1_{t_1,t_2}$ be an indicator that is equal to one whenever t_1 and t_2 are in the same half of time periods, and zero otherwise, and $\tilde{1}_{t_1,t_2} = (-1)^{(1-1_{t_1,t_2})}$. Then we have

$$\begin{aligned}
&\frac{\sqrt{n}}{\sqrt{T}}\big(W_{T,(2)} - \check{W}_{T,(2)} - E[W_{T,(2)}]\big) \\
&= \frac{1}{\sqrt{n}}\sum_i \frac{1}{T\sqrt{T}}\sum_{t_1,t_2}\big(k_1(x_{i,t_1})k_2(x_{i,t_2}) - E[k_1(x_{i,t_1})k_2(x_{i,t_2})]\big) \\
&\quad - \frac{1}{\sqrt{n}}\sum_i \frac{2}{T\sqrt{T}}\sum_{t_1\in\tau_1}\sum_{t_2\in\tau_2}\big(k_1(x_{it_1})k_2(x_{it_2}) + k_1(x_{it_2})k_2(x_{it_1})\big) \\
&= \frac{1}{T\sqrt{nT}}\sum_i\sum_{t_1,t_2}\tilde{1}_{t_1,t_2}\big(k_1(x_{i,t_1})k_2(x_{i,t_2}) - E[k_1(x_{i,t_1})k_2(x_{i,t_2})]\big)
\end{aligned}$$

Consider the variance of the final term

$$\frac{1}{nT^3}Var\Big(\sum_i\sum_{t_1,t_2}\tilde{1}_{t_1,t_2}\big(k_1(x_{i,t_1})k_2(x_{i,t_2}) - E[k_1(x_{i,t_1})k_2(x_{i,t_2})]\big)\Big)$$

$$
\begin{aligned}
&= \frac{1}{nT^3} \sum_{i_1,i_2} \sum_{t_1,t_2,t_3,t_4} E\Big[\big(k_1(x_{i_1,t_1})k_2(x_{i_1,t_2}) - E[k_1(x_{i_1,t_1})k_2(x_{i_1,t_2})]\big) \\
&\times \big(k_1(x_{i_2,t_3})k_2(x_{i_2,t_4}) - E[k_1(x_{i_2,t_3})k_2(x_{i_2,t_4})]\big)\Big] \tilde{1}_{t_1,t_2} \tilde{1}_{t_3,t_4} \\
&= \frac{1}{nT^3} \sum_{i} \sum_{t} E\Big[\big(k_1(x_{i,t})k_2(x_{i,t}) - E[k_1(x_{i,t})k_2(x_{i,t})]\big)^2\Big] \\
&\quad + \frac{1}{nT^3} \sum_{i} \sum_{t_1 \neq t_2} \Big(E[k_1(x_{i,t_1})^2]E[k_2(x_{i,t_2})^2] \\
&\quad + E[k_1(x_{i,t_1})k_2(x_{i,t_1})]E[k_1(x_{i,t_2})k_2(x_{i,t_2})]\Big) \\
&= O(T^{-1})
\end{aligned}
$$

Consequently, $\frac{\sqrt{n}}{\sqrt{T}}\big(W_{T,(2)} - \check{W}_{T,(2)} - E[W_{T,(2)}]\big) = O_p(T^{-1/2})$ as required. Again, the same steps can also be used to show $\frac{\sqrt{n}}{\sqrt{T}}\big(W_{T,(1,1)} - \check{W}_{T,(1,1)} - E[W_{T,(1,1)}]\big) = O_p(T^{-1/2})$, and hence that $\frac{\sqrt{n}}{\sqrt{T}}(B_n - \tilde{B}_{n,1/2} - \mathbf{B}) = O_p(T^{-1/2})$, giving the result of the theorem.

B.4 Proof of Theorem 3

As previously, the proof proceeds in terms of general V-statistics that match the order of those in the relevant expansion terms. We begin by deriving the form of $E[C_n]$ and $E[D_n]$, by taking the expectations of third and fourth-order V-statistics. For a third-order statistic we have

$$
\begin{aligned}
E\big[\sqrt{T}W_{T,(3)}\big] &= \frac{1}{n}\sum_{i} \frac{1}{T} \sum_{t_1,t_2,t_3} E\big[k_1(x_{i,t_1})k_2(x_{i,t_2})k_3(x_{i,t_3})\big] \\
&= \frac{1}{n}\sum_{i} E\big[k_1(x_{i,t})k_2(x_{i,t})k_3(x_{i,t})\big] \\
E\big[\sqrt{T}W_{T,(2,1)}\big] &= \frac{1}{n^2}\sum_{i} E\big[k_1(x_{i,t})k_2(x_{i,t})k_3(x_{i,t})\big] \\
E\big[\sqrt{T}W_{T,(1,1,1)}\big] &= \frac{1}{n^3}\sum_{i} E\big[k_1(x_{i,t})k_2(x_{i,t})k_3(x_{i,t})\big]
\end{aligned}
$$

Clearly $\lim_{n\to\infty} E\big[W_{T,(2,1)}\big] = \lim_{n\to\infty} E\big[W_{T,(1,1,1)}\big] = 0$ and we may ignore these terms. From Lemmas B.1 and B.2 this is true of the jackknife and split-sample versions of these statistics as well, so that we need only focus on $W_{T,(3)}$ and its bias-corrected counterparts.

Computing these expectations gives

$$\begin{aligned}
E\big[\sqrt{T}\tilde{W}_{T,(3)}\big] &= -\frac{T}{(T-1)}\frac{1}{n}\sum_i E\big[k_1(x_{i,t})k_2(x_{i,t})k_3(x_{i,t})\big] \\
&= -\frac{T}{T-1}E\big[\sqrt{T}W_{T,(3)}\big] \\
E\big[\sqrt{T}\check{W}_{T,(3)}\big] &= 2E\big[\sqrt{T}W_{T,(3)}\big] - 4\frac{1}{T}\sum_{t\in\tau_1} E[k_1(x_{it})k_2(x_{it})k_3(x_{it})] \\
&\quad + \sum_{t\in\tau_2} E[k_1(x_{it})k_2(x_{it})k_3(x_{it})] \\
&= -2E\big[\sqrt{T}W_{T,(3)}\big]
\end{aligned}$$

Next we consider the fourth-order V-statistics in D_n. Taking the expectation of $W_{T,(4)}$ gives

$$\begin{aligned}
E[W_{T,(4)}] &= \frac{1}{n}\sum_i \frac{1}{T^2}\sum_{t_1,t_2,t_3,t_4} E\big[k_1(x_{i,t_1})k_2(x_{i,t_1})k_3(x_{i,t_2})k_4(x_{i,t_4})\big] \\
&= \frac{1}{n}\sum_i \frac{1}{T^2}\sum_t E\big[k_1(x_{i,t})k_2(x_{i,t})k_3(x_{i,t})k_4(x_{i,t})\big] \\
&\quad + \frac{1}{n}\sum_i \frac{1}{T^2}\sum_{t_1\neq t_2} E\Big[k_1(x_{i,t_1})k_2(x_{i,t_1})k_3(x_{i,t_2})k_4(x_{i,t_2}) \\
&\quad + k_1(x_{i,t_1})k_2(x_{i,t_2})k_3(x_{i,t_1})k_4(x_{i,t_2}) \\
&\quad + k_1(x_{i,t_1})k_2(x_{i,t_2})k_3(x_{i,t_2})k_4(x_{i,t_1})\Big] \\
&= \frac{T-1}{T}\frac{1}{n}\sum_i \Big(E[k_1(x_{i,t})k_2(x_{i,t})]E[k_3(x_{i,t})k_4(x_{i,t})] \\
&\quad + E[k_1(x_{i,t})k_3(x_{i,t})]E[k_2(x_{i,t})k_4(x_{i,t})] \\
&\quad + E[k_1(x_{i,t})k_4(x_{i,t})]E[k_2(x_{i,t})k_3(x_{i,t})]\Big) + o(1)
\end{aligned}$$

In the Supplementary Appendix it is shown that the fourth-order jackknifed V-statistic has the form

$$\begin{aligned}
\tilde{W}_{i,T,4} &= \frac{-2T+1}{T(T-1)^2}\sum_t k_1(x_{i,t})k_2(x_{i,t})k_3(x_{i,t})k_4(x_{i,t}) \\
&- \frac{T^2-3T+1}{T(T-1)^3}\sum_{t_1\neq t_2}\Big(k_1(x_{i,t_1})k_2(x_{i,t_2})k_3(x_{i,t_2})k_4(x_{i,t_2}) \\
&+ k_1(x_{i,t_2})k_2(x_{i,t_1})k_3(x_{i,t_2})k_4(x_{i,t_2}) + k_1(x_{i,t_2})k_2(x_{i,t_2})k_3(x_{i,t_1})k_4(x_{i,t_2})
\end{aligned}$$

$$
\begin{aligned}
&+ k_1(x_{i,t_2})k_2(x_{i,t_2})k_3(x_{i,t_2})k_4(x_{i,t_1})\Big) \\
&- \frac{T^2-3T+1}{T(T-1)^3}\sum_{t_1\neq t_2}\Big(k_1(x_{i,t_1})k_2(x_{i,t_1})k_3(x_{i,t_2})k_4(x_{i,t_2}) \\
&+ k_1(x_{i,t_1})k_2(x_{i,t_2})k_3(x_{i,t_1})k_4(x_{i,t_2}) + k_1(x_{i,t_1})k_2(x_{i,t_2})k_3(x_{i,t_2})k_4(x_{i,t_1})\Big) \\
&+ \frac{3T-1}{T(T-1)^3}\sum_{t_1\neq t_2\neq t_3}\Big(k_1(x_{i,t_1})k_2(x_{i,t_1})k_3(x_{i,t_2})k_4(x_{i,t_3}) \\
&+ k_1(x_{i,t_1})k_2(x_{i,t_2})k_3(x_{i,t_1})k_4(x_{i,t_3}) + k_1(x_{i,t_1})k_2(x_{i,t_2})k_3(x_{i,t_3})k_4(x_{i,t_1}) \\
&+ k_1(x_{i,t_2})k_2(x_{i,t_1})k_3(x_{i,t_1})k_4(x_{i,t_3}) + k_1(x_{i,t_2})k_2(x_{i,t_1})k_3(x_{i,t_3})k_4(x_{i,t_1}) \\
&+ k_1(x_{i,t_2})k_2(x_{i,t_3})k_3(x_{i,t_1})k_4(x_{i,t_1})\Big) \\
&+ \frac{T^2+3T-1}{T(T-1)^3}\sum_{t_1\neq t_2\neq t_3\neq t_4} k_1(x_{i,t_1})k_2(x_{i,t_1})k_3(x_{i,t_2})k_4(x_{i,t_4})
\end{aligned}
$$

Taking expectations gives

$$
\begin{aligned}
E[\tilde{W}_{T,4}] &= -\frac{T^2-3T+1}{(T-1)^2}\frac{1}{n}\sum_i\Big(E[k_1(x_{i,t})k_2(x_{i,t})]E[k_3(x_{i,t})k_4(x_{i,t})] \\
&+ E[k_1(x_{i,t})k_3(x_{i,t})]E[k_2(x_{i,t})k_4(x_{i,t})] \\
&+ E[k_1(x_{i,t})k_4(x_{i,t})]E[k_2(x_{i,t})k_3(x_{i,t})]\Big) + o(1) \\
&= -\frac{T}{T-1}E[W_{T,4}] + o(1)
\end{aligned}
$$

The fourth order term $\tilde{D}_{n,J}$ also contains fourth-order V-statistics of the form $\tilde{W}_{T,(3,1)}$, $\tilde{W}_{T,(2,2)}$, $\tilde{W}_{T,(2,1,1)}$, and $\tilde{W}_{T,(1,1,1,1)}$, and the same relationship holds for these terms.

The fourth-order split-sample statistic has the form

$$
\begin{aligned}
\check{W}_{i,T,4} = 2W_{i,T,4} - 8\Big(&\frac{1}{T^2}\sum_{t_1,t_2,t_3,t_4\in\tau_1} k_1(x_{it_1})k_2(x_{it_2})k_3(x_{it_3})k_4(x_{it_4}) \\
&+ \frac{1}{T^2}\sum_{t_1,t_2,t_3,t_4\in\tau_2} k_1(x_{it_1})k_2(x_{it_2})k_3(x_{it_3})k_4(x_{it_4})\Big)
\end{aligned}
$$

and taking expectations it is straightforward to see that

$$
E[\check{W}_{T,(4)}] = 2E[W_{T,(4)}] = -2E[W_{T,(4)}] + o(1)
$$

As before, identical results can be shown for the other forms of fourth-order statistics. Combining these results gives the statement in the theorem.

C Appendix for Chapter 3

C.1 Two useful lemmas

In the proofs below, the following notation is used. For some random variable W_i, let $\bar{w}_i = E[W_i|Z_i]$ and $\tilde{w}_i = W_i - \bar{w}_i$. Also, denote $\bar{\mu}_w = \max|\bar{w}_i|$ and $\bar{\sigma}_w^2 = \max_{1\leq i\leq n} Var(W_i|Z_i)$. For J-vectors we use the Euclidean norm $\|x\|^2 \coloneqq \sum_{i=1}^J x_j^2$, and for $J \times K$ matrices the operator norm will be used $\|M\| \coloneqq \sqrt{\lambda_{\max}(M^*M)}$. Let C be a finite constant that differs in different usages.

We will make use of two key results for bounding quadratic forms that have been derived elsewhere. The first is Lemma A1 in Chao et al. (2012), which provides a bound for jackknifed projections, which straightforwardly extends to the vector case and can be applied to numerator terms for our estimators. The second is a bound on the operator norm of matrix-valued U-statistics from Minsker and Wei (2019), which we apply to the jackknifed projections that appear in denominator terms of the estimators.

We begin by repeating both results here

Lemma C.1. *(Lemma A1 of Chao et. al., 2012) If, conditional on Z, (W_i, Y_i), $i = 1, \ldots, n$ are independent with probability one, and if W_i and Y_i are scalars, and P is a symmetric, idempotent matrix of rank K, then there exists a positive constant C such that*

$$E\Big[\Big(\sum_{i\neq j} P_{ij}W_iY_j - \sum_{i\neq j} P_{ij}\bar{w}_i\bar{y}_j\Big)^2|Z\Big] \leq CD_n$$

where $D_n = K\bar{\sigma}_w^2\bar{\sigma}_y^2 + \bar{\sigma}_w^2\bar{y}'\bar{y} + \bar{\sigma}_y^2\bar{w}'\bar{w}$.

When W_i is a J_n-vector, then Lemma C.1 implies that

$$E\big[\|\sum_{i\neq j} P_{ij}W_iY_j - \sum_{i\neq j} P_{ij}\bar{w}_i\bar{y}_j\|^2\big|\mathcal{Z}\big] = O_p(J_nD_n)$$

Before repeating the next result, we introduce some additional notation for the theorem. Define a generalized U-statistic of order 2

$$U_n = \sum_i \sum_{j\neq i} H_{ij}(X_i, X_j)$$

where $X_1, \ldots, X_n$ are a sequence of i.i.d. random variables taking values in a measurable

space $(\mathcal{S}, \mathcal{B})$, and $H_{ij} : \mathcal{S}^2 \mapsto \mathbb{H}^d$ are permutation symmetric (i.e. $H_{ij}(X_i, X_j) = H_{ji}(X_j, X_i)$), and $\mathbb{H}^d$ is the set of $d \times d$ self-adjoint matrices. We also define $\{X_i^{(k)}\}_{i=1}^n$ as independent copies of the X_i sequence. Expectation with respect to $\{X_i^{(k)}\}_{i=1}^n$ only (conditional on other variables) is denoted by $E_{(k)}[\cdot]$.

Lemma C.2. *(Theorem 3.1 of Minsker and Wei, 2019) Let $\{X_i^{(k)}\}_{i=1}^n$, $k = 1, 2$, be $\mathcal{S}$-valued i.i.d. random variables, and let $H_{ij} : \mathcal{S} \times \mathcal{S} \mapsto \mathbb{H}^d$ be permutation-symmetric degenerate kernels. Then, for $r = \log(ed)$*

$$\begin{aligned}\big(E\big[\|U_n\|^2\big]\big)^{1/2} \leq \frac{128}{\sqrt{e}}\Bigg\{ r\|\sum_{i\neq j} E\big[H_{ij}(X_i^{(1)}, X_j^{(2)})^2\big]\|^{1/2} + r\Big(E\|E_2\tilde{G}\tilde{G}^*\|\Big)^{1/2} \\ + 16r^{3/2}E\Big[\max_i \|\sum_{j\neq i} H_{ij}(X_i^{(1)}, X_j^{(2)})^2\|\Big]^{1/2}\Bigg\}\end{aligned}$$

where $\tilde{G}$ is the $nd \times nd$ matrix

$$\tilde{G} \coloneqq \begin{pmatrix} 0 & H_{12}(X_1^{(1)}, X_2^{(2)}) & \cdots & H_{1n}(X_1^{(1)}, X_n^{(2)}) \\ H_{21}(X_2^{(1)}, X_1^{(2)}) & 0 & \cdots & H_{2n}(X_2^{(1)}, X_n^{(2)}) \\ \vdots & \vdots & \ddots & \vdots \\ H_{n1}(X_n^{(1)}, X_1^{(2)}) & H_{n1}(X_n^{(1)}, X_2^{(2)}) & \cdots & 0 \end{pmatrix}$$

C.2 Consistency of estimators

The next lemma applies Lemma C.2 to part of the denominator of the JIVE estimator

Lemma C.3. *Let Assumptions 3.1 to 3.3 hold. Then*

$$\|\sum_{i\neq j} S_n^{-1} V_i P_{ij} V_j' (S_n^{-1})'\| \to 0$$

Proof. To apply the result in Lemma C.2 we write $U_n = \sum_i \sum_{j\neq i} P_{ij} v_i v_j'$, where $v_i = V_i/\sqrt{r_n}$. Using $P_{ij} = P_{ji}$ for a projection matrix, we can then let

$$H_{ij}(X_i, X_j) = \frac{1}{2} P_{ij}(v_i v_j' + v_j v_i')$$

so that $U_n = \sum_i \sum_{j\neq i} H_{ij}(X_i, X_j)$ for a permutation symmetric kernel H_{ij}, as required. Note that, *conditional on Z*, the kernel functions H_{ij} are non-random.

We consider each of the three terms in the bound in turn. Firstly,

$$\begin{aligned}
\|\sum_i \sum_{j\neq i} E\big[H_{ij}(X_i^{(1)}, X_j^{(2)})^2 \big| \mathcal{Z}\big]\|^{1/2} =& \frac{1}{4}\|\sum_i \sum_{j\neq i} E\big[P_{ij}^2\big(v_i v_i' v_j v_j' + v_j v_j' v_i v_i' \\
&+ v_i v_i'(v_j' v_j) + v_j v_j'(v_i' v_i)\big)\big|\mathcal{Z}\big]\|^{1/2} \\
\leq& \frac{1}{4}\Big(\sum_i \sum_{j\neq i} P_{ij}^2\Big(\|E\big[v_i v_i' v_j v_j' + v_j v_j' v_i v_i' \\
&+ v_i v_i'(v_j' v_j) + v_j v_j'(v_i' v_i)\big|\mathcal{Z}\big]\|\Big)^{1/2} \\
=& \frac{1}{4}\Big(\sum_i \sum_{j\neq i} P_{ij}^2\Big(\|\Sigma_i\Sigma_j + \Sigma_j\Sigma_i + \Sigma_i tr(\Sigma_j) + \Sigma_j tr(\Sigma_i)\big]\|\Big)^{1/2} \\
\leq& C\big(\frac{J_n}{r_n^2}\sum_i \sum_{j\neq i} P_{ij}^2\big)^{1/2} \\
\leq& C\frac{\sqrt{J_n K_n}}{r_n}
\end{aligned}$$

where we use a property of projection matrices

$$\sum_i \sum_{j\neq i} P_{ij}^2 \leq \sum_i \sum_j P_{ij}^2 = \sum_i P_{ii} = K_n$$

For the second term, we note that the (i,j) block of $\tilde{G}\tilde{G}^*$ is given by

$$\begin{aligned}
\big(\tilde{G}\tilde{G}^*\big)_{ij} &= \sum_k H_{ik}(X_i^{(1)}, X_k^{(2)}) H_{jk}(X_j^{(1)}, X_k^{(2)}) \\
&= \frac{1}{4}\sum_k P_{ik}P_{jk}\big(v_i v_k' v_j v_k' + v_k v_i' v_j v_k' + v_i v_k' v_k v_j' + v_k v_i' v_k v_j'\big)
\end{aligned}$$

and

$$\big(E_2\tilde{G}\tilde{G}^*\big)_{ij} = \frac{1}{r_n}\frac{1}{4}\sum_k P_{ik}P_{jk}\big(v_i v_j'\Sigma_k + v_i' v_j \Sigma_k + v_i v_j' tr(\Sigma_k) + \Sigma_k v_i v_j'\big)$$

We next make use of the following matrix inequality result (see Hiroshima (2003) and Lin and Wolkowicz (2012)). Let H be a nonnegative definite matrix,

$$H = \begin{bmatrix} A & X \\ X^* & B \end{bmatrix} \geq 0 \implies \|H\| \leq \|A + B\|$$

for a unitarily invariant norm $\|\cdot\|$. Since $\tilde{G}\tilde{G}^*$ is symmetric and non-negative definite, we can apply the inequality to give

$$\begin{aligned}
E\big[\|E_2\tilde{G}\tilde{G}^*\|\big|\mathcal{Z}\big] &\le E\big[\|\sum_i \big(E_2\tilde{G}\tilde{G}^*\big)_{ii}\|\big|\mathcal{Z}\big] \\
&= \frac{1}{r_n}\frac{1}{4}E\big[\|\sum_i\sum_k P_{ik}^2\big(v_iv_i'\Sigma_k + v_i'v_i\Sigma_k + v_iv_i'tr(\Sigma_k) + \Sigma_k v_iv_i'\big)\|\big|\mathcal{Z}\big] \\
&\le C\frac{1}{r_n}\Big(E\big[\|\sum_i\sum_k P_{ik}^2v_iv_i'\|\big|\mathcal{Z}\big] + E\big[\|\sum_i\sum_k P_{ik}^2v_i'v_i\|\big|\mathcal{Z}\big] \\
&+ J_nE\big[\|\sum_i\sum_k P_{ik}^2v_iv_i'\|\big|\mathcal{Z}\big] + E\big[\|\sum_i\sum_k P_{ik}^2v_iv_i'\|\big|\mathcal{Z}\big]\Big) \\
&\le C\frac{1}{r_n}E\big[\|\sum_i P_{ii}v_i'v_i\|\big|\mathcal{Z}\big] + C\frac{J_n}{r_n}E\big[\|\sum_i P_{ii}v_iv_i'\|\big|\mathcal{Z}\big] \\
&\le C\frac{1}{r_n}\sum_i P_{ii}E\big[\|v_i\|^2\big|\mathcal{Z}\big] + C\frac{J_n}{r_n}E\big[\|\sum_i P_{ii}v_iv_i'\|\big|\mathcal{Z}\big] \\
&\le C\frac{J_nK_n}{r_n^2}
\end{aligned}$$

Next, we tackle the final term.

$$\begin{aligned}
\|\sum_{j\neq i} H_{ij}(X_i^{(1)}, X_j^{(2)})^2\| =& \frac{1}{4}\|\sum_{j\neq i} P_{ij}^2\big(v_iv_j'v_iv_j' + v_iv_j'v_jv_i' \\
&+ v_jv_i'v_i'v_j + v_jv_i'v_j'v_i\big)\| \\
=& \frac{1}{4}\|\sum_{j\neq i} P_{ij}^2\big(v_iv_i'v_jv_j' + v_jv_j'v_iv_i' \\
&+ v_iv_i'(v_j'v_j) + v_jv_j'(v_i'v_i)\big)\| \\
\le& \|v_iv_i'\|\cdot\|\sum_{j\neq i} P_{ij}^2v_jv_j'\| + \|v_iv_i'\|\cdot\|\sum_{j\neq i} P_{ij}^2v_j'v_j\| \\
&+ \|v_i\|^2\|\sum_{j\neq i} P_{ij}^2v_jv_j'\| \\
\le& C\frac{J_nK_n}{r_n^2}
\end{aligned}$$

where we use

$$\|\sum_{j\neq i} P_{ij}^2v_jv_j'\| \le \|\sum_{j\neq i} P_{jj}v_jv_j'\| \le CK_n/r_n$$

$$\|\sum_{j\neq i} P_{ij}^2 v_j' v_j\| \leq \sup_j \|v_j\|^2 \cdot \|\sum_{j\neq i} P_{ij}^2\| \leq C J_n / r_n$$

This gives

$$E\Big[\max_i \|\sum_{j\neq i} H_{ij}(X_i^{(1)}, X_j^{(2)})^2\| \Big| \mathcal{Z}\Big] \leq C\frac{J_n K_n}{r_n^2}$$

Combining these results,

$$\begin{aligned}\big(E\big[\|U_n\|^2\big|\mathcal{Z}\big]\big)^{1/2} &\leq \frac{1}{r_n} C\big(r\sqrt{J_n K_n} + r^{3/2}\sqrt{J_n K_n}\big) \\ &\leq C\frac{\log^{3/2} J_n \sqrt{J_n K_n}}{r_n}\end{aligned}$$

Then, we have

$$\begin{aligned}E\big[\|\sum_{i\neq j} S_n^{-1} V_i P_{ij} V_j' (S_n^{-1})'\|^2\big|\mathcal{Z}\big] &\leq r_n^2 \|S_n^{-1}\|^4 E\big[\|\sum_{i\neq j} v_i P_{ij} v_j'\|^2\big|\mathcal{Z}\big] \\ &\leq C\frac{\log^{3/2} J_n \sqrt{J_n K_n}}{r_n} \to 0\end{aligned}$$

Using a conditional Markov inequality, we have for all $c > 0$

$$P\Big(\|\sum_{i\neq j} S_n^{-1} V_i P_{ij} V_j' (S_n^{-1})'\| > c\Big|\mathcal{Z}\Big) \to 0$$

and by a dominated convergence theorem the same is true unconditionally, giving the result in the lemma. □

Lemma C.4. *Let Assumptions 3.1 to 3.3 hold. Then*

$$\|S_n^{-1}\sum_i \sum_{j\neq i} P_{ij} X_i X_j' (S_n')^{-1} - \frac{1}{n}\sum_i (1 - P_{ii})\Pi_0 Z_i Z_i' \Pi_0'\| \to 0$$

Proof. We can decompose the JIVE denominator as

$$S_n^{-1}\sum_i \sum_{j\neq i} P_{ij} X_i X_j' (S_n')^{-1} = \frac{1}{n}\sum_i (1 - P_{ii}) z_i z_i' + \frac{1}{\sqrt{n}}\sum_i (1 - P_{ii}) z_i V_i' (S_n')^{-1}$$

$$+ S_n^{-1}\frac{1}{\sqrt{n}}\sum_i (1-P_{ii})V_i z_i' + S_n^{-1}\sum_i\sum_{j\neq i} P_{ij}V_iV_j'(S_n')^{-1}$$

By Assumption 3.3 we can replace z_i with $\Pi_0 Z_i$, since for example

$$\begin{aligned} E\Big[\|\frac{1}{\sqrt{n}}\sum_i (1-P_{ii})\big(z_i-\Pi' Z_i\big)V_i'(S_n')^{-1}\|^2\Big|\mathcal{Z}\Big] &\leq \frac{1}{nr_n}\sum_i \|\big(z_i-\Pi' Z_i\big)\|^2 E\Big[\|V_i\|^2\Big|\mathcal{Z}\Big] \\ &\leq C\frac{J_n}{r_n}\frac{1}{n}\sum_i \|\big(z_i-\Pi' Z_i\big)\|^2 \to 0 \end{aligned}$$

This allows us to write

$$\begin{aligned} S_n^{-1}\sum_i\sum_{j\neq i} P_{ij}X_iX_j'(S_n')^{-1} &= \frac{1}{n}\sum_i (1-P_{ii})\Pi_0 Z_iZ_i'\Pi_0' + \frac{1}{\sqrt{n}}\sum_i (1-P_{ii})\Pi_0 Z_iV_i'(S_n')^{-1} \\ &+ S_n^{-1}\frac{1}{\sqrt{n}}\sum_i (1-P_{ii})V_iZ_i'\Pi_0' + S_n^{-1}\sum_i\sum_{j\neq i} P_{ij}V_iV_j'(S_n')^{-1} + o_p(1) \end{aligned}$$

The final term is shown to be $o_p(1)$ in the operator norm in Lemma C.3. For the second/third terms

$$\begin{aligned} E\|\frac{1}{\sqrt{n}}\sum_i (1-P_{ii})\Pi_0 Z_iV_i'(S_n')^{-1}\| &\leq E\|\frac{1}{\sqrt{nr_n}}\sum_i \Pi_0 Z_iV_i'\| \\ &\leq C\frac{J_n\log J_n}{\sqrt{nr_n}} \to 0 \end{aligned}$$

by Assumption 3.5, and Rudelson's LLN for matrices (as in Belloni et al. (2015)). This gives the result of the lemma. □

Jackknife consistency

Here we prove the consistency of the jackknife estimator for the parameter vector β in the Euclidean norm. Specifically

$$\|\big(\sum_{i\neq j} X_iP_{ij}X_j'\big)^{-1}\sum_{i\neq j} X_iP_{ij}y_j - \beta_n\| \to 0$$

By Lemma (C.4), since $\frac{1}{n}\sum_i (1-P_{ii})z_iz_i'$ has minimum eigenvalue bounded below by some positive value, with probability approaching one this is also true for $\sum_{i\neq j} S_n^{-1}X_iP_{ij}X_j'(S_n^{-1})'$

and hence the inverse will exist and have largest eigenvalue bounded above. Using this result, we have

$$\begin{aligned}
&\|\big(\sum_{i\neq j} X_i P_{ij} X_j'\big)^{-1} \sum_{i\neq j} X_i P_{ij} y_j - \beta_n\| \\
&\qquad = \|\big(\sum_{i\neq j} X_i P_{ij} X_j'\big)^{-1} \sum_{i\neq j} X_i P_{ij} \varepsilon_j\| \\
&\qquad = \|(S_n^{-1})'\big(\sum_{i\neq j} S_n^{-1} X_i P_{ij} X_j' (S_n^{-1})'\big)^{-1} \sum_{i\neq j} S_n^{-1} X_i P_{ij} \varepsilon_j\| \\
&\qquad \leq C \frac{1}{\sqrt{r_n}} \|\sum_{i\neq j} S_n^{-1} X_i P_{ij} \varepsilon_j\|
\end{aligned}$$

since the largest eigenvalue of S_n^{-1} is bounded by some constant times $r_n^{-1/2}$. Let e_a be a vector that has 1 in the a-th position and zero elsewhere. We may apply Lemma C.1, with $W_i = e_a' S_n^{-1} X_i$ and $Y_i = \varepsilon_i/\sqrt{r_n}$. We have $E[W_i|\mathcal{Z}] = e_a' z_i/\sqrt{n}$, $E[Y_i|\mathcal{Z}] = 0$, $Var(W_i|\mathcal{Z}) \leq C/r_n$ and $Var(Y_i|\mathcal{Z}) \leq C/r_n$. This implies

$$E\Big[\|\sum_{i\neq j} e_a' S_n^{-1} X_i P_{ij} \varepsilon_j/\sqrt{r_n}\|^2 \big|\mathcal{Z}\Big] \leq C\big(\frac{K_n}{r_n^2} + \frac{1}{r_n}\frac{1}{n}\sum_i (e_a' z_i)^2\big)$$

and so, applying the same result to each element of the vector in turn,

$$\begin{aligned}
E\Big[\|\sum_{i\neq j} S_n^{-1} X_i P_{ij} \varepsilon_j/\sqrt{r_n}\|^2 \big|\mathcal{Z}\Big] &\leq C\big(\frac{J_n K_n}{r_n^2} + \frac{1}{r_n}\frac{1}{n}\sum_i z_i' z_i\big) \\
&\to 0
\end{aligned}$$

A conditional Markov inequality and dominated convergence then imply that $\frac{1}{\sqrt{r_n}}\|\sum_{i\neq j} S_n^{-1} X_i P_{ij} \varepsilon_j\| \to 0$, which proves the consistency result.

Consistency of HLIM and HFUL

We follow Hausman et al. (2009) in proving consistency of the HLIM estimator directly. The proof follows theirs, with application of the earlier bounds.

The sample objective function for the HLIM estimator is

$$|\hat{Q}(\beta_0)| = \frac{|\frac{1}{r_n}\sum_i \sum_j P_{ij}\varepsilon_i\varepsilon_j|}{\frac{1}{n}\sum_i \varepsilon_i^2} = O_p(\frac{\sqrt{K_n}}{r_n}) \to 0$$

by Lemma C.1. The HLIM parameter is given by the sample minimizer, and so $|\tilde{\kappa}_n| = |\hat{Q}(\hat{\beta})| \leq |\hat{Q}(\beta_0)|$, and hence the same rate holds for $\tilde{\kappa}_n$.

Next, note that

$$\begin{aligned}\frac{1}{r_n}\sum_i \sum_{j\neq i} P_{ij}\widehat{\varepsilon}_i\widehat{\varepsilon}_j &= \frac{1}{r_n}\sum_i \sum_{j\neq i} P_{ij}(\varepsilon_i + (\beta_0 - \widehat{\beta})'X_i)(\varepsilon_j + (\beta_0 - \widehat{\beta})'X_j) \\ &= \begin{pmatrix}1 & (\beta_0 - \widehat{\beta})'\end{pmatrix} \frac{1}{r_n}\sum_i \sum_{j\neq i} P_{ij} \begin{pmatrix}\varepsilon_i\varepsilon_j & \varepsilon_i X_j' \\ \varepsilon_j X_i & X_i X_j'\end{pmatrix} \begin{pmatrix}1 \\ (\beta_0 - \widehat{\beta})\end{pmatrix}\end{aligned}$$

Focussing on the center matrix, first note that application of Lemma C.1 gives

$$E\big(\frac{1}{r_n}\sum_i \sum_{j\neq i} P_{ij}\varepsilon_i\varepsilon_j\big)^2 = O_p(K_n/r_n^2) \to 0$$

as well as

$$E\|\frac{1}{\sqrt{r_n}}\sum_i \sum_{j\neq i} P_{ij}S_n^{-1}X_i\varepsilon_j\|^2 \leq O_p(J_n/r_n) + O_p(J_nK_n/r_n^2) \to 0$$

Also, be Lemma C.4 we have

$$\|S_n^{-1}\sum_i \sum_{j\neq i} P_{ij}X_iX_j'(S_n')^{-1} - \frac{1}{n}\sum_i (1 - P_{ii})Z_i'\Pi_0\Pi_0'Z_i\| \to 0$$

This then implies that, for $\bar{S}_n^{-1} = diag(r_n^{1/2}, S_n)$

$$\|\bar{S}_n^{-1}\sum_i \sum_{j\neq i} P_{ij} \begin{pmatrix}\varepsilon_i\varepsilon_j & \varepsilon_i X_j' \\ \varepsilon_j X_i & X_i X_j'\end{pmatrix} (\bar{S}_n')^{-1} - diag(0, H_n)\| \to 0$$

for $H_n = \frac{1}{n}\sum_i (1 - P_{ii})Z_i'\Pi_0\Pi_0'Z_i$, and hence w.p.a.1 we have

$$\bar{S}_n^{-1}\sum_i \sum_{j\neq i} P_{ij} \begin{pmatrix}\varepsilon_i\varepsilon_j & \varepsilon_i X_j' \\ \varepsilon_j X_i & X_i X_j'\end{pmatrix} (\bar{S}_n')^{-1} \geq c \times diag(0, I_J)$$

This then gives

$$\frac{1}{r_n}\sum_i\sum_{j\neq i} P_{ij}\widehat{\varepsilon}_i\widehat{\varepsilon}_j \geq C\begin{pmatrix}1 & (\beta_0-\widehat{\beta})'\end{pmatrix}\frac{1}{r_n}\bar{S}_n diag(0,I_J)\bar{S}_n'\begin{pmatrix}1\\(\beta_0-\widehat{\beta})\end{pmatrix}$$
$$= C\frac{1}{r_n}\|(\beta_0-\widehat{\beta})'S_n\|^2$$

Also note that, bounds on the moments of y_i and X_i imply that $|\frac{1}{n}\sum_i \widehat{\varepsilon}_i^2| \leq C(1+\|\widehat{\beta}\|^2)$. Applying the above bound on the HLIM parameter, we find

$$\frac{\frac{1}{r_n}\|(\widehat{\beta}-\beta_0)'S_n\|^2}{1+\|\widehat{\beta}\|^2} \leq C\frac{\frac{1}{r_n}\sum_i\sum_{j\neq i}P_{ij}\widehat{\varepsilon}_i\widehat{\varepsilon}_j}{\frac{1}{n}\sum_i\widehat{\varepsilon}_i^2}$$
$$= C\widehat{Q}(\widehat{\beta}) \leq C\widehat{Q}(\beta_0)\to 0$$

Lemma A0 in Hansen et al. (2008) (inspection of the lemma shows that it is not dependent on the dimension of the parameter vector) then gives consistency of the HLIM parameter estimates.

For the HFUL estimator, note that

$$\begin{aligned}\widehat{\kappa}_n &= \frac{\tilde{\kappa}_n-(1-\tilde{\kappa}_n)C/n}{1-(1-\tilde{\kappa}_n)C/n}\\ &= \tilde{\kappa}_n - \frac{(1-\tilde{\kappa}_n)^2C/n}{1-(1-\tilde{\kappa}_n)C/n}\\ &= \tilde{\kappa}_n + O_p(n^{-1})\end{aligned}$$

For the denominator term, the HFUL differs from HLIM by the term

$$(\widehat{\kappa}_n-\tilde{\kappa}_n)\|S_n^{-1}\sum_i X_iX_i'(S_n')^{-1}\| \leq O_p(n^{-1})\frac{n}{r_n}\|\frac{1}{n}\sum_i X_iX_i'\|\to 0$$

since$\|\frac{1}{n}\sum_i X_iX_i'\| \leq C$ by Assumption XX. Similarly, the numerator terms of the two estimators differ by

$$(\widehat{\kappa}_n-\tilde{\kappa}_n)\|S_n^{-1}\sum_i X_i\varepsilon_i(S_n')^{-1}\| \leq O_p(n^{-1})\frac{n}{r_n}\|\frac{1}{n}\sum_i X_i\varepsilon_i\|$$
$$= O_p(\sqrt{J_n}/r_n)\to 0$$

Denoting A_{FUL} and A_{LIM} as the numerator terms for both estimators, and B_{FUL}, B_{LIM} the denominator terms, we can write

$$\begin{aligned}\|\widehat{\beta}_{HFUL} - \widehat{\beta}_{HLIM}\| \leq \|B_{FUL} - B_{LIM}\|\|A_{FUL} - A_{LIM}\| + \|B_{LIM}\|\|A_{FUL} - A_{LIM}\| \\ + \|B_{FUL} - B_{LIM}\|\|A_{LIM}\|\end{aligned}$$

We have shown both $\|B_{FUL} - B_{LIM}\| \to 0$ and $\|A_{FUL} - A_{LIM}\|$. Also, the HLIM denominator converges to the same quantity as the JIVE estimator since

$$\tilde{\kappa}_n \|S_n^{-1} \sum_i X_i X_i' (S_n')^{-1}\| \leq \tilde{\kappa}_n \frac{n}{r_n} \|\frac{1}{n} \sum_i X_i X_i'\| \to 0$$

from Lemma XX, which gives $\tilde{\kappa}_n = o_p(r_n/n)$. This implies that w.p.a.1 we will have $c \leq \|B_{LIM}\| \leq C$, as we did for the JIVE estimator. Consistency of the HLIM estimator then implies $\|A_{LIM}\| \to 0$. Hence we have

$$\|\widehat{\beta}_{HFUL} - \widehat{\beta}_{HLIM}\| \to 0$$

and so HFUL is consistent.

C.3 Asymptotic normality

We begin by stating a CLT result from Chao et al. (2012) Lemma 2, which will be used in the proofs below.

Lemma C.5. *Let the following hold, conditional on $\mathcal{Z}$: (i) P is a symmetric idempotent matrix with rank K, $P_{ii} \leq C < 1$; $(W_{1n}, U_1, \varepsilon_1), \ldots, (W_{nn}, U_n, \varepsilon_n)$ are independent and $D_n = \sum_i E[W_{in} W_{in}' | \mathcal{Z}]$ is bounded for n sufficiently large; $E[W_{in}|\mathcal{Z}] = 0$, $E[U_i|\mathcal{Z}] = 0$, $E[\varepsilon_i|\mathcal{Z}] = 0$ and there exists a constant C such that $E[\|U_i\|^4|\mathcal{Z}] \leq C$, $E[\varepsilon_i^4|\mathcal{Z}] \leq C$; and $\sum_i E\big[\|W_{in}\|^4\big|\mathcal{Z}\big] \to 0$. Then, for $\Xi_n = \bar{D}_n + \bar{\Sigma}_n$, with $\bar{\Sigma}_n = \frac{1}{K_n} \sum_{i \neq j} P_{ij}^2 \big(E[U_i U_i'|\mathcal{Z}] E[\varepsilon_j^2|\mathcal{Z}] + E[U_i \varepsilon_i|\mathcal{Z}] E[U_i' \varepsilon_i|\mathcal{Z}]\big)$ it follows that*

$$Y_n = \Xi_n^{-1/2} \big(\sum_{i=1}^n W_{in} + \sum_{i \neq j} U_i P_{ij} \varepsilon_j / \sqrt{K_n} \big) \Rightarrow N(0,1)$$

Jackknife asymptotic normality

The object of interest for inference is

$$\theta_n = \alpha_n' \beta_n$$

where we assume $\|\alpha_n\| \leq C < \infty$ for all n.

We begin by considering $c_n' \sum_{i \neq j} S_n^{-1} X_i P_{ij} \varepsilon_j$, for any sequence of vectors c_n such that $\|c_n\| \leq C$. First decompose the numerator term

$$\begin{aligned}
\sum_{i \neq j} S_n^{-1} X_i P_{ij} \varepsilon_j &= \sum_{i \neq j} S_n^{-1} \Upsilon_i P_{ij} \varepsilon_j + \sum_{i \neq j} S_n^{-1} V_i P_{ij} \varepsilon_j \\
&= \frac{1}{\sqrt{n}} \sum_{i \neq j} z_i P_{ij} \varepsilon_j + \sum_{i \neq j} S_n^{-1} V_i P_{ij} \varepsilon_j \\
&= \frac{1}{\sqrt{n}} \sum_{i,j} z_i P_{ij} \varepsilon_j - \frac{1}{\sqrt{n}} \sum_{i} z_i P_{ii} \varepsilon_i + \sum_{i \neq j} S_n^{-1} V_i P_{ij} \varepsilon_j \\
&= \frac{1}{\sqrt{n}} \sum_{i} \Pi' Z_i \varepsilon_i - \frac{1}{\sqrt{n}} \sum_{i} z_i P_{ii} \varepsilon_i + \sum_{i \neq j} S_n^{-1} V_i P_{ij} \varepsilon_j \\
&= \frac{1}{\sqrt{n}} \sum_{i} (1 - P_{ii}) z_i \varepsilon_i + \sum_{i \neq j} S_n^{-1} V_i P_{ij} \varepsilon_j \qquad (9) \\
&\quad - \frac{1}{\sqrt{n}} \sum_{i} \big(z_i - \Pi' Z_i\big) \varepsilon_i
\end{aligned}$$

Note that the first and second terms are uncorrelated, that is

$$E\Big[\Big(\sum_i S_n^{-1}(1 - P_{ii}) \Pi' Z_i \varepsilon_i\Big)\Big(\sum_{j \neq k} S_n^{-1} V_j P_{jk} \varepsilon_k\Big)' \Big| \mathcal{Z}\Big] = 0$$

The third term is due to the approximation of the reduced form z by the observed instruments Z and is negligible asymptotically by Assumption 3.3. Next,

$$\begin{aligned}
E\Big[\|\frac{1}{\sqrt{n}} \sum_i \big(z_i - \Pi' Z_i\big) \varepsilon_i\|^2 \Big| \mathcal{Z}\Big] &\leq \frac{1}{n} \sum_i \|\big(z_i - \Pi' Z_i\big)\|^2 E\Big[\varepsilon_i^2 \Big| \mathcal{Z}\Big] \\
&\leq C \frac{1}{n} \sum_i \|\big(z_i - \Pi' Z_i\big)\|^2 \to 0
\end{aligned}$$

so that a conditional Markov inequality and the dominated convergence theorem will give $\|\frac{1}{\sqrt{n}} \sum_i \big(z_i - \Pi' Z_i\big) \varepsilon_i\| \to 0$.

We now confirm that the remaining quantities in (9) satisfy the conditions of Chao et al (2009) Lemma 2.

Condition (i) holds by P being a projection matrix. For (ii), we have $W_{in} = c_n'(1 - P_{ii})z_i\varepsilon_i/\sqrt{n}$, $U_i = \sqrt{K_n}a'S_n^{-1}V_i$, and $\varepsilon_i = \varepsilon_i$. By assumption these are independent over i and

$$\begin{aligned}\bar{D}_n = \frac{1}{n}\sum_i E\big[(1-P_{ii})^2 c_n' z_i z_i' c_n \varepsilon_i^2 \big| \mathcal{Z}\big] &= \frac{1}{n}\sum_i (1-P_{ii})^2 c_n' z_i z_i' c_n E\big[\varepsilon_i^2 \big| \mathcal{Z}\big] \\ &\leq C\frac{1}{n}\sum_i (1-P_{ii})^2 c_n' z_i z_i' c_n\end{aligned}$$

is bounded by the boundedness of $\lambda_{max}(\sum_i z_i z_i'/n)$ and $\|c_n\|$. The mean zero conditions of (iii) all hold in this setting also, while $E[\varepsilon_i^4|\mathcal{Z}] \leq C$ follows from Assumption 3.7. The same assumption also gives

$$\begin{aligned}E[\|U_i\|^4|\mathcal{Z}] &= K_n^2 E[\|c_n' S_n^{-4} V_i\|^4|\mathcal{Z}] \\ &\leq \frac{K_n^2}{r_n^2} E\Big[\Big(\sum_{j=1}^{J_n} c_{nj} V_{j,i}\Big)^4 \Big| \mathcal{Z}\Big] \\ &= \frac{K_n^2}{r_n^2}\sum_{j=1}^{J_n} c_{n,j}^4 E\big[V_{j,i}^4\big|\mathcal{Z}\big] + \sum_{j\neq k=1}^{J_n} c_{n,j}^2 c_{n,k}^2 E\big[V_{j,i}^2\big|\mathcal{Z}\big] E\big[V_{k,i}^2\big|\mathcal{Z}\big] \\ &\leq C\frac{K_n^2}{r_n^2}\Big(\sum_{j=1}^{J_n} c_{nj}^4 + \sum_{j\neq k=1}^{J_n} c_{nj}^2 c_{nk}^2\Big) \\ &= C\frac{K_n^2}{r_n^2}\|c_n\|^4\end{aligned}$$

which is bounded as long as $K_n/r_n \to c < \infty$, since $\|c_n\|$ is bounded. Finally, from Assumption 3.7 we have

$$\begin{aligned}\sum_i E\big[\|W_{in}\|^4\big|\mathcal{Z}\big] &= \frac{1}{n^2}\sum_i E\big[\|c_n'(1-P_{ii})z_i\varepsilon_i\|^4\big|\mathcal{Z}\big] \\ &\leq \frac{1}{n^2}\sum_i \|c_n'(1-P_{ii})z_i\|^4 E\big[\varepsilon_i^4\big|\mathcal{Z}\big] \\ &\leq C\frac{1}{n^2}\sum_i \|c_n'(1-P_{ii})z_i\|^4\end{aligned}$$

$$
\begin{aligned}
&= C\frac{1}{n^2}\sum_i (1-P_{ii})^4 \|c_n' z_i\|^4 \\
&\leq C\|c_n\|^4 \frac{1}{n^2}\sum_i \|z_i\|^4 \\
&\to 0
\end{aligned}
$$

Now let

$$
\begin{aligned}
\bar{\Sigma}_n &\equiv \frac{1}{K_n}\sum_{i\neq j} P_{ij}^2 \big(E[U_i U_i'|\mathcal{Z}]E[\varepsilon_j^2|\mathcal{Z}] + E[U_i\varepsilon_i|\mathcal{Z}]E[U_i'\varepsilon_i|\mathcal{Z}]\big) \\
&\equiv c_n' \sum_{i\neq j} P_{ij}^2 S_n^{-1}\big(E[V_i V_i'|\mathcal{Z}]E[\varepsilon_j^2|\mathcal{Z}] \\
&\quad + E[V_i\varepsilon_i|\mathcal{Z}]E[V_i'\varepsilon_i|\mathcal{Z}]\big)(S_n^{-1})' c_n \\
\bar{\Xi}_n &\equiv \bar{D}_n + \bar{\Sigma}_n
\end{aligned}
$$

Then, assuming $\bar{\Xi}_n \geq C$, Lemma C.5 implies

$$
\bar{\Xi}_n^{-1/2} c_n' \big(\frac{1}{\sqrt{n}}\sum_{i=1}^n (1-P_{ii})z_i\varepsilon_i + S_n^{-1}\sum_{i\neq j} V_i P_{ij}\varepsilon_j\big) \Rightarrow N(0,1)
$$

and hence by (9)

$$
\bar{\Xi}_n^{-1/2} c_n' \big(\sum_{i\neq j} S_n^{-1} X_i P_{ij}\varepsilon_j\big) \Rightarrow N(0,1)
$$

Now, consider the choice $c_n' = b_n \alpha_n' (S_n^{-1})' H_n^{-1}$, where

$$
\begin{aligned}
H_n &= \frac{1}{n}\sum_i (1-P_{ii}) z_i z_i' \\
b_n &= \|\alpha'(S_n^{-1})'\|^{-1}
\end{aligned}
$$

We have $c \leq \|c_n\| \leq C$ since

$$
\|c_n\| = \|\xi' H^{-1}\|
$$

for ξ a vector with $\|\xi\| = 1$, implying that the norm is bounded by the smallest and largest eigenvalues of H^{-1}, which are by assumption bounded. We must also confirm that $\Xi_n \geq C$ for this choice of c_n. Firstly,

$$\bar{D}_n = c_n'\Big(\frac{1}{n}\sum_i (1-P_{ii})^2 z_i z_i' E\big[\varepsilon_i^2\big|\mathcal{Z}\big]\Big)c_n$$

and since $c \leq \|c_n\| \leq C$ we have $\lambda_{min} \leq D_n \leq \lambda_{max}$ where $\lambda_{min}, \lambda_{max}$ are the minimum and maximum eigenvalues of $\frac{1}{n}\sum_i (1-P_{ii})^2 z_i z_i' E\big[\varepsilon_i^2\big|\mathcal{Z}\big]$ which are bounded above and below by assumption. Next, since $\bar{\Sigma}_n$ is a s.p.d matrix, we have $\bar{D}_n + \bar{\Sigma}_n \geq c$, and

$$\begin{aligned}
\bar{\Sigma}_n &\equiv \sum_{i\neq j} P_{ij}^2 S_n^{-1} c_n'\big(E[V_i V_i'|\mathcal{Z}]E[\varepsilon_j^2|\mathcal{Z}] \\
&\quad + E[V_i\varepsilon_i|\mathcal{Z}]E[V_i'\varepsilon_i|\mathcal{Z}]\big)c_n (S_n^{-1})' \\
&\leq C \sup_i E[\varepsilon_i^2|\mathcal{Z}] \sup_i \lambda_{max}\big(E[V_{ni}V_{ni}'|\mathcal{Z}]\big)\frac{1}{r_n}\sum_{i\neq j} P_{ij}^2 \\
&\leq C\frac{K_n}{r_n} \leq C
\end{aligned}$$

so we also have $\bar{\Xi}_n = \bar{D}_n + \bar{\Sigma}_n \leq C$.

Now, let $A_n = \sum_{i\neq j} S_n^{-1} X_i P_{ij}\varepsilon_j$ and $B_n = \sum_{i\neq j} S_n^{-1} X_i P_{ij} X_j (S_n^{-1})'$, so that $\hat{\beta}_n - \beta_n = (S_n^{-1})' B_n^{-1} A_n$. Also, let $\Xi_n = D_n + \bar{\Sigma}_n$. First note that

$$\begin{aligned}
&b_n \|\alpha_n'(S_n^{-1})'(B_n^{-1} - H_n^{-1})A_n\| \\
&\quad = b_n \|\alpha_n'(S_n^{-1})' H_n^{-1}(H_n B_n^{-1} - I_J)A_n\| \\
&\quad = \|c_n'(H_n B_n^{-1} - I_J)A_n\| \\
&\quad \leq \lambda_{max}\big(H_n B_n^{-1} - I_J\big)\|c_n' A_n\| \\
&\quad = o_p(1) \times O_p(1) \to 0
\end{aligned}$$

since $c_n' A_n = O_p(1)$ and

$$\begin{aligned}
\lambda_{max}\big(H_n B_n^{-1} - I_J\big) &\leq \|H_n B_n^{-1} - I_J\|^2 \\
&\leq \lambda_{max}(H_n)\cdot \|B_n^{-1} - H_n^{-1}\|^2 \to 0
\end{aligned}$$

Finally, note that $\bar{D}_n = c_n' D_n c_n$ and $\bar{\Sigma}_n = c_n'\Sigma_n c_n$ (where D_n and Σ_n are as defined in the paper, i.e. with multiplication by c_n) so that

$$\begin{aligned}
\bar{\Xi}_n &= c_n'(D_n + \Sigma_n)c_n \\
&= b_n^2 \alpha_n'(S_n^{-1})' H_n^{-1}(D_n + \Sigma_n)H_n^{-1} S_n^{-1}\alpha_n
\end{aligned}$$

$$= V_{J,n}$$

Then, using this result we have

$$\begin{aligned} b_n V_{J,n}^{-1/2}(\hat{\theta}_n - \theta_n) &= b_n \bar{\Xi}_n^{-1/2} \alpha_n'(\hat{\beta}_n - \beta_n) \\ &= b_n \bar{\Xi}_n^{-1/2} \alpha_n' (S_n^{-1})' B_n^{-1} A_n \\ &= b_n \bar{\Xi}_n^{-1/2} \alpha_n' (S_n^{-1})' H_n^{-1} A_n \\ &\quad + b_n \bar{\Xi}_n^{-1/2} \alpha_n' (S_n^{-1})' (B_n^{-1} - H_n^{-1}) A_n \\ &= \bar{\Xi}_n^{-1/2} c_n' A_n + o_p(1) \\ &\Rightarrow N(0,1) \end{aligned}$$

HLIM and HFUL asymptotic normality

For HLIM, we have

$$\tilde{\kappa}_n = \frac{\sum_{i \neq j} \varepsilon_i P_{ij} \varepsilon_j}{\sum_i \varepsilon_i^2}$$

The normalized numerator is

$$\begin{aligned} & c_n' S_n^{-1} \sum_{i \neq j} X_i P_{ij} \varepsilon_j - \tilde{\kappa}_n c_n' \sum_i S_n^{-1} X_i \varepsilon_i \\ &= c_n' \sum_i (1 - P_{ii}) z_i \varepsilon_i / \sqrt{n} + c_n' S_n^{-1} \sum_{i \neq j} V_i P_{ij} \varepsilon_j \\ &\quad - \tilde{\kappa}_n \Big(c_n' \sum_i z_i \varepsilon_i / \sqrt{n} + c_n' S_n^{-1} \sum_i V_i \varepsilon_i \Big) \end{aligned}$$

Now, let $\tilde{V}_i = V_i - \delta_n \varepsilon_i$ where $\delta_n = E[V_i \varepsilon_i]/E[\varepsilon_i^2]$. Also let $\hat{\delta} = \sum_i V_i \varepsilon_i / \sum_i \varepsilon_i^2$. Then

$$\begin{aligned} & \tilde{\kappa}_n \Big(c_n' \sum_i z_i \varepsilon_i / \sqrt{n} + c_n' S_n^{-1} \sum_i V_i \varepsilon_i \Big) \\ &= \tilde{\kappa}_n c_n' \sum_i z_i \varepsilon_i / \sqrt{n} + \tilde{\kappa}_n (\sum_i \varepsilon_i^2) c_n' S_n^{-1} \hat{\delta} \\ &= \big(c_n' S_n^{-1} \hat{\delta} \big) \sum_{i \neq j} \varepsilon_i P_{ij} \varepsilon_j + o_p(1) \end{aligned}$$

and substituting into the expression for the numerator gives

$$\begin{aligned}
&c_n' S_n^{-1} \sum_{i \neq j} X_i P_{ij} \varepsilon_j - \tilde{\kappa}_n c_n' \sum_i S_n^{-1} X_i \varepsilon_i \\
&\quad = c_n' \sum_i (1 - P_{ii}) z_i \varepsilon_i / \sqrt{n} + c_n' S_n^{-1} \sum_{i \neq j} \tilde{V}_i P_{ij} \varepsilon_j \\
&\qquad - c_n' S_n^{-1} (\hat{\delta} - \delta_n) \sum_{i \neq j} \varepsilon_i P_{ij} \varepsilon_j + o_p(1) \\
&\quad = c_n' \sum_i (1 - P_{ii}) z_i \varepsilon_i / \sqrt{n} + c_n' S_n^{-1} \sum_{i \neq j} \tilde{V}_i P_{ij} \varepsilon_j + o_p(1)
\end{aligned}$$

since consistency of $\hat{\beta}$ implies consistency for $c_n'(\hat{\delta} - \delta_n)$. Given the above form, we can apply the same steps as in the previous subsection to get the normality result, replacing V_i with $\tilde{V}_i$. For HFUL, we note that $\hat{\kappa}_n = \tilde{\kappa}_n + O_p(n^{-1})$ and so the same result follows from

$$\begin{aligned}
&c_n' S_n^{-1} \sum_{i \neq j} X_i P_{ij} \varepsilon_j - \hat{\kappa}_n c_n' \sum_i S_n^{-1} X_i \varepsilon_i \\
&\quad = c_n' S_n^{-1} \sum_{i \neq j} X_i P_{ij} \varepsilon_j - \tilde{\kappa}_n c_n' \sum_i S_n^{-1} X_i \varepsilon_i + o_p(1)
\end{aligned}$$

C.4 HLIM and HFUL parameter

For HLIM, $\tilde{\kappa}_n$ is equal to the minimized HLIM objective function

$$Q(\hat{\beta}) = \frac{\sum_{i \neq j} P_{ij} (y_i - X_i' \hat{\beta})(y_j - X_j' \hat{\beta})}{\sum_i (y_i - X_i' \hat{\beta})^2}$$

First note that since $\frac{1}{n} \sum_i \varepsilon_i^2 = O_p(1)$ and positive, and $\frac{1}{n} \sum_{i \neq j} P_{ij} \varepsilon_i \varepsilon_j = O_p(\sqrt{K_n}/n)$ by Lemma C.1, we have that $Q(\beta_0) = O_p(\sqrt{K_n}/n)$.

Next, note that

$$\begin{aligned}
\hat{\varepsilon}_i \hat{\varepsilon}_j - \varepsilon_i \varepsilon_j &= (y_i - X_i' \hat{\beta})(y_j - X_j' \hat{\beta}) - (y_i - X_i' \beta_0)(y_j - X_j' \beta_0) \\
&= -(y_i X_j' + y_j X_i')(\hat{\beta} - \beta_0) + (X_i' \hat{\beta})(X_j' \hat{\beta}) - (X_i' \beta_0)(X_j' \beta_0) \\
&= -(\varepsilon_i X_j' + \varepsilon_j X_i')(\hat{\beta} - \beta_0) - (X_i' \beta_0 X_j' + X_j' \beta_0 X_i')(\hat{\beta} - \beta_0) \\
&\quad + (\hat{\beta} - \beta_0)' X_i X_j' (\hat{\beta} - \beta_0) + \beta_0' (X_i X_j' + X_j X_i') \hat{\beta} - 2 (X_i' \beta_0)(X_j' \beta_0) \\
&= (\hat{\beta} - \beta_0)' X_i X_j' (\hat{\beta} - \beta_0) - (\varepsilon_i X_j' + \varepsilon_j X_i')(\hat{\beta} - \beta_0)
\end{aligned}$$

and using this we get that, letting $\alpha_0 = Q(\beta_0)$ and $\hat{\delta} = (\hat{\beta} - \beta_0)$

$$\begin{aligned} Q(\hat{\beta}) - Q(\beta_0) &= \frac{1}{\frac{1}{n}\sum_i \hat{\varepsilon}_i^2} \frac{1}{n} \Big(\sum_{i \neq j} P_{ij} \hat{\varepsilon}_i \hat{\varepsilon}_j - \sum_{i \neq j} P_{ij} \varepsilon_i \varepsilon_j \\ &\quad - Q(\beta_0) \big(\sum_i \hat{\varepsilon}_i^2 - \sum_i \varepsilon_i^2 \big) \Big) \\ &= O_p(1) \frac{1}{n} \Big(\hat{\delta}' \big(\sum_{i \neq j} P_{ij} X_i X_j' - \alpha_0 \sum_i X_i X_i' \big) \hat{\delta} \\ &\quad - 2 \hat{\delta}' \big(\sum_{i \neq j} P_{ij} X_i \varepsilon_j - \alpha_0 \sum_i X_i \varepsilon_i \big) \Big) \end{aligned}$$

Next note that the proof of Theorem 1 implies$\|S_n' \hat{\delta} / \sqrt{r_n}\| \to 0$, so by Lemma C.4,

$$\begin{aligned} \|\hat{\delta}' \sum_{i \neq j} P_{ij} X_i X_j' \hat{\delta}\| &= r_n \|(\hat{\delta}' S_n / \sqrt{r_n}) S_n^{-1} \sum_{i \neq j} P_{ij} X_i X_j' (S_n^{-1})' (S_n' \hat{\delta} / \sqrt{r_n})\| \\ &\leq r_n \|S_n' \hat{\delta} / \sqrt{r_n}\|^2 \|S_n^{-1} \sum_{i \neq j} P_{ij} X_i X_j' (S_n^{-1})'\| \\ &= r_n \|S_n' \hat{\delta} / \sqrt{r_n}\|^2 \|\frac{1}{n} \sum_i (1 - P_{ii}) z_i z_i'\| + o_p(r_n) \end{aligned}$$

by Assumption 2. From Lemma C.1 we also see that

$$\begin{aligned} \|\hat{\delta}' \sum_{i \neq j} P_{ij} X_i \varepsilon_j\| &\leq \sqrt{r_n} \|S_n' \hat{\delta} / \sqrt{r_n}\| \cdot \|S_n^{-1} \sum_{i \neq j} P_{ij} X_i \varepsilon_j\| \\ &= o_p(K_n^{1/2}) \end{aligned}$$

Similarly $\hat{\delta}'(\frac{1}{n}\sum_i X_i X_i')\hat{\delta} = o_p(1)$ and $\hat{\delta}'(\frac{1}{n}\sum_i X_i \varepsilon_i) = o_p(1)$. Combining these results we find

$$\begin{aligned} Q(\hat{\beta}) - Q(\beta_0) &= \frac{1}{n} \big(o_p(r_n) + O_p(\frac{\sqrt{K_n}}{n}) o_p(1) \\ &\quad + o_p(K_n^{1/2}) + O_p(\frac{\sqrt{K_n}}{n}) o_p(1) \big) \\ &= o_p(\frac{r_n}{n}) \end{aligned}$$

since $\sqrt{K_n}/r_n \to 0$. The same result holds for HFUL, since $\widehat{\kappa}_n = \tilde{\kappa}_n + O_p(n^{-1})$.

C.5 Consistency of variance of jackknife estimator

Following the discussion in Section 2 of Chao et al (2009) (CHSNW), note that

$$\begin{aligned}\Xi_n &= E\Big[S_n^{-1}\Big(\sum_{i\neq j} X_i P_{ij}\varepsilon_j\Big)\Big(\sum_{i\neq j} X_i P_{ij}\varepsilon_j\Big)'(S_n^{-1})'\Big|\mathcal{Z}\Big] \\ &= \sum_{i\neq j}\sum_{k\neq l} P_{ij}P_{kl}S_n^{-1}E\big[X_iX_k'\varepsilon_j\varepsilon_l\big|\mathcal{Z}\big](S_n^{-1})' \\ &= \sum_{i,j}\sum_{k\neq\{i,j\}} P_{ik}P_{jk}S_n^{-1}E\big[X_iX_k'\big|\mathcal{Z}\big]E\big[\varepsilon_j^2\big|\mathcal{Z}\big](S_n^{-1})' \\ &\quad + \sum_{i\neq j} P_{ij}^2 S_n^{-1}E\big[X_i\varepsilon_i\big|\mathcal{Z}\big]E\big[X_j'\varepsilon_j\big|\mathcal{Z}\big](S_n^{-1})'\end{aligned}$$

Given this, we first show, for an arbitrary sequence of unit norm vectors c_n, $\|c_n\| = 1$, that $c'\hat{\Xi}_n c - c'\Xi_n c \to 0$, for the estimator

$$\begin{aligned}\hat{\Xi}_n &= S_n^{-1}\Big(\sum_{i,j}\sum_{k\neq\{i,j\}} P_{ik}P_{jk}X_iX_j'\hat{\varepsilon}_k^2 + \sum_{i\neq j} P_{ij}^2 X_i\hat{\varepsilon}_i X_j'\hat{\varepsilon}_j\Big)(S_n^{-1})' \\ &= S_n^{-1}\Big(\sum_{i\neq j\neq k} P_{ik}P_{jk}X_iX_j'\hat{\varepsilon}_k^2 + \sum_{i\neq j} P_{ij}^2\big(X_i\hat{\varepsilon}_i X_j'\hat{\varepsilon}_j + X_iX_i'\hat{\varepsilon}_j^2\big)\Big)(S_n^{-1})' \\ &= \hat{\Sigma}_1 + \hat{\Sigma}_2\end{aligned}$$

Decomposing the first part of this estimator gives

$$\begin{aligned}\hat{\Sigma}_1 &= \sum_{i\neq j\neq k} P_{ij}P_{jk}S_n^{-1}X_iX_k'(S_n^{-1})'\hat{\varepsilon}_j^2 \\ &= \frac{1}{n}\sum_{i\neq j\neq k} P_{ik}P_{jk}z_iz_j'\hat{\varepsilon}_k^2 + \\ &\quad + \frac{1}{\sqrt{n}}\sum_{i\neq j\neq k} P_{ik}P_{jk}z_iV_j'\hat{\varepsilon}_k^2(S_n^{-1})' + \frac{1}{\sqrt{n}}S_n^{-1}\sum_{i\neq j\neq k} P_{ik}P_{jk}V_iz_j'\hat{\varepsilon}_k^2 \\ &\quad + S_n^{-1}\sum_{i\neq j\neq k} P_{ik}P_{jk}V_iV_j'\hat{\varepsilon}_k^2(S_n^{-1})'\end{aligned}$$

while the second part is

$$\hat{\Sigma}_2 = \sum_{i\neq j} P_{ij}^2 S_n^{-1}\big(X_i\hat{\varepsilon}_i X_j'\hat{\varepsilon}_j + X_iX_i'\hat{\varepsilon}_j^2\big)(S_n^{-1})'$$

$$= \frac{1}{n}\sum_{i\neq j} P_{ij}^2\big(z_i\hat{\varepsilon}_i z_j'\hat{\varepsilon}_j + z_i z_i'\hat{\varepsilon}_j^2\big) + \frac{1}{\sqrt{n}}\sum_{i\neq j} P_{ij}^2\big(z_i\hat{\varepsilon}_i V_j'\hat{\varepsilon}_j + z_i V_i'\hat{\varepsilon}_j^2\big)(S_n^{-1})'$$
$$+ S_n^{-1}\frac{1}{\sqrt{n}}\sum_{i\neq j} P_{ij}^2\big(V_i\hat{\varepsilon}_i z_j'\hat{\varepsilon}_j + V_i z_i'\hat{\varepsilon}_j^2\big)$$
$$+ \sum_{i\neq j} P_{ij}^2 S_n^{-1}\big(V_i\hat{\varepsilon}_i V_j'\hat{\varepsilon}_j + V_i V_i'\hat{\varepsilon}_j^2\big)(S_n^{-1})'$$

Lemma A7 in Chao et al. (2012) shows that $\hat{\Sigma}_1 - \dot{\Sigma}_1 = o_p(1)$ and $\hat{\Sigma}_2 - \dot{\Sigma}_2 = o_p(K_n/r_n)$ where $\dot{\Sigma}_1$ and $\dot{\Sigma}_2$ are versions of the respective matrices that replace $\hat{\varepsilon}$ with ε. We will repeat the result for $c'(\hat{\Sigma}_1 - \dot{\Sigma}_1)c$ and $c'(\hat{\Sigma}_2 - \dot{\Sigma}_2)c$

Proof of $c'(\hat{\Sigma}_1 - \dot{\Sigma}_1)c = o_p(1)$

We start by stating an alternate version of Lemma A4 in Chao et al. (2012), which is useful in our setup. It is a special case of the original lemma, in which $W_i = Y_i$, which allows for slightly weaker conditions that are needed.

Recall that $\bar{w}_i = E[W_i|\mathcal{Z}]$, $\tilde{w}_i = W_i - \bar{w}_i$, $\bar{\mu}_W = \max_{1\leq i\leq n}|\bar{w}_i|$, and $\bar{\sigma}_w^2 = \max_{1\leq i\leq n} Var(W_i|\mathcal{Z})$.

Lemma C.6. *[Lemma A4 special case] Suppose that conditional on $\mathcal{Z}$, $(W_1,\eta_1),\ldots,(W_n,\eta_n)$ are independent. Suppose also that $\sqrt{K_n}/r_n \to 0$ as $n\to\infty$, and that there exists a positive constant C such that for n sufficiently large: (i) $\sum_i \bar{w}_i^2 \leq C$ and $E\big[\sum_i \bar{w}_i^4\big] \to 0$, (ii) $\bar{\sigma}_W^2 \leq C/r_n$, and (iii) $E_{\mathcal{Z}}[\bar{\mu}_\eta^2] \leq C$, and $\bar{\sigma}_\eta^2 \leq C$. Then*

$$A_n = E\Big[\sum_{i\neq j\neq k} W_i P_{ik}\eta_k P_{kj} W_j \Big| \mathcal{Z}\Big] = O_p(1)$$
$$\sum_{i\neq j\neq k} W_i P_{ik}\eta_k P_{kj} W_j - A_n \to 0$$

Proof. The proof follows the proof of Lemma A4 and so we only focus on the differences. To show $A_n = O_p(1)$ we follow the steps on pages 48 and 49, except that we may write

$$E\big[\big|\sum_{i,k}\bar{w}_i\bar{y}_i P_{ik}^2\bar{\eta}_k\big|\big] \leq E\big[\sum_{i,k}\big|\bar{w}_i^2 P_{ik}^2\bar{\eta}_k\big|\big]$$
$$\leq E\big[\bar{\mu}_\eta\sum_{i,k}\bar{w}_i^2 P_{ik}^2\big]$$
$$= E\big[\bar{\mu}_\eta\sum_{i}\bar{w}_i^2 P_{ii}\big]$$

$$\leq CE[\bar{\mu}_\eta] \leq C$$

and similarly

$$E\big[\big|\sum_i \bar{w}_i \bar{y}_i P_{ii}^2 \bar{\eta}_k\big|\big] \leq E\big[\bar{\mu}_\eta \sum_i \big|\bar{w}_i^2 P_{ii}^2\big|\big]$$
$$\leq C$$

The other parts of $E[|A_n|]$ are identical to the original proof, and a Markov inequality then gives $A_n = O_p(1)$ as required.

Next we decompose

$$\sum_{i\neq j\neq k} W_i P_{ik} \eta_k P_{kj} W_j - A_n = \sum_{r=1}^{7} \hat{\psi}_r$$

as at the top of page 50. $E[\hat{\psi}_r^2] \to 0$ for $r = \{1,2,3,4,5,7\}$ follows from the assumptions of this lemma identically to the original proof. For $\hat{\psi}_6$ we may write

$$\begin{aligned} E[\hat{\psi}_6^2] &= E\Big[\Big(\sum_{i\neq j\neq k} \bar{w}_i P_{ik} \tilde{\eta}_k P_{kj} \bar{w}_j\Big)^2\Big] \\ &= E\Big[\sum_k E\big[\bar{\eta}_k^2 \big| \mathcal{Z}\big] \Big(\sum_{i\neq k} \sum_{j\neq\{i,k\}} \bar{w}_i P_{ik} P_{kj} \bar{w}_j\Big)^2\Big] \\ &\leq CE\Big[\sum_k \Big(\sum_{i\neq k} \sum_{j\neq\{i,k\}} \bar{w}_i P_{ik} P_{kj} \bar{w}_j\Big)^2\Big] \end{aligned}$$

Next, note that since $\widehat{\bar{w}}_k = \sum_i \bar{w}_i P_{ik}$ is the k-th fitted value from regression of $\bar{w}$ on the columns of P, we have $\sum_k (\sum_i \bar{w}_i P_{ik})^2 \leq \sum_k \bar{w}_k^2$. Similarly, $\widehat{\bar{w}^2}_k = \sum_i \bar{w}_i^2 P_{ik}$ is the fitted value from a regression of $\bar{w}_i^2$ on P so (using $P_{ik} \leq 1$)

$$\begin{aligned} \sum_k \Big(\sum_i \bar{w}_i^2 P_{ik}^2\Big)^2 &\leq \sum_k \Big(\sum_i \bar{w}_i^2 P_{ik}\Big)^2 \\ &= \sum_k (\widehat{\bar{w}^2}_k)^2 \\ &\leq \sum_k \bar{w}_i^4 \end{aligned}$$

Now, using $(\sum_{r=1}^{5} A_r)^2 \leq 5\sum_{r=1}^{5} A_r^2$ and $\sum_k \bar{w}_k^4 \leq (\sum_k \bar{w}_k^2)^2$ we have

$$\begin{aligned}
\sum_k \Big(\sum_{i\neq k} \sum_{j\neq\{i,k\}} \bar{w}_i P_{ik} P_{kj} \bar{w}_j \Big)^2 &= \sum_k \Big(\sum_{i,j} \bar{w}_i P_{ik} P_{kj} \bar{w}_j - \sum_i \bar{w}_i^2 P_{ik}^2 \\
&\quad - \sum_j \bar{w}_k P_{kk} P_{kj} \bar{w}_j - \sum_i \bar{w}_i P_{ik} P_{kk} \bar{w}_k \\
&\quad + 2\bar{w}_k^2 P_{kk}^2 \Big)^2 \\
&\leq 5 \sum_k \Big\{ \Big(\sum_{i,j} \bar{w}_i P_{ik} P_{kj} \bar{w}_j \Big)^2 + \Big(\sum_i \bar{w}_i^2 P_{ik}^2 \Big)^2 \\
&\quad \Big(\sum_j \bar{w}_k P_{kk} P_{kj} \bar{w}_j \Big)^2 + \Big(\sum_i \bar{w}_i P_{ik} P_{kk} \bar{w}_k \Big)^2 \\
&\quad + \Big(2\bar{w}_k^2 P_{kk}^2 \Big)^2 \Big\} \\
&\leq 5 \Big(\sum_k \bar{w}_k^4 + \sum_k \bar{w}_k^4 + 2\sum_k P_{kk}^2 \bar{w}_k^4 + 2\sum_k P_{kk}^4 \bar{w}_k^4 \Big) \\
&\leq C \big(\sum_i \bar{w}_i^4 \big) \to 0
\end{aligned}$$

So we have

$$\begin{aligned}
E[\hat{\psi}_6^2] &\leq CE\Big[\sum_k \Big(\sum_{i\neq k} \sum_{j\neq\{i,k\}} \bar{w}_i P_{ik} P_{kj} \bar{w}_j \Big)^2 \Big] \\
&\leq CE\big[\sum_i \bar{w}_i^4 \big] \\
&= o_p(1)
\end{aligned}$$

Combined with the other results this gives

$$\sum_{i\neq j\neq k} W_i P_{ik} \eta_k P_{kj} W_j - A_n \to 0$$

□

We can now use Lemma C.6 to show $c_n'(\hat{\Sigma}_1 - \dot{\Sigma}_1)c_n \to 0$. Firstly, since

$$\begin{aligned}
\hat{\varepsilon}_i^2 - \varepsilon_i^2 &= -2X_i'(\hat{\beta}_n - \beta_n)\varepsilon_i + \big(X_i'(\hat{\beta}_n - \beta_n) \big)^2 \\
&= -2\Delta_i \varepsilon_i + \Delta_i^2
\end{aligned}$$

we have

$$\begin{aligned}|\hat{\Sigma}_1 - \dot{\Sigma}_1| &= |S_n^{-1} \sum_{i\neq j\neq k} P_{ik}P_{jk}X_iX_j'(\hat{\varepsilon}_k^2 - \varepsilon_k^2)(S_n^{-1})| \\ &\leq 2\sup_i \Delta_i |S_n^{-1} \sum_{i\neq j\neq k} P_{ik}P_{jk}X_iX_j'\varepsilon_k(S_n^{-1})| \\ &\quad + \sup_i \Delta_i^2 |S_n^{-1} \sum_{i\neq j\neq k} P_{ik}P_{jk}X_iX_j'(S_n^{-1})|\end{aligned}$$

We apply the lemma in two parts, taking $W_i = c_n'S_n^{-1}X_i$ in both instances, while $\eta_{1i} = \varepsilon_i$ and $\eta_{2i} = 1$.

The required conditions for Lemma C.6 are:

(i) $\sum_i \bar{w}_i^2 \leq C$ *and* $E\big[\sum_i \bar{w}_i^4\big] \to 0$

Since $W_i = c_n'S_n^{-1}X_i$ we have $\bar{w}_i = c_n'z_i/\sqrt{n}$ so that

$$\sum_i \bar{w}_i^2 = \frac{1}{n}\sum_i c_n'z_iz_i'c_n \leq C$$

by Assumption 3.2 and $\|c_n\| = 1$. Also

$$\begin{aligned}E\big[\sum_i \bar{w}_i^4\big] &= E\big[\frac{1}{n^2}\sum_i (c_n'z_i)^4\big] \\ &\leq \|c_n\|^4 \frac{1}{n^2}\sum_i E\big[\|z_i\|^4\big] \\ &\to 0\end{aligned}$$

by Assumption 3.7.

(ii) $\bar{\sigma}_W^2 \leq C/r_n$

By Assumption 3.4

$$\begin{aligned}Var\big(c_n'S_n^{-1}X_i\big|\mathcal{Z}\big) &= c_n'S_n^{-1}E\big[V_iV_i'\big|\mathcal{Z}\big](S_n^{-1})'c_n \\ &\leq Cc_n'S_n^{-1}(S_n^{-1})'c_n \\ &\leq C/r_n\end{aligned}$$

(iii) $E_{\mathcal{Z}}[\bar{\mu}_\eta^2] \leq C$, *and* $\bar{\sigma}_\eta^2 \leq C$.

For $\eta_{1i} = \varepsilon_i$, we have $\bar{\sigma}_\eta^2 = E[\varepsilon_i^2|\mathcal{Z}] \leq C$ and $\bar{\mu}_\eta = E[\varepsilon_i|\mathcal{Z}] = 0$ by Assumption 3.4. For $\eta_{2i} = 1$, both conditions clearly hold as well.

Then, by Lemma C.6 we can conclude

$$c_n' S_n^{-1} \sum_{i \neq j \neq k} P_{ik} P_{jk} X_i X_j' \varepsilon_i (S_n^{-1}) c_n = O_p(1)$$
$$c_n' S_n^{-1} \sum_{i \neq j \neq k} P_{ik} P_{jk} X_i X_j' (S_n^{-1}) c_n = O_p(1)$$

Now,

$$\begin{aligned} \sup_i |\Delta_i| &= \sup_i |X_i'(\hat{\beta}_n - \beta_n)| \\ &\leq \sup_i \|X_i\| \cdot \|\hat{\beta}_n - \beta_n\| \\ &\leq \zeta_0(X) \|\hat{\beta}_n - \beta_n\| \\ &\to 0 \end{aligned}$$

where the final line results from

$$E\Big[\|\zeta_0(X) \sum_{i \neq j} S_n^{-1} X_i P_{ij} \varepsilon_j / \sqrt{r_n}\|^2 \big| \mathcal{Z}\Big] \leq C \zeta_0^2(X) \big(\frac{J_n K_n}{r_n^2} + \frac{1}{r_n} \frac{1}{n} \sum_i z_i' z_i \big)$$
$$\to 0$$

which implies, from a conditional Markov inequality and DCT that

$$\begin{aligned} \zeta_0(X)^2 \|\hat{\beta}_n - \beta_n\|^2 &\leq \zeta_0(X)^2 C \| \sum_{i \neq j} S_n^{-1} X_i P_{ij} \varepsilon_j / \sqrt{r_n} \|^2 \\ &\to 0 \end{aligned}$$

so that $\sup_i |\Delta_i| = o_p(1)$. Similarly, $\sup_i |\Delta_i^2| \leq \zeta_0(X)^2 \|\hat{\beta}_n - \beta_n\|^2 = o_p(1)$ so that

$$\begin{aligned} |c_n'(\hat{\Sigma}_1 - \dot{\Sigma}_1) c_n| &= |c_n' S_n^{-1} \sum_{i \neq j \neq k} P_{ik} P_{jk} X_i X_j' (\hat{\varepsilon}_k^2 - \varepsilon_k^2) (S_n^{-1})' c_n| \\ &\leq 2 |c_n' S_n^{-1} \sum_{i \neq j \neq k} P_{ik} P_{jk} X_i X_j' \varepsilon_i \Delta_i (S_n^{-1})' c_n| \end{aligned}$$

$$
\begin{aligned}
&+ |c_n' S_n^{-1} \sum_{i \neq j \neq k} P_{ik} P_{jk} X_i X_j' \Delta_i^2 (S_n^{-1})' c_n| \\
&\leq 2 \sup_i \Delta_i |c_n' S_n^{-1} \sum_{i \neq j \neq k} P_{ik} P_{jk} X_i X_j' \varepsilon_i (S_n^{-1})' c_n| \\
&\quad + \sup_i \Delta_i^2 |c_n' S_n^{-1} \sum_{i \neq j \neq k} P_{ik} P_{jk} X_i X_j' (S_n^{-1})' c_n| \\
&= o_p(1) \times O_p(1) = o_p(1)
\end{aligned}
$$

Proof of $c_n'(\hat{\Sigma}_2 - \dot{\Sigma}_2)c_n = o_p(1)$

We next turn to

$$
\begin{aligned}
\hat{\Sigma}_2 - \dot{\Sigma}_2 = S_n^{-1} \sum_{i \neq j} P_{ij}^2 X_i X_j' (\hat{\varepsilon}_i \hat{\varepsilon}_j - \varepsilon_i \varepsilon_j)(S_n^{-1})' \\
+ S_n^{-1} \sum_{i \neq j} P_{ij}^2 X_i X_i' (\hat{\varepsilon}_j^2 - \varepsilon_j^2)(S_n^{-1})'
\end{aligned}
$$

We divide the analysis into the two terms.

Term 1

Using the fact that $\hat{\varepsilon}_i - \varepsilon_i = -\Delta_i$, we may write

$$
\begin{aligned}
& S_n^{-1} \textstyle\sum_{i \neq j} P_{ij}^2 X_i X_j' (\hat{\varepsilon}_i \hat{\varepsilon}_j - \varepsilon_i \varepsilon_j)(S_n^{-1})' \\
&= S_n^{-1} \textstyle\sum_{i \neq j} P_{ij}^2 X_i (\hat{\varepsilon}_i - \varepsilon_i) X_j' \varepsilon_j (S_n^{-1})' \\
&\quad + S_n^{-1} \textstyle\sum_{i \neq j} P_{ij}^2 X_i \varepsilon_i X_j' (\hat{\varepsilon}_j - \varepsilon_j)(S_n^{-1})' \\
&\quad + S_n^{-1} \textstyle\sum_{i \neq j} P_{ij}^2 X_i (\hat{\varepsilon}_i - \varepsilon_i) X_j' (\hat{\varepsilon}_j - \varepsilon_j)(S_n^{-1})' \\
&\leq \sup_i |\Delta_i| S_n^{-1} \textstyle\sum_{i \neq j} P_{ij}^2 X_i X_j' \varepsilon_j (S_n^{-1})' \\
&\quad + \sup_i |\Delta_i| S_n^{-1} \textstyle\sum_{i \neq j} P_{ij}^2 X_i \varepsilon_i X_j' (S_n^{-1})' \\
&\quad + \sup_i |\Delta_i|^2 S_n^{-1} \textstyle\sum_{i \neq j} P_{ij}^2 X_i X_j' (S_n^{-1})'
\end{aligned}
\tag{10}
$$

To bound the first two terms above, the following results will be needed:

(1)

$$
c_n' \frac{1}{n} \sum_{i \neq j} P_{ij}^2 z_i z_j' \varepsilon_j c_n = o_p(1)
$$

(2)

$$
c_n' S_n^{-1} \sum_{i \neq j} P_{ij}^2 V_i V_j' \varepsilon_j (S_n^{-1})' c_n = o_p(1)
$$

(3)

$$c_n' \frac{1}{\sqrt{n}} \sum_{i \neq j} P_{ij}^2 z_i V_j' \varepsilon_j (S_n^{-1})' c_n \leq C + o_p(1)$$

Proof: Let $\widehat{w}_i$ be the projection of w_i onto P, and note that $\widehat{w}_i^2 \leq w_i^2$.
For (1) we have

$$\begin{aligned}
&E\Big[\Big(c_n' \frac{1}{n} \sum_{i \neq j} P_{ij}^2 z_i z_j' \varepsilon_j c_n\Big)^2 \Big| \mathcal{Z}\Big] \\
&\quad = \frac{1}{n^2} \sum_{i,j} \sum_{k \neq \{i,j\}} P_{ik}^2 P_{jk}^2 (c_n' z_i)(c_n' z_j)(c_n' z_k)^2 E[\varepsilon_k^2 | \mathcal{Z}] \\
&\quad \leq C\Big(|\frac{1}{n^2} \sum_{i,j,k} P_{ik}^2 P_{jk}^2 (c_n' z_i)(c_n' z_j)(c_n' z_k)^2| \\
&\qquad + |\frac{2}{n^2} \sum_{i,j} P_{ii}^2 P_{ji}^2 (c_n' z_i)^3 (c_n' z_j)| + |\frac{1}{n^2} \sum_i P_{ii}^4 (c_n' z_i)^4| \Big) \\
&\quad \leq C\Big(|\frac{1}{n^2} \sum_k (c_n' z_k)^2 \big(\sum_i P_{ik} (c_n' z_i) \big) \big(\sum_j P_{jk} (c_n' z_j) \big)| \\
&\qquad + |\frac{2}{n^2} \sum_i P_{ii}^2 (c_n' z_i)^3 \sum_j P_{ji} (c_n' z_j)| + |\frac{1}{n^2} \sum_i P_{ii}^4 (c_n' z_i)^4| \Big) \\
&\quad \leq C \frac{1}{n^2} \sum_k (c_n' z_k)^4 \leq C \frac{1}{n^2} \sum_k \|z_k\|^4 \to 0
\end{aligned}$$

For (2)

$$\begin{aligned}
&E\Big[\Big(c_n' S_n^{-1} \sum_{i \neq j} P_{ij}^2 V_i V_j' \varepsilon_j (S_n^{-1})' c_n\Big)^2 \Big| \mathcal{Z}\Big] \\
&\quad \leq \frac{C}{r_n^2} \sum_{i \neq j} \sum_{s \neq t} P_{ij}^2 P_{st}^2 E\big[(c_n' V_i)(c_n' V_j \varepsilon_j)(c_n' V_s)(c_n' V_t \varepsilon_t) \big| \mathcal{Z}\big] \\
&\quad = \frac{C}{r_n^2} \sum_{i \neq j \neq t} P_{ij}^2 P_{it}^2 E\big[(c_n' V_i)^2 \big| \mathcal{Z}\big] E\big[(c_n' V_j)\varepsilon_j \big| \mathcal{Z}\big] E\big[(c_n' V_t)\varepsilon_t \big| \mathcal{Z}\big] \\
&\qquad + \frac{C}{r_n^2} \sum_{i \neq j} P_{ij}^4 E\big[(c_n' V_i)^2 \varepsilon_i \big| \mathcal{Z}\big] E\big[(c_n' V_j)^2 \varepsilon_j \big| \mathcal{Z}\big] \\
&\qquad + \frac{C}{r_n^2} \sum_{i \neq j} P_{ij}^4 E\big[(c_n' V_i)^2 \big| \mathcal{Z}\big] E\big[(c_n' V_j \varepsilon_j)^2 \big| \mathcal{Z}\big] \\
&\quad \leq C \frac{J_n^2}{r_n^2} \big(\sum_{i \neq j \neq t} P_{ij}^2 P_{it}^2 + 2 \sum_{i \neq j} P_{ij}^4 \big)
\end{aligned}$$

$$\leq C\frac{J_n^2 K_n}{r_n^2} \to 0$$

The results for (1) and (2) then follow from a conditional Markov inequality, combined with the dominated convergence theorem. Finally, for (3)

$$\begin{aligned}
&E\Big[\Big(c_n'\frac{1}{\sqrt{n}}\sum_{i\neq j}P_{ij}^2 z_i V_j'\varepsilon_j (S_n^{-1})' c_n\Big)^2\Big|\mathcal{Z}\Big] \\
&\quad = \frac{1}{n^2}\sum_{i\neq j}\sum_{s\neq t}P_{ij}^2 P_{st}^2 (c_n'z_i)(c_n'z_s)c_n' E[V_j V_t'\varepsilon_j\varepsilon_t|\mathcal{Z}]c_n \\
&\quad = \frac{1}{n^2}\sum_{i\neq j}\sum_{s}\sum_{t\neq\{j,s\}}P_{ij}^2 P_{st}^2 (c_n'z_i)(c_n'z_s)E[c_n'V_j\varepsilon_j|\mathcal{Z}]E[V_t'c_n\varepsilon_t|\mathcal{Z}] \\
&\qquad + \frac{1}{n^2}\sum_{i,s}\sum_{j\neq\{i,s\}}P_{ij}^2 P_{sj}^2 (c_n'z_i)(c_n'z_s)E[c_n'V_j V_j'c_n\varepsilon_j^2|\mathcal{Z}]
\end{aligned}$$

Since $E[c_n'V_j\varepsilon_j|\mathcal{Z}] \leq (c_n'E[V_jV_j'|\mathcal{Z}]c_n)^{1/2}E[\varepsilon_j^2|\mathcal{Z}]^{1/2} \leq C$, we have

$$\begin{aligned}
&\frac{1}{n^2}\sum_{i\neq j}\sum_{s}\sum_{t\neq\{j,s\}}P_{ij}^2 P_{st}^2 (c_n'z_i)(c_n'z_s)E[c_n'V_j\varepsilon_j|\mathcal{Z}]E[V_t'c_n\varepsilon_t|\mathcal{Z}] \\
&\quad \leq C\frac{1}{n^2}\sum_{i,s}(c_n'z_i)(c_n'z_s)\sum_{j\neq\{i\}}P_{ij}^2\sum_{t\neq\{j,s\}}P_{st}^2 \\
&\quad \leq C\frac{1}{n^2}\sum_{i,s}P_{ii}P_{ss}(c_n'z_i)(c_n'z_s) \\
&\quad \leq C\frac{1}{n^2}\sum_{i,s}\frac{1}{2}\big(P_{ii}^2(c_n'z_i)^2 + P_{ss}^2(c_n'z_s)^2\big) \\
&\quad \leq C\frac{1}{n}\sum_{i}(c_n'z_i)^2 \leq C
\end{aligned}$$

Next, we have that $E[\|V_j\|^4|\mathcal{Z}] \leq CJ_n^2$ by Assumption 3.7, and so $E[c_n'V_jV_j'c_n\varepsilon_j^2|\mathcal{Z}] \leq \|c_n\|^2 E[\|V_j\|^4|\mathcal{Z}]^{1/2}E[\varepsilon_j^4|\mathcal{Z}]^{1/2} \leq CJ_n$. Then, we have (repeating the earlier trick of $\widehat{w}_i^2 \leq w_i^2$)

$$\begin{aligned}
&\frac{1}{n^2}\sum_{i,s}\sum_{j\neq\{i,s\}}P_{ij}^2 P_{sj}^2 (c_n'z_i)(c_n'z_s)E[c_n'V_jV_j'c_n\varepsilon_j^2|\mathcal{Z}] \\
&\quad \leq C\frac{J_n}{n^2}\Big(|\sum_{i,j,s}P_{ij}^2 P_{sj}^2 (c_n'z_i)(c_n'z_s)| + |2\sum_{i,j}P_{ij}^2 P_{jj}^2 (c_n'z_i)(c_n'z_j)|
\end{aligned}$$

$$
\begin{aligned}
&\quad + |\sum_i P_{ii}^4 (c_n' z_i)^2| \Big) \\
&\leq C \frac{J_n}{n^2} \Big(\sum_{i,j,s} P_{ij} P_{sj} |c_n' z_i| |c_n' z_s| + 2 \sum_{i,j} P_{ij} P_{jj}^2 |c_n' z_i| |c_n' z_j| \\
&\quad + \sum_i P_{ii}^4 (c_n' z_i)^2 \Big) \\
&\leq C \frac{J_n}{n} \Big(\frac{1}{n} \sum_j (c_n' z_j)^2 + \frac{2}{n} \sum_{i,j} P_{jj}^2 (c_n' z_j)^2 + \frac{1}{n} \sum_i P_{ii}^4 (c_n' z_i)^2 \Big) \\
&\to 0
\end{aligned}
$$

The results above can now be used to bound the first two terms in (10)

$$
\begin{aligned}
&c_n' S_n^{-1} \sum_{i \neq j} P_{ij}^2 X_i X_j' \varepsilon_j (S_n^{-1})' c_n \\
&\quad = c_n' S_n^{-1} \sum_{i \neq j} P_{ij}^2 X_i X_j' \varepsilon_j (S_n^{-1})' c_n \\
&\quad = c_n' \frac{1}{n} \sum_{i \neq j} P_{ij}^2 z_i z_j' \varepsilon_j c_n + c_n' \frac{1}{\sqrt{n}} \sum_{i \neq j} P_{ij}^2 z_i V_j' \varepsilon_j (S_n^{-1})' c_n \\
&\qquad + c_n' \frac{1}{\sqrt{n}} \sum_{i \neq j} P_{ij}^2 S_n^{-1} V_i z_j' \varepsilon_j c_n \\
&\qquad + c_n' S_n^{-1} \sum_{i \neq j} P_{ij}^2 V_i V_j' \varepsilon_j (S_n^{-1})' c_n \\
&\quad \leq C
\end{aligned}
$$

For the third term in (10) we have

$$
\begin{aligned}
&E\Big[\big(S_n^{-1} \sum_{i \neq j} P_{ij}^2 X_i X_j' (S_n^{-1})' \big)^2 \Big| \mathcal{Z} \Big] \\
&\quad = \sum_{i \neq j \neq s \neq t} P_{ij}^2 P_{st}^2 E\big[c_n' S_n^{-1} X_i \big| \mathcal{Z} \big] E\big[c_n' S_n^{-1} X_j \big| \mathcal{Z} \big] \\
&\qquad E\big[c_n' S_n^{-1} X_s \big| \mathcal{Z} \big] E\big[c_n' S_n^{-1} X_t \big| \mathcal{Z} \big] \\
&\qquad + 4 \sum_{i \neq j \neq s} P_{ij}^2 P_{is}^2 E\big[(c_n' S_n^{-1} X_i)^2 \big| \mathcal{Z} \big] E\big[c_n' S_n^{-1} X_j \big| \mathcal{Z} \big] E\big[c_n' S_n^{-1} X_s \big| \mathcal{Z} \big] \\
&\qquad + 2 \sum_{i \neq j} P_{ij}^4 E\big[(c_n' S_n^{-1} X_i)^2 \big| \mathcal{Z} \big] E\big[(c_n' S_n^{-1} X_j)^2 \big| \mathcal{Z} \big]
\end{aligned}
$$

Breaking up the terms:

(1)

$$\begin{aligned}
&\sum_{i\neq j\neq s\neq t} P_{ij}^2 P_{st}^2 E\big[c_n' S_n^{-1} X_i \big| \mathcal{Z}\big] E\big[c_n' S_n^{-1} X_j \big| \mathcal{Z}\big] E\big[c_n' S_n^{-1} X_s \big| \mathcal{Z}\big] E\big[c_n' S_n^{-1} X_t \big| \mathcal{Z}\big] \\
&= \frac{1}{n^2} \sum_{i\neq j\neq s\neq t} P_{ij}^2 P_{st}^2 (c_n' z_i)(c_n' z_j)(c_n' z_s)(c_n' z_t) \\
&\leq \frac{1}{n^2} \sum_{i,j,s,t} P_{ij} P_{st} |c_n' z_i||c_n' z_j||c_n' z_s||c_n' z_t| \\
&\leq \frac{1}{n^2} \sum_{i,s} (c_n' z_i)^2 (c_n' z_s)^2 \leq C
\end{aligned}$$

(2)

$$\begin{aligned}
&\sum_{i\neq j\neq s} P_{ij}^2 P_{is}^2 E\big[(c_n' S_n^{-1} X_i)^2 \big| \mathcal{Z}\big] E\big[c_n' S_n^{-1} X_j \big| \mathcal{Z}\big] E\big[c_n' S_n^{-1} X_s \big| \mathcal{Z}\big] \\
&= \frac{1}{n^2} \sum_{i\neq j\neq s} P_{ij}^2 P_{is}^2 (c_n' z_i)^2 (c_n' z_j)(c_n' z_s) \\
&\quad + \frac{1}{n} \sum_{i\neq j\neq s} P_{ij}^2 P_{is}^2 (c_n' S_n^{-1} E[V_i V_i' | \mathcal{Z}] (S_n^{-1})' c_n)(c_n' z_j)(c_n' z_s) \\
&\leq \frac{1}{n^2} \sum_i (c_n' z_i)^2 \sum_{j,s} P_{ij}^2 P_{is}^2 |c_n' z_j||c_n' z_s| + C \frac{1}{r_n n} \sum_{i,j,s} P_{ij} P_{is} |c_n' z_j||c_n' z_s| \\
&\leq \frac{1}{n^2} \sum_i (c_n' z_i)^4 + \frac{C}{r_n} \frac{1}{n} \sum_i (c_n' z_i)^2 = o_p(1)
\end{aligned}$$

(3)

$$\begin{aligned}
&\sum_{i\neq j} P_{ij}^4 E\big[(c_n' S_n^{-1} X_i)^2 \big| \mathcal{Z}\big] E\big[(c_n' S_n^{-1} X_j)^2 \big| \mathcal{Z}\big] \\
&= \frac{1}{n^2} \sum_{i\neq j} P_{ij}^4 (c_n' z_i)^2 (c_n' z_j)^2 \\
&\quad + 2 \frac{1}{n} \sum_{i\neq j} P_{ij}^4 (c_n' z_i)^2 (c_n' S_n^{-1} E[V_j V_j' | \mathcal{Z}] (S_n^{-1})' c_n) \\
&\quad + \sum_{i\neq j} P_{ij}^4 (c_n' S_n^{-1} E[V_1 V_1' | \mathcal{Z}] (S_n^{-1})' c_n)(c_n' S_n^{-1} E[V_j V_j' | \mathcal{Z}] (S_n^{-1})' c_n)
\end{aligned}$$

$$\leq \frac{1}{n^2}\sum_i (c_n' z_i)^2 \sum_j P_{ij}(c_n' z_j)^2 + \frac{C}{r_n}\frac{1}{n}\sum_i P_{ii}(c_n' z_i)^2$$
$$+ \frac{C}{r_n^2}\sum_i P_{ii}$$
$$\leq \frac{1}{n^2}\sum_i (c_n' z_i)^4 + \frac{C}{r_n}\frac{1}{n}\sum_i (c_n' z_i)^2 + \frac{K_n}{r_n^2} = o_p(1)$$

Combining these results gives

$$E\Big[\Big(c_n' S_n^{-1}\sum_{i\neq j} P_{ij}^2 X_i X_j'(S_n^{-1})' c_n\Big)^2 \Big| \mathcal{Z}\Big] \leq C + o_p(1)$$

and so, by a conditional Markov inequality we have that

$$P(|c_n' S_n^{-1}\sum_{i\neq j} P_{ij}^2 X_i X_j'(S_n^{-1})' c_n| > M \big| \mathcal{Z}) \leq \frac{C + o_p(1)}{M^2}$$

Since the LHS is bounded above, a dominated convergence theorem implies that, for sufficiently large n we have

$$P(|c_n' S_n^{-1}\sum_{i\neq j} P_{ij}^2 X_i X_j'(S_n^{-1})' c_n| > M) \leq \frac{C}{M^2}$$

so that $c_n' S_n^{-1}\sum_{i\neq j} P_{ij}^2 X_i X_j'(S_n^{-1})' c_n = O_p(1)$.

Then, by (10)

$$S_n^{-1}\sum_{i\neq j} P_{ij}^2 X_i X_j'(\hat{\varepsilon}_i\hat{\varepsilon}_j - \varepsilon_i\varepsilon_j)(S_n^{-1})'$$
$$\leq \sup_i |\Delta_i| \times O_p(1)$$
$$= o_p(1)$$

since $\sup_i |\Delta_i| \to 0$ by Assumption 3.8.

Term 2

Recall that $\hat{\varepsilon}_i^2 - \varepsilon_i^2 = -2\Delta_i\varepsilon_i + \Delta_i^2$. Then

$$S_n^{-1}\sum_{i\neq j} P_{ij}^2 X_i X_i'(\hat{\varepsilon}_j^2 - \varepsilon_j^2)(S_n^{-1})'$$

$$
\begin{aligned}
&\leq 2 \sup_i |\Delta_i| ||S_n^{-1} \sum_{i \neq j} P_{ij}^2 X_i X_i' \varepsilon_j (S_n^{-1})'| \\
&\quad + \sup_i |\Delta_i^2| ||S_n^{-1} \sum_{i \neq j} P_{ij}^2 X_i X_i' (S_n^{-1})'| \\
&\leq \sup_i |\Delta_i| O_p(1) + \sup_i |\Delta_i^2| O_p(1) \\
&= o_p(1)
\end{aligned}
$$

where the second inequality follows from the results in part (1) of this subsection.

Combining these results we have $\hat{\Sigma}_2 - \dot{\Sigma}_2 = o_p(1)$ as desired.

C.6 Proof of $c'\hat{\Xi}_n c \to c'\Xi_n c$

We begin with a lemma that will be used below that is similar to Lemma A3 from Chao et al (2009).

Lemma C.7. *Let scalars (W_i, Y_i) be independent over i, conditional on $\mathcal{Z}$. Also, let P be a symmetric, idempotent matrix, and denote $\bar{w}_i = E[W_i|\mathcal{Z}]$, $\tilde{\sigma}_{W_i} = Var(W_i|\mathcal{Z})$, and similarly for Y.*

Then, there exists a constant C such that

$$
\begin{aligned}
E\Big[\Big(\sum_{i \neq j} P_{ij}^2 W_i Y_j - \sum_{i \neq j} P_{ij}^2 \bar{w}_i \bar{y}_j\Big)^2 \Big| \mathcal{Z}\Big] \\
\leq C\big(\sum_{i \neq j} P_{ij}^4 \tilde{\sigma}_{W_i}^2 \tilde{\sigma}_{Y_j}^2 + \big)
\end{aligned}
$$

Proof. To begin, recalling that $\tilde{w} = W - \bar{w}$

$$
\sum_{i \neq j} P_{ij}^2 W_i Y_j - \sum_{i \neq j} P_{ij}^2 \bar{w}_i \bar{y}_j = \sum_{i \neq j} P_{ij}^2 \bar{w}_i \tilde{y}_j + \sum_{i \neq j} P_{ij}^2 \tilde{w}_i \bar{y}_j + \sum_{i \neq j} P_{ij}^2 \tilde{w}_i \tilde{y}_j
$$

We deal with each of the three RHS terms. Firstly

$$
\begin{aligned}
E\Big[\Big(\sum_{i \neq j} P_{ij}^2 \tilde{w}_i \tilde{y}_j\Big)^2 \Big| \mathcal{Z}\Big] &= \sum_{i \neq j} \sum_{s \neq t} P_{ij}^2 P_{st}^2 E[\tilde{w}_i \tilde{y}_j \tilde{w}_s \tilde{y}_t | \mathcal{Z}] \\
&= \sum_{i \neq j} P_{ij}^4 \big(E[\tilde{w}_i \tilde{y}_i | \mathcal{Z}] E[\tilde{w}_j \tilde{y}_j | \mathcal{Z}] + E[\tilde{w}_i^2 | \mathcal{Z}] E[\tilde{y}_j^2 | \mathcal{Z}]\big) \\
&\leq 2 \sum_{i \neq j} P_{ij}^4 \tilde{\sigma}_{W_i}^2 \tilde{\sigma}_{Y_j}^2
\end{aligned}
$$

where $\tilde{\sigma}^2_W = Var(W|\mathcal{Z})$.

Then, since $E[\bar{w}_i\tilde{y}_i|\mathcal{Z}] = 0$

$$\begin{aligned}
E\Big[\Big(\sum_{i\neq j} P^2_{ij}\bar{w}_i\tilde{y}_j\Big)^2\Big|\mathcal{Z}\Big] &= \sum_{i\neq j}\sum_{s\neq t} P^2_{ij}P^2_{st}E[\bar{w}_i\tilde{y}_j\bar{w}_s\tilde{y}_t|\mathcal{Z}] \\
&= \sum_{i,s}\sum_{j\neq\{i,s\}} P^2_{ij}P^2_{sj}\bar{w}_i\bar{w}_s\tilde{\sigma}^2_{Y_j} \\
&\leq |\sum_j \tilde{\sigma}^2_{Y_j}\big(\sum_i P^2_{ij}\bar{w}_i\big)\big(\sum_j P^2_{sj}\bar{w}_s\big)| \\
&\quad + |2\sum_{i,s} P^2_{ii}P^2_{si}\bar{w}_i\bar{w}_s\tilde{\sigma}^2_{Y_i}| + |\sum_i P^4_{ii}\bar{w}^2_i\tilde{\sigma}^2_{Y_i}| \\
&\leq |\sum_j \tilde{\sigma}^2_{Y_j}\bar{w}^2_j| + |2\sum_i P^2_{ii}\bar{w}^2_i\tilde{\sigma}^2_{Y_i}| + |\sum_i P^4_{ii}\bar{w}^2_i\tilde{\sigma}^2_{Y_i}| \\
&\leq C\sum_j \tilde{\sigma}^2_{Y_j}\bar{w}^2_j
\end{aligned}$$

and similarly for

$$\mathrm{E}\Big[\Big(\sum_{i\neq j} P^2_{ij}\tilde{w}_i\bar{y}_j\Big)^2\Big|\mathcal{Z}\Big] \leq C\sum_j \tilde{\sigma}^2_{W_j}\bar{y}^2_j$$

Combining, gives the result. □

From the previous subsection we now have $c_n'\hat{\Xi}_n c_n = c_n'(\dot{\Sigma}_1 + \dot{\Sigma}_2)c_n + o_p(1)$. The Lemma below shows that these quantities approximate certain conditional means.

Lemma C.8. *We have, for* $\sigma^2_i = E[\varepsilon^2_i|\mathcal{Z}]$

$$\begin{aligned}
c_n'\dot{\Sigma}_1 c_n &= \frac{1}{n}\sum_{i\neq j\neq k} P_{ik}P_{jk}c_n'z_iz_j'c_n\sigma^2_k + o_p(1) \\
c_n'\dot{\Sigma}_2 c_n &= \frac{1}{n}\sum_{i\neq j} P^2_{ij}c_n'z_iz_i'c_n\sigma^2_j \\
&\quad + \sum_{i\neq j} c_n'S_n^{-1}P^2_{ij}\big(E[V_iV_i'|\mathcal{Z}]\sigma^2_j + E[V_i\varepsilon_i|\mathcal{Z}]E[V_j'\varepsilon_j|\mathcal{Z}]\big)(S_n^{-1})'c_n \\
&\quad + o_p(1)
\end{aligned}$$

Proof. For the first result, since

$$c_n'\dot{\Sigma}_1 c_n = c_n' S_n^{-1} \sum_{i\neq j\neq k} P_{ik}P_{jk}X_iX_j'\varepsilon_k^2 (S_n^{-1})' c_n$$

we can again apply Lemma C.6, with $W_i = c_n'S_n^{-1}X_i$ and $\eta_i = \varepsilon_i^2$. To confirm the conditions hold note that:

(i) $\bar{w}_i = c_n'z_i/\sqrt{n}$ and so $\sum_i \bar{w}_i^2 \leq C$ by Assumption 3.2, while $E[\sum_i \bar{w}_i^4] \to 0$ by Assumption 3.7.

(ii) $\bar{\sigma}_W^2 = c_n'S_n^{-1}E[V_iV_i'|\mathcal{Z}](S_n')^{-1}c_n \leq C/r_n$ by Assumption 3.4.

(iii) $\bar{\mu}_\eta = E[\varepsilon_i^2|\mathcal{Z}] \leq C$ by Assumption 3.4, so $E_{\mathcal{Z}}[\bar{\mu}_\eta^2] \leq C$ holds. Also $\bar{\sigma}_\eta^2 = Var(\varepsilon_i^2|\mathcal{Z}) \leq E[\varepsilon_i^4|\mathcal{Z}] \leq C$ by Assumption 3.7.

Applying the result of the lemma we have

$$\begin{aligned} c_n'\dot{\Sigma}_1 c_n &= c_n'S_n^{-1} \sum_{i\neq j\neq k} P_{ik}P_{jk}E\big[X_iX_j'\varepsilon_k^2\big|\mathcal{Z}\big](S_n^{-1})'c_n + o_p(1) \\ &= \frac{1}{n}\sum_{i\neq j\neq k} P_{ik}P_{jk}c_n'z_iz_j'c_n\sigma_k^2 + o_p(1) \end{aligned}$$

For the second result, we first decompose the conditional expectation into two parts

$$\begin{aligned} c'\bar{\Sigma}_{21}c &= E\Big[c_n'S_n^{-1}\sum_{i\neq j} P_{ij}^2X_iX_i'\varepsilon_j^2(S_n^{-1})'c_n\big|\mathcal{Z}\Big] \\ &= \frac{1}{n}\sum_{i\neq j} P_{ij}^2c_n'z_iz_i'c_n\sigma_j^2 + \sum_{i\neq j} c_n'S_n^{-1}P_{ij}^2E[V_iV_i'|\mathcal{Z}]\sigma_j^2(S_n^{-1})'c_n \\ c'\bar{\Sigma}_{22}c &= E\Big[c_n'S_n^{-1}\sum_{i\neq j} P_{ij}^2X_i\varepsilon_iX_j'\varepsilon_j(S_n^{-1})'c_n\big|\mathcal{Z}\Big] \\ &= \sum_{i\neq j} c_n'S_n^{-1}P_{ij}^2E[V_i\varepsilon_i|\mathcal{Z}]E[V_j'\varepsilon_j|\mathcal{Z}](S_n^{-1})'c_n \end{aligned}$$

For part 1, letting $W_i = c_n'S_n^{-1}X_iX_i'(S_n^{-1})c_n$ and $Y_i = \varepsilon_i^2$ we have

$$\begin{aligned} E\big[\big(c'(\dot{\Sigma}_{21}-\bar{\Sigma}_{21})c\big)^2\big|\mathcal{Z}\big] &= E\big[\big(\sum_{i\neq j} P_{ij}^2(W_iY_j - \bar{w}_i\bar{y}_j)\big)^2\big|\mathcal{Z}\big] \\ &\leq CE\big[\big(\sum_{i\neq j} P_{ij}^2\tilde{w}_i\bar{y}_j\big)^2\big|\mathcal{Z}\big] + CE\big[\big(\sum_{i\neq j} P_{ij}^2\tilde{y}_i\bar{w}_j\big)^2\big|\mathcal{Z}\big] \end{aligned}$$

$$+ CE\big[\big(\sum_{i\neq j} P_{ij}^2 \tilde{w}_i \tilde{y}_j\big)\big)^2 \big| \mathcal{Z}\big]$$

where

$$\begin{aligned}\bar{w}_i &= \frac{1}{n} c_n' z_i z_i' c_n + c_n' S_n^{-1} E[V_i V_i' | \mathcal{Z}](S_n^{-1})' c_n \\ &\leq \frac{1}{n} c_n' z_i z_i' c_n + C/r_n\end{aligned}$$

$\bar{y}_i = \sigma_i^2$, $\tilde{w}_i = W_i - \bar{w}_i$, and $\tilde{y}_i = Y_i - \bar{y}_i$. Next, note that

$$\begin{aligned}\tilde{\sigma}_{W_i}^2 &\leq E\big[\big(c' z_i/\sqrt{n} + c' S_n^{-1} V_i\big)^4 \big| \mathcal{Z}\big] \\ &\leq C\big(\frac{1}{n^2}(c' z_i)^4 + E\big[(c' S_n^{-1} V_i)^4 \big| \mathcal{Z}\big]\big) \\ &\leq C\big(\frac{1}{n^2}(c' z_i)^4 + \frac{J_n^2}{r_n^2}\big) \\ \tilde{\sigma}_{Y_i}^2 &\leq C\end{aligned}$$

Using this, we have

$$\begin{aligned}E\Big[\Big(\sum_{i\neq j} P_{ij}^2 \bar{w}_i \tilde{y}_j\Big)^2 \Big| \mathcal{Z}\Big] &= \sum_{i\neq j}\sum_{s\neq t} P_{ij}^2 P_{st}^2 E[\bar{w}_i \tilde{y}_j \bar{w}_s \tilde{y}_t | \mathcal{Z}] \\ &= \sum_{i,s}\sum_{j\neq\{i,s\}} P_{ij}^2 P_{sj}^2 \bar{w}_i \bar{w}_s \tilde{\sigma}_{Y_j}^2 \\ &\leq C\frac{1}{n^2}\sum_j \big(\sum_{i\neq j} P_{ij}^2 (c_n' z_i)^2\big)\big(\sum_{s\neq j} P_{sj}^2 (c_n' z_s)^2\big) \\ &\quad + C\frac{1}{r_n^2}\sum_{i,s}\sum_{j\neq\{i,s\}} P_{ij}^2 P_{sj}^2 \\ &\to 0\end{aligned}$$

where the final line uses $\sum_{i,s}\sum_{j\neq\{i,s\}} P_{ij}^2 P_{sj}^2 \leq K_n$,and the fact that, since P is a projection matrix, we have that $\sum_j P_{ij} w_j$ is the projection of w_i onto P and hence $\sum_i w_i \sum_j P_{ij} w_j \leq \sum_i w_i^2$, so that

$$\frac{1}{n^2}\sum_j \big(\sum_{i\neq j} P_{ij}^2 (c_n' z_i)^2\big)\big(\sum_{s\neq j} P_{sj}^2 (c_n' z_s)^2\big) \leq \frac{1}{n^2}\sum_j)^2\big(\sum_i P_{ij} (c_n' z_{ji})^2\big)\big(\sum_s P_{sj} (c_n' z_s)^2\big)$$

$$\leq \frac{1}{n^2}\sum_j (c_n' z_j)^4 \to 0$$

Similarly, since $\sum_i P_{ij}^2 \leq \sum_i P_{ij} = 1$

$$\begin{aligned}
E\Big[\Big(\sum_{i\neq j} P_{ij}^2 \bar{y}_i \tilde{w}_j\Big)^2 \Big| \mathcal{Z}\Big] &= \sum_{i\neq j}\sum_{s\neq t} P_{ij}^2 P_{st}^2 E[\bar{y}_i \tilde{w}_j \bar{y}_s \tilde{w}_t | \mathcal{Z}] \\
&= \sum_{i,s}\sum_{j\neq\{i,s\}} P_{ij}^2 P_{sj}^2 \bar{y}_i \bar{y}_s \tilde{\sigma}_{W_j}^2 \\
&\leq C\sum_{i,s}\sum_{j\neq\{i,s\}} P_{ij}^2 P_{sj}^2 \big(\frac{1}{n^2}(c' z_i)^4 + \frac{J_n^2}{r_n^2}\big) \\
&\leq C\frac{1}{n^2}\sum_j (c_n' z_j)^4 + \frac{K_n J_n^2}{r_n^2} \\
&\to 0
\end{aligned}$$

Finally,

$$\begin{aligned}
E\Big[\Big(\sum_{i\neq j} P_{ij}^2 \tilde{w}_i \tilde{y}_j\Big)^2 \Big| \mathcal{Z}\Big] &= \sum_{i\neq j}\sum_{s\neq t} P_{ij}^2 P_{st}^2 E[\tilde{w}_i \tilde{y}_j \tilde{w}_s \tilde{y}_t | \mathcal{Z}] \\
&= \sum_{i\neq j} P_{ij}^4 \big(E[\tilde{w}_i \tilde{y}_i | \mathcal{Z}] E[\tilde{w}_j \tilde{y}_j | \mathcal{Z}] + E[\tilde{w}_i^2 | \mathcal{Z}] E[\tilde{y}_j^2 | \mathcal{Z}]\big) \\
&\leq 2\sum_{i\neq j} P_{ij}^4 \tilde{\sigma}_{W_i}^2 \tilde{\sigma}_{Y_j}^2 \\
&\leq C\sum_{i\neq j} P_{ij}^4 \big(\frac{1}{n^2}(c' z_i)^4 + \frac{J_n^2}{r_n^2}\big) \to 0
\end{aligned}$$

For part 2, letting $W_i = c_n' S_n^{-1} X_i \varepsilon_i$ and $\bar{w}_i = c_n' S_n^{-1} E[V_i \varepsilon_i | \mathcal{Z}]$ we have

$$\begin{aligned}
E\big[\big(c'(\dot{\Sigma}_{22} - \bar{\Sigma}_{22})c\big)^2 \big| \mathcal{Z}\big] &= E\big[\big(\sum_{i\neq j} P_{ij}^2 (W_i W_j - \bar{w}_i \bar{w}_j)\big)^2 \big| \mathcal{Z}\big] \\
&\leq CE\big[\big(\sum_{i\neq j} P_{ij}^2 \tilde{w}_i \bar{w}_j\big)^2 \big| \mathcal{Z}\big] \\
&\quad + E\big[\big(\sum_{i\neq j} P_{ij}^2 \tilde{w}_i \tilde{w}_j\big)\big)^2 \big| \mathcal{Z}\big]
\end{aligned}$$

where $\tilde{w}_i = W_i - \bar{w}_i$. Note that $\bar{w}_i \leq C/\sqrt{r_n}$, and

$$\begin{aligned}\tilde{\sigma}^2_{W_i} = Var(W_i|\mathcal{Z}) = \frac{1}{n} c_n' z_i z_i' c_n E[\varepsilon_i^2|\mathcal{Z}] \\ c_n' S_n^{-1} E[V_i V_i'|\mathcal{Z}](S_n^{-1})' c_n \\ \leq C(\frac{1}{n} c_n' z_i z_i' c_n + \frac{1}{r_n})\end{aligned}$$

The first term is

$$\begin{aligned}E\Big[\Big(\sum_{i\neq j} P_{ij}^2 \bar{w}_i \tilde{w}_j\Big)^2 \Big|\mathcal{Z}\Big] &= \sum_{i\neq j}\sum_{s\neq t} P_{ij}^2 P_{st}^2 E[\bar{w}_i \tilde{w}_j \bar{w}_s \tilde{w}_t|\mathcal{Z}] \\ &= \sum_{i,s}\sum_{j\neq\{i,s\}} P_{ij}^2 P_{sj}^2 \bar{w}_i \bar{w}_s \tilde{\sigma}^2_{W_j} \\ &\leq \frac{1}{r_n^2}\sum_{i,s}\sum_{j\neq\{i,s\}} P_{ij}^2 P_{sj}^2 + \frac{1}{r_n}\frac{1}{n}\sum_{i,s}\sum_{j\neq\{i,s\}} P_{ij}^2 P_{sj}^2 c_n' z_j z_j' c_n \\ &\leq \frac{K_n}{r_n^2} + \frac{1}{r_n}\frac{1}{n}\sum_i c_n' z_i z_i' c_n \to 0\end{aligned}$$

while the second term is

$$\begin{aligned}E\big[\big(\sum_{i\neq j} P_{ij}^2 \tilde{w}_i \tilde{w}_j\big)\big)^2\big|\mathcal{Z}\big] &= \sum_{i\neq j}\sum_{s\neq t} P_{ij}^2 P_{st}^2 E[\tilde{w}_i \tilde{w}_j \tilde{w}_s \tilde{w}_t|\mathcal{Z}] \\ &\leq 2\sum_{i\neq j} P_{ij}^4 \tilde{\sigma}^2_{W_i}\tilde{\sigma}^2_{W_j} \\ &\leq C\frac{1}{r_n^2}\sum_{i\neq j} P_{ij}^4 + C\frac{1}{r_n}\frac{1}{n}\sum_{i\neq j} P_{ij}^4 c_n' z_j z_j' c_n \\ &\quad + C\frac{1}{n^2}\sum_{i\neq j} P_{ij}^4 (c_n' z_i)^2 (c_n' z_j)^2 \\ &\to 0\end{aligned}$$

where the final line uses the fact that, since P is a projection matrix, we have that $\sum_j P_{ij} w_j$ is the projection of w_i onto P and hence $\sum_i w_i \sum_j P_{ij} w_j \leq \sum_i w_i^2$, so

$$\frac{1}{n^2}\sum_{i\neq j} P_{ij}^4 (c_n' z_i)^2 (c_n' z_j)^2 \leq \frac{1}{n^2}\sum_i (c_n' z_i)^2 \Big(\sum_j P_{ij} (c_n' z_j)^2\Big)$$

$$\leq \frac{1}{n^2}\sum_i (c_n' z_i)^4 \to 0$$

Combining the above results with the triangle inequality gives the result of the Lemma. □

Finishing proof of variance estimator consistency

Note that $\hat{\Xi}_n = S_n^{-1}\hat{\Xi}_J(S_n^{-1})'$, with $\hat{\Xi}_n$ as defined in the main paper. The above results give

$$\begin{aligned}
c_n' S_n^{-1}\hat{\Xi}_J(S_n^{-1})' c_n &= c_n'\hat{\Xi}_n c_n \\
&= S_n^{-1}\Big(\sum_{i\neq j\neq k} P_{ik}P_{jk}X_i X_j'\hat{\varepsilon}_k^2 + \sum_{i\neq j} P_{ij}^2\big(X_i\hat{\varepsilon}_i X_j'\hat{\varepsilon}_j + X_i X_i'\hat{\varepsilon}_j^2\big)\Big)(S_n^{-1})' \\
&= \frac{1}{n}\sum_{i\neq j\neq k} P_{ik}P_{jk}c_n' z_i z_j' c_n \sigma_k^2 + \frac{1}{n}\sum_{i\neq j} P_{ij}^2 c_n' z_i z_i' c_n \sigma_j^2 \\
&+ \sum_{i\neq j} c_n' S_n^{-1} P_{ij}^2\big(E[V_i V_i'|\mathcal{Z}]\sigma_j^2 + E[V_i\varepsilon_i|\mathcal{Z}]E[V_j'\varepsilon_j|\mathcal{Z}]\big)(S_n^{-1})' c_n \\
&+ o_p(1)
\end{aligned}$$

For any $\|c_n\| \leq C$. Now, write $\bar{z}_i = \sum_j P_{ij} z_j$ as the projection of the reduced form z_i onto the observed instruments.

$$\begin{aligned}
&\frac{1}{n}\sum_{i\neq j\neq k} P_{ik}P_{jk}c_n' z_i z_j' c_n \sigma_k^2 \\
&\quad= \frac{1}{n}\sum_k \sigma_k^2 \sum_{j\neq k} P_{jk}c_n' z_j\big(\sum_i P_{ik}c_n' z_i - P_{jk}c_n' z_j - P_{kk}c_n' z_k\big) \\
&\quad= \frac{1}{n}\sum_k \sigma_k^2 \sum_{j\neq k} P_{jk}c_n' z_j\big(c_n'\bar{z}_k - P_{jk}c_n' z_j - P_{kk}c_n' z_k\big) \\
&\quad= \frac{1}{n}\sum_k \sigma_k^2\big(c_n'\bar{z}_k - P_{kk}c_n' z_k\big)\sum_{j\neq k} P_{jk}c_n' z_j \\
&\qquad- \frac{1}{n}\sum_k \sigma_k^2 \sum_{j\neq k} P_{jk}^2 (c_n' z_j)^2
\end{aligned}$$

Next, note that, since $\frac{1}{n}\sum_i \|z_i - \bar{z}_i\|^2 \to 0$, and $\sigma_i^2 \leq C$

$$\frac{1}{n}\sum_k \sigma_k^2\big((c_n'\bar{z}_k)^2 - (c_n' z_k)^2\big) = \frac{1}{n}\sum_k \sigma_k^2\big((c_n'\bar{z}_k) + (c_n' z_k)\big)\big((c_n'\bar{z}_k) - (c_n' z_k)\big)$$

$$\leq C\big(\frac{1}{n}\sum_i (c_n' z_k)^2\big)^{1/2}\big(\frac{1}{n}\sum_k \|z_i - \bar{z}_i\|^2\big)^{1/2} \to 0$$

and similarly $\frac{1}{n}\sum_k \sigma_k^2 P_{kk}\big(c_n'\bar{z}_k - c_n' z_k\big)c_n' z_k \to 0$. So, we may write

$$\begin{aligned}
&\frac{1}{n}\sum_k \sigma_k^2\big(c_n'\bar{z}_k - P_{kk}c_n' z_k\big)\sum_{j\neq k} P_{jk}c_n' z_j \\
&\quad = \frac{1}{n}\sum_k \sigma_k^2\big(c_n'\bar{z}_k - P_{kk}c_n' z_k\big)\big(c_n'\bar{z}_k - P_{kk}c_n' z_k\big) \\
&\quad = \frac{1}{n}\sum_k \sigma_k^2(1-P_{kk})^2 c_n' z_k z_k' c_n \\
&\qquad + \frac{1}{n}\sum_k \sigma_k^2\big((c_n'\bar{z}_k)^2 - (c_n' z_k)^2\big) \\
&\qquad + \frac{1}{n}\sum_k \sigma_k^2 P_{kk}\big(c_n'\bar{z}_k - c_n' z_k\big)c_n' z_k \\
&\quad = \frac{1}{n}\sum_k \sigma_k^2(1-P_{kk})^2 c_n' z_k z_k' c_n + o_p(1)
\end{aligned}$$

from which it follows that

$$\begin{aligned}
c_n' S_n^{-1}\hat{\Xi}_J (S_n^{-1})' c_n &= \frac{1}{n}\sum_i (1-P_{ii})^2 c_n' z_i z_i' c_n \sigma_i^2 \\
&\quad + \sum_{i\neq j} c_n' S_n^{-1} P_{ij}^2\big(E[V_i V_i'|\mathcal{Z}]\sigma_j^2 + E[V_i\varepsilon_i|\mathcal{Z}]E[V_j'\varepsilon_j|\mathcal{Z}]\big)(S_n^{-1})' c_n \\
&\quad + o_p(1) \\
&\quad = c_n'\Xi_n c_n + o_p(1)
\end{aligned}$$

Consistency of the denominator term

$$\begin{aligned}
\|S_n^{-1}\hat{H}_J(S_n^{-1})' - H_n\| &= \|\sum_{i\neq j} S_n^{-1} X_i P_{ij} X_j (S_n^{-1})' - \frac{1}{n}\sum_i (1-P_{ii}) z_i z_i'\| \\
&\to 0
\end{aligned}$$

then, Assumption 3.2, implies

$$b_n^2 \hat{V}_J = b_n^2 \alpha_n' \hat{H}_J^{-1}\hat{\Xi}_J \hat{H}_J^{-1}\alpha_n$$

$$
\begin{aligned}
&= b_n^2 \alpha_n' \big(S_n^{-1} \hat{H}_J (S_n^{-1})'\big)^{-1} S_n^{-1} \hat{\Xi}_J (S_n^{-1})' \big(S_n^{-1} \hat{H}_J (S_n^{-1})'\big)^{-1} \alpha_n \\
&= b_n^2 \alpha_n' (S_n^{-1})' H_n^{-1} S_n^{-1} \hat{\Xi}_J (S_n^{-1})' H_n^{-1} S_n^{-1} \alpha_n + o_p(1) \\
&= c_n' H_n^{-1} S_n^{-1} \hat{\Xi}_J (S_n^{-1})' H_n^{-1} c_n + o_p(1)
\end{aligned}
$$

for $c_n = b_n \alpha_n' (S_n^{-1})'$. Since $\|c_n' H_n^{-1}\| \leq C$, we find

$$
\begin{aligned}
b_n^2 \hat{V}_J &= c_n' H_n^{-1} \Xi_n H_n^{-1} c_n + o_p(1) \\
&= b_n^2 \alpha_n' (S_n^{-1})' H_n^{-1} \Xi_n H_n^{-1} S_n^{-1} \alpha_n + o_p(1) \\
&= V_{J,n} + o_p(1)
\end{aligned}
$$

Then by the continuous mapping theorem and Slutsky's lemma, we have

$$
\hat{V}_n^{-1/2} (\hat{\theta}_n - \theta_n) \Rightarrow N(0,1)
$$

C.7 Consistency of variance of jackknife k-class estimator

Let

$$
\begin{aligned}
\hat{X}_i &= X_i - \hat{\varepsilon}_i \hat{\gamma} \\
\hat{\gamma} &= (\hat{X}' \hat{\varepsilon}) / (\hat{\varepsilon}' \hat{\varepsilon})
\end{aligned}
$$

The variance estimator is

$$
\begin{aligned}
\hat{\Sigma}_{HFUL} = &\sum_{i \neq j \neq k} P_{ik} P_{jk} \hat{X}_i \hat{X}_j' \hat{\varepsilon}_k^2 \\
&+ \sum_{i \neq j} P_{ij}^2 \big(\hat{X}_i \hat{X}_i' \hat{\varepsilon}_j^2 + \hat{X}_i \hat{\varepsilon}_i \hat{X}_j' \hat{\varepsilon}_j \big)
\end{aligned}
$$

This is equivalent to the variance estimator for JIVE, replacing X_i with $\hat{X}_i$, and we prove consistency in a similar manner. Specifically, define

$$
\begin{aligned}
\hat{\Sigma}_1 &= \sum_{i \neq j \neq k} P_{ik} P_{jk} \check{X}_i \check{X}_j' \hat{\varepsilon}_k^2 \\
\hat{\Sigma}_2 &= \sum_{i \neq j} P_{ij}^2 \big(\check{X}_i \check{X}_i' \hat{\varepsilon}_j^2 + \check{X}_i \hat{\varepsilon}_i \check{X}_j' \hat{\varepsilon}_j \big)
\end{aligned}
$$

with $\check{X}_i = S_n^{-1}\hat{X}_i$. Next, we define new versions of these matrices, replacing the estimated quantities with their probability limits. Let $\dot{X}_i = S_n^{-1}(X_i - \varepsilon_i\gamma_n)$ with $\gamma_n = \sum_i E[X_i\varepsilon_i]/\sum_i \sigma_i^2$ and $\sigma_i^2 = E[\varepsilon_i^2]$. Then define

$$\dot{\Sigma}_1 = \sum_{i\neq j\neq k} P_{ik}P_{jk}\dot{X}_i\dot{X}_j'\varepsilon_k^2$$
$$\dot{\Sigma}_2 = \sum_{i\neq j} P_{ij}^2\big(\dot{X}_i\dot{X}_i'\varepsilon_j^2 + \dot{X}_i\varepsilon_i\dot{X}_j'\varepsilon_j\big)$$

Proof of $c'(\hat{\Sigma}_2 - \dot{\Sigma}_2)c \to 0$

Let $d_i = C(1+|\varepsilon_i|+|c'V_i|)$, $\hat{A} = C(1+|c'\hat{\gamma}|+|c'\gamma_n|)$, and $\hat{B} = C|c'(\hat{\gamma}-\gamma_n)|+C\sup_i|\Delta_i|$. Note that $\|c'S_nz_i/\sqrt{n}\|$ must be bounded since $S_n/\sqrt{n}$ has bounded eigenvalues and $\|c\| \leq 1$. $\hat{B} = o_p(1)$ as shown previously. Also, $\hat{A} = O_p(1)$ since $Var(c'X_i)$ is bounded above, $Var(\varepsilon_i^2) > 0$ and γ_n is a vector of regression coefficients, so that $Var(c'\gamma_n\varepsilon_i) = (c'\gamma_n)^2Var(\varepsilon_i^2) \leq Var(c'X_i)$.

and $\hat{B} = o_p(1)$. Using these results we have

$$\begin{aligned}
|c'X_i| &\leq d_i, \quad |c'\dot{X}_i| \leq \hat{A}d_i/\sqrt{r_n}, \quad |\check{X}_i| \leq \hat{A}d_i/\sqrt{r_n} \\
|\hat{\varepsilon}_i| &\leq |\varepsilon_i - \Delta_i| \leq \hat{A}d_i \\
|\hat{\varepsilon}_i^2 - \varepsilon_i^2| &\leq \hat{A}\hat{B}d_i^2 \\
|c'(\check{X}_i\check{X}_i' - \dot{X}_i\dot{X}_i')c| &\leq (|c'\check{X}_i| + |c'\dot{X}_i|)|c'(\check{X}_i - \dot{X}_i)| \\
&\leq \frac{1}{\sqrt{r_n}}\hat{A}d_i \times c'S_n^{-1}\big(\hat{\gamma}(\hat{\varepsilon}_i - \varepsilon_i) + \varepsilon_i(\hat{\gamma} - \gamma_n)\big) \\
&\leq \frac{1}{r_n}\hat{A}^2\hat{B}d_i^2 \\
|c'\dot{X}_i\varepsilon_i| &\leq \hat{A}d_i^2/\sqrt{r_n}, \quad |\check{X}_i\hat{\varepsilon}_i| \leq \hat{A}^2d_i^2/\sqrt{r_n} \\
|c'(\check{X}_i\hat{\varepsilon}_i - \dot{X}_i\varepsilon_i)| &\leq \frac{C}{\sqrt{r_n}}\big(|c'X_i||\hat{\varepsilon}_i - \varepsilon_i| + |c'\hat{\gamma}||\hat{\varepsilon}_i^2 - \varepsilon_i^2| \\
&\quad + |\varepsilon_i^2||c'(\hat{\gamma} - \gamma_n)|\big) \\
&\leq \frac{1}{\sqrt{r_n}}\hat{A}^2\hat{B}d_i^2
\end{aligned}$$

Finally, we also have

$$E\big[\frac{1}{r_n}\sum_{i\neq j}P_{ij}^2 d_i^2 d_j^2\big] \leq C\frac{1}{r_n}\sum_{i\neq j}P_{ij}^2 \leq C\frac{K_n}{r_n}$$

so that, a Markov inequality gives $\frac{1}{r_n}\sum_{i\neq j}P_{ij}^2 d_i^2 d_j^2 = O_p(1)$.

We next decompose the deviation into

$$\begin{aligned}\hat{\Sigma}_2 - \dot{\Sigma}_2 &= \sum_{i\neq j}P_{ij}^2\big(\check{X}_i\check{X}_i'\hat{\varepsilon}_j^2 - \dot{X}_i\dot{X}_i'\varepsilon_j^2\big)\\ &\quad+\sum_{i\neq j}P_{ij}^2\big(\check{X}_i\hat{\varepsilon}_i\check{X}_j'\hat{\varepsilon}_j - \dot{X}_i\varepsilon_i\dot{X}_j'\varepsilon_j\big)\\ &= T_1 + T_2\end{aligned}$$

Beginning with T_1

$$\begin{aligned}|\sum_{i\neq j}P_{ij}^2 c'\big(\check{X}_i\check{X}_i'\hat{\varepsilon}_j^2 - \dot{X}_i\dot{X}_i'\varepsilon_j^2\big)c| &\leq |\sum_{i\neq j}P_{ij}^2 c'\big(\check{X}_i\check{X}_i' - \dot{X}_i\dot{X}_i'\big)c\varepsilon_j^2|\\ &\quad+|\sum_{i\neq j}P_{ij}^2 c'\check{X}_i\check{X}_i'c(\hat{\varepsilon}_j^2-\varepsilon_j^2)|\\ &\leq \frac{1}{r_n}\hat{A}^2\hat{B}|\sum_{i\neq j}P_{ij}^2 d_i^2 d_j^2|\\ &\quad+\frac{1}{r_n}\hat{A}^3\hat{B}|\sum_{i\neq j}P_{ij}^2 d_i^2 d_j^2|\\ &= o_p(1)\end{aligned}$$

Next, for T_2 we have

$$\begin{aligned}&|\sum_{i\neq j}P_{ij}^2 c'\big(\check{X}_i\hat{\varepsilon}_i\check{X}_j'\hat{\varepsilon}_j - \dot{X}_i\varepsilon_i\dot{X}_j'\varepsilon_j\big)c|\\ &\qquad\leq |\sum_{i\neq j}P_{ij}^2(c'\check{X}_i\hat{\varepsilon}_i - c'\dot{X}_i\varepsilon_i)(\check{X}_j'\hat{\varepsilon}_j c + \dot{X}_j'\varepsilon_j c)|\\ &\qquad\leq \frac{1}{r_n}\hat{A}^4\hat{B}|\sum_{i\neq j}P_{ij}^2 d_i^2 d_j^2|\\ &\qquad= o_p(1)\end{aligned}$$

Proof of $c'(\hat{\Sigma}_1 - \dot{\Sigma}_1)c \to 0$

We first decompose $\hat{\Sigma}_1 - \dot{\Sigma}_1$ into seven terms, T_1 to T_7.

$$\begin{aligned}
\hat{\Sigma}_1 - \dot{\Sigma}_1 = & \sum_{i\neq j\neq k} P_{ik}P_{jk}(\check{X}_i - \dot{X}_i)(\check{X}_j - \dot{X}_j)'(\hat{\varepsilon}_k^2 - \varepsilon_k^2) \\
& + \sum_{i\neq j\neq k} P_{ik}P_{jk}\dot{X}_i(\check{X}_j - \dot{X}_j)'(\hat{\varepsilon}_k^2 - \varepsilon_k^2) \\
& + \sum_{i\neq j\neq k} P_{ik}P_{jk}(\check{X}_i - \dot{X}_i)\dot{X}_j'(\hat{\varepsilon}_k^2 - \varepsilon_k^2) \\
& + \sum_{i\neq j\neq k} P_{ik}P_{jk}(\check{X}_i - \dot{X}_i)(\check{X}_j - \dot{X}_j)'\varepsilon_k^2 \\
& + \sum_{i\neq j\neq k} P_{ik}P_{jk}\dot{X}_i\dot{X}_j'(\hat{\varepsilon}_k^2 - \varepsilon_k^2) \\
& + \sum_{i\neq j\neq k} P_{ik}P_{jk}\dot{X}_i(\check{X}_j - \dot{X}_j)'\varepsilon_k^2 \\
& + \sum_{i\neq j\neq k} P_{ik}P_{jk}(\check{X}_i - \dot{X}_i)\dot{X}_j'\varepsilon_k^2
\end{aligned}$$

Firstly, $c'T_2c = c'T_3c$ may be written as

$$\begin{aligned}
& |c' \sum_{i\neq j\neq k} P_{ik}P_{jk}\dot{X}_i(\check{X}_j - \dot{X}_j)'(\hat{\varepsilon}_k^2 - \varepsilon_k^2)c| \\
& \quad \leq \sup_i|2\Delta_i| \cdot |\sum_{i\neq j\neq k} P_{ik}P_{jk}c'\dot{X}_i\big(c'S_n^{-1}\hat{\gamma}X_j'(\hat{\beta} - \beta)\big)\varepsilon_k| \\
& \quad + \sup_i|\Delta_i^2| \cdot |\sum_{i\neq j\neq k} P_{ik}P_{jk}c'\dot{X}_i\big(c'S_n^{-1}\hat{\gamma}X_j'(\hat{\beta} - \beta)\big)| \\
& \quad \leq \|\hat{\gamma}\| \cdot \|\hat{\beta} - \beta\|\Big\{ \sup_i|2\Delta_i| \cdot |\sum_{i\neq j\neq k} P_{ik}P_{jk}c'\dot{X}_ic'S_n^{-1}X_j\varepsilon_k| \\
& \quad + \sup_i|\Delta_i^2| \cdot |\sum_{i\neq j\neq k} P_{ik}P_{jk}c'\dot{X}_ic'S_n^{-1}X_j|\Big\}
\end{aligned}$$

From Lemma A4 of Chao et al (2009), the sums are each $O_p(1)$. Recall that $\|\hat{\gamma}\| = O_p(J_n^{1/2})$ and that $\|\hat{\beta} - \beta\| = o_p(J_n^{-1/2})$. Then, $\|\hat{\gamma}\| \cdot \|\hat{\beta} - \beta\| \to 0$, and hence $c'T_2c = c'T_3c =$

$o_p(1)$. For $c'T_4c$ we have

$$
\begin{aligned}
&|c' \sum_{i\neq j\neq k} P_{ik}P_{jk}(\check{X}_i - \dot{X}_i)(\check{X}_j - \dot{X}_j)'\varepsilon_k^2 c| \\
&\qquad \leq |\sum_{i\neq j\neq k} P_{ik}P_{jk}\big(c'S_n^{-1}\hat{\gamma}X_i'(\hat{\beta}-\beta)\big)\big(c'S_n^{-1}\hat{\gamma}X_j'(\hat{\beta}-\beta)\big)\varepsilon_k^2| \\
&\qquad + 2|\sum_{i\neq j\neq k} P_{ik}P_{jk}c'S_n^{-1}\hat{\gamma}X_i'(\hat{\beta}-\beta)\varepsilon_j(\hat{\gamma}-\gamma_n)'(S_n^{-1})'c\varepsilon_k^2| \\
&\qquad + |\sum_{i\neq j\neq k} P_{ik}P_{jk}\varepsilon_i\varepsilon_j\big(c'S_n^{-1}(\hat{\gamma}-\gamma_n)\big)^2\varepsilon_k^2| \\
&\qquad \leq \|\hat{\gamma}\|^2\cdot\|\hat{\beta}-\beta\|^2|\sum_{i\neq j\neq k} P_{ik}P_{jk}\big(c'S_n^{-1}X_i\big)\big(c'S_n^{-1}X_j'\big)\varepsilon_k^2| \\
&\qquad + 2\|\hat{\gamma}\|\cdot\|\hat{\beta}-\beta\||c'(\hat{\gamma}-\gamma_n)|\cdot|\frac{1}{\sqrt{r_n}}\sum_{i\neq j\neq k} P_{ik}P_{jk}c'S_n^{-1}X_i\varepsilon_j\varepsilon_k^2| \\
&\qquad + |c'(\hat{\gamma}-\gamma_n)|^2\cdot|\frac{1}{\sqrt{r_n}}\sum_{i\neq j\neq k} P_{ik}P_{jk}\varepsilon_i\varepsilon_j\varepsilon_k^2|
\end{aligned}
$$

Applying Lemma C.6 we find that each of the sums are $O_p(1)$ and hence $c'T_4c \to 0$. Similarly, for $c'T_5c$

$$
\begin{aligned}
&|\sum_{i\neq j\neq k} P_{ik}P_{jk}c'\dot{X}_i\dot{X}_j'c(\hat{\varepsilon}_k^2 - \varepsilon_k^2)| \\
&\qquad \leq 2\sup_i|\Delta_i|\cdot|\sum_{i\neq j\neq k} P_{ik}P_{jk}\dot{X}_i\dot{X}_j'\varepsilon_k| \\
&\qquad\quad + 2\sup_i|\Delta_i^2|\cdot|\sum_{i\neq j\neq k} P_{ik}P_{jk}\dot{X}_i\dot{X}_j'|
\end{aligned}
$$

Applying Lemma C.6 we find that both sums are $O_p(1)$ and hence $c'T_5c \to 0$.

For $c'T_6c = c'T_7c$ we have

$$
\begin{aligned}
&|c' \sum_{i\neq j\neq k} P_{ik}P_{jk}(\check{X}_i - \dot{X}_i)\dot{X}_j'\varepsilon_k^2 c| \\
&\leq |\sum_{i\neq j\neq k} P_{ik}P_{jk}c'S_n^{-1}\big(\hat{\gamma}X_i'(\hat{\beta}-\beta) + \varepsilon_i(\hat{\gamma}-\gamma_n)\big)\dot{X}_j'c\varepsilon_k^2| \\
&\leq \|\hat{\gamma}\|\cdot\|\hat{\beta}-\beta\||\sum_{i\neq j\neq k} P_{ik}P_{jk}c'S_n^{-1}X_i\dot{X}_j'c\varepsilon_k^2|
\end{aligned}
$$

$$+ |c'(\hat{\gamma} - \gamma_n)| \cdot |\frac{1}{\sqrt{r_n}} \sum_{i \neq j \neq k} P_{ik} P_{jk} \varepsilon_i \dot{X}_j' c \varepsilon_k^2|$$

Now, by Lemma A4 in Chao et al (2009) we have that the sums are $O_p(1)$ and hence $c'T_6c = c'T_7c = o_p(1)$. Finally, we may show $c'T_1c = o_p(1)$ in the same way as the above, giving the main result.

Proof of $c'(\hat{\Sigma}_1 + \hat{\Sigma}_2)c - c'\Xi_n c \to 0$

We have from the previous subsections that $c'(\hat{\Sigma}_1 + \hat{\Sigma}_2)c - c'(\dot{\Sigma}_1 + \dot{\Sigma}_2)c \to 0$. The remainder of the proof is now identical to that for the JIVE variance, replacing X_i with $\dot{X}_i = X_i - \gamma_n \varepsilon_i$ and V_i with $\tilde{V}_i = V_i - \delta_n \varepsilon_i$.