Essays in Econometrics

Riccardo D'Adamo

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
of
University College London.

Department of Economics
University College London

July 27, 2023

I, Riccardo D'Adamo, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Date: July 27, 2023

Abstract

This thesis presents new methodologies in the field of Econometrics and their application to microeconomic data. In Chapter 2, I develop a framework for estimation of optimal individualized treatment rules in the presence of partial identification. I propose an estimation procedure that ensures Neyman-orthogonality with respect to nuisance components and provide statistical guarantees for its performance. The approach is illustrated using data from the Job Partnership Training Act Study to estimate the optimal participation of workers in a job training programme. Chapter 3 presents a new instrumental variable (IV) estimator for nonlinear models with endogenous covariates. This estimator formalizes the idea that the IVs should be "excluded variables" that have no direct explanatory power for the outcome, and does not require to specify the distribution of the endogenous covariates. The theoretical properties are explored through asymptotic theory and Monte Carlo simulations, and the method is illustrated with two empirical applications. Chapter 4 develops inference methods for linear regression models with many controls and clustering. I show that commonly used cluster-robust standard errors are inconsistent when the number of controls grows proportionally with the sample size. I then propose a new standard error formula that allows to carry out valid inference in high-dimensional regression models. Monte Carlo evidence supports the theoretical results and the proposed method is illustrated with an empirical application that studies the impact of abortion on crime.

Impact Statement

This dissertation develops new econometric methods aimed at improving the robustness of procedures for programme evaluation and data-driven decision making.

Chapter 2 presents a novel framework for estimation of individualized treatment rules. By allowing for partial identification of treatment effects, we can improve the robustness of treatment decisions based on widely available observational data, as well as experimental data suffering from noncompliance, attrition or missing observations. The proposed framework therefore enhances the opportunity for credible data-driven decision making in public policy, medicine and industry settings.

Chapter 3 proposes a new instrumental variable estimator for nonlinear models, including binary, censored and count outcome variables. This estimator leads to a statistical test of relevance of the endogenous regressor which does not rely on parametric assumptions for the distribution of endogenous variables, which are typically not justified by economic theory. Furthermore, it will typically correctly estimate the sign of the coefficient on the endogenous variable. The proposed method therefore provides a useful tool for programme evaluation, where it is often a primary concern whether the effect of an endogenous treatment is different from zero, and what the sign of such effect is.

Chapter 4 provides a new tool for cluster-robust inference in linear regression models with many controls. The proposed methodology will find fruitful application in programme evaluation, where linear regression is routinely used to estimate the effect of a treatment of interest. Researchers often include a large set of covariates to control for observed and unobserved confounders. As a result, the proposed

method can be used for empirical economic research, where cluster-robust standard errors are routinely used to account for potential dependence across units.

Acknowledgements

I am deeply grateful to my supervisor, Prof. Martin Weidner, for his extensive support and academic guidance throughout my studies at UCL. This thesis would not have been written without his inspiring teaching, which motivated me to pursue a PhD, as well as the insightful research discussions and frequent feedback. Most importantly, he has provided me with a precious role model of generous scholar and mentor, which I can aspire to for the rest of my career.

I would like to thank my secondary supervisor, Prof. Toru Kitagawa, for introducing me to statistical decision theory and crucially enriching the academic atmosphere at UCL with the organisation of many reading groups. His thoughtful comments have considerably improved the chapters of this dissertation.

I am deeply indebted to Prof. Whitney Newey for taking an interest in my research, hosting me at MIT, and always being generous with his time. I would also like to thank the econometrics group at MIT for making me feel welcome during my visit despite the challenges brought by the pandemic.

This thesis would not have been written without the input of the many exceptional teachers I have been fortunate to meet throughout my academic journey: Franco Peracchi, Tommaso Proietti, Eleni Katirtzoglou, Whitney Newey, Alberto Abadie, Philippe Rigollet, Alexander Rakhlin, Guillaume Pouliot, Bryan Graham, just to mention a few. Some of them may not know me personally, but they have had a lasting impact on my academic development.

I am immensely grateful to the friends and colleagues who supported me through the difficult moments of the PhD and, most importantly, ensured that I leave London with many great memories: Enrico, Gherardo, Anusha, Morgane,

Nick, Caterina, Yannis, Anna, Andres, Hugo, Guanyi, Vittorio, Edoardo, Matteo, Emma, and many more whom I cannot mention here.

Finally, I would like to thank my parents for always believing in me, and my late grandfather Ulisse for his financial support and encouragement throughout my studies. Although he did not have much formal education, he had an inspiring thirst for knowledge. This thesis is dedicated to his memory.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Econometrics plays a crucial role in empirical research by providing rigorous methods for estimating causal relationships and making evidence-based policy recommendations. In particular, the "credibility revolution" in empirical economics has lead to the increasing popularity of econometric methods for programme evaluation with observational data. However, the validity of such methods typically relies on strong identification and/or functional-form assumptions, which are often not justified by economic theory. This thesis proposes new methodologies to improve the robustness of econometric procedures for programme evaluation and data-driven decision making with observational data.

Chapter 2 studies the problem of choosing an optimal treatment assignment using data. Existing methods often rely on the assumption of point identification of treatment effects, which is not always justifiable in many empirical settings. In this chapter, I extend the framework of empirical welfare maximization (EWM) to handle partial identification of treatment effects. I first introduce ambiguity-robust policies that provide a notion of optimal treatment assignment under partial identification, accommodating different attitudes towards ambiguity. I then propose a procedure for estimating the ambiguity-robust optimal policy and provide theoretical guarantees on its statistical performance. The proposed methodology accommodates the use of machine learning algorithms for estimation of nuisance components, and achieves fast rates of convergence thanks to Neyman-orthogonalization. Finally, I apply the method to experimental data from the Job Training Partnership

Act study to estimate the optimal participation of workers in a job-training programme in the presence of non-compliance.

Chapter 3, co-authored with Martin Weidner and Frank Windmeijer, focuses on instrumental variables (IVs), a powerful tool for estimating causal relationships in models with endogeneity. We introduce the auxiliary IV (AIV) estimator, which generalizes the classical IV estimation approach to models with nonlinear relationships between the outcome and covariates, such as the probit regression model. The AIV estimator is obtained through maximum likelihood estimation by including the instruments as auxiliary regressors. Despite its potential inconsistency, we demonstrate the usefulness of the AIV estimator for carrying out inference on the presence of treatment effects (and their sign) under minimal assumptions on the data-generating process for the endogenous regressors. We provide formal results on the properties of the AIV estimator through asymptotic theory and illustrate its use with two empirical applications.

In Chapter 4, I study the problem of inference on treatment effects in linear regression models with many controls and clustering. In particular, I show that the conventional cluster-robust standard errors by Liang and Zeger (1986) are generally invalid when the number of controls is a non-negligible fraction of the sample size. I then propose a new clustered standard errors formula which is robust to the inclusion of many controls, enabling valid inference in high-dimensional linear regression models, including fixed effects panel data models and the semiparametric partially linear model. The theoretical results are supported by Monte Carlo simulations, illustrating the favourable performance of the proposed standard errors formula in finite samples. Finally, I illustrate the proposed method through an empirical application that re-examines Donohue and Levitt's (2001) study of the impact of abortion on crime.

Chapter 2

Orthogonal Policy Learning Under Ambiguity

The problem of choosing an optimal treatment assignment based on data is ubiquitous in economics and other fields, including medicine and marketing. Individuals often display heterogeneous responses to the same treatment. Decision-makers in policy and industry are therefore interested in leveraging the growing availability of rich granular data to tailor treatment assignment to individuals based on their characteristics. As a result, a fast-growing literature has emerged focused on developing procedures for estimation of individualized treatment rules. While a variety of approaches have been recently established, these typically assume that the available data allow to provide credible point estimates for the effect of the treatment, that is treatment effects are point identified. While of important stylized value, this assumption is often hard to justify in many empirical settings. For example, economists have long been aware that popular quasi-experimental and observational research designs, such as instrumental variables (IVs), allow to point identify treatment effects only for specific sub-populations (Imbens and Angrist, 1994). Even in randomized control trials, point identification of the treatment effects is often precluded due to non-random attrition, e.g. when participants dropout from a program or the researcher is denied information on the outcome variable (Lee, 2009). In such settings, the data may only provide partial knowledge about the treatment response in the form of credible bounds, i.e. the treatment effects are *partially iden-*

tified. As a result, the decision-maker may have ambiguous evidence on whether a candidate policy should be preferred to another, so that only a partial ordering of policies can be deduced in general. While informative from a scientific perspective, a partial ordering of policies is unsatisfying when the ultimate goal of the analysis is to select a single policy to be implemented in the real world. In this scenario, a decision-maker has to confront two sources of ambiguity. The first source concerns ambiguous knowledge of the treatment response τ conditional on knowledge of distribution of the data P, due to partial identification. The second source is the lack of knowledge of the distribution P, which must be estimated from the data.

In this chapter, we develop methods to handle both sources of ambiguity within the framework of "empirical welfare maximization" (Kitagawa and Tetenov, 2018), also referred to as "policy learning" (Athey and Wager, 2021). This approach considers treatment policies that are exogenously constrained to have low complexity in terms of Vapnik-Chervonenkis (VC) dimension. This encompasses many practical settings of interest, as policies often have to satisfy requirements imposed for institutional or practical reasons, such as fairness, budget or interpretability. The empirical welfare maximization (EWM) method selects the optimal policy as the maximizer of the empirical analogue of the population welfare, formulated as the average of the individual outcomes in the target population. The EWM estimation procedure has the convenient structure of an empirical risk minimization problem, which is exploited by Kitagawa and Tetenov (2018) and Athey and Wager (2021) to study its statistical properties.

We extend the EWM framework to settings with partial identification by making several contributions. First, we study the problem of assigning treatment under partial identification at the population level (i.e. where the distribution of the data P is known) from a general perspective. In particular, we show how classic optimality criteria for decision under ambiguity, such as minimax risk and minimax regret, can be applied in the context of welfare maximization. Our unified framework accommodates different attitudes towards ambiguity and a wide range of popular identification assumptions, including Manski (1990) and Manski and Pepper

(2000) bounds. Our analysis delivers several notions of optimal treatment policies, which we refer to as *ambiguity-robust*: they are "robust" in the sense that each of them delivers a notion of single optimal policy in the presence of partial identification, while they all reduce to the same optimal treatment assignment in the special case of point-identification. As part of this analysis, we establish general conditions on the identification sets under which the treatment assignment problem can be expressed in a simplified form, leading to computationally tractable sample analogues. In particular, we show that all ambiguity-robust policies can be represented as maximizers of a "surrogate" welfare, in which identification bounds are combined to form a proxy for the partially identified CATE. The surrogate welfare depends on several nuisance components, and its specific form is determined by the identification assumptions and attitude towards ambiguity held by the decision-maker.

We then propose an algorithm for computing the estimated ambiguity-robust policy and provide statistical guarantees on its performance in terms of the regret convergence of the surrogate welfare. Similarly to Athey and Wager (2021) and Foster and Syrgkanis (2019), our procedure leverages insights from the literature on double/de-biased machine learning (Chernozhukov et al., 2022) by making use of Neyman-orthogonalized estimates of the surrogate welfare. This, coupled with sample-splitting, allows us to guarantee fast rates of convergence for the estimated ambiguity-robust optimal policy while imposing minimal requirements on the estimation of the nuisance components. One unique feature of the partially identified setting studied in this chapter is the restricted degree of smoothness enjoyed by the welfare criterion. In particular, we show that popular choices of identification assumptions and optimality criteria for choice under ambiguity lead to surrogate welfare criteria that are only *directionally differentiable* with respect to the data-generating process. We highlight the importance of this feature for the problem at hand and develop new theoretical results showing how the extent of non-differentiability in the data-generating process affects the statistical properties of the learning procedure. To the best of our knowledge, we are the first to investigate the role of non-differentiabilities in the context of semiparametric statistical learn-

ing problems. Our results are therefore of independent interest and may be relevant beyond the treatment assignment problem presented here.

Finally, we apply the proposed method to experimental data from the Job Training Partnership Act study, a dataset that has been extensively used to study the effect of subsidized job training on labor market outcomes. We study the optimal participation of workers into the job training programme based on their education and previous earnings, and show that allowing for partial identification delivers substantially different programme participation policies compared to existing methods that assume point-identification.

2.0.1 Related literature

The results of this chapter contribute to the recent literature on EWM methods, e.g. Kitagawa and Tetenov (2018), Athey and Wager (2021), Mbakop and Tabord-Meehan (2021), Viviano (2019), Sun (2021), and more broadly to the literature studying statistical treatment choice, including Manski (2004), Dehejia (2005), Hirano and Porter (2009), Stoye (2009), Chamberlain (2011), Christensen et al. (2022), Kitagawa et al. (2022).[1]

Kitagawa and Tetenov (2018) introduced the EWM method and provided theoretical results showing its optimality when implemented with experimental data. Athey and Wager (2021) leverage insights from the recent literature on orthogonal machine learning (Chernozhukov et al., 2022) and propose doubly-robust estimation of the treatment effect which leads to optimal learning rates even with observational data. We build on their work by adopting Neyman-orthogonal estimates while we relax the fundamental assumption that treatment effects are point identified. Cui and Tchetgen (2021) also develop procedures for learning optimal treatments rules with instrumental variables but consider unconstrained policy classes. Similarly to Athey and Wager (2021), they ensure point-identification of treatment response by restricting their analysis to the effect on compliers.

Kasy (2016), Han (2019) and Byambadalai (2022) provide methods for comparing policies in the presence of covariates and partial identification of treatment

[1] See also Hirano and Porter (2020) and referenes therein.

effects. The focus of their work is on characterizing the partial ordering of policies in terms of their associated welfare rather than resolving the ambiguity and estimating an optimal treatment rule.

In a series of papers, Manski (2009, 2010, 2011) studies the problem of a social planner who must choose treatment for a population under partial knowledge of the treatment response in the absence of covariates. He shows that when the sign of the treatment effect is ambiguous, the minimax regret criterion leads to policies that randomize treatment in the population. While our study of the population problem is inspired by Manski's work in this area, the focus of our analysis is on deterministic rules assigning individualized treatment, i.e. based on (potentially continuous) covariates. Stoye (2012), Ishihara and Kitagawa (2021) and Yata (2021) consider treatment assignment under partial identification from a finite-sample minimax perspective, while Christensen et al. (2022) adopt a local-asymptotic approach. However, these works do not consider individualization of the treatment assignment.

More closely related to our work is Kallus and Zhou (2018), who extend the EWM framework to learn an optimal policy in the presence of partially identified treatment effects under violations of unconfoundedness. In particular, they target welfare improvement with respect to a baseline pre-existing policy and consider partial-identification of the welfare criterion within Rosenbaum's sensitivity model (Rosenbaum, 1987). Adjaho and Christensen (2022) and Kido (2022) examine policies with maximin welfare guarantees when the target population lies in a Wasserstein neighborhood of the experimental population. The identification assumptions (and associated estimation procedures) considered in these papers are distinct and do not nest those covered by our framework. As a result, our contributions are complementary to these works.

Russell (2020) considers estimation of the optimal policy under partial identification within a "probably approximately correct" learning framework (Valiant, 1984). His proposed procedure has the advantage of side-stepping direct estimation of the identified set, and can be applied in the context of incomplete models for which the identification bounds cannot typically be obtained in closed form. How-

ever, in the context of the identification assumptions considered in this chapter (e.g. Manski bounds), the theoretical results in Russell (2020) require that the covariates have discrete support. On the other hand, our proposed procedure requires computation of the identification bounds in closed form but accommodates continuous covariates.

In independent work, Pu and Zhang (2021) study policy learning under ambiguity from a classification perspective and derive an optimal policy which coincides with one notion of ambiguity-robust policy studied in this chapter. However, our estimation procedure crucially differs from theirs for the use of Neyman-orthogonalization which, combined with a refined proof-strategy that accounts for the lack of full-differentiability in the welfare criterion, allows us to guarantee considerably faster rates of convergence. In this sense, our results extend and improve those in Pu and Zhang (2021).

Finally, we contribute to a body of literature dealing with estimation and inference for directionally-differentiable functionals. Hirano and Porter (2012) show that if a target estimand is not differentiable in the parameters of the data distributions, then no asymptotically unbiased or regular estimator exists. Ponomarev (2022) studies efficient estimation of directionally differentiable functionals from a local minimax perspective. Fang and Santos (2018) and Kitagawa et al. (2020) provide inference results for directionally differentiable functions from a frequentist and Bayesian perspective, respectively. Also motivated by partial identification, Christensen et al. (2022) consider estimation of non-individualized decision rules when the welfare criterion is only directionally differentiable with respect to a finite-dimensional parameter. Our framework instead involves infinite-dimensional nuisance components and therefore our analysis must account for the lack of differentiability with novel theoretical results that complement those in Christensen et al. (2022). Our approach also differs from Christensen et al. (2022) in that we evaluate the statistical properties of estimation procedures in terms of maximum regret over (τ, P), while they consider maximum regret over the partially identified τ

conditional on P then averaged over the posterior distribution for P.[2]

The rest of this chapter is organized as follows. Section 2.1 introduces the setup. Section 2.2 presents several notions of ambiguity-robust optimal policies. Section 2.3 presents the proposed estimation procedure for the ambiguity-robust optimal policy. Section 2.4 provides statistical guarantees for the estimated optimal policy. Section 2.5 presents an empirical illustration based on the Job Training Partnership Act Study. Section 2.6 concludes. Proofs and extensions are given in the Appendix.

Notation. Throughout the chapter, for $d \in \mathbb{N}$, let $\mathbb{R}^d$ denote the Euclidean space, with $\|\cdot\|_p$ and $\langle\cdot,\cdot\rangle$ being the usual ℓ_p-norm and inner product, respectively. For two vectors $x \in \mathbb{R}^p$ and $y \in \mathbb{R}^q$, $x \subset y$ means that x is a sub-vector of y. For a symmetric matrix A, $\lambda_{\max}(A)$ denotes its largest eigenvalue. Unless otherwise stated, the expectation $\mathbb{E}[\cdot]$, probability $\mathbb{P}(\cdot)$, and variance $\mathrm{Var}(\cdot)$ operators will be taken with respect to the underlying distribution of observables P. Given a random variable $Z \in \mathcal{Z}$ with $\mathcal{Z} \subseteq \mathbb{R}^d$, the associated probability measure P_Z, and a function $f : \mathcal{Z} \to \mathcal{W}$ with $\mathcal{W} \subseteq \mathbb{R}^q$, we define $\|f\|_{Lp(P_Z)} = \left(\mathbb{E}_{P_Z}\left[\|f(Z)\|_p^p\right]\right)^{1/p}$ for $p \in (0\,\infty)$. We extend this definition to $p = \infty$ in the natural way. For a sequence of real numbers x_n and y_n, $x_n = o(y_n)$ and $x_n = O(y_n)$ mean, respectively, that $x_n/y_n \to 0$ and $x_n \leq Cy_n$ for some constant C as $n \to \infty$. For real numbers a, b, $a \lesssim b$ means that there exists a constant C such that $a \leq Cb$. For a positive real number a, $\lfloor a \rfloor$ denotes its nearest smallest integer. The notation $\to_p$ denotes convergence in probability.

2.1 Setup

Let $Y_i \in \mathbb{R}$ be an outcome measuring utility, $D_i \in \{0, 1\}$ a binary treatment, $X_i \in \mathcal{X} \subseteq \mathbb{R}^{k_x}$ a set of pre-treatment covariates for an individual i from an i.i.d. population of interest. We use standard notation to define the potential outcomes

[2] A key advantage of Christensen et al.'s (2022) asymmetric treatment of the ambiguity in τ and P is the additional tractability, which allows them to characterize (asymptotically) optimal decision rules. On the other hand, our fully minimax approach with respect to (τ, P) typically only allows to characterize worst-case rates of convergence for specific estimation procedures.

$Y_i(0), Y_i(1)$. The conditional average treatment effect (CATE) $\tau : \mathcal{X} \to \mathbb{R}$ is then defined as

$$\tau(x) = y_1(x) - y_0(x), \quad y_d(x) = \mathbb{E}[Y_i(d)|X_i = x],\ d = 0, 1,$$

where the expectation is taken with respect to the distribution of the population, and we will henceforth suppress the i-subscript for convenience. The decision-maker (DM) is interested in choosing a deterministic treatment assignment rule (or policy) $\pi : \mathcal{X} \to \{0, 1\}$, which maps from the support of individual pre-treatment covariates to the binary decision "treat" ($\pi(x) = 1$) or "do not treat" ($\pi(x) = 0$). Following Manski (2004), we define the utilitarian social welfare associated with a policy π and a given configuration of the expected potential outcomes $y_0(\cdot), y_1(\cdot)$ as

$$\begin{aligned} W_{y_0,y_1}(\pi) &= \mathbb{E}_{P_X}\left[y_1(X) \cdot \pi(X) + y_0(X) \cdot (1 - \pi(X))\right] \\ &= \underbrace{\mathbb{E}_{P_X}\left[\pi(X) \cdot \tau(X)\right]}_{=: I_\tau(\pi)} + \mathbb{E}_{P_X}\left[y_0(X)\right], \end{aligned} \tag{2.1}$$

where $I_\tau(\pi)$ represents the average impact of policy π. The optimal policy for a given configuration of the CATE function is the one that maximizes the associated welfare:

$$\pi^* = \underset{\pi \in \Pi}{\operatorname{argmax}}\, W_{y_0,y_1}(\pi) = \underset{\pi \in \Pi}{\operatorname{argmax}}\, I_\tau(\pi), \tag{2.2}$$

where Π is a family of candidate policies.[3] The DM has knowledge of the CATE through the distribution $P \in \mathcal{P}$ of observable random variables W, where $(Y, D, X) \subseteq W$. In particular, we denote $\mathcal{T}(P)$ the set of plausible CATE functions associated with a certain distribution of observables. When the DM has perfect knowledge of P and $\mathcal{T}(P)$ is a singleton, i.e. τ is point-identified, she can obtain π^* by solving (2.2).

Suppose now that $\mathcal{T}(P)$ is a non-singleton set, i.e. τ is partially identified. In

[3] We will assume throughout that the maximization problem in (2.2) has at least one solution. If multiple solutions exist, the DM is assumed to arbitrarily pick π^* from the set of maximizers.

that case, even under perfect knowledge of P, there exists a set of plausible values for the impact $I_\tau(\pi)$ of a candidate policy π. Notice that partial identification of the CATE does not necessarily imply that the DM cannot obtain the optimal policy π^*. In particular, it is easy to see that under point-identification of the CATE one has $\pi^*(x) = \mathbb{1}\{\tau(x) \geq 0\}$ when the class of candidate policies Π is unrestricted, so that identification of the sign of the CATE is sufficient to obtain the optimal policy.[4] However, the unrestricted policy class has limited relevance in many practical settings. For example, the policy space Π may be exogenously constrained for institutional reasons, e.g. as policies may be required to satisfy specific requirements for budget, fairness or interpretability. While the DM may still hope that his specification of Π contains the first-best policy $\mathbb{1}\{\tau(X) \geq 0\}$, it is useful to interpret π^* as the "best-in-class" policy for the chosen class Π, when this does not contain the first-best. When Π is constrained, the DM is not able to obtain π^* in general without full knowledge of the CATE, although a partial ordering of policies can still be deduced (see Kasy, 2016; Han, 2019; Byambadalai, 2022).

Under partial identification, the DM therefore faces two sources of ambiguity. First, she does not know the distribution P. However, we assume that she has access to a random sample $(W_i)_{i=1,\dots,n}$ from which she can learn about P. Second, she does not have knowledge about τ within the identified-set $\mathcal{T}(P)$, even under perfect knowledge of P. The broad objective of this chapter is to provide a framework that allows the DM to handle both sources of ambiguity. We will approach the problem in two steps. First, we will study the decision problem faced by the DM under perfect knowledge of P. In particular, we will handle the ambiguity arising from partial identification of τ using well-known optimality criteria for decision under ambiguity. Each of the optimality criteria we consider will deliver a corresponding notion of optimal policy, which we call "ambiguity-robust". The ambiguity-robust optimal policy is a unique treatment assignment rule that is preferred to all other policies in Π according to preferences of the DM, and that coincides with the usual notion of optimal policy π^* in (2.2) in the special case of point identification of the

[4] Cui and Tchetgen (2021) study a case in which sole point-identification of the sign of the CATE via an instrumental variable allows to obtain the optimal policy.

CATE. In the next section, we study several notions of ambiguity-robust optimal policy.

In the second part of our analysis, we study how to handle the ambiguity in P by showing how the random sample $(W_i)_{i=1,\dots,n}$ can be used to obtain an estimate $\widehat{\pi}_n$ for the ambiguity-robust optimal policy. The estimation procedure and the associated statistical guarantees are presented in Section 2.3 and 2.4, respectively.

Remark 2.1.1. *Unrestricted policy classes may also be precluded for practical reasons related to the estimation of the optimal policy. For example, the researcher may need to condition on a large number of covariates X for identification of the treatment effects, but only be interested in assigning treatment based on a restricted set of the covariates $\widetilde{X} \subset X$ (e.g. because she may not observe the full set of covariates when assigning treatment to new individuals from the population). In that case, a practical way to side-step computation of an estimate for the lower-dimensional CATE, $\mathbb{E}[Y_i(1)|\widetilde{X}_i = \widetilde{x}] - \mathbb{E}[Y_i(0)|\widetilde{X}_i = \widetilde{x}]$, is to impose restrictions directly on the policy class Π and estimate the optimal policy based on the estimated higher-dimensional CATE via the sample analogue of* (2.2)*.*

2.2 Ambiguity-robust optimal policies

The study of decision under ambiguity has a long tradition in decision theory and has received considerable attention in the context of treatment assignment problems (see Manski, 2011, for a review). In this section we review some classical optimality criteria for decision under ambiguity and study how they can be applied in the context of the treatment assignment problem at hand, leading to several notions of ambiguity-robust optimal policy.

A well-known optimality criterion for decision under ambiguity is minimax risk (see, e.g. Wald, 1950). In the context of our treatment assignment problem we can interpret welfare as negative risk, and this criterion leads to the optimal *maximin welfare* policy

$$\pi^*_{\texttt{MMW}} = \operatorname*{argmax}_{\pi\in\Pi} \min_{(y_0,y_1)\in\mathcal{Y}(P)} W_{y_0,y_1}(\pi), \tag{2.3}$$

where $\mathcal{Y}(P)$ is the ambiguity set for $(y_0(\cdot), y_1(\cdot))$ identified from the distribution P of observables random variables. The optimal maximin welfare policy maximizes the lowest possible welfare under any configuration of the expected potential outcome functions in the identified set $\mathcal{Y}(P)$. An alternative application of minimax risk optimality in the context of treatment assignment is *maximin impact*, leading to the optimal policy

$$\pi^*_{\texttt{MMI}} = \underset{\pi\in\Pi}{\operatorname{argmax}} \min_{\tau\in\mathcal{T}(P)} I_\tau(\pi), \tag{2.4}$$

where $\mathcal{T}(P)$ denotes the ambiguity set for the CATE function. The optimal maximin impact policy maximizes the lowest possible impact under any configuration of the CATE in the identified set $\mathcal{T}(P)$. Notice that the minimax welfare criterion reflects an extreme degree of pessimism with regards to outcomes associated with both treatment and non-treatment scenarios; on the other hand, the minimax impact criterion reflects an extreme degree of pessimism with regards to the impact of the policy, thus directly raising the threshold for treatment.[5] Despite its intuitive appeal, minimax optimality has been criticised for being too conservative and often delivering decisions that are especially sensitive to changes in the ambiguity set.[6]

An alternative criterion that alleviates some of these concerns is *minimax regret*, with corresponding optimal policy

$$\begin{aligned}\pi^*_{\texttt{MMR}} &= \underset{\pi\in\Pi}{\operatorname{argmin}} \max_{(y_0,y_1)\in\mathcal{Y}(P)} \left[\left(\max_{\pi\,:\,\mathcal{X}\to\{0,1\}} W_{y_0,y_1}(\pi)\right) - W_{y_0,y_1}(\pi)\right] \\ &= \underset{\pi\in\Pi}{\operatorname{argmin}} \max_{\tau\in\mathcal{T}(P)} \left[\left(\max_{\pi\,:\,\mathcal{X}\to\{0,1\}} I_\tau(\pi)\right) - I_\tau(\pi)\right],\end{aligned} \tag{2.5}$$

The minimax regret criterion delivers a policy that minimizes the largest possible distance between attained welfare and the highest level of welfare attainable by the "oracle" treatment rule $\pi^* = \mathbb{I}\{\tau(x) \geq 0\}$ that has knowledge of the true τ. Minimax regret optimality has been advocated by Manski (2004) for its balanced

[5] In the empirical application of Section 2.5, both minimax welfare and minimax impact criteria result in $\pi(x) = 0$ for the entire population.

[6] In his classic textbook, Berger goes as far as saying that "In actually making decisions, the use of the minimax principle is definitely suspect." (Berger, 1985).

consideration of the possible states of nature and for delivering more "reasonable" decisions rules in practice, compared to minimax risk approaches.

Remark 2.2.1. *An alternative version of the minimax regret criterion is minimax regret with respect to the welfare attained by the best-in-class policy in* Π*, resulting in the objective*

$$\pi^*_{MMR2} = \underset{\pi\in\Pi}{\operatorname{argmin}} \max_{\tau\in\mathcal{T}(P)} \left[\left(\max_{\pi\in\Pi} I_\tau(\pi) \right) - I_\tau(\pi) \right]. \tag{2.6}$$

While these two versions of the minimax regret criterion can be expected to enjoy similar properties, the first version we have considered is considerably more tractable. In fact, the innermost maximization in (2.5) has the closed-form solution $\max_{\pi\,:\,\mathcal{X}\to\{0,1\}} I_\tau(\pi) = \mathbb{E}_{P_X}\left[\max\left\{\tau(X), 0\right\}\right]$*. As we show in Proposition 2.2.2 below, this allows to more explicitly characterize the properties of the optimization problem and the resulting optimal policy, as well as reduce the computational burden in solving the empirical analogue of the problem. For this reason we will focus on the version in (2.5) of the criterion. We also note that whenever the class* Π *is "well-specified", in the sense that* $\mathbb{I}\left\{\tau(x) \geq 0\right\} \in \Pi$ *for all* $\tau \in \mathcal{T}(P)$*, the two optimality criteria are equivalent.*

One critical drawback in the application of the optimality criteria just presented to the treatment assignment problem of this chapter is that the optimal policies cannot be obtained in closed form. This is due to the form of (2.3), (2.4) and (2.5) involving several nested optimizations whose solutions cannot be easily characterized at the current level of generality when X includes continuously distributed covariates and Π may be arbitrarily restricted, which are both primary cases of interest in our analysis. To make progress, we impose the following restrictions on the ambiguity sets for the expected potential outcomes and CATE.

Assumption 2.2.1 (Rectangular identified set for (y_0, y_1))**.** *The identified set for* (y_0, y_1) *is rectangular, that is,* $\mathcal{Y}$ *is of the form*

$$\mathcal{Y} = \{(y_0(\cdot), y_1(\cdot)) : (y_0(x), y_1(x)) \in \mathcal{Y}(x)\},$$

where $\mathcal{Y}(x)$ is a compact subset of $\mathbb{R}^2$.

Assumption 2.2.2 (Rectangular identified set for τ)**.** *The identified set for τ is rectangular, that is, $\mathcal{T}$ is of the form*

$$\mathcal{T} = \{\tau(\cdot) : \tau(x) \in [\underline{\tau}(x), \overline{\tau}(x)]\},$$

where $|\overline{\tau}(x)| < \infty$, $|\underline{\tau}(x)| < \infty$ for all $x \in \mathcal{X}$.

Assumptions 2.2.1 and 2.2.2 impose separation of the identified sets for the expected potential outcomes and CATE across the support of the covariates $\mathcal{X}$.[7] They are typically satisfied by identification schemes that do not impose shape restrictions on counterfactual outcomes with respect to the covariates X_i. These assumptions are widely adopted in the partial identification literature, and we refer the reader to Appendix B in Kasy (2016) for an extensive review of identification schemes that result in rectangular identified sets. Below we present three examples of identification schemes for the CATE that satisfy this assumption.

Example 2.2.1 (Manski bounds)**.** *Suppose there exists a binary instrument $Z_i \in \{0, 1\}$ that satisfies the well know exogeneity and exclusion restrictions $Y_i(0), Y_i(1), D_i(0), D_i(1) \perp Z_i|X_i$, where $Y_i(d)$ and $D_i(z)$ denote the counterfactual outcome and treatment functions, respectively. If the instrument Z_i also satisfies the overlap condition*

$$\eta \leq \mathbb{P}(Z_i = 1|X_i) \leq 1 - \eta, \quad \eta > 0,$$

and the monotonicity condition (also known as no-defiers condition):

$$\mathbb{P}\big(D_i(1) \leq D_i(0)|X_i\big) = 1 \quad \text{or} \quad \mathbb{P}\big(D_i(1) \geq D_i(0)|X_i\big) = 1,$$

then seminal work by Imbens and Angrist (1994) shows point-identification of the

[7] Notice that Assumption 2.2.1 implies Assumption 2.2.2, but not viceversa.

conditional local average treatment effect (LATE):

$$\mathbb{E}[Y_i(1) - Y_i(0) \mid D_i(1) \neq D_i(0),\, X_i = x].$$

Let us now assume that $Y \in [Y_L, Y_U]$, i.e. the outcome is bounded, and define

$$\begin{aligned}
h(z,x) &= \mathbb{E}[Y_i | Z_i = z, X_i = x], \\
m(d,z,x) &= \mathbb{E}[Y_i | D_i = d, Z_i = z, X_i = x], \\
p(z,x) &= \mathbb{P}(D_i = 1 | Z_i = z, X_i = x), \\
z(x) &= \mathbb{P}(Z_i = 1 | X_i = x).
\end{aligned}$$

The identified sets for the expected potential outcomes $y_0(x)$ and $y_1(x)$ are contained within the bounds

$$\begin{aligned}
\overline{y}_0(x) &= \min_{z \in \{0,1\}} \big\{ m(0,z,x) \cdot (1 - p(z,x)) + Y_U \cdot p(z,x) \big\}, \\
\underline{y}_0(x) &= \max_{z \in \{0,1\}} \big\{ m(0,z,x) \cdot (1 - p(z,x)) + Y_L \cdot p(z,x) \big\},
\end{aligned}$$

and

$$\begin{aligned}
\overline{y}_1(x) &= \min_{z \in \{0,1\}} \big\{ m(1,z,x) \cdot p(z,x) + Y_U \cdot (1 - p(z,x)) \big\}, \\
\underline{y}_1(x) &= \max_{z \in \{0,1\}} \big\{ m(1,z,x) \cdot p(z,x) + Y_L \cdot (1 - p(z,x)) \big\}.
\end{aligned}$$

The identified set for the CATE is then contained within the bounds

$$\begin{aligned}
\overline{\tau}(x) &= \overline{y}_1(x) - \underline{y}_0(x), \\
\underline{\tau}(x) &= \underline{y}_1(x) - \overline{y}_0(x).
\end{aligned}$$

If no further functional form assumption on the distribution of potential outcomes is made, these bounds are sharp (Heckman and Vytlacil, 2001) and the sharp identified sets for the average potential outcomes and CATE respectively satisfy Assumption 2.2.1 and Assumption 2.2.2.

Example 2.2.2 (Balke-Pearl)**.** *Suppose that the same assumptions as in Example 2.2.1 hold, and additionally the monoticity assumption is strengthened to*

$$\mathbb{P}\big(D_i(1) \geq D_i(0)|X_i\big) = 1,$$

that is, the direction of the monotonicity is known and positive. The bounds for the potential outcomes simplify to

$$\begin{aligned}
\overline{y}_0(x) &= m(0,0,x) \cdot (1 - p(0,x)) + Y_U \cdot p(0,x), \\
\underline{y}_0(x) &= m(0,0,x) \cdot (1 - p(0,x)) + Y_L \cdot p(0,x), \\
\overline{y}_1(x) &= m(1,1,x) \cdot p(1,x) + Y_U \cdot (1 - p(1,x), \\
\underline{y}_1(x) &= m(1,1,x) \cdot p(1,x) + Y_L \cdot (1 - p(1,x)),
\end{aligned}$$

and the CATE is contained within the bounds

$$\begin{aligned}
\overline{\tau}(x) &= h(1,x) - h(0,x) + p(0,x) \cdot \big(m(1,0,x) - Y_L\big) + (1 - p(1,x)) \cdot \big(Y_U - m(0,1,x)\big), \\
\underline{\tau}(x) &= h(1,x) - h(0,x) + p(0,x) \cdot \big(m(1,0,x) - Y_U\big) + (1 - p(1,x)) \cdot \big(Y_L - m(0,1,x)\big).
\end{aligned}$$

where $p(0,x)$ and $1 - p(1,x)$ identify the proportions of always-takers and never-takers at $X_i = x$, respectively. If no further functional form assumption on the distribution of outcomes for non-compliant populations is made, these bounds are sharp (Balke and Pearl, 1997) and the sharp identified sets for the average potential outcomes and CATE respectively satisfy Assumption 2.2.1 and Assumption 2.2.2.

Example 2.2.3 (Manski-Pepper bounds)**.** *Suppose that instead of full exogeneity, the instrumental variable Z_i satisfies the weaker "monotone IV" condition*

$$\mathbb{E}[Y_i(d)|Z_i = 0, X_i] \leq \mathbb{E}[Y_i(d)|Z_i = 1, X_i], \quad d = 0, 1. \tag{2.7}$$

Manski and Pepper (2000) show that when the outcome is bounded one has

$$\begin{aligned}
&\sum_{z=0,1} \mathbb{P}\left(Z_i = z|X_i\right) \\
&\qquad \times \max_{z_1 \leq z} \left\{m(d, z_1, X_i) \cdot \mathbb{P}(D_i = d|Z_i = z_1, X_i) + Y_L \cdot \mathbb{P}(D_i = 1 - d|Z_i = z_1, X_i)\right\} \\
&\qquad\qquad \leq \mathbb{E}[Y_i(d)|X_i] \leq \\
&\sum_{z=0,1} \mathbb{P}\left(Z_i = z|X_i\right) \\
&\qquad \times \min_{z_2 \geq z} \left\{m(d, z_2, X_i) \cdot \mathbb{P}(D_i = d|Z_i = z_2, X_i) + Y_L \cdot \mathbb{P}(D_i = 1 - d|Z_i = z_2, X_i)\right\}.
\end{aligned}$$

Upper (lower) bounds for the CATE are obtained by combining upper (lower) bounds for $\mathbb{E}[Y_i(1)|X_i = x]$ *with the lower (upper) bound for* $\mathbb{E}[Y_i(0)|X_i = x]$*:*

$$\begin{aligned}
\overline{\tau}(x) &= z(x) \cdot \psi_{1,1}(x; Y_U) + (1 - z(x)) \cdot \min\left\{\psi_{0,1}(x; Y_U), \psi_{1,1}(x; Y_U)\right\} \\
&\quad - z(x) \cdot \max\left\{\psi_{0,0}(x; Y_L), \psi_{1,0}(x; Y_L)\right\} - (1 - z(x)) \cdot \psi_{0,0}(x; Y_L), \\
\underline{\tau}(x) &= z(x) \cdot \max\left\{\psi_{0,1}(x; Y_L), \psi_{1,1}(x; Y_L)\right\} + (1 - z(x)) \cdot \psi_{0,1}(x; Y_L) \\
&\quad - z(x) \cdot \psi_{1,1}(x; Y_U) - (1 - z(x)) \cdot \min\left\{\psi_{0,0}(x; Y_U), \psi_{1,0}(x; Y_U)\right\}, .
\end{aligned}$$

where

$$\begin{aligned}
\psi_{z,d}\left(x; Y_{(\cdot)}\right) &= m(d, z, x) \cdot (d \cdot p(z, x) + (1 - d) \cdot (1 - p(z, x)) \\
&\quad + Y_{(\cdot)} \cdot (d \cdot (1 - p(z, x)) + (1 - d) \cdot p(z, x)).
\end{aligned}$$

Under no further assumption on the distribution of potential outcomes, these bounds are sharp (Manski and Pepper, 2000) and satisfy Assumptions 2.2.1 and 2.2.2.

Having restricted the identified sets $\mathcal{Y}$ and $\mathcal{T}$ as in Assumptions 2.2.1-2.2.2, we are now able to provide a simpler characterization of the maximin welfare and maximin impact policies.

Proposition 2.2.1. *Define* $\underline{y}_d(x) = \min_{y_d(x) \in \mathcal{Y}(x)} y_d(x)$ *and* $\overline{y}_d(x) = \max_{y_d(x) \in \mathcal{Y}(x)} y_d(x)$.

Under Assumption 2.2.1 the optimal maximin welfare policy is

$$\pi^*_{\texttt{MMW}} = \underset{\pi\in\Pi}{\operatorname{argmax}}\, \mathbb{E}_{P_X}\left[(2\pi(X)-1)\cdot(\underline{y}_1(X)-\underline{y}_0(X))\right]. \tag{2.8}$$

Furthermore, under Assumption 2.2.2 the optimal maximin impact policy is

$$\pi^*_{\texttt{MMI}} = \underset{\pi\in\Pi}{\operatorname{argmax}}\, \mathbb{E}_{P_X}\left[(2\pi(X)-1)\cdot\underline{\tau}(X)\right]. \tag{2.9}$$

Proposition 2.2.1 shows that the optimal maximin welfare and maximin impact policies maximize surrogate versions of the welfare substituting the unidentified CATE with the difference in the lower bounds of potential outcomes $\underline{y}_1(x)-\underline{y}_0(x)$ and the lower bound for CATE $\underline{\tau}(x)$ at every point in the covariate space, respectively. Notice that when $\mathcal{Y}(x)$ is rectangular with respect to the two potential outcomes, i.e. $y_0(x)\in\mathcal{Y}_0(x), y_1(x)\in\mathcal{Y}_1(x)$ and $\mathcal{Y}(x)=\mathcal{Y}_0(x)\times\mathcal{Y}_1(x)$, we have $\underline{\tau}(x)=\underline{y}_1(x)-\overline{y}_0(x)$, thus highlighting the "pessimistic" nature of the maximin impact criterion.

Remark 2.2.2. *The maximin welfare and maximin impact optimal policies coincide when $y_0(\cdot)$ is point-identified. This case is relevant when $y_0(x)$ represents the (conditional) average outcome under the status-quo in the entire population and is typically point-identified from observational data.*

The simplification of these two maximin problems into single maximisation problems has important advantages for the study of the optimal policies and their estimation from the data. In fact, the sample analogues of optimizations (2.8) and (2.9) are amenable to standard computation procedures for a variety of policy classes Π. Furthermore, their solution can be studied using tools for empirical risk minimisation problems, as discussed in Section 2.3.

Despite the involvement of an additional maximization problem compared to maximin welfare and maximin impact, Assumption 2.2.2 allows to provide a simpler characterization also for the minimax regret optimal policy.

Proposition 2.2.2. *Under Assumption 2.2.2 the optimal minimax welfare regret policy is*

$$\pi^*_{\texttt{MMR}} = \underset{\pi \in \Pi}{\operatorname{argmax}} \, \mathbb{E}_{P_X} \Big[(2\pi(X) - 1) \cdot \widetilde{\tau}(X) \Big] \tag{2.10}$$

where

$$\widetilde{\tau}(x) = \overline{\tau}(x) \cdot \mathbb{1}\big\{\overline{\tau}(x) \geq 0\big\} + \underline{\tau}(x) \cdot \mathbb{1}\big\{\underline{\tau}(x) \leq 0\big\} \tag{2.11}$$

This simpler characterization of the minimax regret problem as a single maximization sheds light on the properties of its associated optimal policy. In particular, we see that the objective function symmetrically treats individuals whose expected treatment effect sign is identified by assigning as surrogate for the CATE their outer bound, i.e. the CATE upper (lower) bound for individuals with identified positive (negative) sign for CATE. Individuals for which the sign of the treatment effect is ambiguous are assigned an intermediate point within their respective CATE bounds. The location of this intermediate point depends on the extent to which the identified set lies in the positive or negative region. Intuitively, the criterion prioritizes correct treatment allocation to individuals who unambiguously benefit from (or are harmed by) the treatment and down-weights the importance of individuals for which the sign of the treatment response is ambiguous within the treatment allocation problem. As an extreme case, individuals with CATE bounds exactly symmetric around 0 (i.e. $\overline{\tau}(x) = -\underline{\tau}(x)$) are given no consideration in the solution of the treatment allocation problem. This intuition can be further supported by noticing that the original welfare maximization under point-identification in (2.2) can be re-casted as the weighted classification problem

$$\pi^* = \underset{\pi \in \Pi}{\operatorname{argmin}} \, \mathbb{E}_{P_X} \Big[\mathbb{1}\big\{(2\pi(X) - 1) \neq \operatorname{sign}(\tau(X))\big\} \cdot |\tau(X)| \Big],$$

of which the minimax welfare regret optimal policy in (2.11) turns out to solve the

minimax version under Assumption 2.2.2:

$$\pi^*_{\text{MMR}} = \underset{\pi\in\Pi}{\operatorname{argmin}} \max_{\tau\in\mathcal{T}} \mathbb{E}_{P_X}\Big[\mathbb{1}\big\{(2\pi(X)-1) \neq \operatorname{sign}(\tau(X))\big\}\cdot|\tau(X)|\Big].$$

It is from this minimax classification risk perspective that Pu and Zhang (2021) obtain and study the minimax regret policy, which they call the "IV-optimal policy".

An alternative version of minimax regret optimality which has been used in the context of treatment choice is minimax regret with respect to a baseline policy. Kallus and Zhou (2018) assume the existence of a fixed policy π_{B} from which the DM does not want to unnecessarily deviate. They define the optimal policy as minimizing regret with respect to this baseline policy:

$$\begin{aligned}\pi^*_{\text{MMRB}} &= \underset{\pi\in\Pi}{\operatorname{argmin}} \max_{\tau\in\mathcal{T}} \{I_\tau(\pi_{\text{B}}) - I_\tau(\pi)\} \\ &= \underset{\pi\in\Pi}{\operatorname{argmax}}\, \mathbb{E}_{P_X}\Big[(2\pi(X)-1)\cdot(\overline{\tau}(X)\cdot\mathbb{1}\{\pi_{\text{B}}(X)\geq 0\} + \underline{\tau}(X)\cdot\mathbb{1}\{\pi_{\text{B}}(X)<0\})\Big],\end{aligned}$$

where the second equality uses Assumption 2.2.2. While potentially appealing in certain settings, e.g. when π_{B} represents the existing standard of care in a medical setting, this optimality criterion suffers the potential drawback of requiring the DM to specify (and motivate) the baseline policy for it to be operational. Adopting the never-treat baseline policy, i.e. $\pi_{\text{B}}(x) = 0,\ \forall x \in \mathcal{X}$, could be seen as an appealing "agnostic" choice, which however makes this criterion default to maximin impact and thus inherit its potentially undesirable properties.

The last notion of ambiguity-robust optimal policy that we present in this section is based on the Hurwicz criterion (Hurwicz, 1951), arguably one of the most widely used in decision-making under ambiguity. In the context of the treatment assignment problem at hand, the Hurwicz criterion leads to the ambiguity-robust

policy

$$\pi^*_{\texttt{HurW},\delta_0,\delta_1} = \underset{\pi\in\Pi}{\operatorname{argmax}}\, \mathbb{E}_{P_X}\Bigg[(2\pi(X)-1)\cdot\big(\{\delta_1\cdot\overline{y}_1(X)+(1-\delta_1)\cdot\underline{y}_1(X)\} - \{\delta_0\cdot\overline{y}_0(X)+(1-\delta_0)\cdot\underline{y}_0(X)\}\big)\Bigg],$$

where $\delta_1 \in [0,1]$ and $\delta_0 \in [0,1]$ are user-defined weights reflecting the degree of optimism with respect to the outcomes under treatment and non-treatment, respectively. It is easy to see that the maximin welfare criterion in (2.3) corresponds to the choice $\delta_1 = 0, \delta_0 = 0$. An analogous notion of optimality focused on impact rather than welfare, leads to the optimal policy

$$\pi^*_{\texttt{HurI},\delta} = \underset{\pi\in\Pi}{\operatorname{argmax}}\, \mathbb{E}_{P_X}\Big[(2\pi(X)-1)\cdot\big(\delta\cdot\overline{\tau}(X)+(1-\delta)\cdot\underline{\tau}(X)\big)\Big],$$

where $\delta \in [0,1]$ controls the degree of optimism with respect to the effect of treatment, with the maximin impact optimal policy corresponding to the choice $\delta = 0$. Under Assumption 2.2.1 and $\mathcal{Y}(x) = \mathcal{Y}_0(x) \times \mathcal{Y}_1(x)$, the Hurwicz Impact criterion is nested into the Hurwicz Welfare for the choice of parameters $\delta = \delta_0 = 1 - \delta_1$; unlike Hurwicz Welfare, however, the Hurwicz Impact criterion is still well-defined under the weaker Assumption 2.2.2. Interestingly, when $\delta = 1/2$ and Π is well-specified, in the sense that it contains the first-best assignment $\mathbb{1}\{\overline{\tau}(x)+\underline{\tau}(x) \geq 0\}$, we have

$$\pi^*_{\texttt{MMR}}(x) = \mathbb{1}\{\widetilde{\tau}(x) \geq 0\} = \mathbb{1}\{\overline{\tau}(x)+\underline{\tau}(x) \geq 0\} = \pi^*_{\texttt{HurI},\frac{1}{2}}(x).$$

Therefore the minimax regret and Hurwicz impact optimal policies coincide under correct specification of Π, as they assign treatment based on the middle point between the upper and lower CATE bounds. When Π is not well-specified, however, minimax regret optimality is not nested into any of the Hurwicz-type criteria just presented, thus highlighting the radically different attitude towards ambiguity implied by minimax regret compared to maximin welfare/impact. In particu-

lar, minimax regret is the only criterion of those presented (along with Hurwicz impact under $\delta = 1/2$) that treats symmetrically individuals with identified sets symmetric around 0, in the sense that $\widetilde{\tau}_1(x) = -\widetilde{\tau}_2(x)$ whenever $\overline{\tau}_1(x) = -\underline{\tau}_2(x)$ and $\underline{\tau}_1(x) = -\overline{\tau}_2(x)$. For this reason, minimax regret does not reflect an optimistic/pessimistic attitude towards ambiguity but rather an "opportunistic" one, in light of its prioritization of correct treatment assignment to individuals whose CATE sign is unambiguously identified.

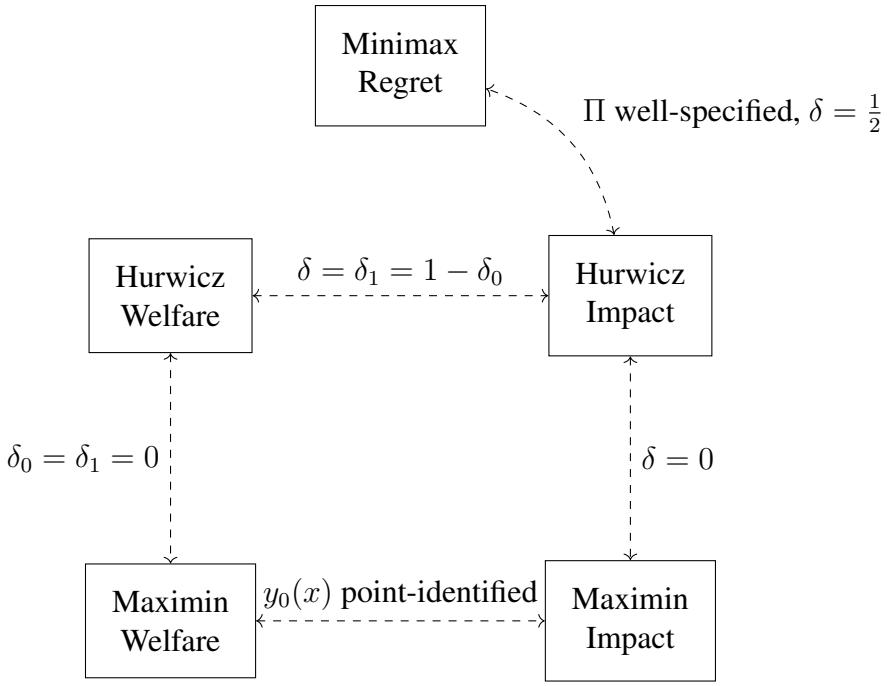


Figure 2.1: Relationship between ambiguity-robust optimal policies

2.2.1 A common framework

While accommodating a wide range of attitudes towards ambiguity, the notions of optimality presented in Section 2.4.2 share a common structure. In fact, by virtue of Assumptions 2.2.1 and 2.2.2, the corresponding optimal policies can all be written as

$$\pi^*(P) = \underset{\pi \in \Pi}{\operatorname{argmax}}\, Q(P;\pi), \qquad Q(P;\pi) := \mathbb{E}_{P_X}\Big[\,(2\pi(X) - 1) \cdot \Gamma(P;X)\Big], \tag{2.12}$$

for a specific score function[8] $\Gamma(P; \cdot)$, where we have highlighted the dependence of the score on the distribution P. The specific dependence on P is determined by the optimality criterion (as summarized in Table 2.1) as well as the identification assumptions (e.g. Balke-Pearl, Manski-Pepper etc.). This common structure also nests the point identified setting as the special case $\Gamma(P; X) = \tau(X)$ and thus suggests that existing estimation procedures for this special case can be extended to the partially identified setting.

However, one peculiar feature of the partially identified setting is the restricted degree of smoothness enjoyed by the objective function, in particular the differentiability of the scores with respect to P. Under point-identification of the CATE via standard unconfoundedness assumptions, one has $\Gamma(P; x) = \mathbb{E}[Y|D = 1, X = x] - \mathbb{E}[Y|D = 0, X = x]$ and the full differentiability of the score with respect to the expectation $\mathbb{E}[Y|D, X]$ is immediately apparent. However, for the minimax regret criterion we notice that the score is *directionally differentiable*[9] with respect to P at $\overline{\tau}(x) = 0$ or $\underline{\tau}(x) = 0$. Even when $\Gamma(P; x)$ depends smoothly on expected outcomes/CATE bounds, lack of full differentiability of the scores can arise through a lack of differentiability of the expected outcomes and CATE bounds themselves. In fact, many popular identification assumptions, including the Manski and Manski-Pepper bounds from Examples 2.2.1 and 2.2.3, deliver bounds that are only directionally differentiable with respect to identified parameters due to the presence of $\min / \max$ operators (see Chernozhukov et al., 2013, and examples therein). Whether a consequence of the optimality criterion or the identification assumptions, lack of full differentiability of the scores is a unique and pervasive feature of the treatment assignment problem under partial identification, one that has not been explicitly acknowledged in the most recent contributions in this area,

[8] The term 'score function' is borrowed from Athey and Wager (2021).

[9] Let $P \in \mathcal{P}$ be a probability distribution on which the function $f : \mathcal{P} \to \mathbb{R}$ depends. We say that f is directionally differentiable at P_0 if the limit

$$\lim_{t \downarrow 0} \frac{f(P_0 + t(h - P_0)) - f(P_0)}{t} = \dot{f}_{P_0}[h]$$

exists for every $h \in \mathcal{P}$, in which case $\dot{f}_{P_0}[\cdot]$ denotes the directional derivative of f at P_0. If it exists, the directional derivative $\dot{f}_{P_0}[\cdot]$ is positively homogeneous of degree one but not necessarily linear. If $\dot{f}_{P_0}[\cdot]$ is linear then f is fully differentiable at P_0.

with the notable exception of Christensen et al. (2022). A major contribution of this chapter is to account for the role played by the lack of full differentiability when we establish procedures for estimating ambiguity-robust optimal policies in Section 2.3.

Table 2.1: Optimality criteria and associated scores

Optimality criterion	$\Gamma(P; x)$
Maximin Welfare	$\underline{y}_1(x) - \underline{y}_0(x)$
Maximin Impact	$\underline{\tau}(x)$
Minimax Regret (oracle)	$\overline{\tau}(x) \cdot \mathbb{1}\{\overline{\tau}(x) \geq 0\} + \underline{\tau}(x) \cdot \mathbb{1}\{\underline{\tau}(x) \leq 0\}$
Minimax Regret (baseline)	$\overline{\tau}(x) \cdot \mathbb{1}\{\pi_{\text{B}}(x) = 1\} + \underline{\tau}(x) \cdot \{\pi_{\text{B}}(x) = 0\}$
Hurwicz (welfare)	$\big(\{\delta_1 \cdot \overline{y}_1(x) + (1-\delta_1) \cdot \underline{y}_1(x)\} - \{\delta_0 \cdot \overline{y}_0(x) + (1-\delta_0) \cdot \underline{y}_0(x)\}\big), \quad \delta_1, \delta_0 \in [0,1]$
Hurwicz (impact)	$\delta \cdot \overline{\tau}(x) + (1-\delta) \cdot \underline{\tau}(x), \quad \delta \in [0,1]$

2.3 Estimation

In this section we present the statistical framework underlying the problem of estimation of optimal treatment rules under partial identification. We will discuss heuristics underlying several features of the estimation problem, and then present our proposed estimation procedure.

We work in a learning setting where the estimand $\pi^*(P)$ is as in (2.12), and we observe an i.i.d. sample $(W_i)_{i=1,\dots,n}$ of size n from the unknown distribution P of the observed random variables $W \in \mathcal{W}$, $X \subset W$. To retain generality of the framework, we do not specify the exact dependence of the functional $\Gamma(P; x)$ on P, which will depend on the choice of optimality criterion for the resolution of ambiguity (maximin welfare, minimax regret etc.) and identification assumptions determining the identification sets $\mathcal{Y}(P), \mathcal{T}(P)$. However, we will assume that the scores depend on P only through a vector of nuisance functions $g : \mathcal{V} \to \mathbb{R}^J$ specified by the moment equations

$$\mathbb{E}[U - g(V) \mid V] = 0, \tag{2.13}$$

where U and V are random vectors with $U \subseteq W$ and $X \subseteq V \subset W$. Furthermore, we will stipulate that the dependence of $\Gamma(g; x)$ on the nuisance functions g from the possibly infinite-dimensional space $\mathcal{G}$ can be reduced as

$$\Gamma(g; x) = \Gamma(\theta(x), x),$$

where, for a fixed x, the parameter $\theta(x) \in \Theta_x \subseteq \mathbb{R}^M$ is a finite-dimensional vector of conditional moments of U deduced from g. This latter restriction rules out scores $\Gamma(g; X)$ that at a single point in the covariate space depend on exhaustive evaluations of the nuisance functions g over continuous supports. This is the case, for example, in versions of the CATE bounds from Examples 2.2.1-2.2.3 featuring instruments with continuous support $\mathcal{Z}$. In those settings, the CATE bounds depend on objects such as $\sup_{z \in \mathcal{Z}} \mathbb{E}[Y \mid Z = z, X = x]$, and are therefore not covered by the results of this chapter. Finally, we will assume that that $\Gamma(\theta(x); x)$ can be

expressed as

$$\Gamma(\theta(x);x) = \varphi_0(\theta(x);x) + \sum_{\ell=1}^{L} a_\ell \cdot \varphi_\ell(\theta(x);x) \cdot \mathbb{1}\left\{\varphi_\ell(\theta(x);x) \geq 0\right\}, \quad a_\ell \in \{-1,1\}, \tag{2.14}$$

where the functions $\varphi_\ell(\theta(x);x) : \Theta_x \times \mathcal{X} \to \mathbb{R}$ are fully differentiable with respect to $\theta(x)$ for all $x \in \mathcal{X}$. While seemingly ad-hoc, this restriction is sufficiently general to accommodate a wide range of popular partial identification assumptions for the CATE as well as optimality criteria for the resolution of ambiguity. In particular, formulation (2.14) accommodates linear combinations of $\min/\max$ operators, which typically feature in many identification bounds for the CATE with discrete instruments. In fact, our framework can be shown to be applicable to any combination of the optimality criteria discussed in Section 2.2 and the identification schemes contained in the recent survey paper by Swanson et al. (2018).[10]

Example 2.2.2 (Continued)**.** *Under the identification assumptions of the Balke-Pearl bounds and resolution of ambiguity via Minimax Regret, we have*

$$\begin{aligned} g &= (h, m, p), \\ \theta(x) &= (h(1,x), h(0,x), m(1,0,x), m(0,1,x), p(1,x), p(0,x)), \end{aligned}$$

and

$$\Gamma(g;x) = \varphi_1(\theta(x);x) \cdot \mathbb{1}\left\{\varphi_1(\theta(x);x) \geq 0\right\} - \varphi_2(\theta(x);x) \cdot \mathbb{1}\left\{\varphi_2(\theta(x);x) \geq 0\right\},$$

where $\varphi_1(\theta(x);x) = \overline{\tau}(\theta(x);x)$*,* $\varphi_2(\theta(x);x) = -\underline{\tau}(\theta(x);x)$ *are differentiable with respect to* $\theta(x)$*.*

In this framework, a natural approach for estimation is via the so-called "empirical risk minimisation" (ERM) principle (Vapnik, 1998), in which the estimate for the optimal policy is obtained as the maximiser of a sample analogue of the

[10] Albeit not directly accommodated by formulation (2.14), our framework and theoretical results also apply to scores that feature a finite number of nested linear combinations of $\min$ / $\max$ operators. We discuss this extension in Appendix A.1.

population objective Q:

$$\widehat{\pi}_n = \operatorname{argmax}\left\{\widehat{Q}_n(\pi) \,:\, \pi \in \Pi\right\}, \qquad \widehat{Q}_n(\pi) = \frac{1}{n}\sum_{i=1}^{n}(2\pi(X_i) - 1)\cdot\widehat{\Gamma}_i \quad (2.15)$$

where $\widehat{\Gamma}_i$ is some suitable estimate for $\Gamma(g; X_i)$. The ERM approach is a cornerstone of statistical learning theory and is at the foundation of many traditional and modern estimation methods in statistics, econometrics and machine learning. The ERM principle has also guided much of the recent literature on individualized treatment rules, where different variations have been applied under the names of "outcome-weighted learning" (Zhao et al., 2012) and "empirical welfare maximization" (Kitagawa and Tetenov, 2018). A major challenge in the implementation of (2.15) comes from the presence of the nuisance functions g, which are typically unknown and thus need to be estimated. Assuming that we have access to appropriate algorithms/nonparametric procedures for estimation the nuisance functions, one simple approach would be to use the sample $(W_i)_{i=1,\dots,n}$ to obtain the estimates $\widehat{g}$ and then form plug-in estimates for the score as $\widehat{\Gamma}_i = \Gamma(\widehat{g}; X_i)$. While seemingly natural, this "naive plug-in" approach has undesirable properties. In particular, policies estimated via the naive plug-in approach can typically only be shown to converge at sub-optimal rates to their population counterparts, unless very restrictive assumptions are imposed on first-stage estimators for the nuisance functions g (see, e.g., Foster and Syrgkanis, 2019).

One crucial reason underlying the undesirable statistical properties of the naive plug-in approach is that the resulting objective function estimate $\widehat{Q}_n$ is overtly sensitive to error in estimating the nuisance functions g. In order to gain intuition, it is useful to consider the following expansion of the population objective function $Q(g; \pi) = \mathbb{E}_{P_X}\left[(2\pi(X) - 1)\cdot\Gamma(g; X)\right]$,

$$Q(\widetilde{g}; \pi) - Q(g; \pi) = \left.\frac{\partial Q(g + t(\widetilde{g} - g); \pi)}{\partial t}\right|_{t=0} + \Delta(\widetilde{g}, g; \pi) + O\left(\|\widetilde{g} - g\|^2_{L_2(P)}\right) \quad (2.16)$$

where

$$\Delta(\widetilde{g}, g; \pi) =$$
$$\mathbb{E}_{P_X}\left[(2\pi(X)-1)\cdot\left(\sum_{\ell=1}^{L} a_\ell\cdot\varphi_\ell(g;X)\cdot(\mathbb{1}\left\{\varphi_\ell(\widetilde{g};X)\geq 0\right\}-\mathbb{1}\left\{\varphi_\ell(g;X)\geq 0\right\})\right)\right].$$

Von Mises expansions like the above are at the heart of the theory of orthogonal machine learning (Chernozhukov et al., 2022). In our setting, it allows to describe the impact of a small deviation from g in the direction $\widetilde{g}-g$ as consisting of three terms. The first term is the so-called "pathwise derivative" of $Q(g;\pi)$ and typically scales with $\|\widetilde{g}-g\|_{L_1(P)}$. The second term $\Delta(\widetilde{g}, g; \pi)$ is due the presence of the type of non-differentiabilities arising under partial identification, and is unique to the framework of this chapter. This term accounts for the bias that arises from misclassifying whether the component functions φ_ℓ are above or below 0, as we move away from g in the direction $\widetilde{g}-g$. The third term is a second-order remainder scaling with the mean-square distance between $\widetilde{g}$ and g. A central feature of our proposed estimation procedure is the construction of a new objective function, called Neyman-orthogonal, with reduced sensitivity to local perturbations away from g. For this purpose, we will assume that there exists functionals $\alpha_\ell(\{g,f\};V)$ such that for every $\widetilde{g}\in\mathcal{G}$

$$\varphi_\ell(\widetilde{g};x)=\mathbb{E}[\langle\alpha_\ell(\{\widetilde{g},f\};V),\widetilde{g}(V)\rangle\mid X=x],\qquad \ell=1,\ldots,L,$$

where $f\in\mathcal{F}$ is a vector of additional nuisance functions defined analogously to g.[11] We then construct Neyman-orthogonal formulations for the component functions as

$$\varphi_\ell^{\text{NO}}(\{g,f\};w)=\varphi_\ell(\theta;x)+\phi_\ell(\{g,f\};w),\quad \phi_\ell(\{g,f\};w)=\langle\alpha_\ell(\{g,f\},v),u-g(v)\rangle.$$

[11] We will also assume that for the j-th entry of the Riesz-representer we have $\alpha_\ell^{(j)}(\{(\widetilde{g}_{-j},\widetilde{g}_j),\widetilde{f}\},x)=\alpha_\ell^{(j)}(\{\widetilde{g}_{-j},\widetilde{f}\},x)$, where $\widetilde{g}_{-j}$ denotes the exclusion of the j-th entry $\widetilde{g}_j$ from the vector of nuisance functions $\widetilde{g}$. This restriction is sufficiently general to accommodate component functions $\varphi_\ell(\theta(x);x)$ that feature linear combinations of products of the parameters $\theta(x)$, thus encompassing all the discussed identification schemes, including Examples 2.2.1-2.2.3.

The functionals α_ℓ are the Riesz-representers of φ_ℓ, while the functionals ϕ_ℓ are their so-called influence function adjustments. We refer the reader to Ichimura and Newey (2022) for their properties and general methods for their calculation, while we provide below their specific form for the Balke-Pearl CATE bounds of Example 2.2.2.[12]

Example 2.2.2 (Continued)**.** *Following Ichimura and Newey (2022), we compute the influence function adjustment $\phi_U(\{g,f\},W_i)$ for the CATE upper bound by taking the Gateaux derivative of $\overline{\tau}(g;X)$, which yields*

$$\begin{aligned}
&\phi_U(\{g,f\};W_i) = \\
&\underbrace{\left[\frac{Z_i}{z(X_i)} - \frac{1-Z_i}{1-z(X_i)}\right]}_{\alpha_U^{(1)}(\{g,f\},V_i)} \cdot \big(Y_i - h(Z_i,X_i)\big) \\
&+ \underbrace{\left[\frac{D_i(1-Z_i)}{1-z(X_i)} + \frac{(1-D_i)Z_i}{z(X_i)}\right]}_{\alpha_U^{(2)}(\{g,f\},V_i)} \cdot \big(Y_i - m(D_i,Z_i,X_i)\big) \\
&+ \underbrace{\left[(m(1,0,X_i) - Y_L)\cdot\frac{1-Z_i}{1-z(X_i)} - (Y_U - m(0,1,X_i))\cdot\frac{Z_i}{z(X_i)}\right]}_{\alpha_U^{(3)}(\{g,f\},V_i)} \cdot \big(D_i - p(Z_i,X_i)\big),
\end{aligned}$$

where the associated Riesz-representer is $\alpha_U(\{g,f\},V_i) = (\alpha_U^{(1)},\alpha_U^{(2)},\alpha_U^{(3)})'$ with $f = z(x)$ and $V_i = (D_i,Z_i,X_i)'$. The influence function and Riesz-representer for the CATE lower bound are obtained by interchanging Y_U and Y_L in the expressions above.

Finally we construct Neyman-orthogonal formulations for the scores as

$$\Gamma^{\text{NO}}(\{g,f\};w) = \varphi_0^{\text{NO}}(\{g,f\};w) + \sum_{\ell=1}^{L} a_\ell \cdot \varphi_\ell^{\text{NO}}(\{g,f\};w)\cdot \mathbb{1}\left\{\varphi_\ell(g;x)\geq 0\right\},$$

[12] See also Kennedy (2022) for a user-friendly discussion of methods for computation influence function adjustments.

which are then used to form the Neyman-orthogonal objective function

$$Q^{\text{NO}}(\{\cdot,\cdot\};\pi) = \mathbb{E}_{P_W}\Big[(2\pi(X)-1)\cdot\Gamma^{\text{NO}}(\{\cdot,\cdot\};W)\Big].$$

Our construction of Neyman-orthogonal scores features the addition of the influence function adjustments ϕ_ℓ to the component functions φ_ℓ outside the indicators, but crucially not inside. Heuristically, the influence function adjustments serve the purpose of reducing the bias induced by the evaluation of the component functions $\varphi_\ell(\cdot;x)$ away from g. Since the indicators vary discontinuously with g it is not possible to linearly approximate the dependence of the indicators on the nuisance functions at the point of discontinuity $\varphi_\ell(g;x)=0$. As a result, it is not possible to reduce the bias induced by the presence of the indicators (represented by the term $\Delta(\widetilde{g},g,\pi)$ in (2.16)) by means of influence function adjustments, whose de-biasing properties implicitly rely on the validity of such linear approximation.[13] Notice that $Q^{\text{NO}}(\{g,f\};\pi) = Q(g;\pi)$ by the mean-zero property of the influence function adjustments, so that orthogonalization of the objective does not change the notion of optimal policy $\pi^*(P)$. Nonetheless, for the orthogonalized objective we have that

$$Q^{\text{NO}}(\{\widetilde{g},\widetilde{f}\};\pi) - Q^{\text{NO}}(\{g,f\};\pi) = \Delta(\widetilde{g},g;\pi) + O\left(\|\widetilde{g}-g\|^2_{L_2(P)} + \|\widetilde{f}-f\|^2_{L_2(P)}\right). \tag{2.17}$$

Comparing the above with (2.16), we see that the von Mises expansion for the orthogonalized objective does not feature the pathwise derivative term, implying that $Q^{\text{NO}}(\,\cdot\,;\pi)$ is less sensitive to deviations away from g compared to the original objective $Q(\,\cdot\,;\pi)$. As shown in Section 2.4, this property will generally translate in improved statistical guarantees for the estimated policy when the nuisance functions have to be learned from the data. It should however be noticed that the term $\Delta(\widetilde{g},g;\pi)$ still appears in the relevant expansion after orthogonalization. The contribution of this term is quantified in Section 2.4, where it is shown to be of first-order importance for the statistical properties of the estimation procedures.

[13] On the contrary, naively adding the influence function adjustments inside the indicators would lead to a bias increase, rather than a reduction.

The second key component of our approach is the use of sample-splitting, which is a commonly employed method in semiparametric inference (Chernozhukov et al., 2022) and statistical learning (Foster and Syrgkanis, 2019). The main purpose of sample-splitting is to reduce the risk of overfitting that generally arises from using the same data to estimate the nuisance functions as well as the optimal policy, as in the naive plug-in approach. Similarly to Athey and Wager (2021), we employ a particular form of sample-splitting known as K-fold cross-fitting (described below). This procedure ensures that in the estimate for $\Gamma^{\text{NO}}(\{g, f\}; W_i)$, the estimates for the nuisance functions $\{g, f\}$ are independent from the data-point W_i for that same unit. This independence property is crucial for the theoretical guarantees of our proposed method.

Our proposed estimation procedure is therefore as follows. We first randomly split the data into K evenly-sized folds and for each fold $k = 1, \ldots, K$ we obtain estimates $\{\widehat{g}^{(-k)}, \widehat{f}^{(-k)}\}$ using data from the remaining $K-1$ folds. These estimates are then used to form cross-fitted Neyman-orthogonal estimates for the scores

$$\widehat{\Gamma}_i = \widehat{\Gamma}^{\text{NO}}\left(\{\hat{g}^{-k(i)}, \hat{f}^{-k(i)}\}; W_i\right), \qquad i = 1, \ldots, n, \tag{2.18}$$

where $k(i) \in \{1, \ldots, K\}$ denotes the fold containing the i-th observation. Finally, the estimated optimal policy rule $\widehat{\pi}_n$ is obtained via the optimization problem (2.15).

2.4 Statistical guarantees for the estimated policy

Let $\widehat{\pi}_n$ be the estimated treatment policy defined in (2.15), with estimated scores as in (2.18). Following Manski (2004), we assess the performance of the estimated policy in terms of (statistical) regret with respect to population optimal policy. Let the population ambiguity-robust optimal policy be $\pi_n^*(P) \in \operatorname{argmax}_{\pi \in \Pi_n} Q(P; \pi)$, where we have included the n-subscript to the policy class Π_n to allow this to depend on the sample size for generality. The statistical regret of an estimated policy $\widehat{\pi}_n$ is defined as

$$R_n(P; \widehat{\pi}_n) = \mathbb{E}_{P_n}\left[Q(P; \pi_n^*) - Q(P; \widehat{\pi}_n)\right] \geq 0, \tag{2.19}$$

where $\mathbb{E}_{P_n}$ is the expectation with respect to the i.i.d. sample of observable random variables $(W_i)_{i=1,\dots,n}$ used to estimate $\widehat{\pi}_n$. The next few subsections build up to a final result providing asymptotic convergence guarantees for $\widehat{\pi}_n$ to π_n^* in terms of statistical regret.

2.4.1 Assumptions

We make the following assumptions.

Assumption 2.4.1 (VC-class)**.** *There exists constants* $0 \leq \nu < 1/2$ *and* $N \geq 1$ *such that* $\text{VC}(\Pi_n) \lesssim n^\nu$ *for all* $n \geq N$.

Assumption 2.4.1 restricts the policy class to have finite VC-dimension, which is a standard requirement for controlling the complexity of a policy class in the classification literature. The VC-dimension of the policy-class Π is defined as the largest interger m such that there exist points $x_1, \dots, x_m$ that are shattered by Π, i.e. where the policy values $\pi(x_1), \dots, \pi(x_m)$ can take on all 2^m possible combinations in $\{0,1\}^m$ (for more on the VC-dimension, see Wainwright, 2019). Several practically relevant classes of treatment rules satisfy this requirement, including the linear-index and quadrant rules used in the empirical application of Section 2.5. Our assumption allows the VC-dimension of the policy class to grow moderately with the sample size, thus allowing the treatment rule to depend on high-dimensional covariates.

Assumption 2.4.2 (Regularity conditions for data-generating process)**.**

(i) There exist constants $\mathcal{C}_{1,\varphi}, \mathcal{C}_{1,\alpha}$ *such that for all* $\{\widetilde{g}, \widetilde{f}\} \in \mathcal{G} \times \mathcal{F}$

$$\begin{aligned} \|\varphi_\ell(\widetilde{g}; X) - \varphi_\ell(g; X)\|_{L_\infty(P_X)} &\leq \mathcal{C}_{1,\varphi} \cdot \|\widetilde{g} - g\|_{L_\infty(P_V)}, \\ \left\|\alpha_\ell(\{\widetilde{g}, \widetilde{f}\}; V) - \alpha_\ell(\{g, f\}; V)\right\|_{L_\infty(P_V)} &\leq \mathcal{C}_{1,\alpha} \cdot \left(\|\widetilde{g} - g\|_{L_\infty(P_V)} + \left\|\widetilde{f} - f\right\|_{L_\infty(P_V)}\right), \end{aligned}$$

for $\ell = 0, \dots, L$.

(ii) There exist constants $\mathcal{C}_{2,\varphi}, \mathcal{C}_{2,\alpha}$ *such that for all* $\{\widetilde{g}, \widetilde{f}\} \in \mathcal{G} \times \mathcal{F}$

$$\begin{aligned}\|\varphi_\ell(\widetilde{g}; X) - \varphi_\ell(g; X)\|_{L_2(P_V)} &\leq \mathcal{C}_{2,\varphi} \cdot \|\widetilde{g} - g\|_{L_2(P_V)}, \\ \left\|\alpha_\ell(\{\widetilde{g}, \widetilde{f}\}; V) - \alpha_\ell(\{g, f\}; V)\right\|_{L_2(P_V)} &\leq \mathcal{C}_{2,\alpha} \cdot \left(\|\widetilde{g} - g\|_{L_2(P_V)} + \left\|\widetilde{f} - f\right\|_{L_2(P_V)}\right),\end{aligned}$$

for $\ell = 0, \ldots, L$.

(iii) There exist constants $\mathcal{C}_{3,\varphi}, \mathcal{C}_{3,\alpha}$ *such that for all* $\{\widetilde{g}, \widetilde{f}\} \in \mathcal{G} \times \mathcal{F}$

$$\begin{aligned}\|\varphi_\ell(\widetilde{g}; X)\|_{L_\infty(P_X)} &\leq \mathcal{C}_{3,\varphi}, \\ \left\|\alpha_\ell(\{\widetilde{f}, \widetilde{g}\}; V)\right\|_{L_\infty(P_V)} &\leq \mathcal{C}_{3,\alpha},\end{aligned}$$

for $\ell = 0, \ldots, L$.

(iv) The irreducible noise $\varepsilon_i := U_i - g(V_i)$ *is a sub-Gaussian vector conditional on* V_i*, with conditional variance* $\mathrm{Var}(\varepsilon_i \mid V_i) = \Sigma(V_i)$ *satisfying* $\|\lambda_{\max}(\Sigma(V))\|_{L_\infty(P_V)} \leq \overline{\lambda} < \infty$.

Assumptions 5(i) and 5(ii) impose Lipschitz continuity of the component functions and Riesz-representers with respect to the nuisance component in the L_∞ and L_2-norm, respectively. These requirements are typically met under mild conditions within the framework of this chapter. For the Balke-Pearl bounds of Example 2.2.2, these assumptions hold under the overlap condition whenever $\mathcal{G}$ and $\mathcal{F}$ are subsets of the space of bounded functions[14], which is automatically satisfied since $U_i = (Y_i, D_i, Z_i)'$ is a vector of random variables with bounded support. Assumption 2.4.2(iii) is a uniform bound on the component functions and Riesz-representers, the former implying uniform boundedness of the scores $\Gamma(g; \cdot)$. Assumption 2.4.2(iv) is a standard requirement in statistical learning theory restricting the tail behaviour of the statistical noise ε_i. It is automatically satisfied when U_i has bounded support, as in the Balke-Pearl bounds, but also allows for outcomes with unbounded support whose conditional distributions have sufficiently thin tails.

[14]That is, there exists a constant $B > 0$ such that $\|\{\widetilde{g}, \widetilde{f}\}\|_{L_\infty(P_V)} \leq B$, $\forall \{\widetilde{g}, \widetilde{f}\} \in \mathcal{G} \times \mathcal{F}$

Together with Assumption 2.4.2(iii), this assumption implies sub-gaussianity of $\Gamma^{\text{NO}}(\{g, f\}, W_i)$.

The next two assumptions impose requirements on the estimators for the nuisance components.

Assumption 2.4.3 (Regularity conditions for fist-step estimators)**.**

(i) The estimators of the nuisance functions $\{\widehat{g}_n, \widehat{f}_n\}$ belong to the function classes $\mathcal{G} \times \mathcal{F}$ with probability 1.

(ii) There exists a constant $\mathcal{C}_4 > 0$ such that

$$\|\widehat{g}_n - g\|_{L_\infty(P_V)} \leq \mathcal{C}_4,$$
$$\left\|\widehat{f}_n - f\right\|_{L_\infty(P_V)} \leq \mathcal{C}_4,$$

with probability approaching 1 as $n \to \infty$.

Part (i) of Assumption 2.4.3 is needed to ensure the validity of the Lipschitz continuity requirements of Assumption 2.4.2 for the component functions and Riesz-representers when evaluated at the first-stage estimates. In the context of the Balke-Pearl bounds, it is satisfied when $\widehat{g}_n, \widehat{f}_n$ are uniformly bounded and the estimated propensity score $\widehat{z}(X_i)$ is uniformly bounded away from 0 and 1, with probability one. The first condition is satisfied by virtually any estimation procedure when the outcomes U_i have bounded support. The second requirement can be guaranteed under appropriate trimming of the estimated propensities. Part (ii) requires that estimation errors for the nuisance components are uniformly bounded, which is satisfied under Assumption 2.4.3(i) when $\mathcal{G} \times \mathcal{F}$ is a subset of the space of bounded functions. When U_i has unbounded support and $\mathcal{G} \times \mathcal{F}$ includes unbounded functions, a more primitive condition for (ii) would be uniform consistency of the first stage estimates, that is $\|\{\widehat{g}_n, \widehat{f}_n\} - \{g, f\}\|_{L_\infty(P_V)} \to_p 0$.[15]

[15]However, it should be noted that the uniform consistency requirement is not completely innocuous when $\{\widehat{g}_n, \widehat{f}_n\}$ are machine learning estimators (Farrell et al., 2021).

Assumption 2.4.4 (L_2 convergence rates)**.** *The estimators of the nuisance functions satisfy*

$$\mathbb{E}_{P_n}\left[\|\widehat{g}_n - g\|^2_{L_2(P_V)}\right] \leq \frac{r_n}{n^{1/2}},$$
$$\mathbb{E}_{P_n}\left[\left\|\widehat{f}_n - f\right\|^2_{L_2(P_V)}\right] \leq \frac{r_n}{n^{1/2}},$$

for some sequence $r_n = o(1)$.

The above requirement on the L_2-convergence rates for the learners of the nuisance functions is a standard assumption in the semiparametric inference literature (see, e.g., Farrell, 2015, and Chernozhukov et al., 2022). It can be shown to provably hold for traditional nonparametric estimation methods such as sieve methods (Chen, 2007) as well as modern black-box machine learning algorithms including Lasso (see, e.g., Farrell, 2015), deep neural networks (Farrell et al., 2021), boosting and others, for which stronger guarantees such as Donsker-type properties are typically not available. The ability to invoke a mild L_2-convergence requirement is a virtue of the combined use of Neyman-orthogonalization and sample-splitting, a key insight brought forward by Chernozhukov et al. (2022) for semiparametric GMM inference, and subsequently leveraged by Athey and Wager (2021) and Foster and Syrgkanis (2019) in the context of statistical learning problems.[16]

Finally, we present an assumption that concerns the distribution of the component functions φ_ℓ at the population level.

Assumption 2.4.5 (Margin)**.** *There exist constants* $\mathcal{C}_m > 0$ *and* $\gamma \geq 0$ *such that*

$$\mathbb{P}_X\Big(0 < |\varphi_\ell(g; X)| \leq t\Big) \leq \mathcal{C}_m \cdot t^\gamma, \quad \forall\, t > 0.$$

for $\ell = 1, \ldots, L$.

The above assumption restricts the extent to which the distribution of the com-

[16]Unlike Athey and Wager (2021), our assumptions do not allow to trade-off accuracy in the estimation across the different nuisance functions. This is because our framework allows for $\varphi_\ell(g; x)$ to be a potentially non-linear functional of the nuisance functions g, as is the case in Examples 2.2.1-2.2.3, thus precluding such double-robustness property.

ponent functions $\varphi_\ell(g; X)$ can concentrate around the point of non-differentiability 0 and it is a form of "margin assumption", first introduced by Mammen and Tsybakov (1999). Such an assumption has been widely used in statistics to obtain fast learning rates in classification problems (see, e.g., Arlot and Bartlett, 2011). Notice that the above formulation for the margin assumption restricts the concentration of probability for the distribution of the components functions in a neighbourhood of 0, but still allows for arbitrary probability mass at 0.

Example 2.4.1 ($\gamma = 1$)**.** *Suppose X contains an absolutely continuous covariate $\tilde{\mathrm{x}}$ and $\varphi_\ell(g; X) \cdot \mathbb{1}\{\tilde{\mathrm{x}} \neq 0\}$ is absolutely continuous with density bounded above by $\overline{f}$ for $\ell = 1, \ldots, L$. Then Assumption (2.4.5) holds with $\gamma = 1$ and $\mathcal{C}_m = 2\overline{f}$.*

Example 2.4.2 ($\gamma = \infty$)**.** *Suppose there exists a $t_0 > 0$ such that $\mathbb{P}_X\left(0 < |\varphi_\ell(X)| < t_0\right) = 0$ for $\ell = 1, \ldots, L$. Then Assumption (2.4.5) holds with $\gamma = \infty$ and some $\mathcal{C}_m > 0$.*

In the context of the Balke-Pearl bounds from Example 2.2.2 with resolution of ambiguity via Minimax Regret, Assumption 2.4.5 restricts the extent to which the CATE bounds $\overline{\tau}, \underline{\tau}$ can concentrate around 0 in the data-generating process. Under $\gamma = \infty$ the support of each CATE bound is required to be fully separated from 0, while $\gamma = 1$ requires that each CATE bound has bounded density in a neighborhood of 0.

In the next section we present our theoretical results based on Assumptions 2.4.1-2.4.5.

2.4.2 Regret convergence rates

In this section, we provide asymptotic rates of convergence for the regret of the estimated policy $R_n(P; \widehat{\pi}_n)$ as defined in (2.19). In line with the existing literature, we study *uniform* regret bounds that are valid for all distributions $P \in \mathcal{P}$ satisfying Assumptions 2.4.1-2.4.5. All results in this section are thus intended to hold uniformly in the above sense, and we will drop the dependence on P for notational convenience.

We begin by noticing that controlling the convergence of $\widehat{\pi}_n$ to the best-in-class policy π_n^* intuitively requires accounting for: 1) the estimation error in the

component functions φ_ℓ and influence function adjustments ϕ_ℓ due to estimation of the nuisance components $\{g, f\}$, 2) the difference between the population Neyman-orthogonal score and true score[17], and 3) the fact that we estimate our policy using a sample from the distribution of the covariates X_i rather than their true distribution. We define the following quantities:

$$\widehat{Q}_n^{\text{NO}}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\big(2\pi(X_i)-1\big)\cdot\Gamma^{\text{NO}}(\{\widehat{g},\widehat{f}\};W_i),$$
$$Q_n^{\text{NO}}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\big(2\pi(X_i)-1\big)\cdot\Gamma^{\text{NO}}(\{g,f\},W_i),$$

and formalize this intuition in the next proposition.

Proposition 2.4.1. *The regret of $\widehat{\pi}_n$ obeys the following bound:*

$$R_n(\widehat{\pi}_n) \leq 2\mathbb{E}\left[\sup_{\pi\in\Pi_n}\left|\widehat{Q}_n^{\text{NO}}(\pi)-Q_n^{\text{NO}}(\pi)\right|\right] + \mathbb{E}\left[\sup_{\pi\in\Pi_n}\left|Q_n^{\text{NO}}(\pi)-Q(\pi)\right|\right]. \quad (2.20)$$

The second term in the above bound accounts for points 2) and 3). $Q_n(\pi) - Q(\pi)$ is a centred (mean-zero) empirical process and therefore its uniform expectation can be shown to be $O\left(\sqrt{\text{VC}(\Pi_n)/n}\right)$ using symmetrization and chaining arguments (see, e.g., Wainwright, 2019, Ch. 5.3). Controlling the first term, which accounts for point 1), is particularly challenging and requires tailored arguments that deal with the particular form of the population scores in (2.14), in particular their lack of full differentiability.

Lemma 2.4.1. *Suppose that Assumptions 2.4.1-2.4.5 hold and define $\kappa_n = \lfloor n(1-1/K)\rfloor$. Then we have*

$$\mathbb{E}_{P_n}\left[\sup_{\pi\in\Pi_n}\left|\widehat{Q}_n^{\text{NO}}(\pi)-Q_n^{\text{NO}}(\pi)\right|\right] = O\left(\frac{r_{\kappa_n}}{\sqrt{n}} + \sqrt{\frac{\text{VC}(\Pi_n)}{n}} + \left(\frac{r_{\kappa_n}}{\sqrt{n}}\right)^{\frac{\gamma+1}{\gamma+2}}\right).$$

Lemma 2.4.1 is the central result of this chapter. It provides an asymptotic rate of convergence to zero of the empirical process $\left|\widehat{Q}_n^{\text{NO}}(\pi)-Q_n^{\text{NO}}(\pi)\right|$ uniformly over

[17] That is, we need to account for the fact that we have added the influence function adjustments to the component functions.

the policy class Π_n, which depends on the VC-dimension of the class and the degree of concentration of the component functions $\varphi_\ell(g; X)$ around 0, as indexed by γ. In order to convey intuition on this result we provide a brief outline of the proof, which is based on the decomposition

$$\begin{aligned}
&\widehat{Q}_n^{\text{NO}}(\pi) - Q_n^{\text{NO}}(\pi) \\
&\quad = \frac{1}{n}\sum_{i=1}^{n}(2\pi(X_i)-1)\cdot\left[\widehat{\Gamma}^{\text{NO}}\left(\{\widehat{g}^{-k(i)},\widehat{f}^{-k(i)}\},W_i\right) - \Gamma^{\text{NO}}\left(\{g,f\},W_i\right)\right] \\
&\quad = A_0(\pi) + \sum_{\ell=1}^{L} a_\ell\cdot\left[A_{1,\ell}(\pi)+A_{2,\ell}(\pi)+A_{3,\ell}(\pi)\right], \qquad (2.21)
\end{aligned}$$

where

$$\begin{aligned}
A_0(\pi) &= \frac{1}{n}\sum_{i=1}^{n}(2\pi(X_i)-1)\cdot\left[\varphi_0^{\text{NO}}(\{\widehat{g}^{-k(i)},\widehat{f}^{-k(i)}\},W_i) - \varphi_0^{\text{NO}}(\{g,f\},W_i)\right], \\
A_{1,\ell}(\pi) &= \frac{1}{n}\sum_{i=1}^{n}(2\pi(X_i)-1)\cdot\left[\varphi_\ell^{\text{NO}}(\{\widehat{g}^{-k(i)},\widehat{f}^{-k(i)}\},W_i) - \varphi_\ell^{\text{NO}}(\{g,f\},W_i)\right] \\
&\qquad\qquad \times \mathbb{1}\left\{\varphi_\ell\left(\widehat{g}^{-k(i)};X_i\right)>0\right\}, \\
A_{2,\ell}(\pi) &= \frac{1}{n}\sum_{i=1}^{n}(2\pi(X_i)-1)\cdot\left[\mathbb{1}\left\{\varphi_\ell(\widehat{g}^{-k(i)};X_i)\geq 0\right\} - \mathbb{1}\left\{\varphi_\ell(g;X_i)\geq 0\right\}\right] \\
&\qquad\qquad \times \phi_\ell(\{g,f\};W_i), \\
A_{3,\ell}(\pi) &= \frac{1}{n}\sum_{i=1}^{n}(2\pi(X_i)-1)\cdot\left[\mathbb{1}\left\{\varphi_\ell(\widehat{g}^{-k(i)};X_i)\geq 0\right\} - \mathbb{1}\left\{\varphi_\ell(g;X_i)\geq 0\right\}\right] \\
&\qquad\qquad \times \varphi_\ell(g;X_i).
\end{aligned}$$

Terms $A_0(\pi)$ and $A_{1,\ell}(\pi)$ can be controlled using similar arguments to Athey and Wager (2021) and are responsible for the $O\left(r_{\kappa_n}/\sqrt{n}\right)$ term in the bound of Lemma 2.4.1. The de-biasing properties of Neyman-orthogonalization combined with sample-splitting play a crucial role in this context, as they ensure that the error in estimating $\varphi_\ell(g;x)$ only has a second-order contribution. As a result, term $A_{1,\ell}(\pi)$ scales with the mean-squared estimation error in the nuisance functions and, under Assumption 2.4.4, its expectation decays faster than $1/\sqrt{n}$ uniformly

over Π_n.[18] If plug-in (non-orthogonalized) estimates for φ_ℓ are instead used to form the score estimates $\widehat{\Gamma}_i$, the estimation error in the nuisance functions has a first-order impact on term $A_{1,\ell}(\pi)$. As a result, its uniform expectation would scale with the L_1 estimation error which, under Assumption 2.4.4, implies the much slower convergence $\mathbb{E}[\sup_{\pi\in\Pi_n} |A_{1,\ell}(\pi)|] = o(n^{1/4})$.

For term $A_{2,\ell}(\pi)$, the mean-zero property of the influence function adjustments together with sample-splitting ensures that this term is a centred empirical process and thus it is responsible for a $O\left(\sqrt{\text{VC}(\Pi_n)/n}\right)$ contribution again by symmetrization and chaining arguments.

Finally, for term $A_{3,\ell}(\pi)$ we show that

$$\mathbb{E}\left[\sup_{\pi\in\Pi_n} A_{3,\ell}(\pi)\right] \leq \mathbb{E}\left[\left|\varphi_\ell(g;X_i)\cdot\left(\mathbb{1}\left\{\varphi_\ell^{-k(i)}(\widehat{g}^{-k(i)};X_i)\geq 0\right\} - \mathbb{1}\left\{\varphi_\ell(g;X_i)\geq 0\right\}\right)\right|\right],$$

where the RHS can be recognized to be the classification loss of an estimator for the sign of $\varphi_\ell(g;x)$ based on thresholding $\varphi_\ell(\widehat{g}^{-k(i)};x)$. Rates of convergence in binary classification problems intuitively depend on the degree of separation of the true regression function from 0, as indexed by γ. We thus leverage results from the literature on classification (Audibert and Tsybakov, 2007) to quantify the contribution of $A_{3,\ell}$ in the bound of Lemma 2.4.1 in terms of γ.

We are now ready to combine the rates of convergence for the three terms in Proposition 2.4.1 to obtain a final regret bound for our proposed estimation procedure.

Theorem 2.4.1. *Suppose Assumptions 2.4.2-2.4.5 hold. Then the regret obeys*

$$R_n(\widehat{\pi}_n) = O\left(\sqrt{\frac{\text{VC}(\Pi_n)}{n}} \vee \left(\frac{r_{\kappa_n}}{\sqrt{n}}\right)^{\frac{\gamma+1}{\gamma+2}}\right).$$

We see that regret convergence for our policy learning procedure happens at a rate corresponding to whichever is the leading term in the asymptotic expansion of Lemma 2.4.1, which depends on ν and γ. When the policy class Π_n has fixed

[18] Notice that, by virtue of sample-splitting, the presence of the indicator $\mathbb{1}\{\varphi_\ell(\widehat{g}^{-k(i)};X_i)\geq 0\}$ is immaterial when controlling the expectation of $A_{1,\ell}(\pi)$ uniformly over Π_n.

VC-dimension ($\nu = 0$), regret convergence happens at rates ranging from $o(n^{1/4})$ in the least favourable case ($\gamma = 0$) to $O(\sqrt{\text{VC}(\Pi)/n})$ in the most favourable case ($\gamma = \infty$). The latter case is in line with existing results for policy learning with point-identified CATE, in which full-differentiability of the scores leads to $\sqrt{\text{VC}(\Pi_n)/n}$ learning rates (see Kitagawa and Tetenov, 2018; Athey and Wager, 2021; Foster and Syrgkanis, 2019). For the intermediate case $\gamma = 1$ of Example 2.4.1 our procedure guarantees regret convergence at rate $o(n^{1/3})$.

It is useful to compare the performance guarantees in this chapter with Pu and Zhang (2021), whose procedure involves the use of non-orthogonalized estimates for the scores with sample-splitting. They show that the regret of a policy estimated via the maximization (2.15) based on cross-fitted non-orthogonalized scores is upper bounded by the L_1-norm of the estimation error in the nuisance functions. Under Assumption 2.4.4, this implies $o(n^{1/4})$ convergence for the regret, which is strictly slower than our rates *for all values* of $\gamma > 0$. The faster speed of convergence guaranteed by our procedure is not just due to a refined proof strategy but crucially depends on the use of Neyman-orthogonalization, as elucidated by our discussion of Lemma 2.4.1.

Remark 2.4.1. *The procedure of Pu and Zhang (2021) also differs from ours in its final implementation, which in their case is carried out via support vector machines (SVM) with Π_n assumed to be a reproducing kernel Hilbert space. While the use of surrogate losses (such as the hinge loss in SVM) to convexify problem (2.15) can bring considerable computational benefits in terms of speed and scalability, it comes at the cost of even slower convergence guarantees than the $o(n^{-1/4})$ discussed above. We stress that our insights regarding the benefits of Neyman-orthogonalization in terms of faster learning rates apply irrespective of the final implementation. Notice also that the use of surrogate loss functions does not guarantee convergence of the estimated optimal policy to the best-in-class π_n^* in general when the policy class Π_n does not contain the "first-best" policy $\mathbb{1}\{\Gamma(g; x) \geq 0\}$, as shown by the recent work of Kitagawa et al. (2021).*

2.5 Empirical application

In this section we apply the methods discussed in this chapter to data from the National Job Training Partnership Act (JTPA) Study. This study randomly selected applicants to receive various training and services, including job-search assistance, for a period of 18 months. The study collected background information on applicants before random assignment and then recorded their earnings in the 30-month period following treatment assignment. Kitagawa and Tetenov (2018) apply their EWM method to a sample of 9,223 adult JTPA applicants to estimate the optimal allocation of *eligibility* into the programme that maximizes individual earnings across the population. In particular, they take total individual earnings in the 30 months after assignment as the welfare outcome measure Y_i, and consider policies that allocate eligibility in the programme based on the individual's observable characteristics. Kitagawa and Tetenov's analysis is from an *intent-to-treat* perspective as they focus on the problem of deciding who should be given eligibility to participate in the programme. Since eligibility in the JTPA study is randomly assigned, the effect of eligibility on earnings is point identified from the data and methods for policy learning under point-identification can be applied in this setting. We depart from Kitagawa and Tetenov (2018) and instead consider optimal assignment of *actual participation* in the training. This analysis would be of interest to a policy-maker that expects to achieve (close to) perfect compliance to her treatment decision, e.g. when participation is made a condition for receipt of a generous unemployment benefit.[19] Compliance in the JTPA study is imperfect as roughly 23% of applicants' participation status $D_i = 0, 1$ deviates from their assigned eligibility status $Z_i = 0, 1$, as shown in Table 2.2. As a result, random assignment of the eligibility instrument Z_i is not sufficient to point-identify the effect of participation in the training, motivating the use of the methods proposed in this chapter.

For partial identification of the CATE we consider the Balke-Pearl scheme of Example 2.2.2, where bounds for the 30-month post-treatment earnings are $Y_L = \$0$

[19] No financial incentive had been put in place to promote compliance in the implementation of the JTPA study.

Table 2.2: Joint distribution of eligibility and participation, JTPA study

	Eligibility (Z_i)		
Participation (D_i)	0	1	Total
0	3047	2118	5165
1	43	4015	4058
Total	3090	6133	9233

Data source: Kitagawa and Tetenov (2018) and Abadie, Angrist, and Imbens (2002).

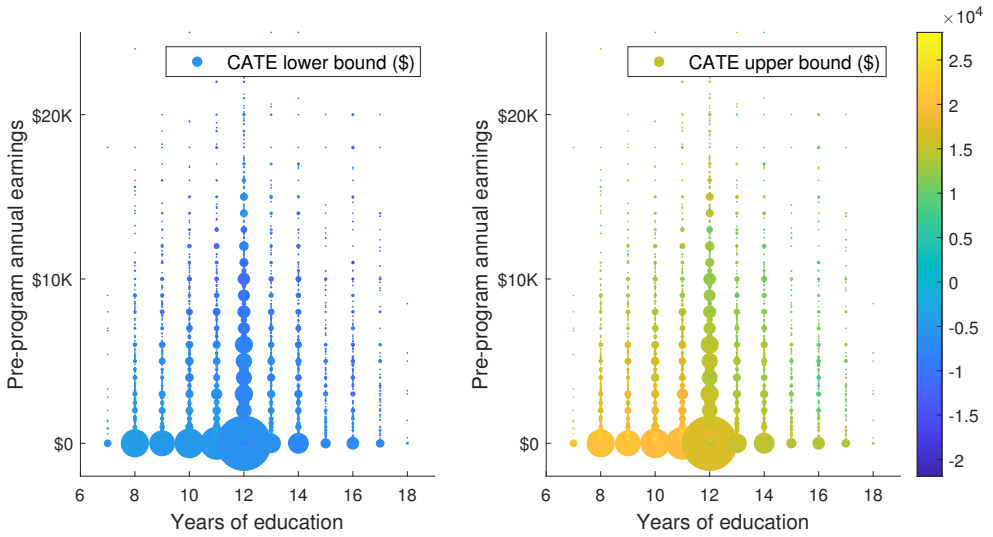
and $Y_U = \$59,640$.[20] We compare this with point-identification of the CATE as the conditional local average treatment effect (LATE), predicated under the assumption of no unobserved heterogeneity. We subtract \$1216 from both the CATE bounds and the conditional LATE; this is the average cost of services per actual treatment, estimated from Table 5 in Bloom et al. (1997). Following Kitagawa and Tetenov (2018), we condition treatment assignment on two pre-treatment variables: the individual's years of education and earnings in the year prior to assignment. Estimation of the optimal policy follows the procedure described in Section 2.3, with $K = 10$ evenly-sized data folds used to form cross-fitted Neyman-orthogonal estimates for the CATE bounds and conditional LATE functions. The nuisance functions are estimated via boosted regression trees, performed by the MATLAB function `fitrensenmble`.[21]

Figure 1 demonstrates cross-fitted plug-in estimates for the CATE bounds (a) and the LATE/minimax regret scores (b), where the size of the dots indicates the number of individuals with different covariate values. We first notice that the estimated CATE lower bounds are negative for the whole sample, and thus the maximin impact optimal policy never assigns treatment in this application. We therefore focus our analysis on minimax regret (MMR).[22] Comparison of the conditional LATE

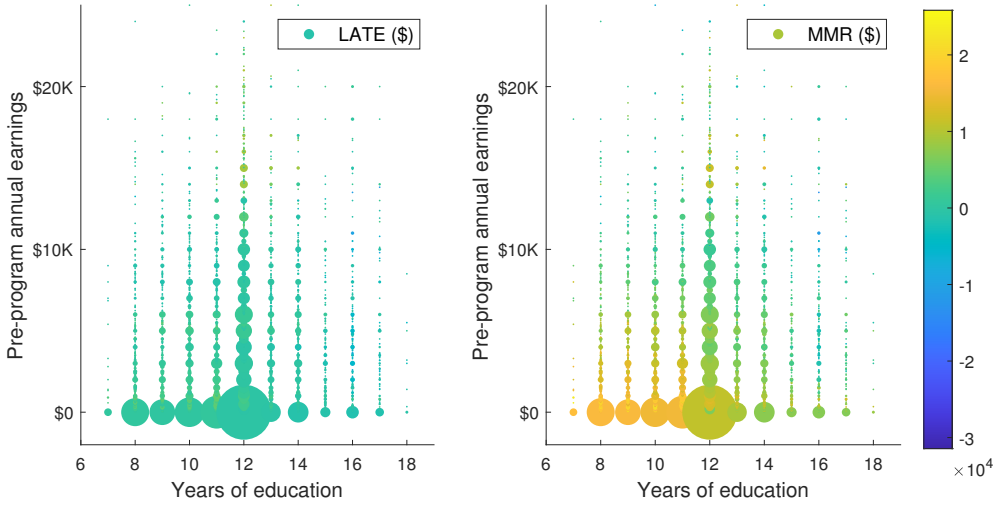
[20]The outcome upper bound corresponds to the 97.5th percentile of the earnings distribution rather than highest recorded value of $\$155,760$. Outcome bounds in Balke-Pearl bounds effectively impute unidentified expected earnings for never-takers and always-takers. Restricting expected earnings to be below such high quantile is in effect a mild requirement which brings considerable identification power.

[21]Tuning parameters have been chosen via cross-validation within each data-fold. For further details on the estimation procedure we refer to the MATLAB documentation for the command.

[22]Maximin welfare also results in no treatment for the whole population in this application.



(a) CATE bounds



(b) Scores

Figure 2.2: JTPA – Plug-in cross-fitted estimates (net of $1216)

and MMR scores highlights how partial identification leads to increased variation of the scores across different levels of education and pre-program earnings. In particular, the MMR scores are considerably higher (and positive) for individuals with fewer years of education and smaller pre-programme earnings. The conditional LATE estimates display less overall variation over the support of the covariates, compared to the MMR scores, but are lower for individuals with 0 pre-programme earnings.

We consider three alternative choices for the candidate policy class Π. The

first is the class of quadrant treatment policies. To be assigned to treatment according to this policy, an individual's education and pre-program earnings have to be above (or below) some specific threshold. Figure 2.3 illustrates the optimal quadrant treatment policies based on Neyman-orthogonalized cross-fitted MMR and LATE scores, where the colored shaded areas indicate individuals required to undertake job training by the respective policies. The optimal MMR policy (green) assigns treatment to individuals with education below 15 years and pre-treatment earnings below \$39,952. The optimal LATE policy (blue) selects the same threshold for education, but selects individuals with pre-treatment earnings above \$200 for treatment. While the two policies appear similar, they substantially differ in the proportion of population assigned to treatment (96% by the MMR policy versus 64% by the LATE policy), as shown in Table 2.3. This is due to the large concentration of individuals with pre-treatment earnings close to (or equal) zero. As a result, 32% of individuals receive a different treatment assignment across the two policies. Figure 2.3 also shows the optimal "naïve" MMR policy based on cross-fitted but non-orthogonalized scores (yellow), which recommends participation into the programme for the entire population.

Table 2.3: Treatment proportions of alternative treatment assignment policies

	Share of Population to be treated	Share of Population receiving same treatment as		
		MMR (naïve)	MMR	LATE
Quadrant Rule				
Minimax Regret (naïve)	1.00	–		
Minimax Regret	0.96	0.96	–	
LATE	0.64	0.68	0.68	–
Linear Index Rule				
Minimax Regret (naïve)	0.99	–		
Minimax Regret	0.96	0.96	–	
LATE	0.69	0.69	0.70	–
Linear Index Rule + $\mathbf{edu^2} + \mathbf{edu^3}$				
Minimax Regret (naïve)	0.99	–		
Minimax Regret	0.96	0.97	–	
LATE	0.75	0.75	0.75	–

The rows labeled "Minimax Regret (naïve)" give information on the estimated optimal minimax regret policy based on the scores in Equation (2.11) with the Balke-Pearl CATE bounds of Example 2.2.2, without Neyman-orthogonalization. The rows labeled "Minimax Regret" give information on the estimated optimal minimax regret policy with Neyman-orthogonalization. The rows labeled "LATE" give information on optimal policy for Neyman-orthogonal scores for the conditional LATE.

Second, we consider the class of linear treatment policies. This class consists of policies that assign treatment to an individual according to whether a linear index in his observable characteristics is above a certain threshold. Figure 2.4 illustrates how the direction of treatment assignment as a function of prior earnings differs between the MMR and LATE policy in a similar fashion to the quadrant rules; contrary to the LATE policy, MMR prioritizes treatment assignment to individuals with lower pre-program earnings. Nonetheless, 70% of the population still receives the same treatment under the two different policies, in light of the relatively low concentration of individuals in the areas of the covariate space where the two policies differ. Similarly to the quadrant policy rule, the MMR policy assigns treatment to a larger share of the population (95%) compared to the LATE policy (69%). The naïve MMR policy is qualitatively similar to the one using Neyman-orthogonolization, but recommends programme participation to a larger share of individuals.

Finally, we consider linear treatment policies that additionally include

quadratic and cubic terms for education. Figure 2.5 shows how the additional flexibility in the policy class leads to rules that are less interpretable but maintain similar qualitative features compared to the more parsimonious classes previously considered.

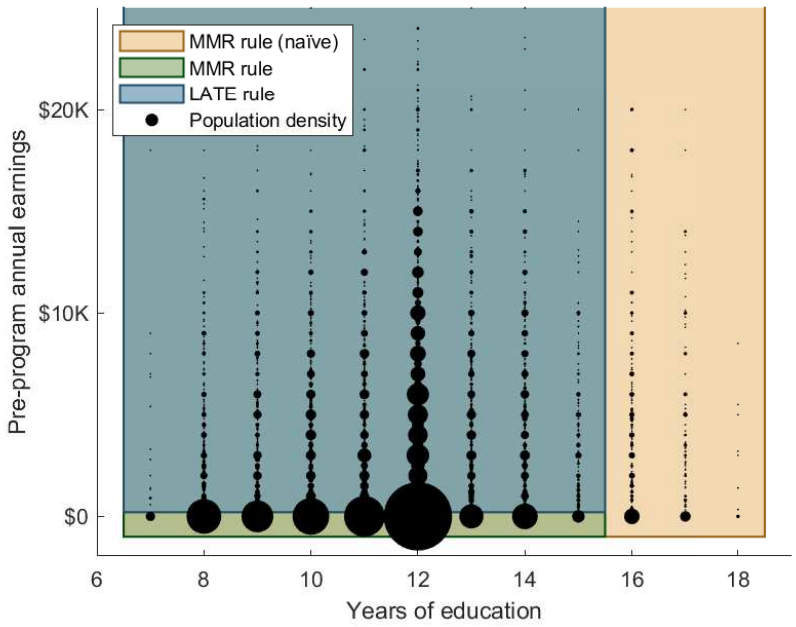


Figure 2.3: Estimated optimal policies from the quadrant policy class conditioning on years of education and pre-programme earnings.

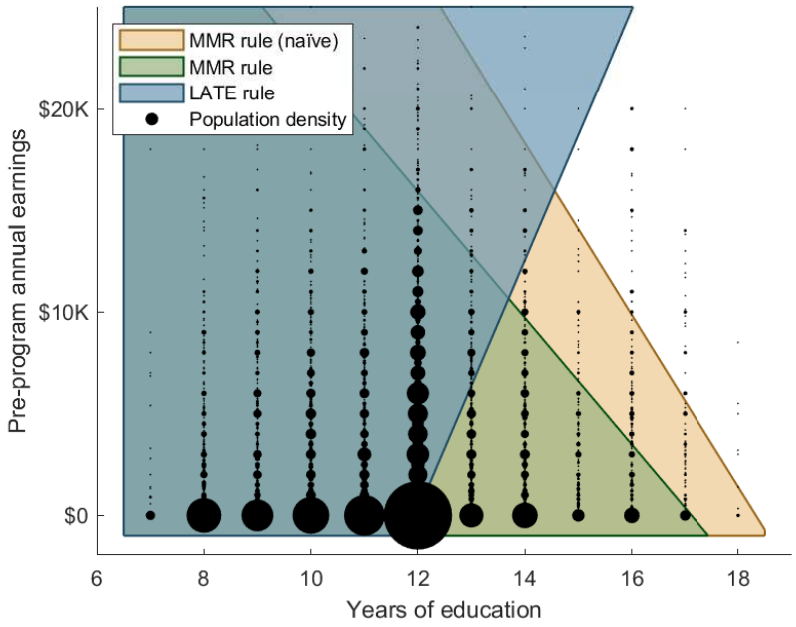


Figure 2.4: Estimated optimal policies from the linear-index policy class

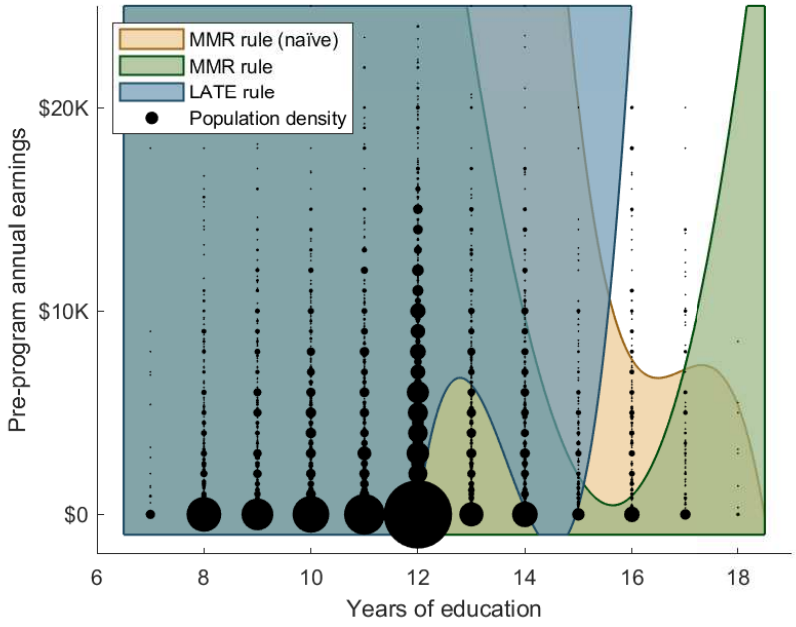


Figure 2.5: Estimated optimal policies from the linear-index policy class conditioning on years of education, $(\text{education})^2$, $(\text{education})^3$, and pre-programme earnings.

2.6 Conclusions

In this chapter, we develop a general policy learning framework for estimation of individualized treatment rules when treatment effects are partially identified. By drawing connections between the treatment assignment problem and classical decision theory, we have characterized several notions of optimal treatment policies in the presence of partial identification. We have shown how partial identification leads to a new policy learning problem where the risk is only directionally-differentiable with respect to a nuisance infinite-dimensional component. We have proposed an estimation procedure that ensures Neyman-orthogonality with respect to the nuisance components and we have provide statistical guarantees that depend on the amount of concentration around the points of non-differentiability in the data-generating process. Our proposed methods are illustrated with an application to the Job Training Partnership Act study, where we have shown that allowing for partial identification delivers substantially different programme participation policies compared to existing methods that assume point-identification.

There are several avenues for future research. First, it would be interesting to extend the theory of this chapter to partial identification via instrumental variables with continuous support. Second, it would be useful to extend the methods to more general identification sets that incorporate smoothness restrictions on unobserved counterfactual quantities, such as those considered in Kim et al. (2018). Finally, it would be interesting to assess the optimality of our proposed estimation procedure by deriving minimax lower bound rates for semiparametric statistical learning problems with directionally-differentiable risk.

Chapter 3

Auxiliary IV Estimation for Nonlinear Models

Instrumental variables (IVs) are an essential tool to estimate causal relationships from observational data. The underlying idea has been around for nearly a century (going back to the appendix in Wright 1928), and the "credibility revolution" in empirical economics has raised attention to IV methods even further in the past few decades (see e.g. Angrist and Pischke 2010). Accordingly, there is a very large literature on the subject, but the majority of both applied and theoretical work focuses on estimating linear regression models. IV estimation of non-linear models is a challenging problem, and, despite a lot of work on the subject (see references below), there is still room for new methods and ideas.

In this chapter we study estimation of non-linear models with endogenous covariates when appropriate IVs are available. However, to explain our estimation approach in the simplest possible setting, consider a linear regression model for a scalar outcomes Y_i, with a vector of (potentially endogenous) covariates X_i, and a vector of instruments Z_i, observed for units $i = 1, \dots, n$. We are interested in the effect of X_i on Y_i, parameterized by the vector β. There are many ways to construct an IV estimator in a linear model, and, at least for the case of (not too many) strong instruments, they are all essentially equivalent (up to some choice of appropriate weight matrix). One of those ways is as follows: Let $\widehat{\gamma}(\beta)$ be the ordinary least squares (OLS) estimator obtained by regressing Z_i on the residuals $Y_i - X_i' \beta$, and

let $\widehat{\beta}$ be obtained by minimizing the objective function $\widehat{\gamma}'(\beta)\,\Omega\,\widehat{\gamma}(\beta)$. Here, Ω is a symmetric positive definite weight matrix. For example, if we set $\Omega = \sum_{i=1}^{N} Z_i\,Z_i'$, then, under standard regularity conditions, it is easy to verify that $\widehat{\beta}$ is equal to the two-stage least squares (2SLS) estimator.[1]

This procedure of obtaining the 2SLS estimator is quite intuitive: We choose β such that Z_i has no explanatory power for the residuals $Y_i - X_i'\,\beta$, or equivalently, such that the regression coefficient of Z_i on $Y_i - X_i'\,\beta$ is (close to) zero. This is one way of formalizing what is meant by the instrument being an "excluded variable".

Depending on the underlying model specification, we might not want to obtain $\widehat{\gamma}(\beta)$ by OLS. For example, Chernozhukov and Hansen (2006) apply this estimation approach for quantile regressions with endogeneity, that is, $\widehat{\gamma}(\beta)$ is obtained by a quantile regression (Koenker and Bassett 1978) of Z_i on $Y_i - X_i'\,\beta$. Similarly, Lee et al. (2012) estimate panel regression models with endogeneity and unobserved factors, and therefore obtain $\widehat{\gamma}(\beta)$ by a panel regression with unobserved factors (Pesaran 2006; Bai 2009). Those ideas are combined by Harding and Lamarche (2014) who obtain $\widehat{\gamma}(\beta)$ by a quantile regression that also controls for unobserved factors.

In all those papers, the relation between Y_i and X_i is still linear. In this chapter, we generalize this estimation approach to models where the relation between Y_i and X_i is non-linear. Our leading example is the binary choice model $Y_i = \mathbb{1}\{X_i'\,\beta + U_i \geq 0\}$, where the distribution of the unobserved error U_i is assumed to be known (e.g. a logit or probit model), and U_i is independent of Z_i, but may be correlated with X_i.

In this model, if X_i were exogenous (i.e. $X_i = Z_i$), then we would simply use the maximum likelihood estimator (MLE) to estimate β. For the case of endogenous covariates, it therefore seems natural to obtain $\widehat{\gamma}(\beta)$ as the MLE of the model $Y_i = \mathbb{1}\{X_i'\,\beta + Z_i'\,\gamma + U_i \geq 0\}$, where β is fixed, and the likelihood function is only maximized over γ. The estimator for β is then obtained by minimizing $\widehat{\gamma}'(\beta)\,\Omega\,\widehat{\gamma}(\beta)$, as before. We denote the resulting estimator for β the "auxiliary IV"

[1] This is a representation of 2SLS as a minimum-distance estimator. Windmeijer (2019) shows that 2SLS can be expressed in a different way as a minimum-distance estimator as well.

(AIV) estimator, because the instrument Z_i is included as an "auxiliary regressor" in the maximum likelihood estimation.

Given the natural and intuitive structure of the AIV estimator, this chapter aims to show that it has interesting theoretical properties and is useful in practice. However, the problem of IV estimation of non-linear models is too complicated to expect that the AIV estimator is a miracle solution that always works well. In particular, under the model assumptions imposed so far, the AIV will generally not be consistent for the true parameter value for β (as $n \to \infty$). This is because the estimator $\widehat{\gamma}(\beta)$ is obtained by maximizing a misspecified likelihood function: When we write down the likelihood for the model $Y_i = \mathbb{1}\{X_i'\beta + Z_i'\gamma + U_i \geq 0\}$, we use the distribution of U_i conditional on Z_i (which is assumed to be known by the model assumptions), but one should really use the distribution of U_i conditional X_i and Z_i (which, however, is unknown to us without further assumptions on the data generating process for the endogenous X_i).

Despite this obvious weakness of the AIV estimator, we argue that it is still a useful estimator, exactly because it is a plausible estimator for β that can be constructed without making any assumptions on the data generating process for X_i. The endogenous regressors can be discrete or continuous, and apart from regularity conditions, can be arbitrarily distributed and arbitrarily correlated with U_i. This should be contrasted with other simple IV estimators for non-linear models like the control function estimator (Rivers and Vuong 1988) or the joint MLE that also fully parameterizes the distribution of X_i. Such distributional assumptions are seldom justified by economic theory, and it is well known that maximum likelihood estimators of bivariate models can be very sensitive to misspecification of the error distribution (Little, 1985; Monfardini and Radice, 2008)

The main reason why we think that the AIV estimator is useful despite being inconsistent in general is the following: If $\beta = 0$ (or more precisely, if the coefficients on the endogenous components of X_i are zero), then the AIV estimator *is* consistent as $n \to \infty$, and it also typically estimates the sign of β correctly within a neighborhood of β. This "local sign consistency" is a very useful property in empir-

ical applications, where it is often a primary concern whether a coefficient is different from zero, and what the sign of a coefficient is. We are, therefore, confident that the AIV estimator is a useful addition to the toolbox of applied researchers, which should be reported alongside other estimation approaches that have complementary properties, as illustrated by the empirical applications in this chapter.

As already mentioned above, there is a large existing literature on IV estimation in both linear and non-linear models. General non-parametric identification results are discussed, for example, in Imbens and Newey (2009), Chesher (2010), and Chesher and Rosen (2017).

Newey (1986) presents a weighted IV estimator for continuous endogenous regressor that requires estimation of the density of the exogenous regressors and instruments, and assumes linearity of the first-stage equation. Yildiz (2013) proposes a matching estimator that is $\sqrt{n}$-consistent for the coefficient of the single binary endogenous variable under non-parametric restrictions on the distribution of the unobservables, but relies on parametric specification of the functional form for the first-stage equation (e.g. a linear index specification). Han and Lee (2019) consider estimation of generalized bivariate probit models under a parametric copula assumption for the errors. The validity of their proposed procedure does not rely on knowledge of the marginal distribution of the errors in the structural and first-stage equations, but requires parametric specification of the functional form of the first-stage equation.[2] Our proposed estimator assumes knowledge of the distribution of the error in structural equation but does not impose any functional form or distributional assumptions on the first-stage equation. It also invariably accommodates continuous and discrete endogenous regressors. As a result, our proposed estimator has complementary properties to those mentioned above.

Abrevaya et al. (2010) provide a consistent test for the relevance and sign of the endogenous regressor under no parametric assumption on the distribution of the errors. Their test is based on a version of Kendall's τ-statistic that uses fitted values from the first-stage equation. Unlike Abrevaya et al. (2010), the validity of the

[2]Han and Lee (2019) also discuss identification in bivariate probit models in the absence of excluded instruments. See also Mourifié and Méango (2014) and Han and Vytlacil (2017).

test of regressor relevance based on the AIV estimator does not rely on parametric assumptions on the functional form of the first-stage.

Mu and Zhang (2018) propose an estimator for triangular binary choice models with binary endogenous regressor based on maximum score Manski (1985). Their proposal relies on the existence of continuous exogenous regressors with large support, in the spirit of Lewbel (2000). Their procedure does not require parametric specification of the distribution of unobservables or the endogenous regressor, but leads to rates of convergence that can be considerably slower than $\sqrt{n}$.

Bhattacharya et al. (2012) show 2SLS with binary outcome and binary endogenous regressor correctly estimates the sign of the average treatment effect. This property of 2SLS however is not guaranteed in the presence of additional exogenous covariates. Our simulations suggest that, unlike 2SLS, the AIV estimator's sign-consistency property is robust to the inclusion of additional regressors.

Our results for the AIV estimator of consistency at $\beta = 0$ and the local sign consistency generalise the results in the epidemiology literature of Dai and Zhang (2015), who show this result for the logit model with a continuous endogenous regressor when it is replaced by its first-stage linear IV prediction.

In the following, we first introduce the model assumptions and AIV estimator in Section 3.1, for the case where the only unknown parameter are the slope coefficients in a single index. The large sample properties of the estimator are then studies in Section 3.2. Generalizations to models with additional parameters are discussed in Section 3.3. Monte Carlos results and empirical applications are presented in Section 3.4 and 3.5, respectively. Finally, Section 3.6 concludes.

3.1 Model and Auxiliary IV estimator

3.1.1 Model

For each unit $i = 1, \ldots, n$ we observe a scalar outcome $Y_i \in \mathcal{Y}$, a vector of covariates X_i, and a vector of instrumental variables Z_i. In practice, often only a subset of the covariates are suspected to be endogenous, in which case the known exogenous covariates are included in Z_i. We denote the dimension of X_i and Z_i by

$k_x \in \{1, 2, \ldots\}$ and $k_z \in \{1, 2, \ldots\}$, respectively.

Assumption 3.1.1 (Model).

(i) The outcomes Y_i are generated from the latent variable model

$$Y_i = g(\omega_{0,i}, U_i), \qquad \omega_{0,i} := X_i' \beta_0,$$

where $U_i \in \mathbb{R}$ are unobserved random variables, the function $g(\cdot, \cdot)$ is known, and β_0 are vectors of unknown parameters.

(ii) The distribution of U_i is independent of Z_i, and U_i has known cumulative distribution function $F_U(\cdot)$.

(iii) (X_i, Z_i, U_i) are independent and identically distributed across $i = 1, \ldots, n$.

For example, for a binary choice model the function $g(\cdot, \cdot)$ in Assumption 3.1.1(i) is given by $g(\omega, u) = \mathbb{1}\{\omega + u \geq 0\}$, that is, in that example we have

$$Y_i = \mathbb{1}\{X_i' \beta_0 + U_i \geq 0\}.$$

In particular, for a binary choice probit model we choose the distribution of U_i to be standard normal, that is, $F_U(\cdot)$ in Assumption 3.1.1(ii) would be equal to the standard normal cumulative distribution function $\Phi(\cdot)$ in that case. Notice that Assumption 3.1.1(ii) imposes independence between the unobserved error U_i and the instrument Z_i, but the covariate X_i may be correlated with U_i. Finally, Assumption 3.1.1(iii) imposes cross-sectional sampling.

The binary choice probit model will be the leading example in this chapter. However, Assumption 3.1.1 also covers, for example, a Poisson model. Moreover, Section 3.3 discusses more general models where $Y_i = g(\omega_{0,i}, U_i)$ is replaced by $Y_i = g(\omega_{0,i}, W_i, U_i, \alpha_0)$, with additional unknown parameters α_0 and additional exogenous covariates W_i. That extension is important to cover models that feature additional unknown parameters beyond the regression coefficients β_0, for example, Tobit models, ordered choice models, or multinomial choice models. However, to

present our main idea and results as clearly as possible we find it convenient to focus on the simpler model structure in Assumption 3.1.1 first, which covers the binary choice model as our leading example.

For the model described by Assumption 3.1.1, let $\ell\left(y \,\middle|\, \omega\right)$ denote the log-likelihood of observing $Y_i = y$ conditional on $\omega_{0,i} = \omega \in \mathbb{R}$, treating U_i and $\omega_{0,i}$ as indepedent. For example, for discrete Y_i we have

$$\ell\left(y \,\middle|\, \omega\right) = \log \Pr\left\{y = g(\omega, U_i)\right\},$$

where the probability is evaluated according to the cdf $F_U(\cdot)$. For all our theoretical results below we will assume that the log-likelihood is strictly concave and continuously differentiable in ω. This is, of course, satisifed for the binary choice probit model where $\ell\left(y \,\middle|\, \omega\right) = y \, \log \Phi(\omega) + (1 - y) \, \log[1 - \Phi(\omega)]$.

3.1.2 AIV estimator

If Assumption 3.1.1 holds with $Z_i = X_i$, then X_i is strictly exogenous and the most natural estimator for β in the model described above is given by the maximum likelihood estimator (MLE)

$$\widehat{\beta}_{\mathrm{MLE}} = \underset{\beta}{\operatorname{argmax}} \sum_{i=1}^{n} \ell\left(Y_i \,\middle|\, X_i'\beta\right).$$

However, if $Z_i \neq X_i$ and (some of) the covariates X_i are endogenous, then $\widehat{\beta}_{\mathrm{MLE}}$ is generally not a good estimator anymore. Some estimation strategy that makes use of the instrumental variables Z_i is required in that case. The auxiliary IV estimator $\widehat{\beta}_{\mathrm{AIV}}$ that we consider in this chapter is defined by

$$\begin{aligned} \widehat{\gamma}(\beta) &= \underset{\gamma \in \mathcal{E}}{\operatorname{argmax}} \sum_{i=1}^{n} \ell\left(Y_i \,\middle|\, X_i'\beta + Z_i'\gamma\right), \\ \widehat{\beta}_{\mathrm{AIV}} &\in \underset{\beta \in \mathcal{B}}{\operatorname{argmin}} \left\|\widehat{\gamma}(\beta)\right\|_{\Omega_{n,\beta}}, \end{aligned} \tag{3.1}$$

where $\mathcal{E} \subset \mathbb{R}^{k_z}$ and $\mathcal{B} \subset \mathbb{R}^{k_x}$ are compact sets, and $\|\gamma\|_\Omega^2 = \gamma' \, \Omega \, \gamma$ is a quadratic distance measure for vectors $\gamma \in \mathbb{R}^{k_z}$, parameterized by a positive definite $k_z \times k_z$

weight matrix $\Omega = \Omega_{n,\beta}$, which might be stochastic and might depend on β. If we choose Ω equal to the identity matrix, then $\|\cdot\|_\Omega$ is simply the Euclidean norm. But having the flexibility to choose more general Ω is useful, for example, by choosing $\Omega = \frac{1}{n}\sum_i Z_i Z_i'$ the estimator $\widehat{\beta}_{\mathrm{AIV}}$ remains unchanged under the transformation $Z_i \mapsto Z_i\, A$, for any invertible $k_z \times k_z$ matrix A.

We introduce the compact sets $\mathcal{E}$ and $\mathcal{B}$ for technical reasons. In our practical implementation we assume that the boundedness conditions imposed through $\mathcal{E}$ and $\mathcal{B}$ are non-binding, that is, in practice we implement $\widehat{\beta}_{\mathrm{AIV}}$ with $\mathcal{E} = \mathbb{R}^{k_z}$ and and $\mathcal{B} = \mathbb{R}^{k_x}$.

For the special case of all regressors known to be exogenous, $Z_i = X_i$, we have $\widehat{\gamma}(\beta) = \widehat{\beta}_{\mathrm{MLE}} - \beta$, and therefore $\widehat{\beta}_{\mathrm{AIV}} = \widehat{\beta}_{\mathrm{MLE}}$. Also, for the linear regression model, $Y_i = X_i'\,\beta_0 + U_i$, with normal errors $U_i \sim \mathcal{N}(0, \sigma_0^2)$ and $\Omega = \frac{1}{n}\sum_i Z_i Z_i'$ one can easily show that $\widehat{\beta}_{\mathrm{AIV}} = \widehat{\beta}_{\mathrm{2SLS}}$, as long as the boundedness conditions imposed through $\mathcal{E}$ and $\mathcal{B}$ are non-binding.

The idea underlying the IV estimator $\widehat{\beta}_{\mathrm{AIV}}$ is as follows: We include the instruments Z_i as "auxiliary regressors" into the model, and for fixed β we maximize the corresponding log-likelihood $\ell\left(Y_i \,\middle|\, X_i'\beta + Z_i'\gamma\right)$ only over the parameters γ that correspond to the exogenous variables Z_i. Intuitively, the instruments Z_i should be "excluded variables" and their coefficient estimates $\widehat{\gamma}(\beta)$ are therefore expected to be close to zero whenever β is close to the true value β_0. Following that intuition we therefore obtain $\widehat{\beta}_{\mathrm{AIV}}$ by minimizing the distance between $\widehat{\gamma}(\beta)$ and zero.

The idea of using instrumental variables as auxiliary regressors and then minimizing their coefficients to find the parameters of interest has previously been used in other contexts. In a quantile regression setting, this method was proposed by Chernozhukov and Hansen (2006). To deal with endogeneity in panel regressions with interactive fixed effects and for the purpose of demand estimation the method was used in Lee et al. (2012) and Moon et al. (2018). However, none of those existing papers consider the type of non-linear models with endogenity that are our focus here. The IV estimator in (3.1) and our theoretical results below are novel in that context.

An interesting alternative characterization of the objective function $\|\widehat{\gamma}(\beta)\|_{\Omega_{n,\beta}}$ for $\widehat{\beta}_{\text{AIV}}$ is provided by the following lemma.

Lemma 3.1.1. *Let $\beta \in \mathbb{R}^{k_x}$. Let $W_{n,\beta} \in \mathbb{R}^{k_z \times k_z}$ be symmetric and positive definite. Assume that the log-likelihood $\ell\left(y \,\middle|\, \omega\right)$ is strictly concave and twice continuously differentiable in $\omega \in \mathbb{R}$, and that the maximizer $\widehat{\gamma}(\beta)$ in* (3.1) *is well-defined. Define the $k_z \times k_z$ matrix*[3]

$$H_n(\beta, \gamma) := \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \ell\left(Y_i \,\middle|\, X_i'\beta + Z_i'\gamma\right)}{\partial \omega^2} Z_i \, Z_i'.$$

Then, there exists $\gamma_*(\beta) \in \mathbb{R}^{k_z}$ such that for

$$\Omega_{n,\beta} = H_n(\beta, \gamma_*(\beta)) \, W_{n,\beta} \, H_n(\beta, \gamma_*(\beta)) \tag{3.2}$$

we have

$$\|\widehat{\gamma}(\beta)\|_{\Omega_{n,\beta}} = \left\| \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \ell\left(Y_i \,\middle|\, X_i'\beta\right)}{\partial \omega} Z_i \right\|_{W_{n,\beta}}.$$

The lemma provides an alternative characterization for the objective function $\|\widehat{\gamma}(\beta)\|_{\Omega_{n,\beta}}$ that is used to define our IV estimator $\widehat{\beta}_{\text{AIV}}$ in (3.1). For matrices $\Omega_{n,\beta}$ and $W_{n,\beta}$ satisfying the relation (3.2), we can use the lemma to express $\widehat{\beta}_{\text{AIV}}$ as

$$\widehat{\beta}_{\text{AIV}} \in \operatorname*{argmin}_{\beta \in \mathcal{B}} \left\| \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \ell\left(Y_i \,\middle|\, X_i'\beta\right)}{\partial \omega} Z_i \right\|_{W_{n,\beta}}. \tag{3.3}$$

The researcher could choose the weight matrix $W_{n,\beta}$ (e.g. a fixed matrix independent of β) and use (3.3) to compute $\widehat{\beta}_{\text{AIV}}$. In that case, (3.1) provides an alternative characterization of the same $\widehat{\beta}_{\text{AIV}}$ as long as (3.2) holds. Or the researcher could choose the weight matrix Ω_β (e.g. a fixed matrix independent of β). Then, if

[3] We use the following notation

$$\frac{\partial^q \ell\left(Y_i \,\middle|\, a_i\right)}{\partial \omega^q} := \left. \frac{\partial^q \ell\left(Y_i \,\middle|\, \omega\right)}{\partial \omega^q} \right|_{\omega = a_i}.$$

$H_n(\beta, \gamma_*(\beta))$ is invertible, (3.3) provides an alternative characterization of the same $\widehat{\beta}_{\mathrm{AIV}}$ as long as $W_{n,\beta} = [H_n(\beta, \gamma_*(\beta))]^{-1} \, \Omega_{n,\beta} \, [H_n(\beta, \gamma_*(\beta))]^{-1}$.

Furthermore, for the exactly identified case, $k_z = k_x$, if a solution $\widehat{\beta}_{\mathrm{AIV}}$ of the method of moment equations

$$\sum_{i=1}^{n} \frac{\partial \ell \left(Y_i \,\middle|\, X_i' \, \widehat{\beta}_{\mathrm{AIV}} \right)}{\partial \omega} Z_i = 0 \tag{3.4}$$

exists, then that solution also solves (3.3). Our assumptions in Section 3.2 guarantee existence of a solution to (3.4) for $k_z = k_x$ in large samples. Notice that (3.4) generalizes the first order condition of the MLE by replacing X_i with Z_i. While (3.4) is conveniently simple, we prefer the more general characterizations (3.1) and (3.3) of the estimator since they are applicable to the overidentified case, $k_z > k_x$, as well.

For both (3.1) and (3.3), the objective function for the minimization over β may not be convex. For computation we refer to Section 3.3. There we show that if only a single regressor is endogenous, then the "outer loop" optimization over β in (3.1) can be transformed into a one-dimensional problem (for which a grid search is computationally feasible), while the "inner loop" optimization over γ in (3.1) always remains a convex problem as long as the log-likelihood is concave.

3.2 Asymptotic results for the IV estimator

We have argued in the last section that the AIV estimator is a quite intuitive and plausible estimator to consider. However, IV estimation in non-linear models is a challenging problem and our relatively simple estimator $\widehat{\beta}_{\mathrm{AIV}}$ does not miraculously fully solve this. Indeed, under the assumptions imposed so far, the IV estimator $\widehat{\beta}_{\mathrm{AIV}}$ is *not* consistent for β_0 in general. Nevertheless, we believe that the estimator $\widehat{\beta}_{\mathrm{AIV}}$ is a useful element in the toolbox of nonlinear IV estimation, and the purpose of the current section is to demonstrate this by deriving some asymptotic properties of $\widehat{\beta}_{\mathrm{AIV}}$. To show consistency and asymptotic normality of $\widehat{\beta}_{\mathrm{AIV}}$ we impose the following additional assumption.

Assumption 3.2.1 (Exogeneity of $X_i' \beta_0$). *U_i is independent of $(X_i'\beta_0, Z_i)$.*

Assumption 3.2.1 is satisfied if for every $k = 1, \ldots, k_x$ we *either* have $\beta_{0,k} = 0$ *or* $X_{i,k}$ is exogenous. Thus, endogenous regressors are allowed for here, as long as the corresponding coefficient is zero. Indeed, we are particularly interested in cases where some of the covariates $X_{i,k}$ are endogenous and the corresponding coefficients $\beta_{0,k}$ are close to zero, but the researcher may not know that the coefficients are close to zero. Those are the cases where the estimator $\widehat{\beta}_{\mathrm{AIV}}$ will be most useful, either to formally test the null hypothesis $H_0 : \beta_{0,k} = 0$, or to simply report and interpret $\widehat{\beta}_{\mathrm{AIV}}$ in a table with multiple other estimators that have complementary properties.

In subsections 3.2.1 we derive consistency and asymptotic normality of $\widehat{\beta}_{\mathrm{AIV}}$ under Assumption 3.2.1. In subsection 3.2.2 we do not impose Assumption 3.2.1 strictly, but instead show that for endogenous $X_{i,k}$ we obtain "local sign consistency" for $\widehat{\beta}_{\mathrm{AIV},k}$ in a neighborhood around $\beta_{0,k} = 0$.

3.2.1 Consistency and asymptotic normality

In addition to the Assumptions 3.1.1 and 3.2.1 imposed so far, we also require some more technical regularity conditions. For this purpose we introduce the matrices

$$\begin{aligned}
G_n(\beta,\gamma) &:= \frac{1}{n}\sum_{i=1}^{n} Z_i X_i' \frac{\partial^2 \ell\left(Y_i \,\middle|\, X_i'\beta + Z_i'\gamma\right)}{\partial \omega^2}, \\
H_n(\beta,\gamma) &:= \frac{1}{n}\sum_{i=1}^{n} Z_i Z_i' \frac{\partial^2 \ell\left(Y_i \,\middle|\, X_i'\beta + Z_i'\gamma\right)}{\partial \omega^2}, \\
G(\beta,\gamma) &:= \mathbb{E}\left[Z_i X_i' \frac{\partial^2 \ell\left(Y_i \,\middle|\, X_i'\beta + Z_i'\gamma\right)}{\partial \omega^2} \right], \\
H(\beta,\gamma) &:= \mathbb{E}\left[Z_i Z_i' \frac{\partial^2 \ell\left(Y_i \,\middle|\, X_i'\beta + Z_i'\gamma\right)}{\partial \omega^2} \right],
\end{aligned} \tag{3.5}$$

and the score function for γ:

$$S_n(\beta,\gamma) = \frac{1}{n}\sum_{i=1}^{n} Z_i \frac{\partial \ell\left(Y_i \mid X_i'\beta + Z_i'\gamma\right)}{\partial \omega}. \tag{3.6}$$

Assumption 3.2.2 (Regularity conditions).

(i) *The parameter sets $\mathcal{B}$ and $\mathcal{E}$ are compact. $\mathcal{B}$ contains β_0 as an interior point. $\mathcal{E}$ contains 0 as an interior point.*

(ii) *For all possible outcomes y, the log-likelihood function $\ell\left(y \,\middle|\, \omega\right)$ is strictly convex in $\omega \in \mathbb{R}$. Furthermore, $\ell\left(Y_i \,\middle|\, X_i'\beta + Z_i'\gamma\right)$ is three times continuously differentiable in (β, γ) with derivatives that in expectation are bounded for all $(\beta, \gamma) \in (\mathcal{B}, \mathcal{E})$.*

(iii) $\sup_{\beta\in\mathcal{B}} \sup_{\gamma\in\mathcal{E}} \|G_n(\beta,\gamma) - G(\beta,\gamma)\| = o_P(1), \quad \sup_{\beta\in\mathcal{B}} \sup_{\gamma\in\mathcal{E}} \|H_n(\beta,\gamma) - H(\beta,\gamma)\| = o_P(1).$

(iv) *For all $(\beta, \gamma) \in (\mathcal{B}, \mathcal{E})$ and $H(\beta, \gamma)$ has full rank k_z, and $G(\beta, 0)$ has full rank k_x.*

(v) *The symmetric matrix $\Omega_{n,\beta}$ is a twice continuously differentiable function in β, and there exists a constant $c > 0$ such that with probability approaching one we have $\Omega_{n,\beta} \geq c$ for all $\beta \in \mathcal{B}$. Furthermore, we have $\sup_{\beta\in\mathcal{B}} \|\Omega_{n,\beta} - \Omega_\beta\| = o_p(1)$ for some non-random symmetric matrix Ω_β which is positive-definite for all $\beta \in \mathcal{B}$.*

Before we discuss these assumptions we first state our main consistency theorem.

Theorem 3.2.1. *Let Assumption 3.1.1, 3.2.1, 3.2.2 hold. Then we have $\widehat{\beta}_{\mathrm{AIV}} = \beta_0 + o_P(1)$, as $n \to \infty$.*

Assumption 3.2.2(i) is a standard technical regularity condition that demands the parameters sets to be compact while also containing the true parameter values – notice that 0 is the "true value" for γ. Assumption 3.2.2(ii) demands the log-likelihood to be strictly convex and sufficiently smooth. Assumption 3.2.2(iii) is a uniform convergence requirement for the second derivatives of the sample likelihood function. Classic primitive conditions for uniform convergence through "dominance conditions" are satisfied under the smoothness assumptions in Assumption 3(ii) whenever $\mathbb{E}\left[\|Z_i'X_i\|\right] < \infty$ and $\mathbb{E}\left[\|Z_i'Z_i\|\right] < \infty$. Assumption 3.2.2(v) is a standard regularity condition on the weight matrix $\Omega_{n,\beta}$.

In Assumption 3.2.2(iv), the condition on $H(\beta, 0)$ is a generalized non-collinearity condition on the instruments Z_i, while the condition on $G(\beta, 0)$ is a generalized relevance condition on the instruments — if the definition of H and G in (3.5) would not contain $\partial^2\ell/\partial\omega^2$, then these would be the standard non-collinearity and relevance conditions. If one only wanted to show "local consistency" for $\mathcal{B}$ being a small neighborhood around β_0, then it would be sufficient to impose Assumption 3.2.2(iv) at β_0 only.

Theorem 3.2.2. *Suppose that Assumptions 3.1.1, 3.2.1, and 3.2.2 hold. Furthermore, assume that* $\sqrt{n}\, S_n(\beta_0, 0) \xrightarrow{d} \mathcal{N}(0, \Sigma)$ *with* $\Sigma := \mathrm{Var}\left[Z_i \frac{d\ell\left(Y_i \,|\, X_i'\beta\right)}{d\omega} \right]$. *Then,*

$$\sqrt{n}\,(\widehat{\beta}_{\mathrm{AIV}} - \beta_0) \xrightarrow{d} \mathcal{N}\left(0,\, (G'\,W\,G)^{-1}\,G'\,W\,\Sigma\,W\,G\,(G'\,W\,G)^{-1}\right),$$

where $G := G(\beta_0, 0)$ *and* $W := H^{-1}\,\Omega\,H^{-1}$, *with* $\Omega := \Omega_{\beta_0}$ *and* $H := H(\beta_0, 0)$.

Asymptotic normality of the score $S_n(\beta_0, 0)$ can be shown using the Lindeberg-Lévy central limit theorem under the moment bound $\mathbb{E}\left[\|Z_i'Z_i\|\right] < \infty$. Apart from that, the assumptions of Theorem 3.2.2 are identical to those of Theorem 3.2.1. From the asymptotic variance formula of the AIV estimator one can deduce the optimal weighting matrix $\Omega^* = H\Sigma^{-1}H$ under which $\mathrm{AVar}(\sqrt{\mathrm{n}}\,\widehat{\beta}_{\mathrm{AIV}}) = (G'\Sigma^{-1}G)^{-1}$. While continuously-updating or feasible two-step procedures would be asymptotically efficient, we find that they bring negligible gains in our simulations compared to the simple choice $\Omega_{n,\beta} = \frac{1}{n}\sum_{i=1}^n Z_i Z_i'$, which is the one we recommend.

3.2.2 Local sign consistency

In this section we consider the case where all regressors are exogenous, except for a single endogenous regressor $X_{i,k}$. We are interested in how the probability limit of the corresponding component $\widehat{\beta}_{\mathrm{AIV},k}$ of our AIV estimator depends on the corresponding true parameter value $\beta_{0,k}$. The red line in Figure 3.1 plots this relationship for one particular data generating process (DGP) that we also employ in our Monte Carlo simulations (the binary choice probit model with a continuous endogenous

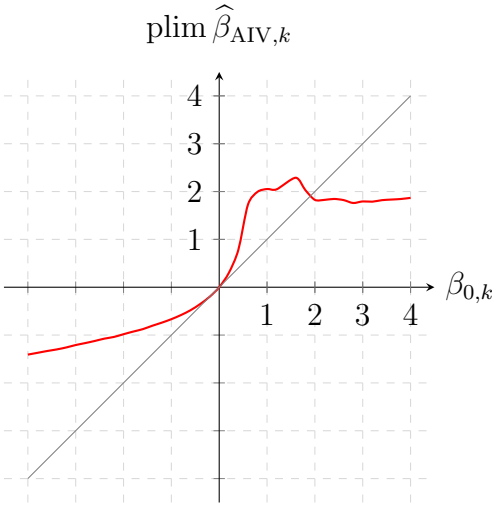


Figure 3.1: Probability limit of $\widehat{\beta}_{\mathrm{AIV}}$ as a function of β_0.

regressor of Table 3.1).[4] The details of this DGP do not matter here. What we are interested in are a couple of qualitative features of Figure 3.1 that are valid more generally:

(i) If the true value $\beta_{0,k}$ of the regression coefficient corresponding to the single endogenous regressor is equal to zero, then $\operatorname{plim} \widehat{\beta}_{\mathrm{AIV},k}$ is also equal to zero. We already know that this is true for all data generating processes that satisfy the conditions in Theorem 3.2.1.

(ii) According to Theorem 3.2.1 we also know that if the regressor $X_{i,k}$ would be exogenous as well, then we would have $\operatorname{plim} \widehat{\beta}_{\mathrm{AIV},k} = \beta_{0,k}$, corresponding to the 45-degree line drawn in grey in Figure 3.1. If the degree of endogeneity would be small, then we would expect only a small deviation from the 45-degree line. If the degree of endogeneity is larger, then we expect a larger deviation from the 45-degree line.

(iii) In Figure 3.1 the sign of $\operatorname{plim} \widehat{\beta}_{\mathrm{AIV},k}$ is always equal to the sign of $\beta_{0,k}$. If this property holds, then we say that $\widehat{\beta}_{\mathrm{AIV}}$ is "globally sign consistent". In our simulations in Section 3.4 we always find global sign consistency for all DGPs that we explore, but we are not able to provide formal conditions under

[4] In that DGP both the variance of the endogenous regressor $X_{i,k}$ and of the error term U_i are equal to one.

which global sign consistency holds in this chapter (apart from exogeneity of $X_{i,k}$). Instead, in the following we want to discuss "local sign consistency", that is, sign consistency in a small neighborhood of $\beta_{0,k} = 0$.

(iv) Local sign consistency of the AIV estimator leads to a test of the null hypothesis $H_0 : \beta_{0,k} = 0$ that is consistent for alternatives in a neighborhood of H_0. Global sign consistency leads to general consistency of the same test. This is particularly useful in applications where a main concern is whether the effect of an endogenous "treatment" variable is zero.

Let $\beta_*(\beta_0)$ be the large n probability limit of $\widehat{\beta}_{\text{AIV}}$. We say that the k'th component of the AIV estimator is *locally sign consistent* if there exists $\delta > 0$ such that

$$\text{sign}\left(\beta_{*,k}(\beta_0)\right) = \text{sign}\left(\beta_{0,k}\right),$$

for all β_0 with $|\beta_{0,k}| < \delta$. Under appropriate smoothness conditions, a sufficient condition for local sign consistency of $\widehat{\beta}_{\text{AIV},k}$ is given by

$$\left.\frac{\partial \beta_{*,k}(\beta_0)}{\partial \beta_{0,k}}\right|_{\beta_{0,k}=0} > 0. \tag{3.7}$$

In the following we give two concrete examples where (3.7) holds. Notice, however, that (3.7) is not a necessary condition for for local sign consistency of $\widehat{\beta}_{\text{AIV},k}$, because one could, for example, have $\frac{\partial \beta_{*,k}(\beta_0)}{\partial \beta_{0,k}} = 0$, at $\beta_{0,k} = 0$, and still achieve local sign consistency via $\frac{\partial^2 \beta_{*,k}(\beta_0)}{\partial^2 \beta_{0,k}} = 0$ and $\frac{\partial^3 \beta_{*,k}(\beta_0)}{\partial^3 \beta_{0,k}} > 0$, at $\beta_{0,k} = 0$.

Example 3.2.1 (Probit control function model)**.** *Consider the generalized probit control function model:*

$$\begin{aligned} &Y_i = \mathbb{1}(X_i'\beta_0 - U_i > 0), \\ &x_i = m(Z_i) + V_i, \quad X_i = (1, x_i), \\ &(U_i, V_i) \,\big|\, Z_i \sim (U_i, V_i) \sim F_{U,V},\ U_i \sim \mathcal{N}(0,1), \end{aligned} \tag{3.8}$$

where x_i, U_i, V_i are all scalar random variables, Z_i is a vector of instruments that includes a constant, and $F_{U,V}$ is absolutely continuous with density $f_{U,V}$. This model

is more general than the one studied in Rivers and Vuong (1988) in that it does not require the conditional distribution $U_i|V_i$ to be linear in V_i nor normal; we also do not require linearity of $m(\cdot)$. In this example, the regressor $X_{i,k}$ for $k = 2$ is endogenous, and one can show (see Appendix) that

$$\left.\frac{\partial \beta_{*,2}(\beta_0)}{\partial \beta_{0,2}}\right|_{\beta_{0,2}=0} = 1,$$

therefore local sign consistency holds.

Example 3.2.2 (Generalized bivariate probit IV)**.** *Consider the bivariate probit IV model:*

$$\begin{aligned} &Y_i = \mathbb{1}(X_i'\beta_0 + U_i > 0), \\ &x_i = \mathbb{1}(m(Z_i) + V_i > 0) \quad X_i = (1, x_i),\ Z_i = (1, z_i), \\ &(U_i, V_i) \,\big|\, Z_i \sim\ (U_i, V_i)\ \sim F_{U,V},\ U_i \sim \mathcal{N}(0,1), \end{aligned} \tag{3.9}$$

where x_i, z_i, U_i, V_i are all scalar random variables, and $m(Z_i)$ is assumed to be a monotonic function of z_i. This model nests the popular bivariate probit model which further assumes joint normality of (U_i, V_i) and linearity of $m(Z_i)$. Again, the regressor $X_{i,k}$ for $k = 2$ is endogenous, and one can show that (3.7) *holds for $k = 2$, that is, local sign consistency holds in this example as well. Unlike Example 3.2.1, the arguments we use to show local sign consistency in this model do not directly generalize to the over-identified case ($k_z > 2$).*

We know that local sign consistency of the AIV estimators holds whenever all the regressors are exogenous. In addition, the above examples provide two concrete data generating processses where a single regressor is endogenous and local sign consistency still holds. We have also verified local sign consistency (in fact, global sign consistency) numerically for all the data generating processes in our Monte Carlo simulations. We therefore conclude that local sign consistency of the AIV estimator holds for a large class of data generating processes.

As mentioned above, an important implication of the local sign consistency property is that a t-test for the hypothesis $H_0 : \beta_{0,k} = 0$ based on our estimator has

non-trivial power — and it is in fact consistent — in a neighbourhood of the null hypothesis. The distribution of this t-test under H_0 is guaranteed by Theorem 3.2.2. For implemention, one just needs to compute the sample analog $\widehat{\mathrm{AVar}}(\sqrt{n}\beta_{\mathrm{AIV}})$ of the asymptotic variance-covariance matrix $(G'\,W\,G)^{-1}\,G'\,W\,\Sigma\,W\,G\,(G'\,W\,G)^{-1}$ given in the theorem, and $n(\widehat{\beta}_{\mathrm{AIV},k})^2/[\widehat{\mathrm{AVar}}(\sqrt{n}\beta_{\mathrm{AIV}})]_{kk}$ will be $\chi^2(1)$ distributes as $n \to \infty$.

3.3 Generalization and implementation

We now want to discuss a generalization of the model and AIV estimator described in Section 3.1.1. Specifically, we now assume that in addition to (Y_i, X_i, Z_i), $i = 1, \ldots, n$, we also observe the additional strictly exogenous covariate W_i. The difference between X_i and W_i is that W_i need not enter the model through the linear single index $\omega_i = X_i'\beta$. Similarly, in addition to the unknown parameters β we now allow for the additional unknown parameters α, which also need not enter the model through the single index ω_i. Examples where this generalization is important are ordered choice models, Tobit models, and negative binomial models. The appropriate generalization of Assumption 3.1.1 is as follows:

Assumption 3.3.1 (Generalized Model).

(i) The outcomes Y_i are generated from the latent variable model

$$Y_i = g(\omega_{0,i}, W_i, U_i, \alpha_0), \qquad \omega_{i,0} := X_i'\,\beta_0,$$

where $U_i \in \mathbb{R}$ are unobserved random variables, the function $g(\cdot,\cdot,\cdot,\cdot)$ is known, and α_0 and β_0 are vectors of unknown parameters.

(ii) The distribution of U_i is independent of (Z_i, W_i), and U_i has known cumulative distribution function of $F_U(\cdot)$.

(iii) (X_i, Z_i, W_i, U_i) are independent and identically distributed across $i = 1, \ldots, n$.

Let $\ell\left(Y_i \,\middle|\, W_i,\, \omega_i,\, \alpha\right)$ be the log-likelihood of Y_i conditional on W_i, $\omega_{0,i} = \omega_i$ and $\alpha_0 = \alpha$. Then, the generalization of the AIV estimator in (3.1) is given by

$$\left(\widehat{\gamma}(\beta), \widehat{\alpha}(\beta)\right) = \underset{(\gamma,\alpha)\in\mathcal{E}}{\operatorname{argmax}} \sum_{i=1}^{n} \ell\left(Y_i \,\middle|\, W_i,\, X_i'\beta + Z_i'\gamma,\, \alpha\right), \quad \widehat{\beta}_{\mathrm{AIV}} \in \underset{\beta\in\mathcal{B}}{\operatorname{argmin}} \left\|\widehat{\gamma}(\beta)\right\|_{\Omega_{n,\beta}},$$
$$\widehat{\alpha}^{\dagger}(\beta) = \underset{\alpha\in\mathcal{E}}{\operatorname{argmax}} \sum_{i=1}^{n} \ell\left(Y_i \,\middle|\, W_i,\, X_i'\beta,\, \alpha\right), \quad \widehat{\alpha}_{\mathrm{AIV}} = \widehat{\alpha}^{\dagger}(\widehat{\beta}_{\mathrm{AIV}}, 0), \quad (3.10)$$

where $\mathcal{B}$ is a compact parameter set for β_0 as before, and $\mathcal{E}$ now is a compact parameter set for (γ, α). Compactness of the parameter sets is again a very helpful technical regularity condition to derive asymptotic results. However, for practical implementation we again assume that the boundedness imposed by $\mathcal{B}$ and $\mathcal{E}$ is not binding, that is, in practice we replace $\mathcal{B}$ by $\mathbb{R}^{k_x}$ and $\mathcal{E}$ by $\mathbb{R}^{k_z+k_\alpha}$, where k_α denotes the dimension of α.

The appropriate generalizations of our consistency result of Theorem 3.2.1 for the AIV estimator in Section 3.2.1 to the model and estimator in Assumption 3.3.1 and display (3.10) are provided in the appendix.

In our Monte Carlo simulations and empirical applications below we focus on the binary choice model for which this extension of the model discussed here is not actually required. However, even for the binary choice model there can be computational advantages in implementing the AIV estimator according to (3.10) instead of (3.1). This is because we can "move" all the regression coefficients that correspond to exogenous covariates from β to α and then implement (3.10) instead of (3.1). The advantage of that implementation is that the "inner loop" optimization over (γ, α) in (3.10) is a convex optimization problem (since we assume the log-likelihood to be a concave function) while the "outer loop" optimization over β is in general a non-convex problem, implying that we generally want the dimension the vector β to be as small as possible for computational reasons.

This computational issue is important in practice (e.g. for our simulation and application sections below) and we therefore want to be explicit about it: Consider the setup of our original Assumption 3.1.1 and decompose $X_i = (X_i^{\mathrm{end}\prime}, X_i^{\mathrm{ex}\prime})'$ and $Z_i = (Z_i^{\mathrm{ex}\prime}, X_i^{\mathrm{ex}\prime})$, where X_i^{end} are the endogenous regressors, X_i^{ex} are the

exogenous regressors, and Z_i^{ex} are the excluded instruments. In most applications we expect X_i^{end} to be low-dimensional (often just a single variable). Let β^{end} and β^{ex} be the regression coefficients corresponding to X_i^{end} and X_i^{ex}. By applying the generalized AIV estimator in (3.10) to this setup with $(X_i, W_i, Z_i, \beta, \alpha)$ equal to $(X_i^{\text{end}}, X_i^{\text{ex}}, Z_i^{\text{ex}}, \beta^{\text{end}}, \beta^{\text{ex}})$ we obtain

$$\begin{aligned}\left(\widehat{\gamma}(\beta^{\text{end}}), \overline{\beta}^{\text{ex}}(\beta^{\text{end}})\right) &= \underset{(\gamma,\,\beta^{\text{ex}})}{\operatorname{argmax}} \sum_{i=1}^{n} \ell\left(Y_i \,\middle|\, X_i^{\text{end}\prime}\beta^{\text{end}} + X_i^{\text{ex}\prime}\beta^{\text{ex}} + Z_i^{\text{ex}\prime}\gamma\right), \\ \widehat{\beta}^{\text{end}} &\in \underset{\beta^{\text{end}}}{\operatorname{argmin}} \left\|\widehat{\gamma}(\beta^{\text{end}})\right\|_{\Omega_{n,\beta}^{\text{end}}}, \\ \widehat{\beta}^{\text{ex}} &= \underset{\beta^{\text{ex}}}{\operatorname{argmax}} \sum_{i=1}^{n} \ell\left(Y_i \,\middle|\, X_i^{\text{end}\prime}\widehat{\beta}^{\text{end}} + X_i^{\text{ex}\prime}\beta^{\text{ex}}\right), \end{aligned} \tag{3.11}$$

where $\Omega_{n,\beta}^{\text{end}}$ now is a positive definite matrix of dimension $\dim(X_i^{\text{end}}) \times \dim(X_i^{\text{end}})$ only.

Again, the key observation here is that the optimization over $(\gamma,\, \beta^{\text{ex}})$ is a convex optimization problem, while the optimization over β^{end} is non-convex but usually low-dimensional (often just one-dimensional which can e.g. be implemented by an initial grid-search followed by, for example, a golden-section search). Implementing the AIV estimator via (3.11) is therefore often computationally preferable to (3.1) and to (3.3), in particular, if k_x is large. Our results in the Appendix show that the two implementations are asymptotically equivalent when $k_z = k_x$. When $k_z > k_x$, then the choice of implementation and weight matrix matters for the (asymptotic) distribution of the resulting estimator, see the the appendix for more details.[5] In practice, we again recommend the simple choice $\Omega_{n,\beta}^{\text{end}} = \frac{1}{n}\sum_{i=1}^{n} Z_i^{\text{ex}} Z_i^{\text{ex}\prime}$.

---

[5] When $k_z > k_x$, the asymptotic distribution for $\widehat{\beta}^{\text{end}}$ is equivalent under the two implementations if

$$\Omega = \begin{bmatrix} \Omega^{\text{end}} & 0 \\ 0 & \Omega^{\text{ex}} \end{bmatrix} \quad \text{and} \quad \Sigma_{\gamma\alpha} := \mathbb{E}\left[Z_i X_i' \frac{\partial^2 \ell\left(Y_i \,\middle|\, X_i^{\text{end}\prime}\beta_0^{\text{end}} + X_i^{\text{ex}\prime}\beta_0^{\text{ex}}\right)}{\partial \omega^2} \right] = 0.$$

Beyond this set of special conditions, the two implementations do not in general lead to asymptotically equivalent estimators under over-identification.

3.4 Monte Carlo simulations

We consider the following data generating process (DGP):

$$
\begin{aligned}
&Y_i = \mathbb{1}\left\{\beta_1 + X_{2,i}\beta_2 + X_{3,i}\beta_3 + U_i \geq 0\right\}, \qquad U_i \sim \mathcal{N}(0,1),\\
&X_{2,i} = \sigma_{X_2}^{-1}\left(Z_i + V_i\right), \quad Z_i \sim (\chi^2(k) - k)/\sqrt{2k}, \quad k = 10,\\
&X_{3,i} = \sigma_{X_3}^{-1}\left(\mathcal{N}(0,1) + 0.5 \cdot Z_i^2\right)\\
&V_i = \varepsilon_i + \delta_{end} \cdot (U_i + \delta_{no\_norm} \cdot (2 \cdot \mathbb{1}\left\{U_i \geq 0\right\} + U_i^2 - 2)), \quad \varepsilon_i \sim \mathcal{N}(0,1),
\end{aligned}
$$

with normalizing constants σ_{X_2} and σ_{X_3} chosen so that $\mathrm{Var}(X_{2,i}) = \mathrm{Var}(X_{3,i}) = 1$.

We set $\beta_1 = 1$, $\beta_3 = -1$ and we document the performance of different procedures in the estimation of β_2 under different configurations of β_2, δ_{end} and δ_{no_norm}. We also report the empirical size of a two-sided t-test for the null hypothesis that β_2 is equal to its true value.

The AIV estimator is implemented as in (3.11), with outer-loop direct search over β_2 initialized at the control function estimate. Standard errors used in the t-test are based on the sample analogue of the asymptotic variance formula in Theorem 3.2.2.

For the control function estimator, the test statistic is based on the standard error formula provided in Rivers and Vuong (1988), which assumes correct specification of the model (including joint normality).[6]

The results are collected in Table 3.1. As expected, MLE is severely biased under endogeneity of the regressor and non-normality of the errors, leading to confidence intervals with no coverage. As predicted by theory, the control function estimator is consistent and provides accurate inference under joint normality of the errors, or in the absence of endogeneity. However, the coverage of its associated confidence intervals is null in the presence of endogeneity and lack of joint normality, due to large biases. The AIV estimator instead enjoys negligible bias under all

[6]Notice that the asymptotic variance formula contained in Rivers and Vuong (1988) is for a different normalization of the variance of U_i compared to MLE, bivariate Probit and the AIV estimators, which all assume $\mathrm{Var}(U_i) = 1$. In order to make the control function estimates comparable with the other methods, we rescale the original control function estimates based on the normalization of Rivers and Vuong (1988) and appropriately adjust standard errors via the Delta method.

configurations considered, at the cost of mild variance increases compared to control function. Remarkably, the resulting rejection probabilities for a two-sided t-test are close to nominal size, including for values of β_2 away from 0. Figure 3.2 reports the power function of a two-sided t-test of regressor relevance ($H_0 : \beta_2 = 0$) based on the AIV estimator under $\delta_{end} = 1$ and $\delta_{no\_norm} = 2$. The sign-consistency property of the AIV estimator results in good power for this test, even though the presence of bias in the estimator for values of β_2 away from 0 leads to non-monotonic power in this DGP.

3.4.1 Simulations with binary endogenous regressor

We consider a modification of the previous DGP which now features a binary endogenous regressor:

$$
\begin{aligned}
&Y_i = \mathbb{1}\left\{\beta_1 + X_{2,i}\beta_2 + X_{3,i}\beta_3 + U_i \geq 0\right\}, \qquad U_i \sim \mathcal{N}(0,1),\\
&X_{2,i} = \left\{\sigma_{X_2}^{-1}\left(Z_i + V_i\right) \geq 0\right\}, \quad Z_i \sim (\chi^2(k) - k)/\sqrt{2k}, \quad k = 10,\\
&X_{3,i} = \sigma_{X_3}^{-1}\left(\mathcal{N}(0,1) + 0.5 \cdot Z_i^2\right)\\
&V_i = \varepsilon_i + \delta_{end} \cdot (U_i + \delta_{no\_norm} \cdot (2 \cdot \mathbb{1}\left\{U_i \geq 0\right\} + U_i^2 - 2)), \quad \varepsilon_i \sim \mathcal{N}(0,1).
\end{aligned}
$$

where $\sigma_{X_2}, \sigma_{X_3}, \beta_3$ are as before, and we set $\beta_1 = 0.4$ to ensure $\mathbb{E}[Y_i] \approx 0.5$.

The results are given in Table 3.2. Curiously, the control function estimator has negligible bias under endogeneity and non-normality.[7] However, the associated rejection probabilities for control function are far from nominal size due to severe underestimation of the standard errors. As expected, the bivariate probit estimator performs well under joint-normality of the errors or exogeneity of the regressors. Under endogeneity and non-normality, the bivariate probit estimator suffers from large bias and considerable size distortions of its associated tests. On the other hand, the AIV estimator has negligible bias, resulting in good size control and high power of the associated two-sided test of regressor relevance, as shown in Figure 3.3. It is interesting to notice that 2SLS is not sign-consistent for the effect of

[7] We have verified that the small bias of property exhibited by the control function estimator in this DGP is coincidental, and does not hold generally.

the endogenous treatment in this DGP, while it known to be sign-consistent in the absence of additional covariates (Bhattacharya et al., 2012). The AIV estimator enjoys sign-consistency in this DGP, suggesting improved robustness of the sign-consistency property to the inclusion of additional covariates compared to 2SLS.

3.5 Empirical applications

In this section we present two empirical applications. In each application we compare estimates of the coefficient on the the binary endogenous regressor of interest based on popular existing estimators and the AIV estimator. In the first application, a test of relevance of the endogenous regressor based on the AIV estimator cautions the researcher about the conclusion that that having health insurance increases the probability that an individual visits a doctor in a given year. In the second application, the AIV estimator confirms the conclusion that smoking habits are transmitted by a mother to her offspring which would be reached using existing methods. Overall, the two applications showcase the usefulness of the AIV estimator as a tool for checking the robustness of inferential conclusions in nonlinear models.

3.5.1 The effect of health insurance on hospital visits (Han and Lee, 2019)

Health insurance coverage is considered an important factor for patients' decisions to use medical services. On the other hand, the decision to acquire health insurance is endogenously determined by an individual's health status, as well as socioeconomic characteristics that are correlated with health outcomes. In this application, we investigate how health insurance coverage affects an individual's choice to visit a doctor. For this purpose, we use a dataset constructed by Han and Lee (2019) which combines data from the 2010 wave of the Medical Expenditure Panel Survey (MEPS) with information from the National Compensation Survey published by the US Bureau of Labor Statistics. The outcome of interest Y_i is a binary variable indicating whether an individual visited a doctor's office in January 2010; the binary endogenous treatment X_i^{end} indicates whether an individual has his/her own private insurance. Two instrumental variables are used following Zimmer (2018): the num-

ber of employees in the firm at which the individual works and a dummy variable that indicates whether a firm has multiple locations. These variables reflect how big the firm is, and the underlying rationale for using these variables as instruments is a that a bigger the firm is more likely it provides fringe benefits including health insurance. The validity of these instruments relies on firm size not directly affecting the decision to visit a doctor. Following Han and Lee (2019), we include a further 23 exogenous variables in the model as additional controls, including demographic characteristics as well as indicators of health status.

Table 3.3 provides estimates for the coefficient β^{end} on the binary treatment using probit MLE, 2SLS, the control function estimator of Rivers and Vuong (1988), the bivariate probit estimator and the AIV estimator. We also report the associated standard errors and the p-value of a two-sided t-test of no effect of health insurance coverage on doctor visits ($H_0 : \beta^{\text{end}} = 0$). Remarkably, all methods deliver positive estimates for β^{end} with similar magnitudes, with the exception of MLE being roughly three times smaller than the other methods considered.[8] The test of regressor relevance based on bivariate probit leads to rejection of the null hypothesis at all conventional levels of significance. On the other hand, a test based on the AIV estimator does not reject the same hypothesis at the 1% level of significance. The difference between p-values in this application is driven by the varying magnitude of the standard errors associated with each method. Standard errors associated with bivariate probit are likely to underestimate the sampling variability of the estimator, as their validity relies on the assumptions of joint normality of the unobserved disturbances and linearity of the first-stage equation. Our theory reassures us that the AIV estimator provides inference that is robust to relaxing those assumptions in this application.

[8] As an estimator for the average partial effect of X^{end} rather than the coefficient β^{end}, only the sign of the 2SLS estimator can be compared to the other estimators. Even though we report results for the control function estimator, its use is not recommended in this application as the endogenous regressor is binary.

3.5.2 The intergenerational transmission of smoking habits (Mu and Zhang, 2018)

Vertical transmission within family is considered a key driver of the persistence of health behaviours. The way in which harmful practices such as smoke are transmitted within a family has therefore important implications for health policies. In this application we apply our proposed methods to the study of the intergenerational transmission of smoking habits using data from British Household Panel Survey. The outcome of interest Y_i is a binary variable indicating whether an adolescent smokes or not; the binary endogenous treatment X_i^{end} indicates whether his/her single mother smokes or not. Following Loureiro et al. (2010) and Mu and Zhang (2018), the instrument used is an indicator for whether the teenagers' grandfather was high-skilled or low-skilled occupation (including unemployed). The underlying rationale for using this variable as an instrument is that the impact of parental socio-economic status on smoking behaviour does not extend beyond one generation, after controlling for the relevant explanatory variables. We include a further 5 exogenous variables in the model as additional controls: the child's age at interview year, the single mother's age at interview year, an indicator for whether the mother has higher education, an indicator for whether the mother is in a high-skilled or low-skilled occupation, and the natural logarithm of monthly household income. Table 3.4 provides estimation results for the coefficient β^{end}. All methods deliver positive estimates for the coefficient β^{end}, implying that a mother's decision to smoke increases the probability that her offspring chooses to be a smoker too. Similarly to the previous empirical application, we find that all methods deliver estimates of similar magnitude, with the exception of the MLE estimate being roughly a third of the bivariate probit and AIV estimators. While the AIV estimator delivers a smaller estimate for β^{end} compared to bivariate probit, two-sided tests based on these two estimators both lead to rejection of the hypothesis $H_0 : \beta^{\text{end}} = 0$ at all conventional levels of significance. As a result, the AIV estimator provides evidence on the robustness of the conclusion that smoking habits are transmitted between generations.

3.6 Conclusions

We have introduced the AIV estimator as a new and simple estimator in non-linear models with endogenous covariates. The estimator translates the concept of an "excluded instruments" into a criterion functions that demands the MLE of the instrument coefficient to be zero when the instruments are includes as covariates. We show that the resulting AIV estimator is consistent if the endogenous regression coefficient is equal to zero. We also demonstrate that, for the case of a single endogenous regressor, the AIV estimator is usually sign-consistent. We have argued that those properties and its simplicity make the estimator useful in practice, as illustrated by our empirical applications. In particular, the estimator is complementary to the control function and the probit IV estimator, because it makes weaker assumptions, but also delivers weaker consistency results.

Table 3.1: Monte Carlo simulations with continuous endogenous regressor, $n = 7000$

			MLE			2SLS			Control Function			Auxiliary IV		
β_2	δ_{end}	δ_{norm}^{no}	Bias	Std.	$\hat{p}$; .05	Bias	Std.	$\hat{p}$; .05	Bias	Std.	$\hat{p}$; .05	Bias	Std.	$\hat{p}$; .05
0	1	0	−0.74	.024	1.00	.016	.009	.415	.000	.031	.047	.000	.033	.047
0	1	2	−1.22	.054	1.00	.041	.024	.385	.550	.030	.999	−.002	.079	.046
−0.1	1	2	−1.22	.063	1.00	.106	.024	.999	.499	.025	1.00	−.006	.077	.047
0.1	1	2	−1.19	.046	1.00	.022	.025	.168	.590	.058	.99	.031	.087	.059
1	0	-	.000	.025	.046	.756	.006	1.00	.000	.032	.049	.001	.043	.047

Notes: Simulation results based on 5000 replications.

Table 3.2: Monte Carlo simulations with binary endogenous regressor, $n = 7000$

			2SLS			Control Function			Bivariate Probit			Auxiliary IV		
β_1	δ_{end}	δ_{norm}^{no}	Bias	Std.	$\hat{p}$; .05	Bias	Std.	$\hat{p}$; .05	Bias	Std.	$\hat{p}$; .05	Bias	Std.	$\hat{p}$; .05
0	1	0	.124	.025	.990	.152	.076	.521	.001	.064	.054	.001	.082	.052
0	1	2	.310	.069	.999	−.020	.399	.513	.325	.178	.494	.006	.209	.048
−0.1	1	2	.370	.068	1.00	−.053	.399	.440	.403	.183	.748	−.029	.209	.054
0.1	1	2	.240	.071	.973	−.013	.412	.521	.281	.121	.170	.046	.216	.042
1	0	-	.690	.039	1.00	.000	.069	.056	.000	.065	.057	.001	.072	.056

Notes: Simulation results based on 5000 replications.

Table 3.3: Effect of Health Insurance on Doctor Visits

	$\widehat{\beta}^{\text{end}}$	Std. Err.	p-value
MLE	.1796	.0404	$<$ .0000
2SLS	.1326	.0459	.0038
Control Function	.5358	.1740	.0020
Bivariate Probit	.4962	.1558	.0014
Auxiliary IV	.5622	.2487	.0238

Sample size $n = 7555$. Data source: Han and Lee (2019).

Table 3.4: Effect of mother's smoking habits on child's smoking habits

	$\widehat{\beta}^{\text{end}}$	Std. Err.	p-value
MLE	.3305	.0347	$<$ .0000
2SLS	.3746	.1243	.0026
Control Function	1.089	.3047	.0004
Bivariate Probit	1.440	.1203	$<$ .0000
Auxiliary IV	1.130	.4269	.0081

Sample size $n = 7053$. Data source: Mu and Zhang (2018).

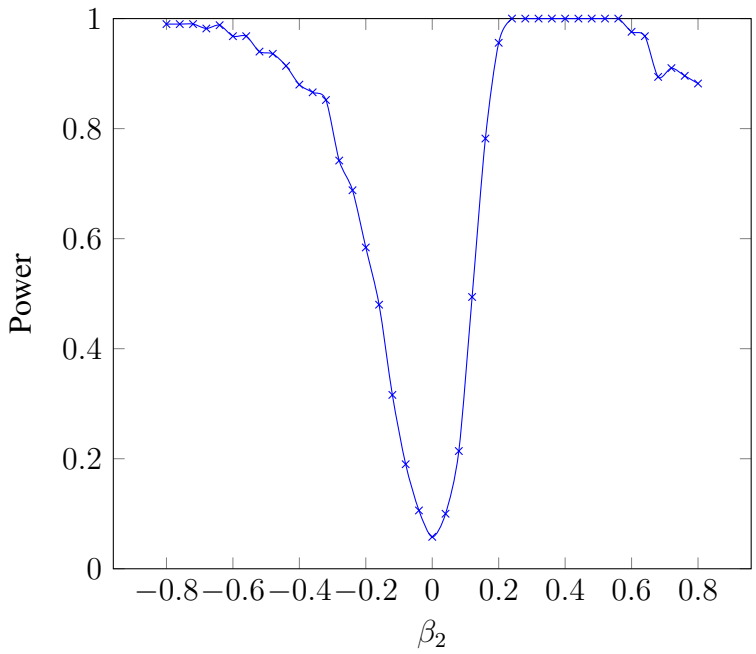


Figure 3.2: Power function of two-sided test with continuous endogenous regressor, $n = 7000$

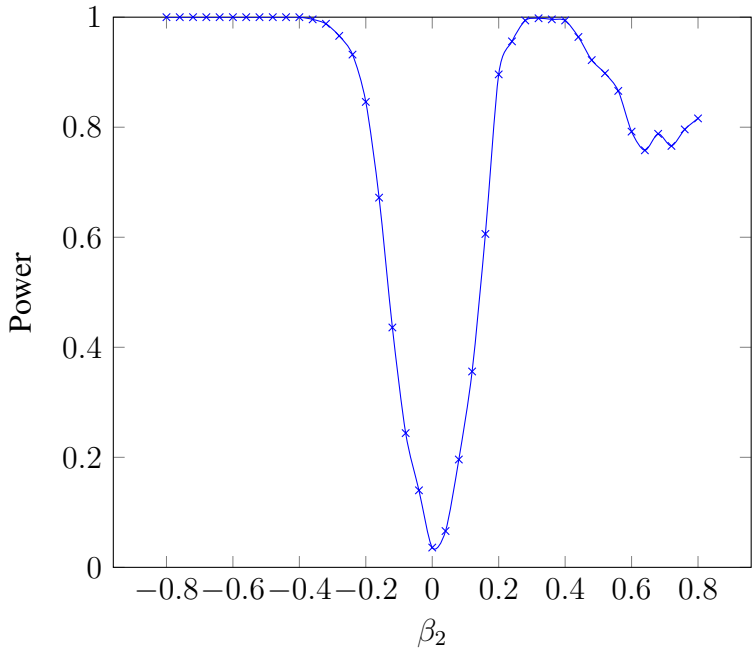


Figure 3.3: Power function of two-sided test with discrete endogenous regressor, $n = 7000$

Chapter 4

Cluster-Robust Standard Errors for Linear Regression Models with Many Controls

It is common practice in empirical work to use standard errors and associated confidence intervals that are robust to heteroskedasticity and/or various forms of dependence. In particular, since Moulton (1986) highlighted the importance of accounting for dependence arising in data with a group structure, researchers often assume that the data are clustered at some economically relevant level, e.g. by individual unit or geographical location.

The justification for this type of inference procedures is asymptotic, in the sense that their validity relies on the assumption that the sample size is large relative to the (fixed) number of parameters in the model. In small samples two issues arise: (i) confidence intervals based on the usual Gaussian approximation become invalid, (ii) robust standard errors are biased. A variety of methods that address these issues in the context of the linear regression model have been proposed in the literature. However, these usually alleviate but do not entirely solve the problem and are not always appealing given their ad hoc nature (see Imbens and Kolesár, 2016, for a discussion).

Furthermore, while modern datasets usually include a large number of observations, the assumption that the number of estimated parameters is negligible relative

to the sample size can still be unattractive, even when the researcher's goal is to conduct inference on a small set of parameters. For example, in many important applications of the linear regression model, the object of interest is $\boldsymbol{\beta}$ in a model of the form

$$y_{i,n} = \boldsymbol{\beta}'\mathbf{x}_{i,n} + \boldsymbol{\gamma}_n'\mathbf{w}_{i,n} + u_{i,n}, \qquad i = 1, \ldots, n, \tag{4.1}$$

where $y_{i,n}$ is a scalar outcome variable, $\mathbf{x}_{i,n}$ is a $d \times 1$ vector of regressors of fixed dimension, $\mathbf{w}_{i,n}$ is a vector of covariates of possibly "large" dimension K_n, and $u_{i,n}$ is an unobserved scalar error term. In many applications of this model, the assumption that $K_n/n \to 0$ is unpalatable or even violated, as researchers often include a large set of covariates in $\mathbf{w}_{i,n}$ in order to control for observed and unobserved confounders (see discussion below).

Motivated by the above observations, this chapter develops inference theory for linear regression models with many controls and clustering. In particular, we first show that the usual cluster-robust standard errors by Liang and Zeger (1986) are inconsistent in general when $K_n/n \not\to 0$. We then propose a new clustered standard error formula that allows to carry out valid inference on $\boldsymbol{\beta}$ under asymptotics in which K_n is allowed (but not required) to grow as fast as the sample size.

The results of this chapter contribute to the long-established literature initiated by White (1984) dealing with cluster-robust inference in a variety of models, a review of which is given by Cameron and Miller (2015); see, e.g., Arellano (1987), Bell and Mccaffrey (2002), Hansen (2007), Cameron et al. (2008), Ibragimov and Müller (2016), Pustejovsky and Tipton (2018) and Canay et al. (2021). In particular, our analysis is related to a literature, reviewed in Imbens and Kolesár (2016), in which bias-reduction modifications of standard errors and particular distributional approximations are proposed with the aim of improving the performance of cluster-robust inference procedures in small samples. We contribute to this literature by studying a new general class of cluster-robust variance estimators that allows to fully correct such "small-sample" bias, while also exploiting the particular structure of the model in (4.1) to circumvent the need for non-Gaussian distributional

approximations.

This chapter also adds to a sizeable body of literature that deals with inference procedures in models that involve the estimation of many incidental parameters; see, e.g., Angrist and Hahn (2004), Hahn and Newey (2004), Stock and Watson (2008), Belloni et al. (2013), Cattaneo, Jansson, and Newey (2018b, 2018a), Verdier (2020), and references therein. In particular, our findings can be seen as generalising those of Cattaneo et al. (2018b), who establish asymptotic normality of the OLS estimator of $\boldsymbol{\beta}$ in (4.1) when $K_n/n \nrightarrow 0$, and provide inference methods under such asymptotics when the errors are independent and heteroskedastic.

The results in this chapter were derived independently of Li (2016), who tackles the problem of cluster-robust variance estimation for the full set of coefficients of a generic high-dimensional linear model and obtains a similar estimator to ours.[1] However, his results are not directly applicable to inference and are silent about the extent to which sufficient conditions for consistent variance estimation restrict the underlying data generating process of the regressors.

The rest of this chapter is organized as follows. Section 2 introduces the framework and illustrates its relevance using three leading examples. Section 3 discusses our assumptions. Section 4 presents our main theoretical results. Section 5 reports the findings of a Monte Carlo study. Section 6 presents an empirical illustration. Section 7 briefly concludes. Proofs and extensions of the results are given in Appendix C.

4.1 Framework and motivation

The main object of interest in our analysis is $\boldsymbol{\beta}$ in (4.1), on which we would like to carry out inference while treating the high-dimensional $\mathbf{w}_{i,n}$ as nuisance covariates. A natural choice of estimator for $\boldsymbol{\beta}$ is the OLS estimator, which can be written as

$$\hat{\boldsymbol{\beta}} = (\sum_{i=1}^{n} \hat{\mathbf{v}}_{i,n}\hat{\mathbf{v}}_{i,n}')^{-1}(\sum_{i=1}^{n} \hat{\mathbf{v}}_{i,n}y_{i,n}), \qquad \hat{\mathbf{v}}_{i,n} = \sum_{j=1}^{n} M_{ij,n}\mathbf{x}_{j,n}, \tag{4.2}$$

[1] I am grateful to Valentin Verdier for alerting me to the existance of Li's (2016) thesis.

where $M_{ij,n} = \mathbb{1}\{i = j\} - \mathbf{w}_{i,n}'(\sum_{k=1}^n \mathbf{w}_{k,n}\mathbf{w}_{k,n}')^{-1}\mathbf{w}_{j,n}$ is the (i,j) entry of the symmetric and idempotent annihilator matrix $\mathbf{M}_n$, with $\mathbb{1}\{\cdot\}$ denoting the indicator function. Defining $\hat{\mathbf{\Gamma}}_n = \sum_{i=1}^n \hat{\mathbf{v}}_{i,n}\hat{\mathbf{v}}_{i,n}'/n$ and $\mathbf{\Sigma}_n$ the (conditional) variance of $\sum_{i=1}^n \hat{\mathbf{v}}_{i,n}u_{i,n}/\sqrt{n}$, it is well-known that, when $n \to \infty$ and K_n is fixed, the asymptotic distribution of $\hat{\boldsymbol{\beta}}_n$ is

$$\mathbf{\Omega}_n^{-1/2}\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_d), \qquad \mathbf{\Omega}_n = \hat{\mathbf{\Gamma}}_n^{-1}\mathbf{\Sigma}_n\hat{\mathbf{\Gamma}}_n^{-1}. \tag{4.3}$$

When the errors are assumed to be correlated only within G_n clusters of bounded size, $\mathbf{\Sigma}_n$ can be estimated consistently with the popular cluster-robust variance estimator by Liang and Zeger (1986, LZ hereafter):

$$\hat{\mathbf{\Sigma}}_n^{\mathrm{LZ}} = \frac{1}{n}\sum_{g=1}^{G_n}\sum_{i,j\in\mathcal{T}_{g,n}} \hat{\mathbf{v}}_{i,n}\hat{\mathbf{v}}_{j,n}'\hat{u}_{i,n}\hat{u}_{j,n}, \qquad \hat{u}_{i,n} = \sum_{j=1}^n M_{ij,n}(y_{j,n} - \hat{\boldsymbol{\beta}}_n'\mathbf{x}_{j,n}), \tag{4.4}$$

where $\mathcal{T}_{g,n}$ denotes the subset of observations contained in cluster g and $\{\mathcal{T}_{g,n} : 1 \leq g \leq G_n\}$ is a partition of the data. As a result, asymptotically valid inference can be carried out using the usual testing procedures based on the distributional approximation $\hat{\boldsymbol{\beta}}_n \overset{a}{\sim} \mathcal{N}(\boldsymbol{\beta}, \hat{\mathbf{\Gamma}}_n^{-1}\hat{\mathbf{\Sigma}}_n^{\mathrm{LZ}}\hat{\mathbf{\Gamma}}_n^{-1}/n)$.

The objective of this chapter is to establish cluster-robust inference procedures for $\boldsymbol{\beta}$ under asymptotics in which $K_n/n \nrightarrow 0$. Allowing the dimension of the nuisance covariates K_n to grow at the same rate as the sample size n enables us to cover many relevant applications of the general model in (4.1).

Example 4.1.1. ***Linear regression model with increasing dimension***

This leading example takes (4.1) as the data generating process, in which $\mathbf{w}_{i,n}$ contains many observable individual characteristics and their nonlinear transformations, dummy variables for many categories such as age group, cohort, geographic location etc. and their interactions with the former. The inclusion of many covariates is motivated in practice by the assumption that the variable of interest $\mathbf{x}_{i,n}$ can be taken as exogenous after controlling for $\mathbf{w}_{i,n}$. Although the study of linear regression models with growing dimension has a long tradition in statistics (see

e.g. Huber, 1973; Mammen, 1993), until recently inference results were exiguous and limited to the case in which the number of regressors in the model is at least a vanishing fraction of the sample size. Cattaneo, Jansson, and Newey (2018b) exploit the separability of model (4.1) to develop valid inference procedures for $\boldsymbol{\beta}$ when $K_n/n \nrightarrow 0$, but their theory only covers the case of homoskedastic and heteroskedastic errors. Li and Müller (2021) develop cluster-robust inference theory in this setting for a scalar β, i.e. $d = 1$; their results allow for $K_n \propto n$ but rely on a strong restriction on $\sum_{i=1}^{n}(\boldsymbol{\gamma}_n'\mathbf{w}_{i,n})^2$, which limits the amount of sample variation of y_i that can be induced by the high-dimensional controls $\mathbf{w}_{i,n}$. Belloni et al. (2013) instead propose an estimation procedure for $\boldsymbol{\beta}$ based on LASSO double-selection and provide inference theory for the case of i.n.i.d. data. While their method can accomodate $K_n \gg n$, it relies on the assumption that the effect of confounders can be controlled for by a small subset of the variables in $\mathbf{w}_{i,n}$ up to some small approximation error ("approximate sparsity").

Example 4.1.2. ***Semiparametric Partially Linear Model***

Researchers often assume that data are generated by the model

$$y_i = \boldsymbol{\beta}'\mathbf{x}_i + g(\mathbf{z}_i) + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{4.5}$$

where both $\mathbf{x}_i$ and $\mathbf{z}_i$ have fixed dimension, but the function $g(\cdot)$ is unknown. The partially linear model is a long-standing area of interest in econometrics (see, e.g., Heckman, 1986, and Robinson, 1988). Estimation of this semiparametric model is often carried out via series-based methods, in which the researcher assumes that the function $g(\cdot)$ can be closely approximated using the polynomial functions $\mathbf{p}_n(\mathbf{z}) = (p^1(\mathbf{z}), \ldots, p^{K_n}(\mathbf{z}))'$, so that $g(\mathbf{z}_i) \approx \boldsymbol{\gamma}_n'\mathbf{p}_n(\mathbf{z}_i)$ for some $\boldsymbol{\gamma}_n$. The series estimator for $\boldsymbol{\beta}$ is the OLS estimator as defined in (4.2), where $\mathbf{w}_{i,n} = \mathbf{p}_n(\mathbf{z}_i)$. When the underlying function $g(\cdot)$ is not sufficiently smooth and/or the dimension of $\mathbf{z}_i$ is relatively large, the inclusion of many polynomial terms might be required, resulting in K_n being non-negligible relative to n. Cattaneo et al. (2018a) are the first to consider asymptotics in which $K_n/n \nrightarrow 0$ in this setting. They establish asymptotic

normality for $\hat{\boldsymbol{\beta}}_n$ and valid inference procedures under homoskedasticity and, in their subsequent paper, heteroskedasticity (CJN, 2018a). However, no results are available for the case of clustering.

Example 4.1.3. ***Multi-way fixed effects panel data models***

Panel data models that use fixed effects are often used in order to control for unobserved heterogeneity, such as the one-way fixed effects panel data regression model

$$Y_{it} = \boldsymbol{\beta}'\mathbf{X}_{it} + \alpha_i + U_{it}, \qquad i = 1, \dots, N, \qquad t = 1, \dots, T, \tag{4.6}$$

where α_i is a scalar individual effect, $\mathbf{X}_{it}$ is a vector of regressors and U_{it} is a scalar error term. This model can be mapped into our baseline specification in (4.1) by setting $n = NT$, $y_{(i-1)T+t,n} = Y_{it}$, $\mathbf{x}_{(i-1)T+t,n} = \mathbf{X}_{it}$, $u_{(i-1)T+t,n} = U_{it}$, $\boldsymbol{\gamma}_n = (\alpha_1, \dots, \alpha_N)$ and $\mathbf{w}_{(i-1)T+t,n}$ equal to the i-th unit vector of dimension N. It follows that $K_n = N$ and $K_n/n = 1/T$, which motivates the asymptotics of this chapter under $N \to \infty$ and T fixed. For this case, Arellano (1987) shows that LZ's variance estimator (1986) is consistent when errors are clustered at the individual level. For the same setting, Stock and Watson (2008) propose a cluster-robust estimator for the variance with additional zero restrictions on the conditional autocovariances of the errors within entities, e.g. when an MA(q) structure is imposed on U_{it}.

In many empirical settings, researchers want to control for multiple terms of unobserved heterogeneity. In the analysis of student/teacher or worker/firm matched data, for example, two-way fixed effects models are commonly used, taking the form

$$Y_{it} = \boldsymbol{\beta}'\mathbf{X}_{it} + \alpha_i + e_{d_{it}} + U_{it}, \qquad i = 1, \dots, N, \qquad t = 1, \dots, T, \tag{4.7}$$

where $e_{d_{it}}$ are unobserved factors common to all observations sharing the same value of the indexing variable $d_{it} \in \{1, \dots, N_d\}$, so that $\mathbf{w}_{(i-1)T+t,n}$ is now a $N + N_d$ vector selecting the relevant fixed effects from $\boldsymbol{\gamma}_n = (\alpha_1, \dots, \alpha_N, e_1, \dots, e_{N_d})$. When T is fixed and only few observations are assigned to each value of d_{it} (i.e. data are sparsely matched), then the number of fixed effects grows proportionally

to the sample size and $K_n \propto n$. Verdier (2020) considers cluster-robust inference under these asymptotics for a new estimation procedure for $\boldsymbol{\beta}$ that accomodates instrumental variables but is generally less efficient than OLS. He also proposes a variance estimator that can be seen as a generalisation of the one proposed in this chapter to multi-way cluster dependence.

To simplify exposition, we present our inference theory for linear regression models with many controls for the case of strictly exogenous regressors. While all the results of this chapter can be well-understood for this special case, their generalisation to (potential) misspecification bias in the model is straightforward and is provided in the Appendix.

4.2 Assumptions

In this section we present a set of assumptions for the special case of strict exogeneity of the regressors. A more general set of assumptions that allows for misspecification bias is given in the Appendix.

Suppose that $\{(y_{i,n}, \mathbf{x}'_{i,n}, \mathbf{w}'_{i,n}) : 1 \leq i \leq n\}$ is generated by (4.1) and set $\mathcal{X}_n = (\mathbf{x}_{1,n}, \ldots, \mathbf{x}_{n,n})$ and $\mathcal{W}_n = (\mathbf{w}_{1,n}, \ldots, \mathbf{w}_{n,n})$. We define the following quantities:

$$\chi_n = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\|\mathbf{Q}_{i,n}\|^2], \qquad \mathbf{Q}_{i,n} = \mathbb{E}[\mathbf{v}_{i,n}|\mathcal{W}_n],$$
$$\hat{\boldsymbol{\Gamma}}_n = \sum_{i=1}^{n} \hat{\mathbf{v}}_{i,n}\hat{\mathbf{v}}'_{i,n}/n, \qquad \boldsymbol{\Sigma}_n = \mathbb{V}[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \hat{\mathbf{v}}_{i,n}u_{i,n}|\mathcal{X}_n, \mathcal{W}_n],$$

where $\mathbf{v}_{i,n} = \mathbf{x}_{i,n} - (\sum_{j=1}^{n} \mathbb{E}[\mathbf{x}_{j,n}\mathbf{w}'_{j,n}])(\sum_{j=1}^{n} \mathbb{E}[\mathbf{w}_{j,n}\mathbf{w}'_{j,n}])^{-1}\mathbf{w}_{i,n}$ is the population counterpart of $\hat{\mathbf{v}}_{i,n}$. Also, letting $\lambda_{\text{min}}(\cdot)$ denote the minimum eigenvalue of its argument, define

$$\mathcal{C}_n = \max_{1\leq i\leq n} \{\mathbb{E}[u_{i,n}^4|\mathcal{X}_n, \mathcal{W}_n] + \mathbb{E}[\|\mathbf{V}_{i,n}\|^4|\mathcal{W}_n] + 1/\mathbb{E}[u_{i,n}^2|\mathcal{X}_n, \mathcal{W}_n]\} + 1/\lambda_{\text{min}}(\mathbb{E}[\tilde{\boldsymbol{\Gamma}}_n|\mathcal{W}_n])$$

where $\mathbf{V}_{i,n} = \mathbf{x}_{i,n} - \mathbb{E}[\mathbf{x}_{i,n}|\mathcal{W}_n]$, $\tilde{\boldsymbol{\Gamma}}_n = \sum_{i=1}^{n} \tilde{\mathbf{V}}_{i,n}\tilde{\mathbf{V}}'_{i,n}/n$ and $\tilde{\mathbf{V}}_{i,n} = \sum_{j=1}^{n} M_{ij,n}\mathbf{V}_{i,n}$.

We impose the following three assumptions:

Assumption 4.2.1. $\max_{1\leq g\leq G_n} \#\mathcal{T}_{g,n} = O(1)$, *where* $\#\mathcal{T}_{g,n}$ *is the cardinality of* $\mathcal{T}_{g,n}$ *and where* $\{\mathcal{T}_{g,n} : 1 \leq g \leq G_n\}$ *is a partition of* $\{1,\ldots,n\}$ *such that* $\{(u_{i,n}, \mathbf{x}'_{i,n}) : i \in \mathcal{T}_{g,n}\}$ *are independent over g conditional on* $\mathcal{W}_n$.

Assumption 4.2.2. $\mathbb{P}[\lambda_{min}(\sum_{i=1}^n \mathbf{w}_{i,n}\mathbf{w}'_{i,n}) > 0] \to 1$, $\limsup_{n\to\infty} K_n/n < 1$, $\mathcal{C}_n = O_p(1)$ *and* $\mathbf{\Sigma}_n^{-1} = O_p(1)$

Assumption 4.2.3. $\mathbb{E}[u_{i,n}|\mathcal{X}_n, \mathcal{W}_n] = 0 \quad \forall i, n$, $\chi_n = O(1)$,*and* $\max_{1\leq i\leq n} \|\hat{\mathbf{v}}_{i,n}\|/\sqrt{n} = o_p(1)$.

Assumption 1 defines the sampling structure, in which we allow for arbitrary dependence within clusters of finite but possibly heterogenous size for both the regressors and the errors. In terms of clustering structure, the resulting asymptotics are the same as the usual ones of White (1984) and Liang and Zeger (1986) in which $n, G_n \to \infty$ and $G_n \propto n$. We expect that the results of this chapter would generalize to asymptotics where cluster sizes are allowed to diverge with n and G_n, as considered in Hansen (2007) and Hansen and Lee (2019). It is likely that such extension would require imposing more restrictive conditions on the regression design and the distributional properties of the errors, e.g. stationarity and/or mixing, and we leave it to future work.

Assumption 2 allows for asymptotics where $K_n/n \nrightarrow 0$, while imposing standard restrictions on the regression design and some bounds on the (conditional) higher-order moments of the structural residuals $u_{i,n}$ and $\mathbf{V}_{i,n}$.

The condition on χ_n in Assumption 3 is a requirement on the quality of the linear approximation for the conditional expectation $\mathbb{E}[\mathbf{x}_{i,n}|\mathcal{W}_n]$. The high-level condition $\max_{1\leq i\leq n} \|\hat{\mathbf{v}}_{i,n}\|/\sqrt{n} = o_p(1)$ also places restrictions on the relationship between $\mathbf{x}_{i,n}$ and $\mathbf{w}_{i,n}$ and has a central importance in establishing asymptotic normality of the OLS estimator for $\boldsymbol{\beta}$ and consistency of our proposed variance estimator. Cattaneo et al. (2018b) show that this restriction holds under mild moment conditions when either (i) $K_n/n \to 0$, or (ii) $\chi_n = o(1)$ or (iii) $\max_{1\leq i\leq n} \sum_{i=1}^n \mathbb{1}\{M_{ij,n} \neq 0\} = o_p(n^{1/3})$. While condition (i) is not the case of

primary interest of this chapter, (ii) and (iii) accomodate $K_n/n \nrightarrow 0$ and can be used to verify the high-level condition $\max_{1\leq i\leq n} \|\hat{\mathbf{v}}_{i,n}\|/\sqrt{n} = o_p(1)$ when $\mathbf{w}_{i,n}$ can be interpreted as approximating functions, dummy/discrete variables or fixed effects. See Cattaneo et al. (2018b) for details.

Remark 4.2.1. *In the general formulation provided in Appendix, the assumptions we consider are analogous to those in Cattaneo et al. (2018b) but we also allow for clustered dependence in the errors. The set of restrictions imposed by this framework allows to cover the three leading examples presented in the previous section. A detailed discussion of conditions that satisfy the assumptions in those particular models is provided in Cattaneo et al. (2018b) and their Supplemental Appendix.*

4.3 Main results

This section presents our main theoretical results for inference in linear regression models with many controls and clustering under the set of simplified assumptions presented in Section 3. Proofs of the theorems and other auxiliary results are given in the Appendix for the general case that allows for misspecification bias in the model.

Our first result extends the asymptotic normality result for $\hat{\boldsymbol{\beta}}$ previously derived by Cattaneo et al. (2018b) to the case of clustering.

Theorem 4.3.1. *Suppose Assumptions 1-3 hold. Then,*

$$\boldsymbol{\Omega}_n^{-1/2}\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_d), \qquad \boldsymbol{\Omega}_n = \hat{\boldsymbol{\Gamma}}_n^{-1}\boldsymbol{\Sigma}_n\hat{\boldsymbol{\Gamma}}_n^{-1},$$

where $\boldsymbol{\Sigma}_n = \frac{1}{n}\sum_{g=1}^{G_n}\sum_{i,j\in\mathcal{T}_{g,n}} \hat{\mathbf{v}}_{i,n}\hat{\mathbf{v}}_{j,n}'\mathbb{E}[u_{i,n}u_{j,n}|\mathcal{X}_n, \mathcal{W}_n]$.

Theorem 1 implies that the asymptotic distribution of $\hat{\boldsymbol{\beta}}$ under $K_n/n \nrightarrow 0$ resembles the standard one obtainable under fixed-K_n.[2] As a result, confidence intervals can be constructed using the usual Gaussian approximation and the problem of conducting valid inference reduces to finding a consistent estimator for $\boldsymbol{\Sigma}_n$ under our asymptotics of interest.

[2]From Assumptions 1-3 it also follows that $\hat{\boldsymbol{\Omega}}_n = O_p(1)$, implying that $\hat{\boldsymbol{\beta}}_n$ is $\sqrt{n}$-consistent.

For our discussion of variance estimation, we introduce a new class of estimators. Let $\mathbf{\Omega}_{u,n} = \mathbb{E}[\mathbf{u}_n\mathbf{u}_n'|\mathcal{X}_n, \mathcal{W}_n]$ be the (conditional) variance-covariance matrix of the errors $\mathbf{u}_n = (u_{1,n}, \ldots, u_{n,n})'$ and $L_n = \sum_{g=1}^{G_n}(\#\mathcal{T}_{g,n})^2$ the number of non-zero elements contained in it. We define a general class of cluster-robust estimators for $\mathbf{\Sigma}_n$ of the form

$$\hat{\mathbf{\Sigma}}_n(\boldsymbol{\kappa}_n) = \frac{1}{n}\sum_{g_1=1}^{G_n}\sum_{g_2=1}^{G_n}\sum_{i_1,j_1\in\mathcal{T}_{g_1,n}}\sum_{i_2,j_2\in\mathcal{T}_{g_2,n}} \kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}\hat{\mathbf{v}}_{i_1,n}\hat{\mathbf{v}}_{j_1,n}'\hat{u}_{i_2,n}\hat{u}_{j_2,n}, \quad (4.8)$$

where $\kappa_{g_1,g_2,i_1,j_1,i_2.j_2,n}$ is an entry of the $L_n \times L_n$ matrix $\boldsymbol{\kappa}_n = \boldsymbol{\kappa}_n(\mathbf{w}_{1,n}, \ldots, \mathbf{w}_{1,n})$.[3] Notice that by setting $\boldsymbol{\kappa}_n = \mathbf{I}_{L_n}$ one obtains the usual cluster-robust estimator by Liang and Zeger (1986):

$$\hat{\mathbf{\Sigma}}_n^{\mathrm{LZ}} \equiv \hat{\mathbf{\Sigma}}_n(\mathbf{I}_{L_n}) = \frac{1}{n}\sum_{g=1}^{G_n}\sum_{i,j\in\mathcal{T}_{g,n}} \hat{\mathbf{v}}_{i,n}\hat{\mathbf{v}}_{j,n}'\hat{u}_{i,n}\hat{u}_{j,n}.$$

The next theorem provides an asymptotic representation for this class of estimators.

Theorem 4.3.2. *Suppose Assumptions 1-3 hold.*
If $\|\boldsymbol{\kappa}_n\|_\infty = \max_{(g_1,i_1,j_1)}\sum_{g_2=1}^{G_n}\sum_{i_2,j_2\in\mathcal{T}_{g_2,n}}|\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}| = O_p(1)$, *then*

$$\begin{aligned}\hat{\mathbf{\Sigma}}_n(\boldsymbol{\kappa}_n) = \frac{1}{n}\sum_{g_1=1}^{G_n}\sum_{g_2=1}^{G_n}\sum_{i_1,j_1\in\mathcal{T}_{g_1,n}}\sum_{i_2,j_2\in\mathcal{T}_{g_2,n}} &\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}\hat{\mathbf{v}}_{i_1,n}\hat{\mathbf{v}}_{j_1,n}' \\ &\times\sum_{g_3=1}^{G_n}\sum_{i_3,j_3\in\mathcal{T}_{g_3,n}} M_{i_2j_3,n}M_{j_2i_3,n}\mathbb{E}[u_{i_3,n}u_{j_3,n}|\mathcal{X}_n,\mathcal{W}_n] + o_p(1).\end{aligned} \quad (4.9)$$

Heuristically, in Theorem 2 consistency of $\hat{\boldsymbol{\beta}}_n$ implies that the estimated residuals $\hat{u}_{i,n}$ asymptotically converge to $\tilde{u}_{i,n} = \sum_{j=1}^n M_{ij,n}u_{j,n}$, which are only affected by the estimation noise due to projecting out the high-dimensional covariates $\mathbf{w}_{i,n}$.

[3] In particular, $\kappa_{g_1,g_2,i_1,j_1,i_2.j_2,n}$ corresponds to the $(h(g_1,i_1,j_1), h(g_2,i_2,j_2))$ entry of $\boldsymbol{\kappa}_n$, where $h(g,i,j) = [\sum_{k=0}^{(g-1)}(\#\mathcal{T}_{k,n})^2 + (\#\mathcal{T}_{g,n})(i-1) + j]$ and we adopt the convention that $\#\mathcal{T}_{0,n} = 0$.

The result of Theorem 2 has a central importance in our analysis. First, it immediately provides an explicit characterization for the asymptotic limit of LZ's estimator, as shown in the following corollary.

Corollary 4.3.1. *Suppose the assumptions of Theorem 2 hold. Then,*

$$\hat{\mathbf{\Sigma}}_n^{\mathrm{LZ}} = \frac{1}{n}\sum_{g_1=1}^{G_n}\sum_{g_2=1}^{G_n}\sum_{i_1,j_1\in\mathcal{T}_{g_1,n}}\sum_{i_2,j_2\in\mathcal{T}_{g_2,n}} \hat{\mathbf{v}}_{i_1,n}\hat{\mathbf{v}}'_{j_1,n}M_{i_1j_2,n}M_{j_1i_2,n}\mathbb{E}[u_{i_2,n}u_{j_2,n}|\mathcal{X}_n,\mathcal{W}_n]+o_p(1).$$

Corollary 1 implies that inference based on LZ's clustered standard errors is invalid in general under asymptotics where $K_n/n \nrightarrow 0$. In fact, $\hat{\mathbf{\Sigma}}_n^{\mathrm{LZ}}$ does not converge to the target $\mathbf{\Sigma}_n$ due to elements of $\mathbf{M}_n$ arising in its asymptotic limit. While the sign of the asymptotic bias of LZ's estimator cannot be determined in general, $\hat{\mathbf{\Sigma}}_n^{\mathrm{LZ}}$ will typically underestimate $\mathbf{\Sigma}_n$.[4] Intuitively, the "asymptotic" regression residuals $\tilde{u}_{i,n}$ tend to be smaller than the true residuals as a result of the overfitting due to the high-dimensional controls. In addition, estimated residuals will tend to have lower intra-cluster correlation than the true errors (Bell and Mccaffrey, 2002). Inference based on $\hat{\mathbf{\Sigma}}_n^{\mathrm{LZ}}$ is therefore expected to be asymptotically liberal in most applications.

Furthermore, Theorem 2 suggests that a particular choice of $\boldsymbol{\kappa}_n$ might set the leading term in the expansion (4.9) equal to the target $\mathbf{\Sigma}_n$. Based on this insight, we define the estimator

$$\hat{\mathbf{\Sigma}}_n^{\mathrm{CR}} \equiv \hat{\mathbf{\Sigma}}(\boldsymbol{\kappa}_n^{\mathrm{CR}}) = \frac{1}{n}\sum_{g_1=1}^{G_n}\sum_{g_2=1}^{G_n}\sum_{i_1,j_1\in\mathcal{T}_{g_1,n}}\sum_{i_2,j_2\in\mathcal{T}_{g_2,n}} \kappa^{\mathrm{CR}}_{g_1,g_2,i_1,j_1,i_2,j_2,n}\hat{\mathbf{v}}_{i_1,n}\hat{\mathbf{v}}'_{j_1,n}\hat{u}_{i_2,n}\hat{u}_{j_2,n},$$

where $\boldsymbol{\kappa}_n^{\mathrm{CR}}$ solves the system of $L_n(L_n-1)/2$ equations

$$\sum_{g_2=1}^{G_n}\sum_{i_2,j_2\in\mathcal{T}_{g_2,n}} \kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}M_{i_2,j_3,n}M_{j_2,i_3,n} = \mathbb{1}\{(g_1,i_1,j_1)=(g_3,i_3,j_3)\},$$

$$1\le g_1,g_3\le G_n, i_1,j_1\in\mathcal{T}_{g_1,n}, i_3,j_3\in\mathcal{T}_{g_3,n}.$$

[4] A particular case in which $\operatorname{plim}\hat{\mathbf{\Sigma}}_n^{\mathrm{LZ}} \le \mathbf{\Sigma}_n$ holds in general is when the true residuals are in fact homoskedastic, which can be shown using arguments from Theorem 1 in Bell and Mccaffrey (2002).

The matrix $\boldsymbol{\kappa}_n^{\mathrm{CR}}$ can be characterized in closed form as

$$\boldsymbol{\kappa}_n^{\mathrm{CR}} = (\mathbf{S}_n'(\mathbf{M}_n \otimes \mathbf{M}_n)\mathbf{S}_n)^{-1},$$

where $\otimes$ denotes the Kronecker product and $\mathbf{S}_n$ is the $n^2 \times L_n$ selection matrix with full column rank such that $\mathbf{S}_n'\mathrm{vec}(\boldsymbol{\Omega}_{u,n})$ is the $L_n \times 1$ vector containing the non-zero elements of $\boldsymbol{\Omega}_{u,n}$.

Remark 4.3.1. *When $\boldsymbol{\Omega}_{u,n}$ is assumed to be diagonal, i.e. errors are independent and (conditionally) heteroskedastic, then $G_n = n$, $\mathcal{T}_{i,n} = \{i\}$, $L_n = n$, $\mathbf{S}_n'(\mathbf{M}_n \otimes \mathbf{M}_n)\mathbf{S}_n = \mathbf{M}_n \odot \mathbf{M}_n$ where $\odot$ denotes the Hadamard product, and our estimator reduces to the heteroskedasticity-robust estimator of Cattaneo et al. (2018b).*

In the next theorem we establish consistency of our proposed estimator.

Theorem 4.3.3. *Suppose Assumptions 1-3 hold.*
If $\mathbb{P}[\lambda_{min}(\mathbf{S}_n'(\mathbf{M}_n \otimes \mathbf{M}_n)\mathbf{S}_n) > 0] \to 1$ and $\|\boldsymbol{\kappa}_n^{\mathrm{CR}}\|_\infty = O_p(1)$, then

$$\hat{\boldsymbol{\Sigma}}_n^{\mathrm{CR}} = \boldsymbol{\Sigma}_n + o_p(1).$$

Since $\mathbf{S}_n'(\mathbf{M}_n \otimes \mathbf{M}_n)\mathbf{S}_n$ is observable, the first high-level condition in Theorem 3 is expected to be verified whenever $\mathbf{S}_n'(\mathbf{M}_n \otimes \mathbf{M}_n)\mathbf{S}_n$ is invertible. The second high-level condition could be verified using Theorem 1 of Varah (1975), which provides a bound for $\|\boldsymbol{\kappa}_n^{\mathrm{CR}}\|_\infty$ under the condition that $\mathbf{S}_n'(\mathbf{M}_n \otimes \mathbf{M}_n)\mathbf{S}_n$ is diagonally dominant. In simulations we find that diagonal dominance typically does not hold but our high-level condition is verified in a wide range of models and designs, as shown in Section 5.[5]

Remark 4.3.2. *Notice that explicit computation of $\boldsymbol{\kappa}_n^{CR}$ is not required for the purpose of variance estimation. Having defined $\hat{\mathbf{V}}_n = (\hat{\mathbf{v}}_{1,n}, \ldots, \hat{\mathbf{v}}_{n,n})'$ and $\mathbf{c}_n = (\mathbf{S}_n'(\mathbf{M}_n \otimes \mathbf{M}_n)\mathbf{S}_n)^{-1}\mathbf{S}_n(\hat{\mathbf{u}}_n \otimes \hat{\mathbf{u}}_n)$, one has $\mathrm{vec}(\hat{\boldsymbol{\Sigma}}_n^{\mathrm{CR}}) = (\hat{\mathbf{V}}_n \otimes \hat{\mathbf{V}}_n)'\mathbf{S}_n\mathbf{c}_n$. As*

[5]Cattaneo et al. (2018b) instead develop their theory under the requirement that $\mathbf{M}_n \odot \mathbf{M}_n$ is diagonally dominant. It would be interesting to investigate whether this requirement could be relaxed in practice.

a result, computing our variance estimator only requires to solve the linear system $(\mathbf{S}_n'(\mathbf{M}_n \otimes \mathbf{M}_n)\mathbf{S}_n)\mathbf{c}_n = \mathbf{S}_n(\hat{\mathbf{u}}_n \otimes \hat{\mathbf{u}}_n)$ *for* $\mathbf{c}_n$.

The structure of our proposed estimator is related to the cluster-robust variance estimator proposed by Bell and Mccaffrey (2002), which corresponds to a particular choice of block-diagonal $\boldsymbol{\kappa}_n$ that sets the the bias of the variance estimator to 0 only in the special case in which $\boldsymbol{\Omega}_{u,n} = \sigma^2\mathbf{I}_n$, i.e. the true residuals are in fact homoskedatic. Differently from Bell and Mccaffrey (2002), our choice of correction matrix $\boldsymbol{\kappa}_n^{\mathrm{CR}}$ induces an averaging over cross-products of estimated residuals not just within but also across clusters, thus allowing to set the leading term in expansion (4.9) equal to $\boldsymbol{\Sigma}_n$ in general.

The results of this chapter can be easily extended to a more general version of the variance estimators, described in Section C.6 of the Appendix, that allows to impose within-cluster zero restrictions on the variace-covariance matrix of the errors. In such form, our proposed estimator reduces to the one of Stock and Watson (2008) in the case of one-way fixed effects panel data models with zero restrictions on the conditional autocovariances of U_{it} within entities. While our results cover a much wider class of models, they also partly improve on Stock and Watson (2008) as we do not require $(\mathbf{X}_{i1}', \ldots, \mathbf{X}_{iT}', U_{i1}, \ldots, U_{iT})$ to be i.i.d. nor we require $(\mathbf{X}_{it}, U_{it})$ to be stationary.

4.3.1 Consistency of Liang and Zeger's estimator

Although consistency of $\hat{\boldsymbol{\Sigma}}_n^{\mathrm{CR}}$ is derived under asymptotic sequences that allow but do not require $K_n/n \nrightarrow 0$, it is still desirable to establish consistency of LZ's estimator under some sufficiently slow rate of growth for K_n. For this purpose, define $\mathbf{w}_{i,n}^* = \hat{\boldsymbol{\Sigma}}_{\mathbf{w},n}^{-1/2}\mathbf{w}_{i,n}$, where $\hat{\boldsymbol{\Sigma}}_{\mathbf{w},n}^{1/2}$ is the unique symmetric positive definite $K_n \times K_n$ matrix such that $\hat{\boldsymbol{\Sigma}}_{\mathbf{w},n}^{1/2}\hat{\boldsymbol{\Sigma}}_{\mathbf{w},n}^{1/2} = \frac{1}{n}\sum_{i=1}^n \mathbf{w}_{i,n}\mathbf{w}_{i,n}'$. The following theorem provides sufficient conditions for consistency of LZ's cluster-robust estimator.

Theorem 4.3.4. *Suppose Assumptions 1-3 hold and that* $\max_{i,j}\mathbb{E}[w_{ij,n}^{*2}] = O(1)$. *If* $K_n^2/n \to 0$, *then*

$$\hat{\boldsymbol{\Sigma}}_n^{\mathrm{LZ}} = \boldsymbol{\Sigma}_n + o_p(1). \tag{4.10}$$

Moreover, if $\mathbb{E}[u_{i,n}^2|\mathcal{X}_n,\mathcal{W}_n] = \sigma_n^2\ \forall i$, *and* $\mathbb{E}[u_{i,n}u_{j,n}|\mathcal{X}_n,\mathcal{W}_n] = 0\ \forall i \neq j$, *then (4.10) holds under* $K_n/n \to 0$.

Although we can only prove consistency of LZ's estimator under $K_n^2/n \to 0$, we speculate that $K_n/n \to 0$ might suffice in general. We leave the refinement of this result for future work.[6]

4.4 Simulations

This section reports the findings of a simulation study that investigates the finite sample behaviour of the cluster-robust variance estimators studied in this chapter. We consider three distinct designs motivated by the empirical examples covered by the theoretical framework of this chapter: the linear regression models with increasing dimension, the semiparametric partially linear model and the fixed effects panel data regression model.

4.4.1 Results - Linear regression model with increasing dimension

The chosen designs for our Monte Carlo experiments closely resemble those of Cattaneo et al. (2018b), also borrowing from specifications in Stock and Watson (2008) and MacKinnon (2013). The data generating process (DGP) for the linear regression model with many covariates is:

$$\begin{aligned}
&y_{gi} = \beta x_{gi} + \boldsymbol{\gamma}_n'\mathbf{w}_{gi} + U_{gi},\\
&x_{gi}|\mathbf{w}_{gi} \sim \mathcal{N}(0,\sigma_{x,gi}^2), \quad \sigma_{x,gi}^2 = \varkappa_x(1+(\boldsymbol{\iota}'\mathbf{w}_i)^2),\\
&U_{gi} = (\rho\mathbb{1}(x_{gi}\geq 0) - \rho(1-\mathbb{1}(x_{gi}\geq 0))U_{g,i-1} + \varepsilon_{gi}, \quad \varepsilon_{gi}\sim\mathcal{N}(0,1),\\
&u_{g1} \sim \mathcal{N}(0,\sigma_{u1}^2), \quad \sigma_{u1}^2 = \varkappa_{u1}(1+(t(x_{g1})+\boldsymbol{\iota}'\mathbf{w}_{g1})^2),\\
&i = 1,\dots,n/G, \quad g = 1,\dots,G, \quad n = 700,
\end{aligned} \tag{4.11}$$

where $\mathbf{w}_{gi} \overset{i.i.d.}{\sim} \mathcal{U}(-1,1)$, $\boldsymbol{\iota} = (1,1,\dots,1)'$, $\beta = 1$, $\boldsymbol{\gamma} = \mathbf{0}$, $\rho = 0.3$, the constants $\varkappa_x$ and $\varkappa_{u1}$ are chosen so that $\mathbb{V}[x_{gi}] = \mathbb{V}[U_{g1}] = 1$ and $t(a) = a\mathbb{1}(-2 \leq a \leq$

[6]Theorem 4 also states that $K_n/n \to 0$ is sufficient for consistency of LZ's estimator in the special case of homoskedastic errors.

$2) + 2\text{sgn}(a)(1 - \mathbb{1}(-2 \leq a \leq 2))$.

Table 4.2 reports the results of our experiment for five dimensions of $\mathbf{w}_{gi}$: $K \in \{1, 71, 141, 211, 281\}$, where the first covariate is an intercept, as well as three different numbers of equal-sized clusters: $G \in \{175, 70, 35\}$. We consider three different estimators for the variance of the OLS estimator $\hat{\beta}$: the unfeasible estimator based on $\hat{\mathbf{\Sigma}}_n^{\texttt{Unf}} = \frac{1}{n}\sum_{g=1}^{G_n}\sum_{i,j\in\mathcal{T}_{g,n}} \hat{\mathbf{v}}_{i,n}\hat{\mathbf{v}}_{j,n}' U_{i,n}U_{j,n}$ that makes use of the true error realizations, the classical estimator by LZ and our proposed cluster-robust formula, as previously defined. For each of these estimators, we report the bias (expressed in percentage), the standard deviation (denoted by Std.) and the empirical coverage probability (denoted by $\hat{p}; \alpha$) of the Gaussian confidence interval of the form:

$$\mathsf{I}_\ell \doteq \left[\hat{\beta} - \Phi^{-1}(1-\alpha/2)\cdot\sqrt{\frac{\hat{\Omega}_\ell}{n}}, \hat{\beta} - \Phi^{-1}(\alpha/2)\cdot\sqrt{\frac{\hat{\Omega}_\ell}{n}}\right], \quad \hat{\Omega}_\ell = \hat{\Gamma}^{-1}\hat{\Sigma}^\ell\hat{\Gamma}^{-1},$$

where Φ^{-1} denotes the inverse of the standard normal cumulative distribution function Φ, $\hat{\Sigma}_\ell$ with $\ell \in \{\texttt{Unf, LZ, CR}\}$ corresponds to the variance estimators already discussed and we set $\alpha = 0.05$.

The findings from this experiment are in line with our theoretical predictions. Firstly, we find that inference based on LZ's clustered standard errors formula is highly inaccurate. In fact, its bias quickly increases with the dimensionality of the model, resulting in substantial undercoverage even for $K/n = 0.101$. On the other hand, our proposed estimator performs well, with negligible bias and close-to-correct empirical coverage even for $K/n = 0.401$. Such improvement in inference accuracy compared to LZ's estimator is achieved in spite of a decrease in relative precision. As expected, the performance of all estimators is adversely affected by a reduction in the number of clusters. In Table 4.6 we also report on the behaviour of $\|\boldsymbol{\kappa}_n^{\text{CR}}\|_\infty$ in this design; we find that $\|\boldsymbol{\kappa}_n^{\text{CR}}\|_\infty$ not only seems to be bounded but even decreasing as n grows.[7]

Analogous results are found for a different version of this experiment that con-

[7]Notice that diagonal dominance of $\mathbf{S}_n'(\mathbf{M}_n \otimes \mathbf{M}_n)\mathbf{S}_n$ does not hold in any of the simulations carried out in this section.

siders independent and discrete controls constructed as $\mathbb{1}\{\mathcal{N}(0,1) \geq 1\}$, as reported in Tables 4.3 and 4.7.

4.4.2 Results - Semiparametric partially linear model

The experimental design chosen for the semiparametric partially linear model takes the form:

$$
\begin{aligned}
&y_{gi} = \beta x_{gi} + g(\mathbf{z}_{gi}) + U_{gi}, \\
&x_{gi} = h(\mathbf{z}_{gi}) + v_i, \quad v_{gi}|\mathbf{z}_{gi} \sim \mathcal{N}(0, \sigma^2_{v,gi}), \quad \sigma^2_{vi} = \varkappa_v(1 + (\boldsymbol{\iota}'\mathbf{z}_{gi})^2), \\
&U_{gi} = (\rho \mathbb{1}(z_{1,gi} \geq 0) - \rho(1 - \mathbb{1}(z_{1,gi} \geq 0))U_{g,i-1} + \varepsilon_{gi}, \quad \varepsilon_{gi} \sim \mathcal{N}(0,1), \\
&u_{g1} \sim \mathcal{N}(0, \sigma^2_{u1}), \quad \sigma^2_{u1} = \varkappa_{u1}(1 + (t(x_{gi}) + \boldsymbol{\iota}'\mathbf{z}_{gi})^2), \\
&i = 1, \ldots, n/G, \quad g = 1, \ldots, G, \quad n = 700,
\end{aligned}
\tag{4.12}
$$

where $\dim(\mathbf{z}_{gi}) = 6$, $\mathbf{z}_{gi} = (z_{1,gi}, \ldots, z_{6,gi})'$ with $z_{\ell,gi} \overset{i.i.d.}{\sim} \mathcal{U}(-1,1)$, $\ell = 1, \ldots, 6$. The unknown regressions functions are set to $g(\mathbf{z}_{gi}) = \exp\big(-\|\mathbf{z}_{gi}\|^{1/2}\big)$ and $h(\mathbf{z}_{gi}) = \exp\big(\|\mathbf{z}_{gi}\|^{1/2}\big)$, and the constants $\varkappa_v$ and $\varkappa_{u1}$ are again chosen so that $\mathbb{V}[x_{gi}] = \mathbb{V}[u_{g1}] = 1$. Similarly to the previous simulation, we set $\beta = 1$ and $\rho = 0.3$.

To construct the covariates $\mathbf{w}_{gi}$ entering the estimated linear regression model $y_{gi} = \boldsymbol{\beta}'\mathbf{x}_{gi} + \boldsymbol{\gamma}_n'\mathbf{w}_{gi} + u_{gi}$, we consider power series expansions. The table below gives a summary of the expansions considered, where $\mathbf{w}_{gi} = \mathbf{p}(\mathbf{z}_{gi}; K)$ for $K \in \{1, 7, 13, 28, 34, 84, 90, 210, 216\}$ is defined as follows:

Table 4.1: Polynomial Basis Expansion: $\dim(\mathbf{z}_{gi}) = 6$ and $n = 700$

K	$\mathbf{p}(\mathbf{z}_{gi}; K)$	K/n
1	1	0.001
7	$(1, z_{1,gi}, z_{2,gi}, z_{3,gi}, z_{4,gi}, z_{5,gi}, z_{6,gi})'$	0.010
13	$(\mathbf{p}(\mathbf{z}_{gi}; 7)', z_{1,gi}^2, z_{2,gi}^2, z_{3,gi}^2, z_{4,gi}^2, z_{5,gi}^2, z_{6,gi}^2)'$	0.019
28	$\mathbf{p}(\mathbf{z}_{gi}; 13)$ + first-order interactions	0.040
34	$(\mathbf{p}(\mathbf{z}_{gi}; 28)', z_{1,gi}^3, z_{2,gi}^3, z_{3,gi}^3, z_{4,gi}^3, z_{5,gi}^3, z_{6,gi}^3)'$	0.049
84	$\mathbf{p}(\mathbf{z}_{gi}; 13)$ + second-order interactions	0.120
90	$(\mathbf{p}(\mathbf{z}_{gi}; 84)', z_{1,gi}^4, z_{2,gi}^4, z_{3,gi}^4, z_{4,gi}^4, z_{5,gi}^4, z_{6,gi}^4)'$	0.129
210	$\mathbf{p}(\mathbf{z}_{gi}; 90)$ + third-order interactions	0.300
216	$(\mathbf{p}(\mathbf{z}_{gi}; 210)', z_{1,gi}^5, z_{2,gi}^5, z_{3,gi}^5, z_{4,gi}^5, z_{5,gi}^5, z_{6,gi}^5)'$	0.309

Source: Cattaneo, Jansson, and Newey (2018b, Supplemental Appendix).

The results for this experiment are given in Table 4.4, in which we only report $K \in \{1, 13, 34, 90, 216\}$ for reasons of parsimony. The numerical findings are largely consistent with those reported for the other two simulation models. Although $\|\boldsymbol{\kappa}_n^{\mathsf{CR}}\|_\infty$ has bigger magnitude in this setting compared to the other simulation models, it still appears to be bounded (see Table 4.8).

The main difference between this setting and the linear model with increasing dimension considered previously is that the unfeasible estimator that uses realizations of the true structural disturbances is free not just from estimation error but also specification error, which in turn affects LZ's and our proposed estimator when K is small; in addition, the degree of heteroskedasticity and dependence in the errors is invariant with respect to the dimensionality of the model, since it only depends on x_{gi} and $\mathbf{z}_{gi}$ but not $\mathbf{w}_{gi}$.

4.4.3 Results - Fixed effects panel data regression model

For fixed effects panel data regression model we consider the following specification:

$$y_{it} = \beta x_{it} + \alpha_i + e_{d_{it}} + u_{it}, \qquad i = 1, \dots, N, \qquad t = 1, \dots, T, \tag{4.13}$$

where α_i is a time-invariant individual effect and $e_{d_{it}}$ are unobserved factors common to all observations sharing the same value of the indexing variable $d_{it} \in \{1, \dots, N_d\}$. This model coincides with the one studied in Verdier (2018), whose theory and simulation results concern the case of two-way clustering. We instead consider the case of one-way clustering at the individual level as we postulate the following DGP:

$$\begin{aligned}
&y_{it} = \beta x_{it} + \alpha_i + e_{d_{it}} + U_{it}, \\
&x_{it}|\mathbf{z}_{it} \sim \mathcal{N}(0, \sigma^2_{x,it}), \quad \sigma^2_{x,gi} = \varkappa_x(1 + (\boldsymbol{\iota}'\mathbf{z}_{it})^2), \\
&U_{it} = (\rho \mathbb{1}(x_{it} \geq 0) - \rho(1 - \mathbb{1}(x_{it} \geq 0))U_{i,t-1} + \varepsilon_{it}, \quad \varepsilon_{it} \sim \mathcal{N}(0,1), \\
&u_{i1} \sim \mathcal{N}(0, \sigma^2_{u1}), \quad \sigma^2_{u1} = \varkappa_{u1}(1 + (t(x_{i1}) + \boldsymbol{\iota}'\mathbf{z}_{i1})^2), \\
&i = 1, \dots, N, \quad t = 1, \dots, T,
\end{aligned} \tag{4.14}$$

where $\dim(\mathbf{z}_{it}) = 6$, $\mathbf{z}_{it} = (z_{1,it}, \dots, z_{6,it})'$ with $z_{\ell,gi} \overset{i.i.d.}{\sim} \text{Uniform}(-1, 1)$, $\ell = 1, \dots, 6$, the constants $\varkappa_x$ and $\varkappa_{u1}$ are chosen so that $\mathbb{V}[x_{it}] = \mathbb{V}[U_{i1}] = 1$, the function $t(\cdot)$ is as previously defined and we set $\beta = 1$ and $\alpha_i = e_{d_{it}} = 0$. For the purpose of estimation, we transform (4.13) by partialling out the individual fixed effects α_i, so that the estimated model $\tilde{y}_{it} = \beta \tilde{x}_{it} + \tilde{e}_{d_{it}} + \tilde{u}_{it}$ has $\dim(\mathbf{w}_i) = N_d$.[8] We consider $G = N = [700/T]$ for $T \in \{4, 10, 20\}$, as well as $N_d = 700/r$ for $r \in \{700, 10, 5, 4, 3\}$, so that the total sample size is always roughly $n = 700$. Tables 4.5 and 4.9 report the numerical findings of this experiment, which are consistent with our theoretical predictions and in line with the results obtained for the other simulation designs.

[8]The motivation for this transformation is that $(\mathbf{S}_n'(\mathbf{M}_n \otimes \mathbf{M}_n)\mathbf{S}_n)$ is not invertible when the controls $\mathbf{w}_{i,n}$ include indicators for the clusters (see, e.g., Stock and Watson, 2008). Notice that partialling out the fixed effects does not affect the correlation structure of the errors.

4.5 Empirical illustration

This section illustrates the use of the inference methods discussed in this chapter by revisiting Donohue and Levitt (2001) study of the impact of abortion on crime rates.

Donohue and Levitt (2001, henceforth DL) put forward the hypothesis that the legalization of abortion in the United States in the 1970s played a major role in explaining the sharp decline in crime observed two decades later. In particular, they describe two causal channels through which abortion might affect crime. The first is that abortion reduces the absolute size of a cohort, resulting in lower crime 15-25 years later, when its members are at the highest risk of engaging in criminal activities. The second channel is ascribed to the increased control over fertility that abortion provides to women. In fact, women may use abortion to optimize the timing of childbearing, thus ensuring that the child grows in a more favourable environment, e.g. when a father is present in the family, the mother is better educated and household income is stable. As a result, increased access to abortion is expected to cause a reduction in crime levels even if fertility rates were to remain constant.

In order to estimate the impact of abortion on crime, Donohue and Levitt (2001) consider state-level yearly data for the period 1985-1997 and propose a model for crime rates whose basic specification is

$$y_{cit} = \beta_c a_{cit} + \boldsymbol{\delta}_c' \mathbf{z}_{it} + \theta_{ci} + \lambda_{ct} + u_{cit}, \tag{4.15}$$

where i indexes states, t indexes the time period, $c \in \{\text{violent, property, murder}\}$ indexes the type of crime, y_{cit} is the crime-rate for crime type c; a_{cit} is measure of abortion rate relevent for crime type c; z_{it} is a set of time-varying state-specific controls consisting of the log of lagged prisoners per capita, the log of lagged police per capita, the unemployment rate, per-capita income, the poverty rate, AFDC generosity at time $t-15$, a dummy for concealed weapons law and beer consumptions; θ_{ci} are state fixed effects; and λ_{ct} are time fixed effects. Further details on data definitions and the institutional background can be found in the original paper.

The results from estimating the baseline model in (4.15) are reported in Table

4.10 and resemble those in Donohue and Levitt (2001), although not identical as we have excluded Washington DC from the sample.[9] Following Donohue and Levitt (2001), we report standard errors clustered at the state level. These estimates indicate a strong (and statistically significant) negative association between abortion and crime, as they imply that an increase in the abortion rate of 100 per 1,000 live births is associated with a reduction in crime rates between 9 and 13 per cent, depending on the type of crime. However, the extent to which this association can be interpreted as causal crucially depends on the assumption that abortion rates can be taken as random after controlling for a national trend, time-invariant state-specific confounders and z_{it}. Even if one believes that abortion rates can be taken as exogenous conditional on the the controls included by Donohue and Levitt (2001), one can still expect the assumption that they enter the structural equation for crime rates linearly as in (4.15) to be too restrictive. For example, Foote and Goetz (2008) have argued that the results in Donohue and Levitt (2001) might not be robust to the inclusion of state-specific trends.[10]

For these reasons, we consider a model for crime rates and abortion in which the controls $\mathbf{z}_{it}$ are allowed to enter in a much more flexible way compared to Donohue and Levitt (2001). In particular, we consider a version of the high-dimensional regression model studied in this chapter where in addition to the controls included by Donohue and Levitt (2001), we include first-order interactions, quadratics, cumulative values and interactions of those variables and their initial values with a quadratic trend; in addition, we also include the interaction between the initial level of abortion and a quadratic trend. Once we stack all these regressors and the time-effects λ_{ct} in the vector $\mathbf{w}_{it}$ and absorb the state-effects, we obtain a regression model of the same form as (4.1):

$$\tilde{y}_{cit} = \beta_c \tilde{a}_{cit} + \boldsymbol{\gamma}_c' \tilde{\mathbf{w}}_{it} + \tilde{u}_{cit}, \tag{4.16}$$

[9] We exclude Washington DC for simplicity, as it produces similar results to Donohue and Levitt (2001) and circumvents the need to introduce the estimation weights used in their paper.

[10] In a response to Foote and Goetz (2008), Donohue and Levitt (2008) reexamine their orginal study and use a longer panel to argue that their original results are robust to the inclusion of state-specific linear time trends.

where the dimension of the high-dimensional controls is $K_n = 105$, resulting in $K_n/n \approx 0.161$. Estimates for the causal effect of abortion on crime based on (4.16) are given in Table 4.10, where we also report standard errors based on the variance estimators considered in this chapter. These estimates are qualitatively similar to those obtained for the baseline model considered in Donohue and Levitt (2001), and interestingly imply an even more sizeable negative effect of abortion on crime rates for all types of crime. The statistical significance of these effects however crucially depends on the choice of standard errors. In fact, clustered standard errors based on the variance estimator proposed in this chapter are between 42 and 74 per cent bigger than the traditional clustered stardard errors by Liang and Zeger (1986), depending on the type of crime. In the case of violent crime, for example, the estimated coefficient for abortion rates has associated p-value below 1 per cent when traditional clustered standard errors are used, while the use of our proposed standard errors leads to failure to reject the hypothesis of no effect of abortion on crime at the 5 per cent level.

This empirical illustration showcases the relevance of the inference methods proposed in this chapter. In this particular application, the inclusion of many controls arises naturally as a way to flexibly control for observable state-level characteristics and trends that are allowed to depend on those characteristics. Our approach in this particular application resembles the one adopted by Belloni et al. (2013), who also re-examine the empirical setting in Donohue and Levitt (2001) to illustrate the use of their proposed inference method for treatment effects with many controls based on LASSO double-selection. They consider a similar specification of the high-dimensional model in (4.16) but allow for an even more flexible specification that includes higher-order interactions of the variables we consider (and a few additional ones, such as initial differences of $\mathbf{z}_{it}$) with cubic trends, which gives $K_n/n \approx 0.500$ in their application.[11] While Belloni et al.'s (2013) method is naturally suited to handle such large number of controls, its validity relies on the assumption that the effect of confounding factors can be controlled for by a small

[11] Interestingly, their estimates imply statistically non-significant impact of abortion on all types of crime.

number of variables ("approximate sparsity"). Our proposed inference procedure therefore offers a valuable alternative to selection-based methods in settings where the inclusion of a relatively large number of controls is expected to yield a reasonable approximation of the structural relationship of interest, while circumventing the need to impose requirements of sparsity on the model.

4.6 Conclusions

This chapter presented inference results for the OLS estimator of a subset of coefficients in linear regression models with many controls and clustering. We show that the usual cluster-robust variance estimator by Liang and Zeger (1986) does not deliver consistent standard errors when the number of controls is a non-vanishing fraction of the sample size, typically resulting in confidence intervals with coverage below the nominal size. We then propose a new clustered standard error formula that is robust to the inclusion of many controls. Monte Carlo evidence supports our theoretical results and shows that our proposed variance estimator performs well in finite samples.

While our results are presented for the case of one-way clustering, we expect that they can be easily adapted to the generalisation of our methods to multi-way clustering proposed by Verdier (2020). It would also be of interest to investigate whether the analysis of this chapter could be extended to cases where variance estimation does not rely on zero restrictions on the covariance matrix of the errors, e.g. when time series or spatial dependence in the errors is assumed.

Table 4.2: Monte Carlo simulations for linear regression model with increasing dimension (continuous controls), $n = 700$

	$\hat{\beta}$		Unfeasible			Classical			Robust		
	Mean	Variance	Bias (%)	Std.	$\hat{p}$; .05	Bias (%)	Std.	$\hat{p}$; .05	Bias (%)	Std.	$\hat{p}$; .05
$\boldsymbol{G = 175}$											
$K/n = 0.001$	1.00	.0042	1.29	.0011	.047	-.101	.0011	.051	.360	.0011	.051
$K/n = 0.101$	1.00	.752	-1.42	.492	.044	-19.3	.375	.080	-4.45	.460	.055
$K/n = 0.201$	.981	2.89	-.578	2.09	.043	-19.6	1.21	.106	-4.41	1.87	.053
$K/n = 0.301$	1.00	6.56	2.04	4.18	.042	-38.0	2.02	.143	-1.71	3.93	.054
$K/n = 0.401$	1.02	11.7	-1.42	7.28	.044	-53.2	2.76	.179	-6.17	6.76	.057
$\boldsymbol{G = 70}$											
$K/n = 0.001$	1.00	.0027	-2.00	7.93×10^{-4}	.053	-4.02	7.73×10^{-4}	.064	-3.84	7.75×10^{-4}	.064
$K/n = 0.101$	1.01	.310	-3.16	.317	.039	-21.4	.235	.072	-7.15	.289	.052
$K/n = 0.201$	1.02	1.10	2.45	1.15	.036	-30.0	.702	.096	-1.80	1.07	.050
$K/n = 0.301$	1.04	2.60	1.77	2.67	.034	-42.5	1.273	.133	-3.32	2.48	.052
$K/n = 0.401$	1.01	4.82	-4.27	4.50	.039	-54.7	1.690	.181	-10.3	4.12	.058
$\boldsymbol{G = 35}$											
$K/n = 0.001$	1.00	.0021	.51	7.00×10^{-4}	.047	-2.86	6.80×10^{-4}	.062	-2.70	6.82×10^{-4}	.062
$K/n = 0.101$	1.00	.156	1.71	.027	.037	-18.3	.194	.065	-4.3	.238	.045
$K/n = 0.201$	1.01	.558	1.73	.858	.027	-31.6	.510	.089	-5.12	.770	.048
$K/n = 0.301$	1.00	1.33	-.16	1.81	.033	-45.6	.840	.140	-9.10	1.61	.070
$K/n = 0.401$	.973	2.41	.70	3.05	.037	-53.7	1.12	.186	-9.50	2.72	.083

Notes: Simulation results based on 5,000 replications. DGP as described in Equation (4.11).

Table 4.3: Monte Carlo simulations for linear regression model with increasing dimension (discrete controls), $n = 700$

	$\hat{\beta}$		Unfeasible			Classical			Robust		
	Mean	Variance	Bias (%)	Std.	$\hat{p}$; .05	Bias (%)	Std.	$\hat{p}$; .05	Bias (%)	Std.	$\hat{p}$; .05
$\boldsymbol{G = 175}$											
$K/n = 0.001$	1.00	.0043	-.46	.0011	.048	-1.70	.0011	.051	-1.50	.0011	.051
$K/n = 0.101$	1.00	.0019	-1.84	3.01×10^{-4}	.052	-18.4	2.60×10^{-4}	.072	-3.03	3.21×10^{-4}	.057
$K/n = 0.201$	1.00	.0020	1.63	3.10×10^{-4}	.048	-20.9	2.39×10^{-4}	.086	-.410	3.60×10^{-4}	.053
$K/n = 0.301$	1.00	.0023	-3.40	3.52×10^{-4}	.055	-34.2	2.41×10^{-4}	.114	-5.20	4.53×10^{-4}	.064
$K/n = 0.401$	1.00	.0026	-1.00	4.05×10^{-4}	.050	-41.5	2.44×10^{-4}	.140	-2.48	5.88×10^{-4}	.061
$\boldsymbol{G = 70}$											
$K/n = 0.001$	1.00	.0026	-.161	8.07×10^{-4}	.046	-2.17	7.88×10^{-4}	.056	-1.99	7.91×10^{-4}	.056
$K/n = 0.101$	1.00	.0018	-1.48	3.72×10^{-4}	.049	-14.4	3.27×10^{-4}	.075	-3.88	4.07×10^{-4}	.059
$K/n = 0.201$	1.00	.0020	-.773	4.00×10^{-4}	.052	-22.9	3.16×10^{-4}	.091	-3.26	4.87×10^{-4}	.062
$K/n = 0.301$	1.00	.0024	-4.56	4.60×10^{-4}	.054	-35.3	3.13×10^{-4}	.124	-8.16	6.11×10^{-4}	.071
$K/n = 0.401$	1.00	.0027	-1.16	5.20×10^{-4}	.053	-42.5	3.13×10^{-4}	.146	-5.96	8.00×10^{-4}	.076
$\boldsymbol{G = 35}$											
$K/n = 0.001$	1.00	.0021	.233	7.26×10^{-4}	.046	-2.97	7.06×10^{-4}	.063	-2.97	7.07×10^{-4}	.063
$K/n = 0.101$	1.00	.0018	1.29	4.85×10^{-4}	.046	-13.0	4.19×10^{-4}	.075	-2.77	5.21×10^{-4}	.062
$K/n = 0.201$	1.00	.0020	.570	5.32×10^{-4}	.046	-23.1	4.11×10^{-4}	.099	-4.24	6.37×10^{-4}	.069
$K/n = 0.301$	1.00	.0023	-.892	6.08×10^{-4}	.052	-34.2	4.07×10^{-4}	.122	-8.16	8.03×10^{-4}	.080
$K/n = 0.401$	1.00	.0028	-4.11	6.97×10^{-4}	.055	-45.4	4.06×10^{-4}	.160	-13.9	11.0×10^{-4}	.102

Notes: Simulation results based on 5,000 replications. DGP as described in Equation (4.11), with $w_{\ell,gi} = \mathbb{1}\{\mathcal{N}(0,1,) \geq 1\}, \quad \forall \ell, g, i.$

Table 4.4: Monte Carlo simulations for semiparametric partially linear model, $n = 700$

	$\hat{\beta}$		Unfeasible			Classical			Robust		
	Mean	Variance	Bias (%)	Std.	$\hat{p}$; .05	Bias (%)	Std.	$\hat{p}$; .05	Bias (%)	Std.	$\hat{p}$; .05
$\boldsymbol{G = 175}$											
$K/n = 0.001$	.984	.0436	-1.42	.0192	.048	-2.84	.0182	.053	-2.70	.0183	.052
$K/n = 0.019$	1.00	.0519	-1.80	.0231	.046	-5.33	.0210	.052	-3.23	.0220	.049
$K/n = 0.049$	.994	.0515	2.22	.0225	.044	-5.14	.0195	.056	.680	.0219	.049
$K/n = 0.129$	.993	.0570	.922	.0231	.050	-15.0	.0174	.078	-.760	.0230	.056
$K/n = 0.309$	.993	.0733	.824	.0287	.049	-31.0	.0170	.103	-1.94	.0304	.056
$\boldsymbol{G = 70}$											
$K/n = 0.001$	.986	.0433	-.89	.0196	.050	-2.31	.0189	.054	-2.20	.0183	.054
$K/n = 0.019$	1.00	.0209	.28	.0132	.042	-4.13	.0120	.052	-2.08	.0125	.050
$K/n = 0.049$	1.00	.0218	-.59	.0136	.041	-8.36	.0117	.0548	-3.07	.0130	.049
$K/n = 0.129$	1.00	.0241	-.04	.0153	.046	-16.5	.0113	.074	-2.87	.0148	.059
$K/n = 0.309$	1.00	.0315	-1.44	.0172	.043	-33.0	.0100	.115	-5.68	.0174	.067
$\boldsymbol{G = 35}$											
$K/n = 0.001$	.988	.0098	-1.37	.0083	.042	-5.01	.0079	.059	-4.88	.0079	.059
$K/n = 0.019$	1.00	.0121	-3.54	.0108	.050	-9.40	.0095	.066	-3.41	.0099	.061
$K/n = 0.049$	1.00	.0116	-.41	.0092	.038	-9.70	.0079	.064	-4.64	.0087	.060
$K/n = 0.129$	1.00	.0127	1.70	.0100	.033	-16.0	.0073	.070	-3.02	.0096	.058
$K/n = 0.309$	1.00	.0169	1.29	.0134	.040	-31.9	.0074	.115	-6.50	.0118	.065

Notes: Simulation results based on 5,000 replications. DGP as described in Equation (4.12).

Table 4.5: Monte Carlo simulations for two-way fixed effects panel data regression model, $n = 700$

	$\hat{\beta}$		Unfeasible			Classical			Robust		
	Mean	Variance	Bias (%)	Std.	$\hat{p}$; .05	Bias (%)	Std.	$\hat{p}$; .05	Bias (%)	Std.	$\hat{p}$; .05
$\boldsymbol{G = 175}$											
$K/n = 0.001$	1.00	.0025	.435	7.02×10^{-4}	.045	-1.28	6.72×10^{-4}	.053	-1.28	6.72×10^{-4}	.053
$K/n = 0.100$	1.00	.0029	-1.26	7.74×10^{-4}	.048	-17.6	6.07×10^{-4}	.076	-3.08	7.82×10^{-4}	.056
$K/n = 0.200$	1.00	.0033	.350	8.30×10^{-4}	.046	-30.4	5.18×10^{-4}	.105	-2.02	8.80×10^{-4}	.056
$K/n = 0.250$	1.00	.0036	-1.22	8.71×10^{-4}	.052	-38.0	5.04×10^{-4}	.127	-3.86	9.95×10^{-4}	.062
$K/n = 0.333$	1.00	.0042	-1.51	9.61×10^{-4}	.051	-48.8	4.63×10^{-4}	.167	-5.18	.0012	.066
$\boldsymbol{G = 70}$											
$K/n = 0.001$	1.00	.0020	-2..30	5.15×10^{-4}	.049	-4.60	5.04×10^{-4}	.059	-4.50	5.05×10^{-4}	.059
$K/n = 0.100$	1.00	.0022	-1.58	5.83×10^{-4}	.049	-15.4	4.89×10^{-4}	.075	-3.99	6.04×10^{-4}	.061
$K/n = 0.200$	1.00	.0024	-.482	6.20×10^{-4}	.046	-25.9	4.41×10^{-4}	.100	-4.13	6.92×10^{-4}	.062
$K/n = 0.250$	1.00	.0026	.498	6.53×10^{-4}	.045	-30.8	4.32×10^{-4}	.111	-3.63	7.66×10^{-4}	.063
$K/n = 0.333$	1.00	.0029	-1.84	7.13×10^{-4}	.051	-41.2	4.11×10^{-4}	.136	-6.55	9.14×10^{-4}	.071
$\boldsymbol{G = 35}$											
$K/n = 0.001$	1.00	.0018	-3.66	5.26×10^{-4}	.049	-7.32	5.07×10^{-4}	.068	-7.19	5.08×10^{-4}	.068
$K/n = 0.100$	1.00	.0020	-1.98	5.87×10^{-4}	.046	-16.1	4.98×10^{-4}	.080	-6.30	4.98×10^{-4}	.066
$K/n = 0.200$	1.00	.0022	-1.21	6.22×10^{-4}	.044	-26.0	4.65×10^{-4}	.099	-6.70	7.20×10^{-4}	.068
$K/n = 0.250$	1.00	.0023	-.057	6.83×10^{-4}	.045	-29.7	4.83×10^{-4}	.109	-5.45	8.41×10^{-4}	.073
$K/n = 0.333$	1.00	.0027	-2.90	7.21×10^{-4}	.050	-40.8	4.43×10^{-4}	.145	-10.7	9.72×10^{-4}	.093

Notes: Simulation results based on 5,000 replications. DGP as described in Equation (4.14).

Table 4.6: Absolute row sum of $\boldsymbol{\kappa}_n$ - Linear regression model with many continuous controls

	$\boldsymbol{G = 140}$			
	$n = 250$	$n = 500$	$n = 750$	$n = 1000$
$K_n/n = 0.200$	4.93	4.23	3.95	3.80
	(0.26)	(0.14)	(0.11)	(0.090)
$K_n/n = 0.300$	9.36	7.66	7.12	6.77
	(0.65)	(0.32)	(0.21)	(0.17)
$K_n/n = 0.400$	18.4	14.5	13.3	12.6
	(1.42)	(0.64)	(0.50)	(0.36)
	$\boldsymbol{G = 70}$			
	$n = 250$	$n = 500$	$n = 750$	$n = 1000$
$K_n/n = 0.200$	8.41	6.71	6.11	5.81
	(0.56)	(0.24)	(0.18)	(0.14)
$K_n/n = 0.300$	17.5	13.2	11.9	11.2
	(1.10)	(0.67)	(0.37)	(0.29)
$K_n/n = 0.400$	37.3	26.6	23.8	22.0
	(3.26)	(1.15)	(0.84)	(0.59)
	$\boldsymbol{G = 35}$			
	$n = 240$	$n = 500$	$n = 740$	$n = 1000$
$K_n/n = 0.200$	18.1	12.4	11.1	10.1
	(1.35)	(0.50)	(0.38)	(0.25)
$K_n/n = 0.300$	41.2	26.1	22.6	20.7
	(2.93)	(1.13)	(0.77)	(0.55)
$K_n/n = 0.400$	100	59.9	47.6	43.0
	(8.30)	(2.64)	(1.77)	(1.20)

Notes: 250 repetitions. Standard deviations in parenthesis. DGP as described in Equation (4.11).

Table 4.7: Absolute row sum of $\boldsymbol{\kappa}_n$ - Linear regression model with many discrete controls

	$\boldsymbol{G = 140}$			
	$n = 250$	$n = 500$	$n = 750$	$n = 1000$
$K_n/n = 0.200$	6.12	4.94	4.53	4.27
	(0.50)	(0.28)	(0.18)	(0.14)
$K_n/n = 0.300$	11.2	8.88	8.04	7.57
	(0.96)	(0.51)	(0.34)	(0.28)
$K_n/n = 0.400$	21.4	16.5	14.9	13.9
	(1.87)	(0.85)	(0.68)	(0.53)
	$\boldsymbol{G = 70}$			
	$n = 250$	$n = 500$	$n = 750$	$n = 1000$
$K_n/n = 0.200$	10.5	8.00	7.10	6.62
	(0.85)	(0.48)	(0.31)	(0.25)
$K_n/n = 0.300$	20.8	15.3	13.4	12.6
	(1.62)	(0.83)	(0.57)	(0.47)
$K_n/n = 0.400$	43.5	30.3	26.3	24.3
	(4.49)	(1.66)	(1.11)	(0.88)
	$\boldsymbol{G = 35}$			
	$n = 240$	$n = 500$	$n = 740$	$n = 1000$
$K_n/n = 0.200$	22.9	15.0	12.8	11.7
	(2.33)	(0.94)	(0.56)	(0.51)
$K_n/n = 0.300$	49.9	30.5	25.8	23.4
	(4.70)	(1.82)	(1.19)	(0.86)
$K_n/n = 0.400$	119	64.6	53.1	47.5
	(11.7)	(4.22)	(2.47)	(1.90)

Notes: 250 repetitions. Standard deviations in parenthesis. DGP as described in Equation (4.11), with $w_{\ell,gi} = \mathbb{1}\{\mathcal{N}(0,1,) \geq 1\}, \quad \forall \ell, g, i.$

Table 4.8: Absolute row sum of $\boldsymbol{\kappa}_n$ - Semiparametric partially linear model

	$\boldsymbol{G = 140}$			
	$n = 250$	$n = 500$	$n = 750$	$n = 1000$
$K_n/n = 0.200$	19.6	18.1	43.3	43.7
	(6.01)	(4.86)	(18.6)	(17.2)
$K_n/n = 0.300$	64.9	142.5	142.4	56.1
	(41.1)	(70.7)	(63.9)	(25.0)
$K_n/n = 0.400$	140	509	140	56.1
	(56.4)	(349)	(65.2)	(25.0)
	$\boldsymbol{G = 70}$			
	$n = 250$	$n = 500$	$n = 750$	$n = 1000$
$K_n/n = 0.200$	33.0	27.5	63.2	63.1
	(11.3)	(8.99)	(28.8)	(25.1)
$K_n/n = 0.300$	109	211	191	80.1
	(46.9)	(104)	(84.7)	(34.8)
$K_n/n = 0.400$	265	700	208	80
	(143)	(532)	(97.4)	(30.4)
	$\boldsymbol{G = 35}$			
	$n = 240$	$n = 500$	$n = 740$	$n = 1000$
$K_n/n = 0.200$	75.1	51.8	101	98
	(26.8)	(15.0)	(41.6)	(33.0)
$K_n/n = 0.300$	260	390	356	134
	(108)	(177)	(161)	(62.5)
$K_n/n = 0.400$	703	1321	356	130
	(342)	(712)	(164)	(48.9)

Notes: 250 repetitions. Standard deviations in parenthesis. DGP as described in Equation (4.12).

Table 4.9: Absolute row sum of $\boldsymbol{\kappa}_n$ - Two-way fixed effects panel data regression model

	$\boldsymbol{G = 140}$			
	$n = 250$	$n = 500$	$n = 750$	$n = 1000$
$K_n/n = 0.200$	3.97	3.90	3.87	3.85
	(.063)	(.039)	(.031)	(0.30)
$K_n/n = 0.333$	11.6	11.1	10.9	10.8
	(.42)	(.30)	(.23)	(.21)
	$\boldsymbol{G = 70}$			
	$n = 250$	$n = 500$	$n = 750$	$n = 1000$
$K_n/n = 0.200$	3.92	3.87	3.83	3.81
	(.040)	(.026)	(.017)	(.015)
$K_n/n = 0.333$	10.6	10.2	10.0	9.90
	(.18)	(.11)	(.098)	(.074)
	$\boldsymbol{G = 35}$			
	$n = 240$	$n = 500$	$n = 740$	$n = 1000$
$K_n/n = 0.200$	8.25	3.89	7.16	3.84
	(1.94)	(.023)	(1.03)	(.011)
$K_n/n = 0.333$	10.7	10.2	9.91	9.77
	(.17)	(.082)	(.062)	(.047)

Notes: 250 repetitions. Standard deviations in parenthesis. DGP as described in Equation (4.14).

Table 4.10: Empirical illustration - Effect of Abortion on Crime

	$\hat{\beta}$	LZ		Robust	
		Std. Error	p-value	Std. Error	p-value
Violent crime					
Baseline	-0.135	0.0422	0.0013	0.0448	0.0025
Many controls	-0.266	0.0842	0.0016	0.1473	0.0718
Property crime					
Baseline	-0.093	0.0146	< 0.00001	0.0149	< 0.00001
Many controls	-0.135	0.0254	< 0.00001	0.0408	0.00091
Murder					
Baseline	-0.134	0.0536	0.0126	0.0551	0.0154
Many controls	-0.197	0.1498	0.1848	0.2117	0.3513

The rows labeled "Baseline estimates" give estimates for the original model in Donohue and Levitt (2001) as in (4.15). The rows labeled "Many controls" give estimates for the high-dimensional model in (4.16) that includes a broader set of controls. Columns under the label "LZ" report clustered standard errors and relative p-values computed with the traditional variance formula by Liang and Zeger (1986). Columns under the label "Robust" report standard errors and relative p-values computed with the cluster-robust variance estimator proposed in this paper. Clustering is at the state level.
Data source: Belloni et al. (2013).

Appendix A

Appendix – Chapter 1

A.1 Extension to nested min/max operators

In this section we describe how the proposed estimation procedure and the theoretical results of Section 2.3 can be extended to scores $\Gamma(g; X)$ that feature nested linear combinations of $\min / \max$ operators. This extension comprises $\min / \max$ operators over multiple components, since $\max\{a, b, c\} = \max\{\max\{a, b\}, c\}$.

We begin by noticing that our proposed estimation described in Section 2.3 can be also defined as follows. First, for each $\min\{a(g; x), b(g; x)\}$ (or $\max$) operator contained in $\Gamma(g; x)$, one substitutes the operator with $a(g; x)$ or $b(g; x)$ based on their cross-fitted plug-in (non-orthogonalized) estimates $\widehat{a}_i := a(\widehat{g}^{-k(i)}; X_i)$ and $\widehat{b}_i := b(\widehat{g}^{-k(i)}; X_i)$. Then, the selected component is estimated ($a(g; x)$ or $b(g; x)$) is estimated by their cross-fitted Neyman-orthogonal analogue ($\widehat{a}_i^{\text{NO}}$ or $\widehat{b}_i^{\text{NO}}$). In the presence of nested $\min / \max$ operators, our estimation is generalized as follows. First, in succession from the most inner to the most outer $\min / \max$, each operator is substituted with their smallest/largest argument based on cross-fitted non-orthogonalized estimates. Then, the selected components are estimated by their cross fitted Neyman-orthogonal analogue. As an illustration, consider the hypothet-

ical score

$$\begin{aligned}\Gamma(g;x) &= d(g;x) + \max\{c(g;x) + \min\{a(g;x), b(g;x)\}, 0\} \\ &= d(g;x) + \max\{\underbrace{c(g;x) + b(g;x) + (a(g;x) - b(g;x)) \cdot \mathbb{1}\{a(g;x) - b(g;x) \leq 0\}}_{\varrho(g;x)}, 0\} \\ &= d(g;x) + \varrho(g;x) \cdot \mathbb{1}\left\{\varrho(g;x) \geq 0\right\}.\end{aligned}$$

Applying the above procedure to this example gives the following expression for the estimated Neyman-orthogonal score:

$$\Gamma^{\text{NO}}(\{\widehat{g}^{-k(i)}, \widehat{f}^{-k(i)}\}, W_i) = \widehat{d}_i^{\text{NO}} + \left[\widehat{c}_i^{\text{NO}} + \widehat{b}_i^{\text{NO}} + (\widehat{a}_i^{\text{NO}} - \widehat{b}_i^{\text{NO}}) \cdot \mathbb{1}\{\widehat{a}_i - \widehat{b}_i \leq 0\}\right] \cdot \mathbb{1}\left\{\widehat{\varrho}_i \geq 0\right\},$$

where

$$\widehat{\varrho}_i = \widehat{c}_i + \widehat{b}_i + (\widehat{a}_i - \widehat{b}_i) \cdot \mathbb{1}\{\widehat{a}_i - \widehat{b}_i \leq 0\}.$$

We will now show how the theoretical results of Section 2.4 can be generalized to this example.[1] Following the arguments of Section 2.4.2, we have

$$\begin{aligned}&\widehat{Q}_n^{\text{NO}}(\pi) - Q_n^{\text{NO}}(\pi) \\ &= \sum_{i=1}^n (2\pi(X_i) - 1) \cdot (\widehat{d}_i^{\text{NO}} - d_i^{\text{NO}}) \\ &+ \sum_{i=1}^n (2\pi(X_i) - 1) \cdot (\widehat{c}_i^{\text{NO}} + \widehat{b}_i^{\text{NO}} - c_i^{\text{NO}} - b_i^{\text{NO}}) \cdot \mathbb{1}\{\widehat{\varrho}_i \geq 0\} \\ &+ \sum_{i=1}^n (2\pi(X_i) - 1) \cdot (\widehat{a}_i^{\text{NO}} - \widehat{b}_i^{\text{NO}}) \cdot \mathbb{1}\{\widehat{a}_i - \widehat{b}_i \leq 0\} \cdot \mathbb{1}\{\widehat{\varrho}_i \geq 0\} \\ &+ \sum_{i=1}^n (2\pi(x) - 1) \cdot (a_i^{\text{NO}} - b_i^{\text{NO}}) \cdot \left[\mathbb{1}\{\widehat{a}_i - \widehat{b}_i \leq 0\} - \mathbb{1}\{a_i - b_i \leq 0\}\right] \cdot \mathbb{1}\{\widehat{\varrho}_i \geq 0\} \\ &+ \sum_{i=1}^n (2\pi(x) - 1) \cdot \varrho_i^{\text{NO}} \cdot \left[\mathbb{1}\{\widehat{\varrho}_i \geq 0\} - \mathbb{1}\{\varrho_i \geq 0\}\right].\end{aligned}$$

The first term in the expansion has the same structure as $A_{0,\ell}$ and thus obeys the

[1] The extension to general scores containing an arbitrary finite number of nested $\min$ / $\max$ operators follows immediately from our discussion of this example.

same bound. The second and third term obey the same bound as $A_{1,\ell}$ since, by virtue of sample-splitting, the indicators $\mathbb{1}\{\widehat{\varrho}_i \geq 0\}$ and $\mathbb{1}\{\widehat{a}_i - \widehat{b}_i \geq 0\}$ are immaterial when controlling the expectation of these term uniformly over Π_n (see arguments in the Proof of Lemma 2.4.1). The fourth term has the same structure as $A_{2,\ell} + A_{3,\ell}$ except for the presence of the indicator $\mathbb{1}\{\widehat{\varrho}_i \geq 0\}$, which again can be shown to be immaterial for controlling the $A_{2,\ell}$-like term by virtue of sample-splitting. For the $A_{3,\ell}$-like term we instead have the bound

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\pi\in\Pi_n}\frac{1}{n}\sum_{i=1}^{n}(2\pi(x)-1)\cdot(c_i-b_i)\cdot\Big(\mathbb{1}\{\widehat{c}_i-\widehat{b}_i\leq 0\}-\mathbb{1}\{c_i-b_i\leq 0\}\Big)\cdot\mathbb{1}\{\widehat{\varrho}_i\geq 0\}\right]\\
&\quad\leq\mathbb{E}\left[\left|(c_i-b_i)\cdot\Big(\mathbb{1}\{\widehat{c}_i-\widehat{b}_i\leq 0\}-\mathbb{1}\{c_i-b_i\leq 0\}\Big)\right|\cdot\left|\mathbb{1}\{\widehat{\varrho}_i\geq 0\}\right|\right]\\
&\quad\leq\mathbb{E}\left[\left|(c_i-b_i)\cdot\Big(\mathbb{1}\{\widehat{c}_i-\widehat{b}_i\leq 0\}-\mathbb{1}\{c_i-b_i\leq 0\}\Big)\right|\right],
\end{aligned}$$

which can be bounded in the same fashion as $A_{3,\ell}$ under a margin assumption on $a(g;X) - b(g;X)$. Finally, the fifth term also has the same structure as $A_{2,\ell} + A_{2,\ell}$, and can be controlled using arguments from Section 2.4.2 under a margin assumption on $\rho(g;X)$.

A.2 Proofs

A.2.1 Proof of Proposition 2.2.1

For the maximin welfare policy we have

$$\begin{aligned}
\min_{(y_0,y_1)\in\mathcal{Y}}\mathbb{E}_{P_X}\left[\pi(X)\cdot y_{\pi(X)}(X)\right] &= \mathbb{E}_{P_X}\left[\min_{(y_0(x),y_1(x))\in\mathcal{Y}(x)}\pi(X)\cdot y_{\pi(X)}(X)\right]\\
&= \mathbb{E}_{P_X}\left[\pi(X)\cdot\underline{y}_1(X)+(1-\pi(X))\cdot\underline{y}_0(X)\right],
\end{aligned}$$

where the first equality is justified by 2.2.1. Thus we have

$$\operatorname*{argmax}_{\pi\in\Pi}\min_{(y_0,y_1)\in\mathcal{Y}}W_\tau(\pi) = \operatorname*{argmax}_{\pi\in\Pi}\mathbb{E}_{P_X}\left[\pi(X)\cdot\underline{y}_1(X)+(1-\pi(X))\cdot\underline{y}_0(X)\right].$$

For the maximin impact policy we have

$$\min_{\tau\in\mathcal{T}} \mathbb{E}_{P_X}\left[\pi(X)\cdot\tau(X)\right] = \mathbb{E}_{P_X}\left[\min_{\tau\in\mathcal{T}} \pi(X)\cdot\tau(X)\right] = \mathbb{E}_{P_X}\left[\pi(X)\cdot\underline{\tau}(X)\right].$$

where the first equality is justified by Assumption 2.2.2. Thus we have

$$\operatorname*{argmax}_{\pi\in\Pi} \min_{\tau\in\mathcal{T}} W_\tau(\pi) = \operatorname*{argmax}_{\pi\in\Pi} \mathbb{E}_{P_X}\left[\pi(X)\cdot\underline{\tau}(X)\right].$$

The statement of the proposition again follows from the invariance of the maximizer to positive affine transformations of the objective function.

A.2.2 Proof of Proposition 2.2.2

We notice that

$$\begin{aligned}
&\max_{\tau\in\mathcal{T}}\left(\max_{\tilde{\pi}\,:\,\mathcal{X}\to\{0,1\}} I_\tau(\tilde{\pi}) - I_\tau(\pi)\right)\\
&= \max_{\tau\in\mathcal{T}} \mathbb{E}_{P_X}\left[\left(\frac{1}{2}+\frac{1}{2}\mathrm{sgn}(\tau(X)) - \pi(X)\right)\cdot\tau(X)\right]\\
&= \mathbb{E}_{P_X}\left[\max_{\tau\in\mathcal{T}}\left(\frac{1}{2}+\frac{1}{2}\mathrm{sgn}(\tau(X)) - \pi(X)\right)\cdot\tau(X)\right]\\
&= \mathbb{E}_{P_X}\Bigg[\underbrace{\max_{\tau\in\mathcal{T}} \mathbb{1}\big\{\underline{\tau}(X)\geq 0\big\}\cdot(1-\pi(X))\cdot\tau(X)}_{=(1-\pi(X))\cdot\mathbb{1}\{\underline{\tau}(X)\geq 0\}\cdot\overline{\tau}(X)}\Bigg]\\
&\qquad + \mathbb{E}_{P_X}\Bigg[\underbrace{\max_{\tau\in\mathcal{T}} \mathbb{1}\big\{\overline{\tau}(X)\leq 0\big\}\cdot -\pi(X)\cdot\tau(X)}_{=-\pi(X)\cdot\mathbb{1}\{\underline{\tau}(X)\geq 0\}\cdot\underline{\tau}(X)}\Bigg]\\
&\qquad + \mathbb{E}_{P_X}\Bigg[\underbrace{\max_{\tau\in\mathcal{T}} \mathbb{1}\big\{\underline{\tau}(X)<0<\overline{\tau}(X)\big\}\cdot\left(\frac{1}{2}+\frac{1}{2}\mathrm{sgn}(\tau(X)) - \pi(X)\right)\cdot\tau(X)}_{=-\pi(X)\cdot\mathbb{1}\{\underline{\tau}(X)<0<\overline{\tau}(X)\}(\overline{\tau}(X)+\underline{\tau}(X))+\mathbb{1}\{\underline{\tau}(X)<0<\overline{\tau}(X)\}\cdot\overline{\tau}(X)}\Bigg]\\
&= -\mathbb{E}_{P_X}\Bigg[\pi(X)\cdot\Bigg(\mathbb{1}\big\{\underline{\tau}(X)\geq 0\big\}\cdot\overline{\tau}(X) + \mathbb{1}\big\{\overline{\tau}(X)\leq 0\big\}\cdot\underline{\tau}(X)\\
&\qquad\qquad + \mathbb{1}\big\{\underline{\tau}(X)<0<\overline{\tau}(X)\big\}\cdot\big(\overline{\tau}(X)+\underline{\tau}(X)\big)\Bigg)\Bigg]\\
&\qquad + \mathbb{E}_{P_X}\left[\overline{\tau}(X)\cdot\mathbb{1}\big\{\underline{\tau}(X)\geq 0\big\} + \mathbb{1}\left\{\underline{\tau}(X)<0<\overline{\tau}(X)\right\}\cdot\overline{\tau}(X)\right],
\end{aligned}$$

where the first equality uses the fact $\operatorname{argmax}_{\pi\,:\,\mathcal{X}\to\{0,1\}} W_\tau(\pi) = \frac{1}{2}+\frac{1}{2}\mathrm{sign}(\tau(X))$, and the second equality uses Assumption 2.2.2. The statement of the proposition

then follows from the invariance of the maximizer to positive affine transformations of the objective function.

A.2.3 Proof of Proposition 2.4.1

We being by decomposing regret as follows:

$$Q(\pi_n^*) - Q(\widehat{\pi}_n) = \Big[Q(\pi_n^*) - Q_n^{\mathrm{NO}}(\pi_n^*)\Big] + \Big[Q_n^{\mathrm{NO}}(\pi_n^*) - \widehat{Q}_n^{\mathrm{NO}}(\widehat{\pi}_n)\Big] + \Big[\widehat{Q}_n^{\mathrm{NO}}(\widehat{\pi}_n) - Q(\widehat{\pi}_n)\Big]. \tag{A.1}$$

The first term is zero in expectation. The second term can be upper bounded as

$$\begin{aligned}\Big[Q_n^{\mathrm{NO}}(\pi_n^*) - \widehat{Q}_n^{\mathrm{NO}}(\widehat{\pi}_n)\Big] &\leq \Big[Q_n^{\mathrm{NO}}(\pi_n^*) - \widehat{Q}_n^{\mathrm{NO}}(\pi_n^*)\Big] + \Big[\widehat{Q}_n^{\mathrm{NO}}(\pi_n^*) - \widehat{Q}_n^{\mathrm{NO}}(\widehat{\pi}_n)\Big] \\ &\leq \sup_{\pi\in\Pi_n}\Big|Q_n^{\mathrm{NO}}(\pi) - \widehat{Q}_n^{\mathrm{NO}}(\pi)\Big|,\end{aligned}$$

where we have used that $\widehat{Q}_n^{\mathrm{NO}}(\pi^*) - \widehat{Q}_n^{\mathrm{NO}}(\widehat{\pi}_n) \leq 0$, which follows from $\widehat{\pi}_n$ being the maximizer of $\widehat{Q}_n^{\mathrm{NO}}(\cdot)$. The third term can be further expanded and upper bounded as follows

$$\widehat{Q}_n^{\mathrm{NO}}(\widehat{\pi}_n) - Q(\widehat{\pi}_n) \leq \sup_{\pi\in\Pi_n}\Big|\widehat{Q}_n^{\mathrm{NO}}(\pi) - Q_n^{\mathrm{NO}}(\pi)\Big| + \sup_{\pi\in\Pi_n}\left|Q_n^{\mathrm{NO}}(\pi) - Q(\pi)\right|.$$

Using the last two displays and taking expectations in (A.1) yields the desired conclusion.

A.2.4 Proof of Lemma 2.4.1

We will establish each of the following bounds in turn:

$$\begin{aligned}\mathbb{E}\left[\sup_{\pi\in\Pi_n}|A_0(\pi)|\right] &= O\left(\sqrt{\mathrm{VC}(\Pi_n)\cdot\frac{r_{\kappa_n}}{n^{3/2}}} + \frac{r_{\kappa_n}}{n^{1/2}}\right), \\ \mathbb{E}\left[\sup_{\pi\in\Pi_n}|A_{1,\ell}(\pi)|\right] &= O\left(\sqrt{\mathrm{VC}(\Pi_n)\cdot\frac{r_{\kappa_n}}{n^{3/2}}} + \frac{r_{\kappa_n}}{n^{1/2}}\right), \\ \mathbb{E}\left[\sup_{\pi\in\Pi_n}|A_{2,\ell}(\pi)|\right] &= O\left(\sqrt{\frac{\mathrm{VC}(\Pi_n)}{n}}\right), \\ \mathbb{E}\left[\sup_{\pi\in\Pi_n}|A_{3,\ell}(\pi)|\right] &= O\left(\left(\frac{r_{\kappa_n}}{n}\right)^{\frac{\gamma+1}{\gamma+2}}\right).\end{aligned}$$

Combining the above through decomposition (2.21) gives the desired final bound.

Bound for A_0 and $A_{1,\ell}$

We prove a bound for $\sup_{\pi\in\Pi_n}|A_{1,\ell}(\pi)|$; it will be immediate that $A_0(\pi)$ obeys the same bound. We begin with the following decomposition

$$\begin{aligned}
A_{1,\ell}(\pi) = \frac{1}{n}\sum_{i=1}^{n}(2\pi(X_i)-1)\cdot\langle\widehat{\alpha}_i-\alpha_i, U_i-g(V_i)\rangle, &\qquad (=B_1(\pi))\,,\\
+\frac{1}{n}\sum_{i=1}^{n}(2\pi(X_i)-1)\cdot\langle\widehat{\alpha}_i-\alpha_i, g_i-\widehat{g}_i\rangle, &\qquad (=B_2(\pi)),\\
+\frac{1}{n}\sum_{i=1}^{n}(2\pi(X_i)-1)\cdot\left(\varphi_\ell(\widehat{g}^{-k(i)};X_i)-\varphi_\ell(g;X_i)+\langle\alpha_i, g_i-\widehat{g}_i\rangle\right), &\qquad (=B_3(\pi)),
\end{aligned}$$

where we have used the shorthand notation $\widehat{g}_i := \widehat{g}^{-k(i)}(V_i)$, $g_i = g(V_i)$, $\widehat{\alpha}_i := (\{\widehat{g}^{-k(i)}, f^{-k(i)}\}; V_i)$, $\alpha_i := (\{g,f\};V_i)$. Starting with $B_1(\pi)$, the contribution of the k-th fold is

$$B_1^{(k)}(\pi) = \frac{1}{n}\sum_{i:k(i)=k}(2\pi(X_i)-1)\cdot\langle\widehat{\alpha}_i-\alpha_i,\varepsilon_i\rangle\cdot\mathbb{1}\{\widehat{\varphi}_i\geq 0\}.$$

The sample-splitting procedure guarantees that $\{\widehat{g}^{-k(i)}, \widehat{f}^{-k(i)}\}$ only depend on data from the remaining $K-1$ folds, and thus conditioning on these estimates for the nuisance components makes $B_1^{(k)}(\pi)$ a sum of independent mean-zero terms, in light of

$$\mathbb{E}\left[U_i - g(V_i)\mid V_i, \widehat{g}^{-k(i)}, \widehat{f}^{-k(i)}\right] = 0.$$

Furthermore, the terms are also sub-Gaussian since it is a linear combination of sub-Gaussian random variables with bounded weights w.p.a 1, in light of

$$\begin{aligned}
\|\mathbb{1}\{\widehat{\varphi}_i\geq 0\}\cdot(\widehat{\alpha}_i-\alpha_i)\|_{L_\infty(P_V)} &\leq \left\|\widehat{g}^{-k}(V)-g(V)\right\|_{L_\infty(P_V)} + \left\|\widehat{f}^{-k}(V)-f(V)\right\|_{L_\infty(P_V)}\\
&\leq 2\cdot\mathcal{C}_2\cdot\mathcal{C}_3,
\end{aligned}$$

w.p.a 1, where the first inequality uses Assumption 2.4.2(i) and the second inequality uses Assumption 2.4.2(iii). Having computed the variance of $B_1^{(k)}(\pi)$ conditional on $\left(\widehat{g}^{-k}, \widehat{f}^{-k}\right)$

$$V_n(k) = \mathbb{E}\left[(\widehat{\alpha}_i^{-k} - \alpha_i)' \Sigma(V_i) (\widehat{\alpha}_i^{-k} - \alpha_i) \cdot \mathbb{1}\{\widehat{\varphi}_i \geq 0\} \mid \widehat{g}^{-k}, \widehat{f}^{-k}\right],$$

we can apply Corollary 3 in Athey and Wager (2021) to establish the bound

$$\frac{n}{n_k}\mathbb{E}\left[\sup_{\pi\in\Pi}\left|B_1^{(k)}(\pi)\right| \mid \hat{g}^{-k}\right] = O\left(\sqrt{V_n(k)\frac{\text{VC}(\Pi_n)}{n_k}}\right), \tag{A.2}$$

where n_k denotes the number of observations in the k-th fold. Using Assumptions 2.4.2(ii) and 2.4.4, we have

$$\begin{aligned}
\mathbb{E}\left[V_n(k)\right] &\leq \mathbb{E}_{P_n}\left[\overline{\lambda} \cdot \|\widehat{\alpha}_i - \alpha_i\|_{L_2(P_V)}^2\right] \\
&\leq 2 \cdot \overline{\lambda} \cdot \mathcal{C}_{2,\alpha}^2 \cdot \mathbb{E}_{P_n}\left[\left\|\widehat{g}^{-k(i)} - g\right\|_{L_2(P_V)}^2 + \left\|\widehat{f}^{-k(i)} - f\right\|_{L_2(P_V)}^2\right] \\
&= O\left(\frac{r_{\kappa_n}}{\sqrt{n}}\right).
\end{aligned} \tag{A.3}$$

Finally, we apply (A.2) repeatedly for each of the K data-folds and using Jensen's Inequality and (A.3) and obtain the final bound

$$\mathbb{E}\left[\sup_{\pi\in\Pi}|B_1(\pi)|\right] = O\left(\sqrt{\text{VC}(\Pi_n) \cdot \frac{r_{\kappa_n}}{n^{3/2}}}\right). \tag{A.4}$$

We now turn to $B_2(\pi)$, for which we have

$$\begin{aligned}
B_2(\pi) &= \frac{1}{n}\sum_{i=1}^{n}(2\pi(X_i)-1)\cdot\langle\widehat{\alpha}_i-\alpha_i,\widehat{g}_i-g_i\rangle\cdot\mathbb{1}\{\widehat{\varphi}_i\geq 0\} \\
&= \sum_{j=1}^{J}\left[\frac{1}{n}\sum_{i=1}^{n}(2\pi(X_i)-1)\cdot(\widehat{\alpha}_i^{(j)}-\alpha_i^{(j)})\cdot(\widehat{g}_i^{(j)}-g_i)\cdot\mathbb{1}\{\widehat{\varphi}_i\geq 0\}\right] \\
&\leq \sum_{j=1}^{J}\left[\frac{1}{n}\sum_{i=1}^{n}|\widehat{\alpha}_i^{(j)}-\alpha_i^{(j)}|\cdot|\widehat{g}_i^{(j)}-g_i|\right] \\
&\leq \sum_{j=1}^{J}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\alpha}_i^{(j)}-\alpha_i^{(j)}\right)^2}\times\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(g_i^{(j)}-g_i\right)^2},
\end{aligned}$$

where the last inequality uses Cauchy-Schwarz inequality. This bound does not depend on π and thus holds uniformly over Π_n. We then apply Cauchy-Schwarz again and use Asuumption 2.4.4 to verify that

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\pi\in\Pi}|B_2(\pi)|\right] \\
&\leq \sum_{j=1}^{J}\mathbb{E}_{P_n}\left[\left\|\widehat{\alpha}_i^{(j)}-\alpha_i^{(j)}\right\|_{L_2(P_V)}^2\right]^{1/2}\times\mathbb{E}_{P_n}\left[\left\|\widehat{g}_i^{(j)}-g_i^{(j)}\right\|_{L_2(P_V)}^2\right]^{1/2}, \\
&\leq J\cdot\mathbb{E}_{P_n}\left[\|\widehat{\alpha}_i-\alpha_i\|_{L_2(P_V)}^2\right]^{1/2}\times\mathbb{E}_{P_n}\left[\|\widehat{g}_i-g_i\|_{L_2(P_V)}^2\right]^{1/2} \\
&\lesssim J\cdot\mathbb{E}_{P_n}\left[\left\|\widehat{f}^{-k(i)}-f\right\|_{L_2(P_V)}^2+\left\|\widehat{g}^{-k(i)}-g\right\|_{L_2(P_V)}^2\right]^{1/2}\times\mathbb{E}_{P_n}\left[\left\|\widehat{g}^{-k(i)}-g\right\|_{L_2(P_V)}^2\right]^{1/2} \\
&= O\left(\frac{r_{\kappa_n}}{\sqrt{n}}\right).
\end{aligned}$$

We now turn to $B_3(\pi)$. We begin by considering the following telescoping

$$\widehat{g}_i-g_i=\sum_{j=1}^{J}\left[(g_i^{(\bullet:j-1)},\widehat{g}_i^{(j:\bullet)})-(g_i^{(\bullet:j)},\widehat{g}_i^{(j+1:\bullet)})\right]=\sum_{j=1}^{J}\left(0,0,\ldots,\widehat{g}_i^{(j)}-g_i^{(j)},0,\ldots,0\right),$$

where $g^{(\bullet:j)}$ and $g^{(j:\bullet)}$ denote, respectively, the first and last j entries of $g(V_i)$, where we adopt the convention $g^{(\bullet:0)}=g^{(J+1:\bullet)}=\emptyset$. We can therefore decompose $B_3(\pi)$

as follows:

$$
\begin{aligned}
B_3(\pi) &= \sum_{j=1}^{J} \frac{1}{n} \sum_{i=1}^{n} (2\pi(X_i) - 1) \\
&\times \left[\varphi((g^{(\bullet:j-1)}, \widehat{g}_i^{(j:\bullet)}); X_i) - \varphi((g_i^{(\bullet:j)}, \widehat{g}_i^{(j+1:\bullet)}); X_i) - \alpha^{(j)}(\{(g^{(\bullet:j-1)}, \widehat{g}^{(j:\bullet)}), f\}) \cdot \left(\widehat{g}_i^{(j)} - g_i^{(j)} \right) \right] \\
&\times \mathbb{1}\{\widehat{\varphi}_i \geq 0\} \\
&- \sum_{j=1}^{J} \frac{1}{n} \sum_{i=1}^{n} (2\pi(X_i) - 1) \cdot \left[\left(\alpha_i^{(j)} - \alpha^{(j)}(\{(g^{(\bullet:j-1}, \widehat{g}^{(j:\bullet)}), f\}) \right) \cdot \left(\widehat{g}_i^{(j)} - g_i^{(j)} \right) \right] \cdot \mathbb{1}\{\widehat{\varphi}_i \geq 0\}.
\end{aligned}
$$

By the definition of the Riesz-representer and cross-fitting we have

$$
\begin{aligned}
\mathbb{E}\Big[&\varphi((g^{(\bullet:j-1)}, \widehat{g}_i^{(j:\bullet)}); X_i) - \varphi((g_i^{(\bullet:j)}, \widehat{g}_i^{(j+1:\bullet)}); X_i) \\
&- \alpha^{(j)}(\{(g^{(\bullet:j-1)}, \widehat{g}^{(j:\bullet)}), f\}, W_i) \cdot \left(\widehat{g}_i^{(j)} - g_i^{(j)} \right) \mid V_i, \widehat{g}^{-k(i)}, \widehat{f}^{-k(i)} \Big] = 0,
\end{aligned}
$$

where we have used the property $\alpha_\ell^{(j)}(\{(\widetilde{g}_{-j}, \widetilde{g}_j), \widetilde{f}\}, x) = \alpha_\ell^{(j)}(\{\widetilde{g}_{-j}, \widetilde{f}\}, x)$. Furthermore, the term within the expectation operator is sub-Gaussian since uniformly bounded by Assumption 2.4.2(iii). Therefore the first term in the expansion of $B_3(\pi)$ can be controlled uniformly using identical arguments as for $B_1(\pi)$ and obeys the same bound. The second term in the expansion of $B_3(\pi)$ can be bounded with identical arguments as for $B_2(\pi)$ and obeys the same bound. We therefore conclude that

$$
\mathbb{E}\left[\sup_{\pi \in \Pi} |B_3(\pi)| \right] = O\left(\sqrt{\mathrm{VC}(\Pi_n) \cdot \frac{r_{\kappa_n}}{n^{3/2}}} + \frac{r_{\kappa_n}}{\sqrt{n}} \right).
$$

Combining the bounds for $B_1(\pi)$,$B_2(\pi)$ and $B_3(\pi)$ via the triangle inequality finally gives the desired bound for $\mathbb{E}\left[\sup_{\pi \in \Pi_n} |A_{1,\ell}(\pi)|\right]$.

Bound for $A_{2,\ell}$

We first notice that

$$
\mathbb{E}\left[\phi_\ell(\{g, f\}; W_i) \cdot \left(\mathbb{1}\left\{ \varphi_\ell(\widehat{g}^{-k(i)}; X_i) \geq 0 \right\} - \mathbb{1}\left\{ \varphi_\ell(g; X_i) \geq 0 \right\} \right) \mid V_i, \widehat{g}^{-k(i)} \right] = 0, \tag{A.5}
$$

by the mean-zero property of the influence function adjustments ϕ_ℓ and cross-fitting. Furthermore, the term inside expectation is sub-Gaussian by uniform boundedness of the Riesz-representer, guaranteed by Assumption 2.4.2(iii). Thus we can use similar arguments to those used for $B_1(\pi)$ to show

$$\mathbb{E}\left[\sup_{\pi\in\Pi_n}|A_{2,\ell}(\pi)|\right] = O\left(\sqrt{\frac{\mathrm{VC}(\Pi_n)}{n}}\right).$$

Bound for $A_{3,\ell}$

We begin by noticing that $\varphi_\ell(g;X_i)\left(\mathbb{1}\left\{\varphi_\ell(\widehat{g}^{-k(i)};X_i)\geq 0\right\} - \mathbb{1}\left\{\varphi_\ell(g;X_i)\geq 0\right\}\right) \leq 0$ and thus, since the "never treat" policy belongs to any policy class Π for which $\mathrm{VC}(\Pi)\geq 1$, we have

$$\sup_{\pi\in\Pi_n} A_{3,\ell}(\pi) = \frac{1}{2n}\sum_{i=1}^{n}\left|\varphi_\ell(g;X_i)\cdot\left(\mathbb{1}\left\{\varphi_\ell(\widehat{g}^{-k(i)};X_i)\geq 0\right\} - \mathbb{1}\left\{\varphi_\ell(g;X_i)\geq 0\right\}\right)\right|,$$

and thus we obtain the uniform bound[2]

$$\mathbb{E}\left[\sup_{\pi\in\Pi_n} A_{3,\ell}(\pi)\right] = \frac{1}{2}\mathbb{E}\left[\left|\varphi_\ell(g;X_i)\cdot\left(\mathbb{1}\left\{\varphi_\ell(\widehat{g}^{-k(i)};X_i)\geq 0\right\} - \mathbb{1}\left\{\varphi_\ell(g;X_i)\geq 0\right\}\right)\right|\right]. \tag{A.6}$$

For the RHS in (A.6), we closely follow Lemma 5.2 in Audibert and Tsybakov (2007), but we report the steps of the proof for completeness. For $\gamma > 0$ and any

[2]For a policy class of zero VC-dimension, (A.6) holds as an inequality.

$t > 0$ we have

$$
\begin{aligned}
&\mathbb{E}\left[\left|\varphi_\ell(g; X_i) \left(\mathbb{1}\left\{\varphi_\ell(\widehat{g}^{-k(i)}; X_i) \geq 0\right\} - \mathbb{1}\left\{\varphi_\ell(g; X_i) \geq 0\right\}\right)\right| \right] \\
&\leq \mathbb{E}\left[|\varphi_\ell(g; X_i)| \cdot \mathbb{1}\left\{\left|\varphi_\ell(\widehat{g}^{-k(i)}; X_i) - \varphi_\ell(g; X_i)\right| \geq |\varphi_\ell(g; X_i)|\right\} \right] \\
&\leq \mathbb{E}\left[|\varphi_\ell(g; X_i)| \cdot \mathbb{1}\left\{\left|\varphi_\ell(\widehat{g}^{-k(i)}; X_i) - \varphi_\ell(g; X_i)\right| \geq |\varphi_\ell(g; X_i)|\right\} \cdot \mathbb{1}\left\{0 < |\varphi_\ell(g; X_i)| \leq t\right\} \right] \\
&\qquad + \mathbb{E}\left[|\varphi_\ell(g; X_i)| \cdot \mathbb{1}\left\{\left|\varphi_\ell(\widehat{g}^{-k(i)}; X_i) - \varphi_\ell(g; X_i)\right| \geq |\varphi_\ell(g; X_i)|\right\} \cdot \mathbb{1}\left\{|\varphi_\ell(g; X_i)| > t\right\} \right] \\
&\leq \mathbb{E}\left[\left|\varphi_\ell(\widehat{g}^{-k(i)}; X_i) - \varphi_\ell(g; X_i)\right| \cdot \mathbb{1}\left\{0 < |\varphi_\ell(g; X_i)| \leq t\right\} \right] \\
&\qquad + \mathbb{E}\left[\left|\varphi_\ell(\widehat{g}^{-k(i)}; X_i) - \varphi_\ell(g; X_i)\right| \cdot \mathbb{1}\left\{\left|\varphi_\ell(\widehat{g}^{-k(i)}; X_i) - \varphi_\ell(g; X_i)\right| > t\right\} \right] \\
&\leq \mathbb{E}\left[\left(\varphi_\ell(\widehat{g}^{-k(i)}; X_i) - \varphi_\ell(g; X_i)\right)^2 \right]^{1/2} \cdot \mathbb{P}\big(0 < |\varphi_\ell(g; X_i)| \leq t\big)^{1/2} \\
&\qquad + \frac{\mathbb{E}\left[\left(\varphi_\ell(\widehat{g}^{-k(i)}; X_i) - \varphi_\ell(g; X_i)\right)^2 \right]}{t} \\
&\leq C_0^{1/2} \mathbb{E}\left[\left(\varphi_\ell(\widehat{g}^{-k(i)}; X_i) - \varphi_\ell(g; X_i)\right)^2 \right]^{1/2} t^{\gamma/2} + \frac{\mathbb{E}\left[\left(\varphi_\ell(\widehat{g}^{-k(i)}; X_i) - \varphi_\ell(g; X_i)\right)^2 \right]}{t},
\end{aligned}
$$

where the penultimate inequality uses Cauchy-Schwarz and Markov inequalities, and the last inequality uses the Margin Assumption. Minimizing the last display over t gives

$$
\begin{aligned}
\mathbb{E}\left[\sup_{\pi \in \Pi_n} A_{3,\ell}(\pi) \right] &\leq (\gamma + 2) \cdot \left(\frac{2}{\gamma}\right)^{\gamma/(\gamma+2)} \cdot \mathcal{C}_m^{1/(\gamma+2)} \cdot \mathbb{E}\left[\left(\varphi_\ell(\widehat{g}^{-k(i)}; X_i) - \varphi_\ell(g; X_i)\right)^2 \right]^{\frac{\gamma+1}{\gamma+2}} \\
&\leq (\gamma + 2) \cdot \left(\frac{2}{\gamma}\right)^{\gamma/(\gamma+2)} \cdot \mathcal{C}_m^{1/(\gamma+2)} \cdot \mathcal{C}_{2,\varphi}^{\frac{2(\gamma+1)}{\gamma+2}} \cdot \mathbb{E}_{P_n}\left[\left\| \widehat{g}^{-k} - g \right\|_{L_2(P_X)}^2 \right]^{\frac{\gamma+1}{\gamma+2}}
\end{aligned}
$$

For $\gamma = 0$, a similar argument gives

$$\begin{aligned}
&\mathbb{E}\left[\left|\varphi_\ell(g;X_i)\left(\mathbb{1}\left\{\varphi_\ell(\widehat{g}^{-k(i)};X_i)\geq 0\right\}-\mathbb{1}\left\{\varphi_\ell(g;X_i)\geq 0\right\}\right)\right|\right]\\
&\leq \mathbb{E}\left[\left|\varphi_\ell(\widehat{g}^{-k(i)};X_i)-\varphi_\ell(g;X_i)\right|\cdot\mathbb{1}\left\{0<|\varphi_\ell(g;X_i)|\leq t\right\}\right]\\
&\qquad+\mathbb{E}\left[\left|\varphi_\ell(\widehat{g}^{-k(i)};X_i)-\varphi_\ell(g;X_i)\right|\cdot\mathbb{1}\left\{\left|\varphi_\ell(\widehat{g}^{-k(i)};X_i)-\varphi_\ell(g;X_i)\right|>t\right\}\right]\\
&\leq 2\mathbb{E}_{P_n}\left[\left\|\varphi_\ell(\widehat{g}^{-k};X)-\varphi_\ell(g;X)\right\|^2_{L_2(P_X)}\right]^{1/2}\\
&\leq 2\cdot\mathcal{C}^2_{2,\varphi}\cdot\mathbb{E}_{P_n}\left[\left\|\widehat{g}^{-k}-g\right\|^2_{L_2(P_X)}\right]^{1/2}.
\end{aligned}$$

Combining the cases $\gamma > 0$ and $\gamma = 0$, and using the L_2-risk bounds for $\widehat{g}^{-k}$ from Assumption 2.4.4 we finally get

$$\mathbb{E}\left[\sup_{\pi\in\Pi_n} A_{3,\ell}(\pi)\right] = O\left(\left(\frac{r_{\kappa_n}}{\sqrt{n}}\right)^{\frac{\gamma+1}{\gamma+2}}\right).$$

A.2.5 Proof of Theorem 2.4.1

The Neyman-orthogonalized score $\Gamma^{\text{NO}}(\{g,f\};W_i)$ satisfies the assumptions of Corollary 3 in Athey and Wager (2021), and thus it can be applied verbatim to show that

$$\mathbb{E}\left[\sup_{\pi\in\Pi_n}|Q^{\text{NO}}_n(\pi)-Q(\pi)|\right] = O\left(\sqrt{\frac{\text{VC}(\Pi_n)}{n}}\right).$$

Combining the above bound with Lemma 2.4.1 via Proposition 2.4.1 gives the statement of the theorem.

Appendix B

Appendix – Chapter 2

B.0.1 Consistency result for generalized model

Here we present a consistency result for the generalized model of Assumption 3.3.1. We make the following assumptions.

Assumption B.0.1 (Exogeneity of $X_i' \beta_0$). *U_i is independent of $(W_i, X_i'\beta_0, Z_i)$.*

Assumption B.0.2 (Regularity conditions).

(i) The parameter sets $\mathcal{B}$ and $\mathcal{E}$ are compact. $\mathcal{B}$ contains β_0 as an interior point. $\mathcal{E}$ contains $(0, \alpha_0)$ as an interior point.

(ii) For all possible outcomes y, the log-likelihood function $\ell\left(y \,\middle|\, w, \omega, \alpha\right)$ is strictly convex in (ω, α) and has Hessian with eigenvalues bounded away from zero, uniformly over (w, ω, α). Furthermore, $\ell\left(Y_i \,\middle|\, W_i, X_i'\beta + Z_i'\gamma, \alpha\right)$ is three times continuously differentiable in (β, γ, α) with derivatives that in expectation are bounded for all $(\beta, \gamma, \alpha) \in \mathcal{B} \times \mathcal{E}$.

(iii) Let $\mathcal{L}(\beta, \gamma, \alpha) := \mathbb{E}\left[\ell\left(Y_i \,\middle|\, W_i, X_i\beta + Z_i\gamma,\, \alpha\right)\right]$ denote the population log-likelihood function. For all $\beta \in \mathcal{B}$ and $\eta := (\gamma, \alpha) \in \mathcal{E}$, we have

$$\operatorname{rank}\left\{\frac{\partial^2 \mathcal{L}(\beta, \gamma, \alpha)}{\partial \eta \partial \eta'}\right\} = k_z + k_\alpha. \tag{B.1}$$

For all $\beta \in \mathcal{B}$ and $(0, \alpha) \in \mathcal{E}$, the matrix

$$A(\beta, 0, \alpha) \coloneqq \begin{pmatrix} \frac{\partial^2 \mathcal{L}(\beta,0,\alpha)}{\partial \gamma \partial \beta'} & \frac{\partial^2 \mathcal{L}(\beta,0,\alpha)}{\partial \gamma \partial \alpha'} \\ \frac{\partial^2 \mathcal{L}(\beta,0,\alpha)}{\partial \alpha \partial \beta'} & \frac{\partial^2 \mathcal{L}(\beta,0,\alpha)}{\partial \alpha \partial \alpha'} \end{pmatrix} \tag{B.2}$$

has full rank $k_x + k_\alpha$.

(iv) The second derivatives of the sample log-likelihood

$$\mathcal{L}_n(\beta, \gamma, \alpha) = \frac{1}{n} \sum_{i=1}^{n} \ell \left(Y_i \mid W_i, X_i'\beta + Z_i'\gamma, \alpha \right)$$

converge in probability to those of the population log-likelihood

$$\mathcal{L}(\beta, \gamma, \alpha) = \mathbb{E} \left[\ell \left(Y_i \mid W_i, X_i'\beta + Z_i'\gamma, \alpha \right) \right]$$

uniformly over $(\beta, \gamma, \alpha) \in \mathcal{B} \times \mathcal{E}$.

(v) The symmetric matrix $\Omega_{n,\beta}$ is a twice continuously differentiable function in β, and there exists a constant $c > 0$ such that with probability approaching one we have $\Omega_{n,\beta} \geq c$ for all $\beta \in \mathcal{B}$. Furthermore, we have $\sup_{\beta \in \mathcal{B}} \|\Omega_{n,\beta} - \Omega_\beta\| = o_p(1)$ for some non-random symmetric matrix Ω_β which is positive-definite for all $\beta \in \mathcal{B}$.

Assumptions B.0.1 and B.0.2 generalize Assumptions 3.2.1 and 3.2.2, respectively, to the case with additional regressors W_i and parameters α. In particular, Assumption B.0.2(iii) imposes generalizations of the non-collinearity and relevance conditions for the instruments. Under (B.1), condition (B.2) is equivalent to requiring

$$\operatorname{rank} \left\{ \frac{\partial^2 \mathcal{L}(\beta, 0, \alpha)}{\partial \gamma \partial \beta'} - \left[\frac{\partial^2 \mathcal{L}(\beta, 0, \alpha)}{\partial \gamma \partial \alpha'} \right] \left[\frac{\partial^2 \mathcal{L}(\beta, 0, \alpha)}{\partial \alpha \partial \alpha'} \right]^{-1} \left[\frac{\partial^2 \mathcal{L}(\beta, 0, \alpha)}{\partial \alpha \partial \beta'} \right] \right\} = k_x. \tag{B.3}$$

Theorem B.0.1. *Let Assumption 3.3.1, B.0.1, B.0.2 hold. Then we have* $(\widehat{\beta}_{\mathrm{AIV}}, \widehat{\alpha}_{\mathrm{AIV}}) = (\beta_0, \alpha_0) + o_P(1)$*, as* $n \to \infty$.

B.0.2 Asymptotic normality result for generalized model

We now present the general result for the asymptotic distribution of the AIV estimator. To do so, we introduce the following notation for the first and second derivatives of the sample and population log-likelihood:

$$\begin{aligned}
\mathcal{L}_{\alpha}(\beta, \gamma, \alpha) &= \mathbb{E}\left[\frac{\partial \ell(Y_i \mid W_i, X_i'\beta + Z_i'\gamma, \alpha)}{\partial \alpha}\right], \\
\mathcal{L}_{n,\alpha}(\beta, \gamma, \alpha) &= \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ell(Y_i \mid W_i, X_i'\beta + Z_i'\gamma, \alpha)}{\partial \alpha}, \\
\mathcal{L}_{\alpha\beta}(\beta, \gamma, \alpha) &= \mathbb{E}\left[\frac{\partial \ell(Y_i \mid W_i, X_i'\beta + Z_i'\gamma, \alpha)}{\partial \alpha\, \partial \beta'}\right], \\
\mathcal{L}_{n,\alpha\beta}(\beta, \gamma, \alpha) &= \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ell(Y_i \mid W_i, X_i'\beta + Z_i'\gamma, \alpha)}{\partial \alpha\, \partial \beta'},
\end{aligned}$$

where we will also use the short-hand $\mathcal{L}_{\alpha\beta} \coloneqq \mathcal{L}_{\alpha\beta}(\beta_0, 0, \alpha_0)$. We also define the matrices

$$\widetilde{H} = \mathcal{L}_{\gamma\gamma} - \mathcal{L}_{\gamma\alpha}\,\mathcal{L}_{\alpha\alpha}^{-1}\,\mathcal{L}_{\alpha\gamma}, \quad \widetilde{G} = \mathcal{L}_{\gamma\beta} - \mathcal{L}_{\gamma\alpha}\,\mathcal{L}_{\alpha\alpha}^{-1}\,\mathcal{L}_{\alpha\beta},$$

and their sample analogues $\widetilde{H}_n, \widetilde{G}_n$ based on $\mathcal{L}_{n,\alpha\alpha}, \mathcal{L}_{n,\alpha\beta}, \ldots$ in the natural way.

Theorem B.0.2. *Let Assumption 3.3.1, B.0.1, B.0.2 hold. Then we have*

$$\begin{aligned}
\sqrt{n}\,(\widehat{\beta}_{\mathrm{AIV}} - \beta_0) &= -(\widetilde{G}'\,\widetilde{W}\,\widetilde{G})^{-1}\,\widetilde{G}'\,\widetilde{W}\,\sqrt{n}\,\left\{\mathcal{L}_{n,\gamma} - \mathcal{L}_{\gamma\alpha}\,\mathcal{L}_{\alpha\alpha}^{-1}\,\mathcal{L}_{n,\alpha}\right\} + o_p(1), \\
\sqrt{n}\,(\widehat{\alpha}_{\mathrm{AIV}} - \alpha_0) &= -\mathcal{L}_{\alpha\alpha}^{-1}\,\mathcal{L}_{\alpha\beta}\,\sqrt{n}\,(\widehat{\beta}_{\mathrm{AIV}} - \beta_0) - \mathcal{L}_{\alpha\alpha}^{-1}\,\sqrt{n}\mathcal{L}_{n,\alpha} + o_p(1) \\
&= \mathcal{L}_{\alpha\alpha}^{-1}\,\mathcal{L}_{\alpha\beta}\,(\widetilde{G}'\,\widetilde{W}\,\widetilde{G})^{-1}\,\widetilde{G}'\,\widetilde{W}\,\sqrt{n}\,\mathcal{L}_{n,\gamma} \\
&\quad - \left\{\mathcal{L}_{\alpha\alpha}^{-1}\,\mathcal{L}_{\alpha\beta}\,(\widetilde{G}'\,\widetilde{W}\,\widetilde{G})^{-1}\,\widetilde{G}'\,\widetilde{W}\,\mathcal{L}_{\gamma\alpha}\mathcal{L}_{\alpha\alpha}^{-1} + \mathcal{L}_{\alpha\alpha}^{-1}\right\}\,\sqrt{n}\,\mathcal{L}_{n,\alpha} + o_p(1),
\end{aligned}$$

where $\widetilde{W} := \widetilde{H}^{-1}\,\Omega_{\beta_0}\,\widetilde{H}^{-1}$.

The asymptotic representation in Theorem B.0.2 can be used to show asymp-

totic normality of the AIV estimator based on

$$\sqrt{n}\begin{pmatrix}\mathcal{L}_{n,\gamma}\\ \mathcal{L}_{n,\alpha}\end{pmatrix} \xrightarrow{d} \mathcal{N}\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\Sigma_\gamma & \Sigma_{\gamma\alpha}\\ \Sigma'_{\gamma\alpha} & \Sigma_\alpha\end{pmatrix}\right]$$

where $\Sigma_\alpha = \text{Var}\left[\frac{\partial \ell(Y_i|W_i, X_i'\beta_0, \alpha_0)}{\partial \alpha}\right]$, $\Sigma_\gamma = \text{Var}\left[Z_i \frac{\partial \ell\left(Y_i \,|\, W_i, X_i'\beta_0, \alpha_0\right)}{\partial \omega}\right]$, and $\Sigma_{\gamma\alpha}$ was defined in the main text.

B.0.3 Local sign consistency: formal results

In this section, we formalize conditions under which the auxiliary IV estimator is sign-consistent and we show that these conditions are verified in two benchmark models. For this purpose we need some additional notation. Let $\gamma(\cdot,\cdot) : \mathbb{R}^{k_x} \times \mathbb{R}^{k_x} \to \mathbb{R}^{k_z}$ be the function implicitly defined by the relationship

$$s(\beta, \gamma(\beta,\beta_0), \beta_0) = 0, \quad s(\beta,\gamma,\beta_0) := \mathbb{E}_{P_{\beta_0}}\left[\frac{\partial \ell\left(Y_i \,\middle|\, X_i'\beta + Z_i'\gamma\right)}{\partial \omega} Z_i\right],$$

where P_{β_0} denotes the true data generating process parametrized by β_0. Our previous results show that $\widehat{\gamma}(\beta)$ defined in (3.1) converges uniformly to $\gamma(\beta,\beta_0)$ under P_{β_0}. Thus, we can define the probability limit of our auxiliary IV estimator as a function of β_0 as

$$\beta^*(\beta_0) = \underset{\beta\in\mathcal{B}}{\text{argmin}} \left\|\gamma(\beta,\beta_0)\right\|_{\Omega(\beta,\beta_0)}, \tag{B.4}$$

where $\Omega(\beta,\beta_0)$ is the probability limit of $\Omega_{n,\beta}$ under P_{β_0}. The next theorem provides a sufficient condition for local sign consistency of the AIV estimator, which relies on some additional regularity conditions.

Assumption B.0.3 (Additional regularity conditions). *There exists an open set around $\beta^{\mathcal{O}} = (\beta^{\mathcal{O}}_{-k}, 0)$ such that*

(i) The function $s(\beta,\gamma,\beta_0)$ is three-times continuously differentiable with uni-

formly bounded derivatives, and second derivatives

$$G(\beta,\gamma,\beta_0) := \frac{\partial^2 s(\beta,\gamma,\beta_0)}{\partial\gamma\partial\beta'} = \mathbb{E}_{P_{\beta_0}}\left[Z_i X_i' \frac{\partial^2 \ell\left(Y_i \,\middle|\, X_i\beta + Z_i\gamma\right)}{\partial\omega^2}\right],$$
$$H(\beta,\gamma,\beta_0) := \frac{\partial^2 s(\beta,\gamma,\beta_0)}{\partial\gamma\partial\beta'} = \mathbb{E}_{P_{\beta_0}}\left[Z_i Z_i' \frac{\partial^2 \ell\left(Y_i \,\middle|\, X_i\beta + Z_i\gamma\right)}{\partial\omega^2}\right],$$

having singular values uniformly bounded away from 0.

(ii) The function $\Omega(\beta,\beta_0)$ is positive-definite with eigenvalues uniformly bounded away from 0 and uniformly bounded entries that have uniformly bounded continuous derivatives up to second-order.

Assumption B.0.3 imposes additional smoothness conditions on the population score function. These guarantee that the AIV estimator solves a convex optimization problem for data generating process in a neighborhood of $\beta^{\mathcal{O}}$, when the optimization is made over a suitably small set.

Theorem B.0.3. *Suppose that Assumptions 3.1.1,3.2.1, 3.2.2, and B.0.3 hold. Then the auxiliary IV estimator that solves* (3.1) *over a suitably small $\mathcal{B}_* \subseteq \mathcal{B}$ is locally sign consistent if*

$$\left.\frac{\partial\beta_{*,k}(\beta_0)}{\partial\beta_{0,k}}\right|_{\beta_0=\beta^{\mathcal{O}}} = \left[\left(G_{\mathcal{O}}' H_{\mathcal{O}}^{-1}\Omega_{\mathcal{O}} H_{\mathcal{O}}^{-1} G_{\mathcal{O}}\right)^{-1} G_{\mathcal{O}}' H_{\mathcal{O}}^{-1}\Omega_{\mathcal{O}} H_{\mathcal{O}}^{-1} \frac{\partial s(\beta^{\mathcal{O}},0,\beta^{\mathcal{O}})}{\partial\beta_0'}\right]_{(k_x,k_x)} > 0,$$

where $G_{\mathcal{O}} = G(\beta^{\mathcal{O}},0,\beta^{\mathcal{O}})$, $H_{\mathcal{O}} = H(\beta^{\mathcal{O}},0,\beta^{\mathcal{O}})$ and $\Omega_{\mathcal{O}} = \Omega(\beta^{\mathcal{O}},\beta^{\mathcal{O}})$.

In the following subsections, we use the above Lemma to verify local sign consistency of the auxiliary IV estimator in the benchmark models of Examples 3.2.1 and 3.2.2.

B.0.3.1 Details for Example 3.2.1 (Control function)

We have

$$
\begin{aligned}
&s(\beta, \gamma, \beta_0) \\
&= \mathbb{E}_{P_{\beta_0}} \left[(Y_i - \Phi(X_i'\beta + Z_i'\gamma)) \cdot \frac{\phi(X_i'\beta + Z_i'\gamma)}{\Phi(X_i'\beta + Z_i'\gamma) \cdot (1 - \Phi(X_i'\beta + Z_i'\gamma))} \, Z_i \right] \\
&= \mathbb{E}_{P_{\beta_0}} \left[\left\{ F_{U|V}(X'\beta_0 \mid V_i) - \Phi(X_i'\beta + Z_i'\gamma) \right\} \cdot \frac{\phi(X_i'\beta + Z_i'\gamma)}{\Phi(X_i'\beta + Z_i'\gamma) \cdot (1 - \Phi(X_i'\beta + Z_i'\gamma))} \, Z_i \right],
\end{aligned}
$$

since

$$
\begin{aligned}
\mathbb{E}_{P_{\beta_0}}[Y_i \mid X_i, Z_i] &= F_{U|X,Z}(X'\beta_0 \mid X_i, Z_i) \\
&= F_{U|V,Z}(X'\beta_0 \mid V_i, Z_i) \\
&= F_{U|V}(X'\beta_0 \mid V_i),
\end{aligned}
$$

which then gives

$$
\frac{\partial s(\beta^{\mathcal{O}}, 0, \beta^{\mathcal{O}})}{\partial \beta_0} = \frac{\phi(\beta_2^{\mathcal{O}})}{\Phi(\beta_2^{\mathcal{O}}) \cdot (1 - \Phi(\beta_2^{\mathcal{O}}))} \cdot \mathbb{E}\left[f_{U|V}(\beta_2^{\mathcal{O}} \mid V_i) \, Z_i X_i' \right].
$$

Having defined $\widetilde{Z}_i := \Omega_{\mathcal{O}}^{-1/2} \, H_{\mathcal{O}}^{-1} \, Z_i$, we have by Theorem B.0.3 that

$$
\left. \frac{d\beta_*(\beta_0)}{d\beta_0} \right|_{\beta_0 = \beta_{\mathcal{O}}} = \frac{1}{\phi(\beta_2^{\mathcal{O}})} \cdot \left[\mathbb{E}[X_i \widetilde{Z}_i'] \cdot \mathbb{E}[\widetilde{Z}_i' \, X_i] \right]^{-1} \mathbb{E}[X_i \widetilde{Z}_i'] \cdot \mathbb{E}\left[f_{U|V}(\beta_2^{\mathcal{O}} \mid V_i) \, \widetilde{Z}_i X_i' \right], \tag{B.5}
$$

It is useful to define $Q := \left[\mathbb{E}[X_i \widetilde{Z}_i'] \cdot \mathbb{E}[\widetilde{Z}_i' \, X_i] \right]$, for which we we have

$$
\begin{aligned}
Q^{-1} &= \frac{1}{\det(Q)} \cdot \begin{pmatrix} Q_{22} & -Q_{12} \\ -Q_{12} & Q_{11} \end{pmatrix}, \\
Q_{11} &= \sum_{j=1}^{k_z} \mathbb{E}[x_i \, \widetilde{Z}_{i,j}]^2, \quad Q_{12} = \sum_{j=1}^{k_z} \mathbb{E}[x_i \, \widetilde{Z}_{i,j}] \cdot \mathbb{E}[\widetilde{Z}_{i,j}], \quad Q_{22} = \sum_{j=1}^{k_z} \mathbb{E}[\widetilde{Z}_{i,j}]^2.
\end{aligned}
$$

We also have

$$\begin{aligned}\mathbb{E}[f_{U|V}(\beta_2^{\mathcal{O}} \mid V_i)\, Z_i x_i] &= \mathbb{E}[f_{U|V}(\beta_2^{\mathcal{O}} \mid V_i)\, m(Z_i)\widetilde{Z}_i] + \mathbb{E}[f_{U|V}(\beta_2^{\mathcal{O}} \mid V_i)\, V_i \widetilde{Z}_i] \\ &= \mathbb{E}[f_{U|V}(\beta_2^{\mathcal{O}} \mid V_i)] \cdot \mathbb{E}[m(Z_i)\widetilde{Z}_i] + \mathbb{E}[f_{U|V}(\beta_2^{\mathcal{O}} \mid V_i)\, V_i] \cdot \mathbb{E}[\widetilde{Z}_i] \\ &= \mathbb{E}[f_{U|V}(\beta_2^{\mathcal{O}} \mid V_i)] \cdot \mathbb{E}[x_i\widetilde{Z}_i] + \mathbb{E}[f_{U|V}(\beta_2^{\mathcal{O}} \mid V_i)\, V_i] \cdot \mathbb{E}[\widetilde{Z}_i] \\ &= \phi(\beta_2^{\mathcal{O}}) \cdot \mathbb{E}[x_i\widetilde{Z}_i] + \mathbb{E}[f_{U|V}(\beta_2^{\mathcal{O}} \mid V_i)\, V_i] \cdot \mathbb{E}[\widetilde{Z}_i],\end{aligned}$$

where we have used the independence between Z_i and V_i. Thus we can express the lower-diagonal entry of (B.5)

$$\begin{aligned}\left.\frac{\partial \beta_{*,2}(\beta_0)}{\partial \beta_{0,2}}\right|_{\beta_0=\beta_{\mathcal{O}}} &= \left[Q^{-1}\mathbb{E}[X_i\widetilde{Z}_i']\right]_{(2,\bullet)} \cdot \mathbb{E}\left[\widetilde{Z}_i X_i\right]_{(\bullet,2)} \\ &\quad + \frac{\mathbb{E}[f_{U|V}(\beta_2^{\mathcal{O}} \mid V_i)\, V_i]}{\phi(\beta_2^{\mathcal{O}})} \cdot \left[Q^{-1}\mathbb{E}[X_i\widetilde{Z}_i']\right]_{(2,\bullet)} \cdot \mathbb{E}[\widetilde{Z}_i].\end{aligned}$$

The first term in the above expansion is equal to 1 and the second term is equal to 0 since

$$\begin{aligned}&\left[Q^{-1}\mathbb{E}[X_i\widetilde{Z}_i']\right]_{(2,\bullet)} \cdot \mathbb{E}[\widetilde{Z}_i] \\ &= \frac{1}{\det(Q)} \cdot \left[- Q_{12} \cdot \left(\sum_{j=1}^{k_z} \mathbb{E}[\widetilde{Z}_{i,j}]^2 \right) + Q_{11} \cdot \left(\sum_{j=1}^{k_z} \mathbb{E}[x_i\widetilde{Z}_{i,j}] \cdot \mathbb{E}[\widetilde{Z}_{i,j}] \right) \right] \\ &= \frac{1}{\det(Q)} \cdot \left[- \left(\sum_{\ell=1}^{k_z} \mathbb{E}[x_i\, \widetilde{Z}_{i,\ell}] \cdot \mathbb{E}[\widetilde{Z}_{i,j}] \right) \cdot \left(\sum_{j=1}^{k_z} \mathbb{E}[\widetilde{Z}_{i,j}]^2 \right) \right. \\ &\qquad\qquad \left. + \left(\sum_{j=1}^{k_z} \mathbb{E}[\widetilde{Z}_{i,j}]^2 \right) \cdot \left(\sum_{\ell=1}^{k_z} \mathbb{E}[x_i\widetilde{Z}_{i,j}] \cdot \mathbb{E}[\widetilde{Z}_{i,j}] \right) \right] \\ &= 0.\end{aligned}$$

Hence we conclude that

$$\left.\frac{\partial \beta_{*,2}(\beta_0)}{\partial \beta_{0,2}}\right|_{\beta_0=\beta_{\mathcal{O}}} = 1.$$

B.0.3.2 Details for Example 3.2.2 (Generalized bivariate Probit)

Standard calculations (see Section 15.7.3 of Wooldridge, 2010) give:

$$\begin{aligned}
&\mathbb{E}_{P_{\beta_0}}[Y_i \mid X_i, Z_i] = \mathbb{E}\left[F_{U|V}(X_i'\beta_0 \mid V_i) \mid X_i, Z_i\right] \\
&= \frac{X_i}{F_V(m(Z_i))} \cdot \int_{-m(Z_i)}^{\infty} F_{U|V}(X_i'\beta_0 \mid V_i) \cdot f_V(v)dv \\
&\qquad\qquad + \frac{1 - X_i}{1 - F_V(m(Z_i))} \cdot \int_{-\infty}^{-m(Z_i)} F_{U|V}(X_i'\beta_0 \mid V_i) \cdot f_V(v)dv,
\end{aligned}$$

which then gives

$$\frac{\partial s(\beta^{\mathcal{O}}, 0, \beta^{\mathcal{O}})}{\partial \beta_{0,2}} = \frac{\phi(\beta_2^{\mathcal{O}})}{\Phi(\beta_2^{\mathcal{O}}) \cdot (1 - \Phi(\beta_2^{\mathcal{O}}))} \cdot \mathbb{E}\left[Z_i \cdot \int_{-m(Z_i)}^{\infty} f_{U|V}(\beta_2^{\mathcal{O}} \mid V_i) \cdot f_V(v)dv\right].$$

Using Theorem B.0.3 we obtain

$$\begin{aligned}
\left.\frac{\partial \beta_{*,2}(\beta_0)}{\partial \beta_{0,2}}\right|_{\beta_0 = \beta^{\mathcal{O}}} &= \frac{1}{\phi(\beta_2^{\mathcal{O}})} \cdot \mathbb{E}[Z_i X_i]^{-1}_{(2,\bullet)} \cdot \mathbb{E}\left[Z_i \cdot \int_{-m(Z_i)}^{\infty} f_{U|V}(\beta_2^{\mathcal{O}} \mid V_i) \cdot f_V(v)dv\right] \\
&= \frac{1}{\phi(\beta_2^{\mathcal{O}})} \cdot \frac{\text{Cov}\left(z_i, \int_{-m(Z_i)}^{\infty} f_{U|V}(\beta_2^{\mathcal{O}} \mid V_i) \cdot f_V(v)dv\right)}{\text{Cov}\left(z_i, F_V(m(Z_i))\right)}.
\end{aligned}$$

The functions $\int_{-Z_i\delta}^{\infty} f_{U|V}(\beta_2^{\mathcal{O}} \mid V_i) \cdot f(v)dv$ and $F_V(m(Z_i))$ are both monotonic increasing (decreasing) in z_i when $m(Z_i)$ is monotonic increasing (decreasing). As a result, the two covariances in the above display have concordant signs and we conclude that the auxiliary IV estimator is sign consistent.

B.0.4 Technical Lemmas

Lemma B.0.1. *Under the Assumptions of Theorem B.0.3, there exists a convex and compact set $\mathcal{B}_*$ containing $\beta^{\mathcal{O}}$ such that the optimization problem*

$$\underset{\beta \in \mathcal{B}_*}{\text{argmin}} \, \|\gamma(\beta, \beta_0)\|_{\Omega(\beta,\beta_0)}$$

is convex for all $\beta_0 \in \mathcal{B}_*$.

B.0.5 Proofs

B.0.5.1 Proof of Lemma 3.1.1

Concavity of the log-likelihood, we have that $\widehat{\gamma}(\beta)$ is uniquely characterized by the FOC

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \ell\left(Y_i \,\middle|\, X_i'\beta + Z_i'\gamma,\, \alpha\right)}{\partial \gamma} Z_i = 0$$

By a mean-value expansion of the LHS in γ around $\gamma = 0$, the last display equation becomes

$$\widehat{\gamma}(\beta) = -H_n(\beta, \gamma_*(\beta))^{-1} \cdot \left[\frac{1}{n}\sum_{i=1}^{n}\frac{d\ell\left(Y_i \,\middle|\, X_i'\beta\right)}{d\omega} Z_i\right].$$

Plugging the above into the objective function $\|\widehat{\gamma}(\beta)\|_{\Omega_{n,\beta}}$ gives the desired equivalence.

B.0.5.2 Proof of Theorem B.0.1

We begin by defining the population (large n limit) analog of (3.10) as

$$\begin{aligned}
(\gamma(\beta), \alpha(\beta)) &= \underset{\gamma,\, \alpha \in \mathcal{E}}{\operatorname{argmax}}\ \mathcal{L}(\beta, \gamma, \alpha), \\
\beta^* &= \left\{\beta : \beta \in \underset{\beta \in \mathcal{B}}{\operatorname{argmin}}\ \|\gamma(\beta)\|_{\Omega_{n,\beta}}\right\}, \\
\alpha^* &= \left\{\alpha(\beta) : \beta \in \beta^*\right\}, \\
\alpha^\dagger(\beta) &= \underset{(\alpha,0) \in \mathcal{E}}{\operatorname{argmax}}\ \mathcal{L}(\beta, 0, \alpha), \\
\alpha_*^\dagger &= \left\{\alpha^\dagger(\beta) : \beta \in \beta^*\right\}.
\end{aligned}$$

The proof consists of two parts. In Part I we show that $\beta^* = \beta_0$ and $\alpha^* = \alpha_0$. In Part II we use the identification result of Part I to show consistency of $(\widehat{\beta}_{\mathrm{AIV}}, \widehat{\alpha}_{\mathrm{AIV}})$.

Part I: Strict concavity of the expected log-likelihood in η (Assumption B.0.2(ii))

guarantee that $(\gamma(\beta), \alpha(\beta))$ are uniquely defined by the FOC

$$\frac{\partial \mathcal{L}(\beta, \eta)}{\partial \eta} = 0,$$

for which we have $\gamma(\beta_0) = 0$, $\alpha(\beta_0) = \alpha_0$ by Assumption B.0.1. Suppose there exists $\check{\beta} \in \beta^*$ with $\check{\beta} \neq \beta_0$, so that

$$\frac{\partial \mathcal{L}(\beta_0, 0, \alpha(\beta_0))}{\partial \eta} = 0,$$
$$\frac{\partial \mathcal{L}(\check{\beta}, 0, \alpha(\check{\beta}))}{\partial \eta} = 0.$$

By a mean value expansion of $\frac{\partial \mathcal{L}(\beta,0,\alpha)}{\partial \eta}$ in (β, α):

$$0 = \frac{\partial \mathcal{L}(\check{\beta}, 0, \alpha)}{\partial \eta} - \frac{\partial \mathcal{L}(\beta_0, 0, \alpha(\check{\beta}))}{\partial \eta} = \underbrace{\begin{pmatrix} \frac{\partial^2 \mathcal{L}(\widetilde{\beta},0,\widetilde{\alpha})}{\partial \gamma \partial \beta'} & \frac{\partial^2 \mathcal{L}(\widetilde{\beta},0,\widetilde{\alpha})}{\partial \gamma \partial \alpha'} \\ \frac{\partial^2 \mathcal{L}(\widetilde{\beta},0,\widetilde{\alpha})}{\partial \alpha \partial \beta'} & \frac{\partial^2 \mathcal{L}(\widetilde{\beta},0,\widetilde{\alpha})}{\partial \alpha \partial \alpha'} \end{pmatrix}}_{=A(\widetilde{\beta},0,\widetilde{\alpha})} \begin{pmatrix} \check{\beta} - \beta_0 \\ \alpha(\check{\beta}) - \alpha_0 \end{pmatrix}$$

where $(\widetilde{\beta}, \widetilde{\alpha})$ is an intermediate value between $(\check{\beta}, \alpha(\check{\beta}))$ and (β_0, α_0). The matrix $A(\widetilde{\beta}, 0, \widetilde{\alpha})$ has full rank by Assumption B.0.2 (iii), and therefore we conclude from the previous display that

$$\check{\beta} = \beta_0, \qquad \alpha(\check{\beta}) = \alpha_0.$$

By similar arguments we have $\alpha^\dagger(\beta_0) = \alpha_0$ and thus $\alpha_*^\dagger = \alpha_0$.

Part II: show $(\widehat{\beta}_{\mathrm{AIV}}, \widehat{\alpha}_{\mathrm{AIV}}) = (\beta_0, \alpha_0) + o_P(1)$**.**

First define

$$\widehat{\eta}(\beta) = \underset{\eta \in \mathcal{E}}{\operatorname{argmax}} \sum_{i=1}^{n} \ell\left(Y_i \,\middle|\, W_i,\, X_i'\beta + Z_i'\gamma,\, \alpha\right),$$
$$\eta(\beta) = \underset{\eta \in \mathcal{E}}{\operatorname{argmax}}\, \mathbb{E}\, \ell\left(Y_i \,\middle|\, W_i,\, X_i'\beta + Z_i'\gamma,\, \alpha\right),$$

Then by Pollard's convexity lemma we know that

$$\sup_{\beta\in\mathcal{B}}\sup_{\gamma\in\mathcal{E}}\left|\frac{1}{n}\sum_{i=1}^{n}\ell\left(Y_i\,\middle|\,W_i,\,X_i'\beta+Z_i'\gamma,\,\alpha\right)-\mathbb{E}\,\ell\left(Y_i\,\middle|\,W_i,\,X_i'\beta+Z_i'\gamma,\,\alpha\right)\right|=o_P(1).$$

Having this, we satisfy all the assumptions of Lemma B.1 in Chernozhukov and Hansen (2006), and therefore conclude

$$\sup_{\beta\in\mathcal{B}}\|\widehat{\eta}(\beta)-\eta(\beta)\|=o_P(1).$$

It directly follows that the objective function $\|\widehat{\gamma}(\beta)\|_{\Omega_{n,\beta}}$ converges uniformly to $\|\gamma(\beta)\|_{\Omega}$, which together with the continuity of $\|\gamma(\beta)\|_{\Omega}$ and β_0 being its unique minimizer over the compact set $\mathcal{B}$ (Part I) ensures that standard conditions for consistency of extremum estimators are satisfied (see, e.g., Theorem 2.1 in Newey and McFadden, 1994). We thus conclude

$$\widehat{\beta}_{\mathrm{AIV}}=\beta_0+o_P(1).$$

By analogous arguments we have $\sup_{(\beta,\gamma)\in(\mathcal{B},\mathcal{E})}\|\widehat{\alpha}^{\dagger}(\beta,\gamma)-\alpha^{\dagger}(\beta,\gamma)\|=o_p(1)$. Furthermore, consistency of $\widehat{\beta}_{\mathrm{AIV}}$ and continuity of $\alpha^{\dagger}(\beta,\gamma)$ imply that $\alpha^{\dagger}(\widehat{\beta}_{\mathrm{AIV}},0)=\alpha(\beta_0)+o_P(1)$. This, together with the uniform consistency of $\widehat{\alpha}^{\dagger}(\beta,\gamma)$, guarantees that

$$\widehat{\alpha}_{\mathrm{AIV}}=\alpha_0+o_P(1).$$

B.0.5.3 Proof of Theorem B.0.2

The proof is in three parts. First, we show that

$$\|\widehat{\gamma}(\beta)\|_{\Omega_{n,\beta}}=\|s_*(\beta)\|_{W_{n,\beta}},\tag{B.6}$$

with

$$s_*(\beta) = \mathcal{L}_{n,\gamma}(\beta, 0, \alpha_0) - \mathcal{L}_{n,\gamma\alpha}(\beta, \eta_*(\beta))\, \mathcal{L}_{n,\alpha\alpha}(\beta, \eta_*(\beta))^{-1}\, \mathcal{L}_{n,\alpha}(\beta, 0, \alpha_0).$$
$$W_{n,\beta} = \widetilde{H}_n(\beta, \eta_*(\beta))^{-1}\, \Omega_{n,\beta}\, \widetilde{H}_n(\beta, \eta_*(\beta))^{-1},$$

where $\eta_*(\beta) = (\gamma_*(\beta), \alpha_*(\beta))$ lies on the line between $(\widehat{\gamma}(\beta), \widehat{\alpha}(\beta))$ and $(0, \alpha_0)$. In Part II, we use the result from Part I to derive the asymptotic representation for $\widehat{\beta}_{\text{AIV}}$. In Part III, we use the result from Part II to derive the asymptotic representation for $\widehat{\alpha}_{\text{AIV}}$.

Part I: Strict concavity of the sample log-likelihood in η guarantees that $\widehat{\eta}(\beta) = (\widehat{\gamma}(\beta), \widehat{\alpha}(\beta))$ are uniquely defined by the FOC

$$\mathcal{L}_{n,\eta}(\beta, \widehat{\eta}(\beta)) = 0.$$

A mean-value expansion the above around $(\gamma, \alpha) = (0, \alpha_0)$ gives

$$\begin{aligned} &\mathcal{L}_{n,\eta}(\beta, 0, \alpha_0) + \mathcal{L}_{n,\eta\eta}(\beta, \eta_*(\beta)) \cdot \widehat{\eta}(\beta) = 0 \\ &\quad \implies \widehat{\eta}(\beta) = -\mathcal{L}_{n,\eta\eta}(\beta, \eta_*(\beta))^{-1} \cdot \mathcal{L}_{n,\eta}(\beta, 0, \alpha_0). \end{aligned}$$

Using the partitioned inverse formula we obtain

$$\widehat{\gamma}(\beta) = -\widetilde{H}_n(\beta, \eta_*(\beta))^{-1} \cdot s_*(\beta)$$

Plugging the above into $\|\widehat{\gamma}(\beta)\|_{\Omega_{n,(\beta,\alpha)}}$ gives (B.6).

Part II: Define

$$\widehat{\beta}^\dagger := \beta_0 - \left(\widetilde{G}'\, \widetilde{W}\, \widetilde{G}\right)^{-1}\, \widetilde{G}'\, \widetilde{W}\, s(\beta), \quad s(\beta) = \mathcal{L}_{n,\gamma}(\beta, 0, \alpha_0) - \mathcal{L}_{\gamma\alpha}\, \mathcal{L}_{\alpha\alpha}^{-1}\, \mathcal{L}_{n,\alpha}(\beta, 0, \alpha_0).$$

By definition, $\widehat{\beta} := \widehat{\beta}_{\text{AIV}}$ minimizes $s_*(\beta)'\, W_{n,\beta}\, s_*(\beta)$. Therefore,

$$s_*(\widehat{\beta})'\, W_{n,\widehat{\beta}}\, s_*(\widehat{\beta}) \leq s_*(\widehat{\beta}^\dagger)'\, W_{n,\widehat{\beta}^\dagger}\, s_*(\widehat{\beta}^\dagger). \tag{B.7}$$

Uniform convergence of $\widehat{\eta}(\beta)$ to $\eta(\beta)$ (see proof of Theorem B.0.1), along with consistency of $\widehat{\beta}$, implies $\eta_*(\widehat{\beta}) = (0, \alpha_0) + o_P(1)$. This, together with uniform consistency of the second derivatives of $\mathcal{L}_n$ and $\Omega_{n,\beta}$ implies that $W_{n,\widehat{\beta}} = \widetilde{W} + o_p(1)$. Uniform convergence of $\widetilde{G}_n$ to $\widetilde{G}$ justifies the expansions

$$
\begin{aligned}
s_*(\widehat{\beta}) &= s(\beta_0) + \widetilde{G}\,(\widehat{\beta} - \beta_0) + o_P\left(\|\widehat{\beta} - \beta_0\|\right), \\
s_*(\widehat{\beta}^\dagger) &= s(\beta_0) + \widetilde{G}\,(\widehat{\beta}^\dagger - \beta_0) + o_P\left(\|\widehat{\beta}^\dagger - \beta_0\|\right) = s(\beta_0) + \widetilde{G}\,(\widehat{\beta}^\dagger - \beta_0) + o_P\left(\frac{1}{\sqrt{n}}\right),
\end{aligned}
$$

where we have used that $\sqrt{n}(\widehat{\beta}^\dagger - \beta_0) = O_p(1)$. Plugging the expansions into (B.7) and using $\widetilde{W}_{n,\widehat{\beta}} = \widetilde{W} + o_P(1)$ gives for the LHS

$$
\begin{aligned}
&\left[s(\beta_0) + \widetilde{G}\,(\widehat{\beta} - \beta_0) + o_P\left(\|\widehat{\beta} - \beta_0\|\right)\right]' \widetilde{W} \left[s(\beta_0) + \widetilde{G}\,(\widehat{\beta} - \beta_0) + o_P\left(\|\widehat{\beta} - \beta_0\|\right)\right] \\
&\qquad\qquad + \quad R(\widehat{\beta}),
\end{aligned}
$$

with

$$
\begin{aligned}
R(\widehat{\beta}) = o_P(1) \cdot \Big[& s(\beta_0)' s(\beta_0) + (\widehat{\beta} - \beta_0)' \widetilde{G}' \widetilde{G} (\widehat{\beta} - \beta_0) + o_P(\|\widehat{\beta} - \beta_0\|^2) \\
& + 2\, s(\beta_0)' \widetilde{G} (\widehat{\beta} - \beta_0) + 2\, s(\beta_0)'\, o_P\left(\|\widehat{\beta} - \beta_0\|\right) + 2\,(\widehat{\beta} - \beta_0)\, \widetilde{G}' \widetilde{G}\, o_P\left(\|\widehat{\beta} - \beta_0\|\right) \Big] \\
= o_P & \left(\|\widehat{\beta} - \beta_0\|^2 + \frac{1}{\sqrt{n}} \|\widehat{\beta} - \beta_0\| + \frac{1}{n}\right),
\end{aligned}
$$

where we have used $s(\beta_0) = O_P(1/\sqrt{n})$. Similarly, for the RHS we have

$$
\begin{aligned}
&\left[s(\beta_0) + \widetilde{G}\,(\widehat{\beta}^\dagger - \beta_0) + o_P\left(\|\widehat{\beta}^\dagger - \beta_0\|\right)\right]' \widetilde{W} \left[s(\beta_0) + \widetilde{G}\,(\widehat{\beta}^\dagger - \beta_0) + o_P\left(\|\widehat{\beta}^\dagger - \beta_0\|\right)\right] \\
&\qquad\qquad + \quad R(\widehat{\beta}^\dagger),
\end{aligned}
$$

with

$$
\begin{aligned}
R(\widehat{\beta}^\dagger) &= o_P\left(\|\widehat{\beta}^\dagger - \beta_0\|^2 + \frac{1}{\sqrt{n}} \|\widehat{\beta}^\dagger - \beta_0\| + \frac{1}{n}\right) \\
&= o_P\left(\frac{1}{n}\right).
\end{aligned}
$$

Combining the previous results with the inequality (B.7) gives

$$\begin{aligned}
&\left[s(\beta_0) + \widetilde{G}\,(\widehat{\beta} - \beta_0)\right]' \widetilde{W} \left[s(\beta_0) + \widetilde{G}\,(\widehat{\beta} - \beta_0)\right] \\
&\quad \leq \left[s(\beta_0) + \widetilde{G}\,(\widehat{\beta}^{\dagger} - \beta_0)\right]' \widetilde{W} \left[s(\beta_0) + \widetilde{G}\,(\widehat{\beta}^{\dagger} - \beta_0)\right] \\
&\qquad\qquad + o_P\left(\|\widehat{\beta} - \beta_0\|^2 + \frac{1}{\sqrt{n}}\|\widehat{\beta} - \beta_0\| + \frac{1}{n}\right). \qquad \text{(B.8)}
\end{aligned}$$

We now decompose $s(\beta) = A_1 + A_2$, where

$$A_1 = \widetilde{G}(\widetilde{G}'\widetilde{W}\widetilde{G})^{-1}\widetilde{G}'\widetilde{W}s(\beta), \qquad A_2 = \left[\mathbb{I} - \widetilde{G}(\widetilde{G}'\widetilde{W}\widetilde{G})^{-1}\widetilde{G}'\widetilde{W}\right] s(\beta).$$

Because $\widetilde{G}'\widetilde{W}A_2 = 0$, we find that the contributions of A_2 on both sides of the inequality (B.8) are identical and thus drop out. Also plugging in the definition of $\widehat{\beta}^{\dagger}$, this inequality becomes

$$\begin{aligned}
&\left[(\widehat{\beta} - \beta^0) + \underbrace{(\widetilde{G}'\widetilde{W}\widetilde{G})^{-1}\widetilde{G}'\widetilde{W}s(\beta_0)}_{:=L}\right]' \widetilde{G}'\widetilde{W}\widetilde{G} \\
&\quad \times \left[(\widehat{\beta} - \beta^0) + (\widetilde{G}'\widetilde{W}\widetilde{G})^{-1}\widetilde{G}'\widetilde{W}s(\beta_0)\right] \leq o_P\left(\|\widehat{\beta} - \beta_0\|^2 + \frac{1}{\sqrt{n}}\|\widehat{\beta} - \beta_0\| + \frac{1}{n}\right).
\end{aligned}$$

Because $\widetilde{G}'\widetilde{W}\widetilde{G}$ has full rank (since $\widetilde{W} > 0$ and $\mathrm{rank}(\widetilde{G}) = k_x$) we have that

$$\begin{aligned}
\|(\widehat{\beta} - \beta_0) + L\|^2 &\leq o_P(1) \cdot \left(\|\widehat{\beta} - \beta_0\|^2 + \frac{1}{\sqrt{n}}\|\widehat{\beta} - \beta_0\| + \frac{1}{n}\right) \\
&\leq o_P(1) \cdot \left(\|\widehat{\beta} - \beta_0 + L\|^2 + \frac{1}{\sqrt{n}}\|\widehat{\beta} - \beta_0 + L\| + \|L\|^2 + \frac{1}{\sqrt{n}}\|L\| + \frac{1}{n}\right) \\
&\leq o_P\left(\frac{1}{\sqrt{n}}\right) \cdot \|\widehat{\beta} - \beta_0 + L\| + o_P\left(\frac{1}{n}\right),
\end{aligned}$$

where we have used $L = O_P(1/\sqrt{n})$. Denoting $\xi_n := o_P\left(\frac{1}{\sqrt{n}}\right) \cdot \|\widehat{\beta} - \beta_0 + L\|$ we can re-write the above as

$$\left(\|\widehat{\beta} - \beta_0 + L\| - \xi_n\right)^2 \leq \xi_n^2 + o_P\left(\frac{1}{n}\right),$$

from which we conclude

$$\sqrt{n}(\widehat{\beta} - \beta_0) = \sqrt{n}L + o_P(1).$$

Part III: Consider the decomposition

$$\widehat{\alpha}^\dagger(\widehat{\beta}) - \alpha_0 = [\widehat{\alpha}^\dagger(\widehat{\beta}) - \widehat{\alpha}^\dagger(\beta_0)] + [\widehat{\alpha}^\dagger(\beta_0) - \alpha(\beta_0)]. \tag{B.9}$$

For the first term we consider the mean-value expansion:

$$\widehat{\alpha}^\dagger(\widehat{\beta}) - \widehat{\alpha}^\dagger(\beta_0) = \frac{d\widehat{\alpha}^\dagger(\beta)}{d\beta}\Big|_{\beta=\widetilde{\beta}} \cdot (\widehat{\beta} - \beta_0),$$

for $\widetilde{\beta}$ between β_0 and $\widehat{\beta}$. By the implicit function theorem, we have that in a neighbourhood of $(\beta_0, \widehat{\alpha}^\dagger(\beta_0))$

$$\frac{\partial\widehat{\alpha}^\dagger(\beta)}{\partial\beta'} = -\mathcal{L}_{n,\alpha\alpha}(\beta, 0, \widehat{\alpha}^\dagger(\beta))^{-1} \cdot \mathcal{L}_{n,\alpha\beta}(\beta, 0, \widehat{\alpha}^\dagger(\beta)).$$

Using $\widetilde{\beta} = \beta_0 + o_p(1)$ and the usual uniform convergence arguments we obtain

$$\widehat{\alpha}(\widehat{\beta}) - \widehat{\alpha}(\beta_0) = -\mathcal{L}_{\alpha\alpha}^{-1}\,\mathcal{L}_{\alpha\beta}\,(\widehat{\beta} - \beta_0) + o_p\left(\|\widehat{\beta} - \beta_0\|\right). \tag{B.10}$$

Furthermore, classical likelihood results give

$$\widehat{\alpha}^\dagger(\beta_0) - \alpha_0 = -\mathcal{L}_{\alpha\alpha}^{-1}\,\mathcal{L}_{n,\alpha} + o_p(1/\sqrt{n}). \tag{B.11}$$

Plugging (B.10) and (B.11) into (B.9) gives the asymptotic representation for $\widehat{\alpha}$.

B.0.5.4 Proof of Theorem B.0.3

The proof is made of three parts. In Part I we derive the formula for $\frac{d\beta_*(\beta^{\mathcal{O}})}{d\beta_0}$. In Part II we argue that $\frac{d\beta_*(\beta_0)}{d\beta_0}$ is bounded and continuous around $\beta^{\mathcal{O}}$. In Part III we finally show local sign consistency of the AIV estimator defined as the minimizer of the objective function over a suitably small closed ball around $\beta^{\mathcal{O}}$.

Part I: The function $s(\beta, \gamma, \beta_0)$ is thrice continuously differentiable, and thus by the implicit function theorem the function $\gamma(\beta, \beta_0)$ is thrice-continuously differentiable in an open set containing $(\beta, \beta_0) = (\beta^{\mathcal{O}}, \beta^{\mathcal{O}})$ with first-derivatives equal to

$$\begin{aligned}\frac{\partial \gamma(\beta, \beta_0)}{\partial \beta'} &= \left[H(\beta, \gamma(\beta, \beta_0), \beta_0)\right]^{-1} G(\beta, \gamma(\beta, \beta_0), \beta_0), \\ \frac{\partial \gamma(\beta, \beta_0)}{\partial \beta_0'} &= -\left[H(\beta, \gamma(\beta, \beta_0), \beta_0)\right]^{-1} \cdot \frac{s(\beta, \gamma(\beta, \beta_0), \beta_0)}{\partial \beta_0'}.\end{aligned}$$

Thrice-differentiability of $\gamma(\beta, \beta_0)$ together with Technical Lemma B.0.1 implies that the limit of the AIV estimator $\beta^*(\beta_0)$ is characterized around $\beta^{\mathcal{O}}$ by the FOC of the minimisation in (B.4):

$$\Pi(\beta, \beta_0) := 2\frac{\partial \gamma(\beta, \beta_0)}{\partial \beta'}' \Omega(\beta, \beta_0)\gamma(\beta, \beta_0) + \sum_{i,j} \gamma_i(\beta, \beta_0) \cdot \gamma_j(\beta, \beta_0) \cdot \frac{\partial \Omega_{i,j}(\beta, \beta_0)}{\partial \beta} = 0, \tag{B.12}$$

when the estimator maximizes the objective function over the closed ball $B_{\infty,\epsilon}(\beta^{\mathcal{O}})$. We now apply the implicit function theorem to (B.12), where thrice-differentiability of $\gamma(\beta, \beta_0)$ implies that $\beta^*(\beta_0)$ is twice-differentiable with

$$\begin{aligned}\frac{d\beta_*(\beta^{\mathcal{O}})}{\beta_0'} &= -\left[\frac{\partial \Pi(\beta^{\mathcal{O}}, 0, \beta^{\mathcal{O}})}{\partial \beta'}\right]^{-1} \frac{\partial \Pi(\beta^{\mathcal{O}}, 0, \beta^{\mathcal{O}})}{\partial \beta_0'} \\ &= -\left[\frac{\partial \gamma(\beta^{\mathcal{O}}, \beta^{\mathcal{O}})}{\partial \beta'}' \Omega_{\mathcal{O}} \frac{\partial \gamma(\beta^{\mathcal{O}}, \beta^{\mathcal{O}})}{\partial \beta'}\right]^{-1} \frac{\partial \gamma(\beta^{\mathcal{O}}, \beta^{\mathcal{O}})}{\partial \beta'}' \Omega_{\mathcal{O}} \frac{\partial \gamma(\beta^{\mathcal{O}}, \beta^{\mathcal{O}})}{\partial \beta_0'} \\ &= \left(G_{\mathcal{O}}' H_{\mathcal{O}}^{-1} \Omega_{\mathcal{O}} H_{\mathcal{O}}^{-1} G_{\mathcal{O}}\right)^{-1} G_{\mathcal{O}}' H_{\mathcal{O}}^{-1} \Omega_{\mathcal{O}} H_{\mathcal{O}}^{-1} \frac{\partial s(\beta^{\mathcal{O}}, 0, \beta^{\mathcal{O}})}{\partial \beta_0'},\end{aligned}$$

where we have used that $\gamma(\beta^{\mathcal{O}}, \beta^{\mathcal{O}}) = 0$.

Part II: Applying the implicit function theorem twice to (B.12) shows, after some simple but tedious algebra, that $\frac{d^2\beta_*(\beta_0)}{(d\beta_0)^2}$ is bounded and continuous in a neighborhood of $\beta^{\mathcal{O}}$ when H,G and Ω are bounded and have singular values bounded away from zero, uniformly in (β, γ, β_0), which we assume.

Part III: We consider the Taylor expansion of $\beta_{*,k}(\beta_0)$ with respect to $\beta_{0,k}$ around

$\beta^{\mathcal{O}}$:

$$\beta_{*,k}(\beta_0) = \frac{\partial \beta_{*,k}(\beta^{\mathcal{O}})}{\partial \beta_{0,k}} \cdot \beta_{0,k} + \frac{\partial^2 \beta_{*,k}(\tilde{\beta})}{(\partial \beta_{0,k})^2} \cdot (\beta_{0,k})^2$$

where $\tilde{\beta}$ is an intermediate point between $(\beta^{\mathcal{O}}_{-k}, \beta_{0,k})$ and $(\beta^{\mathcal{O}}_{-k}, 0)$, and we have used that $\beta^{\mathcal{O}}_{0,k} = 0$. Multiplying the above by $\beta^{0,k}$ we obtain

$$\beta_{*,k}(\beta_0) \cdot \beta_{0,k} = \frac{\partial \beta_{*,k}(\beta^{\mathcal{O}})}{\partial \beta_{0,k}} \cdot \beta^2_{0,k} + \frac{\partial^2 \beta_{*,k}(\tilde{\beta})}{(\partial \beta_{0,k})^2} \cdot \beta^3_{0,k}. \tag{B.13}$$

Continuity of $\frac{\partial^2 \beta_{*,k}(\beta_0)}{(\partial \beta_{0,k})^2}$ implies that for an arbitrary $\varepsilon > 0$ there exists a $\delta_\varepsilon > 0$ such that for any $|\beta_{0,k}| < \delta_\varepsilon$ one has $\left|\frac{\partial^2 \beta_{*,k}(\beta_0)}{(\partial \beta_{0,k})^2}\right| < C_\varepsilon := \left|\frac{\partial^2 \beta_{*,k}(\beta^{\mathcal{O}})}{(\partial \beta_{0,k})^2}\right| + \varepsilon$. Fixing such ε, and using that $\frac{\partial \beta_{*,k}(\beta^{\mathcal{O}})}{\partial \beta_{0,k}} > 0$, we have that $\beta_{*,k}(\beta_0) \cdot \beta_{0,k} > 0$ for any $\beta_{0,k}$ small enough to satisfy the requirements of the implicit function theorem in Part I and II and

$$0 < |\beta_{0,k}| < \min \left\{ \delta_\epsilon, \delta_\varepsilon, \frac{\partial \beta_{*,k}(\beta^{\mathcal{O}})}{\partial \beta_{0,k}} / C_\varepsilon \right\}.$$

B.0.5.5 Proof of Lemma B.0.1

We want to find a convex set $\mathcal{B}_*$ containing $\beta^{\mathcal{O}}$ for which the objective function $\|\gamma(\beta, \beta_0)\|_{\Omega(\beta,\beta_0)}$ is convex in $\beta \in \mathcal{B}_*$ for all $\beta_0 \in \mathcal{B}_*$. By the continuous twice-differentiability of $\gamma(\beta, \beta_0)$ and Ω_{β,β_0} wrt (β, β_0), for every $\epsilon > 0$ there exists a δ_ϵ such that for $\|(\beta, \beta_0) - (\beta^{\mathcal{O}}, \beta^{\mathcal{O}})\|_\infty \leq \delta_\epsilon$ we have

$$\left\| \frac{\partial \, \|\gamma(\beta, \beta_0)\|_{\Omega_{(\beta,\beta_0)}}}{\partial \beta \, \partial \beta'} - \frac{\partial \gamma(\beta^{\mathcal{O}}, \beta^{\mathcal{O}})'}{\partial \beta} \Omega_{(\beta^{\mathcal{O}}, \beta^{\mathcal{O}})} \frac{\partial \gamma(\beta^{\mathcal{O}}, \beta^{\mathcal{O}})}{\partial \beta} \right\| \leq \epsilon.$$

Denote $\mathcal{C}_\lambda$ the minimum eigenvalue of $\frac{\partial \gamma(\beta^{\mathcal{O}}, \beta^{\mathcal{O}})'}{\partial \beta} \Omega_{(\beta^{\mathcal{O}}, \beta^{\mathcal{O}})} \frac{\partial \gamma(\beta^{\mathcal{O}}, \beta^{\mathcal{O}})}{\partial \beta}$, which is bounded away from 0 by Assumption 3.2.2. By Weyl's Inequality we have

$$\left| \lambda_{\min} \left(\frac{\partial \, \|\gamma(\beta, \beta_0)\|_{\Omega_{(\beta,\beta_0)}}}{\partial \beta \, \partial \beta'} \right) - \mathcal{C}_\lambda \right| \leq \epsilon$$

Choosing $\epsilon < \mathcal{C}_\lambda$ ensures that objective function is convex with respect to β over the convex set $B_{\infty,\epsilon}(\beta^{\mathcal{O}})$ for all $\beta_0 \in B_{\infty,\epsilon}(\beta^{\mathcal{O}})$, where $B_{\infty,\epsilon}(\beta^{\mathcal{O}})$ denotes a closed ball around $\beta^{\mathcal{O}}$ with respect to the ℓ_∞-norm.

Appendix C

Appendix – Chapter 3

This appendix is organized as follows. Section C.1 presents the assumptions and the variance estimators studied in Chapter 4 for the fully general case that allows for misspecification bias in the model. Section C.2 presents the main results of Chapter 4 under the setup described in Section C.1. Section C.3 presents the technical lemmas needed to establish the main results of the chapter. Section C.4 provides the proofs for the main results. Section C.5 provides the proofs for the technical lemmas. Section C.6 presents an extension of the variance estimators studied in Chapter 4 which allows to impose within-cluster zero restrictions on the variance-covariance matrix of the errors.

C.1 Setup - general case

C.1.1 Assumptions

Suppose that $\{(y_{i,n}, \mathbf{x}'_{i,n}, \mathbf{w}'_{i,n}) : 1 \leq i \leq n\}$ is generated by

$$y_{i,n} = \boldsymbol{\beta}'\mathbf{x}_{i,n} + \boldsymbol{\gamma}'_n\mathbf{w}_{i,n} + u_{i,n}, \qquad i = 1, \ldots, n, \tag{C.1}$$

for which $\mathcal{W}_n$ is a collection of random variables such that $\mathbb{E}[\mathbf{w}_{i,n}|\mathcal{W}_n] = \mathbf{w}_{i,n}$, and we set $\mathcal{X}_n = (\mathbf{x}_{1,n}, \ldots, \mathbf{x}_{n,n})$. We define the following quantities:

$$\begin{aligned}
\varrho_n &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[R_{i,n}^2], \qquad R_{i,n} = \mathbb{E}[u_{i,n}|\mathcal{X}_n, \mathcal{W}_n], \\
\rho_n &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[r_{i,n}^2], \qquad r_{i,n} = \mathbb{E}[u_{i,n}|\mathcal{W}_n], \\
\chi_n &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\mathbf{Q}_{i,n}\|^2], \qquad \mathbf{Q}_{i,n} = \mathbb{E}[\mathbf{v}_{i,n}|\mathcal{W}_n], \\
\hat{\mathbf{\Gamma}}_n &= \sum_{i=1}^{n}\hat{\mathbf{v}}_{i,n}\hat{\mathbf{v}}_{i,n}'/n, \qquad \mathbf{\Sigma}_n = \mathbb{V}[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\mathbf{v}}_{i,n}U_{i,n}|\mathcal{X}_n, \mathcal{W}_n],
\end{aligned} \tag{C.2}$$

where $\mathbf{v}_{i,n} = \mathbf{x}_{i,n} - (\sum_{j=1}^{n}\mathbb{E}[\mathbf{x}_{j,n}\mathbf{w}_{j,n}'])(\sum_{j=1}^{n}\mathbb{E}[\mathbf{w}_{j,n}\mathbf{w}_{j,n}'])^{-1}\mathbf{w}_{i,n}$ is the population counterpart of $\hat{\mathbf{v}}_{i,n}$. Also, letting $\lambda_{\text{min}}(\cdot)$ denote the minimum eigenvalue of its argument, define

$$\mathcal{C}_n = \max_{1\leq i\leq n}\{\mathbb{E}[U_{i,n}^4|\mathcal{X}_n, \mathcal{W}_n] + \mathbb{E}[\|\mathbf{V}_{i,n}\|^4|\mathcal{W}_n] + 1/\mathbb{E}[U_{i,n}^2|\mathcal{X}_n, \mathcal{W}_n]\} + 1/\lambda_{\text{min}}(\mathbb{E}[\tilde{\mathbf{\Gamma}}_n|\mathcal{W}_n]), \tag{C.3}$$

where $U_{i,n} = y_{i.n} - \mathbb{E}[y_{i,n}|\mathcal{X}_n, \mathcal{W}_n]$, $\mathbf{V}_{i,n} = \mathbf{x}_{i,n} - \mathbb{E}[\mathbf{x}_{i,n}|\mathcal{W}_n]$, $\tilde{\mathbf{\Gamma}}_n = \sum_{i=1}^{n}\tilde{\mathbf{V}}_{i,n}\tilde{\mathbf{V}}_{i,n}'/n$ and $\tilde{\mathbf{V}}_{i,n} = \sum_{j=1}^{n}M_{ij,n}\mathbf{V}_{i,n}$.

We impose the following three assumptions:

Assumption C.1.1. $\max_{1\leq g\leq G_n}\#\mathcal{T}_{g,n} = O(1)$, *where* $\#\mathcal{T}_{g,n}$ *is the cardinality of* $\mathcal{T}_{g,n}$ *and where* $\{\mathcal{T}_{g,n} : 1 \leq g \leq G_n\}$ *is a partition of* $\{1,\ldots,n\}$ *such that* $\{(U_{i,n}, \mathbf{x}_{i,n}') : i \in \mathcal{T}_{g,n}\}$ *are independent over g conditional on* $\mathcal{W}_n$.

Assumption C.1.2. $\mathbb{P}[\lambda_{min}(\sum_{i=1}^{n}\mathbf{w}_{i,n}\mathbf{w}_{i,n}') > 0] \to 1$, $\limsup_{n\to\infty}K_n/n < 1$, $\mathcal{C}_n = O_p(1)$ *and* $\mathbf{\Sigma}_n^{-1} = O_p(1)$

Assumption C.1.3. $\chi_n = O(1)$, $\varrho_n + n(\varrho_n - \rho_n) + n\chi_n\varrho_n = o(1)$, *and* $\max_{1\leq i\leq n}\|\hat{\mathbf{v}}_{i,n}\|/\sqrt{n} = o_p(1)$.

The only difference with the simplified set of assumptions presented in Section 4.2 of this paper is that we now allow for misspecification bias, i.e. $\mathbb{E}[u_i|\mathcal{X}_n, \mathcal{W}_n] \neq$

0. In particular, Assumption C.1.3 now also includes conditions on ϱ_n and ρ_n, which are requirements on the quality of the linear approximation for the conditional expectations $\mathbb{E}[y_{i,n}|\mathcal{X}_n, \mathcal{W}_n]$ and $\mathbb{E}[y_{i,n}|\mathcal{W}_n]$, respectively. The misspecification bias is required to vanish asymptotically, thus ruling out the presence of lagged outcomes in the model. Notice that when no misspecification bias is present one gets $\varrho_n = \rho_n = 0$ and this set of assumptions reduces to the one presented in Section 4.2.

C.1.2 Variance estimators

Let $\mathbf{\Omega}_{U,n} = \mathbb{E}[\mathbf{U}_n\mathbf{U}_n'|\mathcal{X}_n, \mathcal{W}_n]$ be the (conditional) variance-covariance matrix of the errors $\mathbf{U}_n = (U_{1,n}, \ldots, U_{n,n})'$ and $L_n = \sum_{g=1}^{G_n}(\#\mathcal{T}_{g,n})^2$ the number of non-zero elements contained in it. We define a general class of cluster-robust estimators for $\mathbf{\Sigma}_n$ of the form

$$\hat{\mathbf{\Sigma}}_n(\boldsymbol{\kappa}_n) = \frac{1}{n}\sum_{g_1=1}^{G_n}\sum_{g_2=1}^{G_n}\sum_{i_1,j_1\in\mathcal{T}_{g_1,n}}\sum_{i_2,j_2\in\mathcal{T}_{g_2,n}} \kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}\hat{\mathbf{v}}_{i_1,n}\hat{\mathbf{v}}_{j_1,n}'\hat{u}_{i_2,n}\hat{u}_{j_2,n}, \tag{C.4}$$

where $\kappa_{g_1,g_2,i_1,j_1,i_2.j_2,n}$ is an entry of the $L_n \times L_n$ symmetric matrix $\boldsymbol{\kappa}_n = \boldsymbol{\kappa}_n(\mathbf{w}_{1,n}, \ldots, \mathbf{w}_{1,n})$.[1]

Furthermore, define

$$\boldsymbol{\kappa}_n^{\mathrm{CR}} = (\mathbf{S}_n'(\mathbf{M}_n \otimes \mathbf{M}_n)\mathbf{S}_n)^{-1}, \tag{C.5}$$

where $\otimes$ denotes the Kronecker product and $\mathbf{S}_n$ is the $n^2 \times L_n$ selection matrix with full column rank such that $\mathbf{S}_n'\mathrm{vec}(\mathbf{\Omega}_{U,n})$ is the $L_n \times 1$ vector containing the non-zero elements of $\mathbf{\Omega}_{U,n}$. Our proposed cluster-robust estimator is then defined as

$$\hat{\mathbf{\Sigma}}_n^{\mathrm{CR}} \equiv \hat{\mathbf{\Sigma}}(\boldsymbol{\kappa}_n^{\mathrm{CR}}) = \frac{1}{n}\sum_{g_1=1}^{G_n}\sum_{g_2=1}^{G_n}\sum_{i_1,j_1\in\mathcal{T}_{g_1,n}}\sum_{i_2,j_2\in\mathcal{T}_{g_2,n}} \kappa^{\mathrm{CR}}_{g_1,g_2,i_1,j_1,i_2,j_2,n}\hat{\mathbf{v}}_{i_1,n}\hat{\mathbf{v}}_{j_1,n}'\hat{u}_{i_2,n}\hat{u}_{j_2,n}. \tag{C.6}$$

[1] In particular, $\kappa_{g_1,g_2,i_1,j_1,i_2.j_2,n}$ corresponds to the $(h(g_1,i_1,j_1), h(g_2,i_2,j_2))$ entry of $\boldsymbol{\kappa}_n$, where $h(g,i,j) = [\sum_{k=0}^{(g-1)}(\#\mathcal{T}_{k,n})^2 + (\#\mathcal{T}_{g,n})(i-1) + j]$ and we adopt the convention that $\#\mathcal{T}_{0,n} = 0$.

C.2 Main results - general case

In this section we present the generalisation of the main results of the chapter to the case of potential misspecification bias in the model.

The first theorem establishes asymptotic normality of the OLS estimator for $\boldsymbol{\beta}_n$.

Theorem C.2.1. *Suppose Assumptions C.1.1-C.1.3 hold. Then,*

$$\boldsymbol{\Omega}_n^{-1/2}\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_d), \qquad \boldsymbol{\Omega}_n = \hat{\boldsymbol{\Gamma}}_n^{-1}\boldsymbol{\Sigma}_n\hat{\boldsymbol{\Gamma}}_n^{-1},$$

where $\boldsymbol{\Sigma}_n = \frac{1}{n}\sum_{g=1}^{G_n}\sum_{i,j\in\mathcal{T}_{g,n}} \hat{\mathbf{v}}_{i,n}\hat{\mathbf{v}}_{j,n}'\mathbb{E}[U_{i,n}U_{j,n}|\mathcal{X}_n, \mathcal{W}_n]$.

The second theorem provides an asymptotic representation for the general class of variance estimators defined in (C.4).

Theorem C.2.2. *Suppose Assumptions C.1.1-C.1.3 hold.*
If $\|\boldsymbol{\kappa}_n\|_\infty = \max_{(g_1,i_1,j_1)}\sum_{g_2=1}^{G_n}\sum_{i_2,j_2\in\mathcal{T}_{g_2,n}}|\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}| = O_p(1)$, *then*

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{\kappa}_n) = \frac{1}{n}\sum_{g_1=1}^{G_n}\sum_{g_2=1}^{G_n}\sum_{i_1,j_1\in\mathcal{T}_{g_1,n}}\sum_{i_2,j_2\in\mathcal{T}_{g_2,n}} \kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}\hat{\mathbf{v}}_{i_1,n}\hat{\mathbf{v}}_{j_1,n}' \\ \times\sum_{g_3=1}^{G_n}\sum_{i_3,j_3\in\mathcal{T}_{g_3,n}} M_{i_2j_3,n}M_{j_2i_3,n}\mathbb{E}[U_{i_3,n}U_{j_3,n}|\mathcal{X}_n,\mathcal{W}_n] + o_p(1).\end{aligned} \tag{C.7}$$

The following corollary characterizes the asymptotic limit of LZ's estimator.

Corollary C.2.1. *Suppose the assumptions of Theorem C.2.2 hold. Then,*

$$\hat{\boldsymbol{\Sigma}}_n^{\mathrm{LZ}} = \frac{1}{n}\sum_{g_1=1}^{G_n}\sum_{g_2=1}^{G_n}\sum_{i_1,j_1\in\mathcal{T}_{g_1,n}}\sum_{i_2,j_2\in\mathcal{T}_{g_2,n}} \hat{\mathbf{v}}_{i_1,n}\hat{\mathbf{v}}_{j_1,n}'M_{i_1j_2,n}M_{j_1i_2,n}\mathbb{E}[U_{i_2,n}U_{j_2,n}|\mathcal{X}_n,\mathcal{W}_n]+o_p(1).$$

The follwing theorem establishes consistency of our proposed estimator.

Theorem C.2.3. *Suppose Assumptions C.1.1-C.1.3 hold.*

If $\mathbb{P}[\lambda_{min}(\mathbf{S}_n'(\mathbf{M}_n \otimes \mathbf{M}_n)\mathbf{S}_n) > 0] \to 1$ *and* $\|\boldsymbol{\kappa}_n^{\mathrm{CR}}\|_\infty = O_p(1)$*, then*

$$\hat{\boldsymbol{\Sigma}}_n^{\mathrm{CR}} = \boldsymbol{\Sigma}_n + o_p(1).$$

Finally, the fourth theorem provides sufficient conditions for consistency of LZ's estimator. For this purpose, define $\mathbf{w}_{i,n}^* = \hat{\boldsymbol{\Sigma}}_{\mathbf{w},n}^{-1/2}\mathbf{w}_{i,n}$, where $\hat{\boldsymbol{\Sigma}}_{\mathbf{w},n}^{1/2}$ is the unique symmetric positive definite $K_n \times K_n$ matrix such that $\hat{\boldsymbol{\Sigma}}_{\mathbf{w},n}^{1/2}\hat{\boldsymbol{\Sigma}}_{\mathbf{w},n}^{1/2} = \frac{1}{n}\sum_{i=1}^n \mathbf{w}_{i,n}\mathbf{w}_{i,n}'$.

Theorem C.2.4. *Suppose Assumptions C.1.1-C.1.3 hold and that* $\max_{i,j}\mathbb{E}[w_{ij,n}^{*2}] = O(1)$*. If* $K_n^2/n \to 0$*, then*

$$\hat{\boldsymbol{\Sigma}}_n^{\mathrm{LZ}} = \boldsymbol{\Sigma}_n + o_p(1). \tag{C.8}$$

Moreover, if $\mathbb{E}[U_{i,n}^2|\mathcal{X}_n, \mathcal{W}_n] = \sigma_n^2\ \forall i$*, and* $\mathbb{E}[U_{i,n}U_{j,n}|\mathcal{X}_n, \mathcal{W}_n] = 0\ \forall i \neq j$*, then (C.8) holds under* $K_n/n \to 0$*.*

C.3 Technical Lemmas

Here we present the technical lemmas needed to establish the main results of the chapter.[2]

The first lemma can be used to approximate $\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{\kappa}_n)$ by means of $\tilde{\boldsymbol{\Sigma}}_n(\boldsymbol{\kappa}_n)$, where

$$\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{\kappa}_n) = \frac{1}{n}\sum_{(g_1,i_1,j_1)}\sum_{(g_2,i_2,j_2)} \kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}\hat{\mathbf{v}}_{i_1,n}\hat{\mathbf{v}}_{j_1,n}'\hat{u}_{i_2,n}\hat{u}_{j_2,n},$$

$$\tilde{\boldsymbol{\Sigma}}_n(\boldsymbol{\kappa}_n) = \frac{1}{n}\sum_{(g_1,i_1,j_1)}\sum_{(g_2,i_2,j_2)} \kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}\hat{\mathbf{v}}_{i_1,n}\hat{\mathbf{v}}_{j_1,n}'\tilde{U}_{i_2,n}\tilde{U}_{j_2,n}, \qquad \tilde{U}_{i,n} = \sum_{j=1}^n M_{ij,n}U_{j,n}$$

Lemma C.3.1. *Suppose Assumptions C.1.1-C.1.3 hold. If* $\|\boldsymbol{\kappa}_n\|_\infty = O_p(1)$*, then*

$$\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{\kappa}_n) = \mathbb{E}[\tilde{\boldsymbol{\Sigma}}_n(\boldsymbol{\kappa}_n)|\mathcal{X}_n, \mathcal{W}_n] + o_p(1).$$

[2]Throughout the Technical Lemmas we adopt the notational convention $\sum_{(g,i,j)} \equiv \sum_{g=1}^{G_n}\sum_{i,j\in\mathcal{T}_{g,n}}$

The second lemma can be combined with Lemma 1 to show consistency of $\hat{\mathbf{\Sigma}}_n(\boldsymbol{\kappa}_n)$ under a high-level condition.

Lemma C.3.2. *Suppose Assumption C.1.2 holds. If*

$$\max_{(g_1,i_1,j_1)} \Big\{ \Big| \sum_{(g_2,i_2,j_2)} \kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n} M_{i_1 j_2,n} M_{j_1 i_2,n} - 1 \Big| + \sum_{(g_3,i_3,j_3)\neq(g_1,i_1,j_1)} \Big| \sum_{(g_2,i_2,j_2)} \kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n} M_{i_3,j_2,n} M_{j_3,i_2,n} \Big| \Big\} = o_p(1),$$

then $\mathbb{E}[\tilde{\mathbf{\Sigma}}_n(\boldsymbol{\kappa}_n)|\mathcal{X}_n,\mathcal{W}_n] = \mathbf{\Sigma}_n + o_p(1)$.

The third lemma gives sufficient conditions for the condition of Lemma 2 for our proposed estimator $\hat{\mathbf{\Sigma}}_n^{\text{CR}}$.

Lemma C.3.3. *Suppose Assumption C.1.2 holds. If* $\mathbb{P}[\lambda_{min}(\mathbf{S}_n'(\mathbf{M}_n \otimes \mathbf{M}_n)\mathbf{S}_n) > 0] \to 1$, *then*

$$\mathbb{E}[\tilde{\mathbf{\Sigma}}_n(\boldsymbol{\kappa}_n^{\text{CR}})|\mathcal{X}_n,\mathcal{W}_n] = \mathbf{\Sigma}_n + o_p(1).$$

with $\boldsymbol{\kappa}_n^{\text{CR}} = (\mathbf{S}_n'(\mathbf{M}_n \otimes \mathbf{M}_n)\mathbf{S}_n)^{-1}$.

The fourth lemma finds sufficient conditions for the condition of Lemma 2 for LZ's estimator.

Lemma C.3.4. *Suppose Assumption C.1.2 holds and* $\boldsymbol{\kappa}_n = \mathbf{I}_{L_n}$. *Also define* $\mathbf{w}_{i,n}^* = \hat{\mathbf{\Sigma}}_{\mathbf{w},n}^{-1/2}\mathbf{w}_{i,n}$, *where* $\hat{\mathbf{\Sigma}}_{\mathbf{w},n}^{1/2}$ *is the unique symmetric positive definite* $K_n \times K_n$ *matrix such that* $\hat{\mathbf{\Sigma}}_{\mathbf{w},n}^{1/2}\hat{\mathbf{\Sigma}}_{\mathbf{w},n}^{1/2} = \frac{1}{n}\sum_{i=1}^n \mathbf{w}_n\mathbf{w}_n'$. *If* $\max_{i,j}\mathbb{E}[w_{ij,n}^{*2}] = O(1)$ *and* $K_n = o(n^{1/2})$, *then*

$$\mathbb{E}[\tilde{\mathbf{\Sigma}}_n(\boldsymbol{\kappa}_n)|\mathcal{X}_n,\mathcal{W}_n] = \mathbf{\Sigma}_n + o_p(1).$$

Finally, the fifth lemma establishes sufficient conditions for the condition of Lemma 2 for LZ's estimator for the special case of homoskedastic errors.

Lemma C.3.5. *Suppose Assumption C.1.2 holds and* $\boldsymbol{\kappa}_n = \mathbf{I}_{L_n}$. *If* $\max_{i,j}\mathbb{E}[w_{ij,n}^{*2}] = O(1)$, $K_n = o(n)$, $\mathbb{E}[U_{i,n}^2|\mathcal{X}_n,\mathcal{W}_n] = \sigma_n^2 \quad \forall i$ *and* $\mathbb{E}[U_{i,n}U_{j,n}|\mathcal{X}_n,\mathcal{W}_n] = 0 \quad \forall i \neq j$,

then

$$\mathbb{E}[\tilde{\mathbf{\Sigma}}_n(\boldsymbol{\kappa}_n)|\mathcal{X}_n, \mathcal{W}_n] = \mathbf{\Sigma}_n + o_p(1).$$

C.4 Proof of Main Results

Theorem C.2.1 follows from Lemma SA-1 and Lemma SA-2 in Cattaneo et al. (2018b), combined with the fact that $\mathbf{\Sigma}_n^{-1} = O_p(1)$ in Assumption C.1.2. Theorem C.2.2 follows from Theorem C.2.1 combined with Lemma C.3.1. Theorem C.2.3 follows from Theorem C.2.2 combined with Lemma C.3.2 and C.3.3. Theorem C.2.4 follows from Theorem C.2.2 combined with Lemma C.3.3, C.3.4 and C.3.5.

C.5 Proofs of Technical Lemmas

Here we provide the proofs for the technical lemmas. To simplify notation, throughout the proofs we assume $d = 1$ without loss of generality.

C.5.1 Proof of Lemma C.3.1

It suffices to show that $\hat{\Sigma}_n(\boldsymbol{\kappa}_n) = \tilde{\Sigma}_n(\boldsymbol{\kappa}_n) + o_p(1)$ and that $\tilde{\Sigma}_n(\boldsymbol{\kappa}_n) = \mathbb{E}[\tilde{\Sigma}_n(\boldsymbol{\kappa}_n)|\mathcal{X}_n, \mathcal{W}_n] + o_p(1)$.

First,

$$\tilde{\Sigma}_n(\boldsymbol{\kappa}_n) = \frac{1}{n}\sum_{1\leq i\leq G_n} c_{ii,n} + \frac{2}{n}\sum_{1\leq i,j\leq G_n, i<j} c_{ij,n},$$

$$c_{ij,n} = \sum_{s\in\mathcal{T}_i, t\in\mathcal{T}_j}\sum_{(g_1,i_1,j_1)}\sum_{(g_2,i_2,j_2)} \kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}\hat{v}_{i_1,n}\hat{v}_{j_1,n}M_{i_2s,n}M_{j_2t,n}U_{s,n}U_{t,n},$$

where $\sum_{1\leq i,j\leq G_n}\mathbb{V}[c_{ij,n}|\mathcal{X}_n\mathcal{W}_n]=o_p(n^2)$ because

$$
\begin{aligned}
&\mathbb{V}[c_{ij,n}|\mathcal{X}_n,\mathcal{W}_n]\\
&\leq (\#\mathcal{T}_{i,n})(\#\mathcal{T}_{j,n})\sum_{s\in\mathcal{T}_{i,n},t\in\mathcal{T}_{j,n}}\Big(\sum_{(g_1,i_1,j_1)}\sum_{(g_2,i_2,j_2)}\kappa_{g_1,g_2,i_1,j_1,i_2,j_2}\hat{v}_{i_1,n}\hat{v}_{j_1,n}M_{i_2s,n}M_{j_2t,n}\Big)^2\\
&\qquad\qquad\times\mathbb{V}[U_{s,n}U_{t,n}|\mathcal{X}_n\mathcal{W}_n]\\
&\leq \mathcal{C}_{\mathcal{T},n}^2\mathcal{C}_{U,n}\sum_{s\in\mathcal{T}_{i,n},t\in\mathcal{T}_{j,n}}\Big(\sum_{(g_1,i_1,j_1)}\sum_{(g_2,i_2,j_2)}\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}\hat{v}_{i_1,n}\hat{v}_{j_1,n}M_{i_2s,n}M_{j_2t,n}\Big)^2\\
&\leq \mathcal{C}_{\mathcal{T},n}^2\mathcal{C}_{U,n}\sum_{1\leq s,t\leq n}\Big(\sum_{(g_1,i_1,j_1)}\sum_{(g_2,i_2,j_2)}\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}\hat{v}_{i_1,n}\hat{v}_{j_1,n}M_{i_2s,n}M_{j_2t,n}\Big)^2\\
&= \mathcal{C}_{\mathcal{T},n}^2\mathcal{C}_{U,n}\sum_{1\leq s,t\leq n}\sum_{(g_1,i_1,j_1)}\sum_{(g_2,i_2,j_2)}\sum_{(g_3,i_3,j_3)}\sum_{(g_4,i_4,j_4)}\\
&\qquad\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}\kappa_{g_3,g_4,i_3,j_3,i_4,j_4,n}\hat{v}_{i_1,n}\hat{v}_{j_1,n}\hat{v}_{i_3,n}\hat{v}_{j_3,n}M_{i_2s,n}M_{j_2t,n}M_{i_4s,n}M_{j_4t,n}\\
&= \mathcal{C}_{\mathcal{T},n}^2\mathcal{C}_{U,n}\sum_{(g_1,i_1,j_1)}\sum_{(g_2,i_2,j_2)}\sum_{(g_3,i_3,j_3)}\sum_{(g_4,i_4,j_4)}\\
&\qquad\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}\kappa_{g_3,g_4,i_3,j_3,i_4,j_4,n}\hat{v}_{i_1,n}\hat{v}_{j_1,n}\hat{v}_{i_3,n}\hat{v}_{j_3,n}M_{i_2i_4,n}M_{j_2j_4,n}\\
&\leq \mathcal{C}_{\mathcal{T},n}^2\mathcal{C}_{U,n}\sum_{(g_1,i_1,j_1)}\sum_{(g_2,i_2,j_2)}\sum_{(g_3,i_3,j_3)}\sum_{(g_4,i_4,j_4)}\\
&\qquad|\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}||\kappa_{g_3,g_4,i_3,j_3,i_4,j_4,n}||\hat{v}_{i_1,n}||\hat{v}_{j_1,n}||\hat{v}_{i_3,n}||\hat{v}_{j_3,n}||M_{i_2i_4,n}||M_{j_2j_4,n}|,
\end{aligned}
$$

where $\mathcal{C}_{\mathcal{T},n} = \max_{1\leq i\leq G_n} \#(\mathcal{T}_{i,n})$, $\mathcal{C}_{U,n} = 1 + \max_{1\leq i\leq n} \mathbb{E}[U_{i,n}^4|\mathcal{X}_n, \mathcal{W}_n]$, and

$$
\begin{aligned}
&\sum_{(g_1,i_1,j_1)} \sum_{(g_2,i_2,j_2)} \sum_{(g_3,i_3,j_3)} \sum_{(g_4,i_4,j_4)} \\
&\qquad |\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}||\kappa_{g_3,g_4,i_3,j_3,i_4,j_4,n}||\hat{v}_{i_1,n}||\hat{v}_{j_1,n}||\hat{v}_{i_3,n}||\hat{v}_{j_3,n}||M_{i_2i_4,n}||M_{j_2j_4,n}| \\
&\leq (\max_{1\leq i\leq n} |\hat{v}_{i,n}|)^2 \sum_{(g_1,i_1,j_1)} \sum_{(g_2,i_2,j_2)} \sum_{(g_3,i_3,j_3)} \sum_{(g_4,i_4,j_4)} \\
&\qquad |\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}||\kappa_{g_3,g_4,i_3,j_3,i_4,j_4,n}||\hat{v}_{i_3,n}||\hat{v}_{j_3,n}||M_{i_2i_4,n}||M_{j_2j_4,n}| \\
&\leq (\max_{1\leq i\leq n} |\hat{v}_{i,n}|)^2 \|\boldsymbol{\kappa}_n\|_\infty \sum_{(g_2,i_2,j_2)} \sum_{(g_3,i_3,j_3)} \sum_{(g_4,i_4,j_4)} |\kappa_{g_3,g_4,i_3,j_3,i_4,j_4,n}||\hat{v}_{i_3,n}||\hat{v}_{j_3,n}||M_{i_2i_4,n}||M_{j_2j_4,n}| \\
&\leq (\max_{1\leq i\leq n} |\hat{v}_{i,n}|)^2 \|\boldsymbol{\kappa}_n\|_\infty \mathcal{C}_{\mathcal{T},n} \sum_{(g_3,i_3,j_3)} \sum_{(g_4,i_4,j_4)} |\kappa_{g_3,g_4,i_3,j_3,i_4,j_4,n}||\hat{v}_{i_3,n}||\hat{v}_{j_3,n}| \\
&\leq (\max_{1\leq i\leq n} |\hat{v}_{i,n}|)^2 \|\boldsymbol{\kappa}_n\|_\infty^2 \mathcal{C}_{\mathcal{T},n} \sum_{(g_3,i_3,j_3)} |\hat{v}_{i_3,n}||\hat{v}_{j_3,n}| \\
&\leq n^2 (\frac{\max_{1\leq i\leq n} |\hat{v}_i|}{\sqrt{n}})^2 \|\boldsymbol{\kappa}_n\|_\infty^2 \mathcal{C}_{\mathcal{T},n}^2 (\frac{1}{n} \sum_{1\leq i\leq n} \hat{v}_{i,n}^2) = o_p(n^2),
\end{aligned}
$$

where the third inequality uses

$$
\begin{aligned}
\sum_{g_2=1}^{G_n} \sum_{i_2,j_2\in\mathcal{T}_{g_2,n}} |M_{i_2i_4,n}||M_{j_2j_4,n}| &\leq \sqrt{(\sum_{g_2=1}^{G_n} \sum_{i_2,j_2)\in\mathcal{T}_{g_2,n}} M_{i_2i_4,n}^2)(\sum_{g_2=1}^{G_n} \sum_{(i_2,j_2)\in\mathcal{V}_{g_2,n}} M_{j_2j_4,n}^2)} \\
&\leq \sqrt{\big(\mathcal{C}_{\mathcal{T},n} \sum_{k=1}^{n} M_{ki_4,n}^2\big)\big(\mathcal{C}_{\mathcal{T},n} \sum_{l=1}^{n} M_{lj_4,n}^2\big)} \\
&= \mathcal{C}_{\mathcal{T},n} \sqrt{M_{i_4i_4,n} M_{j_4j_4,n}} \\
&\leq \mathcal{C}_{\mathcal{T},n},
\end{aligned}
$$

and the last inequality similarly uses[3]

$$\begin{aligned}\sum_{g_3=1}^{G}\sum_{i_3,j_3\in\mathcal{T}_{g_2,n}}|\hat{v}_{i_3,n}||\hat{v}_{j_3,n}| &\leq \sqrt{\big(\sum_{g_3=1}^{G}\sum_{i_3,j_3\in\mathcal{T}_{g_3,n}}\hat{v}_{i_3,n}^2\big)\big(\sum_{g_3=1}^{G}\sum_{i_3,j_3\in\mathcal{T}_{g_3,n}}\hat{v}_{j_3,n}^2\big)}\\ &\leq \sqrt{\big(\mathcal{C}_{\mathcal{T},n}\sum_{k=1}^{n}\hat{v}_{k,n}^2\big)\big(\mathcal{C}_{\mathcal{T},n}\sum_{l=1}^{n}\hat{v}_{l,n}^2\big)}\\ &= \mathcal{C}_{\mathcal{T},n}(\sum_{i=1}^{n}\hat{v}_{i,n}^2).\end{aligned}$$

As a consequence,

$$\begin{aligned}\mathbb{V}[\frac{1}{n}\sum_{1\leq i\leq G_n}c_{ii,n}|\mathcal{X}_n,\mathcal{W}_n] &= \frac{1}{n^2}\sum_{1\leq i\leq G_n}\mathbb{V}[c_{ii,n}|\mathcal{X}_n,\mathcal{W}_n]\\ &\leq \frac{1}{n^2}\sum_{1\leq i,j\leq G_n}\mathbb{V}[c_{ij,n}|\mathcal{X}_n,\mathcal{W}_n]\\ &= o_p(1),\end{aligned}$$

and

$$\begin{aligned}\mathbb{V}[\frac{1}{n}\sum_{1\leq i,j\leq G_n,i<j}c_{ij,n}|\mathcal{X}_n,\mathcal{W}_n] &= \frac{1}{n^2}\sum_{1\leq i,j\leq G_n,i<j}\mathbb{V}[c_{ij,n}|\mathcal{X}_n,\mathcal{W}_n]\\ &\leq \frac{1}{n^2}\sum_{1\leq i,j\leq G_n}\mathbb{V}[c_{ij,n}|\mathcal{X}_n,\mathcal{W}_n]\\ &= o_p(1).\end{aligned}$$

[3]We also make use of the bound $\frac{1}{n}\sum_{i=1}^{n}\hat{v}_i^2 = O_p(1)$, as shown in Lemma SA-1 of Cattaneo et al. (2018b, Supplemental Appendix).

In particular, $\tilde{\Sigma}_n(\boldsymbol{\kappa}_n) = \mathbb{E}[\tilde{\Sigma}_n(\boldsymbol{\kappa}_n)|\mathcal{X}_n, \mathcal{W}_n] + o_p(1)$, where

$$
\begin{aligned}
&|\mathbb{E}[\tilde{\Sigma}_n(\boldsymbol{\kappa}_n)|\mathcal{X}_n, \mathcal{W}_n]| \\
&\leq \frac{1}{n} \sum_{(g_1,i_1,j_1)} \sum_{(g_2,i_2,j_2)} \sum_{(g_3,i_3,j_3)} |\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}||\hat{v}_{i_1,n}||\hat{v}_{j_1,n}||M_{i_2j_3,n}||M_{j_2i_3,n}||\mathbb{E}[U_{i_3,n}U_{j_3,n}|\mathcal{X}_n, \mathcal{W}_n]| \\
&\leq \mathcal{C}_{U,n} \frac{1}{n} \sum_{(g_1,i_1,j_1)} \sum_{(g_2,i_2,j_2)} \sum_{(g_3,i_3,j_3)} |\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}||\hat{v}_{i_1,n}||\hat{v}_{j_1,n}||M_{i_2j_3,n}||M_{j_2i_3,n}| \\
&\leq \mathcal{C}_{U,n}\mathcal{C}_{\mathcal{T},n} \frac{1}{n} \sum_{(g_1,i_1,j_1)} \sum_{(g_2,i_2,j_2)} |\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}||\hat{v}_{i_1,n}||\hat{v}_{j_1,n}| \\
&\leq \mathcal{C}_{U,n}\mathcal{C}_{\mathcal{T},n} \|\kappa_n\|_\infty \frac{1}{n} \sum_{(g_1,i_1,j_1)} |\hat{v}_{i_1,n}||\hat{v}_{j_1,n}| \\
&\leq \mathcal{C}_{U,n}\mathcal{C}^2_{\mathcal{T},n} \|\kappa_n\|_\infty (\frac{1}{n} \sum_{1\leq i\leq n} \hat{v}^2_{i,n}) \\
&= O_p(1).
\end{aligned}
$$

We have therefore established that $\tilde{\Sigma}_n(\boldsymbol{\kappa}_n) = \mathbb{E}[\tilde{\Sigma}_n(\boldsymbol{\kappa}_n)|\mathcal{X}_n, \mathcal{W}_n] + o_p(1)$. It remains to show that $\hat{\Sigma}_n(\boldsymbol{\kappa}_n) = \tilde{\Sigma}_n(\boldsymbol{\kappa}_n) + o_p(1)$.
By using that $\hat{u}_{i,n} - \tilde{U}_{i,n} = \tilde{R}_{i,n} - \hat{v}_{i,n}(\hat{\beta}_n - \beta)$, where $\tilde{R}_{i,n} = \sum_{j=1}^n M_{ij,n}R_{j,n}$, we obtain

$$
\begin{aligned}
&\hat{\Sigma}_n(\boldsymbol{\kappa}_n) - \tilde{\Sigma}_n(\boldsymbol{\kappa}_n) = \frac{1}{n} \sum_{(g_1,i_1,j_1)} \sum_{(g_2,i_2,j_2)} \kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}\hat{v}_{i_1,n}\hat{v}_{j_1,n}(\hat{u}_{i_2,n}\hat{u}_{j_2,n} - \tilde{U}_{i_2,n}\tilde{U}_{j_2,n}) \\
&= \frac{1}{n} \sum_{(g_1,i_1,j_1)} \sum_{(g_2,i_2,j_2)} \kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}\hat{v}_{i_1,n}\hat{v}_{j_1,n} \\
&\qquad \times [(\tilde{R}_{i_2,n} - \hat{v}_{i_2,n}(\hat{\beta}_n - \beta) + \tilde{U}_{i_2,n})(\tilde{R}_{j_2,n} - \hat{v}_{j_2,n}(\hat{\beta}_n - \beta) + \tilde{U}_{j_2,n}) - \tilde{U}_{i_2,n}\tilde{U}_{j_2,n}].
\end{aligned}
$$

By the Cauchy-Schwarz inequality, it suffices to show that

$$
\begin{aligned}
&\frac{1}{n} \sum_{(g_1,i_1,j_1)} \sum_{(g_2,i_2,j_2)} \kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}\hat{v}^2_{i_1,n}(\tilde{R}_{i_2,n} - \hat{v}_{i_2,n}(\hat{\beta}_n - \beta))^2 = o_p(1), \\
&\frac{1}{n} \sum_{(g_1,i_1,j_1)} \sum_{(g_2,i_2,j_2)} \kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}\hat{v}^2_{i_1,n}\tilde{U}^2_{j_2,n} = O_p(1).
\end{aligned}
$$

The latter can be straightforwardly shown by means of the arguments previously

used to show $\tilde{\Sigma}_n(\boldsymbol{\kappa}_n) = O_p(1)$. For the former, since $\hat{v}_{j,n} = \tilde{V}_{j,n} + \tilde{Q}_{j,n}$, where $\tilde{Q}_{i,n} = \sum_{j=1}^n M_{ij,n} Q_{j,n}$, and $\tilde{R}_{i,n} = \tilde{r}_{i,n} + (\tilde{R}_{i,n} - \tilde{r}_{i,n})$, where $\tilde{r}_{i,n} = \sum_{j=1}^n M_{ij,n} r_{j,n}$ it suffices to show that

$$\frac{1}{n} \sum_{(g_1,i_1,j_1)} \sum_{(g_2,i_2,j_2)} |\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}| \tilde{Q}_{i_1,n}^2 \tilde{R}_{i_2,n}^2 = o_p(1),$$

$$\frac{1}{n} \sum_{(g_1,i_1,j_1)} \sum_{(g_2,i_2,j_2)} |\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}| \tilde{V}_{i_1,n}^2 \tilde{r}_{i_2,n}^2 = o_p(1),$$

$$\frac{1}{n} \sum_{(g_1,i_1,j_1)} \sum_{(g_2,i_2,j_2)} |\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}| \tilde{V}_{i_1,n}^2 |\tilde{R}_{i_2,n} - \tilde{r}_{i_2,n}|^2 = o_p(1),$$

$$(\hat{\beta}_n - \beta)^2 \frac{1}{n} \sum_{(g_1,i_1,j_1)} \sum_{(g_2,i_2,j_2)} |\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}| \hat{v}_{i_1,n}^2 \hat{v}_{i_2,n}^2 = o_p(1).$$

First, $n^{-1} \sum_{(g_1,i_1,j_1)} \sum_{(g_2,i_2,j_2)} |\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}| \tilde{V}_{i_1,n}^2 \tilde{r}_{i_2,n}^2 = o_p(1)$ because

$$\begin{aligned}
&\mathbb{E}\big[\frac{1}{n} \sum_{(g_1,i_1,j_1)} \sum_{(g_2,i_2,j_2)} |\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}| \tilde{V}_{i_1,n}^2 \tilde{r}_{i_2,n}^2 | \mathcal{W}_n\big] \\
&= \frac{1}{n} \sum_{(g_2,i_2,j_2)} \tilde{r}_{i_2,n}^2 \sum_{(g_1,i_1,j_1)} |\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}| \mathbb{E}\big[\tilde{V}_{i_1,n}^2 | \mathcal{W}_n\big] \\
&\leq \mathcal{C}_{V,n} \mathcal{C}_{\mathcal{T},n} \frac{1}{n} \sum_{(g_2,i_2,j_2)} \tilde{r}_{i_2,n}^2 \sum_{(g_1,i_1,j_1)} |\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}| \\
&\leq \mathcal{C}_{V,n} \mathcal{C}_{\mathcal{T},n} \|\boldsymbol{\kappa}_n\|_\infty (\frac{1}{n} \sum_{(g_2,i_2,j_2)} \tilde{r}_{i_2,n}^2) \\
&\leq \mathcal{C}_{V,n} \mathcal{C}_{\mathcal{T},n}^2 \|\boldsymbol{\kappa}_n\|_\infty (\frac{1}{n} \sum_{i=1}^n \tilde{r}_{i,n}^2) \\
&= O_p(\rho_n) = o_p(1),
\end{aligned}$$

where the first inequality uses the fact that $\mathbb{E}[\tilde{V}_{i,n} | \mathcal{W}_n] \leq \mathcal{C}_{\mathcal{T},n} \mathcal{C}_{V,n}$, with $\mathcal{C}_{V,n} = 1 + \max_{1 \leq i \leq n}, \mathbb{E}[\|V_{i,n}\|^4 | \mathcal{W}_n]$ as shown in Cattaneo et al. (2018b, Supplemental

Appendix). Next,

$$\begin{aligned}
&\frac{1}{n}\sum_{(g_1,i_1,j_1)}\sum_{(g_2,i_2,j_2)}|\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}|\tilde{V}_{i_1,n}^2|\tilde{R}_{i_2,n}-\tilde{r}_{i_2,n}|^2\\
&\leq n\|\boldsymbol{\kappa}_n\|_\infty(\frac{1}{n}\sum_{(g_1,i_1,j_1)}\tilde{V}_{i_1,n}^2)(\frac{1}{n}\sum_{(g_2,i_2,j_2)}|\tilde{R}_{i_2,n}-\tilde{r}_{i_2,n}|^2)\\
&\leq n\|\boldsymbol{\kappa}_n\|_\infty\mathcal{C}_{\mathcal{T},n}^2(\frac{1}{n}\sum_{i=1}^n\tilde{V}_{i,n}^2)(\frac{1}{n}\sum_{i=1}^n|\tilde{R}_{i,n}-\tilde{r}_{i,n}|^2)\\
&=O_p[n(\varrho_n-\rho_n)]=o_p(1)
\end{aligned}$$

and

$$\begin{aligned}
\frac{1}{n}\sum_{(g_1,i_1,j_1)}\sum_{(g_2,i_2,j_2)}|\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}|\tilde{Q}_{i_1,n}^2\tilde{R}_{i_2,n}^2 &\leq n\|\boldsymbol{\kappa}_n\|_\infty(\frac{1}{n}\sum_{(g_1,i_1,j_1)}\tilde{Q}_{i_1,n}^2)(\frac{1}{n}\sum_{(g_2,i_2,j_2)}\tilde{R}_{i_2,n}^2)\\
&\leq n\|\boldsymbol{\kappa}_n\|_\infty\mathcal{C}_{\mathcal{T},n}^2(\frac{1}{n}\sum_{i=1}^n\tilde{Q}_{i,n}^2)(\frac{1}{n}\sum_{i=1}^n\tilde{R}_{i,n}^2)\\
&=O_p(n\chi_n\varrho_n)=o_p(1)
\end{aligned}$$

Finally,

$$(\hat{\beta}_n-\beta)^2\frac{1}{n}\sum_{(g_1,i_1,j_1)}\sum_{(g_2,i_2,j_2)}|\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}|\hat{v}_{i_1,n}^2\hat{v}_{i_2,n}^2=o_p(1)$$

because$\sqrt{n}(\hat{\beta}_n-\beta)=O_p(1)$ and

$$\begin{aligned}
&\frac{1}{n^2}\sum_{(g_1,i_1,j_1)}\sum_{(g_2,i_2,j_2)}|\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}|\hat{v}_{i_1,n}^2\hat{v}_{i_2,n}^2\\
&\leq(\max_{1\leq i\leq n}|\hat{v}_{i,n}|)^2\frac{1}{n^2}\sum_{(g_1,i_1,j_1)}\sum_{(g_2,i_2,j_2)}|\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}|\hat{v}_{i_2,n}^2\\
&\leq(\frac{\max_{1\leq i\leq n}|\hat{v}_{i,n}|}{\sqrt{n}})^2\|\boldsymbol{\kappa}_n\|_\infty(\frac{1}{n}\sum_{(g_2,i_2,j_2)}\hat{v}_{i_2,n}^2)\\
&\leq(\frac{\max_{1\leq i\leq n}|\hat{v}_{i,n}|}{\sqrt{n}})^2\|\boldsymbol{\kappa}_n\|_\infty\mathcal{C}_{\mathcal{T},n}(\frac{1}{n}\sum_{i=1}^n\hat{v}_{i,n}^2)=o_p(1),
\end{aligned}$$

which concludes the proof.

C.5.2 Proof of Lemma C.3.2

Let us define

$$d_{i_1 j_1, i_3 j_3, n} = \sum_{(g_2, i_2, j_2)} \kappa_{g_1, g_2, i_1, j_1, i_2, j_2, n} M_{i_3, j_2, n} M_{j_3, i_2, n} - \mathbb{1}\{(i_1, j_1) = (i_3, j_3)\}.$$

We hence have

$$\mathbb{E}[\tilde{\Sigma}_n(\boldsymbol{\kappa}_n)|\mathcal{X}_n, \mathcal{W}_n] - \Sigma_n = \frac{1}{n} \sum_{(g_1, i_1, j_1)} \sum_{(g_3, i_3, j_3)} d_{i_1 j_1, i_3 j_3, n} \hat{v}_{i_1, n} \hat{v}_{j_1, n} \mathbb{E}[U_{i_3, n} U_{j_3, n} | \mathcal{X}_n, \mathcal{W}_n],$$

so if $\max_{(g_1, i_1, j_1)} \sum_{(g_3, i_3, j_3)} |d_{i_1 j_1, i_3 j_3, n}| = o_p(1)$, then

$$\begin{aligned}
|\mathbb{E}[\tilde{\Sigma}_n(\boldsymbol{\kappa}_n)|\mathcal{X}_n, \mathcal{W}_n] - \Sigma_n| &\leq \frac{1}{n} \sum_{(g_1, i_1, j_1)} \sum_{(g_3, i_3, j_3)} |d_{i_1 j_1, i_3 j_3, n}| |\hat{v}_{i_1, n}| |\hat{v}_{j_1, n}| |\mathbb{E}[U_{i_3, n} U_{j_3, n} | \mathcal{X}_n, \mathcal{W}_n]| \\
&\leq \mathcal{C}_{U,n} \frac{1}{n} \sum_{(g_1, i_1, j_1)} \sum_{(g_3, i_3, j_3)} |d_{i_1 j_1, i_3 j_3, n}| |\hat{v}_{i_1, n}| |\hat{v}_{j_1, n}| \\
&\leq \mathcal{C}_{U,n} \Big(\frac{1}{n} \sum_{(g_1, i_1, j_1)} |\hat{v}_{i_1, n}| |\hat{v}_{j_1, n}|\Big) \Big(\max_{(g_1, i_1, j_1)} \sum_{(g_3, i_3, j_3)} |d_{i_1 j_1, i_3 j_3, n}| \Big) \\
&\leq \mathcal{C}_{U,n} \mathcal{C}_{\mathcal{T},n} \Big(\frac{1}{n} \sum_{i=1}^{n} \hat{v}_{i,n}^2\Big) \Big(\max_{(g_1, i_1, j_1)} \sum_{(g_3, i_3, j_3)} |d_{i_1 j_1, i_3 j_3, n}| \Big) = o_p(1).
\end{aligned}$$

C.5.3 Proof of Lemma C.3.3

If $\lambda_{\min}(\mathbf{S}_n'(\mathbf{M}_n \otimes \mathbf{M}_n)\mathbf{S}_n) > 0$, then

$$\begin{aligned}
&\Big| \sum_{(g_2, i_2, j_2)} \kappa^{\mathrm{CR}}_{g_1, g_2, i_1, j_1, i_2, j_2, n} M_{i_1 j_2, n} M_{j_1 i_2, n} - 1 \Big| \\
&\qquad + \sum_{(g_3, i_3, j_3) \neq (g_1, i_1, j_1)} \Big| \sum_{(g_2, i_2, j_2)} \kappa^{\mathrm{CR}}_{g_1, g_2, i_1, j_1, i_2, j_2} M_{i_3 j_2, n} M_{j_3 i_2, n} \Big| = 0,
\end{aligned}$$

which combined with Lemma 2 gives $\mathbb{E}[\tilde{\boldsymbol{\Sigma}}_n(\boldsymbol{\kappa}_n^{\mathrm{CR}})|\mathcal{X}_n, \mathcal{W}_n] = \boldsymbol{\Sigma}_n + o_p(1)$.

C.5.4 Proof of Lemma C.3.4

Recall that for LZ's estimator we have

$$\begin{aligned}\sum_{g_2,i_2,j_2}|d_{i_1,j_1,i_2,j_2,n}| &= |M_{i_1i_1,n}M_{j_1j_1,n}-1| + \sum_{(g_2,i_2,j_2)\neq(g_1,i_1,j_1)}|M_{i_1j_2,n}||M_{j_1i_2,n}| \\ &= (1-M_{i_1i_1,n}M_{j_1j_1,n}) + M_{i_1i_1,n}(\sum_{\substack{i_2\in g_1\\ i_2\neq j_1}}|M_{j_1i_2,n}|) \\ &\quad + M_{j_1j_1,n}(\sum_{\substack{j_2\in g_1\\ j_2\neq i_1}}|M_{i_1j_2,n}|) + \sum_{\substack{(g_2,i_2,j_2)\\ i_2\neq j_1,j_2\neq i_1}}|M_{i_1,j_2,n}||M_{j_1,i_2,n}|.\end{aligned}$$

Defining $\mathcal{M}_n = 1-\min_{1\leq i\leq n}M_{ii,n}$, we have that $\mathcal{M}_n = O_p(\frac{K_n}{n})$ (see Cattaneo et al., 2018b) and

$$\max_{\substack{i,j\\ i\neq j}}|M_{ij,n}| \leq \max_{\substack{i,j\\ i\neq j}}\frac{1}{n}\sum_{l=1}^{K_n}|w^*_{il,n}||w^*_{jl,n}| \leq \max_i \frac{1}{n}\sum_{l=1}^{K_n}w^{*2}_{il,n} = O_p(\frac{K_n}{n}).$$

As a result, we have

$$\begin{aligned}\max_{(g_1,i_1,j_1)}\sum_{g_2,i_2,j_2}|d_{i_1,j_1,i_2,j_2,n}| &\leq 2\mathcal{M}_n + 2(\mathcal{C}_{\mathcal{T},n}-1)O_p(\frac{K_n}{n}) + \mathcal{C}^2_{\mathcal{T},n}G_nO_p(\frac{K_n^2}{n^2}) \\ &\leq O_p(\frac{K_n}{n}) + 2(\mathcal{C}_{\mathcal{T},n}-1)O_p(\frac{K_n}{n}) + \mathcal{C}^2_{\mathcal{T},n}O(n)O_p(\frac{K_n^2}{n^2}) \\ &= O_p(\frac{K_n^2}{n}),\end{aligned}$$

which combined with Lemma 2 gives $\mathbb{E}[\tilde{\Sigma}_n(\mathbf{I}_{L_n})|\mathcal{X}_n,\mathcal{W}_n] = \Sigma_n + o_p(1)$.

C.5.5 Proof of Lemma C.3.5

Under homoskedasticity one has

$$\begin{aligned}\mathbb{E}[\tilde{\Sigma}_n(\mathbf{I}_{L_n})|\mathcal{X}_n,\mathcal{W}_n] &= \frac{\sigma_n^2}{n}\sum_{(g_1,i_1,j_1)}\sum_{k=1}^{n}\hat{v}_{i_1,n}\hat{v}_{j_1,n}M_{i_1k,n}M_{j_1k,n} \\ &= \frac{\sigma_n^2}{n}\sum_{(g_1,i_1,j_1)}\hat{v}_{i_1,n}\hat{v}_{j_1,n}M_{i_1j_1,n},\end{aligned}$$

and

$$\Sigma_n = \frac{\sigma_n^2}{n}\sum_{i=1}^{n}\hat{v}_{i,n}^2.$$

As a result, we have

$$\begin{aligned}
&|\mathbb{E}[\tilde{\Sigma}_n(\mathbf{I}_{L_n})|\mathcal{X}_n,\mathcal{W}_n] - \Sigma_n| \\
&\quad\leq \frac{\sigma_n^2}{n}\sum_{i=1}^{n}\hat{v}_{i,n}^2|M_{ii.n}-1| + \frac{\sigma_n^2}{n}\sum_{\substack{(g_1,i_1,j_1)\\ i_1\neq j_1}}|\hat{v}_{i_1,n}||\hat{v}_{j_1,n}||M_{i_1j_1,n}| \\
&\quad\leq \mathcal{C}_{U,n}\mathcal{M}_n(\frac{1}{n}\sum_{i=1}^{n}\hat{v}_{i,n}^2) + \mathcal{C}_{U,n}\mathcal{C}_{\mathcal{T},n}(\max_{\substack{i,j\\ i\neq j}}|M_{ij,n}|)(\frac{1}{n}\sum_{i=1}^{n}\hat{v}_{i,n}^2) \\
&\quad\leq \mathcal{C}_{U,n}O_p(\frac{K_n}{n})(\frac{1}{n}\sum_{i=1}^{n}\hat{v}_{i,n}^2) + \mathcal{C}_{U,n}\mathcal{C}_{\mathcal{T},n}O_p(\frac{K_n}{n})(\frac{1}{n}\sum_{i=1}^{n}\hat{v}_{i,n}^2) \\
&\quad= O_p(\frac{K_n}{n}),
\end{aligned}$$

and therefore $\mathbb{E}[\tilde{\Sigma}_n(\mathbf{I}_{L_n})|\mathcal{X}_n,\mathcal{W}_n] = \Sigma_n + o_p(1)$.

C.6 Extension to within-cluster restrictions

In this section, we present an extension of the class of estimators studied in this paper that allows to impose zero restrictions on the variance-covariance matrix of the errors within clusters.

Define the sets

$$\begin{aligned}
\mathcal{V}_{g,n} &= \{(i,j)\in\mathcal{T}_{g,n}\times\mathcal{T}_{g,n} : \mathbb{E}[U_{i,n}U_{j,n}|\mathcal{X}_n,\mathcal{W}_n]\neq 0\}, \\
\mathcal{R}_{g,i,n} &= \{j\in\mathcal{T}_{g,n} : \mathbb{E}[U_{i,n}U_{j,n}|\mathcal{X}_n,\mathcal{W}_n]\neq 0\},
\end{aligned}$$

and let L_n be the number of non-zero elements contained in $\mathbf{\Omega}_{U,n} = \mathbb{E}[\mathbf{U}_n\mathbf{U}_n'|\mathcal{X}_n,\mathcal{W}_n]$. The generalized version of our proposed class of cluster-robust variance estimators reads:

$$\hat{\mathbf{\Sigma}}_n(\boldsymbol{\kappa}_n) = \frac{1}{n}\sum_{g_1=1}^{G_n}\sum_{g_2=1}^{G_n}\sum_{(i_1,j_1)\in\mathcal{V}_{g_1,n}}\sum_{(i_2,j_2)\in\mathcal{V}_{g_2,n}}\kappa_{g_1,g_2,i_1,j_1,i_2,j_2,n}\hat{\mathbf{v}}_{i_1,n}\hat{\mathbf{v}}_{j_1,n}'\hat{u}_{i_2,n}\hat{u}_{j_2,n},$$

where $\kappa_{g_1,g_2,i_1,j_1,i_2.j_2,n}$ corresponds to the $(h(g_1,i_1,j_1),h(g_2,i_2,j_2))$ entry of the $L_n \times L_n$ symmetric matrix $\boldsymbol{\kappa}_n$, where $h(g,i,j) = [\sum_{k=0}^{(g-1)}(\#\mathcal{V}_{k,n}) + \sum_{k=0}^{i-1}(\#\mathcal{R}_{g,k,n}) + \overline{j(i)_{g,n}}]$ with $\overline{j(i)_{g,n}} = \#\{k \in \mathcal{T}_{g,n} : \mathbb{E}[U_{i,n}U_{j,n}|\mathcal{X}_n,\mathcal{W}_n] \neq 0 \quad \text{and} \quad k \leq j\}$ and we adopt the convention that $\#\mathcal{V}_{0,n} = 0$ and $\#\mathcal{R}_{g,0,n} = 0 \quad \forall g$.

A consistent estimator under Assumptions C.1.1-C.1.3 is then defined as $\hat{\boldsymbol{\Sigma}}(\boldsymbol{\kappa}_n^{\text{CR}})$, where $\boldsymbol{\kappa}_n^{\text{CR}} = (\mathbf{S}_n'(\mathbf{M}_n \otimes \mathbf{M}_n)\mathbf{S}_n)^{-1}$.

Bibliography

ABADIE, A., J. ANGRIST, AND G. IMBENS (2002): "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, 70, 91–117.

ABREVAYA, J., J. A. HAUSMAN, AND S. KHAN (2010): "Testing for Causal Effects in a Generalized Regression Model With Endogenous Regressors," *Econometrica*, 78, 2043–2061.

ADJAHO, C. AND T. CHRISTENSEN (2022): "Externally Valid Treatment Choice," .

ANGRIST, J. AND J. HAHN (2004): "When to Control for Covariates? Panel Asymptotics for Estimates of Treatment Effects," *The Review of Economics and Statistics*, 86, 58–72.

ANGRIST, J. D. AND J.-S. PISCHKE (2010): "The credibility revolution in empirical economics: How better research design is taking the con out of econometrics," *Journal of economic perspectives*, 24, 3–30.

ARELLANO, M. (1987): "PRACTITIONERS' CORNER: Computing Robust Standard Errors for Within-groups Estimators*," *Oxford Bulletin of Economics and Statistics*, 49, 431–434.

ARLOT, S. AND P. L. BARTLETT (2011): "Margin-adaptive model selection in statistical learning," *Bernoulli*, 17, 687 – 713.

ATHEY, S. AND S. WAGER (2021): "Policy Learning With Observational Data," *Econometrica*, 89, 133–161.

AUDIBERT, J.-Y. AND A. B. TSYBAKOV (2007): "Fast learning rates for plug-in classifiers," *The Annals of Statistics*, 35, 608 – 633.

BAI, J. (2009): "Panel data models with interactive fixed effects," *Econometrica*, 77, 1229–1279.

BALKE, A. AND J. PEARL (1997): "Bounds on Treatment Effects from Studies with Imperfect Compliance," *Journal of the American Statistical Association*, 92, 1171–1176.

BELL, R. AND D. MCCAFFREY (2002): "Bias reduction in standard errors for linear regression with multi-stage samples," *Survey Methodology*, 28, 169–181.

BELLONI, A., V. CHERNOZHUKOV, AND C. HANSEN (2013): "Inference on Treatment Effects after Selection among High-Dimensional Controls†," *The Review of Economic Studies*, 81, 608–650.

BERGER, J. O. (1985): *Statistical Decision Theory and Bayesian Analysis*, Springer Series in Statistics, 2nd edition.

BHATTACHARYA, J., A. M. SHAIKH, AND E. VYTLACIL (2012): "Treatment effect bounds: An application to Swan–Ganz catheterization," *Journal of Econometrics*, 168, 223–243.

BLOOM, H. S., L. L. ORR, S. H. BELL, G. CAVE, F. DOOLITTLE, W. LIN, AND J. M. BOS (1997): "The Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Training Partnership Act Study," *The Journal of Human Resources*, 32, 549–576.

BYAMBADALAI, U. (2022): "Identification and Inference for Welfare Gains without Unconfoundedness," .

CAMERON, A. C., J. B. GELBACH, AND D. L. MILLER (2008): "Bootstrap-Based Improvements for Inference with Clustered Errors," *The Review of Economics and Statistics*, 90, 414–427.

CAMERON, A. C. AND D. L. MILLER (2015): "A Practitioner's Guide to Cluster-Robust Inference," *Journal of Human Resources*, 50, 317–372.

CANAY, I. A., A. SANTOS, AND A. M. SHAIKH (2021): "The Wild Bootstrap with a "Small" Number of "Large" Clusters," *The Review of Economics and Statistics*, 103, 346–363.

CATTANEO, M. D., M. JANSSON, AND W. K. NEWEY (2018a): "Alternative Asymptotics and the Partially Linear Model with Many Regressors," *Econometric Theory*, 34, 277–301.

——— (2018b): "Inference in Linear Regression Models with Many Covariates and Heteroscedasticity," *Journal of the American Statistical Association*, 113, 1350–1361.

CHAMBERLAIN, G. (2011): "1011 Bayesian Aspects of Treatment Choice," in *The Oxford Handbook of Bayesian Econometrics*, Oxford University Press.

CHEN, X. (2007): "Large Sample Sieve Estimation of Semi-Nonparametric Models," Elsevier, vol. 6 of *Handbook of Econometrics*, 5549–5632.

CHERNOZHUKOV, V., J. C. ESCANCIANO, H. ICHIMURA, W. K. NEWEY, AND J. M. ROBINS (2022): "Locally Robust Semiparametric Estimation," *Econometrica*, 90, 1501–1535.

CHERNOZHUKOV, V. AND C. HANSEN (2006): "Instrumental quantile regression inference for structural and treatment effect models," *Journal of Econometrics*, 132, 491–525.

CHERNOZHUKOV, V., S. LEE, AND A. M. ROSEN (2013): "Intersection Bounds: Estimation and Inference," *Econometrica*, 81, 667–737.

CHESHER, A. (2010): "Instrumental variable models for discrete outcomes," *Econometrica*, 78, 575–601.

CHESHER, A. AND A. M. ROSEN (2017): "Generalized instrumental variable models," *Econometrica*, 85, 959–989.

CHRISTENSEN, T., H. R. MOON, AND F. SCHORFHEIDE (2022): "Optimal Discrete Decisions when Payoffs are Partially Identified," .

CUI, Y. AND E. T. TCHETGEN (2021): "A Semiparametric Instrumental Variable Approach to Optimal Treatment Regimes Under Endogeneity," *Journal of the American Statistical Association*, 116, 162–173, pMID: 33994604.

DAI, J. Y. AND X. C. ZHANG (2015): "Mendelian randomization studies for a continuous exposure under case-control sampling," *American Journal of Epidemiology*, 181, 440–449.

DEHEJIA, R. H. (2005): "Program evaluation as a decision problem," *Journal of Econometrics*, 125, 141–173, experimental and non-experimental evaluation of economic policy and models.

DONOHUE, JOHN J., I. AND S. D. LEVITT (2001): "The Impact of Legalized Abortion on Crime*," *The Quarterly Journal of Economics*, 116, 379–420.

DONOHUE, J. J. AND S. D. LEVITT (2008): "Measurement Error, Legalized Abortion, and the Decline in Crime: A Response to Foote and Goetz," *The Quarterly Journal of Economics*, 123, 425–440.

FANG, Z. AND A. SANTOS (2018): "Inference on Directionally Differentiable Functions," *The Review of Economic Studies*, 86, 377–412.

FARRELL, M. H. (2015): "Robust inference on average treatment effects with possibly more covariates than observations," *Journal of Econometrics*, 189, 1–23.

FARRELL, M. H., T. LIANG, AND S. MISRA (2021): "Deep Neural Networks for Estimation and Inference," *Econometrica*, 89, 181–213.

FOOTE, C. L. AND C. F. GOETZ (2008): "The Impact of Legalized Abortion on Crime: Comment," *The Quarterly Journal of Economics*, 123, 407–423.

FOSTER, D. J. AND V. SYRGKANIS (2019): "Orthogonal Statistical Learning," .

HAHN, J. AND W. NEWEY (2004): "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models," *Econometrica*, 72, 1295–1319.

HAN, S. (2019): "Optimal Dynamic Treatment Regimes and Partial Welfare Ordering," .

HAN, S. AND S. LEE (2019): "Estimation in a generalization of bivariate probit models with dummy endogenous regressors," *Journal of Applied Econometrics*, 34, 994–1015.

HAN, S. AND E. J. VYTLACIL (2017): "Identification in a generalization of bivariate probit models with dummy endogenous regressors," *Journal of Econometrics*, 199, 63–73.

HANSEN, B. E. AND S. LEE (2019): "Asymptotic theory for clustered samples," *Journal of Econometrics*, 210, 268–290.

HANSEN, C. B. (2007): "Asymptotic properties of a robust variance matrix estimator for panel data when T is large," *Journal of Econometrics*, 141, 597–620.

HARDING, M. AND C. LAMARCHE (2014): "Estimating and testing a quantile regression model with interactive effects," *Journal of Econometrics*, 178, 101–113.

HECKMAN, J. J. AND E. J. VYTLACIL (2001): "Instrumental variables, selection models, and tight bounds on the average treatment effect," in *Econometric Evaluation of Labour Market Policies*, Springer, 1–15.

HIRANO, K. AND J. R. PORTER (2009): "Asymptotics for Statistical Treatment Rules," *Econometrica*, 77, 1683–1701.

——— (2012): "IMPOSSIBILITY RESULTS FOR NONDIFFERENTIABLE FUNCTIONALS," *Econometrica*, 80, 1769–1790.

——— (2020): "Asymptotic analysis of statistical decision rules in econometrics," in *Handbook of Econometrics, Volume 7A*, ed. by S. N. Durlauf, L. P. Hansen, J. J. Heckman, and R. L. Matzkin, Elsevier, vol. 7 of *Handbook of Econometrics*, 283–354.

HUBER, P. J. (1973): "Robust Regression: Asymptotics, Conjectures and Monte Carlo," *The Annals of Statistics*, 1, 799 – 821.

HURWICZ, L. (1951): "The Generalised Bayes-Minimax Principle: A Criterion for Decision-Making Under Uncertainty," .

IBRAGIMOV, R. AND U. K. MÜLLER (2016): "Inference with Few Heterogeneous Clusters," *The Review of Economics and Statistics*, 98, 83–96.

ICHIMURA, H. AND W. K. NEWEY (2022): "The influence function of semiparametric estimators," *Quantitative Economics*, 13, 29–61.

IMBENS, G. W. AND J. D. ANGRIST (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.

IMBENS, G. W. AND M. KOLESÁR (2016): "Robust Standard Errors in Small Samples: Some Practical Advice," *The Review of Economics and Statistics*, 98, 701–712.

IMBENS, G. W. AND W. K. NEWEY (2009): "Identification and estimation of triangular simultaneous equations models without additivity," *Econometrica*, 77, 1481–1512.

ISHIHARA, T. AND T. KITAGAWA (2021): "Evidence Aggregation for Treatment Choice," .

KALLUS, N. AND A. ZHOU (2018): "Confounding-Robust Policy Improvement," .

KASY, M. (2016): "Partial Identification, Distributional Preferences, and the Welfare Ranking of Policies," *The Review of Economics and Statistics*, 98, 111–131.

KENNEDY, E. H. (2022): "Semiparametric doubly robust targeted double machine learning: a review," .

KIDO, D. (2022): "Distributionally Robust Policy Learning with Wasserstein Distance," .

KIM, W., K. KWON, S. KWON, AND S. LEE (2018): "The identification power of smoothness assumptions in models with counterfactual outcomes," *Quantitative Economics*, 9, 617–642.

KITAGAWA, T., S. LEE, AND C. QIU (2022): "Treatment Choice with Nonlinear Regret," .

KITAGAWA, T., J. L. MONTIEL OLEA, J. PAYNE, AND A. VELEZ (2020): "Posterior distribution of nondifferentiable functions," *Journal of Econometrics*, 217, 161–175.

KITAGAWA, T., S. SAKAGUCHI, AND A. TETENOV (2021): "Constrained Classification and Policy Learning," .

KITAGAWA, T. AND A. TETENOV (2018): "Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice," *Econometrica*, 86, 591–616.

KOENKER, R. AND G. BASSETT (1978): "Regression quantiles," *Econometrica: journal of the Econometric Society*, 33–50.

LEE, D. S. (2009): "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *The Review of Economic Studies*, 76, 1071–1102.

LEE, N., H. R. MOON, AND M. WEIDNER (2012): "Analysis of interactive fixed effects dynamic linear panel regression with measurement error," *Economics Letters*, 117, 239–242.

LEWBEL, A. (2000): "Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables," *Journal of Econometrics*, 97, 145–177.

LI, C. M. (2016): "Inference in Regressions with Many Controls," .

LI, C. M. AND U. K. MÜLLER (2021): "Linear regression with many controls of limited explanatory power," *Quantitative Economics*, 12, 405–442.

LIANG, K.-Y. AND S. L. ZEGER (1986): "Longitudinal data analysis using generalized linear models," *Biometrika*, 73, 13–22.

LITTLE, R. J. A. (1985): "A Note About Models for Selectivity Bias," *Econometrica*, 53, 1469–1474.

LOUREIRO, M. L., A. SANZ-DE GALDEANO, AND D. VURI (2010): "Smoking Habits: Like Father, Like Son, Like Mother, Like Daughter?*," *Oxford Bulletin of Economics and Statistics*, 72, 717–743.

MACKINNON, J. G. (2013): *Thirty Years of Heteroskedasticity-Robust Inference*, New York, NY: Springer New York, 437–461.

MAMMEN, E. (1993): "Bootstrap and Wild Bootstrap for High Dimensional Linear Models," *The Annals of Statistics*, 21, 255 – 285.

MAMMEN, E. AND A. B. TSYBAKOV (1999): "Smooth discrimination analysis," *The Annals of Statistics*, 27, 1808 – 1829.

MANSKI, C. F. (1985): "Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator," *Journal of Econometrics*, 27, 313–333.

——— (1990): "Nonparametric Bounds on Treatment Effects," *The American Economic Review*, 80, 319–323.

——— (2004): "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica*, 72, 1221–1246.

——— (2009): "Diversified treatment under ambiguity," *International Economic Review*, 50, 1013–1041.

——— (2010): "Vaccination with partial knowledge of external effectiveness," *Proceedings of the National Academy of Sciences*, 107, 3953–3960.

——— (2011): "Choosing Treatment Policies Under Ambiguity," *Annual Review of Economics*, 3, 25–49.

MANSKI, C. F. AND J. V. PEPPER (2000): "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 68, 997–1010.

MBAKOP, E. AND M. TABORD-MEEHAN (2021): "Model Selection for Treatment Choice: Penalized Welfare Maximization," *Econometrica*, 89, 825–848.

MONFARDINI, C. AND R. RADICE (2008): "Testing Exogeneity in the Bivariate Probit Model: A Monte Carlo Study*," *Oxford Bulletin of Economics and Statistics*, 70, 271–282.

MOON, H. R., M. SHUM, AND M. WEIDNER (2018): "Estimation of random coefficients logit demand models with interactive fixed effects," *Journal of Econometrics*, 206, 613–644.

MOULTON, B. R. (1986): "Random group effects and the precision of regression estimates," *Journal of Econometrics*, 32, 385–397.

MOURIFIÉ, I. AND R. MÉANGO (2014): "A note on the identification in two equations probit model with dummy endogenous regressor," *Economics Letters*, 125, 360–363.

MU, B. AND Z. ZHANG (2018): "Identification and estimation of heteroscedastic binary choice models with endogenous dummy regressors," *The Econometrics Journal*, 21, 218–246.

NEWEY, W. K. (1986): "Linear instrumental variable estimation of limited dependent variable models with endogenous explanatory variables," *Journal of Econometrics*, 32, 127–141.

PESARAN, M. H. (2006): "Estimation and inference in large heterogeneous panels with a multifactor error structure," *Econometrica*, 74, 967–1012.

PONOMAREV, K. (2022): "Efficient Estimation of Directionally Differentiable Functionals," .

PU, H. AND B. ZHANG (2021): "Estimating optimal treatment rules with an instrumental variable: A partial identification learning approach," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83, 318–345.

PUSTEJOVSKY, J. E. AND E. TIPTON (2018): "Small-Sample Methods for Cluster-Robust Variance Estimation and Hypothesis Testing in Fixed Effects Models," *Journal of Business & Economic Statistics*, 36, 672–683.

RIVERS, D. AND Q. H. VUONG (1988): "Limited information estimators and exogeneity tests for simulataneous probit models," *Journal of Econometrics*, 39, 347–366.

ROSENBAUM, P. R. (1987): "Sensitivity Analysis for Certain Permutation Inferences in Matched Observational Studies," *Biometrika*, 74, 13–26.

RUSSELL, T. M. (2020): "Policy Transforms and Learning Optimal Policies," .

STOCK, J. H. AND M. W. WATSON (2008): "Heteroskedasticity-Robust Standard Errors for Fixed Effects Panel Data Regression," *Econometrica*, 76, 155–174.

STOYE, J. (2009): "Minimax regret treatment choice with finite samples," *Journal of Econometrics*, 151, 70–81.

——— (2012): "Minimax regret treatment choice with covariates or with limited validity of experiments," *Journal of Econometrics*, 166, 138–156, annals Issue on "Identification and Decisions", in Honor of Chuck Manski's 60th Birthday.

SUN, L. (2021): "Empirical Welfare Maximization with Constraints," .

VALIANT, L. G. (1984): "A Theory of the Learnable," in *Proceedings of the Sixteenth Annual ACM Symposium on Theory of Computing*, New York, NY, USA: Association for Computing Machinery, STOC '84, 436–445.

VAPNIK, V. N. (1998): *Statistical Learning Theory*, New York: Wiley.

VARAH, J. (1975): "A lower bound for the smallest singular value of a matrix," *Linear Algebra and its Applications*, 11, 3–5.

VERDIER, V. (2020): "Estimation and Inference for Linear Models with Two-Way Fixed Effects and Sparsely Matched Data," *The Review of Economics and Statistics*, 102, 1–16.

VIVIANO, D. (2019): "Policy Targeting under Network Interference," .

WAINWRIGHT, M. J. (2019): *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

WALD, A. (1950): *Statistical Decision Functions*, Wiley.

WHITE, H. (1984): *Asymptotic Theory for Econometricians*, San Diego: Academic Press. University of North Carolina.

WINDMEIJER, F. (2019): "Two-stage least squares as minimum distance," *The Econometrics Journal*, 22, 1–9.

WRIGHT, P. G. (1928): *Tariff on animal and vegetable oils*, Macmillan Company, New York.

YATA, K. (2021): "Optimal Decision Rules Under Partial Identification," .

YILDIZ, N. (2013): "Estimation of binary choice models with linear index and dummy endogenous variables," *Econometric Theory*, 29, 354–392.

ZHAO, Y., D. ZENG, A. J. RUSH, AND M. R. KOSOROK (2012): "Estimating Individualized Treatment Rules Using Outcome Weighted Learning," *Journal of the American Statistical Association*, 107, 1106–1118.

ZIMMER, D. (2018): "Using copulas to estimate the coefficient of a binary endogenous regressor in a Poisson regression: Application to the effect of insurance on doctor visits," *Health Economics*, 27, 545–556.

Statement of Conjoint Work

Note on the joint work in Riccardo D'Adamo's thesis "Essays in Econometrics".

Chapter 1, "Orthogonal Policy Learning Under Ambiguity", is single-authored by Riccardo D'Adamo.
Chapter 2, "Auxiliary IV Estimation for Nonlinear Models", was undertaken as joint work with Martin Weidner and Frank Windmeijer.
Chapter 3, "Cluster-Robust Standard Errors for Linear Regression Models with Many Controls", is single-authored by Riccardo D'Adamo.
Contributions from the authors were equal in the case of the coauthored chapters.
Signatures from coauthors: