Title
Essays on Nonparametric and High-Dimensional Econometrics

Author
Soerensen, Jesper Riis-Vestergaard

Publication Date
2018

UNIVERSITY OF CALIFORNIA

Los Angeles

Essays on Nonparametric and High-Dimensional Econometrics

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Economics

by

Jesper Riis-Vestergaard Soerensen

2018

ABSTRACT OF THE DISSERTATION

Essays on Nonparametric and High-Dimensional Econometrics

by

Jesper Riis-Vestergaard Soerensen
Doctor of Philosophy in Economics
University of California, Los Angeles, 2018
Professor Denis Nikolaye Chetverikov, Co-Chair
Professor Jinyong Hahn, Co-Chair

This dissertation studies questions related to identification, estimation, and specification testing of nonparametric and high-dimensional econometric models. The thesis is composed by two chapters.

In Chapter 1, I propose specification tests for two formally distinct but related classes of econometric models: (1) semiparametric conditional moment restriction models dependent on conditional expectation functions, and (2) a class of high-dimensional unconditional moment restriction models dependent on high-dimensional best linear predictors. These classes may be motivated by economic models in which agents make choices under uncertainty and therefore have to predict payoff-relevant variables such as the behavior of other agents. The proposed tests are shown to be both asymptotically correctly sized and consistent. Moreover, I establish a bound on the rate of local alternatives for which the test for high-dimensional unconditional moment restriction models is consistent. These results allow researchers to test the specification of their models without introducing additional parametric, typically ad hoc, assumptions on expectations.

In Chapter 2, I show that it is possible to identify and estimate a generalized panel regression model (GPRM) without imposing any parametric structure on (1) the function of observable explanatory variables, (2) the systematic function through which the function of observable explanatory variables, fixed effect, and disturbance term generate the outcome variable, or (3) the distribution of unobservables. I proceed with estimation using a series maximum rank correlation estimator (SMRCE) of the function of observable explanatory variables and provide conditions under which L^2–consistency is achieved. I also provide conditions under which both L^2 and uniform convergence rates of the SMRCE may be derived.

The dissertation of Jesper Riis-Vestergaard Soerensen is approved.

Marek Biskup

Zhipeng Liao

Adriana Lleras-Muney

Denis Nikolaye Chetverikov, Committee Co-Chair

Jinyong Hahn, Committee Co-Chair

University of California, Los Angeles

2018

Contents

Acknowledgments

I owe a heartfelt debt of gratitude to my primary thesis advisors, Denis Chetverikov and Jinyong Hahn. Both contributed much patience and guidance through my years as a graduate student. Our routine meetings were the apex of my time at UCLA. Denis encouraged me to have the fortitude to expand my skill set. I only hope his work ethic has rubbed off on me. Jin taught me how to make progress on seemingly impossible problems. He was always there to answer endless questions.

A special thanks goes out to Rosa Matzkin, whose teaching and research first stimulated my interest in econometrics. Her comments heavily influenced Chapter 2 of this thesis. I also thank Marek Biskup, Bo Honoré, Zhipeng Liao, Adriana Lleras-Muney, Kathleen McGarry, Anders Munk-Nielsen, Rasmus Søndergaard Pedersen, Anders Rahbek, Andres Santos, and Shuyang Sheng for helpful discussions. Comments from seminar participants at the 2017 CAM-CEBI Research Workshop at University of Copenhagen, 2015 and 2017 DAEiNA Meetings, UCLA econometrics proseminar, Simon Fraser University and Mannheim University helped shape this dissertation as well.

My fellow students have affected my life and research ideas for the past six years. I thank both my current and former office mates (Omer Ali, El-Hadi Çaoui, Tiago Caruso, Alex Fon, Carlos Cantú Garcia, Renato Giroldo, Akina Ikudo, Byeonghyeon Jeong, Greg Kubitz, Box Sean, Ksenia Shahkgildyan, and Kyle Woodward) for making our office a fun yet productive place to be. I am proud to have served in Bunche 9360.

Finally, I thank Matt Miller, Gianni Nicolò, Andrés Schneider, Liyan Shi and the rest of my classmates for their continued friendship since we arrived at UCLA in 2012.

Vita: Jesper Riis-Vestergaard Soerensen (Sørensen)

Education

University of California, Los Angeles	C.Phil. Economics, 2014
Los Angeles, CA, USA	M.A. Economics, 2014
University of Copenhagen	M.S. Economics, 2014
Copenhagen, Denmark	B.S. Economics, 2010
Cornell University,	Visiting Graduate Student, 2011–12
Ithaca, NY, USA	

Fellowships and Awards

• Dissertation Year Fellowship, UCLA Graduate Division	2017–18
• Alumni Association Fellowship, UCLA Economics	2016–17
• Welton Graduate Prize in Economics, UCLA Economics	2015
• Teaching Assistant Award, UCLA Economics	Winter 2015
• Graduate Student Fellowship, UCLA Economics	2012–16

Journal Refereeing

- Quantitative Economics.

Research Assistance

For Denis Chetverikov, UCLA Economics, 2017

- “On Cross-Validated Lasso” (with Zhipeng Liao and Victor Chernozhukov).

For Adriana Lleras-Muney, UCLA Economics, 2014

- “Estimation and Inference using Imperfectly Matched Data” (with Bo Honoré).

Conference Presentations and Invited Seminars

2018 Mannheim U. (Germany), Bates–White Economic Consulting (USA), Simon Fraser U. (Canada).

2017 CAM-CEBI Workshop (U. of Copenhagen, Denmark), 6th Annual DAEiNA Meeting (at Washington and Lee U., USA).

2015 4th Annual DAEiNA Meeting (at U. of Copenhagen, Denmark).

Teaching Assistance

Ph.D.-level courses (all UCLA):

• Teaching College Economics	AY 2015–16, AY 2016–17
• Econometrics I	Fall 2014

Undergraduate-level courses (all UCLA):

• Introduction to Econometrics	Spring 2014
• Statistics for Economists	Spring & Winter 2017, Fall & Spring 2016 Spring & Winter 2015, Winter 2014
• Principles of Economics (Macro)	Fall 2013

Introduction

Economists form theories and formulate models on the behavior and interaction of economic agents and how economies work. Econometrics is an economic discipline which deals with the application of statistical methods and mathematical economics to economic data. Through observation or experimentation, econometrics aims to give empirical content to economic relations and theories. This discipline is used, in part, to obtain helpful estimates for diligent policymakers.

Unlike researchers in the physical sciences, economists are rarely able to conduct controlled experiments. Econometricians therefore face the challenge of quantifying economic relationships using data generated by complex systems of related equations, in which many variables may change simultaneously. Theoretical econometrics relies on economic and mathematical reasoning, theoretical statistics, and numerical methods to argue that a new formula may have the ability to correct inferences outside of controlled environments.

The methodology of economics generally consists of four steps:

1. Suggest a theory to interpret existing data.

2. Develop a model that captures the body of the theory one wishes to test.

3. Use relevant statistical procedures to estimate the unknown parameters of the model.

4. Determine whether the model makes economic sense through hypothesis testing.

The end result of this process, if all goes well, is a tool that can be used to assess the empirical validity of an abstract economic model.

This thesis expands the toolkits of empirical economists used in Steps 3 and 4 by providing methods for estimation and testing under more general conditions than were previously available. Specifically, in Chapters 1 and 2, I develop methods for *testing* the overall accuracy of the employed model (Step 4) and *estimating* unknown model parameters (Step 3), respectively, under *weaker functional form assumptions* than previously invoked. While economic theory may predict the relationship between two economic variables to take a particular shape (e.g., monotonicity or concavity), it rarely takes a stand on a particular parametric functional form (e.g., linearity). Thus, when available, econometric procedures that do *not* hinge on parametric functional form assumptions ought to be favored to those that do, as the former type of procedures align more closely with the underlying economic theory.

In the following two sections, I provide a more detailed introduction to the particular testing and estimation problems considered in this thesis as well as the specific functional form assumptions I have relaxed.

Chapter 1: Consistent Specification Testing in Semiparametric and High-Dimensional Moment Models

Empirical work in economics typically relies on the use of *econometric models*, i.e., simplified, statistical constructs serving the purpose of illustrating complex processes. Any modelling process should be accompanied by a measure of model accuracy, sometimes referred to as performing model *specification tests* (or diagnostics). Specification testing is critical since the usefulness of a model hinges on the precision at which it reflects the relationships it aims to understand.

Econometric models are often indexed by a mix of *parametric* (i.e., fixed- and finite-dimensional) and *nonparametric* (i.e., infinite-dimensional) components. Such models are therefore said to be *semiparametric*.

Semiparametric models occur naturally in settings where agents make choices under uncertain conditions. Decision-making under uncertainty is pervasive in economics and covers both single-agent models and models with strategic interactions (i.e., games). For example, a high-school graduate decides to attend college not knowing if he/she will be able to complete college, perhaps due to financial constraints (or a host of other variables). Similarly, firms decide on whether to enter a new market not knowing the entry decision of their competitors.

A feature of decision-making under uncertainty is that agents have to form *expectations* over payoff-relevant variables unobserved at the time of decision, i.e, they must assess the likelihood of uncertain variables taking on various outcomes in the future. In the college-decision example, whether a high-school graduate obtains a college degree matters for their future employability and, thus, their wage trajectory. Likewise, in the firm-entry example, the profitability of a firm entering a new market depends on the level of competition it stands to face.

Economic theory typically provides little guidance towards the functional form of expectations generated by agents. It therefore seems reasonable to view expectations formed by agents as nonparametric objects both when fitting the resulting (semiparametric) model and evaluating its accuracy.

The literature on model specification testing in econometrics is voluminous and dates back to at least the early 1980s. The work of Herman Bierens is particularly relevant (see, e.g., Bierens, 1982). However, most of the existing literature concerns testing the specification of classes of parametric models or tests tailored to particular instances of semi- or nonparametric models. In contrast, in Chapter 1, I develop a general method for testing the specification of a class of partially or even fully nonparametrically specified models.

Depending on the nature of the decision, expectations formed by agents may depend on a number of variables ranging from only a few to numerous. For example, it may be reasonable to assume that a financially constrained high-school graduate assesses their likelihood of college completion based on just a few variables such as their current wealth and borrowing limit (or lack thereof). In contrast, a firm deciding on entry will in general have to consider not only the characteristics of the market but also all of its competitors, leading to a potentially large set of variables.

When the number of variables is relatively small, one may allow expectations to be nonparametric and leave it to the data to determine their functional forms. To this end, one may employ classical nonparametric methods such as kernel or series estimation. However, when expectations depend on *many* sources, classical nonparametric approaches may break down. As a middle ground between the restrictive low-dimensional, parametric setup and the infeasible infinite-dimensional nonparametric framework, one may entertain a *high-dimensional* setting. A high-dimensional specification allows the number of candidate inputs—and, thus, the number of parameters to be estimated—to be large and, in fact, potentially much larger than the sample size available to the researcher. Under an assumption of *sparsity*, modern machine learning techniques such as the LASSO (Tibshirani, 1996) work well even with a high-dimensional number of parameters to be estimated. Sparsity means that from the potentially very large collection of candidate variables only a few (a priori unknown) variables actually matter.

Building on the insights of Bierens (1982), in Chapter 1, I construct test statistics for models involving either nonparametric or high-dimensional expectations. The chapter is divided into two parts as these different modelling environments, as well as the different estimation techniques employed to construct the test statistics, necessitate substantially different arguments in order to establish the large-sample probabilistic behavior of the proposed test statistics. Nonetheless, drawing on results from the statistics literature for "functional central limit theorems" (see, e.g., van der Vaart and Wellner, 1996) and recently developed "high-dimensional central limit theorems" (see Chernozhukov, Chetverikov, and Kato, 2013), I construct testing procedures that, at least in large samples, are able to distinguish between correctly and incorrectly specified models.

The practical usefulness of the results in Chapter 1 is to provide researchers with general tools that allow them to test the accuracy of their models without having to impose additional parametric, typically ad hoc, assumptions on expectations.

Chapter 2: Identification and Estimation of a Generalized Panel Regression Model

Many empirical applications in economics involve *limited dependent variables*. Variables may be inherently unobservable, limited due to (optimal) choice or mechanically limited. For example, when studying the labor market participation of married females, one only observes whether or not the married female participates and not their underlying, inherently unobservable, willingness to participate. If the object of interest is determinants of wages, then one faces the problem that wages are observed only for those who choose to work. In addition, studying determinants of wealth, one may face the problem of data censoring such as interval or top coding, perhaps due to privacy concerns. In the case of interval coding wealth is in principle observable, but the researcher only observes wealth up to a bracket (e.g., $100,000–$125,000).

Economists often have access to *panel data*, i.e., repeated observations of the same units (e.g., the same individuals in multiple years). Access to panel data allows researchers to control for time-invariant unit-specific effects such as individual ability or taste by inspecting the same unit across time. This feature makes panel data analysis compelling relative to analysis based on a single cross section.

The traditional approach to fitting limited dependent variable models for panel data has been to specify parametric functional forms for all model unknowns. Parametric assumptions assist the researcher in determining what can be learned from the model under the thought experiment of having access to unlimited data, known as *identification* analysis. They also facilitate estimation as the number of unknowns to be quantified using the available data has been greatly reduced by the assumption of parametric functional forms.

However, misspecification of one or more model components may lead to undesirable behavior of standard estimators. Even worse, incorrectly specified parametric functional forms may lead to a lack of identification altogether.

To avoid having identification driven by parametric functional form assumptions, in Chapter 2, I analyze *nonparametric* versions of a collection of panel data models including typically invoked limited dependent variable models. The class of models considered assume a *mono-*

tonic relationship, at least on average, between the rank of the outcome of interest and the rank of the variables used to explained said outcome but does not impose any parametric functional form on this relationship. I show that this type of “rank correlation” assumption can lead to identification of elements of interest in such panel data models. The constructive nature of my identification result suggests natural estimators and I derive their statistical properties.

The results in this chapter provide researchers with tools for estimation of a class of panel data models under weaker functional form assumptions. By comparing the resulting nonparametric estimates with estimates obtained under parametric assumptions, the results in this chapter may also be used to assess the sensitivity of the analysis to the latter set of assumptions.

Chapter 1

Consistent Specification Testing in Semiparametric and High-Dimensional Moment Models

1.1 Introduction

This paper concerns testing the specification of a class of semiparametric conditional moment restriction (CMR) models and a class of high-dimensional unconditional moment restriction (UMR) models. The two classes of models both allow parameterizations to involve flexibly specified predictions: In the CMR models predictions are fully nonparametric, while in the high-dimensional UMR models predictions are high-dimensionally linear. Simple examples of members of these two classes are the partially linear regression model and the linear treatment model with a high-dimensional number of controls, respectively. However, the presence of predictions in the model parameterizations is often intended to capture expectation formation made by economic agents operating within an uncertain environment. Flexible specification of these predictions is then motivated by the fact that researchers typically have limited information on how such expectations are formed.

Econometric models often involve one or multiple agents acting optimally within an uncertain environment. Optimal choice under uncertainty requires decision makers to predict payoff-relevant variables unknown at the time of their decision given their available information. For example, high school graduates decide on whether to attend college not knowing if they will be able to complete college, e.g., due to financial limitations. They must therefore form an opinion about whether they will obtain a degree should they enroll (see, e.g., Manski 1991). Similarly, firms decide on whether to enter a new market not knowing the entry decision of their competitors and must therefore predict whether their competitors will enter (see, e.g., Bajari, Hong, Krainer, and Nekipelov 2010). Economic theory typically provides little guidance towards the functional form of these predictions or expectations, which consequently should be specified in a flexible manner. Even when these predictions are flexibly specified, the models employed may yield a poor approximation to actual behavior. In order to know if conclusions derived from the analysis of such econometric models can be trusted, it is necessary to statistically test whether the model specification is consistent with the data to which it is applied. In other words, does the data reject the model? In this paper, I provide tools for addressing this question.

A fully nonparametric approach to the predictions leads to “classical” semiparametric econometrics, while adopting a high-dimensional linear form may be thought of as “modern” high-dimensional econometrics. Due to the different econometric environments, this paper is divided into two parts. In the first part I consider a class of semiparametric CMR models whose parameterizations involve conditional expectation functions (CEFs). These CMR models are “semiparametric” in the sense that, while the model may impose parametric restrictions, the CEFs are left nonparametric. While a nonparametric *specification* of CEFs remains true to the economic model, in some applications (fully) nonparametric *estimation* may not be practically feasible due to the curse of dimensionality. For example, in the market entry game a firm must in general predict the entry of a competitor as a function of the observable characteristics of *all* firms, which may lead to a sizable state space.

Acknowledging that a fully nonparametric treatment of predictions may be too flexible for practical considerations and that a simple linear model in a few of the state variables may miss important conditioning information, one may be willing to adopt the more parsimonious yet still flexible assumption of high-dimensional linearity. In the second part of this paper I consider a class of high-dimensional moment models. These moment models are “high-dimensional” for two reasons: (1) the number of unconditional moment restrictions to be tested may grow with and possibly exceed the sample size available to the researcher, and (2) the parameterization of these models may itself involve high-dimensional components. The high-dimensional components are here taken to be best linear predictors. Loosely speaking, these models involve numerous moments, each of which may depend on very many “regressors.” To make effective use of the high-dimensional number of regressors, I will rely on the important structure of approximate sparsity (see, e.g., Belloni and Chernozhukov 2011, Belloni, Chen, Chernozhukov, and Hansen 2012, Belloni, Chernozhukov, and Hansen 2014a). In the context of this paper, approximate sparsity refers to the condition that each high-dimensional best linear predictor by and large depends on a small (but a priori unknown) set of regressors.

The high-dimensional best linear predictors may be thought of as finite-dimensional but adaptable approximations to the predictions made by the agents in an underlying economic model. Alternatively, one may justify focusing on best linear predictors by means of *bounded rationality*. The conditional expectation is the optimal predictor under mean-square loss. A fully rational agent computes the optimal predictor and uses it to solve their decision problem. In contrast, a boundedly rational decision-maker may find the computation of a conditional expectation intractable or too time consuming and may find that the more manageable best linear predictors yield a suboptimal yet satisfactory solution.

The first contribution of this paper is to propose a class of specification tests that apply generally to CMR models involving nonparametrically specified CEFs and show that the proposed specification tests are both asymptotically correctly sized and consistent. These results add to the existing literature on consistent specification testing in CMR models. Recall that a test is called *consistent* if its power against any deviation from the null hypothesis approaches one as the sample size grows without bound. The first consistent test for the specification of functional form of cross-sectional regression models was proposed by Bierens (1982) and is sometimes referred to as the Integrated Conditional Moment (ICM) test (Bierens and Ploberger, 1997) or the Bierens Test (de Jong, 1996). Bierens' key observation was that the null hypothesis of a CMR may be equivalently expressed as a testable null hypothesis involving possibly infinitely many UMRs constructed by interacting the model residual with carefully chosen weight functions depending on the conditioning variables. The properties required of these weight function are characterized by Stinchcombe and White (1998). Bierens (1984), de Jong (1996) and Bierens and Ploberger (1997) extended the ICM test to allow for time series regression.

Following Bierens' original paper, two strands of literature emerged. One strand of the literature further developed ICM-type (or related) consistent tests of conditional expectation (mainly regression) models (Stute 1997, Stinchcombe and White 1998, Boning and Sowell 1999, Fan and Li 2000, Whang 2001, and Escanciano 2006). Moreover, Stinchcombe and White (1998) and Whang (2001) extended Bierens' approach to testing to a more general parametric context than the standard regression framework. In particular, their treatments allowed for general parametric CMR models.[1]

A different strand of the literature constructed tests by comparing estimates imposing parametric functional forms with nonparametric or semiparametric estimates (Wooldridge 1992, Yatchew 1992, Hardle and Mammen 1993, Gozalo 1993, Horowitz and Härdle 1994,

[1] Donald, Imbens, and Newey (2003) developed consistent specification tests for parametric CMRs based on generalized empirical likelihood ratio test statistics using a finite but growing number of UMRs. While their tests are not of the ICM-type, they are similar in spirit.

Hong and White (1995), Li and Wang 1998, Zheng 1996, and Lavergne and Vuong 2000, among others). Given that the majority of the latter collection of papers employ kernel regression or smoothing methods, I will refer to these tests as "kernel-based." Stinchcombe and White (1998) unifies the seemingly dissimilar ICM and kernel-based approaches to testing models consistently by showing that they are, in some sense, dual treatments of the same problem. Moreover, Fan and Li (2000) have shown that a particular version of the ICM test may be viewed as a kernel-based tests albeit with a fixed bandwidth.

In this paper I provide a framework for testing the specification of not just regression or other conditional expectation models, but a *class* of CMR models. Building on Bierens' key observation, I recast the null hypothesis of a correctly specified CMR as a testable collection of UMRs. Consequently, my test statistic is of the ICM type. While the above references require *parametrically* specified models, I allow the parameterization of such CMR models to include *non*parametrically specified CEFs. This added flexibility allows a researcher to test the specification of their model without introducing ad hoc assumptions with respect to expectation formation.

The second contribution of this paper is to propose a method for testing the specification of high-dimensional UMR models and show that this method provides asymptotic size control. In addition, I establish an upper bound on the rate of local alternatives for which the test is consistent. To the best of my knowledge, this paper provides the first specification test for high-dimensional econometric models.

Lastly, in this paper I make an additional contribution of potential independent interest by providing low-level conditions under which the Lasso (Tibshirani, 1996) can be used for estimation of potentially very many high-dimensional best linear predictors. The properties of the Lasso for estimation of a single or potentially very many CEFs are well understood (see, e.g., Bickel, Ritov, and Tsybakov 2009, Belloni and Chernozhukov 2011, Belloni et al. 2012, Belloni and Chernozhukov 2013, and Belloni et al. 2014a). In this paper I contribute to the literature by establishing properties of the Lasso when the targets of estimation are instead numerous high-dimensional *best linear predictors*.

The remainder of this paper is organized as follows. I give some motivational examples in Section 1.2 and provide an overview of the main results in Section 1.3. Sections 1.4 and 1.5 contain a formal presentation of the results for semiparametric CMR and high-dimensional UMR models, respectively, and the assumptions under which they are proven. Results on the properties of the Lasso for estimation of a multitude of best linear predictors, and details on implementation have all been relegated to the appendices so as not to interrupt the flow of the paper. The appendices also contain additional motivational examples, verification of assumptions in examples, and some extensions to the settings studied in the main text.

Proofs of main results are in the appendices, while proofs of supporting lemmas may be found in the supplement.

Notation

Section 1.4 concerns independent and identically distributed (i.i.d.) data $\{Z_i\}_1^\infty$ with Z denoting a generic element. For these sections, $c, C, C_1, C_2, \ldots$ denote finite, positive constants independent of n, which may change from place to place. Here $a \lesssim b$ means that $a \leqslant Cb$, and $a \lesssim_{\mathrm{P}} b$ means that $a = O_{\mathrm{P}}(b)$.

In Section 1.5 I work with triangular array data $\{\{Z_{i,n}\}_{i=1}^n\}_{n=1}^\infty$ defined on some common probability space. For each $n \in \mathbf{N}$, $Z_{i,n}, i \in \{1, \ldots, n\}$, are i.i.d. across i, but their common law may change with n. Consequently, all objects that are defined using the distribution of $Z_{i,n}$ are implicitly indexed by the sample size n, but I omit the index n in what follows to simplify notation and let Z denote a generic element. For these sections $a \lesssim b$ is reserved for $a \leqslant Ab$, where A denotes an absolute constant.

Throughout I use the average notation $\mathbb{E}_n[f(Z_i)] \coloneqq n^{-1}\sum_{i=1}^n f(Z_i)$, i.e., $\mathbb{E}_n(\cdot)$ abbreviates $n^{-1}\sum_{i=1}^n(\cdot)$. For $f : \mathbf{R}^K \to \mathbf{R}^L$ differentiable, $\partial_{x^\top} f$ is short for the $L \times K$ matrix of partial derivatives $\partial f_l / \partial x_k$. For a symmetric, real matrix A, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the smallest and largest eigenvalues of A, respectively. The ℓ_1 norm and ℓ_2 (i.e., Euclidean) norm of vectors are denoted by $\|\cdot\|_1$ and $\|\cdot\|$, respectively. The "ℓ_0 norm" $\|\cdot\|_0$ is given by the number of nonzero components of a vector, while $\|\cdot\|_\infty$ denotes the maximal absolute element of a vector. The empirical L^2-norm $L^2(\mathbb{P}_n)$ is given by $\|f\|_{\mathbb{P}_n,2} \coloneqq \{\mathbb{E}_n[f(Z_i)^2]\}^{1/2}$ and for a function $f : \mathcal{X} \to \mathbf{R}$ I write $\|f\|_{\mathcal{X}} \coloneqq \sup_{x \in \mathcal{X}} |f(x)|$. Given a vector $\delta \in \mathbf{R}^p$ and a set of indices $T \subset \{1, \ldots, p\}$, I write δ_T for the vector satisfying $\delta_{Tj} = \delta_j$ if $j \in T$ and zero otherwise. Complements are relative to the index set: $T^c \coloneqq \{1, \ldots, p\} \backslash T$. I denote $a \vee b \coloneqq \max\{a, b\}$ and $a \wedge b \coloneqq \min\{a, b\}$.

1.2 Motivational Examples

The following example illustrates how CMRs involving nonparametrically specified CEFs may arise from an economic model. Section 1.B contains additional motivational examples such as partial and high-dimensional linear regression and discrete choice under uncertainty.

Example 1.1 (**An Entry Game with Incomplete Information**). Consider a simple entry game where J firms consider entering a particular market. These firms may be thought of as major US airlines deciding on whether to enter a particular large metropolitan airport. Bajari, Hong, Krainer, and Nekipelov (2010) analyzed entry and more general static, discrete games under incomplete information. This example builds on their Section 2.1. Each firm $j \in \{1, \ldots, J\}$ must choose an action $a_j \in \{0, 1\}$.[2] Let $A_j = 1$ denote the decision to enter a particular market and and $A_j = 0$ the decision not to enter the same market. At the time of decision, the payoff- and belief-relevant state variables V and W, respectively, are publicly known. In the airline industry example, these variables could include the nearby population or publicly available measures of airline operating costs. Each firm also holds private information $\varepsilon_j \coloneqq (\varepsilon_j(0), \varepsilon_j(1))$, which may be thought of as capturing shocks to the firm's own profitability. Firm j's (ex post) payoff from choosing $a_j \in \{0, 1\}$ is

$$u_j(a_j, a_{-j}, v) = \pi_j(a_j, a_{-j}, v) + \varepsilon_j(a_j).$$

Note that the payoff of a firm j is allowed to depend on not only their own action a_j but also on the actions of others, $a_{-j} \coloneqq (a_1, \ldots, a_{j-1}, a_{j+1}, \ldots, a_J)$. This feature makes the model one of strategic interaction. Let the private information $\varepsilon_j \coloneqq (\varepsilon_j(0), \varepsilon_j(1))$ be distributed according to some cdf $F(\varepsilon_0, \varepsilon_1; \gamma_0)$ independently across firms and independently of the public state variables, where F is known up to the parameter γ_0. Parameterize the "deterministic" part of the payoff as

$$\pi_j(a_j, a_{-j}, v) = \begin{cases} v^\top \theta_0 + \delta_0 \sum_{k \neq j} a_k, & a_j = 1 \\ 0, & a_j = 0, \end{cases}$$

thus normalizing this part of the payoff zero when the firm chooses not to enter. Suppose that the researcher observes actions $A = (A_1, \ldots, A_J)$ of all firms and the public state variables (V, W). Suppose further that the game is played according to a Bayesian Nash Equilibrium (BNE), such that every firm maximizes their expected payoff given their beliefs, and everyone's beliefs turn out to be correct. Then in a BNE the conditional entry probability of firm j equals

$$\mathrm{P}(A_j = 1 | V, W) = G\Big(V^\top \theta_0 + \delta_0 \sum_{k \neq j} \mathrm{P}(A_k = 1 | W); \gamma_0\Big),$$

where $G(u; \gamma_0) = \int \mathbf{1}(\varepsilon_0 < u + \varepsilon_1)\, \mathrm{d}F(\varepsilon_0, \varepsilon_1; \gamma_0)$. The previous display rearranges to the CMR

$$\mathrm{E}\Big[A_j - G\Big(V^\top \theta_0 + \delta_0 \sum_{k \neq j} \mathrm{P}(A_k = 1 | W); \gamma_0\Big)\Big| V, W\Big] = 0. \tag{1.2.1}$$

[2] This example may to some extent be extended to allow for a more general static discrete game with incomplete information similar to Bajari et al. (2010, Section 2).

The implied residual for firm j, $A_j - G(V^\top\theta + \delta\sum_{k\neq j}\mathrm{P}\left(A_k = 1|\, V, W\right); \gamma)]$ depends on the $J-1$ conditional expectations $\mathrm{E}\left(A_k|\, W\right) = \mathrm{P}\left(A_k = 1|\, W\right)$, i.e., the conditional entry probabilities of firm j's competitors. In the special case where $\{\varepsilon_j\left(a_j\right)\}_{a_j,j}$ are distributed Type 1 extreme value independently across both actions and firms and independently of the public state variables, the conditional entry probability of firm j takes the logit form,

$$\mathrm{P}\left(A_j = 1|\, V, W\right) = \mathrm{logistic}\Big[V^\top\theta_0 + \delta_0\sum_{k\neq j}\mathrm{P}\left(A_k = 1|\, W\right)\Big],$$

where $\mathrm{logistic}\left(u\right) = \mathrm{e}^u/\left(1+\mathrm{e}^u\right)$. The previous display rearranges to produce the CMR

$$\mathrm{E}\Big\{A_j - \mathrm{logistic}\Big[V^\top\theta_0 + \delta_0\sum_{k\neq j}\mathrm{P}\left(A_k = 1|\, W\right)\Big]\Big|V, W\Big\} = 0. \tag{1.2.2}$$

(See Appendix 1.B and in particular (1.B.3) for the analogous expression in a single-agent discrete choice model.) Regardless of the choice of distribution, one may have misspecified the payoff function, omitted payoff- or belief-relevant state variables, or settled on the wrong distribution for the private information. In addition, one may have chosen to work with an inadequate equilibrium concept.

The following example is a high-dimensional analog of Example 1.1.

Example 1.2 (**A High-Dimensional Model of Entry with Incomplete Information**). Suppose that instead of maximizing their expected payoff, each firm maximizes their *projected* payoff given very many state variables (V, W). Projected payoff maximization may occur, for example, if firms are boundedly rational. Then in a "Bayesian" Nash equilibrium where all firms maximize their projected payoffs subject to their beliefs, and all beliefs turn out correct, the conditional entry probability of firm j takes the form

$$\mathrm{P}\left(A_j = 1|\, V, W\right) = G\Big(V^\top\theta_0 + \delta_0\sum_{k\neq j}L\left(A_k|\, W\right); \gamma_0\Big),$$

where $L\left(A_k|\, W\right) \coloneqq W^\top h_{k*}$ is the best linear predictor of A_k given W with coefficients given by $h_{k*} \coloneqq [\mathrm{E}(WW^\top)]^{-1}\mathrm{E}(WA_k)$. Interacting the implied residual by the vector $X = (V^\top, W^\top)^\top$ of q instruments, we arrive at the UMRs

$$\mathrm{E}\Big\{\Big[A_j - G\Big(V^\top\theta_0 + \delta_0\sum_{k\neq j}W^\top h_{k*}; \gamma_0\Big)\Big]X\Big\} = \mathbf{0}_{q\times 1}.$$

1.3 Overview

In this section of the paper I informally present the test procedures developed in this paper and provide an overview of the main results. Sections 1.4 and 1.5 contain a more technical presentation of these results and the assumptions under which they are proven.

1.3.1 Overview: Semiparametric Conditional Moments

The null hypothesis is the CMR $\mathrm{E}\left[\rho\left(Z,\beta_*,h_*\left(W\right)\right)\middle|\, X\right]=0$,[3] where ρ denotes a residual function capturing the econometric model, Z denotes all observables, $\beta_*\in\mathbf{R}^d$ a parameter, $h_*\left(W\right)\coloneqq\mathrm{E}\left(Y\middle|\, W\right)$ is a nonparametrically specified conditional expectation depending on a vector W of regressors, and X is a vector of conditioning (instrumental) variables including W. Both β_* and X (thus W) are treated as objects of fixed and low dimension. For the ease of presentation, h_* is here treated as real-valued, although the theory readily extends to cover the case where $h_*\left(W\right)$ represents a vector of conditional expectation functions (cf. Section 1.F.1).

Following Bierens's (1982) approach to specification testing in regression models, I convert the single CMR $\mathrm{E}\left[\rho\left(Z,\beta_*,h_*\left(W\right)\right)\middle|\, X\right]=0$ into a possibly infinite collection of UMRs, $\mathrm{E}\left[\rho\left(Z,\beta_*,h_*\left(W\right)\right)\omega\left(t,X\right)\right]=0, t\in\mathcal{T}$, where ω and $\mathcal{T}$ denote a weight function and index set suitably chosen by the researcher (see Stinchcombe and White, 1998). Applying a functional to the function $t\mapsto\mathrm{E}\left[\rho\left(Z,\beta_*,h_*\left(W\right)\right)\omega\left(t,X\right)\right]$, one may aggregate these UMRs. For simplicitly, I focus on the case where $\mathcal{T}$ is operated out by integrating the squared deviations from zero against an appropriately chosen continuous distribution function μ on $\mathcal{T}$, such that the null holds if and only if

$$\left\|\mathrm{E}\left[\rho\left(Z,\beta_*,h_*\left(W\right)\right)\omega\left(\cdot,X\right)\right]\right\|_{\mu,2}^2\coloneqq\int_{\mathcal{T}}\left\{\mathrm{E}\left[\rho\left(Z,\beta_*,h_*\left(W\right)\right)\omega\left(t,X\right)\right]\right\}^2\mathrm{d}\mu\left(t\right)=0.$$

Given a random sample $\{Z_i\}_1^n$ of size n and estimators $\widehat{\beta}$ and $\widehat{h}$, the previous display suggests basing a test of the null hypothesis on the feasible sample analog

$$T_n\coloneqq\int_{\mathcal{T}}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n\rho(Z_i,\widehat{\beta},\widehat{h}\left(W_i\right))\omega\left(t,X_i\right)\right]^2\mathrm{d}\mu\left(t\right),$$

where I have scaled by $\sqrt{n}$ in anticipation of an application of a central limit theorem. When $\rho(Z_i,\widehat{\beta},\widehat{h}\left(W_i\right))$ is the nonlinear least squares residual $Y_i-f(X_i,\widehat{\beta})$, the previous display becomes the ICM test statistic of Bierens and Ploberger (1997).

[3] Throughout this section I omit the qualifier "with probability one" in making conditional statements.

I take a series approach to estimation of h_*, which is motivated by the fact that h_* is a CEF with a small number of arguments. Its estimator $\widehat{h}$ may therefore be constructed using standard regression tools.

I show that the stochastic process $n^{-1/2}\sum_{i=1}^{n}[\rho(Z_i, \widehat{\beta}, \widehat{h}\,(W_i))\omega\,(\cdot, X_i)]$ driving the behavior of the test statistic T_n is asymptotically equivalent to a process $\sqrt{n}\mathbb{E}_n\,[f_*\,(\cdot, Z_i)]$, where the summand $f_*\,(t, Z_i)$ involves two adjustment terms due to estimation of β_* and h_*. The probabilistic behavior of T_n may therefore be approximated by that of $\|n^{-1/2}\sum_{i=1}^{n} f_*\,(\cdot, Z_i)\|_{\mu,2}^2$. Under the null, by means of a functional central limit theorem (FCLT) I show that $n^{-1/2}\sum_{i=1}^{n} f_*(\cdot, Z_i)$ converges in distribution to a zero-mean Gaussian process G_0. The continuous mapping theorem then implies

$$T_n \xrightarrow{d} \|G_0\|_{\mu,2}^2 = \int_{\mathcal{T}} G_0\,(t)^2\,\mathrm{d}\mu\,(t)\,.$$

The limiting null distribution cannot be tabulated. To obtain critical values I make use of a multiplier bootstrap. To fix ideas, let $\xi_i, i \in \{1, 2, \dots\}$, be i.i.d. standard normal and independent of the data, and define the multiplier process

$$G_n^*\,(t) \coloneqq \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i\,\{f_*\,(t, Z_i) - \mathbb{E}_n\,[f_*\,(t, Z_i)]\}\,, \quad t \in \mathcal{T},$$

Given that $\sqrt{n}\mathbb{E}_n\,[f_*\,(\cdot, Z_i)]$ satisfies a FCLT, so does the multiplier process (conditional on the data). Under the null, the integrated squares of the multiplier process converge in distribution to those of the null process. Consequently, for a given significance level $\alpha \in (0, 1)$ one may use as a critical value

$$c_n^*\,(\alpha) \coloneqq (1-\alpha)\text{-quantile of } \|G_n^*\|_{\mu,2}^2 \text{ given } \{Z_i\}_1^n\,.$$

However, the function f_* is generally unknown, which renders $c_n^*\,(\alpha)$ infeasible. I show that replacing f_* by a feasible analog $\widehat{f}$ is asymptotically equivalent to knowing f_*. As a result, one may construct a feasible critical value $\widehat{c}\,(\alpha)$ using the previous two displays by (1) substituting $\widehat{f}$ for f_*, and (2) simulating the multipliers $\{\xi_i\}_1^n$ holding the data constant. The main results of this section are that the test that rejects the null if and only if $T_n > \widehat{c}\,(\alpha)$ is asymptotically of size α and consistent.

1.3.2 Overview: High-Dimensional Unconditional Moments

The null hypothesis is that $\mathrm{E}[\rho(Z, \beta_*, W^\top h_*)X] = \mathbf{0}_{q\times 1}$, where ρ denotes a residual function capturing the econometric model, Z denotes all observables for a single observation,

$\beta_* \in \mathbf{R}^d$ is a low-dimensional parameter, and $W^\top h_*$ is the best linear predictor of Y based on a random vector W of p "regressors," and X is a random vector of q instruments.[4] The moments $\mathrm{E}[\rho(Z, \beta_*, W^\top h_*)X]$ are "high-dimensional" in two ways as I allow both the number of regressors p and the number of instruments q to grow with as well as potentially greatly exceed the sample size n available to the researcher. Best linear predictors may occur in econometric models due to bounded rationality or as flexible, linear approximation to complex, nonlinear conditional expectation functions. Although an extension to multiple best linear predictors is theoretically possible (see Section 1.F.2), for simplicity of notation of notation I here consider the case of a single best linear predictor.

Given an i.i.d. sample $\{Z_i\}_1^n$ of size n and estimators $\widehat{\beta}$ and $\widehat{h}$, the null hypothesis may be heuristically tested by inspecting whether $n^{-1}\sum_{i=1}^n \rho(Z_i, \widehat{\beta}, W_i^\top \widehat{h})X_i \approx \mathbf{0}_{q\times 1}$, or, equivalently, using the maximal deviation from zero, whether

$$\max_{1\leqslant k\leqslant q} \left| \frac{1}{n} \sum_{i=1}^n \rho(Z_i, \widehat{\beta}, W_i^\top \widehat{h}) X_{ik} \right| \approx 0.$$

An intuitively appealing test statistic $\widetilde{T}$ is therefore defined by

$$\widetilde{T} \coloneqq \max_{1\leqslant k\leqslant q} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \rho(Z_i, \widehat{\beta}, W_i^\top \widehat{h}) X_{ik} \right|,$$

where I have scaled by $\sqrt{n}$ in anticipation of an application of some central limit theorem to be discussed below.

When h_* is high-dimensional, the number of free parameters exceeds the sample size, and one must necessarily make use of some machine learning method (e.g., regularization methods such as the Lasso or Ridge regression) to estimate h_*. In the context of two-step semiparametric estimation, Belloni et al. (2012) (henceforth: BCCH) and Belloni, Chernozhukov, and Hansen (2014b) have shown that when using a machine learning estimator in a first step, in order to obtain valid inference about parameters of interest it is important to use locally robust moments. Moments are said to be *locally robust* (to the first step) when they have a zero derivative with respect to the first step (see, e.g., Chernozhukov, Escanciano, Ichimura, and Newey, 2016).[5] The findings of BCCH and Belloni et al. (2014b) apply to the

[4] These "regressors" may be technical in nature in in the sense of being generated as (many) transformations of underlying basic regressors. Similarly, the instruments may have been generated by underlying conditioning (i.e., instrumental) variables.

[5] Some authors refer to locally robust moments as *debiased, first-order insensitive, immunized, Neyman orthogonalized,* or simply *orthogonalized* moments (see Chernozhukov et al. 2016 and Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins 2017). I use these terms synonymously.

present context of specification testing, where the "parameter" of interest are the moments themselves. I therefore transform the original moment functions to ensure local robustness.

Locally robust moment functions (LRMFs) may be constructed by adding to the original moment functions terms that adjust for estimation in a first step. In this paper the adjustments are done moment by moment.[6] The LRMFs thus created are equal in mean to the original moment functions, which makes them equally suitable for specification testing. However, the resulting orthogonalized moments are less insensitive to estimation of h_*.

Given that each of the very many moments are adusted for estimation of very many parameters, the orthogonalization procedure leading to the LRMFs introduces a high-dimensional number of nuisance parameters to be estimated.[7] However, the LRMFs are constructed in a manner that also ensures local robustness with respect to these additional nuisance parameters.

Denote the q LRMFs by $\psi_k(z, \beta_*, w^\top h_*, w^\top \mu_{k*}), k \in \{1, \ldots, q\}$, where μ_{k*} denotes moment-specific orthogonalization parameters. Further endowed with an estimator of the μ_{k*}'s, the null hypothesis may now be tested using the locally robust test statistic

$$T \coloneqq \max_{1 \leqslant k \leqslant q} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_k(Z_i, \widehat{\beta}, W_i^\top \widehat{h}, W_i^\top \widehat{\mu}_k) \right| .$$

In this paper I estimate both h_* and μ_* using Lasso procedures. Under some assumptions, which include an approximate sparsity condition, I show that that the probabilistic behavior of T may be approximated by that of a random variable T_* taking the form

$$T_* = \max_{1 \leqslant k \leqslant q} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} f_{k*}\left(Z_i\right) \right| ,$$

where each summand $f_{k*}(Z)$ has finite variance and is mean-zero under the null. The finite-sample distribution of T_* cannot be tabulated due to its dependence on the generally unknown f_{k*}'s. To obtain critical values I therefore employ a Gaussian multiplier bootstrap.

To fix ideas, let $\{\xi_i\}_1^n$ denote i.i.d. standard normal random variables independent of the data and define the Gaussian-symmetrized version $\mathcal{W}_*$ of T_* by

$$\mathcal{W}_* \coloneqq \max_{1 \leqslant k \leqslant q} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} f_{k*}\left(Z_i\right) \xi_i \right| .$$

[6] See Chernozhukov et al. (2017) for an alternative orthogonalization procedure that adjusts all moments simultaneously.

[7] The same comment applies to the orthogonalization procedures in Chernozhukov et al. (2017).

Under the null, T_* equals the maximum of an exact average of mean-zero vector. I may therefore rely on a Gaussian approximation, or "high-dimensional central limit theorem," and approximate quantiles of T_* by the corresponding conditional quantiles of $\mathcal{W}_*$ conditional on the data $\{Z_i\}_1^n$ (see Chernozhukov, Chetverikov, and Kato, 2013). Hence, *if* the f_{k*}'s were known, then one may obtain a critical value by simulating the multipliers $\{\xi_i\}_1^n$ and calculated the desired quantile of $\mathcal{W}_*$ holding the data constant. This method for obtaining a critical value is sometimes referred to as the *Gaussian multiplier* (or *Wild*) *bootstrap.*

While the f_{k*}'s are unknown in general, a feasible critical value arises from replacing the unknown f_{k*}'s by consistent estimators $\widehat{f}_k$'s. For given $\widehat{f}_k$'s, one may define the feasible analog $\mathcal{W}$ of $\mathcal{W}_*$ by

$$\mathcal{W} \coloneqq \max_{1 \leqslant k \leqslant q} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widehat{f}_k \left(Z_i\right) \xi_i \right| .$$

Hence, for a given significance level $\alpha \in (0,1)$, a feasible critical value $c_{\mathcal{W}}(\alpha)$ is given by

$$c_{\mathcal{W}}(\alpha) \coloneqq (1-\alpha)\text{-quantile of } \mathcal{W} \text{ conditional on } \left\{Z_i\right\}_1^n .$$

Building on results in Chernozhukov, Chetverikov, and Kato (2013) for the Gaussian multiplier bootstrap, I show that this critical value leads to *uniform size control* in the sense that as $n \to \infty$ and possibly $p = p_n \to n$ and $q = q_n \to \infty$,

$$\sup_{\alpha \in (0,1)} \left| \mathrm{P}\left(T > c_{\mathcal{W}}(\alpha) ; \mathrm{H}_0\right) - \alpha \right| \leqslant C n^{-c} \to 0$$

for some $c > 0$ and $C > 0$ independent of n. In particular, the previous display implies that the test that rejects if and only if $T > c_{\mathcal{W}}(\alpha)$ is asymptotically of correct size. For given estimators $\{\widehat{f}_k\}_1^q$, the critical value $c_{\mathcal{W}}(\alpha)$ may be calculated via simulation of the Gaussian multipliers $\{\xi_i\}_1^n$. An novel feature of this size control result is that it does not rely on knowledge of the limiting null distribution of T. In fact, the test yields approximately correct size in finite sample even in settings where the limiting null distribution of T is complicated, unknown, or fails to exist (even after suitable standardization).

To quantify the degree to which the null is violated, define

$$v_q \coloneqq \max_{1 \leqslant k \leqslant q} \left| \mathrm{E}\left[\rho\left(Z, \beta_*, W^\top h_*\right) X_k\right] \right| .$$

(Here v connotes "violation.") I show that the test that rejects if and only if $T > c_{\mathcal{W}}(\alpha)$ is consistent for any alternative satisfying $v_q^{-1} \ln(q) / \sqrt{n} \to 0$. Failure of the condition

$v_q^{-1}\ln(q)/\sqrt{n} \to 0$ may be interpreted as the alternative being “too local” to the null or that the instruments X are “too weak.”

1.4 Semiparametric Conditional Moment Models

In this section I formally present my main results on specification testing in a class of semiparametric CMR models.

1.4.1 Null Hypothesis

The *null hypothesis* is

$$\mathrm{H}_0 : \exists \beta_0 \in \mathcal{B} \text{ s.t. } \mathrm{E}\left[\rho\left(Z, \beta, h_*(W)\right) \middle| X\right] = 0 \text{ a.s. at } \beta = \beta_0,$$

where ρ is a residual function which depends on data Z, a finite-dimensional parameter β belonging to a given parameter space $\mathcal{B} \subset \mathbf{R}^d$, and a vector of conditional expectations $h_*(W) \coloneqq \mathrm{E}(Y|W)$, and X is a collection of conditioning variables, which includes W as a subvector. The *alternative hypothesis* is the negation of the null,

$$\mathrm{H}_1 : \forall \beta \in \mathcal{B} : \mathrm{P}\left(\mathrm{E}\left[\rho\left(Z, \beta, h_*(W)\right) \middle| X\right] = 0\right) < 1.$$

The vector Z includes both Y and X (and thus W) as subvectors. The dependence on elements of X in ρ may be trival, thus allowing for the presence of excluded exogenous variables, i.e., “instrumental” variables. The model, which is implicit in the residual, may be semiparametric as long as the infinite-dimensional component is composed by CEFs. This structure is fairly common as illustrated by the range of examples in Sections 1.2 and 1.B. While econometric models typically involve multiple conditional expectations (see Examples 1.1, 1.7 and 1.8), to simplify the presentation and ease notation I will here focus on the case where $h_*(W)$ is scalar valued. The discussion of vector-valued $h_*(W)$ is deferred to Section 1.F.1.

1.4.2 Recasting the Null Hypothesis

In this section, through a sequence of steps, I transform the null hypothesis into an equivalent expression which suggests a test statistic.

1.4.2.1 Recasting using Pseudo-True Parameters

As a first step, let β_* in $\mathcal{B}$ be such that (a) β_* may be consistently estimated (irrespective of the null being true or not), and (b) $\beta_* = \beta_0$ under the null. Because β_* agrees with β_0 under the null, β_* is called a *pseudo-true parameter*, or the *pseudo-truth* for short.[8] Using the pseudo-truth we may rewrite the null as

$$\mathrm{H}_0 : \mathrm{E}\left[\rho\left(Z, \beta_*, h_*\left(W\right)\right)\middle|\, X\right] = 0 \text{ a.s.}$$

The following example illustrates how one may obtain a pseudo-true parameter.

Example 1.1 (continued) In the entry game, denote $X \coloneqq (V, W)$ and let $r\left(X\right)$ be a $(d_\theta + 1 + d_\gamma)$-vector of instruments generated by the state variables V and W. Appealing to the CMR (1.2.1), a pseudo-truth $\beta_* \coloneqq (\theta_*, \delta_*, \gamma_*)$ may be taken as the assumed unique root of the map

$$(\theta, \delta, \gamma) \mapsto \mathrm{E}\Big\{\Big[A - G\Big(V^\top\theta + \delta\sum_{k\neq j}\mathrm{P}\left(A_k = 1\middle|\, W\right); \gamma\Big)\Big] r\left(X\right)\Big\}, \quad (\theta, \delta, \gamma) \in \mathbf{R}^{d_\theta + 1 + d_\gamma}.$$

A root of such a map exists under regularity conditions. Uniqueness amounts to an identification condition. To see that β_* is pseudo-true, suppose that the null hypothesis holds for this model. Then there exists $\beta_0 \coloneqq (\theta_0^\top, \delta_0, \gamma_0^\top)^\top$ such that

$$\mathrm{E}\Big[A - G\Big(V^\top\theta_0 + \delta_0\sum_{k\neq j}\mathrm{P}\left(A_k = 1\middle|\, W\right); \gamma_0\Big)\Big| V, W\Big] = 0.$$

The uniqueness assumption and iterated expectations therefore shows that $\beta_* = \beta_0$ under the null. Building on the general framework developed by Newey (1990), Bajari et al. (2010) provide conditions under which a two-step GMM estimator of β_* based on nonparametric (sieve) first-step estimation of conditional choice probabilities is $\sqrt{n}$-asymptotically normal.

Remark 1.1. The assumption of the existence of a *unique* pseudo-true parameter β_* implicitly invokes a *point* identification condition for β_0 under the null. A weaker condition would be to require that the parameterization of the model is *partially identified* under the null as in, e.g., Santos (2012), who studied inference in nonparametric instrumental variables with partial identification. If β_0 is allowed to be *partially* identified under the null, and belongs to the potentially non-singleton identified set $\mathcal{B}_0 \subset \mathcal{B}$, then one needs to find a potentially

[8] There may be more than one option for a pseudo-true parameter, cf. the continuations of Examples 1.8 and 1.1 below. Here I assume that the researcher has settled on a particular option.

non-singleton subset $\mathcal{B}_*$ of $\mathcal{B}$ such that $\mathcal{B}_* = \mathcal{B}_0$ under the null. While I consider allowing for partial identification an important extension, I do not pursue it at present.

1.4.2.2 Recasting using the Nuisance Parameter Approach

To make further progress towards operationalizing the null, let $\mathcal{X} \coloneqq \operatorname{supp}(X)$ denote the support of the conditioning variables X, and let $\omega : \mathcal{T} \times \mathcal{X} \to \mathbf{R}$ be a known function with the property that for any integrable random variable V,

$$\mathrm{E}\left(V \middle| X\right) = 0 \text{ a.s. if and only if } \mathrm{E}\left[V\omega\left(t, X\right)\right] = 0 \text{ for all } t \in \mathcal{T}. \tag{1.4.1}$$

One may then express the null as

$$\mathrm{H}_0 : \; \mathrm{E}\left[\rho\left(Z, \beta_*, h_*\left(W\right)\right)\omega\left(t, X\right)\right] = 0 \text{ for all } t \in \mathcal{T}. \tag{1.4.2}$$

The weight function ω allows us to transform a single CMR into a possibly infinite collection of UMRs indexed by the "nuisance parameter" t through the weight $\omega(t, X)$. While an extension to unknown but consistently estimable weight functions and nuisance parameter spaces is possible, I treat both of these quantities as known.

Remark 1.2 (On the nuisance approach, direct and indirect tests). The "nuisance parameter approach" dates back to Bierens (1982) and was considered by, e.g., Bierens (1990); Bierens and Ploberger (1997); Stinchcombe and White (1998); and Santos (2012). In the language of Stinchcombe and White (1998), $\{x \mapsto \omega(t, x) | \, t \in \mathcal{T}\}$ is a collection of "test functions," which are chosen such that they have the ability to "reveal" departures from zero of the function $\mathrm{E}[\rho(Z, \beta_*, h_*(W))|X = \cdot]$ under the inner product $\langle f_1, f_2 \rangle = \mathrm{E}[f_1(X)f_2(X)]$.

A *direct* test of the null hypothesis (1.4.2) involves estimating the conditional expectation of the residual $x \mapsto \mathrm{E}[\rho(Z, \beta_*, h_*(W))|X = x]$ and checking whether or not the result is the zero function. This method is sometimes referred to as the "nonparametric approach" (Stinchcombe and White, 1998)) or the "kernel-based approach" (Fan and Li, 2000). In contrast, the nuisance approach focuses on estimating the residual function $z \mapsto \rho\left(z, \beta_*, h_*\left(w\right)\right)$ itself, which yields an *indirect* test of the null hypothesis. Focus on the residual function itself instead of its conditional expectation is justified by the law of iterated expectations

$$\mathrm{E}\left[\rho\left(Z, \beta_*, h_*\left(W\right)\right)\omega\left(t, X\right)\right] = \mathrm{E}\left\{\mathrm{E}\left[\rho\left(Z, \beta_*, h_*\left(W\right)\right)\middle| X\right]\omega\left(t, X\right)\right\} \text{ for all } t \in \mathcal{T}$$

and property (1.4.1) of the weight function ω. The nonparametric approach works because it rests on a class of functions which can approximate any function. The nuisance approach works because it rests on a class of functions whose span can approximate any function

(Stinchcombe and White, 1998, p. 298).

Under suitable asssumptions, the function $t \mapsto \mathrm{E}\left[\rho\left(Z, \beta_*, h_*\left(W\right)\right)\omega\left(t, X\right)\right]$ is square integrable with respect to some absolutely continuous, strictly positive, finite measure μ on $\mathcal{T}$. Equation (1.4.2) then allows us to recast the null as

$$\mathrm{H}_0 : \int_{\mathcal{T}} \left\{\mathrm{E}\left[\rho\left(Z, \beta_*, h_*\left(W\right)\right)\omega\left(t, X\right)\right]\right\}^2 \mathrm{d}\mu\left(t\right) = 0. \tag{1.4.3}$$

1.4.3 Test Statistic

Granted a random sample $\{Z_i\}_1^n$, (1.4.3) suggests statistics of the form:

$$\int_{\mathcal{T}} \left\{\mathbb{E}_n\left[\rho\left(Z_i, \beta_*, h_*\left(W_i\right)\right)\omega\left(t, X_i\right)\right]\right\}^2 \mathrm{d}\mu\left(t\right). \tag{1.4.4}$$

Such statistics were considered by Bierens (1982; 1990) and were by Bierens and Ploberger (1997) later named Integrated Conditional Moment (ICM) test statistics. These statistics also resemble the Cramér-von Mises criterion for judging the goodness of fit of a given CDF compared to the empirical distribution function. Different options for the probability measure μ are available. However, Andrews and Ploberger (1994) have shown that the uniform probability measure on $\mathcal{T}$ is optimal in the sense of maximizing average local power (as defined by the same authors). Applying a different functional to $t \mapsto \mathbb{E}_n\left[\rho\left(Z_i, \beta_*, h_*\left(W_i\right)\right)\omega\left(t, X_i\right)\right]$ than an L^2-norm would yield an alternative test statistic. For example, the supremum norm implies the equally valid test statistic

$$\sup_{t \in \mathcal{T}} \left|\sqrt{n}\mathbb{E}_n\left[\rho\left(Z_i, \beta_*, h_*\left(W_i\right)\right)\omega\left(t, X_i\right)\right]\right|.$$

However, some choices of $(\mu, \mathcal{T})$ allow for calculation of the test statistic in closed form.

While the statistic in (1.4.3) cannot be used for testing due to its dependence on the unknowns β_* and h_*, further endowed with estimators $\widehat{\beta}$ and $\widehat{h}$, one may construct the test statistic

$$T_n \coloneqq \int_{\mathcal{T}} \left\{\sqrt{n}\mathbb{E}_n\left[\rho(Z_i, \widehat{\beta}, \widehat{h}\left(W_i\right))\omega\left(t, X_i\right)\right]\right\}^2 \mathrm{d}\mu\left(t\right). \tag{1.4.5}$$

To control the influence of estimation of $\widehat{\beta}$, I make

Assumption 1.1 (**Parametric Estimator**). *For each $n \in \mathbf{N}, \widehat{\beta}$ is a random element of $\mathcal{B} \subset \mathbf{R}^d$, where $\mathcal{B}$ is a compact subset of $\mathbf{R}^d$. Further, there exists $s_* : \mathcal{Z} \to \mathbf{R}^d$ such that*

$$\sqrt{n}(\widehat{\beta} - \beta_*) = \sqrt{n}\mathbb{E}_n\left[s_*\left(Z_i\right)\right] + o_{\mathrm{P}}(1), \tag{1.4.6}$$

where β_* is interior to $\mathcal{B}$, and $s_*(Z)$ is mean zero and square integrable.

Assumption 1.1 requires that $\widehat{\beta}$ is confined to a compact set and that the centered and scaled estimator $\widehat{\beta}$ is *asymptotically linear* with *influence function* s_*. Given the assumption of asymptotic linearity, $\widehat{\beta}$ must eventually belong to a shrinking neighborhood of β_*, and the assumption of compactness may be relaxed.

Asymptotic linearity is a high-level condition. However, as illustrated by Example 1.3 below, for particular classes of estimators it is possible to obtain asymptotic linearity through more primitive assumptions. While primitive, easy-to-verify conditions are desirable, Assumption 1.1 leaves freedom in choice beyond the two-step GMM estimator of Example 1.3. For example, (1.4.6) allows for other or more general two-step (or multi-step) estimation procedures, such as two-step extremum estimation. Such procedures typically estimate the nonparametric component in a first step, use its estimate to contruct a criterion function, and maximize or minimize over β in order to produce a second-step estimator $\widehat{\beta}$. Specifically, one may let $\widehat{\beta}$ be a sieve minimum distance (SMD) estimator (Ai and Chen, 2003) or a penalized sieve minimum distance (PSMD) estimator (Chen and Pouzo, 2009; 2012).

Example 1.3 (**Asymptotic Linearity of Two-Step GMM**)**.** Suppose that β_* satisfies $\mathrm{E}[m(Z,\beta,h_*(W))] = \mathbf{0}_{d\times 1}$ with $h_*(W) = \mathrm{E}(Y|\, W)$ scalar. Define $\widehat{\beta}$ as the minimizer of $\beta \mapsto \|\mathbb{E}_n[m(Z_i,\beta,\widehat{h}(W_i)]\|^2$, where $\widehat{h}$ is some nonparametric estimator of h_*. The estimator $\widehat{\beta}$ is known as a *two-step GMM estimator* based on a nonparametric first step. Newey (1994, Lemma 5.3) provides conditions under which such a two-step GMM estimator of β_* based on a nonparametric first step is $\sqrt{n}$-asymptotically normal.[9] An inspection of Newey's proof reveals that the same set of conditions yield the slightly stronger result of asymptotic linearity. Specifically, under Newey's conditions

$$\sqrt{n}(\widehat{\beta}-\beta_*) = -\left(M_*^\top M_*\right)^{-1} M_*\sqrt{n}\mathbb{E}_n\left[m\left(Z_i,\beta_*,h_*\left(W_i\right)\right)+\alpha_*\left(Z_i\right)\right]+o_{\mathrm{P}}\left(1\right), \qquad (1.4.7)$$

where $M_* = \mathrm{E}\left[\partial_{\beta^\top} m\left(Z,\beta_*,h_*\left(W\right)\right)\right]$ is a Jacobian term, and α_* is an adjustment to the moment function due to estimation of h_*. Because h_* is a CEF, Newey (1994, Proposition 4) shows that, irrespective of the choice of nonparametric estimator, the adjustment is of the form

$$\alpha_*\left(z\right) = \left[y-h_*\left(w\right)\right]\delta_*\left(w\right), \quad \delta_*\left(W\right) \coloneqq \mathrm{E}\left[\partial_v m\left(Z,\beta_*,h_*\left(W\right)\right)\middle|\, W\right] \in \mathbf{R}^d. \qquad (1.4.8)$$

[9] Newey (1994) studies the more general framework, where the nonparametric component h_* need not be a CEF, and the moment functions may depend on the entire function $h_*(\cdot)$ and not necessarily just their values $h_*(w)$.

The influence function is therefore given by

$$s_*(z) = -\left(M_*^\top M_*\right)^{-1} M_* \left\{m\left(z, \beta_*, h_*(w)\right) + \delta_*(w)\left[y - h_*(w)\right]\right\},$$

with δ_* provided by (1.4.8).[10] When the moment function $m(z, \beta, h_*(w))$ depends on the values of a *vector* of CEFs $h_* \coloneqq (h_{1*}, \ldots, h_{L*})$ given by $h_{\ell *}(w_\ell) \coloneqq \mathrm{E}\left(Y_\ell \middle| W_\ell = w_\ell\right)$, then Newey (1994, p. 1357) shows the total adjustment to the moment function is given by adding up the individual adjustment terms,

$$\begin{aligned}\alpha_*(z) &= \sum_{\ell=1}^{L} \alpha_{\ell *}(z) = \sum_{\ell=1}^{L} \left[y_\ell - h_{\ell *}(w_\ell)\right] \delta_\ell(w_\ell),\\ \delta_{\ell *}(W_\ell) &\coloneqq \mathrm{E}\left[\partial_{v_\ell} m\left(Z, \beta_*, h_*(W)\right) \middle| W_\ell\right],\end{aligned} \tag{1.4.9}$$

where ∂_{v_ℓ} denotes differentiation with respect to the value of $h_{\ell *}$.

In what follows I use (1.4.7) and (1.4.9) to derive the influence function of two-step GMM estimators based on Example 1.1.

Example 1.1 (continued) Denote $X \coloneqq (V, W)$ and suppose for the sake of illustration that the $\varepsilon_j(a_j)$'s are Type 1 Extreme Value distributed independently across firms and actions. Let $r(X)$ be a $(d_\theta+1)$-vector of instruments. Let $\widehat{\beta} \coloneqq (\widehat{\theta}, \widehat{\delta})$ be a two-step GMM estimator based on the moment function $m(z, \beta, h_*(w)) = [a_j - \mathrm{logistic}(v^\top \theta + \delta \sum_{k \neq j} h_{k*}(w))] r(x)$ and some nonparametric estimators of $h_{k*}(W) = \mathrm{E}\left(A_k \middle| W\right) = \mathrm{P}\left(A_k = 1 \middle| W\right), k \neq j$. Using the notation of Example 1.3, differentiation implies that

$$\begin{aligned}M_* &= -\mathrm{E}\Big\{f\Big(V^\top \theta_* + \delta_* \sum_{k \neq j} h_{k*}(W)\Big) r(X) \left[V^\top, \sum_{k \neq j} h_{k*}(W)\right]\Big\},\\ \delta_{k*}(W) &= -\delta_* \mathrm{E}\Big[f\Big(V^\top \theta_* + \delta_* \sum_{k \neq j} h_{k*}(W)\Big) r(X) \Big| W\Big].\end{aligned}$$

where $f \coloneqq \mathrm{logistic}(1 - \mathrm{logistic})$ denotes the partial derivative of the logistic function. Given that the $h_{k*}(W)$'s, $k \neq j$, enter the residual only through their sum, $\delta_{k*}(W)$ does not depend on k. Using (1.4.7) and (1.4.9), it therefore follows that

$$s_*(z) = -\left(M_*^\top M_*\right)^{-1} M_* \Big\{[a_j - \mathrm{logistic}(v^\top \theta + \delta \sum_{k \neq j} h_{k*}(w))] r(x)$$

[10] See also Chen, Linton, and Van Keilegom (2003), who extend Newey's (1994) results on two-step GMM estimation to allow for nonsmooth moment functions.

$$- \delta_* \mathrm{E}\Big[f\Big(V^\top \theta_* + \delta_* \sum_{k \neq j} h_{k*}(W) \Big) r(X) \Big| W = w \Big] \sum_{k \neq j} [a_k - h_{k*}(w)] \Big\}.$$

To control the influence of estimation of h_*, I employ a series approach. For any nonnegative integer k, let

$$w \mapsto p^k(w) \coloneqq (p_{1k}(w), \ldots, p_{kk}(w))^\top$$

be a k-vector of known approximating functions $\{p_{jk} | \, j \in \{1, \ldots, k\}\}$ which may change with k. Then the *series estimator* $\widehat{h} \coloneqq \widehat{h}_{k_n}$ of h_* is the regression function $w \mapsto p^{k_n}(w)^\top \widehat{\pi}$ arising from a regression of Y_i on $p^{k_n}(W_i)$ using observations $i \in \{1, \ldots, n\}$, where $\{k_n\}_1^\infty$ denotes a sequence of positive integers satisfying $k_n \to \infty$ as $n \to \infty$, $\widehat{\pi}$ the regression coefficients

$$\widehat{\pi} \coloneqq \widehat{\pi}_{k_n} \coloneqq \{\mathbb{E}_n[p^{k_n}(W_i) p^{k_n}(W_i)^\top]\}^- \mathbb{E}_n[p^{k_n}(W_i) Y_i],$$

and $(\cdot)^-$ represents the (unique) Moore-Penrose generalized inverse of a matrix.[11] The estimand h_* may be viewed as the (essentially unique) projection of Y onto $\mathcal{G} \coloneqq \{g | \, \mathrm{E}[g(W)^2] < \infty\}$, the space of all measurable functions of $W = w$ with finite mean-square,

$$h_* = \operatorname*{argmin}_{g \in \mathcal{G}} \mathrm{E}\{[Y - g(W)]^2\}.$$

Define $\mathcal{G}_k := \{p^{k\top}\pi | \pi \in \mathbf{R}^k\}$. Under the conditions stated below, each $\mathcal{G}_k$ is a finite-dimensional subset of $\mathcal{G}$. The estimator $\widehat{h}$ is the sample projection onto $\mathcal{G}_{k_n}$,[12]

$$\widehat{h} \in \operatorname*{argmin}_{g \in \mathcal{G}_{k_n}} \mathbb{E}_n\{[Y_i - g(W_i)]^2\}.$$

The idea of series estimation is that $\widehat{h}$ should approximate h_* provided k_n is allowed to grow with the sample size n. Essential to this approximation are the requirements that (i) each $p_{jk}, j \in \{1, \ldots, k\}$, belongs to $\mathcal{G}$, and (ii) that the functions $\{p_{jk} | \, j \in \{1 \ldots, k\}\}$ span $\mathcal{G}$ as k grows without bound, in the sense that for any $g \in \mathcal{G}$, k can be chosen large enough to ensure that there exists a linear form $p^{k\top}\pi \in \mathcal{G}_k$ which is arbitrarily close to g in mean-square. When (i) holds, the coefficients $\widehat{\pi}$ may be viewed as an estimate of

$$\pi_{k_n} \coloneqq \{\mathrm{E}[p^{k_n}(W) p^{k_n}(W)^\top]\}^{-1} \mathrm{E}[p^{k_n}(W) Y] = \{\mathrm{E}[p^{k_n}(W) p^{k_n}(W)^\top]\}^{-1} \mathrm{E}[p^{k_n}(W) h_*(W)],$$

[11] The choice of generalized inverse is asymptotically irrelevant, as the matrix $\mathbb{E}_n[p^k(W_i) p^k(W_i)^\top]$ is asymptotically nonsingular (under the conditions stated below).

[12] Under the conditions stated below, the problem "minimize $\mathbb{E}_n\{[Y_i - g(W_i)]^2\}$ subject to $g \in \mathcal{G}_{k_n}$" will asymptotically have a unique solution.

i.e., the coefficients arising from the mean-square projection $h_{k_n} \coloneqq p^{k_n \top} \pi_k$ of h_* onto $\mathcal{G}_{k_n}$, and $\widehat{h}$ estimates h_{k_n}. Under (ii) h_{k_n} approximates h_*, so when both (i) and (ii) hold, $\widehat{h}$ ought to be close to h_*. For detailed discussions of the properties of least-squares series estimators, see Newey (1995; 1997), Chen (2007), and Belloni, Chernozhukov, Chetverikov, and Kato (2015).

1.4.4 Limiting Behavior of Test Statistic

In order to characterize the asymptotic behavior of T_n, I impose the following assumption on the choice of weight function.

Assumption 1.2 (**Weight Function**). *The function $\omega : \mathcal{T} \times \mathcal{X} \to \mathbf{R}$ is continuous and bounded, and has the property (1.4.1) for some nonempty, compact subset $\mathcal{T}$ of $\mathbf{R}^{d_t}$. Moreover, for each $x \in \mathcal{X}, t_1, t_2 \in \mathcal{T}, |\omega(t_1, x) - \omega(t_2, x)| \leqslant C \|t_1 - t_2\|$.*

Bierens (1990) showed that if X is a bounded random variable, then (1.4.1) holds for $\omega(t, x) = \exp\left(x^\top t\right)$ provided $\mathcal{T} \subset \mathbf{R}^{d_x}$ is of positive Lebesgue measure (e.g., $\mathcal{T} = [0,1]^{d_x}$), where d_x denotes the dimension of X. Bierens also showed that the boundedness requirement is innocuous: for any X unbounded we may choose $\omega(t, x) \coloneqq \exp[\Phi(x)^\top t]$ for some bounded, one-to-one transformation $\Phi : \mathbf{R}^{d_x} \to \mathbf{R}^{d_x}$. The one-to-one property of Φ ensures that conditioning on X and $\Phi(X)$ are equivalent, i.e., there is no “loss of information” in using Φ. An application of the mean value theorem shows that the Lipschitz requirement in Assumption 1.2 holds for any such Φ. Stinchcombe and White (1998) provided several other examples of weight functions and index sets satisfying (1.4.1). In particular, Stinchcombe and White (1998, Corollary 3.9) showed that for $t = (t_0, t_1) \in \mathbf{R}^{1+d_x}$, and $G : \mathbf{R} \to \mathbf{R}$ analytic and nonpolynomial, the function $\omega(t, x) \coloneqq G(t_0 + \Phi(x)^\top t_1)$ satisfies (1.4.1) for any $\mathcal{T} \subset \mathbf{R}^{1+d_x}$ of positive Lebesgue measure. (See their paper for definitions.) Additional valid choices may be found using Bierens and Ploberger (1997, Theorem 1) with its addendum in Bierens (2016, Chapter 5), and Stinchcombe and White (1998, Theorem 3.10).

I next impose conditions on the residual function. For this purpose, let $\mathcal{Z} \coloneqq \operatorname{supp}(Z)$ and $\mathcal{W} \coloneqq \operatorname{supp}(W)$.

Assumption 1.3 (**Residual**). *The residual function satisfies:*

1. *For each $z \in \mathcal{Z}, v \in \mathbf{R}, \beta \mapsto \rho(z, \beta, v)$ is continuous on $\mathcal{B}$ and continuously differentiable on an open neighborhood $\mathcal{N}_*$ of β_*. Moreover, there exist $c \in (0, \infty)$ and $L_1 : \mathcal{Z} \to \mathbf{R}_+$ integrable such that for each $z \in \mathcal{Z}, \beta \in \mathcal{N}_*, v \in \mathbf{R}$,*

$$\|\partial_\beta \rho(z, \beta, v) - \partial_\beta \rho(z, \beta, h_*(w))\| \leqslant L_1(z) |v - h_*(w)|^c.$$

2. *For each $z \in \mathcal{Z}, v \mapsto \rho(z, \beta_*, v)$ is continuously differentiable on $\mathbf{R}$. Moreover, there exists $\gamma \in (0,1]$, such that for each $z \in \mathcal{Z}, v \in \mathbf{R}$,*

$$|\partial_v \rho(z, \beta_*, v) - \partial_v \rho(z, \beta_*, h_*(w))| \leqslant R(z)\, |v - h_*(w)|^\gamma,$$

   where $\mathrm{E}[R(Z)] \sqrt{n} \|\widehat{h} - h_*\|_{\mathcal{W}}^{1+\gamma} \to_{\mathrm{P}} 0$.

3. $|\rho(Z, \beta_*, h_*(W))|$, $\sup_{\beta \in \mathcal{N}_*} \|\partial_\beta \rho(Z, \beta, h_*(W))\|$ *and* $|\partial_v \rho(Z, \beta_*, h_*(W))|^2$ *are integrable.*

Assumptions 1.3.1 and 1.3.2 involve smoothness conditions which allow for a linearization around (β_*, h_*) to extract the dominant component of test statistic. The assumption of everywhere differentiability may be relaxed to accommodate nondifferentiable residual (as in quantile regression) or even discontinuous residuals (as in Pakes and Pollard, 1989 and Chen et al., 2003). While I consider extensions to nonsmooth residuals of great value, I leave them for future research.

Assumption 1.3.2 generally requires $\widehat{h}$ to converge to h_* sufficiently fast with respect to the supremum metric. If $\gamma = 1$ (the leading case), then for $\sqrt{n}\|\widehat{h} - h_*\|_{\mathcal{W}}^{1+\gamma} \to_{\mathrm{P}} 0$ it typically suffices that the convergence rate is $o(n^{-1/4})$. This rate requirement often boils down to assuming that the estimand is sufficiently smooth. To illustrate, suppose for the moment that (i) h_* belongs to a Hölder ball $\Sigma(s, L, \mathcal{W})$ with Hölder exponent s, radius L, and domain $\mathcal{W}$; and, (ii) $\widehat{h}$ achieves the Stone (1982) optimal rate of convergence with respect to the supremum metric, i.e., $\|\widehat{h} - h_*\|_{\mathcal{W}} \lesssim_{\mathrm{P}} n^{-s/(2s+d_w)}$ (up to a $\ln n$ factor). Then $n^{-s/(2s+d_w)} = o(n^{-1/4})$ is equivalent to $s > d_w/2$. In words, when $\gamma = 1$, the linearization-in-h requirement holds whenever the target function h_* is sufficiently smooth relative to its number of arguments. Allowing for a general $\gamma \in (0,1]$, the requirement becomes $s\gamma > d_w/2$. Thus, what matters is the *composite* smoothness $s\gamma$, which is given by the smoothness s of h_* scaled by the smoothness γ as h_* passes through the residual function.

While the previous assumptions allow for *general* nonparametric estimation methods, the following regularity conditions are tailored to *series* estimators. The first assumption is prevalent in the series estimation literature (see, e.g., Stone 1985; Newey 1994,1997; and Belloni et al. (2015)).

Assumption 1.4 (**Variance**). $\mathrm{var}(Y|\,W)$ *is bounded.*

The second assumption imposes regularity conditions on the approximating functions in $p^k = (p_{1k}, \ldots, p_{kk})^\top$.

Assumption 1.5 (**Eigenvalues**). *The eigenvalues of* $\mathrm{E}[p^k(W)\, p^k(W)^\top]$ *are bounded from above and away from zero uniformly over* $k \in \mathbf{N}$.

Assumption 1.5 imposes a condition on the design matrix,

$$Q_k \coloneqq \mathrm{E}[p^k(W)p^k(W)^\top], \tag{1.4.10}$$

which, loosely speaking, requires that the "regressors" $p_{1k}(W_i), \dots, p_{kk}(W_i)$ are not too colinear. It may be necessary to apply a nonsingular linear transformation of the approximating functions in order to satisfy the requirements of Assumption 1.5. Note that such nonsingular linear transformations do not alter the estimator. If power series are used as approximating functions, then these may be orthonormalized with respect to some weight function. Similarly, B-splines may be used in place of ordinary splines in order to lower multicollinearity.

Example 1.4 (Stability of Bounds on Eigenvalues). If the W has distribution F, and the $\{p_{jk}\}_{j=1}^k$ are orthonormal on $(\mathcal{V}, \nu)$ for some measure ν, then Assumption 1.5 holds provided $\mathrm{d}F/\mathrm{d}\nu$ is bounded from above and away from zero (Belloni et al., 2015, Proposition 2.1).[13] For example, if W is continuously distributed on $\mathcal{V}$ and $\{p_{jk}\}_{j=1}^k$ are orthonormal with respect to Lebesgue measure on $\mathcal{V}$, then for Assumption 1.5 to hold it suffices that the density of W is bounded from above and away from zero. Specifically, if W is uniformly distributed on $[-1, 1]$, then an orthogonalization of the power series $w \mapsto p_{jk}(w) = w^{j-1}, j \in \{1, \dots, k\}$, with respect to Lebesgue measure leads to the Legendre polynomials.[14]

Assumptions 1.4 and 1.5 are used to control the variance of the series estimator, but do not provide control over the bias arising from approximating the unknown h_* by a linear form. The bias—or, approximation error—will be stated in terms of the supremum metric. The following assumption restricts the quality of the approximation provided by the approximating functions relative to this metric.

Assumption 1.6 (Approximation). *h_* is bounded. Moreover, there exists a constant $\alpha \in (0, \infty)$ such that for each $k \in \mathbf{N}$ there is a $\widetilde{\pi}_k \in \mathbf{R}^k$ such that $\|\widetilde{h}_k - h_*\|_{\mathcal{W}} \lesssim k^{-\alpha}$ for the linear form $\widetilde{h}_k \coloneqq p^{k\top}\widetilde{\pi}_k$.*

Assumption 1.6 is a high-level asssumption, but it is satisfied in many cases. The integer α usually depends on the smoothness of h_* and its number of arguments. When h_* can be viewed as a member of some smooth class of functions, then α is typically available from the approximation theory literature. For example, if h_* belongs to a Hölder ball with

[13] Here $\mathrm{d}F/\mathrm{d}\nu$ denotes the Radon-Nikodym derivative of F with respect to ν.

[14] The jth order Legendre polynomial $\widetilde{p}_j$ satisfies $\int_{-1}^1 \widetilde{p}_j(w)^2 \,\mathrm{d}w = 2/(2j+1), j \in \{0, 1, 2, \dots\}$. Orthonormal Legendre polynomials therefore follow from the formula $p_j \coloneqq \widetilde{p}_{j-1}\sqrt{[2(j-1)+1]/2} = \widetilde{p}_{j-1}\sqrt{(2j-1)/2}, j \in \mathbf{N}$.

Hölder exponent s (sometimes referred to as h_* being "s-smooth," cf. Chen 2007, p. 5570), then Assumption 1.6 holds with $\alpha = s/d_w$, provided p^k is constructed using either power series (see, e.g., Timan, 1963, Section 5.3.2; Lorentz, 1966, Theorem 8) or splines (see, e.g., Schumaker, 2007; DeVore and Lorentz, 1993).

Assumption 1.5 is a normalization that restricts the magnitude of the series terms. The theory to follow will also require that the size of p^k does not grow too fast relative to the sample size, where size is quantified by

$$\zeta_k \coloneqq \sup_{w \in \mathcal{W}} \left\| p^k(w) \right\|. \tag{1.4.11}$$

Bounds on ζ_k are available for specific choices of approximating functions. For example, for power series $\zeta_k \lesssim k$, and for regression splines $\zeta_k \lesssim \sqrt{k}$ (cf. Newey, 1997). See also Belloni et al. (2015, Section 3) for a comprehensive list of examples.

Remark 1.3 (Smallest Size of Approximating Functions). Given that the eigenvalues of Q_k are bounded away from zero (Assumption 1.5), Q_k^{-1} exists and has eigenvalues bounded from above, such that $\mathrm{E}[p^k(W)^\top Q_k^{-1} p^k(W)] \leqslant C\mathrm{E}[\|p^k(W)\|^2]$. Given that $\mathrm{E}[p^k(W)^\top Q_k^{-1} p^k(W)] = \mathrm{tr}\{Q_k^{-1}\mathrm{E}[p^k(W)\, p^k(W)^\top]\} = \mathrm{tr}(I_k) = k$, we must have

$$\zeta_k^2 \geqslant \mathrm{E}[\|p^k(W)\|^2] \geqslant (1/C)\,\mathrm{E}[p^k(W)^\top Q_k^{-1} p^k(W)] = (1/C)\,k,$$

Hence, under Assumption 1.5, one necessarily has $\zeta_k \gtrsim \sqrt{k}$, and $\sqrt{k}$ is the smallest order of size ζ_k for p^k.

The probabilistic behavior of the test statistic T_n depends crucically on the probabilistic behavior of the stochastic process $\{\sqrt{n}\mathbb{E}_n[\rho(Z_i, \widehat{\beta}, \widehat{h}(W_i))\omega(t, X_i)] | t \in \mathcal{T}\}$. An expansion around (β_*, h_*) shows that this process is asymptotically equivalent to the stochastic process $\{\sqrt{n}\mathbb{E}_n[f_*(t, Z_i)] | t \in \mathcal{T}\}$, where

$$f_*(t,z) \coloneqq \rho(z, \beta_*, h_*(w))\,\omega(t,x) + b_*(t)^\top s_*(z) + \delta_*(t,w)\,[y - h_*(w)], \tag{1.4.12}$$

$$b_*(t) \coloneqq \mathrm{E}\left[\omega(t,X)\,\partial_\beta \rho(Z, \beta_*, h_*(W))\right], \tag{1.4.13}$$

$$\delta_*(t,W) \coloneqq \mathrm{E}\left[\omega(t,X)\,\partial_v \rho(Z, \beta_*, h_*(W)) \middle| W\right]. \tag{1.4.14}$$

Here $b_*(t)^\top s_*(z)$ and $\delta_*(t,w)[y - h_*(w)]$ are adjustments to the (optimal) tth moment function $z \mapsto \rho(z, \beta_*, h_*(w))\,\omega(t,x)$ due to estimation of β_* and h_*, respectively. The form of the adjustment term due to estimation of β_* follows from a mean-value expansion, with $b_*(t)$ being the tth element of the Jacobian. The form of the adjustment term due to estimation of h_* is similar to the adjustment to the influence function of a two-step GMM estimation with a nonparametric first step (Newey, 1994; see also Example 1.3). Specifically, the adjustment $\delta_*(t,w)[y-h_*(w)]$ in (1.4.12) follows from the (1.4.8). The main difference is that, while two-step semiparametric GMM estimation requires adjustment of the finite number of moments used in defining the GMM criterion function, I here need to adjust a possibly infinite collection of moment functions $\{z\mapsto\rho(z,\beta_*,h_*(w))\,\omega(t,x)|\,t\in\mathcal{T}\}$ for estimation of h_*.

The following assumption imposes rate conditions whose primary purpose is to ensure the errors arising from approximating the stochastic process $\{\sqrt{n}\mathbb{E}_n[\rho(Z_i,\widehat{\beta},\widehat{h}(W_i))\omega(t,X_i)]|t\in\mathcal{T}\}$ by $\{\sqrt{n}\mathbb{E}_n[f_*(t,Z_i)]\,|t\in\mathcal{T}\}$ are asymptotically negligible. For the purpose of stating these conditions, define the mean-square projection coefficients

$$\pi_{h,k}\coloneqq\operatorname*{argmin}_{\pi\in\mathbf{R}^k}\mathrm{E}\{[p^k(W)^\top\pi-h_*(W)]^2\},\tag{1.4.15}$$

$$\pi_{\delta,k}(t)\coloneqq\operatorname*{argmin}_{\pi\in\mathbf{R}^k}\mathrm{E}\{[p^k(W)^\top\pi-\delta_*(t,W)]^2\},\tag{1.4.16}$$

and their induced mean-square errors

$$r_{h,k}^2\coloneqq\mathrm{E}\{[p^k(W)^\top\pi_{h,k}-h_*(W)]^2\}=\min_{\pi\in\mathbf{R}^k}\mathrm{E}\{[p^k(W)^\top\pi-h_*(W)]^2\},\tag{1.4.17}$$

$$r_{\delta,k}^2(t)\coloneqq\mathrm{E}\{[p^k(W)^\top\pi_{\delta,k}(t)-\delta_*(t,W)]^2\}=\min_{\pi\in\mathbf{R}^k}\mathrm{E}\{[p^k(W)^\top\pi-\delta_*(t,W)]^2\},\tag{1.4.18}$$

$$R_{\delta,k}^2\coloneqq\mathrm{E}\big[\|p^k(W)^\top\pi_{\delta,k}(\cdot)-\delta_*(\cdot,W)\|_{\mathcal{T}}^2\big].\tag{1.4.19}$$

Assumption 1.7 (**Rate Conditions**). *For α provided by Assumption 1.6,*

$$\zeta_{k_n}r_{h,k_n}\to0,\qquad nr_{h,k_n}^2\|r_{\delta,k_n}\|_{\mathcal{T}}^2\to0,\qquad \zeta_{k_n}^2k_n\ln(k_n)/n\to0,$$

$$R_{\delta,k_n}\to0,\quad R_{\delta,k_n}\sqrt{\ln(k_n/R_{\delta,k_n})}\to0,\quad\Big(\sum_{j=1}^{k_n}\|p_{jk_n}\|_{\mathcal{W}}^2\Big)^{1/2}(\sqrt{k_n/n}+k_n^{-\alpha})\to0.$$

Given that $\zeta_k\leqslant(\sum_{j=1}^k\|p_{jk}\|_{\mathcal{W}}^2)^{1/2}$, the latter rate condition ensures that $\zeta_{k_n}(\sqrt{k_n/n}+k_n^{-\alpha})\to0$, which I use to argue uniform consistency. Note that the presence of ζ_k in the rate conditions requires one to use approximating functions that are bounded on $\mathcal{W}$.

Observe that the mean-square error r_{h,k_n} resulting from approximating h_* by linear forms is *not* required to go to zero at a rate faster than $n^{-1/2}$. Such a condition would otherwise require choosing k_n larger than what would maximize its rate of convergence—a phenomenon referred to as "undersmoothing." Instead Assumption 1.7 requires the *product* of r_{h,k_n} and the maximal approximation mean-square error $\|r_{\delta,k_n}\|_{\mathcal{T}}$ to be $o(n^{-1/2})$. This property arises from the orthogonality property of mean-square projections, where, for the projections h_k

and $\delta_k(t,\cdot)$ of h_* and $\delta_*(t,\cdot)$, respectively, the bias term $\mathrm{E}\{\delta_*(t,W)[h_k(W)-h_*(W)]\}$ is equal to $\mathrm{E}\{[\delta_k(t,W)-\delta_*(t,W)][h_k(W)-h_*(W)]\}$ for each $t\in\mathcal{T}$. As a consequence, if the family $\{\delta_*(t,\cdot)|\, t\in T\}$ can be sufficiently well approximated by linear forms, then there is no need to "undersmooth."[15] Newey (1994) shows that a similar feature arises in the context of two-step GMM estimation with a first step based on series estimation of projection functionals.

The expression "sufficiently well approximated" can be quantified by assuming that $\{\delta_*(t,\cdot)|\, t\in\mathcal{T}\}$ belongs to a space of sufficiently smooth functions.

Example 1.5 (**Undersmoothing and Smooth Functions**). Suppose that $h_*\in\Sigma(s_h,L_h,\mathcal{W})$, a Hölder space of functions on $\mathcal{W}$ with smoothness $s_h>0$ and Lipschitz constant L_h, and that $\delta_*(t,\cdot)\in\Sigma(s_\delta,L_\delta,\mathcal{W})$ for all $t\in\mathcal{T}$, where $\Sigma(s_\delta,L_\delta,\mathcal{W})$ denotes a Hölder space of functions on $\mathcal{W}$ with smoothness $s_\delta>0$ and Lipschitz constant L_δ. If p^k is constructed using *power series* then

$$\inf_{\pi\in\mathbf{R}^k}\|p^{k\top}\pi-h_*\|_{\mathcal{W}}\leqslant Ck^{-s/d_w},$$

$$\sup_{t\in\mathcal{T}}\inf_{\pi\in\mathbf{R}^k}\|p^k(\cdot)^\top\pi-\delta_*(t,\cdot)\|_{\mathcal{W}}\leqslant Ck^{-s_\delta/d_w},$$

where the constant C in the second equation does not depend on t. Hence, for $nr_{h,k_n}^2\|r_{\delta,k_n}\|_{\mathcal{T}}^2\to 0$ to hold it suffices that $\sqrt{n}k_n^{-(s_h+s_\delta)/d_w}\to 0$. Assuming for the moment that k_n is chosen to maximize the uniform rate of convergence of $\widehat{h}$ to h_*, i.e., $k_n\asymp n^{d_w/(2s+d_w)}$ (up to a $\ln n$ factor).[16] Then $\sqrt{n}k_n^{-(s_h+s_\delta)/d_w}\to 0$ if and only if $n^{1/2-(s+s_\delta)/(2s+d_w)}\to 0$, which, in turn, is equivalent to $s_\delta>d_w/2$. Thus, if the functions $\delta_*(t,\cdot), t\in\mathcal{T}$, are sufficiently smooth ($s_\delta>d_w/2$), then one may indeed pick the number of series terms k_n in a uniform rate optimal fashion, and "undersmoothing" is unnecessary.

The previous assumptions suffice for the following lemma.

Lemma 1.1 (**Asymptotic Equivalence**). *If Assumptions 1.1–1.7 hold, then for f_* defined in (1.4.12),*

$$\|\sqrt{n}\mathbb{E}_n[\rho(Z_i,\widehat{\beta},\widehat{h}(W_i))\omega(\cdot,X_i)-\sqrt{n}\mathbb{E}_n[f_*(\cdot,Z_i)]\|_{\mathcal{T}}\xrightarrow{\mathrm{P}}0.$$

Lemma 1.1 shows that the stochastic processes $\{\sqrt{n}\mathbb{E}_n[\rho(Z_i,\widehat{\beta},\widehat{h}(W_i))\omega(t,X_i)]|t\in\mathcal{T}\}$ and $\{\sqrt{n}\mathbb{E}_n[f_*(t,Z_i)]\,|t\in\mathcal{T}\}$ are asymptotically equivalent. By the triangle inequality and

[15] While undersmoothing may not be necessary to achieve the claimed asymptotic approximation, it may be "optimal" in the sense of minimizing the remainder resulting from this approximation as shown by Donald and Newey (1994) in the context of partially linear regression.

[16] While the implicit logarithmic factor diverges to infinity with n, it will eventually be dominated by n^c for any $c>0$ and may therefore be ignored in this discussion.

continuous mapping theorem, the lemma implies that the probabilistic behavior of T_n can be approximated by that of $\|\sqrt{n}\mathbb{E}_n\left[f_*\left(\cdot, Z_i\right)\right]\|_{\mu,2}^2$.

Recall that a class $\mathcal{F}$ of real-valued functions f is called a *Donsker class* if the sequence of empirical processes $\left\{\sqrt{n}\left(\mathbb{E}_n - \mathrm{E}\right)\left[f\left(Z_i\right)\right] \middle| f \in \mathcal{F}\right\}, n \in \mathbf{N}$, induced by $\mathcal{F}$—viewed as random elements of the space of real-valued, bounded functions on $\mathcal{F}$—converges weakly to a zero-mean Gaussian process $\left\{\mathbb{G}\left(f\right) \middle| f \in \mathcal{F}\right\}$ with covariance function $\mathrm{E}\left[\mathbb{G}\left(f_1\right)\mathbb{G}\left(f_2\right)\right] = \mathrm{E}\left[f_1\left(Z\right)f_2\left(Z\right)\right] - \mathrm{E}\left[f_1\left(Z\right)\right]\mathrm{E}\left[f_2\left(Z\right)\right], f_1, f_2 \in \mathcal{F}$ (see, for example, van der Vaart and Wellner, 1996, pp. 81-82).

The same assumptions then also show:

Lemma 1.2 (**Donsker Class**). *If Assumptions 1.1–1.7 hold, then* $\mathcal{F} \coloneqq \{f_*\left(t, \cdot\right) : \mathcal{Z} \to \mathbf{R} | t \in \mathcal{T}\}$ *is Donsker.*

Lemma 1.2 implies that the sequence of stochastic processes $\left\{\sqrt{n}\left(\mathbb{E}_n - \mathrm{E}\right)\left[f_*\left(t, Z_i\right)\right] \middle| t \in \mathcal{T}\right\}, n \in \mathbf{N}$—now viewed as random elements of the space of real-valued, bounded functions on $\mathcal{T}$—converges weakly to a zero-mean Gaussian process, i.e., $t \mapsto \sqrt{n}\left(\mathbb{E}_n - \mathrm{E}\right)\left[f_*\left(t, Z_i\right)\right]$ satisfies a functional central limit theorem (FCLT). Noting that

$$\begin{aligned}\mathrm{E}\left[f_*\left(t, Z\right)\right] &= \mathrm{E}\left[\rho\left(Z, \beta_*, h_*\left(W\right)\right)\omega\left(t, X\right)\right] + b_*\left(t\right)^\top \mathrm{E}\left[s_*\left(Z\right)\right] + \mathrm{E}\left\{\delta_*\left(t, W\right)\left[Y - h_*\left(W\right)\right]\right\} \\ &= \mathrm{E}\left[\rho\left(Z, \beta_*, h_*\left(W\right)\right)\omega\left(t, X\right)\right] \qquad (1.4.20)\end{aligned}$$

uniformly in $t \in \mathcal{T}$, one obtains the asympotic behavior of the test statistic.

Theorem 1.1 (**Asymptotic Behavior of Test Statistic**). *Let Assumptions 1.1–1.7 hold. Then (1) under* H_0

$$T_n \xrightarrow{d} \int_{\mathcal{T}} G_0\left(t\right)^2 \mathrm{d}\mu\left(t\right),$$

for a centered Gaussian process G_0 *with covariance function* $\mathrm{E}[f_*(t_1, Z)f_*(t_2, Z)], t_1, t_2 \in \mathcal{T}$*; (2) while under* H_1,

$$T_n/n \xrightarrow{\mathrm{P}} \int_{\mathcal{T}} \left\{\mathrm{E}\left[\rho\left(Z, \beta_*, h_*\left(W\right)\right)\omega\left(t, X\right)\right]\right\}^2 \mathrm{d}\mu\left(t\right) > 0.$$

1.4.5 Critical Values and the Multiplier Bootstrap

The asymptotic results of Theorem 1.1 cannot be implemented for inference without a consistent estimator for the appropriate critical values. For this purpose, I employ a *Gaussian multiplier bootstrap* procedure.

By Theorem 1.1, the limiting law of T_n under the null hypothesis is given by $\|G_0\|_{\mu,2}^2$. To acquiring a consistent bootstrap it therefore suffices to estimate the law of the Gaussian process G_0 on $\mathcal{T}$. Toward this end, let $\{\xi_i\}_1^\infty$ be i.i.d. standard normal random variables independent of the stream of data $\{Z_i\}_1^\infty$. To fix ideas, consider the *multiplier process* G_n^* defined by

$$G_n^*(t) \coloneqq \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\xi_i - \overline{\xi}\right) f_*(t, Z_i), \quad t \in \mathcal{T}, \tag{1.4.21}$$

where $\overline{\xi} \coloneqq \mathbb{E}_n(\xi_i)$. By independence, the summands of G_n^* are centered even if the $f_*(t, Z)$'s are not. The purpose of including $\overline{\xi}$ in (1.4.21) is to take into account that the $f_*(t, Z_i)$ may not be centered with respect to the empirical distribution even if the null is true. Rearranging, this connection can be made explicit:

$$G_n^*(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i \left\{ f_*(t, Z_i) - \mathbb{E}_n\left[f_*(t, Z_i)\right]\right\}.$$

This sample-centering ultimately leads to less conservative critical values in finite sample by correctly accounting for sample variation.

The following discussion requires the notion of weak convergence in probability. The multiplier process G_n^* is said to *converge weakly in probability* to G_*, written $G_n^* \rightsquigarrow_{\mathrm{P},\xi} G_*$ in $\ell^\infty(\mathcal{T})$,[17] if their distance as measured by the bounded Lipschitz metric

$$d_{\mathrm{BL}}(G_n^*, G_*) \coloneqq \sup_{h \in \mathrm{BL}_1(\ell^\infty(\mathcal{T}))} \left| \mathrm{E}\left[h(G_n^*) \mid \{Z_i\}_1^n\right] - \mathrm{E}\left[h(G_*)\right]\right|$$

goes to zero in probability.[18] Given that $\mathcal{F}$ is Donsker (Lemma 1.2), the multiplier process satisfies a "conditional FCLT" in the sense that G_n converges weakly in probability to a centered Gaussian process G_* with covariance function $(t_1, t_2) \mapsto \mathrm{E}[f_*(t_1, Z) f_*(t_2, Z)] - \mathrm{E}[f_*(t_1, Z)]\mathrm{E}[f_*(t_2, Z)]$ (Kosorok, 2008, Theorem 10.4). Under the null, $t \mapsto \mathrm{E}[f_*(t, Z)] = \mathrm{E}[\rho(Z, \beta_*, h_*(W))\omega(t, X)]$ is the zero function, and the covariance function of G_* coincides with that of G_0. Given that the two processes G_* and G_0 are Gaussian, they must therefore be identically distributed under the null. This observation suggests using the critical value

$$c_n^*(\alpha) \coloneqq (1-\alpha)\text{-quantile of } \|G_n^*\|_{\mu,2}^2 \text{ conditional on } \{Z_i\}_1^n$$

[17] For detailed treatments of the topics of weak convergence, conditional weak convergence, and bootstrapping empirical processes, see van der Vaart and Wellner (1996) and Kosorok (2008).

[18] Here $\mathrm{BL}_1(\ell^\infty(\mathcal{T}))$ denotes the space of functionals $h : \ell^\infty(\mathcal{T}) \to \mathbf{R}$ whose Lipschitz norm is bounded by one, i.e., functionals satisfying $\|h\|_{\ell^\infty(\mathcal{T})} \leqslant 1$ and $|h(f) - h(g)| \leqslant \|f - g\|_{\mathcal{T}}$ for all $f, g \in \ell^\infty(\mathcal{T})$.

to approximate

$$c_*(\alpha) \coloneqq (1-\alpha)\text{-quantile of } \|G_*\|_{\mu,2}^2.$$

Of course, f_* is generally unknown, which renders the above procedure infeasible. However, endowed with an estimator $\widehat{s}$ of the influence function s_*, one may estimate f_* and define the *bootstrap process* $\widehat{G}$ as the feasible analog of G_n^*,

$$\widehat{G}(t) \coloneqq \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\xi_i - \overline{\xi}\right)\widehat{f}(t, Z_i), \quad t \in \mathcal{T}, \tag{1.4.22}$$

$$\widehat{f}(t,z) \coloneqq \rho(z,\widehat{\beta},\widehat{h}(w))\omega(t,x) + \widehat{b}(t)^{\top}\widehat{s}(z) + \widehat{\delta}(t,w)\left[y - \widehat{h}(w)\right], \tag{1.4.23}$$

$$\widehat{b}(t) \coloneqq \mathbb{E}_n[\omega(t,X_i)\,\partial_\beta\rho(Z_i,\widehat{\beta},\widehat{h}(W_i))], \tag{1.4.24}$$

$$\widehat{\delta}(t,w) \coloneqq p^{k_n}(w)^{\top}\left(\mathbb{E}_n[p^{k_n}(W_i)\,p^{k_n}(W_i)^{\top}]\right)^{-}\mathbb{E}_n[p^{k_n}(W_i)\,\omega(t,X_i)\,\partial_v\rho(Z_i,\widehat{\beta},\widehat{h}(W_i))]. \tag{1.4.25}$$

Note that $\widehat{\delta}(t,\cdot)$ is the regression function from a regression of $\omega(t,X_i)\,\partial_v\rho(Z_i,\widehat{\beta},\widehat{h}(W_i))$ on $p^{k_n}(W_i)$. Replacing the multiplier process with the bootstrap process, we arrive at a feasible critical value

$$\widehat{c}(\alpha) \coloneqq (1-\alpha)\text{-quantile of } \|\widehat{G}\|_{\mu,2}^2 \text{ conditional on } \{Z_i\}_1^n.$$

For a given significance level $\alpha \in (0,1)$, the critical value $\widehat{c}(\alpha)$ may be obtained through simulation of the Gaussian multipliers $\{\xi_i\}_1^n$ holding the data constant, and integrating over $t \in \mathcal{T}$. Moreover, for some choices of the weight function ω and nuisance parameter space $\mathcal{T}$, both the test statistic T_n and $\|\widehat{G}\|_{\mu,2}^2$ are available in closed form.

The only potentially difficult part of this bootstrap procedure is constructing $\widehat{s}$. For specific estimators, $\widehat{s}$ can often be formed by obtaining a formula for s_* and replacing unknown components by estimates. For example, if s_* is a function $s(\cdot,\beta_*,h_*)$ depending on β_* and h_*, then we may construct $\widehat{s}$ as $\widehat{s}(\cdot) \coloneqq s(\cdot,\widehat{\beta},\widehat{h})$. For such estimators it is possible to give primitive conditions under which $\widehat{s}$ is consistent for s_*. However, at the level of generality considered in this section it does not appear possible to do more than simply assume consistency as in the following assumption.

Assumption 1.8 (**Bootstrap Conditions**). *(1) For each $z \in \mathcal{Z}, \beta \in \mathcal{N}_*, v \mapsto \rho(z,\beta,v)$ is continuously differentiable on $\mathbf{R}$. Moreover, there exists $R' : \mathcal{Z} \to \mathbf{R}_+$ such that for each $z \in \mathcal{Z}, \beta \in \mathcal{N}_*, v \in \mathbf{R}$,*

$$|\partial_v\rho(z,\beta,v) - \partial_v\rho(z,\beta_*,h_*(w))| \leqslant R'(z)\left(\|\beta-\beta_*\| + |v - h_*(w)|\right),$$

where $\mathrm{E}\left[R'\left(Z\right)\right]\sqrt{n}\|\widehat{h}-h_*\|_{\mathcal{W}}^2 \to_{\mathrm{P}} 0$*; (2)* $\|\widehat{s}-s_*\|_{\mathbb{P}_n,2} \to_{\mathrm{P}} 0$*; and, (3)* $\zeta_{k_n}\sqrt{k_n}(\sqrt{k_n/n}+k_n^{-\alpha}) \to 0$.

For the additional rate condition in Assumption 1.8 to hold, we must necessarily have $\zeta_k k^{1/2-\alpha} \to 0$ as $k \to \infty$. When the approximating functions satisfy $\zeta_k \asymp \sqrt{k}$ (see Remark 1.3), $\zeta_k k^{1/2-\alpha} \to 0$ is equivalent to $\alpha > 1$. If h_* is s-smooth (see the discussion following Assumption 1.6), then the latter requirement translates into the smoothness requirement $s > d_w$.

With the help of Assumption 1.8, we obtain the following equivalence result.

Lemma 1.3 (**Bootstrap Equivalence**)**.** *If Assumptions 1.1–1.8 hold, then* $\|\widehat{G}-G_n^*\|_{\mathcal{T}} \to_{\mathrm{P}} 0$.

Lemma 1.3 establishes that the unknown character of f_* is asymptotically irrelevant. Given that G_n^* converges weakly in probability to G_*, by the lemma, so must its feasible analog $\widehat{G}$. Given that its limit G_* is Gaussian, $\|G_*\|_{\mu,2}^2$ is continuously distributed on the positive reals provided not every random variable $f_*\left(t, Z\right), t \in \mathcal{T}$, is degenerate. To rule out this—somewhat unrealistic—scenario and to ensure that the distribution of $\|G_*\|_{\mu,2}^2$ has no mass point at zero, I make the high-level assumption:

Assumption 1.9 (**Nondegeneracy**)**.** $\sup_{t\in\mathcal{T}} \mathrm{var}\left[f_*\left(t, Z\right)\right] > 0$.

Given the continuous nature of the weak in-probability limit $\|G_*\|_{\mu,2}^2$ of $\|\widehat{G}\|_{\mu,2}^2$, convergence of their quantiles now follows.

Theorem 1.2 (**Quantile Consistency**)**.** *If Assumptions 1.1–1.9 hold, then for each* $\alpha \in (0,1), \widehat{c}\left(\alpha\right) \to_{\mathrm{P}} c_*\left(\alpha\right) \in (0,\infty)$.

1.4.6 Limiting Behavior of Test

Theorem 1.1 shows that $T_n \to_d \int_{\mathcal{T}} G_0\left(t\right)^2 \mathrm{d}\mu\left(t\right)$ under the null. Theorem 1.2 shows that $\widehat{c}\left(\alpha\right) \to_{\mathrm{P}} c_*\left(\alpha\right) \in (0,\infty)$, which is equal to the $(1-\alpha)$-quantile of $\sup_{t\in\mathcal{T}}\left|G_0\left(t\right)\right|$ under the null. These observations lead to the following result.

Theorem 1.3 (**Size Control**)**.** *If Assumptions 1.1–1.9 hold, then for each* $\alpha \in (0,1)$,

$$\mathrm{P}\left(T_n > \widehat{c}\left(\alpha\right); \mathrm{H}_0\right) \to \alpha.$$

Theorem 1.3 is the first main result of this paper. The theorem formally establishes that the test which rejects the null hypothesis if and only if $T_n > \widehat{c}\left(\alpha\right)$ is correctly sized.

The next result shows that the test which rejects the null hypothesis if and only if $T_n > \widehat{c}(\alpha)$ is also consistent: For any fixed alternative, this test will reject the null with probability approaching one.

Theorem 1.4 (Consistency). *If Assumptions 1.1–1.9 hold, then for each $\alpha \in (0,1)$,*

$$\mathrm{P}\left(T_n > \widehat{c}\left(\alpha\right); \mathrm{H}_1\right) \to 1.$$

Theorem 1.4 is the second main result of this paper. The argument used in establishing consistency is summarized as follows. Given the asymptotic equivalence result of Lemma 1.1 and a continuity argument one may show that

$$T_n/n = \left\|\mathbb{E}_n\left[f_*\left(t, Z_i\right)\right]\right\|_{\mu,2}^2 + o_{\mathrm{P}}(n^{-1}).$$

By a uniform law of large numbers and (1.4.20),

$$\begin{aligned} &\left\|\mathbb{E}_n\left[f_*\left(\cdot, Z_i\right)\right] - \mathrm{E}\left[\rho\left(Z, \beta_*, h_*\left(W\right)\right)\omega\left(\cdot, X\right)\right]\right\|_{\mu,2} \\ &\leqslant \left\|\mathbb{E}_n\left[f_*\left(\cdot, Z_i\right)\right] - \mathrm{E}\left[\rho\left(Z, \beta_*, h_*\left(W\right)\right)\omega\left(\cdot, X\right)\right]\right\|_{\mathcal{T}} \xrightarrow{\mathrm{P}} 0.\end{aligned}$$

Combining the previous two displays, we therefore get

$$T_n/n \xrightarrow{\mathrm{P}} \left\|\mathrm{E}\left[\rho\left(Z, \beta_*, h_*\left(W\right)\right)\omega\left(\cdot, X\right)\right]\right\|_{\mu,2}^2 \overset{\mathrm{H}_1}{>} 0,$$

where the inequality follows from property (1.4.1) of the weight function and the choice of probability measure μ. This inequality implies that $T_n \to_{\mathrm{P}} \infty$ under the alternative. Given that $\widehat{c}\left(\alpha\right) \to_{\mathrm{P}} c_*\left(\alpha\right) \in (0, \infty)$ (Theorem 1.2), T_n must exceed $\widehat{c}\left(\alpha\right)$ with probability approaching one under the alternative.

1.5 High-Dimensional Unconditional Moment Models

In this section I formally present my main results on specification testing in a class of high-dimensional UMR models.

1.5.1 Null Hypothesis

The *null hypothesis* is

$$\mathrm{H}_0 : \exists \beta_0 \in \mathcal{B} \text{ s.t. } \forall k \in \{1, \ldots, q\}, \mathrm{E}\left[\rho\left(Z, \beta, L_*\left(W\right)\right)X_k\right] = 0 \text{ at } \beta = \beta_0, \tag{1.5.1}$$

where $\mathcal{B} \subset \mathbf{R}^d$ is a pre-specified parameter space, $X = (X_1, \ldots, X_q)^\top$ is a q-dimensional vector of "instruments," and $L_* (W) \coloneqq W^\top h_*$ for $W = (W_1, \ldots, W_p)^\top$ a p-vector of X-measurable "regressors," and $h_* \in \mathbf{R}^p$ defined as

$$h_* \coloneqq \left[\mathrm{E}\left(WW^\top\right)\right]^{-1} \mathrm{E}\left(WY\right), \tag{1.5.2}$$

The vector Z includes both X, Y and W as subvectors. I allow for both p and q to grow without bound with and be (potentially much) larger than the sample size n available to the researcher, i.e., I allow both $p = p_n \to \infty$ and $q = q_n \to \infty$ as well as $p \gg n$ and $q \gg n$. I will therefore treat both W and X as *high-dimensional random vectors* and h_* as a *high-dimensional parameter*. In contrast, I will treat the parameter space $\mathcal{B}$ as "low-dimensional" in the sense that the dimension d is fixed and small relative to both n and q.

The *alternative hypothesis* is the negation of the null:

$$\mathrm{H}_1 : \forall \beta \in \mathcal{B}, \exists k \in \{1, \ldots, q\} \text{ s.t. } \mathrm{E}\left[\rho\left(Z, \beta, L_*(W)\right) X_k\right] \neq 0.$$

The term $L_* (W)$ is a *linear predictor* for the outcome variable Y. In fact, given that h_* is the (assumed unique) solution to the first order condition of the convex problem "minimize $\mathrm{E}[(Y - W^\top h)^2]$ subject to $h \in \mathbf{R}^p$,"

$$\mathrm{E}\left[\left(Y - W^\top h\right) W\right] = \mathbf{0}_{p\times 1}, \tag{1.5.3}$$

$L_* (W)$ is the *best* linear predictor of Y in the sense of minimizing mean-squared error. The reader may find it helpful to think of the high-dimensional best linear predictor $L_* (W)$ as a surrogate for the conditional expectation $\mathrm{E}(Y|W)$, This CEF, in turn, captures the expectation formed by an agent operating in a uncertation environment as illustrated by the collection of examples in Sections 1.2 and 1.B.

On one hand, one may rightfully view $L_* (W)$ as only an *approximation* to $\mathrm{E}(Y|W)$ on which the agents based their decisions.[19] On the other hand, one may argue that not much is lost from using a high-dimensional best linear predictor, since a high-dimensional linear function $w \mapsto \sum_{j=1}^{p} h_{*j} w_j$ is numerically indistinguishable from a truly nonparametric function $w \mapsto \sum_{j=1}^{\infty} \gamma_{*j} w_j$ with the true values γ_{*j} of the infinite-dimensional parameter $\{\gamma_j\}_1^\infty$ decaying sufficiently fast (say, $\gamma_{*j} = 1/j$).

If Y is a binary (e.g., 0/1 or "No/Yes") random variable, then $\mathrm{E}(Y|W)$ is a conditional

[19] An alternative interpretation is that the agents themselves based their decision on the best linear prediction and not necessarily the conditional expectation. Such an interpretation may be justified by appealing to bounded rationality.

choice probability, and (1.5.3) may be viewed as adopting a predictive high-dimensional linear probability model. With a binary Y one may want to use another link function than the linear link implicitly used in (1.5.3). For example, one may want to use a logistic link function. I intend to explore alternative link functions for the binary-Y case—as well as more general models capturing the relationship between Y and W in the general-Y case—in future work.

Examples 1.2, 1.9 and 1.10 all involve *multiple* high-dimensional best linear predictors. However, to avoid cluttering notation I will here focus on the case where $L_*(W)$ is scalar valued and defer the discussion of a vector-valued $L_*(W)$ to Section 1.F.2.

1.5.2 Recasting the Null Hypothesis

In this section I recast the null hypothesis in a manner that suggests a natural, yet biased, preliminary test statistic. Using an orthogonalization procedure, I then show how the preliminary test statistic may be debiased to arrive at a final test statistic.

1.5.2.1 Recasting using Pseudo Truth

I assume that there exists $\beta_* \in \mathcal{B}$ such that (1) β_* is consistently estimable; and, (2) $\beta_0 = \beta_*$ under the null.[20] With β_* available, the null hypothesis simplifies to

$$\mathrm{H}_0 : \forall k \in \{1, \ldots, q\}, \mathrm{E}\left[\rho\left(Z, \beta_*, L_*(W)\right) X_k\right] = 0.$$

Because β_* and β_0 coincide under the null, I will refer to the available β_* as a "pseudo true" parameter or simply the "pseudo truth."[21] The purpose of introducing a pseudo true parameter is to obtain an estimand which is well-defined under both the null and alternative. Example 1.6 illustrates how one may obtain a pseudo true parameter in the present context.

Example 1.6 (Obtaining a Pseudo Truth). A pseudo truth β_* may be constructed as follows. Let $X[T]$ denote the subvector $X[T] \coloneqq (X_k |\, k \in T)$ arising from selecting the elements of X corresponding to the coordinates $T \subset \{1, \ldots, q\}$. Fix a selection T_d of d coordinates, e.g., the first d elements of X. (This selection presupposes $q \geqslant d$.) Then we may let β_* be the (assumed) unique root of $\beta \mapsto \mathrm{E}\{\rho(Z, \beta, L_*(W)) X[T_d]\}$ defined on $\mathbf{R}^d$. A root of such a map exists under regularity conditions. Uniqueness amounts to

[20] Given that this section of the paper deals with triangular array data, β_* may depend on n. I suppress this dependence throughout.

[21] There may be more than one option available for the pseudo truth. Here I assume that the researcher has settled on one option.

an identification condition. Now, *if* the null is true then there exists $\beta_0 \in \mathcal{B}$ such that $\mathrm{E}[\rho(Z, \beta_0, L_*(W)) X_k] = 0$ for all $k \in \{1, \ldots, q\}$. In particular, $\mathrm{E}[\rho(Z, \beta_0, L_*(W)) X_k] = 0$ for all k in the subset T_d. By the assumption of uniqueness, β_0 and β_* must coincide.

With the pseudo truth β_* available, we may write the null in a compact manner by letting β_* play the role of β_0 and taking the maximum deviation of the moments from zero:

$$\mathrm{H}_0 : \max_{1 \leqslant k \leqslant q} |\mathrm{E}[\rho(Z, \beta_*, L_*(W)) X_k]| = 0. \tag{1.5.4}$$

This formulation of the null hypothesis involves aggregating the moments by taking the supremum (i.e., ℓ^∞) norm of the vector of moments $(\mathrm{E}[\rho(Z, \beta_*, L_*(W)) X_k])_{k=1}^q$,

$$\|\mathrm{E}[\rho(Z, \beta_*, L_*(W)) X]\|_\infty = \max_{1 \leqslant k \leqslant q} |\mathrm{E}[\rho(Z, \beta_*, L_*(W)) X_k]|.$$

In principle, one may restate the null hypothesis using *any* norm of $\mathrm{E}[\rho(Z, \beta_*, L_*(W)) X_k]$ including the ℓ^2 norm.[22] The reason I choose to work with the supremum norm is that it allows me to draw upon general results for Gaussian approximations and multiplier bootstrap procedures for maxima of sums of high-dimensional random vectors when analyzing the behavior of the test comprised of the test statistic from Section 1.5.3 and the critical value from Section 1.5.5. Such results were recently developed by Chernozhukov, Chetverikov, and Kato (2013). To the best of my knowledge, there exists no general results on Gaussian approximations or multiplier bootstrap procedures for ℓ^r norms ($r \in [1, \infty]$) of high-dimensional random vectors except for the supremum norm ($r = \infty$).

1.5.2.2 Valid Post-Selection and Post-Regularization Inference

Let $\{Z_i\}_1^n$ denote a random sample of Z available to the researcher for estimation and testing purposes. Suppose for the moment that β_* is a known quantity in order to focus on the consequences of estimation of h_*. With an estimator $\widehat{h}$ of h_* available, one could in principle consider testing the null hypothesis based on the 'plug-in' test statistic

$$\max_{1 \leqslant k \leqslant q} \left|\mathbb{E}_n[\rho(Z_i, \beta_*, \widehat{L}(W_i)) X_{ik}]\right|,$$

where $\widehat{L}(w) = w^\top \widehat{h}$, which is a feasible version of the left-hand side of (1.5.4). If p exceeds n, then $\widehat{h}$ must estimate a high-dimensional object. To estimate h_* one therefore generally needs

[22] In fact, one may use any function $f : \mathbf{R}^q \to \mathbf{R}_+$ satisfying $f(x) = 0$ if and only if $x = \mathbf{0}_{q \times 1}$, as this is the only property of norms that I invoke to recast the null.

a machine learning method, such as the Lasso or ridge regression, or some other regularization method that allows the number of parameters to exceed the available sample size. As discussed in Chernozhukov, Hansen and Spindler (2015a; 2015b), such a ‘plug-in’ approach does in general not lead to correct inference in the presence of a high-dimensional (nuisance) parameter, which is estimated using selection or regularization methods. Intuitively, while the Lasso does well in finding strong predictors, it may miss out on predictors with small yet nonzero coefficients. The work of Leeb and Pötscher (2008) shows that exclusion of such predictors may have a detrimental impact on inference procedures.

Chernozhukov, Hansen, and Spindler (2015b) show that in order to obtain valid inference following machine learning estimation of $\widehat{h}$ it is important to use moments that are robust to small mistakes in estimation of h_*. Chernozhukov et al. (2016) construct locally robust moments (i.e., moments that are not invalidated by small mistakes in learning h_*) via orthogonalization methods for different classes of econometric models and develop general results based on these moments. In Section 1.5.2.3 I construct locally robust moments using a particular orthogonalization procedure.

1.5.2.3 Neyman Orthogonalization

In this section I transform the original moment functions $\{\rho\left(z,\beta,w^\top h\right)x_k\}_{k=1}^q$ in such a way that the resulting moments are locally robust to irregular estimation of h_*. By an “irregular” estimator I mean an estimator that converges to its estimand at a slower-than-$\sqrt{n}$ rate as n grows without bound.[23] Let $\partial_v\rho(z,\beta_*,L_*(w))$ denote the derivative calculated with respect to the values of the best linear predictor L_*, i.e., $\partial_v\rho\left(z,\beta_*,L_*(w)\right) \coloneqq \partial_v\rho\left(z,\beta_*,v\right)|_{v=L_*(w)}$. (The subscript v connotes “value.”) To construct locally robust moments, define $L_{k*}(W)$ as the best as best linear predictor of $X_k\partial_v\rho\left(Z,\beta_*,L_*(W)\right)$ using W,

$$L_{k*}(w) \coloneqq w^\top\mu_{k*},\ \mu_{k*} \coloneqq [\mathrm{E}(WW^\top)]^{-1}\mathrm{E}\left[WX_k\partial_v\rho(Z,\beta_*,L_*(W))\right],\ k\in\{1,\ldots,q\}\,. \quad (1.5.5)$$

Define the *orthogonalized moment function* $\psi_k(z,\beta,w^\top h,w^\top\mu_k)$ by

$$\psi_k\left(z,\beta,w^\top h,w^\top\mu_k\right) \coloneqq \rho\left(z,\beta,w^\top h\right)x_k + (y-w^\top h)w^\top\mu_k.$$

By definition of h_*, $\mathrm{E}[(Y-W^\top h_*)W] = \mathbf{0}_{p\times 1}$, so the second term on the right-hand side is mean-zero when evaluated at $h=h_*$. It follows from the two previous displays that the two sets of moment functions are equal in mean when the kth moment is evaluated at

[23] More precisely, an estimator is defined as “irregular” if the *distance* between the estimator and its estimand vanishes at a slower-than-$\sqrt{n}$ rate as $n\to\infty$. This definition subsumes the definition in the main text while allowing the estimand itself to change with the sample size.

$(\beta, h, \mu) = (\beta_*, h_*, \mu_{k*})$, i.e.,

$$\mathrm{E}[\psi_k(Z, \beta_*, h_*, \mu_{k*})] = \mathrm{E}[\rho(Z, \beta_*, L_*(W)) X_k] \text{ for all } k \in \{1, \ldots, q\}.$$

I may therefore recast the null using the ψ_k's instead of the original moment functions:

$$\mathrm{H}_0: \max_{1 \leqslant k \leqslant q} |\mathrm{E}[\psi_k(Z, \beta_*, L_*(W), L_{k*}(W))]| = 0. \tag{1.5.6}$$

If one may interchange the order of differentiation and integration, then both

$$\begin{aligned} \partial_h \mathrm{E}[\rho(Z, \beta_*, W^\top h) X_k]|_{h=h_*} &= \mathrm{E}[X_k \partial_v \rho(Z, \beta_*, L_*(W)) W], \\ \partial_h \mathrm{E}[(Y - W^\top h) W^\top]|_{h=h_*} &= -\mathrm{E}(WW^\top). \end{aligned}$$

The previous two displays show that, unlike the original moment functions, the ψ_k's satisfy

$$\begin{aligned} &\partial_h \mathrm{E}\left[\psi_k\left(Z, \beta_*, W^\top h, L_{k*}(W)\right)\right]\big|_{h=h_*} \\ &= \mathrm{E}[X_k \partial_v \rho(Z, \beta_*, L_*(W)) W] - \mathrm{E}(WW^\top)\mu_{k*} = \mathbf{0}_{p\times 1}. \end{aligned} \tag{1.5.7}$$

That is, missing the true value h_* by a small amount does not violate the moment conditions. It is due to the orthogonality property (1.5.7) that the ψ_k's are said to be *locally robust to* irregular estimation of h_*.

The orthogonalization method given above is inspired by Neyman (1959), who used orthogonalized scores to obtain his celebrated $C(\alpha)$ test statistic in a parametric likelihood setting. Chernozhukov et al. (2016) constructed locally robust two-step GMM estimators by adding to their original moments functions an adjustment term for first-step nonparametric estimation. This adjustment ensures that the resulting moments have zero derivative with respect to the first step, and their locally robust moment conditions may be viewed as semiparametric analogs of Neyman's (1959) scores.[24]

Given that all the μ_{k*}'s are p-dimensional and generally unknown, immunization of the ψ_k's with respect to the single high-dimensional parameter h_* comes at the cost of q additional high-dimensional parameters to be estimated. However, since each ψ_k is linear in μ,

$$\partial_\mu \mathrm{E}\left[\psi_k\left(Z, \beta_*, L_*(W), W^\top \mu\right)\right] = \mathrm{E}\left[(Y - W^\top h_*) W\right] = \mathbf{0}_{p\times 1}. \tag{1.5.8}$$

for *any* $\mu \in \mathbf{R}^p$ and therefore also when evaluated at $\mu = \mu_{k*}$. The ψ_k's are therefore

[24] See also Wooldridge (1991), Bera, Montes-Rojas, and Sosa-Escudero (2010), Lee (2005) and Chernozhukov et al. (2015b) for extensions of the $C(\alpha)$ test to parametric nonlikelihood settings.

also immunized against irregular estimation μ_{k*}'s. Hence, while the use of orthogonalized moments may come at an increased computational cost, no new bias issues arise.

Examples 1.2, 1.9 and 1.10 all involve *multiple* high-dimensional best linear predictors. When the residual depends on high-dimensional linear projections of Y_ℓ on a collection of regressors W_ℓ with projection coefficients $h_{\ell *}$ given by

$$h_{\ell *} \coloneqq [\mathrm{E}(W_\ell W_\ell^\top)]^{-1}\mathrm{E}\left(W_\ell Y_\ell\right), \quad \ell \in \{1, \ldots, L\},$$

then the kth (orthogonalized) moment function ψ_k is defined by adding up the invidual adjustment terms,

$$\psi_k(z, \beta, (w_\ell^\top h_\ell)_1^L, (w_\ell^\top \mu_{k\ell})_{\ell=1}^L) = \rho\left(z, \beta, (w_\ell^\top h_\ell)_1^L\right) x_k + \sum_{\ell=1}^{L} \left(y_\ell - w_\ell^\top h_\ell\right) w^\top \mu_{k\ell}. \tag{1.5.9}$$

Here the "true" $\mu_{k\ell *}$'s are given by the projection coefficients

$$\mu_{k\ell *} \coloneqq [\mathrm{E}(W_\ell W_\ell^\top)]^{-1}\mathrm{E}\left[W_\ell X_k \partial_{v_\ell}\rho\left(Z, \beta_*, (W_\ell^\top h_{\ell *})_1^L\right)\right], \quad \ell \in \{1, \ldots, L\}, \tag{1.5.10}$$

and ∂_{v_ℓ} denotes differentiation with respect to $w^\top h_\ell$. Equations analogous to (1.5.7) and (1.5.8) show that the ψ_{k*}'s thus defined are immunized against irregular estimation of the $h_{\ell *}$'s (and the $\mu_{k\ell *}$'s).

In some cases some of the μ_{k*}'s are known or at least known up to β_* and h_*. Such μ_{k*} I choose to estimate using the plug-in method. Moreover, in special cases where the residual is affine in the best linear predictors, the orthogonalization procedure may reduce the effective number of moments employed for testing by setting some moment functions to zero. (Both of these points are illustrated in the case of the high-dimensional linear model in Section 1.E.) However, one will in general have as many orthogonalized moment functions as original moment functions.

1.5.3 Test Statistic

In this section I construct a test statistic, which constitutes one half of the specification test. The other half—the critical value—is given in Section 1.5.5. For the purpose of constructing a test statistic, let $\{Z_i\}_1^n$ denote a random sample of Z available to the researcher. For a given regular estimator $\widehat{\beta}$, I test the null hypothesis (1.5.6) using the test statistic

$$T \coloneqq \max_{1 \leqslant k \leqslant q} \left|\sqrt{n}\mathbb{E}_n[\psi_k(Z_i, \widehat{\beta}, \widehat{L}\left(W_i\right), \widehat{L}_k\left(W_i\right))]\right|, \tag{1.5.11}$$

where $\widehat{L}(w) := w^\top \widehat{h}$ and $\widehat{L}_k(w) := w^\top \widehat{\mu}_k$ estimate L_* and L_{k*}, respectively. Although the theory to follow could be modified to allow for the use of other machine learning methods, I will estimate both the h_* and the μ_{k*}'s using the Lasso (see Section 1.G.3).

The Lasso estimator $\widehat{h}$ of h_* is defined as any solution to the penalized least squares problem

$$\widehat{h} \in \operatorname*{argmin}_{h \in \mathbf{R}^p} \left\{ \mathbb{E}_n[(Y_i - W_i^\top h)^2] + \frac{\lambda_h}{n} \|\widehat{\Upsilon}_h h\|_1 \right\}, \tag{1.5.12}$$

where $\lambda_h \geqslant 0$ is a penalty level and $\widehat{\Upsilon}_h := \operatorname{diag}(\widehat{\gamma}_{h1}, \ldots, \widehat{\gamma}_{hp})$ a diagonal matrix specifying penalty loadings resulting in an $\widehat{\Upsilon}_h$-weighted ℓ_1-norm $\|\widehat{\Upsilon}_h h\|_1 = \sum_{j=1}^p \widehat{\gamma}_{hj} |h_j|$. The choice of both penalty level and loadings required to implement this Lasso are given in Section 1.H.2. The estimated high-dimensional best linear predictor $\widehat{L}$ is defined as $\widehat{L}(w) := w^\top \widehat{h}$.

The Lasso estimator $\widehat{\mu}_k$ of μ_{k*} is defined as any solution to the penalized least squares problem

$$\widehat{\mu}_k \in \operatorname*{argmin}_{\mu \in \mathbf{R}^p} \left\{ \mathbb{E}_n\{[\partial_v \rho(Z_i, \widehat{\beta}, \widehat{L}(W_i)) X_{ik} - W_i^\top \mu]^2\} + \frac{\lambda_\mu}{n} \|\widehat{\Upsilon}_{\mu k} \mu\|_1 \right\}, \tag{1.5.13}$$

where $\widehat{\beta}$ and $\widehat{L}$ are the estimators from above, $\lambda_\mu \geqslant 0$ is a penalty level common to all $k \in \{1, \ldots, q\}$ minimization problems, and $\widehat{\Upsilon}_{\mu k} := \operatorname{diag}(\widehat{\gamma}_{\mu k1}, \ldots, \widehat{\gamma}_{\mu kp})$ a problem-specific diagonal matrix specifying penalty loadings resulting in an $\widehat{\Upsilon}_{\mu k}$-weighted ℓ_1-norm $\|\widehat{\Upsilon}_{\mu k}\mu\|_1 = \sum_{j=1}^p \widehat{\gamma}_{\mu kj} |\mu_j|$. The choice of both penalty level and loadings required to implement these Lasso are given in Section 1.H.2.

Note that, in contrast to the observable outcome Y in (1.5.12), the "outcome variables" $\{X_k \partial_v \rho(Z, \beta_*, L_*(W))\}_1^q$ used in defining the μ_{k*}'s in (1.5.5) are generally not observable to the researcher due to their dependence on the unknowns β_* and h_*. To construct feasible estimators $\{\widehat{\mu}_k\}_1^q$, I therefore replace each function $z \mapsto x_k \partial_v \rho(z, \beta_*, w^\top h_*)$ with an estimate $z \mapsto x_k \partial_v \rho(z, \widehat{\beta}, w^\top \widehat{h})$. The extension of the theory for Lasso estimation to many high-dimensional best linear predictors also accommodates estimated outcomes (see Section 1.G).

Remark 1.4 (Linear Combinations and Plug-In Estimates). An exception to the Lasso estimation procedure for the μ_{k*}'s outline above occurs when $X_k \partial_v \rho(Z, \beta_*, L_*(W))$ can be written as a linear combination $\sum_{j=1}^p a_{kj*} W_j + b_{k*} Y$ of the W_j's and Y with coefficients $a_{kj*} := a_{kj}(\beta_*, h_*)$ and $b_{k*} := b_k(\beta_*, h_*)$ being *known* functions a_{kj} and b_k of (β_*, h_*). This special structure occurs in the high-dimensional linear model of Example 1.9. (See also Section 1.E.) In this case linear algebra yields $\mu_{k*} = \sum_{j=1}^p a_{kj*} e_j + b_{k*} h_*$ with $e_j \in \mathbf{R}^p$ denoting the jth elementary vector. Instead of using the Lasso to estimate such μ_{k*}'s, I choose to use the plug-in method and set $\widehat{\mu}_k \coloneqq \sum_{j=1}^{p} \widehat{a}_{kj} e_j + \widehat{b}_k \widehat{h}$ with $\widehat{a}_{kj} \coloneqq a_{kj}(\widehat{\beta}, \widehat{h})$ and $\widehat{b}_k \coloneqq b_k(\widehat{\beta}, \widehat{h})$.

1.5.4 Large Sample Behavior of Test Statistic

In order to characterize the probabilistic behavior of T, I impose a list of conditions. For the purpose of stating these conditions, let c_1, C_1, c_2 and C_2 be some given set of strictly positive, finite constants independent of n. The nonasymptotic, high-probability bounds obtained in this paper will depend on these constants.[25] I assume the following regarding estimation of the low-dimensional parameter.

Assumption 1.10 (Low-Dimensional Parameter). *$\beta_* \in \mathbf{R}^d, d \leqslant C_1$, and for each $n \in \mathbf{N}$, $\widehat{\beta}$ is a $\{Z_i\}_1^n$-measurable, random element of $\mathbf{R}^d$. Moreover, there exists $s_* : \mathcal{Z} \to \mathbf{R}^d$ and a strictly positive sequence $\{a_n\}_1^\infty$ such that $\mathrm{E}[s_*(Z)] = \mathbf{0}_{d\times 1}, \|s_*(Z)\| \leqslant C_1$, $a_n \to 0$ and*

$$\mathrm{P}\left(\|\sqrt{n}(\widehat{\beta} - \beta_*) - \sqrt{n}\mathbb{E}_n\left[s_*(Z_i)\right]\| > a_n\right) \leqslant C_2 n^{-c_2}. \tag{1.5.14}$$

Assumption 1.10 requires that the centered and scaled estimator $\widehat{\beta}$ can be approximated by a $\sqrt{n}$-scaled average at least with high-probability. This assumption is comparable to Assumption 1.1 in that (1.5.14) combined with $a_n \to 0$ implies that $\sqrt{n}(\widehat{\beta} - \beta_*)$ is asymptotically linear with influence function s_*.[26] Assumption 1.10 makes a stronger, finite-sample statement and requires that probability of error declines polynomially fast with n. The assumption of a bounded influence function allows me to control the tail behavior of $s_*(Z)$ in a relatively simple manner (e.g., using Höeffding's inequality for bounded random variables). Boundedness may be replaced by another assumption on tail behavior such as subgaussianity.[27]

The high-dimensional best linear predictors $(L_*, \{L_{k*}\}_1^q)$ can be estimated well by the Lasso under the assumption of *sparsity*. For the sake of illustration, suppose that each best linear predictor depends on at most $s \ll n$ regressors. Then there exists $h_0 \in \mathbf{R}^p$ and $\{\mu_{k0}\}_1^q \subset \mathbf{R}^p$ such that

$$L_*(w) = w^\top h_0, \quad L_{k*}(w) = w^\top \mu_{k0}, \quad k \in \{1, \dots, q\},$$

[25] In principle, one may allow each of the conditions below to have their own set of constants and let the bounds depend on all these constants. To simplify the exposition, I reuse notation for constants that play a qualitatively similar role.

[26] Let $X_n \coloneqq \|\sqrt{n}(\widehat{\beta} - \beta_*) - \sqrt{n}\mathbb{E}_n\left[s_*(Z_i)\right]\|$. For $\varepsilon > 0$ arbitrary, the union bound implies $\mathrm{P}(X_n > \varepsilon) \leqslant \mathbf{1}(a_n > \varepsilon) + \mathrm{P}(X_n > a_n)$. Taking limits now shows that $X_n \to_{\mathrm{P}} 0$.

[27] A random variable X with mean $\mu \coloneqq \mathrm{E}(X)$ is said to be *subgaussian* if there exists $\sigma \in \mathbf{R}_+$ such that $\mathrm{E}\left[\mathrm{e}^{t(X-\mu)}\right] \leqslant \mathrm{e}^{t^2\sigma^2/2}$ for all $t \in \mathbf{R}$. If so, X is said to have *subgaussianity parameter* (at most) σ.

$$\|h_0\|_0 \vee \max_{1\leqslant k\leqslant q}\|\mu_{k0}\|_0 = \sum_{j=1}^{p}\mathbf{1}\left(h_{0j}\neq 0\right) \vee \max_{1\leqslant k\leqslant q}\sum_{j=1}^{p}\mathbf{1}\left(\mu_{k0j}\neq 0\right) \leqslant s \ll n.$$

Note the identity of each active set of regressors $T_0 = \operatorname{supp}(h_0) = \{j\in\{1,\ldots,p\}|h_{0j}\neq 0\}$ and $T_{k0} = \operatorname{supp}(\mu_{k0}) = \{j\in\{1,\ldots,p\}|\mu_{k0j}\neq 0\}$ may differ (across k) as well as be unknown to the researcher.

While this *exact sparsity* assumption is useful for illustration purposes, it is unlikely to hold in practice and unnecessarily restrictive. I will instead assume that the best linear predictors are *approximately sparse.*

Assumption 1.11 (Approximately Sparse Best Linear Predictors). *There exists $h_0\in\mathbf{R}^p$ and $\{\mu_{k0}\}_1^q\subset\mathbf{R}^p$ such that each best linear predictor is well-approximated by a linear function of $s\geqslant 1$ unknown regressors in the sense that*

$$\|h_0\|_0 \vee \max_{1\leqslant k\leqslant q}\|\mu_{k0}\|_0 \leqslant s \ll n \quad \textit{and} \quad \mathrm{P}\left(c_s > C_1\sqrt{s/n}\right) \leqslant C_2 n^{-c_2},$$

where

$$c_s \coloneqq \sqrt{\mathbb{E}_n\{[W_i^\top(h_0-h_*)\}} \vee \max_{1\leqslant k\leqslant q}\sqrt{\mathbb{E}_n\{[W_i^\top(\mu_{k0}-\mu_{k*})]^2\}}.$$

Assumption 1.11 requires that at most s regressors are able to approximate each best linear predictor function up to an approximation error, which is small with high probability. Defining the *sparse linear predictors*

$$w\mapsto L_0(w) \coloneqq w^\top h_0, \quad w\mapsto L_{k0}(w) \coloneqq w^\top\mu_{k0}, \quad k\in\{1,\ldots,q\},$$

we may express c_s as $c_s = \|L_0 - L_*\|_{\mathbb{P}_n,2} \vee \max_{1\leqslant k\leqslant q}\|L_{k0}-L_{k*}\|_{\mathbb{P}_n,2}$, which emphasizes that c_s is an error arising from approximating *best* linear predictors by *sparse* linear predictors. Here c_s is considered "small" when it is not essentially larger than the size $\sqrt{s/n}$ of the estimation error arising from the infeasible least squares estimator that knows the identity of the most important regressors. One may view L_0 and the L_{k0}'s as surrogates for the ultimate estimands $(L_*, \{L_{k*}\}_1^q)$.

Assumption 1.16 assumption roughly amounts to assuming that many of the elements of each β_{k*} are close to zero, i.e., that few regressors truly matter for prediction purposes. Note that this assumption allows for the identity of the most important regressors

$$T_0 \coloneqq \operatorname{supp}(h_0), \quad T_{k0} \coloneqq \operatorname{supp}(\mu_{k0}), \quad k\in\{1,\ldots,q\}, \tag{1.5.15}$$

to be a priori unknown to the researcher as well as differ across k. BCCH used an assumption almost identical to Assumption 1.16 in the context of estimation of CEFs. A detailed motivation and discussion of this type of assumption may be found in BCCH as well as Belloni and Chernozhukov (2011; 2013).

Let $M_j \coloneqq \sup \operatorname{supp}(|W_j|)$. I impose the following boundedness and moment conditions on the outcome Y, the instruments X (hence regressors W) and the projection error $\varepsilon \coloneqq Y - L_*(W)$.

Assumption 1.12 (Observables). *$|X_k| \leqslant C_1, |Y| \leqslant C_1$, W is X-measurable, $c_1 \leqslant M_j \leqslant C_1$, $c_1^2 \leqslant \lambda_{\min}(\mathrm{E}(WW^\top)) \leqslant \lambda_{\max}(\mathrm{E}(WW^\top)) \leqslant C_1^2$, $\mathrm{E}(\varepsilon^2 W_j^2) \geqslant c_1^2$, $\|h_*\|_1 \leqslant C_1$ and*

$$\mathrm{P}\Big(\max_{1 \leqslant j \leqslant p} \big| \max_{1 \leqslant i \leqslant n} |W_{ij}| - M_j \big| > C_2 n^{-c_2} \Big) \leqslant C_2 n^{-c_2}. \tag{1.5.16}$$

The condition that the population Gram matrix $\mathrm{E}(WW^\top)$ has eigenvalues bounded from above and away from zero is quite standard in the econometrics literature; see, for example, Newey (1997) and Belloni et al. (2015). For the sake of analyzing the Lasso, the assumptions of a bounded outcome and error are less standard but may be substantially relaxed at the expense of longer proofs. Specifically, boundedness of the ε may be replaced by some “tail bound” making extreme events unlikely. An example of random variables satisfying a tail bound is the class of *subgaussian* random variables, whose tails are no fatter than normal random variables.

The assumption of bounded regressors ($M_j \leqslant C_1$) appears essential to establishing that the penalty loadings constructed via Algorithms 1.2 and 1.3 are close to being (conservatively or truly) ideal with high probability. This dependence on boundedness stems from the appearance of $\max_{1 \leqslant i \leqslant n} |W_{ij}|$ in the conservatively ideal penalty loadings (1.G.6), which are used as target for the penalty loadings used to initiate each of these algorithms. It may be possible to devise an algorithm that does not rely on boundedness of the regressors, but such a task is beyond the scope of this paper.

The requirement that the lower bound inside the probability statement of (1.5.16) is equal to the right-hand side bound of the same equation is immaterial; were the two bounds to differ, then one may always proceed with the largest of the two bounds. This requirement may be satisfied even when p grows exponentially fast with n, cf. Example 1.11.

Let ε_k denote the projection error $\varepsilon_k \coloneqq X_k \partial_v \rho(Z, \beta_*, L_*(W)) - L_{k*}(W)$.

Assumption 1.13 (Residual). *The residual function ρ satisfies:*

1. *For each $z \in \mathcal{Z}, v \in \mathbf{R}, \beta \mapsto \rho(z, \beta, v)$ is differentiable on $\mathbf{R}^d$, and for each $(z, \beta, v) \in$*

$\mathcal{Z} \times \mathbf{R}^{d+1}$ its derivative satisfies $\|\partial_\beta \rho(Z, \beta_*, L_*(W))\| \leqslant C_1$. and

$$\|\partial_\beta \rho(z, \beta, v) - \partial_\beta \rho(z, \beta_*, L_*(w))\| \leqslant C_1 (\|\beta - \beta_*\| + |v - L_*(w)|),$$

2. *For each $z \in \mathcal{Z}$, $v \mapsto \rho(z, \beta_*, v)$ is differentiable on $\mathbf{R}$, and for each $(z, \beta, v) \in \mathcal{Z} \times \mathbf{R}^{d+1}$ its derivative satisfies $|\partial_v \rho(Z, \beta_*, L_*(W))| \leqslant C_1$ and*

$$|\partial_v \rho(z, \beta, v) - \partial_v \rho(z, \beta_*, L_*(w))| \leqslant C_1 (\|\beta - \beta_*\| + |v - L_*(w)|).$$

3. $\mathrm{E}[\rho(Z, \beta_*, L_*(W))^4] \leqslant C_1^4$*,* $\mathrm{E}(\varepsilon_k^2 W_j^2) \geqslant c_1^2$ *for all k such that $X_k \partial_v \rho(Z, \beta_*, L_*(W)) \notin$* $\operatorname{span}(Y, W)$*, and $\|\mu_{k*}\|_1 \leqslant C_1$.*

Assumption 1.13.1 and 1.13.2 involve smoothness—specifically, Lipschitz—conditions which allow me to linearize around (β_*, L_*) to obtain the dominant component of the test statistic. These assumptions are comparable but stronger than Assumptions 1.3.1 and 1.3.2, and implicitly impose the somewhat crude restriction that the Lipschitz "constants" of $(\beta, v) \mapsto \partial_\beta \rho(z, \beta, v)$ and $v \mapsto \partial_v \rho(z, \beta_*, v)$—both functions of z, in general—may be bounded by an expression independent of z. These restrictions may be replaced by less strict bounds on the tail behavior of the implied random variables at the expense of longer proofs.[28]

If the $X_k \partial_v \rho(Z, \beta_*, L_*(W))$ lies in the span of Y and W, then this "outcome" variable may be described as an exact linear combination of Y and the W_j's with coefficients depending only on (β_*, h_*). In this case ε_k is identically zero, and the ideal penalty loadings in both (1.G.6) and (1.G.7) vanish. However, recall that in such cases, the best linear predictor is estimated using the plug-in approach and not the Lasso. (See also Remark 1.4.)

The following assumption imposes growth conditions, which assist in extracting the dominant component of the test statistic.

Assumption 1.14 (Growth Conditions). *s, q, and a_n satisfy the growth condition*

$$\frac{s \ln^5(pqn)}{n} + \frac{s^2 \ln^4(pqn)}{n} + a_n \sqrt{\ln q} \leqslant C_2 n^{-c_2} \quad \text{and} \quad \ln(pqn) \leqslant n^{1-c_2'},$$

where $c_2' \in \left(\frac{2}{3}, 1\right)$.

[28] However, several elements of the proof of the size control theorem below (Theorem 1.5) rely on Talagrand's deviation inequality for bounded random variables (see Lemma 1.35). It may be possible to relax some or all of these boundedness conditions by means of a deviation inequality allowing for unbounded random variables, e.g., Chernozhukov, Chetverikov, and Kato (2014b, Theorem 5.1), but I leave such potential improvements for future research.

The requirement $\ln(pqn) \leqslant C_2 n^{1-c_2'}$ for some $c_2' \in \left(\frac{2}{3}, 1\right)$ implies that while p and q may grow exponentially fast with n, they cannot grow *too* fast. Although the requirement $c_2' \in \left(\frac{2}{3}, 1\right)$ was not explicitly stated in BCCH, it appears necessary in order to guarantee the validity of moderat deviation inequalities for self-normalized sums (see Section 1.P.1 and, in particular, the proof of Lemma 1.44).

The general result for Lasso estimation of many high-dimensional best linear predictors with estimated outcomes Theorem 1.7 now implies the following result for the Lasso estimator $\widehat{L}$ from (1.5.12) and the Lasso estimators $\{\widehat{L}_k\}$ from (1.5.13).

Lemma 1.4 (Nonasymptotic, Polynomially Valid Bound for Lasso Estimation of Many Best Linear Predictors). *Suppose that Assumptions 1.10–1.14 hold and that the penalty levels λ_h and λ_μ specified as in (1.H.1) for some $c_0 > 1$ and $c_0' > 0$ with the number of best linear predictors set to 1 and q, respectively. Consider any conservatively or truly polynomially valid penalty loadings $\widehat{\Upsilon}_h$ and $\{\widehat{\Upsilon}_{\mu k}\}_1^q$, for example, the penalty loadings resulting from Algorithms 1.2 and 1.3, respectively. Then there exists c, C, C' and n_0 depending only on $c_0, c_0', c_1, C_1, c_2, C_2$ and c_2' such that for all $n \geqslant n_0$, with probability $\geqslant 1 - Cn^{-c}$,*

$$\|\widehat{L} - L_*\|_{\mathbb{P}_n,2} + \max_{1 \leqslant k \leqslant q} \|\widehat{L}_k - L_{k*}\|_{\mathbb{P}_n,2} \leqslant C' \sqrt{\frac{s \ln(pqn)}{n}}. \tag{1.5.17}$$

Provided $s\ln(pqn)/n \to 0$, the nonasymptotic, high-probability bound in Lemma 1.4 implies the rate of convergence result

$$\|\widehat{L} - L_*\|_{\mathbb{P}_n,2} + \max_{1 \leqslant k \leqslant q} \|\widehat{L}_k - L_{k*}\|_{\mathbb{P}_n,2} \lesssim_{\mathrm{P}} \sqrt{\frac{s \ln(qn)}{n}},$$

which is similar to BCCH's rate of convergence result for Lasso estimation of many CEFs (their Theorem 1).

Assumption 1.14 implies that $s\ln(pqn)/n$ is at most polynomial in n. Consequently, under the conditions of Lemma 1.4, we see that for some constants c' and C' and sufficiently large n,

$$\mathrm{P}\Big(\|\widehat{L} - L_*\|_{\mathbb{P}_n,2} + \max_{1 \leqslant k \leqslant q} \|\widehat{L}_k - L_{k*}\|_{\mathbb{P}_n,2} > C' n^{-c'}\Big) \leqslant Cn^{-c}. \tag{1.5.18}$$

As an alternative to the Lasso estimators considered in this paper, the previous diplay may be taking as a high-level condition on the choice of machine learning estimators $\widehat{L}$ and $\{\widehat{L}_k\}_1^q$.[29]

[29] Such a high-level condition would in general entail a lower bound on the constant c'. Under the smoothness assumptions in Assumption 1.13, $c' \geqslant \frac{1}{4}$ ought to suffice.

In what follows $\widehat{L}$ and the $\widehat{L}_k$'s are understood to be the Lasso estimates defined using any conservatively or truly polynomially valid penalty loadings $\widehat{\Upsilon}_h$ and $\{\widehat{\Upsilon}_{\mu k}\}_1^q$, respectively. The penalty loadings resulting from Algorithms 1.2 and 1.3 may be used. The probabilistic behavior of T depends crucially on the probabilistic behavior of the stochastic process $\{\sqrt{n}\mathbb{E}_n[\psi_k(Z_i, \widehat{\beta}, \widehat{L}(W_i), \widehat{L}_k(W_i))] | k \in \{1, \ldots, q\}\}$. The previous assumptions suffice to show that, with probability approaching one polynomially fast, this stochastic process is approximately equivalent to the stochastic process $\{\sqrt{n}\mathbb{E}_n[f_{k*}(Z_i)] | k \in \{1, \ldots, q\}\}$, where

$$f_{k*}(z) \coloneqq \psi_k(z, \beta_*, L_*(w), L_{k*}(w)) + b_{k*}^\top s_*(z), \tag{1.5.19}$$

$$b_{k*} \coloneqq \mathrm{E}[X_k \partial_\beta \rho(Z, \beta_*, L_*(W))]. \tag{1.5.20}$$

Here $b_{k*}^\top s_*(z)$ is an adjustment to the moment function $z \mapsto \psi_k(z, \beta_*, L_*(w), L_{k*}(w))$ due to estimation of β_*. Given that the ψ_k's are locally robust (see Section 1.5.2.3), no further adjustments are needed. The approximate equivalence between the stochastic processes $\{\sqrt{n}\mathbb{E}_n[\psi_k(Z_i, \widehat{W\beta}, \widehat{L}(W_i), \widehat{L}_k(W_i))] | k \in \{1, \ldots, q\}\}$ and $\{\sqrt{n}\mathbb{E}_n[f_{k*}(Z_i)] | k \in \{1, \ldots, q\}\}$ translates into approximate equivalence between the test statistic T and the random variable

$$T_* \coloneqq \max_{1 \leqslant k \leqslant q} \left| \sqrt{n}\mathbb{E}_n[f_{k*}(Z_i)] \right|.$$

Lemma 1.5 (Approximate Equivalence). *If Assumptions 1.10–1.14 hold, then there exist c, C, C' and n_0 depending only on $c_0, c_0', c_1, C_1, c_2, C_2$ and c_2' such that for $n \geqslant n_0$,*

$$\mathrm{P}\big(|T - T_*| > \zeta_1\big) \leqslant C n^{-c} \tag{1.5.21}$$

where

$$\zeta_1 \coloneqq C' \max\left\{ \sqrt{\frac{s^2 \ln^3(pqn)}{n}}, \frac{n^{-c_2/4}}{\sqrt{\ln(pqn)}}, a_n \right\}. \tag{1.5.22}$$

Lemma 1.5 shows that the probabilistic behavior of the test statistic T may be approximated by that of T_*, and that the accuracy of this approximation is polynomially valid. The lemma implies that T and T_* are asymptotically equivalent in the sense that $|T - T_*| \to_{\mathrm{P}} 0$. However, as n grows without bound, T_* may involve taking the maximum over an increasing number of elements, which need not be connected through some (equi-)continuity condition. Consequently, even under the null and even after proper standardization, T_* may not converge in distribution. However, the potential lack of convergence does not prevent one from approximating the *finite-sample* null distribution of T_* (and therefore of T) by a known distribution and using this known distribution to compute a critical value.

1.5.5 Critical Value and Gaussian Multiplier Bootstrap

Unlike T, which is defined as the maximum of an approximate average, T_* is the maximum of an *exact* average. From (1.5.19) we see that $\mathrm{E}[f_{k*}(Z)] = \mathrm{E}[\psi_k(Z, \beta_* L_*(W), L_{k*}(W))]$, which, in turn, equals $\mathrm{E}[\rho(Z, \beta_*, L_*(W))X_k]$. Hence, under the null, the $f_{k*}(Z)$'s are mean-zero, and T_* is a maximum of a *mean-zero, exact average.* These two features allows one to approximate the null distribution of T_* using Gaussian approximation results for maxima of non-Gaussian vectors recently developed by Chernozhukov, Chetverikov, and Kato (2013) for potentially high-dimemsional vectors. Via Lemma 1.5 such a Gaussian approximation in turn allows for a Gaussian finite-sample approximation to the null distribution of the test statistic itself.

To define the Gaussian approximation to T_*, let $\{g_i\}_1^n$ be independent, centered, Gaussian random vectors with common covariance $\mathrm{E}[f_*(Z)f_*(Z)^\top]$, where $f_*(Z) \coloneqq (f_{1*}(Z), \ldots, f_{q*}(Z))^\top$. Under the null, $\mathrm{E}[f_*(Z)] = \mathbf{0}_{q\times 1}$, and the g_i's are *Gaussian analogs* of $f_*(Z)$. The g_i's induce a Gaussian analog $\mathcal{Z}_*$ of T_* given by

$$\mathcal{Z}_* \coloneqq \max_{1\leqslant k\leqslant q} \left|\sqrt{n}\mathbb{E}_n(g_i)\right|. \tag{1.5.23}$$

Chernozhukov, Chetverikov, and Kato (2013) show that, under suitable (moment) assumptions, as $n \to \infty$ and possibly $q = q_n \to \infty$, under the null, the distributions of T_* and $\mathcal{Z}_*$ are close in the sense that

$$\sup_{t\in\mathbf{R}} \left|\mathrm{P}\left(T_* \leqslant t\right) - \mathrm{P}\left(\mathcal{Z}_* \leqslant t\right)\right| \leqslant Cn^{-c} \to 0,$$

for constants $c > 0$ and $C > 0$ not depending on n. If the covariance matrix $\mathrm{E}[f_*(Z)f_*(Z)^\top]$ is *known*, then this Gaussian approximation result suggests using the $(1-\alpha)$-quantile of $\mathcal{Z}_*$ as a critical value for the test statistic T. When $\mathrm{E}[f_*(Z)f_*(Z)^\top]$ is known, this critical value may be calculated via simulation of the g_i's.

The convergence in the previous display is sometimes referred to as $n^{-1/2}\sum_{i=1}^n f_*(Z_i)$ satisfying a *high-dimensional central limit theorem* (under the null). The terminology is potentially confusing since no pass to a limit is made. In fact, an advantage of the Gaussian approximation method is that it applies even in cases where a limiting distribution of $\mathcal{Z}_*$ does not exists, or when the limiting distribution exists but is unknown or complicated.

The covariance of $f_*(Z)$, is generally unknown, which renders the previous strategy for obtaining a critical value infeasible. As a step towards an feasible critical value, suppose instead that while the covariance $\mathrm{E}[f_*(Z)f_*(Z)^\top]$ is unknown, the functions $\{f_{k*}\}_1^q$ are known. While the Gaussian approximation $\mathcal{Z}_*$ to T_* is not feasible, we may define the *Gaussian-symmetrized version* $\mathcal{W}_*$ of T_* by multiplying the $f_*(Z_i)$'s with i.i.d. standard normal random variables $\{\xi_i\}_1^n$:

$$\mathcal{W}_* \coloneqq \max_{1\leqslant k\leqslant q} \left|\sqrt{n}\mathbb{E}_n\left[f_{k*}(Z_i)\,\xi_i\right]\right|. \tag{1.5.24}$$

Chernozhukov et al. (2013) also show that the conditional quantiles of $\mathcal{W}_*$ given the data $\{Z_i\}_1^n$ are able to estimate the corresponding quantiles of $\mathcal{Z}_*$. Due to the Gaussian approximation linking $\mathcal{Z}_*$ and T_* and the probabilistic link between T_* and the test statistic T, for a given significance level $\alpha \in (0,1)$, the $(1-\alpha)$-conditional quantile of $\mathcal{W}_*$ may be used as a critical value for T. This method of inference is often referred to as a *Gaussian multiplier* (or *Wild*) *bootstrap.*

Neither method of inference proposed above is directly applicable when not only the covariance $\mathrm{E}[f_*(Z)f_*(Z)^\top]$ but f_* itself is unknown. However, a feasible critical value arises from replacing the unknown f_* by a consistent estimator $\widehat{f}$. For given $\widehat{f}_k$'s one may define a feasible analog $\mathcal{W}$ of $\mathcal{W}_*$ by

$$\mathcal{W} \coloneqq \max_{1\leqslant k\leqslant q} \left|\sqrt{n}\mathbb{E}_n\left[\widehat{f}_k(Z_i)\,\xi_i\right]\right|. \tag{1.5.25}$$

A feasible critical value $c_{\mathcal{W}}(\alpha)$ then follows from

$$c_{\mathcal{W}}(\alpha) \coloneqq (1-\alpha)\text{-quantile of } \mathcal{W} \text{ conditional on } \{Z_i\}_1^n.$$

Note that for given estimators $\{\widehat{f}_k\}_1^q$, the critical value $c_{\mathcal{W}}(\alpha)$ may be calculated via simulation of the multipliers $\{\xi_i\}_1^n$.

The $\widehat{f}_k$'s will in general depend on an estimate $\widehat{s}$ of the influence function s_* from Assumption 1.10. For specific models it may be possible to provide primitive conditions for $\mathcal{W}$ and $\mathcal{W}_*$ to be close in a probabilistic sense, but at this level of generality it seems impossible to give more than the high-level condition:

Assumption 1.15 (**Bootstrap Conditions**). *$\widehat{s}$ is a $\{Z_i\}_1^n$-measurable random element of $\{f : \mathbf{R}^{d_z} \to \mathbf{R}^d\}$. Moreover, there exists a strictly positive sequence b_n such that*

$$\mathrm{P}\left(\|\widehat{s} - s_*\|_{\mathbb{P}_n,2} > b_n\right) \leqslant C_2 n^{-c_2},$$

and $b_n \ln(qn) \leqslant C_2 n^{-c_2}$.

With estimators of the f_{k*}'s available, $\mathcal{W}$ is well defined. The following lemma shows that $\mathcal{W}$ may be used to approximate the probabilistic behavior of $\mathcal{W}_*$.

Lemma 1.6 (**Approximate Bootstrap Equivalence**). *If Assumptions 1.10–1.15 hold, then there exists c, C, C' and n_0 such that for $n \geqslant n_0$,*

$$\mathrm{P}\left(\mathrm{P}\big(\, |\mathcal{W} - \mathcal{W}_*| > \zeta_1' |\, \{Z_i\}_1^n\,\big) > Cn^{-c}\right) \leqslant Cn^{-c},$$

where

$$\zeta_1' \coloneqq C' \max\left\{\sqrt{\frac{s\ln^2(pqn)}{n}}, \frac{s^{3/2}\ln^{3/2}(pqn)}{n}, \frac{n^{-c_2/4}}{\sqrt{\ln(qn)}}, b_n\sqrt{\ln n}\right\}.$$

1.5.6 Large Sample Behavior of Test

The approximations from the previous two sections imply the third main result.

Theorem 1.5 (**Size Control**). *If Assumptions 1.10–1.15 hold, $\mathrm{E}[f_{k*}(Z)^2] \geqslant c_1^2$ for all k such that f_{k*} is not the zero function, and $\ln^7(qn) \leqslant C_2 n^{1-c_2}$, then there exists c, C and n_0 depending only on $c_0, c_0', c_1, C_1, c_2, C_2$ and c_2' such that for $n \geqslant n_0$,*

$$\sup_{\alpha\in(0,1)} |\mathrm{P}(T > c_{\mathcal{W}}(\alpha); \mathrm{H}_0) - \alpha| \leqslant Cn^{-c}.$$

Theorem 1.5 implies that the test that rejects if and only if $T > c_{\mathcal{W}}(\alpha)$ is asymptotically of size α. While exact size control is only guaranteed in the limit (as $Cn^{-c} \to 0$), the theorem establishes the stronger conclusion that size is approximately preserved in finite sample with an error in size decaying polynomially fast.

The proof of may be sketched as follows. Lemmas 1.5 and 1.6 provide the heuristics $T \approx T_*$ and $\mathcal{W}_* \approx \mathcal{W}$, respectively. The additional assumption that the f_{k*}'s are mean-square bounded away from zero is used to establish the validity of the Gaussian multiplier bootstrap under the null, thus providing the link $T_* \approx \mathcal{W}_*$ (under the null). Given that all approximations are done in finite-sample, at no point is convergence of T under the null required. The proviso "for all k such that f_{k*} is not the zero function" is used to allow cases where the Neyman orthogonalization procedure reduces the effective number of moments by setting some ψ_{k*}'s to zero (see Example 1.9).

To state the fourth and final main result, define

$$v_q \coloneqq \max_{1\leqslant k\leqslant q} |\mathrm{E}[\rho(Z, \beta_*, L_*(W)) X_k]|.$$

The number v_q is a measure of the degree to which the null hypothesis is violated.

Theorem 1.6 (Consistency). *If Assumptions 1.10–1.15 hold and* $\max_{1\leqslant k\leqslant q}\mathrm{E}[f_{k*}(Z)^2]\geqslant c_1^2$*, then for each* $\alpha\in(0,1)$*,*

$$\mathrm{P}\left(T>c_{\mathcal{W}}(\alpha);\mathrm{H}_1\right)\to 1,$$

provided $v_q^{-1}(\ln q)/\sqrt{n}\to 0$*.*

Theorem 1.6 states that the test that rejects the null if and only if $T>c_{\mathcal{W}}(\alpha)$ is consistent for any sequence of alternatives satisfying the condition $v_q^{-1}(\ln q)/\sqrt{n}\to 0$. This condition may be interpreted as the alternative not being too close to the null. (An alternative interpretation is that the instruments X are not too weak.) The theorem therefore states that under any sequence of alternatives which are well-separated from the null, the test will reject the null with probability approaching one. Sequences of alternatives failing to satisfy $v_q^{-1}(\ln q)/\sqrt{n}\to 0$ are not well-separated from the null, and may therefore go undetected.

1.6 Conclusion

In this paper I have proposed specification tests for two classes of econometric models: (1) semiparametric conditional moment restriction models depending on conditional expectation functions, and (2) a class of high-dimensional unconditional moment restriction models depending on high-dimensional best linear predictors. These classes may be motivatived by economic models in which agents make choices under uncertainty and therefore have to predict payoff-relevant variables such as prospects unknown at the time of the decision or the behavior of other agents. The proposed tests are shown to be both asymptotically correctly sized and consistent. Moreover, I establish a bound on the rate of local alternatives for which the test for high-dimensional unconditional moment restriction models is consistent. Both classes of models impose a minimum of structure on the predictions entering their paramerizations. These results therefore allow researchers to test the specification of their models without introducing ad hoc assumptions on expectation formation.

Chapter 1 Appendices

1.A Appendix Abbreviations

In statements and proofs appearing in the appendices, to conserve space I use the abbreviations CS, H, J, M and T for the Cauchy-Schwarz, Hölder, Jensen, Markov and triangle inequalities, respectively. CMT is short for the continuous mapping theorem. MVT and MVE are short for the mean-value theorem and a mean-value expansion, respectively. I also abbreviate "with probability approaching one" by wp $\to 1$ and "with probability at least a" by wp $\geqslant a$.

1.B Additional Motivational Examples

Example 1.7 (**Partially Linear Regression**). Consider the partially linear regression model (PLRM) of Robinson (1988),

$$Y = \beta_0 D + g_0 (W) + \varepsilon, \quad \mathrm{E} (\varepsilon | \, D, W) = 0, \tag{1.B.1}$$

where D represents a treatment whose effect on the outcome Y we are interested in quantifying, and W denotes covariates. Poterba, Venti and Wise (1994; 1995) analyzed the effect of 401(k) retirement savings plan eligibility on savings as measured by net financial assets. These authors (essentially) argued that, while working for a firm that offers access to a 401(k) plan cannot be viewed as randomly assigned, after controlling for income, 401(k) eligibility may be thought of as exogenous. One may therefore be willing to adopt a PLRM for their analysis. Taking expectations in (1.B.1) conditional on W and subtracting the resulting equation from (1.B.1), we may "partial out" $g(W)$ to arrive at

$$Y - \mathrm{E} (Y | \, W) = \beta_0 [D - \mathrm{E} (D | \, W)] + \varepsilon, \quad \mathrm{E} (\varepsilon | \, D, W) = 0. \tag{1.B.2}$$

Taking expectations conditional on (D, W) and rearranging, we are lead to the CMR

$$\mathrm{E}\left\{Y - \mathrm{E}\left(Y|\, W\right) - \beta_0 \left[D - \mathrm{E}\left(D|\, W\right)\right]|\, D, W\right\} = 0,$$

whose implied residual $Y - \mathrm{E}\left(Y|\, W\right) - \beta\left[D - \mathrm{E}\left(D|\, W\right)\right]$ depends on the conditional expectations $\mathrm{E}\left(Y|\, W\right)$ and $\mathrm{E}\left(D|\, W\right)$. While the PLRM controls for W in a flexible manner, it rules out any interaction effect.[30] In Poterba, Venti and Wise (1994; 1995) this condition would correspond to imposing that the partial effect of 401(k) eligibility on savings does not depend on income. As their empirical analysis demonstrates, the no-interaction condition is restrictive, and one may want to subject the PLRM to a specification test.

Example 1.8 (**Discrete Choice Under Uncertainty**)**.** Consider the following simplified version of the static discrete choice model of Manski (1991). An agent must make a (for simplicity) binary (i.e., "Yes/No") decision such as going to college or not, or whether to enter the labor force. The (ex post) utility from choosing alternative $j \in \{0, 1\}$ is

$$u\left(j, V, Y, \varepsilon\right) = \pi\left(j, V, Y\right) + \varepsilon\left(j\right),$$

where $\varepsilon \coloneqq \left(\varepsilon\left(0\right), \varepsilon\left(1\right)\right)$. The realizations of V and the ε are known to the agent at the time of decision, and may therefore be considered as payoff-relevant state variables. However, the (for simplicity) scalar variable Y is realized only after the time of decision and represents a future to the agent. The distribution of Y may depend on the belief-relevant state variables W as well as the decision. The agent holds the subjective probability distribution (belief) $P^s\left(y|\, w, j\right)$ capturing the perceived probability distribution of Y should the agent observing $W = w$ decide on j. The agent chooses the alternative that yields the highest subjective expected utility

$$\begin{aligned} E^s\left[u\left(j, v, Y, \varepsilon\right)|\, w, j\right] &\coloneqq \int u\left(j, v, y, \varepsilon\right) \mathrm{d}P^s\left(y|\, w, j\right) \\ &= \int \pi\left(j, v, y\right) \mathrm{d}P^s\left(y|\, w, j\right) + \varepsilon\left(j\right) \eqqcolon \Pi\left(j, v, w\right) + \varepsilon\left(j\right). \end{aligned}$$

Denote the agent's *optimal* action by

$$A \coloneqq A\left(V, W, \varepsilon\right) \coloneqq \underset{j\in\{0,1\}}{\operatorname{argmax}} \left\{\Pi\left(j, V, W\right) + \varepsilon\left(j\right)\right\}$$

[30] The PLRM framework may be extended to multiple treatments, $Y = D^\top \beta_0 + g\left(W\right) + \varepsilon, \mathrm{E}(\varepsilon|D, W) = 0$, which allows for the possibility of (parametric) interaction between the primary treatment of interest and some or all of the covariates.

The researcher observes the vector (A, V, W, Y) composed by the binary A indicating the decision, payoff-relevant state variables V, belief-relevant state variables W, and the future Y. Parameterize the "deterministic" part of the payoffs as

$$\pi(j, v, y) = \begin{cases} v^\top \theta_0 + \delta_0 y, & j = 1, \\ 0, & j = 0, \end{cases}$$

such that the outside option $(j = 0)$ is normalized to zero. Then

$$\Pi(j, v, w) = \begin{cases} v^\top \theta_0 + \delta_0 E^s(Y|w, j), & j = 1, \\ 0, & j = 0. \end{cases}$$

Suppose that expectations are fulfilled, such that $P^s = \mathrm{P}$ and therefore $E^s = \mathrm{E}$, where P and E denote the population probability distribution and expectation, respectively.[31] Assume for simplificty that the "stochastic" part of the payoffs $(\varepsilon(0), \varepsilon(1))$ are distributed i.i.d. Type 1 extreme value independently of the observable state variables, such that their difference is logistic. A calculation then shows that the conditional choice probability of the agent takes the form

$$\mathrm{P}(A = 1|V, W) = \frac{\exp[\Pi(1, V, W)]}{1 + \exp[\Pi(1, V, W)]} =: \text{logistic}[\Pi(1, V, W)],$$

which may be rearranged to yield the CMR

$$\mathrm{E}\left\{ A - \text{logistic}\left[V^\top \theta_0 + \delta_0 \mathrm{E}(Y|\, W, D = 1)\right] \middle|\, V, W \right\} = 0. \tag{1.B.3}$$

The implied residual $A - \text{logistic}\left(V^\top \theta + \delta \mathrm{E}(Y|\, W, A = 1)\right)$ depends on $\mathrm{E}(Y|\, W, A = 1) = \mathrm{E}(AY|\, W)/\mathrm{E}(A|\, W)$, a ratio of conditional expectations. The extreme value assumption is primarily used to express the CMR in a simple form. If we were to instead specify $(\varepsilon(0), \varepsilon(1))$ to be conditionally distributed according to a cdf $F(\varepsilon_0, \varepsilon_1|\, v, w; \gamma_0)$ known up to the parameter γ_0, then we would still be lead to a CMR. In any case, one may have misspecified the utility, omitted or confused payoff- and belief-relevant state variables, or inadequately specified the distribution of the individual heterogeneity. Due to these observations, it seems desirable to conduct a specification test.

[31] For example, studies of human capital investment may assume that expectations of the returns to schooling are fulfilled (see, e.g., Willis and Rosen 1979 and Fuller, Manski, and Wise 1982). Self-fulfilling expectations may be derived from the more primitive conditions: (a) the population involves a continuum of agents, (b) the realizations of the futures are independent across agents, and (c) expectations are rational in the sense that agents know the stochastic processes driving their environments (cf. Manski, 1991, p. 263).

The following examples are high-dimensional analogs of Examples 1.7, 1.8 and 1.1.

Example 1.9 (High-Dimensional Linear Regression). Consider the linear predictive model

$$Y = \beta_0 D + W^\top \delta_0 + \varepsilon, \quad \mathrm{E}\left[\varepsilon\left(D, W^\top\right)^\top\right] = \mathbf{0}_{(1+p)\times 1}, \tag{1.B.4}$$

where p denotes the dimension of W and δ_0. I allow for p to be (potentially much) larger than the sample size n available to the researcher, $p \gg n$, thus making (1.B.4) a high-dimensional linear model (HDLM). Here W may be thought of as a high-dimensional collection of transformations $P(\mathcal{W})$ of some underlying vector of control variables $\mathcal{W}$, and the term $W^\top\delta_0$ may therefore be thought of as a flexible way of controlling for $\mathcal{W}$ in measuring the effect of D on Y. In the 401(k)-savings setting of Poterba, Venti and Wise (1994; 1995) discussed in Example 1.7, $W^\top\delta_0$ corresponds to controlling for income using a flexible parametric form. Projecting Y onto W we arrive at

$$W^\top\left[\mathrm{E}\left(WW^\top\right)\right]^{-1}\mathrm{E}\left(WY\right) = \beta_0 W^\top\left[\mathrm{E}\left(WW^\top\right)\right]^{-1}\mathrm{E}\left(WD\right) + W^\top\delta_0,$$

and subtracting the result from (1.B.4) we get

$$Y - W^\top h_{1*} = \beta_0\left(D - W^\top h_{2*}\right) + \varepsilon, \quad \mathrm{E}[\varepsilon(D, W^\top)^\top] = \mathbf{0}_{(1+p)\times 1},$$

where I have defined $h_{1*} \coloneqq [\mathrm{E}(WW^\top)]^{-1}\mathrm{E}\left(WY\right)$ and $h_{2*} \coloneqq [\mathrm{E}(WW^\top)]^{-1}\mathrm{E}\left(WD\right)$. The previous display corresponds to the 'partialling out' step (1.B.2) of Example 1.7, the only difference being that I have swapped conditional expectations with linear projections. The previous display may be written as

$$Y = \beta_0 D + W^\top h_{1*} + \left(-\beta_0\right) W^\top h_{2*} + \varepsilon, \quad \mathrm{E}\left[\varepsilon\left(D, W^\top\right)^\top\right] = \mathbf{0}_{(1+p)\times 1},$$

which shows that, in estimating β_0, it is important to control for both variables that matter for predicting Y $\left(W^\top h_{1*}\right)$ and variables that matter for predicting D $\left(W^\top h_{2*}\right)$. A similar rationale underlies the double-selection approach to estimation of β_0 in the PLRM (Belloni et al., 2014b).[32] Given that $\mathrm{E}[\varepsilon(D - W^\top h_{2*})] = \mathrm{E}(\varepsilon D) - \mathrm{E}(\varepsilon W^\top)h_{2*} = 0$, defining $X \coloneqq$

[32] The post-double-selection estimator of β_0 arises from (1) using a variable selector to select the most important variables in a regression of Y on W; (2) using a variable selector to select the most important variables in a regression of D on W; and, (3) regressing Y on D and the union of W-variables selected in the two selection steps.

$(D, W^\top)^\top$ (1.B.4) we arrive at the high-dimensional UMR

$$\mathrm{E}\left\{\left[Y - W^\top h_{1*} - \beta_0 \left(D - W^\top h_{2*}\right)\right] X\right\} = \mathbf{0}_{(1+p)\times 1}.$$

Note that if $\mathrm{E}(\varepsilon|D, W) = 0$, such that (1.B.4) has a structural interpretation, then the previous display holds for not only $X = (D, W^\top)^\top$ but for any vector X of instruments, i.e., any transformation of (D, W). For the purpose of specification testing, one may employ a high-dimensional number of instruments potentially different than the regressors (D, W).

Example 1.10 (A High-Dimensional Model of Discrete Choice Under Uncertainty). A high-dimensional analog of Example 1.8 may be obtained from (1.B.3) in a manner similar to the one in which we obtained the high-dimensional linear model of Example 1.9 from the PLRM of Example 1.7, i.e., by replacing conditional expectations by high-dimensional best linear predictors. Adopting the high-dimensional linear predictive models (see Example 1.9)

$$\begin{aligned}\mathrm{E}\left[\left(AY - W^\top h_{1*}\right) W\right] &= \mathbf{0}_{p\times 1},\\ \mathrm{E}\left[\left(A - W^\top h_{2*}\right) W\right] &= \mathbf{0}_{p\times 1},\end{aligned}$$

inserting the high-dimensional linear predictors $W^\top h_{1*}$ and $W^\top h_{2*}$ into (1.B.3) in place of $\mathrm{E}\left(YD|\, W\right)$ and $\mathrm{E}\left(D|\, W\right)$, respectively, and replacing the act of conditioning by multiplication with a high-dimensional vector X of q instruments, we arrive at

$$\mathrm{E}\left(\left\{D - \text{logistic}\left[V^\top \theta_0 + \delta_0 \left(W^\top h_{1*}/W^\top h_{2*}\right)\right]\right\} X\right) = \mathbf{0}_{q\times 1}.$$

1.C Obtaining Pseudo True Parameters

In this section I illustrate how one may obtain pseudo-true parameters in the examples provided in Section 1.B.

1.C.1 Semiparametric Conditional Moment Models

Example 1.7 (continued) In the partially linear model, denote $\widetilde{Y} \coloneqq Y - \mathrm{E}\left(Y|\, W\right)$ and $\widetilde{D} \coloneqq D - \mathrm{E}\left(D|\, W\right)$. These residuals arise from mean-square projections of Y and D, respectively, onto square-integrable functions of W. Multiplying both sides of (1.B.2) by $\widetilde{D}$, computing the expectation and solving for β_0, we see that $\beta_0 = [\mathrm{E}(\widetilde{D}^2)]^{-1}\mathrm{E}(\widetilde{D}\widetilde{Y})$. Assuming D is not fully determined by W such that $\mathrm{E}(\widetilde{D}^2) = \mathrm{E}\{[D - \mathrm{E}\left(D|\, W\right)]^2\} > 0$, we may therefore take $\beta_* = [\mathrm{E}(\widetilde{D}^2)]^{-1}\mathrm{E}(\widetilde{D}\widetilde{Y})$. Clearly, $\beta_* = \beta_0$ under the null. Given consistent estimators of $\mathrm{E}(Y|W)$ and $\mathrm{E}(D|W)$, under some conditions one may construct a $\sqrt{n}$-consistent, asymptotically normal estimator of β_* (see Robinson 1988, Belloni et al. 2014b, and Chernozhukov et al. 2017).

Example 1.8 (continued) Denote $X \coloneqq (V, W)$ and let $r(X)$ be a $(d_\theta + 1)$-vector of instruments generated by the state variables V and W. Appealing to the conditional moment restriction (1.B.3), a pseudo-truth $\beta_* \coloneqq (\theta_*, \delta_*)$ may be taken as the assumed unique root of the map

$$(\theta, \delta) \mapsto \mathrm{E}\left[\left\{A - \text{logistic}\left[V^\top \theta + \delta \mathrm{E}\left(Y \middle| W, D = 1\right)\right]\right\} r(X)\right], \quad (\theta, \delta) \in \mathbf{R}^{d_\theta + 1}.$$

A root of such a map exists under regularity conditions. Uniqueness amounts to an identification condition. To see that β_* is pseudo-true, suppose that the null hypothesis holds for this model. Then there exists $\beta_0 \coloneqq (\theta_0, \delta_0)$ such that

$$\mathrm{E}\left\{A - \text{logistic}\left[V^\top \theta_0 + \delta_0 \mathrm{E}\left(Y \middle| W, D = 1\right)\right] \middle| X\right\} = 0.$$

By the assumption of a unique root and using iterated expectations it now follows that $\beta_* = \beta_0$. A consistent estimator for β_* may be constructed using the generalized method of moments (GMM) approach with a nonparametric first-step estimator for the conditional expectation function. See Newey (1994) and Chen, Linton, and Van Keilegom (2003) for regularity conditions ensuring $\sqrt{n}$-consistency and asymptotic normality.

1.D Obtaining Influence Functions

In this section I show how one may obtain the influence function of the parametric estimators in the examples provided in Section 1.B.

Example 1.7 (continued) In the PLR model, let $\widehat{\beta}$ be a two-step GMM estimator based on the moment function $m(z, \beta, h_*(w)) = \{y - h_{1*}(w) - \beta[d - h_{2*}(w)]\}[d - h_{2*}(w)]$ and some nonparametric estimators of $h_{1*}(W) = \mathrm{E}\left(Y \middle| W\right)$ and $h_{2*}(W) = \mathrm{E}\left(D \middle| W\right)$. Using the notation of Example 1.3, straightforward differentiation implies that

$$M_* = -\mathrm{E}\{[D - h_{2*}(W)]^2\}, \quad \delta_{1*}(W) = 0, \quad \delta_{2*}(W) = 0.$$

By (1.4.7) and (1.4.9), no adjustment for estimation of (h_{1*}, h_{2*}) is neeeded, so

$$s_*(z) = \left(\mathrm{E}\{[D - h_{2*}(W)]^2\}\right)^{-1} \{y - h_{1*}(w) - \beta_*[d - h_{2*}(w)]\}[d - h_{2*}(w)].$$

Example 1.8 (continued) Denote $X \coloneqq (V, W)$ and let $r(X)$ be a $(d_\theta + 1)$-vector of instruments. Let $\widehat{\beta} \coloneqq (\widehat{\theta}, \widehat{\delta})$ be a two-step GMM estimator based on the moment function $m(z, \beta, h_*(w)) = \{a - \text{logistic}[v^\top \theta + \delta h_{1*}(w)/h_{2*}(w)]\} r(x)$ and some nonparametric estimators of $h_{1*}(W) = \mathrm{E}(AY | W)$ and $h_{2*}(W) = \mathrm{E}(A| W)$. Using the notation of Example 1.3, differentiation implies that

$$\begin{aligned}
M_* &= -\mathrm{E}\Big\{ f'\big(V^\top \theta_* + \delta_* h_{1*}(W)/h_{2*}(W)\big) r(X) \left[V^\top, h_{1*}(W)/h_{2*}(W)\right] \Big\}, \\
\delta_{1*}(W) &= -\delta_* \mathrm{E}\Big[\ f'\big(V^\top \theta_* + \delta_* h_{1*}(W)/h_{2*}(W)\big) h_{2*}(W)^{-1} r(X) \Big| W \Big], \\
\delta_{2*}(W) &= \delta_* \mathrm{E}\Big\{ \ f'\big(V^\top \theta_* + \delta_* h_{1*}(W)/h_{2*}(W)\big) [h_{1*}(W)/h_{2*}^2(W)] r(X) \Big| W \Big\},
\end{aligned}$$

where f' denotes the derivative of the logistic function, $f'(u) \coloneqq \mathrm{e}^{-u}/(1+\mathrm{e}^{-u})^2 = \text{logistic}(u)[1-\text{logistic}(u)]$. Using (1.4.7) and (1.4.9), it follows that

$$\begin{aligned}
s_*(z) = -\left(M_*^\top M_*\right)^{-1} M_* \Big(& \left\{a - \text{logistic}\left[v^\top \theta + \delta h_{1*}(w)/h_{2*}(w)\right]\right\} r(x) \\
& + \delta_{1*}(w)\left[ay - h_{1*}(w)\right] + \delta_{2*}(w)\left[a - h_{2*}(w)\right] \Big).
\end{aligned}$$

1.E Orthogonalization in the High-Dimensional Linear Model

Example 1.9 (continued) In the case of the HDLM, the original moment functions are $[y - w^\top h_1 - \beta(d - w^\top h_2)]x_k$, where $x_1 = d$ and $x_k = w_{k-1}, k \in \{2, \ldots, 1+p\}$. Given that $\partial_{v_1}\rho(z, \beta_*, v_1, v_2) = -1$ and $\partial_{v_2}\rho(z, \beta_*, v_1, v_2) = \beta_*$, letting $e_k \in \mathbf{R}^{1+p}$ denote the kth elementary vector and noting that $W^\top e_k = W_k$, by (1.5.10) we must have

$$\begin{aligned}
\mu_{11*} &= [\mathrm{E}(WW^\top)]^{-1}\mathrm{E}\left[WD(-1)\right] && = -h_{2*}, \\
\mu_{12*} &= [\mathrm{E}(WW^\top)]^{-1}\mathrm{E}\left(WD\beta_*\right) && = \beta_* h_{2*}, \\
\mu_{k1*} &= [\mathrm{E}(WW^\top)]^{-1}\mathrm{E}\left[WW_{k-1}(-1)\right] && = -e_{k-1}, \quad k \in \{2, \ldots, 1+p\}, \\
\mu_{k2*} &= [\mathrm{E}(WW^\top)]^{-1}\mathrm{E}\left(WW_{k-1}\beta_*\right) && = \beta_* e_{k-1}, \quad k \in \{2, \ldots, 1+p\}.
\end{aligned}$$

Following (1.5.9) the first orthogonalized moment function evaluated at $(\beta_*, w^\top h_{1*} w^\top h_{2*}, w^\top \mu_{1*}, w^\top \mu_{2*})$ is given by

$$\psi_1\left(z, \beta_*, (w^\top h_{\ell*})_1^2, (w^\top \mu_{1\ell*})_{\ell=1}^2\right)$$

$$
\begin{aligned}
&= \rho(z,\beta_*,(w^\top h_{\ell *})_1^2)d + \left(y - w^\top h_{1*}\right) w^\top \left(-h_{2*}\right) + \left(d - w^\top h_{2*}\right) w^\top \left(\beta_* h_{2*}\right) \\
&= \rho(z,\beta_*,(w^\top h_{\ell *})_1^2)d - \left[y - w^\top h_{1*} - \beta_* \left(d - w^\top h_{2*}\right)\right] w^\top h_{2*} \\
&= \left[y - w^\top h_{1*} - \beta_* \left(d - w^\top h_{2*}\right)\right] \left(d - w^\top h_{2*}\right).
\end{aligned}
$$

Orthogonalized moment functions $k \in \{2, \ldots, 1+p\}$ are even simpler as

$$
\begin{aligned}
&\psi_k \left(z, \beta_*, (w^\top h_{\ell *})_1^2, (w^\top \mu_{k\ell *})_{\ell=1}^2\right) \\
&\quad = \left[y - w^\top h_{1*} - \beta_* \left(d - w^\top h_{2*}\right)\right] w_{k-1} \\
&\qquad + \left(y - w^\top h_{1*}\right) w^\top \left(-e_{k-1}\right) + \left(d - w^\top h_{2*}\right) w^\top \left(\beta_* e_{k-1}\right) = 0.
\end{aligned}
$$

The collapse of the latter collection of orthogonalization moment functions is due to the linearity of the residual function in the values $w^\top h_1$ and $w^\top h_2$. This linearity structure is very special, and one will in general have as many orthogonalized moment functions as original moment functions.

1.F Extensions

1.F.1 Semiparametric Conditional Moment Models

In this section I extend the framework of Section 1.4 to accommodate multiple CEFs as well as multiple CMRs.

1.F.1.1 Multiple Conditional Expectations

As illustrated by Examples 1.1, 1.7 and 1.8, models that involve a CEF, often involve *multiple* CEFs. For example, a firm considering entry in Example 1.1 typically must form an expectation with regards to more than one competitor. With multiple CEFs, the $h_*(W)$ appearing in the residual should be interpreted as a vector of conditional expectations $h_{\ell *}(W) = \mathrm{E}\left(Y_\ell | W_\ell\right), \ell \in \{1, \ldots, L\}$, where the Y_ℓ's and W_ℓ's are subvectors of Z and W now denotes the union of unique elements of the W_ℓ's. To accommodate a vector of conditional expectations, Assumptions 1.3 and 1.4 are modified as follows.

Assumption 1.3' *The residual function satisfies:*

1. *For each $z \in \mathcal{Z}, v \in \mathbf{R}^L, \beta \mapsto \rho(z, \beta, v)$ is continuous on $\mathcal{B}$ and continuously differentiable on an open neighborhood $\mathcal{N}_*$ of β_*. Moreover, there exists $c \in (0, \infty)$ and*

$L_1 : \mathcal{Z} \to \mathbf{R}_+$ integrable such that for each $z \in \mathcal{Z}, \beta \in \mathcal{N}_*, v \in \mathbf{R}^L$,

$$\|\partial_\beta \rho(z, \beta, v) - \partial_\beta \rho(z, \beta, h_*(w))\| \leqslant L_1(z) \|v - h_*(w)\|^c.$$

2. *For each $z \in \mathcal{Z}, v \mapsto \rho(z, \beta_*, s)$ is continuously differentiable on $\mathbf{R}^L$. Moreover, there exists $\gamma \in (0, 1]$ and $R : \mathcal{Z} \to \mathbf{R}_+$ such that for each $z \in \mathcal{Z}, v \in \mathbf{R}^L$,*

$$\|\partial_v \rho(z, \beta_*, v) - \partial_v \rho(z, \beta_*, h_*(w))\| \leqslant R(z) \|v - h_*(w)\|^\gamma,$$

where $\mathrm{E}[R(Z)] \sqrt{n} \max_{1 \leqslant \ell \leqslant L} \|\widehat{h}_\ell - h_{\ell *}\|_{\mathcal{W}}^{1+\gamma} \to_{\mathrm{P}} 0$.

3. *$|\rho(Z, \beta_*, h_*(W))|$, $\sup_{\beta \in \mathcal{N}_*} \|\partial_\beta \rho(Z, \beta, h_*(W))\|$ and $\|\partial_v \rho(Z, \beta_*, h_*(W))\|^2$ are integrable.*

Assumption 1.4' *$\max_{1 \leqslant \ell \leqslant L} \mathrm{var}(Y_\ell | W_\ell)$ is bounded.*

Each CEF $w_\ell \mapsto h_{\ell *}(w_\ell)$ may require its own set of approximating functions $w_\ell \mapsto p_\ell^k(w_\ell)$. Assumption 1.5 therefore becomes:

Assumption 1.5' *The eigenvalues of $\mathrm{E}[p_\ell^k(W_\ell) p_\ell^k(W_\ell)^\top]$ are bounded from above and away from zero uniformly over $\ell \in \{1, \ldots, L\}$ and $k \in \mathbf{N}$.*

Each CEF and associated approximating functions must now satisfy an approximation requirement.

Assumption 1.6' *The $h_{\ell *}$'s are bounded. Moreover, for each $\ell \in \{1, \ldots, L\}$ there exists a constant $\alpha_\ell \in (0, \infty)$ such that for each $k \in \mathbf{N}$ there is $\widetilde{\pi}_{\ell k} \in \mathbf{R}^k$ such that $\|\widetilde{h}_{\ell k} - h_{\ell *}\|_{\mathcal{W}_\ell} \lesssim k^{-\alpha_\ell}$ for the linear form $\widetilde{h}_{\ell k} \coloneqq p_\ell^{k\top} \widetilde{\pi}_{\ell k}$.*

Quantify the size of the ℓth set of approximating functions by

$$\zeta_{\ell,k} \coloneqq \sup_{w_\ell \in \mathcal{W}_\ell} \|p_\ell^k(w_\ell)\|,$$

where $\mathcal{W}_\ell$ denotes the support of W_ℓ. For the purpose of stating rate conditions, define

$$\delta_{\ell *}(t, W_\ell) \coloneqq \mathrm{E}\left[\omega(t, X) \partial_{v_\ell} \rho(Z, \beta_*, h_*(W)) | W_\ell\right],$$

where ∂_{v_ℓ} denotes the partial derivative with respect to the values of $h_{\ell *}(W_\ell)$. Define the mean-square projection coefficients

$$\pi_{h_\ell, k} \coloneqq \operatorname*{argmin}_{\pi \in \mathbf{R}^k} \mathrm{E}\{[p_\ell^k(W_\ell)^\top \pi - h_{\ell *}(W_\ell)]^2\},$$

$$\pi_{\delta_\ell,k}(t) \coloneqq \operatorname*{argmin}_{\pi\in\mathbf{R}^k} \mathrm{E}\{[p_\ell^k(W_\ell)^\top\pi - \delta_{\ell*}(t,W_\ell)]^2\},$$

and their induced mean-square errors

$$\begin{aligned}
r_{h_\ell,k}^2 &\coloneqq \mathrm{E}\{[p_\ell^k(W_\ell)^\top\pi_{h_\ell,k} - h_{\ell*}(W_\ell)]^2\} = \min_{\pi\in\mathbf{R}^k}\mathrm{E}\{[p_\ell^k(W_\ell)^\top\pi - h_{\ell*}(W_\ell)]^2\},\\
r_{\delta_\ell,k}^2(t) &\coloneqq \mathrm{E}\{[p_\ell^k(W_\ell)^\top\pi_{\delta_\ell,k}(t) - \delta_{\ell*}(t,W_\ell)]^2\} = \min_{\pi\in\mathbf{R}^k}\mathrm{E}\{[p_\ell^k(W_\ell)^\top\pi - \delta_{\ell*}(t,W_\ell)]^2\},\\
R_{\delta_\ell,k}^2 &\coloneqq \mathrm{E}\big[\|p_\ell^k(W_\ell)^\top\pi_{\delta_\ell,k}(\cdot) - \delta_{\ell*}(\cdot,W_\ell)\|_{\mathcal{T}}^2\big].
\end{aligned}$$

Assumption 1.7’ *For each $\ell\in\{1,\dots,L\}$ and the α_ℓ’s from Assumption 1.6’,*

$$\begin{aligned}
\zeta_{\ell,k_{\ell,n}} r_{h_\ell,k_{\ell,n}} &\to 0, & R_{\delta_\ell,k_{\ell,n}}\sqrt{\ln\left(k_{\ell,n}/R_{\delta_\ell,k_{\ell,n}}\right)} &\to 0,\\
R_{\delta_\ell,k_{\ell,n}} &\to 0, & \zeta_{\ell,k_n}^2 k_{\ell,n}\ln\left(k_{\ell,n}\right)/n &\to 0,\\
n r_{h_\ell,k_{\ell,n}}^2\|r_{\delta_\ell,k_{\ell,n}}\|_{\mathcal{T}}^2 &\to 0, & \Big(\sum_{j=1}^{k_{\ell,n}}\|p_{\ell jk_{\ell,n}}\|_{\mathcal{W}_\ell}^2\Big)^{1/2}\Big(\sqrt{k_{\ell,n}/n} + k_{\ell,n}^{-\alpha_\ell}\Big) &\to 0.
\end{aligned}$$

Lastly, I modify the bootstrap conditions as follows:

Assumption 1.8’ *(1) For each $z\in\mathcal{Z},\beta\in\mathcal{N}_*,v\mapsto\rho(z,\beta,v)$ is continuously differentiable on $\mathbf{R}^L$. Moreover, for each $z\in\mathcal{Z},\beta\in\mathcal{N}_*,v\in\mathbf{R}^L$,*

$$\|\partial_v\rho(z,\beta,v) - \partial_v\rho(z,\beta_*,h_*(w))\| \leqslant R'(z)\left(\|\beta-\beta_*\| + \|v-h_*(w)\|\right),$$

where $\mathrm{E}[R'(Z)]\sqrt{n}\max_{1\leqslant\ell\leqslant L}\|\widehat{h}_\ell - h_{\ell*}\|_{\mathcal{W}_\ell}^2 \to_{\mathrm{P}} 0$, (2) $\mathbb{E}_n[\|\widehat{s}(Z_i) - s_*(Z_i)\|^2] \to_{\mathrm{P}} 0$, and (3) $\max_{1\leqslant\ell\leqslant L}\zeta_{\ell,k_{\ell,n}}\sqrt{k_{\ell,n}}(\sqrt{k_{\ell,n}/n} + k_{\ell,n}^{-\alpha_\ell}) \to 0$.

Interpreting $\widehat{h}(w)$ as a vector of $\widehat{h}_\ell(w_\ell)$’s, we may define a test statistic T_n as in (1.4.5). Using Assumptions 1.1 and 1.2 and Assumptions 1.3’–1.8’ one may extend the argument used in proving Lemma 1.1 to show that the process $\{\sqrt{n}\mathbb{E}_n[\rho(Z_i,\widehat{\beta},\widehat{h}(W_i))\omega(t,X_i)]|t\in\mathcal{T}\}$ guiding the behavior of the resulting test statistic is asymptotically equivalent to $\{\sqrt{n}\mathbb{E}_n[f_*(t,Z_i)]|t\in\mathcal{T}\}$ with f_* redefined as

$$f_*(t,z) \coloneqq \rho(z,\beta_*,h_*(w))\,\omega(t,x) + b_*(t)^\top s_*(z) + \sum_{\ell=1}^L\delta_{\ell*}(t,w)\left[y_\ell - h_{\ell*}(w_\ell)\right].$$

One may now modify the Gaussian multiplier bootstrap of Section 1.4.5 to construct critical values, and the arguments of Theorems 1.2, 1.3, and 1.4 may be extended to establish size control and consistency of the resulting test. I omit the formal proofs for the vector case as they parallel their scalar counterparts.

1.F.1.2 Multiple Conditional Moment Restrictions

The null hypothesis in (1.4.3) is based on a single CMR. However, the underlying model may imply more than one CMR. For example, the entry game of Example 1.1 implies a collection of CMRs—one for each firm. By focusing on a single restriction, the test procedure presented above may fail to reject an inadequate econometric model. It therefore seems desirable to be able to test a finite collection of CMRs.

When the underlying implies M ($M \in \mathbf{N}$) CMRs, after passing to a pseudo-true parameter, the null hypothesis becomes

$$\mathrm{H}_0 : \forall m \in \{1, \ldots, M\}, \mathrm{E}\left[\rho_m\left(Z, \beta_*, h_{m*}\left(W_m\right)\right) \middle| X_m\right] = 0,$$

where each X_m is a subvector of Z containing W_m. With the help of weight functions $\{\omega_m\}_1^M$ and nuisance parameter spaces $\{\mathcal{T}_m\}_1^M$ all satisfying Assumption 1.2 and absolutely continuous, strictly positive, finite probability measures $\{\mu_m\}_1^M$, we may transform this null hypothesis into

$$\mathrm{H}_0 : \max_{1\leqslant m\leqslant M} \int_{\mathcal{T}_m} \left\{\mathrm{E}\left[\rho_m\left(Z, \beta_*, h_{m*}\left(W_m\right)\right)\omega_m\left(t_m, X_m\right)\right]\right\}^2 \mathrm{d}\mu_m\left(t_m\right) = 0.$$

Under a set of assumptions similar to Assumptions 1.1–1.7 we may therefore show that, under the null, the test statistic

$$T_n \coloneqq \max_{1\leqslant m\leqslant M} \int_{\mathcal{T}_m} \left\{\sqrt{n}\mathbb{E}_n\left[\rho_m(Z_i, \widehat{\beta}, \widehat{h}_m\left(W_{mi}\right))\omega_m\left(t_m, X_{mi}\right)\right]\middle|\right\}^2 \mathrm{d}\mu_m\left(t_m\right)$$

converges in distribution to $\max_{1\leqslant m\leqslant M} \int_{\mathcal{T}_m} G_{m0}\left(t_m\right)^2 \mathrm{d}\mu_m\left(t_m\right)$, where $\{G_{m0}\}_1^M$ denotes centered Gaussian processes with potentially different covariances functions. By an equivalence result similar to Lemma 1.1 one may extend the Gaussian multiplier bootstrap of Section 1.4.5 to obtain a correctly sized and consistent test.[33]

[33] Such an extension would have to take into account the dependence structure of the G_{m0}'s, which will in general not be independent.

1.F.2 High-Dimensional Unconditional Moment Models

In this section I extend the framework of Section 1.5 to accommodate multiple best linear predictors as well as multiple residuals.

1.F.2.1 Multiple High-Dimensional Predictors

Examples 1.7, 1.8 and 1.1 show that models that involve agents forming a prediction (in the examples: conditional expectation), typically include *multiple* predictions. For example, in the entry game (Example 1.1), the number of predictions needed to evaluate the residual function from the perspective of any particular firm equals the number of its competitors. Being the high-dimensional analogs of these examples, Examples 1.9, 1.10 and 1.2 involve multiple high-dimensional best linear predictors. With multiple high-dimensional best linear predictors, the $L_*(W)$ appearing in the residual should be interpreted as a vector of best linear predictors $L_{\ell*}(W_\ell) \coloneqq L(Y_\ell | W_\ell), \ell \in \{1, \ldots, L\}$, where the Y_ℓ's and W_ℓ's are subvectors of Z, W denotes the union of distinct elements of the W_ℓ's. Denote $p_\ell \coloneqq \dim(W_\ell)$ and $p \coloneqq \dim(W)$.

Let $M_j \coloneqq \sup \operatorname{supp}(|W_j|)$. To accommodate a vector of high-dimensional best linear predictors, I impose the following boundedness and moment conditions on the outcomes Y_ℓ, instruments X, regressors W, and the projection errors $\varepsilon_\ell \coloneqq Y_\ell - L_{\ell*}(W_\ell)$.

Assumption ' $|X_k| \leqslant C_1, |Y_\ell| \leqslant C_1$, $c_1^2 \leqslant \lambda_{\min}(\mathrm{E}(W_\ell W_\ell^\top)) \leqslant \lambda_{\max}(\mathrm{E}(W_\ell W_\ell^\top)) \leqslant C_1^2$, $c_1 \leqslant M_j \leqslant C_1$, $\mathrm{E}(\varepsilon_\ell^2 W_{\ell j}^2) \geqslant c_1^2$, *and*

$$\mathrm{P}\Big(\max_{1 \leqslant j \leqslant p} \big| \max_{1 \leqslant i \leqslant n} |W_{ij}| - M_j \big| > C_2 n^{-c_2} \Big) \leqslant C_2 n^{-c_2}.$$

Define the BLPs $\{L_{k\ell*}\}_{k,\ell}$ by $L_{k\ell*}(W_\ell) \coloneqq W_\ell^\top \mu_{k\ell*}$, where the projection coefficients are given by

$$\mu_{k\ell*} \coloneqq \left[\mathrm{E}(W_\ell W_\ell^\top)\right]^{-1} \mathrm{E}\left[W_\ell X_k \partial_{v_\ell} \rho(Z, \beta_*, L_*(W))\right],$$

and ∂_{v_ℓ} denotes the partial deritative with respect to the ℓth value $L_{\ell*}(w_\ell)$. Let $\varepsilon_{k\ell}$ denote the induced projection error $\varepsilon_{k\ell} \coloneqq X_k \partial_{v_\ell} \rho(Z, \beta_*, L_*(W)) - L_{k\ell*}(W_\ell)$. To accommodate a vector of high-dimensional best linear predictors, Assumption 1.13 now reads:

Assumption 1.13' *The residual function ρ satisfies:*

1. *For each $z \in \mathcal{Z}, v \in \mathbf{R}^L, \beta \mapsto \rho(z, \beta, v)$ is differentiable on $\mathbf{R}^d$, and for each $(z, \beta, v) \in$*

$\mathcal{Z} \times \mathbf{R}^{d+L}$ its derivative satisfies $\|\partial_\beta \rho(Z, \beta_*, L_*(W))\| \leqslant C_1$. and

$$\|\partial_\beta \rho(z, \beta, v) - \partial_\beta \rho(z, \beta_*, L_*(w))\| \leqslant C_1 (\|\beta - \beta_*\| + \|v - L_*(w)\|),$$

2. *For each $z \in \mathcal{Z}, v \mapsto \rho(z, \beta_*, v)$ is differentiable on $\mathbf{R}^L$, and for each $(z, v) \in \mathcal{Z} \times \mathbf{R}^L$ its derivative satisfies $|\partial_{v_\ell} \rho(Z, \beta_*, L_*(W))| \leqslant C_1$ and*

$$\|\partial_v \rho(z, \beta_*, v) - \partial_v \rho(z, \beta_*, L_*(w))\| \leqslant C_1 \|v - L_*(w)\|.$$

3. $\mathrm{E}[\rho(Z, \beta_*, L_*(W))^4] \leqslant C_1^4$, $\mathrm{E}(\varepsilon_{k\ell}^2 W_{\ell j}^2) \geqslant c_1^2$ *for all (k, ℓ) such that $X_k \partial_{v_\ell} \rho(Z, \beta_*, L_*(W)) \notin$* $\operatorname{span}(Y_l, W_l)$*, and $\|\mu_{k\ell*}\|_1 \leqslant C_1$.*

With multiple best linear predictors entering the residual, the Neyman orthogonalization now includes an adjustment for each best linear predictor,

$$\psi_k(z, \beta, \{w_\ell^\top h_\ell\}_1^L, \{w_\ell^\top \mu_{k\ell}\}_{k=1}^q) := \rho(z, \beta, \{w_\ell^\top h_\ell\}_1^L) x_k + \sum_{\ell=1}^{L} (y_\ell - w_\ell^\top h_\ell) w_\ell^\top \mu_{k\ell}.$$

This type of adjustment may be justified using the chain rule. Given an estimator $\widehat{\beta}$ and (Lasso) estimators of the $L_{\ell*}$'s and the $L_{k\ell*}$'s, the test statistic is defined exactly as in (1.5.11):

$$T := \max_{1 \leqslant k \leqslant q} \left|\sqrt{n}\mathbb{E}_n[\psi_k(Z_i, \widehat{\beta}, \{\widehat{L}_\ell(W_{\ell i})\}_{\ell=1}^L, \{\widehat{L}_{k\ell}(W_{\ell i})\}_{\ell=1}^L)]\right|.$$

Under a set of growth conditions similar to Assumption 1.14 (possibly modified to allow for growing L), this test statistic may be shown to be approximately equivalent to the maximum of an exact average, which is mean zero under the null. One may therefore obtain a critical value using the Gaussian multiplier bootstrap described in Section 1.5.5.

1.F.2.2 Multiple Residual Functions

The null hypothesis in (1.5.1) is based on a single residual. However, the underlying econometric model may imply more than candidate for a residual. For example, the entry game of Example 1.1 implies a collection of CMRs, thus yielding one residual for each firm. By focusing on the high-dimensional UMR implied by a single residual, the test procedure presented above may fail to reject an inadequate econometric model. It therefore seems desirable to be able to test a finite collection of high-dimensional UMRs.

When the underlying model implies M ($M \in \mathbf{N}$) candidate residuals, after passing to a pseudo-true parameter, the null hypothesis becomes

$$\mathrm{H}_0 : \forall m \in \{1, \ldots, M\}, \mathrm{E}\left[\rho_m\left(Z, \beta_*, L_{m*}\left(W_m\right)\right) X_m\right] = \mathbf{0}_{q_m \times 1},$$

where $L_{m*}(W_m)$ is the BLP of Y_m given W_m, both subvectors of Z, and X_m is a subvector of Z of length q_m. Taking the supremum norm, we may transform this null hypothesis into

$$\mathrm{H}_0 : \max_{1 \leqslant m \leqslant M} \max_{1 \leqslant k \leqslant q_m} \left|\mathrm{E}\left[\rho_m\left(Z, \beta_*, L_{m*}\left(W_m\right)\right) X_{mk}\right]\right| = 0.$$

A Neyman orthogonalization suggests the moment functions

$$\psi_{mk}(z, \beta, w_m^\top h_m,) \coloneqq \rho_m(z, \beta, w_m^\top h_m) x_{mk} + (y_m - w_m^\top h_m) w_m^\top \mu_{mk},$$

where the true $w_m^\top \mu_{mk*}$'s are given by

$$L_{mk*}\left(w_m\right) \coloneqq w_m^\top \mu_{mk*}, \quad \mu_{mk*} \coloneqq \left[\mathrm{E}(W_m W_m^\top)\right]^{-1} \mathrm{E}\left[W_m X_{mk} \partial_v \rho_m\left(Z, \beta_*, L_{m*}\left(W_m\right)\right)\right].$$

Given an estimator $\widehat{\beta}$ and (Lasso) estimators of the h_{m*}'s and μ_{mk*}'s, these orthogonalized moments may be used to construct a test statistic

$$T \coloneqq \max_{1 \leqslant m \leqslant M} \max_{1 \leqslant k \leqslant q_m} \left|\sqrt{n}\mathbb{E}_n[\psi_{mk}(Z_i, \widehat{\beta}, \widehat{L}_m\left(W_{mi}\right), \widehat{L}_{mk}\left(W_{mi}\right))]\right|.$$

Under a set of assumptions similar to Assumptions 1.10–1.14, one may show that T can be approximated by the maximum of an exact average vector of length $q \coloneqq \sum_{m=1}^M q_m$, which is mean-zero under the null. Using a high-dimensional central limit theorem for this exact average, one may therefore modify the Gaussian multiplier bootstrap procedure to obtain an approximately correctly sized and consistent test.

1.G Sparse Methods for Many Best Linear Predictors

In this section I develop a general framework for modeling best linear predictors by means of sparsity and propose a Lasso method for estimating very many best linear predictors. The results of this section are drawn upon in analyzing the procedure for testing a high-dimensional moment condition proposed in Section 1.5 but may also be of independent interest.

1.G.1 The Projection Model and Statement of the Problem

Let $Y = (Y_k)_1^q$ be an $\mathbf{R}^q$-valued random variable with $\mathrm{E}(Y_k^2) < \infty$ for all $k \in \{1, \ldots, q\}$ and $W = (W_j)_1^p$ an $\mathbf{R}^p$-valued random variable with $\mathrm{E}(W_j^2) < \infty$ for all $j \in \{1, \ldots, p\}$. Throughout this section I assume that $\{(Y_i, W_i)\}_1^n$ are n i.i.d. copies of (Y, W). I allow both p and q to depend on as well as exceed n. While the distribution of (Y, W) may depend on n, I will suppress such dependence throughout. Suppose that $\mathrm{E}(WW^\top)$ is invertible. Then one may define the *best linear predictor* $L_{k*}(W)$ of Y_k based on W by

$$L_k(w) \coloneqq w^\top \beta_{k*}, \quad \beta_{k*} \coloneqq \left[\mathrm{E}\left(WW^\top\right)\right]^{-1} \mathrm{E}\left(WY_k\right).$$

Given that β_{k*} is the unique solution to the the convex problem "minimize $\mathrm{E}[(Y_k - W^\top\beta)^2]$ subject to $\beta \in \mathbf{R}^p$," it must satisfy the first-order sufficient condition for a minimum: $\mathrm{E}[(Y_k - W^\top\beta_{k*})W] = \mathbf{0}_{p\times 1}$. If we define the *prediction errors* $\{\varepsilon_k\}_1^q$ by

$$\varepsilon_k \coloneqq Y_k - L_{k*}(W), \quad k \in \{1, \ldots, q\},$$

we therefore arrive at the tautologically true linear projection models:

$$Y_k = L_{k*}(W) + \varepsilon_k, \quad \mathrm{E}(\varepsilon_k W) = \mathbf{0}_{p\times 1}, \quad k \in \{1, \ldots, q\}. \tag{1.G.1}$$

In the special case where Y_k lies in the span of W, $Y_k \in \mathrm{span}(W)$, the best linear predictor of Y_k follows from the relevant linear combination. When these coefficients are known or can be solved for, this predictor can be estimated by itself. In this section I rule out $Y_k \notin \mathrm{span}(W)$ and focus on estimation of the q unknown best linear predictors.

1.G.2 Sparse Models for Best Linear Predictors

The potentially many regressors $\{W_j\}_1^p$ can be successfully employed under the key assumption of *sparsity.* For the sake of illustration, suppose that each best linear predictor function L_{k*} depends only on $s \ll n$ regressors. Then there exists $\beta_{k0} \in \mathbf{R}^p, k \in \{1, \ldots, q\}$, such that

$$L_{k*}(w) = w^\top \beta_{k0},\ k \in \{1, \ldots, q\},$$
$$\max_{1\leqslant k\leqslant q} \|\beta_{k0}\|_0 = \max_{1\leqslant k\leqslant q} \sum_{j=1}^p \mathbf{1}\left(\beta_{k0j} \neq 0\right) \leqslant s \ll n.$$

Note that the identity of each set of "active" regressors $T_{k0} \coloneqq \mathrm{supp}(\beta_{k0}) = \{j \in \{1, \ldots, p\}| \beta_{k0j} \neq 0\}$ may differ across k as well as be unknown to the researcher.

While this *exact sparsity* assumption is useful for illustration purposes, it is unlikely to hold in practice and unnecessarily restrictive. I will instead assume that the best linear predictors are *approximately sparse*. For the purpose of stating the assumption of approximate sparsity as well as assumptions to follow, let c_1, C_1, c_2 and C_2 be some given set of strictly positive, finite constants independent of n. The nonasymptotic, high-probability bounds obtained in this paper will depend on these constants.[34]

Assumption 1.16 (Approximately Sparse Best Linear Predictors). *There exists* $\{\beta_{k0}\}_1^q \subset \mathbf{R}^p$ *such that each best linear predictor is well-approximated by a function of unknown* $s \geqslant 1$ *regressors in the sense that*

$$\max_{1\leqslant k\leqslant q} \|\beta_{k0}\|_0 \leqslant s \ll n \quad \textit{and} \quad \mathrm{P}\left(c_s > C_1\sqrt{s/n}\right) \leqslant C_2 n^{-c_2},$$
$$\textit{where} \quad c_s \coloneqq \max_{1\leqslant k\leqslant q} \sqrt{\mathbb{E}_n\{[W_i^\top(\beta_{k0} - \beta_{k*})]^2\}}.$$

Assumption 1.16 requires that at most s regressors are able to approximate each best linear predictor function up to an approximation error, which is small with high probability. Defining the *sparse linear predictors*

$$w \mapsto L_{k0}(w) \coloneqq w^\top \beta_{k0}, \quad k \in \{1, \ldots, q\}, \tag{1.G.2}$$

we may express c_s as $c_s = \max_{1\leqslant k\leqslant q} \|L_{k0} - L_{k*}\|_{\mathbb{P}_n,2}$, which emphasizes that c_s is an error resulting from approximating *best* linear predictors by *sparse* linear predictors. Here c_s is considered "small" when it is not essentially larger than the size $\sqrt{s/n}$ of the estimation error arising from the infeasible least squares estimator that knows the identity of the most "important" regressors. One may view the L_{k0}'s as surrogate functions for the target functions $\{L_{k*}\}_1^q$.

Assumption 1.16 assumption roughly amounts to assuming that many of the elements of each β_{k*} are close to zero, i.e., that few regressors truly matter for prediction purposes. Note that this assumption allows for the identity of the most important regressors,

$$T_{k0} \coloneqq \operatorname{supp}(\beta_{k0}), \quad k \in \{1, \ldots, q\}, \tag{1.G.3}$$

to be unknown to the researcher as well as differ across k.

BCCH used an assumption almost identical to Assumption 1.16 in the context of esti-

[34] In principle, one may allow each of the conditions below to have their own set of constants and let the bounds depend on all these constants. To simplify the exposition, I reuse notation for constants that play a qualitatively similar role.

mation of conditional expectations. A detailed motivation and discussion of this type of assumption may be found in BCCH as well as Belloni and Chernozhukov (2011; 2013).

Defining the *approximation errors* $\{r_k\}_1^q$ by

$$r_k \coloneqq r_k(W) \coloneqq L_{k*}(W) - L_{k0}(W), \quad r_{ik} \coloneqq r(W_i), \quad k \in \{1, \ldots, q\},$$

we arrive at the approximately sparse linear models

$$Y_k = L_{k0}(W) + r_k + \varepsilon_k, \quad \mathrm{E}(\varepsilon_k W) = \mathbf{0}_{p\times 1}, \quad k \in \{1, \ldots, q\},$$

where $\max_{1 \leqslant k \leqslant q}[\mathbb{E}_n(r_{ik}^2)]^{1/2} \leqslant C_1\sqrt{s/n}$ wp $\geqslant 1 - C_2 n^{-c_2}$ by Assumption 1.16.

1.G.3 Lasso Estimation of Many Best Linear Predictors with Estimated Outcomes

Suppose that Y is *not* observable, but that each Y_i may be estimated by some $\widehat{Y}_i$. In the notation of Section 1.5, $Y_k = X_k \partial_v \rho(Z, \beta_*, W^\top h_*)$, which may be estimated by substituting estimators $\widehat{\beta}$ and $\widehat{h}$ for the unknowns (β_*, h_*). Defining the *outcome estimation error* e_{ik} by $e_{ik} \coloneqq \widehat{Y}_{ik} - Y_{ik}$, we may write

$$\widehat{Y}_{ik} = L_{k0}(W_i) + e_{ik} + r_{ik} + \varepsilon_{ik}, \quad \mathrm{E}(\varepsilon_{il} x_i) = \mathbf{0}_{p\times 1}, \quad k \in \{1, \ldots, q\}. \tag{1.G.4}$$

I will make use of the following high-level assumption in order to control the error arising from using the estimated and not necessarily true outcomes.

Assumption 1.17 (Outcome Estimation). $|e_{ik}| \leqslant C_1$ *and*

$$\mathrm{P}\left(\Delta > C_1\sqrt{s\ln(pqn)/n}\right) \leqslant C_2 n^{-c_2}, \quad \textit{where} \quad \Delta \coloneqq \max_{1\leqslant k \leqslant q} \sqrt{\mathbb{E}_n(e_{ik}^2)}.$$

The estimation error term Δ in Assumption 1.17 plays a role qualitatively similar to the approximation error term c_s from Assumption 1.16.

Given that the number p of parameters in each β_{k*} may exceed the sample size n, the use of machine learning or regularization methods appears unavoidable. A particularly popular machine learning method is the Lasso (Tibshirani, 1996), which uses regularization to simultaneously carry out estimation and variable selection in the context of regression.[35] For each $k \in \{1, \ldots, q\}$, the Lasso estimator $\widehat{L}_k$ of L_{k0} (and thus of L_{k*}) is defined by $\widehat{L}_k(w) \coloneqq w^\top \widehat{\beta}_k$,

[35] The name "Lasso" is an acronym for *L*east *a*bsolute *s*hrinkage and *s*election *o*perator.

where $\widehat{\beta}_k$ is a solution to the penalized least squares problem

$$\widehat{\beta}_k \in \operatorname*{argmin}_{\beta \in \mathbf{R}^p} \Big\{ \mathbb{E}_n[(\widehat{Y}_{ik} - W_i^\top \beta)^2] + \frac{\lambda}{n} \|\widehat{\Upsilon}_k \beta\|_1 \Big\}, \tag{1.G.5}$$

$\lambda \geqslant 0$ is a penalty level common to all q Lasso problems, and each $\widehat{\Upsilon}_k \coloneqq \operatorname{diag}(\widehat{\gamma}_{k1}, \ldots, \widehat{\gamma}_{kp})$ a diagonal matrix specifying penalty loadings to be described in further detail below.

The analysis will be centered around the "conservatively ideal" penalty loadings

$$\widehat{\Upsilon}_k^* \coloneqq \operatorname{diag}(\widehat{\gamma}_{k1}^*, \ldots, \widehat{\gamma}_{kp}^*),\ \widehat{\gamma}_{kj}^* \coloneqq \sqrt{\mathbb{E}_n\left(\varepsilon_{ik}^2\right)} \max_{1 \leqslant i \leqslant n} |W_{ij}|\,,\ (j,k) \in \{1, \ldots, p\} \times \{1, \ldots, q\}\,, \tag{1.G.6}$$

and the "truly ideal" penalty loadings

$$\widehat{\Upsilon}_k^{**} \coloneqq \operatorname{diag}(\widehat{\gamma}_{k1}^{**}, \ldots, \widehat{\gamma}_{kp}^{**}),\ \widehat{\gamma}_{kj}^{**} \coloneqq \sqrt{\mathbb{E}_n\left(\varepsilon_{ik}^2 W_{ij}^2\right)},\ (j,k) \in \{1, \ldots, p\} \times \{1, \ldots, q\}\,, \tag{1.G.7}$$

Use of "ideal" penalty loadings leads to sharp theoretical bounds on estimation risk. Neither the conservatively nor truly ideal penalty loadings are feasible since the ε_k's are unobservable. In practice one can estimate the ideal loadings using preliminary, conservative loadings and then inserting the resulting residuals in place of the ε_{ik}'s to obtain refined loadings. A procedure for initial and refined estimation of the penalty loadings is given in Algorithm 1.1.

The idea behind the "truly ideal" penalty loadings is to introduce self-normalization to the first-order condition of the Lasso minimization problem by using data-dependent penalty loadings. Self-normalization, in turn, allows use of moderate deviation inequalities for self-normalized sums as in de la Pena, Lai, and Shao (2009). Self-normalization via penalty loadings was first introduced by BCCH in the context of estimation of conditional expectations using the Lasso or Post-Lasso (least squares following Lasso selection).

In the present context $\varepsilon_k = Y_k - L_{k*}(W)$, where $L_{k*}(W)$ is the best linear predictor of Y_k. The best linear predictor need not coincide with the conditional expectation $\mathrm{E}(Y_k|W)$. The reason for not immediately focusing on "truly ideal" penalty loadings is that the definition of a best linear predictor as a linear projection does not suggest a conservative initial estimate of $\mathbb{E}_n(\varepsilon_{ik}^2 W_{ij}^2)$. The definition of a best linear predictor *does*, however, suggest a conservative initial estimate of $\mathbb{E}_n(\varepsilon_{ik}^2)$, since $\mathrm{E}(\varepsilon_k^2) = \mathrm{E}[(Y_k - W^\top \beta_{k*})^2] \leqslant \mathrm{E}[(Y_k - W^\top \beta)^2]$ for all $\beta \in \mathbf{R}^p$ by definition. This observation, in turn, suggests bounding a "truly ideal" penalty $[\mathbb{E}_n(\varepsilon_{ik}^2 W_{ij}^2)]^{1/2}$ by $[\mathbb{E}_n(\varepsilon_{ik}^2)]^{1/2} \max_{1 \leqslant i \leqslant n} |W_{ij}|$. If the best linear predictor were to coincide with the conditional expectation $\mathrm{E}(Y_k|W)$, then one could obtain a conservative initial estimate of $\mathbb{E}_n(\varepsilon_{ik}^2 W_{ij}^2)$ exploiting the inequality $\mathrm{E}(\varepsilon_k^2 W_j^2) = \mathrm{E}\{[Y_k - \mathrm{E}(Y_k|W)]^2 W_j^2\} \leqslant$

$\mathrm{E}\{[Y_k - f(W)]^2 W_j^2\}$ for any $j \in \{1, \ldots, p\}$ and any function f of W (as in BCCH).

The moderate deviation inequalities of de la Pena, Lai, and Shao (2009) yield a bound on maximal element of the $\widehat{\Upsilon}_k^*$-normalized score vectors defined by

$$S_k^* \coloneqq 2(\widehat{\Upsilon}_k^*)^{-1}\mathbb{E}_n\left(W_i \varepsilon_{ik}\right), \quad k \in \{1, \ldots, q\}, \tag{1.G.8}$$

which capture the "noise" of the estimation problem. Specifically, the moderate deviation theory implies that for any significance level $\alpha \in (0,1)$, there exists a finite constant C such that for n sufficiently large,

$$\mathrm{P}\left(\max_{1\leqslant k\leqslant q} \|\sqrt{n} S_k^*/2\|_\infty > \Phi^{-1}\left(1 - \alpha/\left(2pq\right)\right)\right) \leqslant C\alpha. \tag{1.G.9}$$

To guarantee good behavior of the Lasso estimators, one must choose a penalty level λ/n that overrules the noise from all score vectors $\{S_k^*\}_1^q$ simultaneously such that $\lambda/n \geqslant c_0 \max_{1\leqslant k\leqslant q} \|S_k^*\|_\infty$ with high probability for some constant $c_0 > 1$. Expressing the significance level as a polynomially decreasing function $n^{-c_0'}$ in n, the previous display shows that

$$\mathrm{P}\left(\lambda/n < c_0 \max_{1\leqslant k\leqslant q} \|S_k^*\|_\infty\right) \leqslant C n^{-c_0'}$$

can be achieved at least for n sufficiently large by setting the penalty level

$$\lambda \coloneqq 2c_0\sqrt{n}\Phi\left(1 - n^{-c_0'}/\left(2pq\right)\right), \tag{1.G.10}$$

where $c_0 > 1$ and $c_0' > 0$ are user-specified constants.[36] Given that $\widehat{\gamma}_{kj}^{**2} = \mathbb{E}_n(\varepsilon_{ik}^2 W_{ij}^2) \leqslant \mathbb{E}_n(\varepsilon_{ik}^2) \max_i W_{ij}^2 = \widehat{\gamma}_{kj}^{*2}$, we must have $|S_{kj}^*| \geqslant |S_{kj}^{**}|$, where the S_k^{**}'s are the $\widehat{\Upsilon}_k^{**}$-normalized score vectors given by

$$S_k^{**} \coloneqq 2(\widehat{\Upsilon}_k^{**})^{-1}\mathbb{E}_n\left(W_i \varepsilon_{ik}\right), \quad k \in \{1, \ldots, q\}. \tag{1.G.11}$$

Consequently, whenever the penalty level overrules the noise stemming from $\{S_k^*\}_1^q$, it also overrules the noise stemming from $\{S_k^{**}\}_1^q$.

[36] BCCH recommend setting $c_0 = 1.1$. Further simulation evidence is needed to determine a reasonable value for $c_0' > 0$.

1.G.4 Regularity Conditions for Estimating Best Linear Predictors

The performance of the Lasso estimators hinge crucially on the empirical Gram matrix $\mathbb{E}_n(W_iW_i^\top)$ being well-behaved. Given that the rank of this matrix is bounded by $p \wedge n$, when $p > n$, the $p \times p$ matrix $\mathbb{E}_n(W_iW_i^\top)$ must necessarily be singular. However, due to the assumption of approximate sparsity (Assumption 1.16), good performance of the Lasso estimator only requires that the empirical Gram matrix is well-behaved for "small" submatrices. To formalize this idea, define the *compatibility constant* $\kappa(a)$ by

$$\kappa(a) \coloneqq \min_{1 \leqslant |T| \leqslant s} \min_{\substack{\delta \neq \mathbf{0} \\ \|\delta_{T^c}\|_1 \leqslant a\|\delta_T\|_1}} \frac{\sqrt{s}[\delta^\top \mathbb{E}_n(W_iW_i^\top)\delta]^{1/2}}{\|\delta_T\|_1}, \quad a > 0,$$

where s is the sparsity index from Assumption 1.16, and T is understood to be a subset of $\{1, \ldots, p\}$. The compatibility constant $\kappa(a)$ may depend on n, although this dependence is suppressed. Good performance of the Lasso estimates can be ensured provided that the compatibility constant is bounded away from zero at least with high probability and for a suitable choice of $a > 0$. For the purposes of bounding the compatibility constant away from zero, define the *maximal* and *minimal sparse eigenvalues* of the empirical Gram matrix by

$$\phi_{\max}(m) \coloneqq \max_{1 \leqslant \|\delta\|_0 \leqslant m} \frac{\delta^\top \mathbb{E}_n(W_iW_i^\top)\delta}{\|\delta\|_2^2}, \tag{1.G.12}$$

$$\phi_{\min}(m) \coloneqq \min_{1 \leqslant \|\delta\|_0 \leqslant m} \frac{\delta^\top \mathbb{E}_n(W_iW_i^\top)\delta}{\|\delta\|_2^2}. \tag{1.G.13}$$

Under some conditions on the design, compatibility constant may be bounded away from zero using the minimal and maximal sparse eigenvalues (see Lemma 1.30). To state these design conditions, denote $M_j \coloneqq \sup \operatorname{supp}(|W_j|)$. I impose the following boundedness and moment conditions on the outcomes Y, regressors W and projection errors ε_k.

Assumption 1.18 (Observables and Errors). *$|Y_k| \leqslant C_1$,$c_1 \leqslant M_j \leqslant C_1$, $c_1^2 \leqslant \lambda_{\min}(\mathrm{E}(WW^\top)) \leqslant \lambda_{\max}(\mathrm{E}(WW^\top)) \leqslant C_1^2$, $|\varepsilon_k| \leqslant C_1$, $\mathrm{E}(\varepsilon_k^2 W_j^2) \geqslant c_1^2$, and*

$$\mathrm{P}\Big(\max_{1 \leqslant j \leqslant p} \Big| \max_{1 \leqslant i \leqslant n} |W_{ij}| - M_j \Big| > C_2 n^{-c_2}\Big) \leqslant C_2 n^{-c_2}. \tag{1.G.14}$$

The condition that the population Gram matrix $\mathrm{E}(WW^\top)$ has eigenvalues bounded from above and away from zero is common in the econometrics literature; see, e.g., Newey (1997) and Belloni et al. (2015). For the sake of analyzing the Lasso, the assumptions of bounded outcomes and errors are less standard but may be substantially relaxed at the expense of longer proofs. Specifically, boundedness of the ε_k's may be replaced by some tail bound making extreme events unlikely. An example of random variables satisfying a tail bound is the class of *subgaussian* random variables, whose tails are no fatter than normal random variables.

The assumption of bounded regressors ($M_j \leqslant C_1$) appears essential to establishing that the penalty loadings constructed via Algorithm 1.1 are close to being ideal with high probability. This dependence on boundedness stems from the appearance of $\max_{1 \leqslant i \leqslant n} |W_{ij}|$ in the conservatively ideal penalty loadings (1.G.6), which are used as target for the penalty loadings used to initiate the algorithm. It may be possible to devise an algorithm that does not rely on boundedness of the regressors, but such a task is beyond the scope of this paper.

The requirement that the lower bound inside the probability statement of (1.G.14) is equal to the right-hand side bound of the same equation is immaterial; were the two bounds to differ, then one may always proceed with the largest of the two bounds. The following example shows that the requirement in (1.G.14) can be satisfied even when p grows exponentially fast with n.

Example 1.11 (**Plausibility of Absolute Order Statistic Convergence**)**.** Suppose that $\{W_i\}_1^n$ are independent across i, $W_{ij} \sim \mathrm{U}[0,1]$ for all (i,j), and $p \geqslant 2$. Suppose also that there exists $c \in (0,1)$ and $C > 0$ such that $\ln p \leqslant Cn^{1-c}$. Then $M_j = 1$ for all j and $\max_{1 \leqslant i \leqslant n} |W_{ij}|$ equals the order statistic $W_{(n)j} \coloneqq \max_{1 \leqslant i \leqslant n} W_{ij}$, which is $\mathrm{Beta}(n,1)$ distributed for all j. In particular, $\mathrm{E}[W_{(n)j}] = n/(n+1)$. A $\mathrm{Beta}(\alpha,\beta)$ random variable is subgaussian with (optimal) subgaussianity parameter $\sigma^2(\alpha,\beta)$ bounded by $1/[4(\alpha+\beta+1)]$ (cf. Lemma 1.42). From this bound it follows that $W_{(n)j}$ is subgaussian with (optimal) subgaussianity $\sigma^2(n)$ parameter bounded by $1/[4(n+2)]$. By a maximal inequality for subgaussian random variables (Lemma 1.43) we therefore see that

$$\mathrm{E}\left[\max_{1 \leqslant j \leqslant p}\left|W_{(n)j} - \frac{n}{n+1}\right|\right] \leqslant \sigma(n)\sqrt{2\ln(2p)} \leqslant \frac{1}{4(n+2)}\sqrt{4\ln p} \leqslant \frac{1}{2}\sqrt{\frac{\ln p}{n}},$$

where I have used $p \geqslant 2$. Given that $|n/(n+1) - 1| \leqslant 1/n$, by the triangle inequality it follows that

$$\mathrm{E}\left[\max_{1 \leqslant j \leqslant p}\left|W_{(n)j} - 1\right|\right] \leqslant \frac{1}{2}\sqrt{\frac{\ln p}{n}} + \frac{1}{n} \leqslant \frac{\sqrt{C}}{2}n^{-c/2} + \frac{1}{n} \leqslant C'n^{-c/2},$$

where $C' \coloneqq \sqrt{C}/2 + 1$. Markov's inequality therefore shows that for any $\alpha \in (0, c/2)$ and any $D > 0$,

$$\mathrm{P}\left(\max_{1\leqslant j\leqslant p}\left|W_{(n)j}-1\right| > Dn^{-\alpha}\right) \leqslant \frac{C'}{D}n^{-(c/2-\alpha)},$$

which declines polynomially fast in n. Choosing $\alpha = c/4$ and $D = \sqrt{C'}$, we see that $\max_{1\leqslant j\leqslant p}|W_{(n)j}-1| \leqslant \sqrt{C'}n^{-c/4}$ with probability $\geqslant 1-\sqrt{C'}n^{-c/4}$.

The properties of the Lasso estimators relies on the following growth conditions.

Assumption 1.19 (**Lasso Growth Conditions**). *$s\ln^5(pqn) \leqslant C_2 n^{1-c_2}$ and $\ln(pqn) \leqslant C_2 n^{1-c_2'}$ for some $c_2' \in \left(\frac{2}{3},1\right)$.*

The requirement $\ln(pqn) \leqslant C_2 n^{1-c_2'}$ for some $c_2' \in \left(\frac{2}{3},1\right)$ implies that while p and q may grow exponentially fast with n, they cannot grow *too* fast. Although the requirement $c_2' \in \left(\frac{2}{3},1\right)$ was not explicitly stated in BCCH, it appears necessary in order to guarantee the validity of moderat deviation inequalities for self-normalized sums (see Appendix 1.P.1 and, in particular, the proof of Lemma 1.44).

1.G.5 Results on Lasso for Estimating Many Best Linear Predictors

In this section I consider the Lasso estimators of best linear predictors defined in (1.G.5) in the potentially high-dimensional system of projection equations (1.G.1) as well as estimated outcomes. These results extend the previous results of BCCH and Belloni and Chernozhukov (2011; 2013) for Lasso estimation of CEFs with nongaussian and heteroskedatic structural errors. Moreover, throughout the analysis I account for the fact that I am simultaneously estimating a potentially high-dimensional number (q) of equations. In particular, in establishing the validity of the penalty loadings constructed using Algorithm 1.1, I account for the dependence of my results on q.[37]

To state the rate results for the Lasso, call a set of penalty loadings $\widehat{\Upsilon}_k = \operatorname{diag}(\widehat{\gamma}_{k1},\ldots,\widehat{\gamma}_{kp})$ *conservatively polynomially valid* if there exists ℓ, u, c, C and n_0 depending only on c_0, c_0', c_1, C_1, c_2, C_2 and c_2' such that for all $n \geqslant n_0$,

$$\ell\widehat{\gamma}_{kj}^* \leqslant \widehat{\gamma}_{kj} \leqslant u\widehat{\gamma}_{kj}^* \text{ for all } j \in \{1,\ldots,p\} \tag{1.G.15}$$

[37] BCCH also account for the dependence of their results on the number of regressions (their k_e) but treated this number as fixed in establishing the validity of their penalty loadings (see their proof of their Lemma 11).

with probability $\geqslant 1 - Cn^{-c}$, where $0 < \ell \leqslant 1 \leqslant u$ and both $\ell \to 1$ and $u \to u' \geqslant 1$ polynomially fast. The initial penalty loadings arising from Algorithm 1.1 satisfy this requirement uniformly over $k \in \{1, \ldots, q\}$.

Similarly, call a set of penalty loadings $\widehat{\Upsilon}_k = \operatorname{diag}(\widehat{\gamma}_{k1}, \ldots, \widehat{\gamma}_{kp})$ *truly polynomially valid* if there exists ℓ, u, c, C and n_0 depending only on $c_0, c_0', c_1, C_1, c_2, C_2$ and c_2' such that for all $n \geqslant n_0$,

$$\ell \widehat{\gamma}_{kj}^{**} \leqslant \widehat{\gamma}_{kj} \leqslant u \widehat{\gamma}_{kj}^{**} \text{ for all } j \in \{1, \ldots, p\} \tag{1.G.16}$$

with probability $\geqslant 1 - Cn^{-c}$, where $0 < \ell \leqslant 1 \leqslant u$ and now both $\ell \to 1$ and $u \to 1$ polynomially fast. The refined penalty loadings arising from Algorithm 1.1 satisfy this requirement uniformly over $k \in \{1, \ldots, q\}$.

The reason for calling the loadings in (1.G.15) and (1.G.16) "conservatively" respectively "truly" valid is that they come close to the conservatively and truly ideal loadings from (1.G.6) and (1.G.7). The truly ideal penalty loadings induce self-normalization of the Lasso first order conditions, while the conservatively ideal loadings deflate the Lasso first order conditions by *more* than what would induce self-normalization. [See also the discussion following (1.G.6) and (1.G.7).]

The following theorem characterizes the behavior of the Lasso.

Theorem 1.7 (**Nonasymptotic, Polynomially Valid Bound for Lasso Estimation of Many Best Linear Predictors**). *Suppose that Assumptions 1.16–1.19 hold and that the penalty level λ is specified as in (1.G.10) for some c_0 and c_0'. Consider any conservatively or truly polynomially valid penalty loadings $\{\widehat{\Upsilon}_k\}_1^q$, for example, the penalty loadings resulting from Algorithm 1.1. Then there exists c, C, C' and n_0 depending only on $c_0, c_0', c_1, C_1, c_2, C_2$ and c_2' such that for all $n \geqslant n_0$, with probability $\geqslant 1 - Cn^{-c}$,*

$$\max_{1 \leqslant k \leqslant q} \|\widehat{L}_k - L_*\|_{\mathbb{P}_n,2} \leqslant C' \sqrt{\frac{s \ln(pqn)}{n}}.$$

Theorem 1.7 provides a nonasymptotic, high-probability bound for Lasso estimation of many best linear predictors. Provided $s \ln(pqn)/n \to 0$, the theorem implies the rate of convergence result

$$\max_{1 \leqslant k \leqslant q} \|\widehat{L}_k - L_*\|_{\mathbb{P}_n,2} \lesssim_{\mathrm{P}} \sqrt{\frac{s \ln(pqn)}{n}}. \tag{1.G.17}$$

The theorem and its rate corollary (1.G.17) parallel the nonasymptotic bound and rate result in BCCH (their Lemma 6 and Theorem 1, respectively). The main difference between the two pairs of results is that, compared to CEFs, having best linear predictors as estimands place less structure on the problem. Specifically, in the algorithm proposed for constructing penalty loadings (Algorithm 1.1) I take into account that the projection errors $\{\varepsilon_k\}_1^q$ do not necessarily exhibit a conditional-mean-zero property, $\mathrm{E}(\varepsilon_k|W) = 0$. In the proof of Lemmas 1.7 and 1.8, which establish the validity of the penalty loadings arising from Algorithm 1.1, I explicitly take into account that the number of estimands (best linear predictors) q may be high-dimensional (see also Footnote 37).[38]

1.H Implementation Details

In this appendix I present implementation algorithms for the Lasso procedures described in Sections 1.G and 1.5.

1.H.1 Implementation Details for Section 1.G

For any $m \in \mathbf{N}$, let $[m]$ denote the set $[m] \coloneqq \{1, \ldots, m\}$. Feasible options for setting the penalty level and loadings for $(j, k) \in [p] \times [q]$ are:

$$\text{Level: } \lambda \coloneqq 2c_0\Phi^{-1}(1 - n^{-c_0'}/(2pq)), \tag{1.H.1}$$

$$\text{Initial Loadings: } \widehat{\gamma}_{kj} \coloneqq \sqrt{\mathbb{E}_n\{[\widehat{Y}_{ik} - \mathbb{E}_n(\widehat{Y}_{ik})]^2\}} \max_{1 \leqslant i \leqslant n} |W_{ij}|\,, \quad (j,k) \in [p] \times [q]\,, \tag{1.H.2}$$

$$\text{Refined Loadings: } \widehat{\gamma}_{kj} \coloneqq \sqrt{\mathbb{E}_n(\widehat{\varepsilon}_{ik}^2 W_{ij}^2)}, \quad (j,k) \in [p] \times [q]\,, \tag{1.H.3}$$

where $\widehat{\varepsilon}_{ik}$ is an estimate of ε_{ik}, and $\widehat{Y}_{ik} = Y_{ik}$ if Y_{ik} is observed. Here $c_0 > 1$ and $c_0' \in (0, 1)$ are user-specified constants. BCCH (2012, p. 2380) recommend setting the constant $c_0 = 1.1$.

Algorithm 1.1 (Penalty Loadings for Lasso Estimation of Many BLPs). ***Step 0 (initiate):*** *Choose an integer $M \geqslant 1$, specify the penalty level λ as in (1.H.1) and the penalty loadings as in (1.H.2). Use this initial specification to compute the q Lasso estimators $\{\widehat{\beta}_k^{(0)}\}_1^q$ as in (1.G.5), and compute residuals $\widehat{\varepsilon}_{ik}^{(0)} \coloneqq \widehat{Y}_{ik} - W_i^\top \widehat{\beta}_k^{(0)}, (i, k) \in [n] \times [q]$.* ***Step m+1 (update):*** *Given residuals from Step $m < M$, $\{\widehat{\varepsilon}_{ik}^{(m)}\}_{ik}$, update the penalty loadings according to the refined option in (1.H.3), compute the Lasso estimators $\{\widehat{\beta}_k^{(m+1)}\}_{k=1}^q$ based on these refined penalty loadings, and compute residuals $\widehat{\varepsilon}_{ik}^{(m+1)} \coloneqq \widehat{Y}_{ik} - W_i^\top \widehat{\beta}_k^{(m+1)}, (i, k) \in [n] \times [q]$. Increment m and repeat this step until $m = M$ or tolerance is met.*

[38] Another difference is that I show that the nonasymptotic bound in Theorem 1.7 holds not just with probability approaching one but with probability approaching one *polynomially fast*.

Algorithm 1.1 is essentially Algorithm A.1 in Belloni, Chen, Chernozhukov, and Hansen (2012) with the inital step modified to allow for the estimands to be best linear predictors but not necessarily conditional expectations. (See also Section 1.G.3 for the necessity of this modification.)

The following lemmas establishes the conservative and true polynomial validity of the initial and refined penalty loadings, respectively.

Lemma 1.7 (**Conservative Polynomial Validity of Initial Penalty Loadings**). *Under Assumptions 1.17 and 1.18 and the growth condition $\ln^4(q) \leqslant C_2 n^{1-c_2}$, the initial penalty loadings $\{\widehat{\Upsilon}_k\}_1^q$ specified in (1.H.2) are conservatively polynomially valid uniformly over $k \in \{1, \ldots, q\}$.*

Lemma 1.8 (**True Polynomial Validity of Refined Penalty Loadings**). *Let $\{\widehat{\gamma}_{kj}\}$ denote the refined penalty loadings specified in (1.H.1), where the estimated residuals $\widehat{\varepsilon}_{ik} = \widehat{Y}_{ik} - W_i^\top \widehat{\beta}_k$ are based on estimators $\{\widehat{\beta}_k\}$ for which there exists c, C, C' and n_0 depending only on $c_0, c_0', c_1, C_1, c_2, C_2$ and c_2' such that for all $n \geqslant n_0$,*

$$\max_{1 \leqslant k \leqslant q} \|\widehat{\beta}_k - \beta_{k*}\|_{2,n} \leqslant C' \sqrt{\frac{s \ln(pqn)}{n}} \quad \text{wp} \geqslant 1 - Cn^{-c}. \tag{1.H.4}$$

Then, under Assumption 1.18 and the growth requirements $s \ln(pqn) \leqslant C_2 n^{1-c_2}$ and $\ln^4(pq) \leqslant C_2 n^{1-c_2}$, the refined penalty loadings $\{\widehat{\Upsilon}_k\}_1^q$ arising from (1.H.3) and Algorithm 1.1 are truly polynomially valid uniformly over $k \in \{1, \ldots, q\}$.

1.H.2 Implementation Details for Section 1.5

Algorithm 1.2 specializes Algorithm 1.1 to estimation of h_* by setting $q = 1$ and $\widehat{Y}_{i1} = Y_i$. Feasible options for setting the penalty level and loadings for the purpose of estimation of h_* are

$$\text{Level: } \lambda_h \coloneqq 2c_0 \Phi^{-1}(1 - n^{-c_0'}/(2p)), \tag{1.H.5}$$

$$\text{Initial Loadings: } \widehat{\gamma}_{hj} \coloneqq \sqrt{\mathbb{E}_n\{[Y_i - \mathbb{E}_n(Y_i)]^2\}} \max_{1 \leqslant i \leqslant n} |W_{ij}|, \quad j \in [p], \tag{1.H.6}$$

$$\text{Refined Loadings: } \widehat{\gamma}_{kj} \coloneqq \sqrt{\mathbb{E}_n(\widehat{\varepsilon}_i^2 W_{ij}^2)}, \quad j \in [p], \tag{1.H.7}$$

where $\widehat{\varepsilon}_{ik}$ is an (updated) estimate of ε_{ik} and $c_0 > 1$ and $c_0' \in (0,1)$ are user-specified constants.

Algorithm 1.2 (**Penalty Loadings for Lasso estimation of h_***). ***Step 0 (initiate):*** *Choose an integer $M \geqslant 1$, specify the penalty level λ_h as in (1.H.5) and the penalty loadings*

$\widehat{\Upsilon}_h$ as in (1.H.6). Use this initial specification to compute the single Lasso estimator $\widehat{h}$ as in (1.5.12), and compute residuals $\widehat{\varepsilon}_i^{(0)} \coloneqq Y_i - W_i^\top \widehat{h}^{(0)}, i \in \{1, \dots, n\}$. ***Step m+1 (update):*** *Given residuals from Step $m < M$, $\{\widehat{\varepsilon}_i^{(m)}\}_{i=1}^n$, update the penalty loadings according to the refined option in (1.H.7), compute the Lasso estimator $\widehat{h}^{(m+1)}$ based on these refined penalty loadings, and compute residuals $\widehat{\varepsilon}_i^{(m+1)} \coloneqq Y_i - W_i^\top \widehat{h}^{(m+1)}, i \in \{1, \dots, n\}$. Increment m and repeat this step until $m = M$ or tolerance is met.*

Lemma 1.9 (**Conservative Polynomial Validity of Penalty Loadings for Lasso Estimation of** h_*)**.** *Suppose that Assumption 1.12 holds. Then the penalty loadings $\widehat{\Upsilon}_h$ arising from the initial step of Algorithm 1.2 are conservatively polynomially valid.*

Lemma 1.10 (**True Polynomial Validity of Penalty Loadings for Lasso Estimation of** h_*)**.** *Suppose that Assumptions 1.10–1.14 hold. Then the penalty loadings $\widehat{\Upsilon}_h$ arising from Algorithm 1.2 with $M \geqslant 2$ are truly polynomially valid.*

Similarly, Algorithm 1.3 specializes Algorithm 1.1 to estimation of the μ_{k*}'s. Feasible options for setting the penalty level and loadings for the purpose of estimation of the μ_{k*}'s are

$$\text{Level: } \lambda_\mu \coloneqq 2c_0\Phi^{-1}(1 - n^{-c_0'}/(2pq)), \tag{1.H.8}$$

$$\text{Initial Loadings: } \widehat{\gamma}_{\mu kj} \coloneqq \sqrt{\mathbb{E}_n\{[\widehat{Y}_i - \mathbb{E}_n(\widehat{Y}_i)]^2\}} \max_{1 \leqslant i \leqslant n} |W_{ij}|\,, \quad (j,k) \in [p] \times [q]\,, \tag{1.H.9}$$

$$\text{Refined Loadings: } \widehat{\gamma}_{\mu kj} \coloneqq \sqrt{\mathbb{E}_n(\widehat{\varepsilon}_i^2 W_{ij}^2)}, \quad (j,k) \in [p] \times [q]\,, \tag{1.H.10}$$

$\widehat{Y}_{ik} = X_{ik}\partial_v\rho(Z_i, \widehat{\beta}, \widehat{L}(W_i))$ with $\widehat{L}(w) = w^\top \widehat{h}$ and $\widehat{\beta}$ given by Assumption 1.10, $\widehat{\varepsilon}_{ik}$ is an (updated) estimate of ε_{ik}, and $c_0 > 1$ and $c_0' \in (0,1)$ are user-specified constants.

Algorithm 1.3 (**Penalty Loadings for Lasso Estimation of** μ_{k*}**'s**)**.** ***Step 0 (initiate):*** *Choose an integer $M \geqslant 1$, specify the penalty level λ_μ as in (1.H.8) and the penalty loadings as in (1.H.9). Use this initial specification to compute the q Lasso estimators $\{\widehat{\mu}_k^{(0)}\}_1^q$ as in (1.5.13), and compute residuals $\widehat{\varepsilon}_{ik}^{(0)} \coloneqq \widehat{Y}_{ik} - W_i^\top \widehat{\mu}_k^{(0)}, (i,k) \in [n] \times [q]$.* ***Step m+1 (update):*** *Given residuals from Step $m < M$, $\{\widehat{\varepsilon}_{ik}^{(m)}\}_{ik}$, update the penalty loadings according to the refined option in (1.H.10), compute the Lasso estimators $\{\widehat{\beta}_k^{(m+1)}\}_{k=1}^q$ based on these refined penalty loadings, and compute residuals $\widehat{\varepsilon}_{ik}^{(m+1)} \coloneqq \widehat{Y}_{ik} - W_i^\top \widehat{\mu}_k^{(m+1)}, (i,k) \in [n] \times [q]$. Increment m and repeat this step until $m = M$ or tolerance is met.*

Lemma 1.11 (**Conservative Polynomial Validity of Penalty Loadings for Lasso Estimation of** μ_{k*}**'s**)**.** *Suppose that Assumptions 1.10–1.14 hold. Then the penalty loadings $\{\widehat{\Upsilon}_{\mu k}\}_{k=1}^q$ arising from Algorithm 1.3 with $M \geqslant 2$ are conservatively polynomially valid uniformly over $k \in \{1, \dots, q\}$.*

Lemma 1.12 (True Polynomial Validity of Penalty Loadings for Lasso Estimation of μ_{k*}'s). *Suppose that Assumptions 1.10–1.14 hold. Then the penalty loadings $\{\widehat{\Upsilon}_{\mu k}\}_{k=1}^{q}$ arising from Algorithm 1.3 with $M \geqslant 2$ are truly polynomially valid uniformly over $k \in \{1, \ldots, q\}$.*

1.I Proofs for Section 1.4

1.I.1 Proofs for Section 1.4.4

Lemma 1.13. *If Assumption 1.3 holds, then for any $z \in \mathcal{Z}$ and any $h : \mathcal{W} \to \mathbf{R}$*

$$|\rho_* (z, h (w)) - \rho_* (z, h_* (w)) - \partial_v \rho_* (z, h_* (w)) [h (w) - h_* (w)]| \leqslant L_1 (z) \|h - h_*\|_{\mathcal{W}}^{1+\gamma},$$

where $\rho_* (z, v) \coloneqq \rho (z, \beta_*, v)$.

Proof. Let $z \in \mathcal{Z}, h : \mathcal{W} \to \mathbf{R}$ be arbitrary. Then $h (w) \in \mathbf{R}$, so by Assumption 1.3 and a MVE of $v \mapsto \rho(z, \beta_*, v)$ at $h(w)$ around $h_*(w)$ yields

$$\begin{aligned}
&|\rho (z, \beta_*, h (w)) - \rho (z, \beta_*, h_* (w)) - \partial_v \rho (z, \beta_*, h_* (w)) [h (w) - h_* (w)]| \\
&= \left| \left[\partial_v \rho(z, \beta_*, \widetilde{h} (w)) - \partial_v \rho (z, \beta_*, h_* (w))\right] [h (w) - h_* (w)] \right| \\
&\leqslant L_1 (z) |\widetilde{h} (w) - h_* (w)|^{\gamma} |h (w) - h_* (w)| \leqslant L_1 (z) |h (w) - h_* (w)|^{1+\gamma} \\
&\leqslant L_1 (z) \|h - h_*\|_{\mathcal{W}}^{1+\gamma},
\end{aligned}$$

where $\widetilde{h}(w)$ lies on the line segment connecting $h(w)$ and $h_*(w)$, thus satisfying $|\widetilde{h}(w) - h_*(w)| \leqslant |h(w) - h_*(w)|$. □

Abbreviate the processes appearing in Lemma 1.1 by

$$\widehat{B} \coloneqq t \mapsto \sqrt{n} \mathbb{E}_n [\rho(Z_i, \widehat{\beta}, \widehat{h} (W)) \omega (t, X_i)], \quad t \in \mathcal{T}, \tag{1.I.1}$$

$$B_n^* \coloneqq t \mapsto \sqrt{n} \mathbb{E}_n [f_* (t, Z_i)], \quad t \in \mathcal{T}. \tag{1.I.2}$$

The following result is the crucial step in proving Lemma 1.1.

Lemma 1.14. *If Assumptions 1.1–1.7 hold, then*

$$\begin{aligned}
&\|\widehat{B} - B_n^*\|_{\mathcal{T}} = \|\sqrt{n} \mathbb{E}_n \big[\rho(Z_i, \widehat{\beta}, \widehat{h} (W_i)) \omega (\cdot, X_i) \big] - \sqrt{n} \mathbb{E}_n [f_* (\cdot, Z_i)]\|_{\mathcal{T}} \\
&\lesssim_{\mathrm{P}} \mathrm{E} [R (Z)] \sqrt{n} \|\widehat{h}_n - h_*\|_{\mathcal{W}}^{1+\gamma} + \Big(\sum_{j=1}^{k_n} \|p_{jk_n}\|_{\mathcal{W}}^2 \Big)^{1/2} \Big(\sqrt{k_n / n} + k_n^{-\alpha} \Big)
\end{aligned}$$

$$+ \sqrt{n} r_{h,k_n} \sup_{t \in \mathcal{T}} r_{\delta,k_n}(t) + \sqrt{\zeta_{k_n}^2 k_n \ln(k_n)/n} + R_{\delta,k_n} \sqrt{\ln(k_n/R_{\delta,k_n})} + \zeta_{k_n} r_{h,k_n} + o_{\mathrm{P}}(1).$$

Proof of Lemma 1.14. The proof proceeds in a number of steps.

Main

Let $t \in \mathcal{T}$ be arbitrary. Assumption 1.1 and M implies that $\|\widehat{\beta}_n - \beta_*\| \lesssim_{\mathrm{P}} n^{-1/2} \to 0$. Let $\mathcal{N}_*$ be the open neighborhood provided by Assumption 1.3 such that $\widehat{\beta}_n \in \mathcal{N}_*$ wp $\to 1$. To simplify notation and ensure that objects are globally well defined, in what follows I will—without loss of generality—assume that $\widehat{\beta}_n \in \mathcal{N}_*$ with *probability one for all n.* Then by Assumption 1.3, for any $z \in \mathcal{Z}, v \in \mathbf{R}$ we may conduct a mean value expansion of $\beta \mapsto \rho(z, \beta, v)$ at $\widehat{\beta}_n$ around β_* to get

$$\widehat{B}_n(t) = \sqrt{n}\mathbb{E}_n[\omega(t, X_i)\rho(Z_i, \beta_*, \widehat{h}_n(W_i))] + \mathrm{I}_n(t)^\top \sqrt{n}(\widehat{\beta}_n - \beta_*),$$

where

$$\mathrm{I}_n(t) \coloneqq \mathbb{E}_n\left[\omega(t, X_i)\partial_\beta \rho(Z_i, \overline{\beta}_n, \widehat{h}_n(W_i))\right], \quad t \in \mathcal{T},$$

and $\overline{\beta}_n$ lies on the line segment connecting $\widehat{\beta}_n$ and β_*, thus satisfying $\|\overline{\beta}_n - \beta_*\| \leqslant \|\widehat{\beta}_n - \beta_*\| \to_{\mathrm{P}} 0$. Recall that $b_*(t) = \mathrm{E}[\omega(t, X)\partial_\beta \rho(Z, \beta_*, h_*(W))]$, which is well defined on $\mathcal{T}$ since β_* belongs to the open set $\mathcal{N}_*$ (Assumption 1.3). Step 1.I.1 below shows that $\sup_{t \in \mathcal{T}} \|\mathrm{I}_n(t) - b_*(t)\| \to_{\mathrm{P}} 0$, and that b_* is bounded on $\mathcal{T}$, so Assumption 1.1 and the previous display combine to yield

$$\widehat{B}_n(t) = \sqrt{n}\mathbb{E}_n[\omega(t, X_i)\rho(Z_i, \beta_*, \widehat{h}_n(W_i))] + b_*(t)^\top \sqrt{n}\mathbb{E}_n[s_*(Z_i)] + o_{\mathrm{P}}(1), \qquad (1.\mathrm{I}.3)$$

uniformly on $\mathcal{T}$.

The remainder of the proof is about adjusting for the use of $\widehat{h}_n$ as an estimator for h_*. Given that β_* is held fixed throughout this argument, I will suppress the β argument and write $\rho_*(z, s) \coloneqq \rho(z, \beta_*, s)$.

For the purpose of adjusting for estimation of h_*, denote the first term on the right-hand side of (1.I.3)

$$\widehat{B}_n^*(t) \coloneqq \sqrt{n}\mathbb{E}_n[\omega(t, X_i)\rho_*(Z_i, \widehat{h}_n(W_i))],$$

and conduct a MVE of $v \mapsto \rho_*(Z_i, v)$ at $\widehat{h}_n(W_i)$ around $h_*(W_i)$ to arrive at

$$\widehat{B}_n^*(t) = \sqrt{n}\mathbb{E}_n\left(\omega(t, X_i)\left\{\rho_*(Z_i, h_*(W_i)) + \partial_v\rho_*(Z_i, \overline{h}_n(W_i))[\widehat{h}_n(W_i) - h_*(W_i)]\right\}\right),$$

where $\overline{h}_n(W_i)$ lies on the line segment connecting $\widehat{h}_n(W_i)$ and $h_*(W_i)$. Such an expansion is justified by Assumption 1.3. A decomposition of the right-hand side yields

$$\begin{aligned}\widehat{B}_n^*(t) &= \sqrt{n}\mathbb{E}_n\{\omega(t, X_i)\rho_*(Z_i, h_*(W_i)) + \delta_*(t, W_i)[Y_i - h_*(W_i)]\}\\ &\quad + \sqrt{n}\mathbb{E}_n\left\{\omega(t, X)\left[\partial_v\rho_*(Z_i, \overline{h}_n(W_i)) - \partial_v\rho_*(Z_i, h_*(W_i))\right][\widehat{h}_n(W_i) - h_*(W_i)]\right\}\\ &\quad + \mathbb{G}_n[\omega(t, X_i)\partial_v\rho_*(Z_i, h_*(W_i))][\widehat{h}_n(W_i) - h_*(W_i)]\\ &\quad + \sqrt{n}\Big(\mathrm{E}_Z\Big[\omega(t, X)\partial_v\rho_*(Z, h_*(W))[\widehat{h}_n(W) - h_*(W)]\Big]\\ &\qquad - \mathbb{E}_n\{\delta_*(t, W_i)[Y_i - h_*(W_i)]\}\Big)\\ &=: \sqrt{n}\mathbb{E}_n\{\omega(t, X_i)\rho_*(Z_i, h_*(W_i)) + \delta_*(t, W_i)[Y_i - h_*(W_i)]\}\\ &\quad + \mathrm{II}_n(t) + \mathrm{III}_n(t) + \mathrm{IV}_n(t),\end{aligned} \tag{1.I.4}$$

where $\mathrm{E}_Z(\cdot)$ denotes integration with respect to the distribution of Z. Recall the $k \times k$ matrix design matrix $Q_k = \mathrm{E}[p^k(W)p^k(W)^\top]$, which is invertible by Assumption 1.5. Let h_k and $\delta_k(t, \cdot)$ denote the mean-square projections of h_* and $\delta_*(t, \cdot)$, respectively, onto the span of $\{p_{jk} | j \in \{1, \ldots, k\}\}$, i.e.,

$$\begin{aligned}h_k(\cdot) &\coloneqq p^k(\cdot)^\top Q_k^{-1}\mathrm{E}[p^k(W)h_*(W)],\\ \delta_k(t, \cdot) &\coloneqq p^k(\cdot)^\top Q_k^{-1}\mathrm{E}[p^k(W)\delta_*(t, W)],\ t \in \mathcal{T}.\end{aligned}$$

Note that... $h_k = p^{k\top}\pi_{h,k}$ and $\delta_k(t, \cdot) = p^k(\cdot)^\top\pi_{\delta,k}(t)$, where $\pi_{h,k}$ and $\pi_{\delta,k}$ are defined in (1.4.15) and (1.4.16), respectively. Consequently, $\mathrm{E}\{[h_k(W) - h_*(W)]^2\} = r_{h,k}^2$, $\mathrm{E}\{[\delta_k(t, W) - \delta_*(t, W)]^2\} = r_{\delta,k}^2(t)$, and $\mathrm{E}\{\|\delta_k(\cdot, W) - \delta_*(\cdot, W)\|^2\} = R_{\delta,k}^2$, where $r_{h,k}$, $r_{\delta,k}(t)$ and $R_{\delta,k}$ are defined in (1.4.17), (1.4.18) and (1.4.19), respectively. Steps 1.I.1, 1.I.1 and 1.I.1 below show that the three remainder terms in the decomposition (1.I.4) satisfy:

$$\begin{aligned}\|\mathrm{II}_n\|_{\mathcal{T}} &\lesssim_{\mathrm{P}} \mathrm{E}[R(Z)]\sqrt{n}\|\widehat{h}_n - h_*\|_{\mathcal{W}}^{1+\gamma},\\ \|\mathrm{III}_n\|_{\mathcal{T}} &\lesssim_{\mathrm{P}} \Big(\sum_{j=1}^{k_n}\|p_{jk_n}\|_{\mathcal{W}}^2\Big)^{1/2}\Big(\sqrt{k_n/n} + k_n^{-\alpha}\Big),\\ \text{and } \|\mathrm{IV}_n\|_{\mathcal{T}} &\lesssim_{\mathrm{P}} \sqrt{n}r_{h,k_n}\sup_{t\in\mathcal{T}} r_{\delta,k_n}(t) + \sqrt{\zeta_{k_n}^2 k_n \ln(k_n)/n}\\ &\quad + R_{\delta,k_n}\sqrt{\ln(k_n/R_{\delta,k_n})} + \zeta_{k_n}r_{h,k_n}.\end{aligned}$$

Plug (1.I.4) into (1.I.3), apply T and use the definition of B_n^* in (1.I.2) to get

$$\begin{aligned}\|\widehat{B}_n - B_n^*\|_T &\leqslant \|\mathrm{II}_n\|_{\mathcal{T}} + \|\mathrm{III}_n\|_{\mathcal{T}} + \|\mathrm{IV}_n\|_{\mathcal{T}} + o_{\mathrm{P}}(1) \\ &\lesssim_{\mathrm{P}} \mathrm{E}\left[R(Z)\right]\sqrt{n}\|\widehat{h}_n - h_*\|_{\mathcal{W}}^{1+\gamma} + \Big(\sum_{j=1}^{k_n}\|p_{jk_n}\|_{\mathcal{W}}^2\Big)^{1/2}\Big(\sqrt{k_n/n} + k_n^{-\alpha}\Big) \\ &\quad + \sqrt{n}r_{h,k_n}\sup_{t\in T} r_{\delta,k_n}(t) + \sqrt{\zeta_{k_n}^2 k_n \ln(k_n)/n} \\ &\quad + R_{\delta,k_n}\sqrt{\ln(k_n/R_{\delta,k_n})} + \zeta_{k_n} r_{h,k_n} + o_{\mathrm{P}}(1),\end{aligned}$$

as claimed.

I_{n} **and** b_*

In this step I show that

$$\text{(a)}\ \sup_{t\in\mathcal{T}}\|\mathrm{I}_n(t) - b_*(t)\| \xrightarrow{\mathrm{P}} 0 \quad \text{and} \quad \text{(b)}\ \sup_{t\in\mathcal{T}}\|b_*(t)\| < \infty.$$

Decompose $\mathrm{I}_n(t)$ as

$$\begin{aligned}\mathrm{I}_n(t) &= \mathbb{E}_n\left[\omega(t,X_i)\,\partial_\beta\rho(Z_i,\overline{\beta}_n,h_*(W))\right] \\ &\quad + \mathbb{E}_n\Big\{\omega(t,X_i)\Big[\partial_\beta\rho(Z_i,\overline{\beta}_n,\widehat{h}_n(W_i)) - \partial_\beta\rho(Z_i,\overline{\beta}_n,h_*(W_i))\Big]\Big\} =: \mathrm{I}_{a,n}(t) + \mathrm{I}_{b,n}(t).\end{aligned}$$

Since $\|\overline{\beta}_n - \beta_*\| \leqslant \|\widehat{\beta}_n - \beta_*\|$ and $\widehat{\beta}_n \in \mathcal{N}_*$, we must have $\overline{\beta}_n \in \mathcal{N}_*$ wp → 1, so using T and Assumptions 1.2 and 1.3 , we get

$$\begin{aligned}&\sup_{t\in\mathcal{T}}\Big\|\mathbb{E}_n\Big\{\omega(t,X_i)\Big[\partial_\beta\rho(Z_i,\overline{\beta}_n,\widehat{h}_n(W_i)) - \partial_\beta\rho(Z_i,\overline{\beta}_n,h_*(W_i))\Big]\Big\}\Big\| \\ &\lesssim \mathbb{E}_n\left[L_1(Z_i)\right]\|\widehat{h}_n - h_*\|_{\mathcal{W}}^c.\end{aligned}$$

Now, $\mathbb{E}_n[L_1(Z_i)] \lesssim_{\mathrm{P}} 1$ by M, so by $\|\widehat{h}_n - h_*\|_{\mathcal{W}} \to_{\mathrm{P}} 0$ (Lemma 1.22 and Assumptions 1.4–1.7) the right-hand side $\to_{\mathrm{P}} 0$, and—as a consequence—$\|\mathrm{I}_{b,n}\|_T \to_{\mathrm{P}} 0$.

Given that β_* is interior to $\mathcal{N}_*$ (Assumption 1.1) there is an $r > 0$ such that the open ball $B_r(\beta_*)$ in $\mathbf{R}^{d_\beta}$ centered at β_* with radius r is contained in $\mathcal{N}_*$. Let $\overline{B} \coloneqq \overline{B}_{r/2}(\beta_*)$ denote the closed ball in $\mathbf{R}^{d_\beta}$ with the same center but half the radius. Given that $\overline{B}$ is a closed and bounded subset of a finite-dimensional Euclidean space, by the Heine–Borel theorem it is compact. Assumptions 1.2 and 1.3 imply that $(t,\beta) \mapsto \omega(t,x)\,\partial_\beta\rho(z,\beta,h_*(w))$ is continuous on $\mathcal{T}\times\mathcal{N}_*$ for each $z \in \mathcal{Z}$, hence on the subset $\mathcal{T}\times\overline{B}$, and this function is is dominated by an integrable function depending on z only. Moreover, via Tychonoff's theorem (cf. Aliprantis and Border, 2006, Theorem 2.61), $\mathcal{T}$ and $\overline{B}$ compact imply that is $\mathcal{T} \times \overline{B}$ compact. Combining these observations with the fact that the data are i.i.d., Newey and McFadden (1994, Lemma 2.4) Lemma 2.4 tells us that

$$\text{(i)}\ \ (t,\beta) \mapsto \mathrm{E}\left[\omega\left(t,X\right)\partial_\beta \rho\left(Z,\beta,h_*\left(W\right)\right)\right] \text{ is continuous on } \mathcal{T} \times \overline{B},$$

$$\text{(ii)}\ \sup_{(t,\beta)\in\mathcal{T}\times\overline{B}} \|(\mathbb{E}_n - \mathrm{E})\left[\omega\left(t,X_i\right)\partial_\beta\rho(Z_i,\beta,h_*\left(W_i\right))\right]\| \xrightarrow{\mathrm{P}} 0.$$

Given (i) and $\mathcal{T} \times \overline{B}$ compact, we must have (cf. Rudin, 1976, Theorem 4.19) that

$$\text{(iii)}\ (t,\beta) \mapsto \mathrm{E}\left[\omega(t,X)\partial_\beta\rho\left(Z,\beta,h_*\left(W\right)\right)\right] \text{ is uniformly continuous on } \mathcal{T} \times \overline{B}.$$

Let $\widetilde{\beta}_n$ be an arbitrary consistent estimator of β_*. Then $\widetilde{\beta}_n \in \overline{B}$ wp → 1, and, on this event,

$$\begin{aligned}
&\sup_{t\in\mathcal{T}} \left\|\mathbb{E}_n\left[\omega\left(t,X_i\right)\partial_\beta\rho(Z_i,\widetilde{\beta}_n,h_*\left(W_i\right))\right] - b_*\left(t\right)\right\| \\
&\leqslant \sup_{t\in\mathcal{T}} \left\|(\mathbb{E}_n - \mathrm{E}_Z)\left[\omega\left(t,X_i\right)\partial_\beta\rho(Z_i,\widetilde{\beta}_n,h_*\left(W_i\right))\right]\right\| \\
&\quad + \sup_{t\in\mathcal{T}} \left\|\mathrm{E}_Z\left[\omega\left(t,X\right)\partial_\beta\rho(Z,\widetilde{\beta}_n,h_*\left(W\right))\right] - b_*\left(t\right)\right\| \\
&\leqslant \sup_{(t,\beta)\in\mathcal{T}\times\overline{B}} \|(\mathbb{E}_n - \mathrm{E})\left[\omega\left(t,X_i\right)\partial_\beta\rho(Z_i,\beta,h_*\left(W_i\right))\right]\| \\
&\quad + \sup_{t\in\mathcal{T}} \left\|\mathrm{E}_Z\left[\omega\left(t,X\right)\partial_\beta\rho(Z,\widetilde{\beta}_n,h_*\left(W\right))\right] - b_*\left(t\right)\right\| \xrightarrow{\mathrm{P}} 0,
\end{aligned}$$

where the first inequality is due to T, the second uses $\{\widetilde{\beta}_n \in \overline{\mathcal{N}}\}$, and we have used (ii) uniform convergence and (iii) uniform continuity. Invoking the conclusion of the previous display for the mean value $\widetilde{\beta}_n \coloneqq \overline{\beta}_n$ we see that $\sup_{t\in\mathcal{T}}\|\mathrm{I}_{a,n}\left(t\right) - b_*\left(t\right)\| \to_{\mathrm{P}} 0$, which combined with $\sup_{t\in\mathcal{T}}\|\mathrm{I}_{b,n}\left(t\right)\| \to_{\mathrm{P}} 0$ and T establishes Part (a).

Continuity and $\mathcal{T}\times\overline{B}$ compact also imply $(t,\beta) \mapsto \mathrm{E}[\omega\left(t,X\right)\partial_\beta\rho(Z,\beta,h_*\left(W\right))]$ is *bounded* on $\mathcal{T} \times \overline{B}$ (cf. Rudin, 1976, Theorem 4.15). Part (b) then follows from $\beta_* \in \overline{B}$.

$\|\mathbf{II}_n\|_{\mathcal{T}}$

In this step I show that

$$\|\mathrm{II}_n\|_{\mathcal{T}} \lesssim_{\mathrm{P}} \mathrm{E}\left[R\left(Z\right)\right]\sqrt{n}\|\widehat{h}_n - h_*\|_{\mathcal{W}}^{1+\gamma}.$$

Using T, Assumptions 1.2 and 1.3 imply that

$$
\begin{aligned}
\|\mathrm{II}_n\|_T &\leqslant \|\omega\|_{\mathcal{T}\times\mathcal{X}}\sqrt{n}\mathbb{E}_n\left[\left|\partial_v\rho_*(Z_i,\overline{h}_n\left(W_i\right)) - \partial_v\rho_*(Z_i,h_*\left(W_i\right))\right|\left|\widehat{h}_n\left(W_i\right) - h_*\left(W_i\right)\right|\right] \\
&\lesssim \sqrt{n}\mathbb{E}_n[R\left(Z_i\right)|\overline{h}_n\left(W_i\right) - h_*\left(W_i\right)|^{\gamma}|\widehat{h}_n\left(W_i\right) - h_*\left(W_i\right)|] \\
&\leqslant \sqrt{n}\mathbb{E}_n[R\left(Z_i\right)|\widehat{h}_n\left(W_i\right) - h_*\left(W_i\right)|^{1+\gamma}] \\
&\leqslant \mathbb{E}_n\left[R\left(Z_i\right)\right]\sqrt{n}\|\widehat{h}_n - h_*\|_{\mathcal{W}}^{1+\gamma} \lesssim_{\mathrm{P}} \mathrm{E}\left[R\left(Z\right)\right]\sqrt{n}\|\widehat{h}_n - h_*\|_{\mathcal{W}}^{1+\gamma},
\end{aligned}
$$

where $\widetilde{h}_n(W_i)$ is on the line segment connecting $\widehat{h}_n(W_i)$ and $h_*\left(W_i\right)$, thus satisfying $|\overline{h}_n\left(W_i\right) - h_*\left(W_i\right)| \leqslant |\widehat{h}_n\left(W_i\right) - h_*\left(W_i\right)|$, and $\mathbb{E}_n\left[R\left(Z_i\right)\right] \lesssim_{\mathrm{P}} \mathrm{E}\left[R\left(Z\right)\right]$ follows from M.

$\|\mathbf{III}_n\|_{\mathcal{T}}$

In this step I show that

$$
\|\mathrm{III}_n\|_T \lesssim_P \Big(\sum_{j=1}^{k_n}\|p_{jk_n}\|_{\mathcal{W}}^2\Big)^{1/2}\Big(\sqrt{k_n/n} + k_n^{-\alpha}\Big).
$$

For square-integrable maps $h : \mathcal{W} \to \mathbf{R}$, define the map D by

$$
D\left(t,z,h\right) \coloneqq \omega\left(t,x\right)\partial_v\rho_*\left(z,h_*\left(w\right)\right)h\left(w\right) \tag{1.I.5}
$$

such that $h \mapsto D\left(t,z,h\right)$ is a linear functional for given $\left(t,z\right) \in \mathcal{T}\times\mathcal{Z}$. Let Δ denote the centered version of D, i.e.,

$$
\Delta\left(t,z,h\right) \coloneqq \omega\left(t,x\right)\partial_v\rho_*\left(z,h_*\left(w\right)\right)h\left(w\right) - \mathrm{E}_Z\left[\omega\left(t,X\right)\partial_v\rho_*\left(z,h_*\left(W\right)\right)h\left(W\right)\right], \tag{1.I.6}
$$

which is also linear in h. Letting $\widetilde{h}_k = p^{k\top}\widetilde{\pi}_k$ be as in Assumption 1.6, by linearity we may write

$$
\begin{aligned}
\mathrm{III}_n\left(t\right) &= \sqrt{n}\mathbb{E}_n\left[\Delta(t,Z_i,\widehat{h} - h_*)\right] \\
&= \sqrt{n}\mathbb{E}_n\left[\Delta(t,Z_i,\widehat{h} - \widetilde{h}_{k_n})\right] + \sqrt{n}\mathbb{E}_n\left[\Delta(t,Z_i,\widetilde{h}_{k_n} - h_*)\right] \\
&\eqqcolon \mathrm{III}_{a,n}\left(t\right) + \mathrm{III}_{b,n}\left(t\right).
\end{aligned}
$$

Given that $\zeta_k = \sup_{w\in\mathcal{W}}\|p^k\left(w\right)\| = \sup_{w\in\mathcal{W}}[\sum_{j=1}^{k}p_{jk}\left(w\right)^2]^{1/2}$, $\zeta_{k_n} \to \infty$ (see Remark 1.3) implies $\sum_{j=1}^{k_n}\|p_{jk_n}\|_{\mathcal{W}}^2 \to \infty$. In particular, $\sum_{j=1}^{k_n}\|p_{jk_n}\|_{\mathcal{W}}^2$ is bounded away from zero as

$n \to \infty$. By T it therefore suffices to show that

$$\begin{aligned} \|\mathrm{III}_{a,n}\|_{\mathcal{T}} &\lesssim_{\mathrm{P}} \Big(\sum_{j=1}^{k_n} \|p_{jk_n}\|_{\mathcal{W}}^2\Big)^{1/2} \Big(\sqrt{k_n/n} + k_n^{-\alpha}\Big), \\ \text{and} \quad \|\mathrm{III}_{b,n}\|_{\mathcal{T}} &\lesssim_{\mathrm{P}} k_n^{-\alpha}. \end{aligned}$$

$\|\boldsymbol{III}_{a,n}\|_{\mathcal{T}}$

In this step I show that

$$\|\mathrm{III}_{a,n}\|_{\mathcal{T}} \lesssim_{\mathrm{P}} \Big(\sum_{j=1}^{k_n} \|p_{jk_n}\|_{\mathcal{W}}^2\Big)^{1/2} \Big(\sqrt{k_n/n} + k_n^{-\alpha}\Big).$$

Let $\Delta_i^k(t) \coloneqq (\Delta(t, Z_i, p_{1k}), \ldots, \Delta(t, Z_i, p_{kk}))^\top$. Then CS implies

$$\begin{aligned} \|\mathrm{III}_{a,n}\|_{\mathcal{T}} &= \sup_{t \in \mathcal{T}} \left| \sqrt{n} \mathbb{E}_n \left[\Delta(t, Z_i, p^{k_n \top}(\widehat{\pi} - \widetilde{\pi}_{k_n})) \right] \right| \\ &= \sup_{t \in \mathcal{T}} \left| \sqrt{n} \left\{ \mathbb{E}_n[\Delta_i^{k_n}(t)] \right\}^\top (\widehat{\pi} - \widetilde{\pi}_{k_n}) \right| \\ &\leqslant \|\widehat{\pi} - \widetilde{\pi}_{k_n}\| \sup_{t \in \mathcal{T}} \left\| \sqrt{n} \mathbb{E}_n[\Delta_i^{k_n}(t)] \right\|. \end{aligned}$$

Lemma 1.22 tells us that $\|\widehat{\pi} - \widetilde{\pi}_{k_n}\| \lesssim_{\mathrm{P}} \sqrt{k_n/n} + k_n^{-\alpha}$, so it remains to show that

$$\sup_{t \in \mathcal{T}} \left\| \sqrt{n} \mathbb{E}_n[\Delta_i^{k_n}(t)] \right\| \lesssim_{\mathrm{P}} \Big(\sum_{j=1}^{k_n} \|p_{jk_n}\|_{\mathcal{W}}^2\Big)^{1/2}.$$

By M it suffices to show the finite-sample moment bound, for any $k \in \mathbf{N}$,

$$\mathrm{E}\left[\sup_{t \in \mathcal{T}} \left\| \sqrt{n} \mathbb{E}_n[\Delta_i^k(t)] \right\|^2\right] \lesssim \sum_{j=1}^{k} \|p_{jk}\|_{\mathcal{W}}^2.$$

Given that

$$\mathrm{E}\left[\sup_{t \in \mathcal{T}} \left\| \sqrt{n} \mathbb{E}_n[\Delta_i^k(t)] \right\|^2\right] \leqslant \sum_{j=1}^{k} \mathrm{E}\left\{\sup_{t \in \mathcal{T}} \left[\sqrt{n} \mathbb{E}_n[\Delta(t, Z_i, p_{jk})\right]^2\right\},$$

it suffices to show that

$$\mathrm{E}\left\{\sup_{t \in \mathcal{T}} \left[\sqrt{n} \mathbb{E}_n[\Delta(t, Z_i, p_{jk})\right]^2\right\} \lesssim \|p_{jk}\|_{\mathcal{W}}^2, \quad j \in \{1, \ldots, k\}.$$

To this end, fix $j \in \{1, \ldots, k\}$ and consider the function class $\mathcal{F}_{jk} \coloneqq \mathcal{F}_{jk}(\mathcal{T}) := \{f : z \mapsto \Delta(t, z, p_{jk}) \,|\, t \in \mathcal{T}\}$. For $f_1 \coloneqq f(\cdot; t_1), f_2 \coloneqq f(\cdot; t_2) \in \mathcal{F}_{jk}$ arbitrary, by T, J and Assumptions 1.2 and 1.3,

$$\begin{aligned}
|f_1(z) - f_2(z)| &= |\left[\omega(t_1, x) - \omega(t_2, x)\right] \partial_v \rho_*(z, h_*(w))\, p_{jk}(w) \\
&\quad - \mathrm{E}\left\{\left[\omega(t_1, X) - \omega(t_2, X)\right] \partial_v \rho_*(Z, h_*(W))\, p_{jk}(W)\right\}| \\
&\leqslant |\omega(t_1, x) - \omega(t_2, x)| \Big|\partial_v \rho_*(z, h_*(w))\Big| |p_{jk}(w)| \\
&\quad + \mathrm{E}\left[|\omega(t_1, X) - \omega(t_2, X)| \left|\partial_v \rho_*(Z, h_*(W))\right| |p_{jk}(W)|\right] \\
&\lesssim \Big(\left|\partial_v \rho_*(z, h_*(w))\right| |p_{jk}(w)| + \mathrm{E}\Big[\left|\partial_v \rho_*(Z, h_*(W))\right| |p_{jk}(W)|\Big]\Big) \|t_1 - t_2\| \\
&\leqslant \Big\{\Big|\partial_v \rho_*(z, h_*(w))\Big| + \mathrm{E}\left[\left|\partial_v \rho_*(Z, h_*(W))\right|\right]\Big\} \|p_{jk}\|_{\mathcal{W}} \|t_1 - t_2\| \\
&= L_1(z) \|p_{jk}\|_{\mathcal{W}} \|t_1 - t_2\|,
\end{aligned}$$

such that we may write

$$|f_1(z) - f_2(z)| \leqslant F_{1,jk}(z) \|t_1 - t_2\|,$$

for $F_{1,jk}(z) \coloneqq C_1 L_1(z) \|p_{jk}\|_{\mathcal{W}}$ and some constant $C_1 \in (0, \infty)$. Similarly, for $f \coloneqq f(\cdot; t) \in \mathcal{F}_{jk}$ arbitrary, by T, J and Assumptions 1.2 and 1.3,

$$\begin{aligned}
|f(z)| &= |\omega(t, x) \partial_v \rho_*(z, h_*(w))\, p_{jk}(w) - \mathrm{E}_Z\left[\omega(t, X) \partial_v \rho_*(Z, h_*(W))\, p_{jk}(W)\right]| \\
&\lesssim L_1(z) \|p_{jk}\|_{\mathcal{W}},
\end{aligned}$$

such that we may write

$$|f(z)| \leqslant F_{2,jk}(z)$$

for $F_{2,jk}(z) \coloneqq C_2 L_1(z) \|p_{jk}\|_{\mathcal{W}}$ and some constant $C_2 \in (0, \infty)$. Let $C_3 \coloneqq C_1 \vee C_2$ and

$$F_{jk}(z) \coloneqq C_3 L_1(z) \|p_{jk}\|_{\mathcal{W}}.$$

Then $\|F_{jk}\|_{P,2} \lesssim \|p_{jk}\|_{\mathcal{W}}$, so F_{jk} is an square-integrable envelope for $\mathcal{F}_{jk}$ satisfying

$$|f_1(z) - f_2(z)| \leqslant F_{jk}(z) \|t_1 - t_2\|.$$

Given that $\mathcal{T}$ is compact (Assumption 1.2), we must have $\operatorname{diam}(\mathcal{T}) < \infty$. Pollard (1990, Lemma 4.1) and the fact that covering numbers are bounded by packing numbers (cf. van der

Vaart and Wellner, 1996, p. 98) therefore combine to yield $N\left(\varepsilon, \mathcal{T}, \|\cdot\|\right) \leqslant \left(3\mathrm{diam}\left(\mathcal{T}\right)/\varepsilon\right)^{d_t}$ for $\varepsilon \in (0, \mathrm{diam}\left(\mathcal{T}\right)]$. Hence, by van der Vaart and Wellner (1996, Theorem 2.7.11) and the previous display,

$$N_{[\,]}(\varepsilon\|F_{jk}\|_{P,2}, \mathcal{F}_{jk}, L^2\left(P\right)) \leqslant N\left(\varepsilon/2, \mathcal{T}, \|\cdot\|\right) \leqslant \left(6\mathrm{diam}\left(\mathcal{T}\right)/\varepsilon\right)^{d_t} \leqslant \left(C/\varepsilon\right)^{d_t}$$

for $\varepsilon \in (0, \mathrm{diam}\left(\mathcal{T}\right)]$. The bracketing integral of $\mathcal{F}_{jk}$ therefore satisfies the bound

$$J_{[\,]}\left(\delta, \mathcal{F}_{jk}, L^2\left(P\right)\right) \leqslant \int_0^\delta \sqrt{1 + C\ln\left(1/\varepsilon\right)}\mathrm{d}\varepsilon.$$

Note that the right-hand side depends on neither j nor k. In particular, $J_{[\,]}\left(1, \mathcal{F}_{jk}, L^2\left(P\right)\right)$ is bounded uniformly in $j \in \{1, \ldots, k\}, k \in \mathbf{N}$. By construction, $\mathrm{E}[f(Z)] = \mathrm{E}[\Delta(t, Z, p_{jk})] = 0$ for any $f \in \mathcal{F}_{jk}$, so we may view the stochastic process $\{\sqrt{n}\mathbb{E}_n[\Delta\left(t, Z_i, p_{jk}\right)] \,|t \in \mathcal{T}\}$ as an *empirical* process $\{\mathbb{G}_n(f) \,|f \in \mathcal{F}_{jk}\}$. van der Vaart and Wellner (1996, Theorem 2.14.2) therefore implies the finite-sample bound

$$\mathrm{E}(\|\mathbb{G}_n\|_{\mathcal{F}_{jk}}) \lesssim J_{[\,]}\left(1, \mathcal{F}_{jk}, L^2\left(P\right)\right)\|F_{jk}\|_{P,2} \lesssim \|F_{jk}\|_{P,2} \lesssim \|p_{jk}\|_{\mathcal{W}}.$$

van der Vaart and Wellner (1996, Theorem 2.14.5) shows that

$$[\mathrm{E}(\|\mathbb{G}_n\|^2_{\mathcal{F}_{jk}})]^{1/2} \lesssim \mathrm{E}(\|\mathbb{G}_n\|_{\mathcal{F}_{jk}}) + \|F_{jk}\|_{P,2} \lesssim \|p_{jk}\|_{\mathcal{W}},$$

which is the desired bound.

$\|\mathbf{III}_{b,n}\|_T$

In this step I show that

$$\|\mathrm{III}_{b,n}\|_{\mathcal{T}} \lesssim_{\mathrm{P}} k_n^{-\alpha_d}.$$

For this purpose, fix $k \in \mathbf{N}$ and consider the function class $\mathcal{F}_k \coloneqq \mathcal{F}_k\left(\mathcal{T}\right) \coloneqq \{f \colon z \mapsto \Delta(t, z, \widetilde{h}_k - h_*)|t \in \mathcal{T}\}$. For $f \coloneqq f(\cdot, t), f_1 \coloneqq f(\cdot, t_1), f_2 \coloneqq f(\cdot, t_2) \in \mathcal{F}_k$ arbitrary, arguments analogous to the ones from Step 1.I.1 establish that

$$\begin{aligned} |f_1\left(z\right) - f_2\left(z\right)| &\leqslant C_1 L_1\left(z\right)\|\widetilde{h}_k - h_*\|_{\mathcal{W}}\|t_1 - t_2\|, \\ |f\left(z\right)| &\leqslant C_2 L_1\left(z\right)\|\widetilde{h}_k - h_*\|_{\mathcal{W}}. \end{aligned}$$

Define $C_3 \coloneqq C_1 \vee C_2$ and $F_k\left(z\right) \coloneqq C_3 L_1\left(z\right)\|\widetilde{h}_k - h_*\|_{\mathcal{W}}$. Then $\|F_k\|_{P,2} = C_4\|\widetilde{h}_k - h_*\|_{\mathcal{W}} \lesssim k^{-\alpha}$ by Assumption 1.6. Hence F_k is an square-integrable envelope for $\mathcal{F}_k$, and arguments analogous to the ones from Step 1.I.1 show that the resulting bracketing integral $J_{[\,]}\left(\delta, \mathcal{F}_k, L^2\left(P\right)\right)$ is bounded by a constant independent of k. van der Vaart and Wellner (1996, Theorem 2.14.2) therefore implies

$$\mathrm{E}(\|\mathbb{G}_n\|_{\mathcal{F}_k}) \lesssim J_{[\,]}\left(1, \mathcal{F}_k, L^2\left(P\right)\right)\|F_k\|_{P,2} \lesssim \|F_k\|_{P,2} \lesssim k^{-\alpha}.$$

and the claim follows from M.

$\|\mathbf{IV}_n\|_{\mathcal{T}}$

In this step I show that

$$\|\mathrm{IV}_n\|_{\mathcal{T}} \lesssim_P \sqrt{n} r_{h,k_n} \sup_{t\in\mathcal{T}} r_{\delta,k_n}\left(t\right) + \sqrt{\zeta_{k_n}^2 k_n \ln\left(k_n\right)/n} + R_{\delta,k_n}\sqrt{\ln\left(k_n/R_{\delta,k_n}\right)} + \zeta_{k_n} r_{h,k_n}.$$

Recall that h_k and $\delta_k\left(t,\cdot\right)$ are the mean-square projections of $h_*$and $\delta_*\left(t,\cdot\right)$, respectively, onto the linear span of $\{p_{jk} | \, j \in \{1,\ldots,k\}\}$, and $r_{h,k}^2$ and $r_{\delta,k}^2\left(t\right)$ are the mean-square errors resulting from these projections. Define $\psi_k\left(t\right) \coloneqq \mathrm{E}\left[\delta_*\left(t,W\right)p^k\left(W\right)\right]$. Assumption 1.5 implies that the population least-square coefficients $\pi_k = Q_k^{-1}\mathrm{E}[p^k\left(W\right)Y]$ are well defined for any $k \in \mathbf{N}$. Applying Lemma 1.17 with $A_n \coloneqq Q_{k_n}$ and $B_n \coloneqq \widehat{Q}_{k_n}$, we see that the inverse of $\widehat{Q}_{k_n}$ exists wp → 1. As a consequence, the sample least-squares coefficients take the form $\widehat{\pi}_n = \widehat{Q}_{k_n}^{-1}\mathbb{E}_n[p^{k_n}\left(W_i\right)Y_i]$ wp → 1. Assuming—without loss of generality—that $\widehat{Q}_{k_n}^{-1}$ exists with probability one for all n,

$$\begin{aligned}
\sqrt{n}\mathrm{E}_W\{\delta_*\left(t,W\right)[\widehat{h}\left(W\right) - h_{k_n}\left(W\right)]\} &= \sqrt{n}\mathrm{E}_W\{\delta_*\left(t,W\right)p^{k_n}\left(W\right)^\top\left(\widehat{\pi} - \pi_{k_n}\right)\} \\
&= \psi_{k_n}\left(t\right)^\top\sqrt{n}(\widehat{\pi} - \pi_{k_n}) \\
&= \psi_{k_n}\left(t\right)^\top\sqrt{n}\left(\widehat{Q}_{k_n}^{-1}\mathbb{E}_n\left[p^{k_n}(W_i)Y_i\right] - \pi_{k_n}\right) \\
&= \psi_{k_n}\left(t\right)^\top\widehat{Q}_{k_n}^{-1}\sqrt{n}\left(\mathbb{E}_n\left[p^{k_n}(W_i)Y_i\right] - \widehat{Q}_{k_n}\pi_{k_n}\right) \\
&= \psi_{k_n}\left(t\right)^\top\widehat{Q}_{k_n}^{-1}\sqrt{n}\mathbb{E}_n\left[p^{k_n}\left(W_i\right)\right][Y_i - h_{k_n}\left(W_i\right)],
\end{aligned}$$

where $\mathrm{E}_W\left(\cdot\right)$ denotes integration with respect to the distribution of W. By definition of $\delta_*\left(t,W\right)$ [see (1.4.14)] and iterated of expectations, for any measurable function h of W alone,

$$\mathrm{E}\omega\left(t,X\right)\partial_v\rho_*\left(Z,h_*\left(W\right)\right)h\left(W\right)\Big] = \mathrm{E}\left[\delta_*\left(t,W\right)h\left(W\right)\right].$$

Using the previous two displays and adding and subtracting

$$\begin{aligned}&\sqrt{n}\mathbb{E}_n\left\{\delta_{k_n}(t,W_i)\left[Y_i-h_{k_n}(W_i)\right]\right\}\\&=\sqrt{n}\mathbb{E}_n\left\{p^{k_n}(W_i)^\top Q_{k_n}^{-1}\mathrm{E}[p^{k_n}(W)\,\delta_*(t,W)]\left[Y_i-h_{k_n}(W_i)\right]\right\}\\&=\psi_{k_n}(t)^\top Q_{k_n}^{-1}\sqrt{n}\mathbb{E}_n\left\{p^{k_n}(W_i)\left[Y_i-h_{k_n}(W_i)\right]\right\},\end{aligned}$$

we may decompose $\mathrm{IV}_n(t)$ as

$$\begin{aligned}\mathrm{IV}_n(t)&=\sqrt{n}\mathrm{E}_W\{\delta_*(t,W)\,[\widehat{h}(W)-h_*(W)]\}-\sqrt{n}\mathbb{E}_n\left\{\delta_*(t,W_i)\left[Y_i-h_*(W_i)\right]\right\}\\&=\sqrt{n}\mathrm{E}_W\{\delta_*(t,W)\,[h_{k_n}(W)-h_*(W)]\}+\sqrt{n}\mathrm{E}_W\{\delta_*(t,W)\,[\widehat{h}(W)-h_{k_n}(W)]\}\\&\quad+\sqrt{n}\mathbb{E}_n\left\{\delta_*(t,W_i)\left[Y_i-h_*(W_i)\right]\right\}\\&=\sqrt{n}\mathrm{E}_W\{\delta_*(t,W)\,[h_{k_n}(W)-h_*(W)]\}\\&\quad+\psi_{k_n}(t)^\top(\widehat{Q}_{k_n}^{-1}-Q_{k_n}^{-1})\sqrt{n}\mathbb{E}_n\left\{p^{k_n}(W_i)\left[Y_i-h_{k_n}(W_i)\right]\right\}\\&\quad+\sqrt{n}\mathbb{E}_n\left\{\delta_{k_n}(t,W_i)\left[Y_i-h_{k_n}(W_i)\right]-\delta_*(t,W_i)\left[Y_i-h_*(W_i)\right]\right\}\\&=:\mathrm{IV}_{a,n}(t)+\mathrm{IV}_{b,n}(t)+\mathrm{IV}_{c,n}(t).\end{aligned}$$

By T it therefore suffices to show that

$$\begin{aligned}&\|\mathrm{IV}_{a,n}\|_{\mathcal{T}}\leqslant\sqrt{n}r_{h,k_n}\sup_{t\in\mathcal{T}}r_{\delta,k_n}(t),\\&\|\mathrm{IV}_{b,n}\|_{\mathcal{T}}\lesssim_{\mathrm{P}}\sqrt{\zeta_{k_n}^2k_n\ln(k_n)/n},\\\text{and}\quad&\|\mathrm{IV}_{c,n}\|_{\mathcal{T}}\lesssim_{\mathrm{P}}R_{\delta,k_n}\sqrt{\ln(k_n/R_{\delta,k_n})}+\zeta_{k_n}r_{h,k_n}.\end{aligned}$$

$\|\mathbf{IV}_{a,n}\|_{\mathcal{T}}$

In order to establish the inequality

$$\|\mathrm{IV}_{a,n}\|_{\mathcal{T}}\leqslant\sqrt{n}r_{h,k_n}\sup_{t\in\mathcal{T}}r_{\delta,k_n}(t),$$

recall that h_k is the mean-square projection of h_* onto the span of $\{p_{jk}|\,j\in\{1,\ldots,k\}\}\}$, so by orthogonality of projections we have $\mathrm{E}\{\delta_k(t,W)\,[h_k(W)-h_*(W)]\}=0$ for each $t\in\mathcal{T}$. Now J followed by CS yield

$$\begin{aligned}\|\mathrm{IV}_{a,n}\|_{\mathcal{T}}&=\sqrt{n}\sup_{t\in\mathcal{T}}\left|\mathrm{E}\left\{\delta_*(t,W)\left[h_{k_n}(W)-h_*(W)\right]\right\}\right|\\&=\sqrt{n}\sup_{t\in\mathcal{T}}\left|\mathrm{E}\left\{\left[\delta_{k_n}(t,W)-\delta_*(t,W)\right]\left[h_{k_n}(W)-h_*(W)\right]\right\}\right|\end{aligned}$$

$$\leqslant \sqrt{n}\,\|h_{k_n} - h^*\|_{P,2} \sup_{t\in\mathcal{T}} \|\delta_{k_n}(t,\cdot) - \delta_*(t,\cdot)\|_{P,2} = \sqrt{n} r_{h,k_n} \sup_{t\in\mathcal{T}} r_{\delta,k_n}(t)\,.$$

$\|\mathbf{IV}_{b,n}\|_{\mathcal{T}}$

In this step I show that

$$\|\mathrm{IV}_{b,n}\|_{\mathcal{T}} \lesssim_{\mathrm{P}} \sqrt{\zeta_{k_n}^2 k_n \ln(k_n)/n}.$$

Using the fact that mean-square projections are $L^2(P)$-contractions followed by Assumptions 1.2 and 1.3, we see that

$$\begin{aligned}\psi_k(t)^\top Q_k^{-1}\psi_k(t) &= \{Q_k^{-1}\mathrm{E}[p^k(W)\,\delta_*(t,W)]\}^\top Q_k\{Q_k^{-1}\mathrm{E}[p^k(W)\,\delta_*(t,W)]\}\\ &= \mathrm{E}[\delta_k(t,W)] \leqslant \mathrm{E}[\delta_*(t,W)^2] = \mathrm{E}[\omega(t,W)^2\,\partial_v\rho_*(Z,h_*(W))^2]\\ &\lesssim \mathrm{E}[\partial_v\rho_*(Z,h_*(W))^2] < \infty,\end{aligned}$$

with an upper bound that depends on neither t nor k. By the Min-Max Theorem, Assumption 1.5, and the previous display, it follows that

$$\begin{aligned}\|\psi_k(t)\,Q_k^{-1}\|^2 = [\psi_k(t)\,Q_k^{-1/2}]^\top Q_k^{-1}[Q_k^{-1/2}\psi_K(t)] &\lesssim \|\psi_k(t)\,Q_k^{-1/2}\|^2\\ &\leqslant \sup_{k\in\mathbf{N},t\in\mathcal{T}}|\psi_k(t)^\top Q_k^{-1}\psi_k(t)| < \infty,\end{aligned}$$

thus implying $\sup_{k\in\mathbf{N},t\in\mathcal{T}}\|\psi_k(t)\,Q_k^{-1}\| < \infty$. By Lemma 1.21 we have $\|\widehat{Q}_{k_n} - Q_{k_n}\|_{\mathrm{op}} \lesssim_{\mathrm{P}} [\zeta_{k_n}^2 \ln(k_n)/n]^{1/2} \to 0$, where $\to 0$ follows from Assumption 1.7. Moreover, Lemma 1.17 applied with $A_n \coloneqq Q_{k_n}$ and $B_n \coloneqq \widehat{Q}_{k_n}$ shows that $\|\widehat{Q}_{k_n}^{-1}\|_{\mathrm{op}} \lesssim_{\mathrm{P}} 1$. Using these observations and the previous display,

$$\begin{aligned}\sup_{t\in\mathcal{T}}\|\psi_{k_n}(t)^\top \widehat{Q}_{k_n}^{-1} - \psi_{k_n}(t)^\top Q_k^{-1}\| &= \sup_{t\in\mathcal{T}}\|\psi_{k_n}(t)^\top Q_{k_n}^{-1}(Q_{k_n} - \widehat{Q}_{k_n})\widehat{Q}_{k_n}^{-1}\|\\ &\leqslant \|(Q_{k_n} - \widehat{Q}_{k_n})\widehat{Q}_{k_n}^{-1}\|_{\mathrm{op}} \sup_{t\in\mathcal{T}}\|\psi_{k_n}(t)^\top Q_{k_n}^{-1}\|\\ &\leqslant \|\widehat{Q}_{k_n} - Q_{k_n}\|_{\mathrm{op}}\|\widehat{Q}_{k_n}^{-1}\|_{\mathrm{op}} \sup_{t\in\mathcal{T}}\|\psi_{k_n}(t)^\top Q_{k_n}^{-1}\|\\ &\lesssim_{\mathrm{P}} \sqrt{\zeta_{k_n}^2 \ln(k_n)/n} \to 0.\end{aligned}$$

From the previous display and $\sup_{k\in\mathbf{N},t\in\mathcal{T}}\|\psi_k(t)\,Q_k^{-1}\| < \infty$ it follows that

$$\sup_{t\in\mathcal{T}}\|\psi_{k_n}(t)^\top \widehat{Q}_{k_n}^{-1}\| \lesssim_{\mathrm{P}} 1.$$

Observe also that, by the Assumption 1.5, the Min-Max theorem, and the fact that $\mathrm{E}\{p^k(W)[Y-h_k(W)]\} = \mathbf{0}_{k\times 1}$ (which follows from h_k being the mean-square projection of h_*),

$$\begin{aligned}
&\mathrm{E}\big[\|Q_k^{-1}\sqrt{n}\mathbb{E}_n\left\{p^k\left(W_i\right)\left[Y_i-h_k\left(W_i\right)\right]\right\}\|^2\big] \\
&\lesssim \mathrm{E}\big[\|Q_k^{-1/2}\sqrt{n}\mathbb{E}_n\left\{p^k\left(W_i\right)\left[Y_i-h_k\left(W_i\right)\right]\right\}\|^2\big] \\
&= \mathrm{E}\big\{p^k\left(W\right)^\top Q_k^{-1}p^k\left(W\right)\left[Y-h_k\left(W\right)\right]^2\big\} \\
&= \mathrm{E}\big[U^2p^k\left(W\right)^\top Q_k^{-1}p^k\left(W\right)\big] + \mathrm{E}\big\{p^k\left(W\right)^\top Q_k^{-1}p^k\left(W\right)\left[h_k\left(W\right)-h_*\left(W\right)\right]^2\big\},
\end{aligned}$$

where I have used $U = Y - h_*(W)$. By Assumption 1.4, $\mathrm{E}(U^2|W)$ is bounded, so

$$\mathrm{E}[U^2p^k\left(W\right)^\top Q_k^{-1}p^k\left(W\right)] = \mathrm{E}[\mathrm{E}(U^2|W)p^k\left(W\right)^\top Q_k^{-1}p^k\left(W\right)] \lesssim \mathrm{E}[p^k\left(W\right)^\top Q_k^{-1}p^k\left(W\right)] = k.$$

Moreover,

$$\mathrm{E}\{p^k\left(W\right)^\top Q_k^{-1}p^k\left(W\right)\left[h_k\left(W\right)-h_*\left(W\right)\right]^2\} \lesssim \mathrm{E}\{\|p^k\left(W\right)\|^2[h_k\left(W\right)-h_*\left(W\right)]^2\} \leqslant \zeta_k^2 r_{h,k}^2.$$

Given Assumption 1.7, $\zeta_k^2 r_{h,k}^2 = (\zeta_k r_{h,k})^2 \to 0$ as $k \to \infty$, so

$$\mathrm{E}\big[\|Q_k^{-1}\sqrt{n}\mathbb{E}_n\left\{p^k\left(W_i\right)\left[Y_i-h_k\left(W_i\right)\right]\right\}\|^2\big] \lesssim k.$$

M now implies

$$\|Q_k^{-1}\sqrt{n}\mathbb{E}_n\left\{p^{k_n}\left(W_i\right)\left[Y_i-h_{k_n}\left(W_i\right)\right]\right\}\| \lesssim_{\mathrm{P}} \sqrt{k_n}.$$

Using CS we therefore arrive at

$$\begin{aligned}
\|\mathrm{IV}_{b,n}\|_{\mathcal{T}} &= \sup_{t\in\mathcal{T}}\left|\psi_{k_n}\left(t\right)^\top \widehat{Q}_{k_n}^{-1}(Q_{k_n}-\widehat{Q}_{k_n})Q_{k_n}^{-1}\sqrt{n}\mathbb{E}_n\left\{p^{k_n}\left(W_i\right)\left[Y_i-h_{k_n}\left(W_i\right)\right]\right\}\right| \\
&\leqslant \left\|Q_{k_n}^{-1}\sqrt{n}\mathbb{E}_n\left\{p^{k_n}\left(W_i\right)\left[Y_i-h_{k_n}\left(W_i\right)\right]\right\}\right\| \sup_{t\in\mathcal{T}}\|\psi_{k_n}\left(t\right)^\top \widehat{Q}_{k_n}^{-1}(Q_{k_n}-\widehat{Q}_{k_n})\| \\
&\leqslant \left\|Q_{k_n}^{-1}\sqrt{n}\mathbb{E}_n\left\{p^{k_n}\left(W_i\right)\left[Y_i-h_{k_n}\left(W_i\right)\right]\right\}\right\| \|\widehat{Q}_{k_n}-Q_{k_n}\|_{\mathrm{op}} \sup_{t\in\mathcal{T}}\|\psi_{k_n}\left(t\right)^\top \widehat{Q}_{k_n}^{-1}\| \\
&\lesssim_{\mathrm{P}} \sqrt{k_n}\sqrt{\zeta_{k_n}^2\ln\left(k_n\right)/n}.
\end{aligned}$$

$\|\mathbf{IV}_{c,n}\|_{\mathcal{T}}$

In this section I show that

$$\|\mathrm{IV}_{c,n}\|_{\mathcal{T}} \lesssim_{\mathrm{P}} R_{\delta,k_n}\sqrt{\ln\left(k_n/R_{\delta,k_n}\right)} + \zeta_{k_n}r_{h,k_n}.$$

Letting $U_i \coloneqq Y_i - h^*(W_i)$, we may decompose $\mathrm{IV}_{c,n}(t)$ as

$$\begin{aligned}\mathrm{IV}_{c,n}(t) &= \sqrt{n}\mathbb{E}_n\left\{U_i\left[\delta_{k_n}(t, W_i) - \delta_*(t, W_i)\right]\right\} - \sqrt{n}\mathbb{E}_n\left\{\delta_{k_n}(t, W_i)\left[h_{k_n}(W_i) - h_*(W_i)\right]\right\} \\ &\eqqcolon \mathrm{IV}_{d,n}(t) + \mathrm{IV}_{e,n}(t).\end{aligned}$$

By T it therefore suffices to show that

$$\|\mathrm{IV}_{d,n}\|_{\mathcal{T}} \lesssim_{\mathrm{P}} R_{\delta,k_n}\sqrt{\ln(k_n/R_{\delta,k_n})} \quad \text{and} \quad \|\mathrm{IV}_{e,n}\|_T \lesssim_{\mathrm{P}} \zeta_{k_n} r_{h,k_n}.$$

For the purpose of bounding $\|\mathrm{IV}_{d,n}\|_{\mathcal{T}}$, consider the function class $\mathcal{F}_k \coloneqq \mathcal{F}_k(\mathcal{T}) \coloneqq \{f : z \mapsto [y - h_*(w)][\delta_k(t, w) - \delta^*(t, w)] \,|\, t \in \mathcal{T}\}$. Note that $\mathrm{E}[f(Z)] = 0$ for any $f \in \mathcal{F}_k$, so we may view the stochastic process $\{\mathrm{IV}_{d,n}(t) | t \in T\}$ as an empirical process $\{\mathbb{G}_n(f) \,|\, f \in \mathcal{F}_k\}$. For any $t_1, t_2 \in \mathcal{T}$, by J we have

$$\begin{aligned}|\delta_*(t, w) - \delta_*(t, w)| &= |\mathrm{E}\left\{[\omega(t_1, X) - \omega(t_2, X)]\,\partial_v\rho_*(Z, h_*(W))\right\}| \\ &\lesssim \mathrm{E}\left[|\partial_v\rho_*(Z, h_*(W))| \,|W = w\right] \|t_1 - t_2\|.\end{aligned}$$

Consequently, using Assumption 1.3 and the fact that conditional expectations are $L^2(P)$ contractions,

$$\begin{aligned}\mathrm{E}\{[\delta_*(t_1, W) - \delta_*(t_2, W)]^2\} &\lesssim \mathrm{E}(\{\mathrm{E}\left[|\partial_v\rho_*(Z, h_*(W))| \,|W = w\right]\}^2)\, \|t_1 - t_2\|^2 \\ &\leqslant \mathrm{E}\Big[\partial_v\rho_*(Z, h_*(W))^2\Big]\, \|t_1 - t_2\|^2 \lesssim \|t_1 - t_2\|^2.\end{aligned}$$

Given that mean-square projections are also $L^2(P)$ contractions,

$$\begin{aligned}&\|Q_k^{-1/2}\mathrm{E}\left\{p^k(W)\left[\delta_*(t_1, W) - \delta_*(t_2, W)\right]\right\}\|^2 \\ &= \mathrm{E}\Big[\big(p^k(W)^\top Q_k^{-1}\mathrm{E}\left\{p^k(W)\left[\delta_*(t_1, W) - \delta_*(t_2, W)\right]\right\}\big)^2\Big] \\ &\leqslant \mathrm{E}\{[\delta_*(t_1, W) - \delta_*(t_2, W)]^2\}\end{aligned}$$

so by CS and the previous two displays,

$$\begin{aligned}|\delta_k(t_1, w) - \delta_k(t_2, w)| &= |p^k(w)^\top Q_k^{-1}\mathrm{E}\left\{p^k(W)\left[\delta_*(t_1, W) - \delta_*(t_2, W)\right]\right\}| \\ &\leqslant \|p^k(w)^\top Q_k^{-1/2}\|\|Q_k^{-1/2}\mathrm{E}\left\{p^k(W)\left[\delta_*(t_1, W) - \delta_*(t_2, W)\right]\right\}\| \\ &\lesssim \|p^k(w)^\top Q_k^{-1/2}\|\, \|t_1 - t_2\|. \qquad (1.\mathrm{I}.7)\end{aligned}$$

Thus, for any $f_1 \coloneqq f(\cdot, t_1), f_2 \coloneqq f(\cdot, t_2) \in \mathcal{F}_k$, by T,

$$
\begin{aligned}
|f_1(z) - f_2(z)| &\leqslant |y - h_*(w)| \left[|\delta_k(t_1, w) - \delta_k(t_2, w)| + |\delta_*(t_1, w) - \delta_*(t_2, w)|\right] \\
&\leqslant C\,|y - h_*(w)| \left\{\|p^k(w)^\top Q_k^{-1/2}\| + \mathrm{E}\left[|\partial_v \rho_*(Z, h_*(W))|\,|W = w\right]\right\} \|t_1 - t_2\| \\
&\eqqcolon F_{1,k}(z)\,\|t_1 - t_2\|.
\end{aligned}
$$

Moreover, for any $f \coloneqq f(\cdot, t) \in \mathcal{F}_k$,

$$
|f(z)| = |y - h_*(w)|\,|\delta_k(t, w) - \delta_*(t, w)| \leqslant |y - h_*(w)|\,\|\delta_k(\cdot, w) - \delta_*(\cdot, w)\|_T \eqqcolon F_{2,k}(z).
$$

Using Assumptions 1.3 and 1.4, the inequality $(a + b)^2 \leqslant 2a^2 + 2b^2$, and the fact that conditional expectations are $L^2(P)$ contractions, we see that

$$
\begin{aligned}
\mathrm{E}[F_{1,k}(Z)^2] &\lesssim \mathrm{E}\Big(U^2 \left\{\|p^k(W)^\top Q_k^{-1/2}\| + \mathrm{E}\left[|\partial_v \rho_*(Z, h_*(W))|\,|\,W\right]\right\}^2\Big) \\
&\lesssim \mathrm{E}[\|p^k(W)^\top Q_k^{-1/2}\|^2] + \mathrm{E}\left(\{\mathrm{E}\left[|\partial_v \rho_*(Z, h_*(W))|\,|\,W\right]\}^2\right) \\
&\leqslant k + \mathrm{E}\left[\partial_v \rho_*(Z, h_*(W))^2\right] \lesssim k \ \text{ as } k \to \infty.
\end{aligned}
$$

Given Assumptions 1.4 and 1.7, we get

$$
\mathrm{E}[F_{2,k}(Z)^2] = \mathrm{E}\{U^2 \|\delta_k(\cdot, W) - \delta_*(\cdot, W)\|_{\mathcal{T}}^2\} \lesssim \mathrm{E}\{\|\delta_k(\cdot, W) - \delta_*(\cdot, W)\|_{\mathcal{T}}^2\} = R_{\delta,k}^2 \to 0
$$

as $k \to \infty$. Thus, defining $F_k \coloneqq F_{1,k} + F_{2,k}$ we must have

$$
\mathrm{E}[F_k(Z)^2] \lesssim k + R_{\delta,k}^2 \lesssim k \ \text{ as } k \to \infty,
$$

and it follows that F_k is a square-integrable envelope for $\mathcal{F}_k$ satisfying

$$
|f_1(z) - f_2(z)| \leqslant F_k(z)\,\|t_1 - t_2\| \quad \text{and} \quad \|F_k\|_{P,2} \lesssim k^{1/2} \ \text{ as } k \to \infty.
$$

Using $\mathcal{T}$ compact and the previous display, van der Vaart and Wellner (1996, 2.7.11) implies that

$$
N_{[\,]}(\varepsilon \|F_k\|_{P,2}, \mathcal{F}_k, L^2(P)) \leqslant (C/\varepsilon)^{d_t}, \quad \varepsilon \in (0, 1],
$$

and thus

$$
J_{[\,]}\left(\delta, \mathcal{F}_k, L^2(P)\right) \leqslant \int_0^\delta \sqrt{1 + d_t \ln(C/\varepsilon)}\mathrm{d}\varepsilon, \quad \delta > 0.
$$

where the right-hand side does not depend on k. In particular, $J_{[\,]}\left(1, \mathcal{F}_{k_n}, L^2(P)\right) \lesssim 1$

Defining

$$\sigma_n^2 \coloneqq \sup_{f \in \mathcal{F}_{k_n}} \mathbb{E}_n\left(f^2\right)$$

we see that

$$\sigma_n^2 = \sup_{t \in \mathcal{T}} \mathbb{E}_n\{U_i^2 \left[\delta_{k_n}(t, W_i) - \delta_*(t, W_i)\right]^2\} \leqslant \mathbb{E}_n\{U_i^2 \|\delta_{k_n}(\cdot, W_i) - \delta_*(\cdot, W_i)\|_{\mathcal{T}}^2\}$$

such that

$$\mathrm{E}(\sigma_n^2) \leqslant \mathrm{E}\{U^2 \|\delta_{k_n}(\cdot, W) - \delta_*(\cdot, W)\|_{\mathcal{T}}^2\} \lesssim \mathrm{E}\{\|\delta_{k_n}(\cdot, W) - \delta_*(\cdot, W)\|_{\mathcal{T}}^2\} = R_{\delta, k_n}^2.$$

There are two cases: (1) $R_{\delta,k_n}/\|F_{k_n}\|_{P,2} \to 0$ and (2) $\mathbf{R}_{\delta,k_n}/\|F_{k_n}\|_{P,2} \nrightarrow 0$.

Case 1: $R_{\delta,k_n}/\|F_{k_n}\|_{P,2} \to 0$. Given that $\sqrt{\mathrm{E}\left(\sigma_n^2\right)} \leqslant C_1 R_{\delta,k_n}$, by the change of variables $\varepsilon' := \varepsilon/C_1$ we have

$$\begin{aligned}
J_{[\,]}\left(\sqrt{\mathrm{E}\left(\sigma_n^2\right)}/\|F_{k_n}\|_{P,2}, \mathcal{F}_{k_n}, L^2(P)\right) &\leqslant J_{[\,]}\left(C_1 R_{\delta,k_n}/\|F_{k_n}\|_{P,2}, \mathcal{F}_{k_n}, L^2(P)\right) \\
&= C_1 \int_0^{R_{\delta,k_n}/\|F_{k_n}\|_{P,2}} \sqrt{1 + d_t \ln\left(C_3/\varepsilon'\right)} \mathrm{d}\varepsilon' \\
&=: C_1 \overline{J}_{[\,]}\left(\Delta_{k_n}/\|F_{k_n}\|_{P,2}\right) \qquad (1.\mathrm{I}.8)
\end{aligned}$$

van der Vaart and Wellner (2011, p. 196) establishes the maximal inequality

$$\mathrm{E}(\|\mathbb{G}_n\|_{\mathcal{F}_{k_n}}) \lesssim J_{[\,]}\left(\sqrt{\mathrm{E}\left(\sigma_n^2\right)}/\|F_{k_n}\|_{P,2}, \mathcal{F}_{k_n}, L^2(P)\right) \|F_{k_n}\|_{P,2}.$$

The previous two displays show that

$$\mathrm{E}(\|\mathbb{G}_n\|_{\mathcal{F}_{k_n}}) \lesssim \overline{J}_{[\,]}\left(\Delta_{k_n}/\|F_{k_n}\|_{P,2}\right) \|F_{k_n}\|_{P,2}$$

and from van der Vaart and Wellner (1996, p. 239) we know that an integral of the form $\int_0^\delta [1+\ln(1/u)]^{1/2} \mathrm{d}u$—as in (1.I.8)—satisfies $\int_0^\delta [1+\ln(1/u)]^{1/2} \mathrm{d}u \lesssim \delta\sqrt{\ln(1/\delta)}$ as $\delta \downarrow 0$. Since $R_{\delta,k_n}/\|F_{\delta,k_n}\|_{P,2} \to 0$ holds by hypothesis, the previous display combined with $\|F_{k_n}\|_{P,2} \lesssim \sqrt{k_n}$ and M yields

$$\begin{aligned}
\|\mathbb{G}_n\|_{\mathcal{F}_{k_n}} &\lesssim_{\mathrm{P}} \left(R_{\delta,k_n}/\|F_{k_n}\|_{\mathrm{P},2}\right) \sqrt{\ln\left(\|F_{k_n}\|_{P,2}/R_{\delta,k_n}\right)} \|F_{k_n}\|_{P,2} \\
&= R_{\delta,k_n} \sqrt{\ln\left(\|F_{k_n}\|_{P,2}/R_{\delta,k_n}\right)} \lesssim R_{\delta,k_n} \sqrt{\ln\left(k_n/R_{\delta,k_n}\right)}.
\end{aligned}$$

Case 2. $R_{\delta,k_n}/\|F_{k_n}\|_{P,2} \nrightarrow 0$. Given that $R_{\delta,k_n} \to 0$ (Assumption 1.7), we must have $\|F_{k_n}\|_{P,2} \lesssim R_{\delta,k}$. van der Vaart and Wellner (1996, Theorem 2.14.2) and $J_{[\,]}\left(1, \mathcal{F}_{k_n}, L^2(P)\right) \lesssim 1$ yield

$$\mathrm{E}(\|\mathbb{G}_n\|_{\mathcal{F}_{k_n}}) \lesssim J_{[\,]}\left(1, \mathcal{F}_{k_n}, L^2(P)\right) \|F_{k_n}\|_{P,2} \lesssim \|F_{k_n}\|_{P,2} \lesssim R_{\delta,k_n} \lesssim R_{\delta,k_n}\sqrt{\ln\left(k_n/R_{\delta,k_n}\right)}.$$

M now yields the same rate as in Case 1. In either case, $\|\mathrm{IV}_{d,n}\|_{\mathcal{T}} \lesssim_{\mathrm{P}} R_{\delta,k_n}\sqrt{\ln\left(k_n/R_{\delta,k_n}\right)}$.

For the purpose of bounding $\|\mathrm{IV}_{e,n}\|_{\mathcal{T}}$, consider the function class $\mathcal{F}_k \coloneqq \{f : z \mapsto \delta_k(t,w)\left[h_k(w) - h_*(w)\right] | t \in \mathcal{T}\}$. Note that, by orthogonality of mean–square projections we have $\mathrm{E}[f(Z)] = 0$ for any $f \in \mathcal{F}_k$, so we may view the stochastic process $\{\mathrm{IV}_{e,n}(t) |\, t \in \mathcal{T}\}$ as an empirical process $\{\mathbb{G}_n(f) |\, f \in \mathcal{F}_{k_n}\}$. For any $t_1, t_2 \in \mathcal{T}$, using the bound in (1.I.7) we have that $f_1 \coloneqq f(\cdot; t_1), f_2 \coloneqq f(\cdot; t_2) \in \mathcal{F}_k$, satisfy

$$\begin{aligned} |f_1(z) - f_2(z)| &= |\delta_k(t_1,w) - \delta_k(t_2,w)|\, |h_k(w) - h_*(w)| \\ &\lesssim \|p^k(w)^\top Q_k^{-1/2}\|\, |h_k(w) - h_*(w)|\, \|t_1 - t_2\| \\ &\lesssim \zeta_k\, |h_k(w) - h_*(w)|\, \|t_1 - t_2\|. \end{aligned}$$

The previous display implies

$$|f_1(z) - f_2(z)| \leqslant F_{1,k}(z)\, \|t_1 - t_2\|,$$

for $F_{1,k}(z) \coloneqq C_1 \zeta_k\, |h_k(w) - h_*(w)|$ and some $C_1 \in (0,\infty)$. Since conditional expectations are $L^2(P)$ contractions, by Assumptions 1.2 and 1.3,

$$\begin{aligned} \mathrm{E}[\delta_*(t,W)^2] &= \mathrm{E}\Big(\mathrm{E}\left[\omega(t,X)\,|\partial_v \rho_*(Z, h_*(W))|\,|\, W\right]^2\Big) \leqslant \mathrm{E}\left[\omega(t,X)^2\, |\partial_v \rho_*(Z, h_*(W))|^2\right] \\ &\lesssim \mathrm{E}\Big[\,|\partial_v \rho_*(Z, h_*(W))|^2\Big] < \infty, \end{aligned}$$

thus implying $\sup_{t\in\mathcal{T}} \mathrm{E}[\delta_*(t,W)^2] < \infty$. By CS and using that mean–square projections are $L^2(P)$ contractions as well, we get

$$\begin{aligned} |\delta_k(t,w)| &= |p^k(w)^\top Q_k^{-1}\mathrm{E}[p^k(W)\,\delta_*(t,W)]| \leqslant \|p^k(w)^\top Q_k^{-1/2}\| \|Q_k^{-1/2}\mathrm{E}[p^k(W)\,\delta_*(t,W)]\| \\ &\lesssim \|p^k(w)\|\mathrm{E}[\delta^*(t,W)^2] \lesssim \zeta_k, \end{aligned}$$

which implies that for any $f \coloneqq f(\cdot;t) \in \mathcal{F}_k$,

$$|f(z)| = |\delta_k(w;t)|\, |h_k(w) - h^*(w)| \lesssim \zeta_k\, |h_k(w) - h^*(w)|.$$

The previous diplay shows that $|f(z)| \leqslant F_{2,k}(z)$ for $F_{2,k}(z) \coloneqq C_2\zeta_k |h_k(w) - h^*(w)|$ and some $C_2 \in (0,\infty)$. Let $C_3 \coloneqq C_1 \vee C_2$, and define $F_k(z) \coloneqq C_3\zeta_k |h_k(w) - h^*(w)|$. Then by Assumption 1.7,

$$\|F_k\|_{P,2} = C_3\zeta_k \|h_k - h^*\|_{P,2} = C_3\zeta_k r_{h,k} \to 0 \text{ as } k \to \infty,$$

In particular, $\|F_k\|_{P,2} \lesssim 1$. Now, F_k is a square-integrable envelope for $\mathcal{F}_k$ satisfying

$$|f_1(z) - f_2(z)| \leqslant F_k(z) \|t_1 - t_2\|.$$

Using $\mathcal{T}$ compact and the previous display, by van der Vaart and Wellner (1996, Theorem 2.7.11) we see that

$$N_{[\,]}(\varepsilon\|F_k\|_{P,2}, \mathcal{F}_k, L^2(\mathrm{P})) \leqslant (C/\varepsilon)^{d_t}, \quad \varepsilon \in (0,1],$$

and thus

$$J_{[\,]}\left(\delta, \mathcal{F}_k, L^2(P)\right) \leqslant \int_0^\delta \sqrt{1 + d_t \ln(C/\varepsilon)} \mathrm{d}\varepsilon,$$

where the right-hand side does not depend on k. In particular, $J_{[\,]}(1, \mathcal{F}_k, L^2(P)) \lesssim 1$. Using van der Vaart and Wellner (1996, Theorem 2.14.2) $J_{[\,]}(1, \mathcal{F}_{k_n}, L^2(P)) \lesssim 1$, we arrive at

$$\mathrm{E}(\|\mathbb{G}_n\|_{\mathcal{F}_{k_n}}) \lesssim J_{[\,]}\left(1, \mathcal{F}_{k_n}, L^2(P)\right) \|F_{k_n}\|_{P,2} \lesssim \|F_{k_n}\|_{P,2} \lesssim \zeta_{k_n} r_{h,k_n},$$

so $\|\mathrm{IV}_{e,n}\|_{\mathcal{T}} \lesssim_{\mathrm{P}} \zeta_{k_n} r_{h,k_n}$ by M. □

Proof of Lemma 1.1. The claim follows from Lemma 1.14 and Assumption 1.7. □

Proof of Lemma 1.2. Given that β_* and h_* are held fixed throughout the argument, abbreviate $\rho_{**}(z) \coloneqq \rho(z, \beta_*, h_*(w))$, $\partial_\beta \rho_{**}(Z) \coloneqq \partial_\beta \rho(z, \beta_*, h_*(w))$ and $\partial_v \rho_{**}(Z) \coloneqq \partial_v \rho(z, \beta_*, h_*(w))$. By Assumption 1.2 and J we have both $\|b_*(t)\| \leqslant \mathrm{E}[|\omega(t,X)| \|\partial_\beta \rho_{**}(Z)\|] \lesssim \mathrm{E}[\|\partial_\beta \rho_{**}(Z)\|]$ and $|\delta_*(t,w)| \leqslant \mathrm{E}[|\omega(t,X)| |\partial_v \rho_{**}(Z)| | W = w] \lesssim \mathrm{E}[|\partial_v \rho_{**}(Z)| | W = w]$. Letting $f(t,\cdot) \in \mathcal{F}$ be arbitrary, T and CS therefore imply

$$\begin{aligned} |f(t,z)| &\leqslant |\omega(t,x)| |\rho_{**}(z)| + \|b_*(t)\| \|s_*(z)\| + |\delta_*(t,w)| |y - h_*(w)| \\ &\leqslant C_1 |\rho_{**}(z)| + \mathrm{E}[\|\partial_\beta \rho_{**}(Z)\|] \|s_*(z)\| + \mathrm{E}[|\partial_v \rho_{**}(Z)| | W = w] |y - h_*(w)| \eqqcolon F_1(z). \end{aligned}$$

Taking the expectation and using the inequality $(a+b)^2 \leqslant 2a^2 + 2b^2$ repeatedly alongside the integrability and boundedness parts of Assumptions 1.1 and 1.3, we see that $F_1(Z)^2$ is integrable. Hence, F is a square-integrable envelope for $\mathcal{F}$. Let $f(t_1,\cdot), f(t_2,\cdot) \in \mathcal{F}$ be arbitrary. Then by T and CS, followed by J, CS and Assumption 1.2,

$$\begin{aligned} |f(t_1,z)-f(t_2,z)| &\leqslant |\omega(t_1,x)-\omega(t_2,x)|\,|\rho_{**}(z)| + \|b^*(t_1)-b^*(t_2)\|\,\|s_*(z)\| \\ &\quad + |y-h_*(w)|\,|\delta_*(t_1,w)-\delta_*(t_2,w)| \\ &\leqslant C_2\Big(|\rho_{**}(z)| + \mathrm{E}[\|\partial_\beta\rho_{**}(Z)\|]\,\|s_*(z)\| \\ &\quad + |y-h_*(w)|\,\mathrm{E}[|\partial_v\rho_{**}(Z)|\,|W=w]\Big)\|t_1-t_2\| =: F_2(z)\,\|t_1-t_2\| \end{aligned}$$

Defining $F \coloneqq F_1 \vee F_2$, we see that F is a square-integrable envelope for $\mathcal{F}$ satisfying

$$|f(t_1,z)-f(t_2,z)| \leqslant F(z)\,\|t_1-t_2\|\,.$$

Given that $\mathcal{T}$ is compact (Assumption 1.2), we thus have

$$N_{[\,]}(\varepsilon\,\|F\|_{P,2}\,, \mathcal{F}, L^2(P)) \leqslant N(\varepsilon, \mathcal{T}, \|\cdot\|) \leqslant (\operatorname{diam}(\mathcal{T})/\varepsilon)^{d_t} \leqslant (C/\varepsilon)^{d_t}\,,\ \varepsilon \in (0, \operatorname{diam}(\mathcal{T})],$$

so using $\|F\|_{P,2} < \infty$,

$$N_{[\,]}(\varepsilon, \mathcal{F}, L^2(P)) \leqslant (C/\varepsilon)^{d_t}\,, \quad \varepsilon > 0.$$

The previous display implies

$$\int_0^\infty \sqrt{\ln(N_{[\,]}(\varepsilon, \mathcal{F}, L^2(P)))}\mathrm{d}\varepsilon \leqslant \sqrt{d_t}\int_0^\infty \sqrt{\ln(C/\varepsilon)}\mathrm{d}\varepsilon < \infty.$$

The desired conclusion now follows from van der Vaart (2000, Theorem 19.5), which uses the Ossiander (1987) sufficient condition for $\mathcal{F}$ to be Donsker. □

Proof of Theorem 1.1. To prove (1), observe first that under the null, $\mathrm{E}[f_*(\cdot,Z)]$ is the zero function on $\mathcal{T}$, and B_n^* equals the empirical process $\{\mathbb{G}_n(f)\,|f \in \mathcal{F}\}$. $\mathcal{F}$ being Donsker (Lemma 1.2) is equivalent to $\mathbb{G}_n \rightsquigarrow \mathbb{G}_0$ in $\ell^\infty(\mathcal{F})$ for a centered Gaussian process $\mathbb{G}_0$ with covariance function $\mathrm{E}[f_1(Z)f_2(Z)], f_1, f_2 \in \mathcal{F}$, which, by definition of $\mathcal{F}$, is equivalent to $B_n^* \rightsquigarrow G_0$ in $\ell^\infty(\mathcal{T})$ for a centered Gaussian process G_0 with covariance function $\mathrm{E}[f(t_1,Z)f(t_2,Z)], t_1, t_2 \in \mathcal{T}$. By T and Lemma 1.1,

$$\big|T_n^{1/2} - \|B_n^*\|_{\mu,2}\big| = |\|\widehat{B}_n\|_{\mu,2} - \|B_n^*\|_{\mu,2}| = || \leqslant \|\widehat{B}_n - B_n^*\|_{\mu,2} \xrightarrow{\mathrm{P}} 0.$$

so Part 1 follows from the CMT.

To prove (2), note that by the previous display

$$\left|T_n/n - \|B_n^*/n\|_{\mu,2}\right| = \left|T_n - \|B_n^*\|_{\mu,2}\right|/n \xrightarrow{\mathrm{P}} 0.$$

The proof of Lemma 1.2 shows that $t \mapsto f_*(t,Z)$ is continuous at each $t \in \mathcal{T}$ with probability one, and $\mathcal{F}$ admits a square-integrable envelope. Given that the data are i.i.d., and $\mathcal{T}$ is compact (Assumption 1.2), Newey and McFadden (1994, Lemma 2.4) implies that

$$\sup_{t\in\mathcal{T}}\left|B_n^*(t)/\sqrt{n} - \mathrm{E}\left[\rho(Z,\beta_*,h_*(W))\,\omega(t,X)\right]\right| = \sup_{t\in\mathcal{T}}\left|(\mathbb{E}_n - \mathrm{E})\left[f_*(t,Z)\right]\right| \xrightarrow{\mathrm{P}} 0.$$

Given that $\left|T_n/n - \|B_n^*/n\|_{\mu,2}\right| = \left|T_n - \|B_n^*\|_{\mu,2}\right|/n \to_{\mathrm{P}} 0$ (Lemma 1.1), the previous diplay implies $T_n \to_{\mathrm{P}} \int_{\mathcal{T}}\left\{\mathrm{E}\left[\rho(Z,\beta_*,h_*(W))\,\omega(t,X)\right]\right\}^2 \mathrm{d}\mu(t)$, which is strictly positive under the alternative by the choice of weight function (Assumption 1.2) and measure.

1.I.2 Proofs for Section 1.4.5

□

Define the stochastic processes $\widehat{G}^u$ and G_n^{*u} by

$$\widehat{G}^u(t) \coloneqq \frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i \widehat{f}(t,Z_i),$$
$$G_n^{*u}(t) \coloneqq \frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i f_*(t,Z_i).$$

which are the "*u*ncentered" versions of $\widehat{G}$ and G_n^*, respectively, i.e., the displayed processes are not centered at the sample mean. The following lemma shows that the uncentered processes are asymptotically equivalent.

Lemma 1.15. *If Assumptions 1.1–1.8 hold, then* $\|\widehat{G}^u - G_n^{*u}\|_{\mathcal{T}} \to_{\mathrm{P}} 0$.

Proof of Lemma *1.15.*

Main

For fixed $t \in \mathcal{T}$ a decomposition yields

$$\begin{aligned}\widehat{G}^u(t) - G_n^{*u}(t) &= \sqrt{n}\mathbb{E}_n\left\{\xi_i[\widehat{f}(t,Z_i) - f_*(t,Z_i)]\right\}\\ &= \sqrt{n}\mathbb{E}_n\left\{\xi_i\omega(t,X_i)\left[\rho(Z_i,\widehat{\beta},\widehat{h}(W_i)) - \rho(Z_i,\beta_*,h_*(W_i))\right]\right\}\end{aligned}$$

$$
\begin{aligned}
&- [\widehat{b}(t) - b_*(t)]^\top \sqrt{n}\mathbb{E}_n [\xi_i s_*(Z_i)] \\
&- \widehat{b}(t)^\top \sqrt{n}\mathbb{E}_n \big\{\xi_i[\widehat{s}(Z_i) - s_*(Z_i)]\big\} \\
&+ \sqrt{n}\mathbb{E}_n \big(\xi_i\{\widehat{\delta}(t, W_i)[Y_i - \widehat{h}(W_i)] - \delta_*(t, W_i) U_i\}\big). \\
=:\ & \mathrm{I}_n(t) + \mathrm{II}_n(t) + \mathrm{III}_n(t) + \mathrm{IV}_n(t).
\end{aligned}
$$

The following steps show that the four remainder terms $\to_{\mathrm{P}} 0$ uniformly over $\mathcal{T}$. The claim therefore follows from T.

$\|\mathrm{I}_n\|_{\mathcal{T}}$

Assumption 1.1 and M implies that $\|\widehat{\beta}-\beta_*\| \lesssim_{\mathrm{P}} n^{-1/2} \to 0$. Let $\mathcal{N}_*$ be the open neighborhood provided by Assumption 1.3. Then $\widehat{\beta} \in \mathcal{N}_*$ wp $\to$ 1. To simplify notation and ensure that objects are globally well defined, in what follows I will—without loss of generality—assume that $\widehat{\beta} \in \mathcal{N}_*$ with probability one for all n. A mean value expansion of $\beta \mapsto \rho(Z_i, \beta, \widehat{h}(W_i))$ at $\widehat{\beta}$ around β_* and CS show that

$$
\begin{aligned}
\|\mathrm{I}_n\|_{\mathcal{T}} &\leqslant \sup_{t\in\mathcal{T}} \left|\sqrt{n}\mathbb{E}_n \left\{\xi_i \omega(t, X_i) \left[\rho(Z_i, \beta_*, \widehat{h}(W_i)) - \rho(Z_i, \beta_*, h_*(W_i))\right]\right\}\right| \\
&\quad + \sqrt{n}\|\widehat{\beta} - \beta_*\| \sup_{t\in\mathcal{T}} \left\|\mathbb{E}_n \left[\xi_i \omega(t, X_i)\, \partial_\beta \rho(Z_i, \overline{\beta}, \widehat{h}(W_i))\right]\right\| \\
&=: \|\mathrm{I}_{a,n}\|_{\mathcal{T}} + \sqrt{n}\|\widehat{\beta} - \beta_*\| \|\mathrm{I}_{b,n}\|_{\mathcal{T}},
\end{aligned}
$$

where $\overline{\beta}$ satisfies $\|\overline{\beta} - \beta_*\| \leqslant \|\widehat{\beta} - \beta_*\|$ such that $\overline{\beta} \in \mathcal{N}_*$ for n sufficiently large. Since $\sqrt{n}\|\widehat{\beta} - \beta_*\| \lesssim_{\mathrm{P}} 1$ it suffices to show that $\|\mathrm{I}_{a,n}\|_{\mathcal{T}}$ and $\|\mathrm{I}_{b,n}\|_{\mathcal{T}} \to_{\mathrm{P}} 0$.

$\|\mathrm{I}_{a,n}\|_{\mathcal{T}}$ Abbreviate $(z, v) \mapsto \rho(z, \beta_*, v)$ by ρ_*. By a mean value expansion of $s \mapsto \rho_*(Z_i, s)$ at $\widehat{h}(W_i)$ around $h_*(W_i)$ and T we may be bound $\|\mathrm{I}_{a,n}\|_{\mathcal{T}}$ by

$$
\begin{aligned}
&\sup_{t\in\mathcal{T}} \left|\sqrt{n}\mathbb{E}_n \left\{\xi_i \omega(t, X_i) \left[\partial_v \rho_*(Z_i, \overline{h}(W_i)) - \partial_v \rho_*(Z_i, h_*(W_i))\right] \left[\widehat{h}(W_i) - h_*(W_i)\right]\right\}\right| \\
&\quad + \sup_{t\in\mathcal{T}} \left|\sqrt{n}\mathbb{E}_n \left\{\xi_i \omega(t, X_i)\, \partial_v \rho_*(Z_i, h_*(W_i))[\widehat{h}(W_i) - h_*(W_i)]\right\}\right| =: \|\mathrm{I}_{a,1,n}\|_{\mathcal{T}} + \|\mathrm{I}_{a,2,n}\|_{\mathcal{T}}.
\end{aligned}
$$

By T and Assumptions 1.2 and 1.8

$$
\begin{aligned}
\|\mathrm{I}_{a,1,n}\|_{\mathcal{T}} &\lesssim \sqrt{n}\mathbb{E}_n \left\{|\xi_i| R'(Z_i)\, [\widehat{h}(W_i) - h_*(W_i)]^2\right\} \Big| \\
&\leqslant \mathbb{E}_n [|\xi_i|\, R'(Z_i)]\, \sqrt{n}\|\widehat{h}_n - h_*\|_{\mathcal{W}}^2 \\
&\lesssim_{\mathrm{P}} \mathrm{E}[R'(Z)]\sqrt{n}\|\widehat{h} - h_*\|_{\mathcal{W}}^2 \to_{\mathrm{P}} 0.
\end{aligned}
$$

$\|\mathrm{I}_{a,2,n}\|_{\mathcal{T}}$: Let $\widetilde{h}_k \coloneqq p^{k\top}\widetilde{\pi}_k$ for $\widetilde{\pi}_k$ provided by Assumption 1.6. Then we may bound $\|\mathrm{I}_{a,2,n}\|_{\mathcal{T}}$ by

$$\begin{aligned}\|\mathrm{I}_{a,2,n}\|_{\mathcal{T}} &\leqslant \sup_{t\in\mathcal{T}}\left|\sqrt{n}\mathbb{E}_n\left\{\xi_i\omega\left(t,X_i\right)\partial_v\rho_*(Z_i,h_*\left(W_i\right))[\widehat{h}\left(W_i\right)-\widetilde{h}_{k_n}\left(W_i\right)]\right\}\right| \\ &\quad+\sup_{t\in T}\left|\sqrt{n}\mathbb{E}_n\left\{\xi_i\omega\left(t,X_i\right)\partial_v\rho_*(Z_i,h_*\left(W_i\right))[\widetilde{h}_{k_n}\left(W_i\right)-h_*\left(W_i\right)]\right\}\right| \\ &\eqqcolon \|\mathrm{I}_{a,2,1,n}\|_{\mathcal{T}}+\|\mathrm{I}_{a,2,2,n}\|_{\mathcal{T}}.\end{aligned}$$

I consider $\|\mathrm{I}_{a,2,1,n}\|_{\mathcal{T}}$ and $\|\mathrm{I}_{a,2,2,n}\|_{\mathcal{T}}$ in turn. By CS $\|\mathrm{I}_{a,2,1,n}\|_{\mathcal{T}}$ is bounded by

$$\begin{aligned}\|\mathrm{I}_{a,2,1,n}\|_{T} &\leqslant \|\widehat{\pi}-\widetilde{\pi}_{k_n}\|\sup_{t\in\mathcal{T}}\left\|\sqrt{n}\mathbb{E}_n\left\{\xi_i\omega\left(t,X_i\right)\partial_v\rho_*(Z_i,h_*\left(W_i\right))p^{k_n}\left(W_i\right)\right\}\right\| \\ &\leqslant \|\widehat{\pi}-\widetilde{\pi}_{k_n}\|\Big(\sum_{j=1}^{k_n}\sup_{t\in\mathcal{T}}\left\{\sqrt{n}\mathbb{E}_n\left[\xi_i\omega\left(t,X_i\right)\partial_v\rho_*(Z_i,h_*\left(W_i\right))p_{jk_n}\left(W_i\right)\right]\right\}^2\Big)^{1/2}.\end{aligned}$$

Fix k and let

$$\mathcal{F}'_{jk} \coloneqq \left\{f:(s,z)\mapsto s\omega\left(t,x\right)\partial_v\rho_*\left(z,s\right)p_{jk}\left(w\right)\middle|\, t\in\mathcal{T}\right\}.$$

Note $\mathrm{E}[f(\xi,Z)]=0$ for every $f\in\mathcal{F}'_{jk}$, so $\{\sqrt{n}\mathbb{E}_n\left[f\left(\xi_i,Z_i\right)\right]|f\in\mathcal{F}'_{jk}\}$ is an empirical process. For $f\coloneqq f\left(\cdot,t\right), f_1\coloneqq f\left(\cdot;t_1\right), f_2\coloneqq f\left(\cdot;t_2\right)\in\mathcal{F}'_{jk}$ arbitrary, by Assumption 1.2 we have

$$\begin{aligned}\left|f\left(s,z\right)\right| &\leqslant C_1\left|s\right|\left|\partial_v\rho_*\left(z,h_*\left(w\right)\right)\right|\|p_{jk}\|_{\mathcal{W}} \\ \left|f_1\left(s,z\right)-f_2\left(s,z\right)\right| &\leqslant C_2\left|s\right|\left|\partial_v\rho_*\left(z,h_*\left(w\right)\right)\right|\|p_{jk}\|_{\mathcal{W}}\|t_1-t_2\|.\end{aligned}$$

By Assumption 1.8, CS and the previous display we see that

$$F'_{jk}\left(s,z\right)\coloneqq\left(C_1\vee C_2\right)\left|s\right|\left|\partial_v\rho_*\left(z,h_*\left(w\right)\right)\right|\|p_{jk}\|_{\mathcal{W}}$$

is an envelope for $\mathcal{F}'_{jk}$ satisfying $\mathrm{E}[F'_{jk}(\xi,Z)^2]\propto\|p_{jk}\|^2_{\mathcal{W}}$, which is finite for every (j,k) by Assumption 1.7. Moreover, by compactness of $\mathcal{T}$ (Assumption 1.2) and the previous display,

$$N_{[\,]}(\varepsilon(\mathrm{E}[F'_{jk}(\xi,Z)^2])^{1/2},\mathcal{F}'_{jk},L^2\left(\xi,Z\right))\leqslant N\left(\varepsilon,\mathcal{T},\|\cdot\|\right)\lesssim\varepsilon^{-d_t},\quad \varepsilon\in(0,1].$$

It follows that the bracketing entropy integral $J_{[\,]}(1,\mathcal{F}'_{jk},L^2\left(\xi,Z\right))$ is bounded by a constant independent of j or k, so by van der Vaart and Wellner (1996, Theorem 2.14.2)

$$\mathrm{E}(\|\mathbb{G}_n\|_{\mathcal{F}'_{jk}})\lesssim J_{[\,]}(1,\mathcal{F}'_{jk},L^2\left(\xi,Z\right))\mathrm{E}[F'_{jk}(\xi,Z)^2])^{1/2}\lesssim\mathrm{E}[F'_{jk}(\xi,Z)^2])^{1/2}\propto\|p_{jk}\|_{\mathcal{W}}.$$

van der Vaart and Wellner (1996, Theorem 2.14.5) and the previous display show that

$$[\mathrm{E}(\|\mathbb{G}_n\|^2_{\mathcal{F}'_{jk}})]^{1/2} \lesssim \mathrm{E}(\|\mathbb{G}_n\|_{\mathcal{F}'_{jk}}) + \mathrm{E}[F'_{jk}(\xi, Z)^2])^{1/2} \lesssim \|p_{jk}\|_{\mathcal{W}},$$

Allowing $k = k_n$, the previous display, in turn, implies

$$\mathrm{E}\Big(\sum_{j=1}^{k_n} \|\mathbb{G}_n\|^2_{\mathcal{F}'_{jk_n}}\Big) = \sum_{j=1}^{k_n} \mathrm{E}(\|\mathbb{G}_n\|^2_{\mathcal{F}'_{jk_n}}) \lesssim \sum_{j=1}^{k_n} \|p_{jk_n}\|^2_{\mathcal{W}},$$

so by M we get

$$\sum_{j=1}^{k_n} \|\mathbb{G}_n\|^2_{\mathcal{F}'_{jk_n}} \lesssim_{\mathrm{P}} \sum_{j=1}^{k_n} \|p_{jk_n}\|^2_{\mathcal{W}}.$$

From Lemma 1.22, M and Assumption 1.7 it now follows that

$$\|\mathrm{I}_{a,2,1,n}\|_{\mathcal{T}} \leqslant \|\widehat{\pi}_n - \widetilde{\pi}_{k_n}\| \Big(\sum_{j=1}^{k_n} \|\mathbb{G}_n\|^2_{\mathcal{F}'_{jk_n}}\Big)^{1/2} \lesssim_{\mathrm{P}} (\sqrt{k_n/n} + k_n^{-\alpha}) \Big(\sum_{j=1}^{k_n} \|p_{jk_n}\|^2_{\mathcal{W}}\Big)^{1/2} \to 0.$$

Similarly, fix k and let

$$\mathcal{F}'_k \coloneqq \{f : (s, z) \mapsto s\omega(t, x)\, \partial_v \rho_*(z, h_*(w))\, [\widetilde{h}_k(w) - h_*(w)] \big| t \in \mathcal{T}\}.$$

Note $\mathrm{E}[f(\xi, Z)] = 0$ for every $f \in \mathcal{F}'_k$, so $\{\sqrt{n}\mathbb{E}_n [f(\xi_i, Z_i)] \,| f \in \mathcal{F}'_k\}$ is an empirical process. For $f \coloneqq f(\cdot; t), f_1 \coloneqq f(\cdot; t_1), f_2 \coloneqq f(\cdot; t_2) \in \mathcal{F}'_{jk}$ arbitrary, by Assumption 1.2 we have

$$|f(s, z)| \leqslant C_1 |s| |\partial_v \rho_*(z, h_*(w))| \|\widetilde{h}_k - h_*\|_{\mathcal{W}},$$
$$|f_1(s, z) - f_2(s, z)| \leqslant C_2 |s| |\partial_v \rho_*(z, h_*(w))| \|\widetilde{h}_k - h_*\|_{\mathcal{W}} \|t_1 - t_2\|.$$

By Assumption 1.8, CS and the previous display we see that

$$F'_k(s, z) \coloneqq (C_1 \vee C_2)\, |s|\, |\partial_v \rho_*(z, h_*(w))|\, \|\widetilde{h}_k - h_*\|_{\mathcal{W}}$$

is an envelope for $\mathcal{F}'_k$ satisfying $\mathrm{E}[F'_k(\xi, Z)^2] \propto \|\widetilde{h}_k - h_*\|^2_{\mathcal{W}}$, which by Assumption 1.6 is finite for every (j, k). Moreover, by compactness of $\mathcal{T}$ (Assumption 1.2) and the previous display,

$$N_{[\,]}(\varepsilon(\mathrm{E}[F'_k(\xi, Z)^2])^{1/2}, \mathcal{F}'_k, L^2(\xi, Z)) \leqslant N(\varepsilon, \mathcal{T}, \|\cdot\|) \lesssim \varepsilon^{-d_t}, \quad \varepsilon \in (0, 1].$$

which implies that the bracketing entropy integral $J_{[\,]}(1, \mathcal{F}'_k, L^2(\xi, Z))$ is bounded by a constant independent of j or k. Using van der Vaart and Wellner (1996, Theorem 2.14.2) and Assumption 1.6, we therefore get

$$\mathrm{E}(\|\mathbb{G}_n\|_{\mathcal{F}'_k}) \lesssim J_{[\,]}(1, \mathcal{F}'_k, L^2(\xi, Z))\mathrm{E}[F'_k(\xi, Z)^2])^{1/2} \lesssim \mathrm{E}[F'_k(\xi, Z)^2])^{1/2} \propto \|\widetilde{h}_k - h^*\|_{\mathcal{W}} \lesssim k^{-\alpha}.$$

By M it follows that $\|\mathrm{I}_{a,2,2,n}\|_{\mathcal{T}} = \|\mathbb{G}_n\|_{\mathcal{F}'_{k_n}} \lesssim_{\mathrm{P}} k_n^{-\alpha} \to 0$, which completes the proof of $\|\mathrm{I}_{a,2,n}\|_{\mathcal{T}} \to_{\mathrm{P}} 0$ and therefore $\|\mathrm{I}_{a,n}\|_{\mathcal{T}} \to_{\mathrm{P}} 0$.

$\|\mathrm{I}_{b,n}\|_{\mathcal{T}}$ By T we may bound $\|\mathrm{I}_{b,n}\|_{\mathcal{T}} \equiv \sup_{t\in\mathcal{T}}\|\mathbb{E}_n[\xi_i \omega(t, X_i)\, \partial_\beta \rho(Z_i, \overline{\beta}, \widehat{h}(W_i))]\|$ by

$$\begin{aligned}&\sup_{t\in\mathcal{T}} \left\|\mathbb{E}_n\left[\xi_i \omega(t, X_i)\, \partial_\beta \rho(Z_i, \overline{\beta}, h_*(W_i))\right]\right\| \\ &\quad + \sup_{t\in\mathcal{T}} \left\|\mathbb{E}_n\left\{\xi_i \omega(t, X_i)\left[\partial_\beta \rho(Z_i, \overline{\beta}, \widehat{h}(W_i)) - \partial_\beta \rho(Z_i, \overline{\beta}, h_*(W_i))\right]\right\}\right\| =: \|\mathrm{I}_{b,1,n}\|_{\mathcal{T}} + \|\mathrm{I}_{b,2,n}\|_{\mathcal{T}}.\end{aligned}$$

The second term $\|\mathrm{I}_{b,2,n}\|_{\mathcal{T}}$ satisfies

$$\begin{aligned}\|\mathrm{I}_{b,2,n}\|_{\mathcal{T}} &\lesssim \mathbb{E}_n\left[|\xi_i|\, L_1(Z_i)\, |\widehat{h}(W_i) - h_*(W_i)|^c\right] \\ &\leqslant \mathbb{E}_n\left[|\xi_i|\, L_1(Z_i)\right] \|\widehat{h} - h_*\|^c_{\mathcal{W}} \lesssim_{\mathrm{P}} \|\widehat{h} - h_*\|^c_{\mathcal{W}} \xrightarrow{\mathrm{P}} 0,\end{aligned}$$

where the $\lesssim$ follows from Assumptions 1.2 and 1.3, the $\lesssim_{\mathrm{P}}$ from the ξ_i's being i.i.d., zero mean, unit variance (hence having finite first moment) and independent of the data, and the $\to_{\mathrm{P}} 0$ stems from Lemma 1.22 and Assumption 1.7.

To show that $\|\mathrm{I}_{b,1,n}\|_{\mathcal{T}} \to_{\mathrm{P}} 0$, observe that the $\{(\xi_i, Z_i)\}_1^n$ are i.i.d., the map $(t, \beta) \mapsto \xi\omega(t, X)\, \partial_\beta \rho(Z, \beta, h_*(W))$ is continuous on $\mathcal{T} \times \mathcal{N}_*$ (Assumptions 1.2 and 1.3) and therefore continuous on the product $\mathcal{T} \times \overline{B}$, where $\overline{B} \subset \mathcal{N}_*$ is a closed ball with center β_* and sufficiently small radius (Assumption 1.1). Moreover, $\mathcal{T} \times \overline{B}$ is compact (Assumption 1.2), and $\sup_{\mathcal{T}\times\overline{B}}\|\xi\omega(t, X)\, \partial_\beta \rho(Z, \beta, h_*(W))\| \lesssim |\xi| \sup_{\overline{B}}\|\partial_\beta \rho(Z, \beta, h_*(W))\|$, where by independence, CS, and Assumption 1.3,

$$\mathrm{E}\Big[|\xi| \sup_{\beta\in\overline{B}} \|\partial_\beta \rho(Z, \beta, h_*(W))\|\Big] \leqslant \mathrm{E}\Big[\sup_{\beta\in\overline{B}} \|\partial_\beta \rho(Z, \beta, h_*(W))\|\Big] < \infty,$$

Given that the ξ_i's are centered and independent of the data, Newey and McFadden (1994, Lemma 2.4) shows that

$$\sup_{\mathcal{T}\times\overline{B}} \|\mathbb{E}_n[\xi_i \omega(t, X_i)\, \partial_\beta \rho(Z_i, \beta, h_*(W_i))]\| \xrightarrow{\mathrm{P}} 0.$$

$\|\mathrm{I}_{b,n}\|_{\mathcal{T}} \to_{\mathrm{P}} 0$ now follows from $\overline{\beta} \in \overline{B}$ wp → 1 and the previous display.

$\widehat{b}$ In this step I show that

$$\text{(a) } \sup_{t\in\mathcal{T}} \|\widehat{b}\,(t) - b_*\,(t)\| \xrightarrow{\mathrm{P}} 0 \quad \text{and} \quad \text{(b) } \sup_{t\in\mathcal{T}} \|\widehat{b}\,(t)\| \lesssim_{\mathrm{P}} 1,$$

To show (a), note that the argument in Section 1.I.1 of the proof of Lemma 1.14 shows that

$$\begin{aligned} &(t,\beta) \mapsto \mathrm{E}\left[\omega\,(t,X)\,\partial_\beta \rho\,(Z,\beta,h_*\,(W))\right] \text{ is uniformly continuous on } \mathcal{T} \times \overline{B}, \\ &\text{and } \sup_{\mathcal{T}\times\overline{B}} \|(\mathbb{E}_n - \mathrm{E})\,\omega\,(t,X_i)\,\partial_\beta \rho\,(Z_i,\beta,h_*\,(W_i))\| \xrightarrow{\mathrm{P}} 0, \end{aligned}$$

where $\overline{B} \subset \mathcal{N}_*$ is a closed ball with center β_* and sufficiently small radius (Assumption 1.1). By T we have

$$\begin{aligned} \sup_{t\in\mathcal{T}} \|\widehat{b}\,(t) - b_*\,(t)\| \leqslant{} & \sup_{t\in\mathcal{T}} \left\| \mathbb{E}_n \left\{ \omega\,(t,X_i) \left[\partial_\beta \rho(Z_i, \widehat{\beta}, \widehat{h}\,(W_i)) - \partial_\beta \rho(Z_i, \widehat{\beta}, h_*\,(W_i)) \right] \right\} \right\| \\ & + \sup_{t\in\mathcal{T}} \left\| (\mathbb{E}_n - \mathrm{E}_Z) \left[\omega\,(t,X_i)\,\partial_\beta \rho(Z_i, \widehat{\beta}, h_*\,(W_i)) \right] \right\| \\ & + \sup_{t\in\mathcal{T}} \left\| \mathrm{E}_Z \left[\omega\,(t,X)\,\partial_\beta \rho(Z, \widehat{\beta}, h_*\,(W)) \right] - b_*\,(t) \right\|. \end{aligned}$$

Given that $\widehat{\beta} \in \overline{B}$ wp → 1, the second and third term on the right $\to_{\mathrm{P}} 0$ due to uniform convergence and uniform continuity, respectively. By T and Assumptions 1.2 and 1.3, the first term is bounded by a constant multiple of

$$\mathbb{E}_n \left[L_1\,(Z_i)\,|\widehat{h}\,(Z_i) - h_*\,(Z_i)|^c \right] \leqslant \mathbb{E}_n\left[L_1\,(Z_i)\right] \|\widehat{h} - h_*\|^c_{\mathcal{W}} \lesssim_{\mathrm{P}} \|\widehat{h} - h_*\|^c_{\mathcal{W}} \xrightarrow{\mathrm{P}} 0,$$

where the $\lesssim_{\mathrm{P}}$ follows from M and the $\to_{\mathrm{P}} 0$ from Lemma 1.22. The previous display finishes the proof of (a).

To show (b), note that the argument in Section 1.I.1 of the proof of Lemma 1.14 also shows that $\sup_{t\in\mathcal{T}} \|b_*\,(t)\| < \infty$. Two applications of T yield

$$\left| \sup_{t\in\mathcal{T}} \|\widehat{b}\,(t)\| - \sup_{t\in\mathcal{T}} \|b_*\,(t)\| \right| \leqslant \sup_{t\in\mathcal{T}} \left| \|\widehat{b}\,(t)\| - \|b_*\,(t)\| \right| \leqslant \sup_{t\in\mathcal{T}} \|\widehat{b}\,(t) - b_*\,(t)\| \xrightarrow{\mathrm{P}} 0,$$

which combined with $\sup_{t\in\mathcal{T}} \|b_*\,(t)\| < \infty$ implies $\sup_{t\in\mathcal{T}} \|\widehat{b}\,(t)\| \lesssim_{\mathrm{P}} 1$.

$\|\mathrm{II}_n\|_{\mathcal{T}}$. By CS, Step 1.I.1 and

$$\|\mathrm{II}_n\|_{\mathcal{T}} \leqslant \left\|\sqrt{n}\mathbb{E}_n\left[\xi_i s_*\left(Z_i\right)\right]\right\| \sup_{t\in\mathcal{T}}\|\widehat{b}\left(t\right) - b_*\left(t\right)\|,$$

it suffices to show $\|\sqrt{n}\mathbb{E}_n\left[\xi_i s_*\left(Z_i\right)\right]\| \lesssim_{\mathrm{P}} 1$. For this purpose, note that by the ξ_i's being i.i.d., zero-mean, unit variance and independent of the data we have

$$\mathrm{E}\left[\left\|\sqrt{n}\mathbb{E}_n\left[\xi_i s_*\left(Z_i\right)\right]\right\|^2 \middle| \left\{Z_i\right\}_1^n\right] = \mathbb{E}_n\left[\|s_*\left(Z_i\right)\|^2\right].$$

The desired $\|\sqrt{n}\mathbb{E}_n\left[\xi_i s_*\left(Z_i\right)\right]\| \lesssim_{\mathrm{P}} 1$ now follows from iterated expectations, Assumption 1.1 and M.

$\|\mathrm{III}_n\|_{\mathcal{T}}$.

By CS, Step 1.I.1 and

$$\|\mathrm{III}_n\|_{\mathcal{T}} \leqslant \left\|\sqrt{n}\mathbb{E}_n\left\{\xi_i[\widehat{s}\left(Z_i\right) - s_*\left(Z_i\right)]\right\}\right\| \sup_{t\in\mathcal{T}}\|\widehat{b}\left(t\right)\|,$$

it suffices to show that $\|\sqrt{n}\mathbb{E}_n\left\{\xi_i[\widehat{s}\left(Z_i\right) - s_*\left(Z_i\right)]\right\}\| \to_{\mathrm{P}} 0$. To this end, note that by the ξ_i's being i.i.d., zero-mean, unit variance and independent of the data, and $\widehat{s}$ being $\{Z_i\}_1^n$-measurable (Assumption 1.8), we have

$$\mathrm{E}\left[\|\sqrt{n}\mathbb{E}_n\left\{\xi_i[\widehat{s}\left(Z_i\right) - s_*\left(Z_i\right)]\right\}\| \,\middle|\, \left\{Z_i\right\}_1^n\right] = \mathbb{E}_n\left[\|\widehat{s}\left(Z_i\right) - s_*\left(Z_i\right)\|^2\right] = \|\widehat{s} - s_*\|_{\mathbb{P}_n,2}^2.$$

By Assumption 1.8, the right-hand side $\to_{\mathrm{P}} 0$, so Lemma 1.23 implies

$$\|\sqrt{n}\mathbb{E}_n\left\{\xi_i[\widehat{s}\left(Z_i\right) - s_*\left(Z_i\right)]\right\}\|^2 \xrightarrow{\mathrm{P}} 0$$

and therefore $\|\sqrt{n}\mathbb{E}_n\left\{\xi_i[\widehat{s}\left(Z_i\right) - s_*\left(Z_i\right)]\right\}\| \to_{\mathrm{P}} 0$. This finishes the proof of $\|\mathrm{III}_n\|_{\mathcal{T}} \to_{\mathrm{P}} 0$.

Recall that $t \mapsto \psi_k\left(t\right) = \mathrm{E}[p^k\left(W\right)\delta_*\left(t, W\right)]$, so by the LOIE

$$\psi_k\left(\cdot\right) = \mathrm{E}[p^k\left(W\right)\omega\left(\cdot, X\right)\partial_v\rho(Z, \beta_*, h_*\left(W\right))].$$

I estimate ψ_k by

$$t \mapsto \widehat{\psi}_k\left(t\right) \coloneqq \mathbb{E}_n[p^k\left(W_i\right)\omega\left(t, X_i\right)\partial_v\rho(Z_i, \widehat{\beta}, \widehat{h}\left(W_i\right))].$$

Note that this definition allows us to write $(t,w) \mapsto \widehat{\delta}(t,w)$ as

$$(t,w) \mapsto \widehat{\delta}(t,w) = p^{k_n}(w)^\top \widehat{Q}_{k_n}^{-} \widehat{\psi}_{k_n}(t).$$

$\widehat{\psi}_{k_n}$ and $\widehat{Q}_{k_n}^{-}$

In this step I show that

$$\begin{aligned} &\text{(a)}\ \sup_{t\in\mathcal{T}} \|\widehat{\psi}_{k_n}(t) - \psi_{k_n}(t)\| \lesssim_{\mathrm{P}} \Big[\zeta_{k_n}(\sqrt{k_n/n} + k_n^{-\alpha}) + \Big(\sum_{j=1}^{k_n} \|p_{jk_n}\|_{\mathcal{W}}^2\Big)^{1/2}/\sqrt{n}\Big] \to 0,\\ &\text{(b)}\ \sup_{t\in\mathcal{T}} \|\widehat{\psi}_{k_n}(t)^\top \widehat{Q}_{k_n}^{-} - \psi_{k_n}(t)^\top Q_{k_n}^{-1}\| \xrightarrow{\mathrm{P}} 0,\\ \text{and}\quad &\text{(c)}\ \sup_{t\in\mathcal{T}} \|\widehat{\psi}_{k_n}(t)^\top \widehat{Q}_{k_n}^{-}\| \lesssim_{\mathrm{P}} 1. \end{aligned}$$

To show (a), recall $\Delta(t,z,h)$ from (1.I.6)

$$\Delta(t,z,h) = \omega(t,x)\,\partial_v\rho(z,\beta_*,h_*(w))\,h(w) - \mathrm{E}_Z[\omega(t,X)\,\partial_v\rho(z,\beta_*,h_*(W))\,h(W)].$$

Letting $\Delta_i^k(t) \coloneqq (\Delta(t,Z_i,p_{1k}),\ldots,\Delta(t,Z_i,p_{kk}))^\top$, by T we have

$$\begin{aligned} \sup_{t\in\mathcal{T}} \|\widehat{\psi}_{k_n}(t) - \psi_{k_n}(t)\| \leqslant\, &\sup_{t\in\mathcal{T}} \|\mathbb{E}_n\{\omega(t,X_i)[\partial_v\rho(Z_i,\widehat{\beta},\widehat{h}(W_i)) - \partial_v\rho(Z_i,\beta_*,h_*(W_i))]p^{k_n}(W_i)\}\|\\ &+ \sup_{t\in\mathcal{T}} \|(\mathbb{E}_n - \mathrm{E})\,\Delta_i^{k_n}(t)\|. \end{aligned}$$

By Assumptions 1.1, 1.2 and 1.8 and T followed by CS

$$\begin{aligned} &\sup_{t\in\mathcal{T}} \|\mathbb{E}_n\{\omega(t,X_i)[\partial_v\rho(Z_i,\widehat{\beta},\widehat{h}(W_i)) - \partial_v\rho(Z_i,\beta_v,h_*(W_i))]p^{k_n}(W_i)\}\|\\ &\lesssim \mathbb{E}_n\{\|p^{k_n}(W_i)\|L_2(Z_i)[\|\widehat{\beta} - \beta_*\| + |\widehat{h}(W_i) - h_*(W_i)|]\}\\ &\lesssim_{\mathrm{P}} \zeta_{k_n}(\mathrm{E}[L_2(Z)^2]\})^{1/2}(n^{-1/2} + \|\widehat{h} - h_*\|_{\mathbb{P}_n,2})\\ &\lesssim_{\mathrm{P}} \zeta_{k_n}(n^{-1/2} + \|\widehat{h}_n - h_*\|_{\mathbb{P}_n,2}) \lesssim_{\mathrm{P}} \zeta_{k_n}(\sqrt{k_n/n} + k_n^{-\alpha})\\ &\leqslant \Big(\sum_{j=1}^{k_n} \|p_{jk_n}\|_{\mathcal{W}}^2\Big)^{1/2}(\sqrt{k_n/n} + k_n^{-\alpha}) \to 0, \end{aligned}$$

where the last $\lesssim_{\mathrm{P}}$ follows from Lemma 1.22 and the $\to 0$ from Assumption 1.7.

Moreover, the argument of Section 1.I.1 shows that

$$\sup_{t\in\mathcal{T}} \|\mathbb{E}_n\{\Delta_i^{k_n}(t)\}\| \lesssim_{\mathrm{P}} \Big(\sum_{j=1}^{k_n} \|p_{jk_n}\|_{\mathcal{W}}^2\Big)^{1/2}/\sqrt{n}.$$

Lemmas 1.17 and 1.21 and Assumptions 1.5 and 1.7 show that $\widehat{Q}_{k_n}$ is invertible wp → 1 and $\underline{\lambda}(\widehat{Q}_{k_n})^{-1} \lesssim_{\mathrm{P}} 1$. To ease notation I will (without loss of generality) assume that $\widehat{Q}_{k_n}^{-1}$ exists with probability one for all n, such that $\widehat{Q}_{k_n}^{-} = \widehat{Q}_{k_n}^{-1}$. The argument in Section 1.I.1 shows that $\sup_{\mathcal{T}}\|\psi_{k_n}(t)^\top Q_{k_n}^{-1}\| \lesssim 1$, so by (a) and T,

$$\begin{aligned}
&\sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}(t)^\top \widehat{Q}_{k_n}^{-1} - \psi_{k_n}(t)^\top Q_{k_n}^{-1}\| \\
&\leqslant \sup_{t\in\mathcal{T}}\|[\widehat{\psi}_{k_n}(t) - \psi_{k_n}(t)]^\top \widehat{Q}_{k_n}^{-1}\| + \sup_{t\in\mathcal{T}}\|\psi_{k_n}(t)^\top (\widehat{Q}_{k_n}^{-1} - Q_{k_n}^{-1})\| \\
&\leqslant \|\widehat{Q}_{k_n}^{-1}\|_{\mathrm{op}} \sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}(t) - \psi_{k_n}(t)\| + \sup_{t\in\mathcal{T}}\|\psi_{k_n}(t)^\top Q_{k_n}^{-1}(\widehat{Q}_{k_n} - Q_{k_n})\widehat{Q}_{k_n}^{-1}\| \\
&\leqslant \|\widehat{Q}_{k_n}^{-1}\|_{\mathrm{op}}\Big(\sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}(t) - \psi_{k_n}(t)\| + \|\widehat{Q}_{k_n} - Q_{k_n}\|_{\mathrm{op}} \sup_{t\in\mathcal{T}}\|\psi_{k_n}(t)^\top Q_{k_n}^{-1}\|\Big) \xrightarrow{\mathrm{P}} 0,
\end{aligned}$$

which shows (b). Part (c) follows from (b) and $\sup_{t\in\mathcal{T}}\|\psi_{k_n}(t)^\top Q_{k_n}^{-1}\| \lesssim 1$.

$\|\mathrm{IV}_n\|_{\mathcal{T}}$

Denoting $U_i = Y_i - h_*(W_i)$, by T we get

$$\begin{aligned}
\|\mathrm{IV}_n\|_{\mathcal{T}} &= \sup_{t\in\mathcal{T}}\big|\sqrt{n}\mathbb{E}_n\big(\xi_i\{\widehat{\delta}(t,W_i)[Y_i - \widehat{h}(W_i)] - \delta_*(t,W_i)U_i\}\big)\big| \\
&\leqslant \sup_{t\in\mathcal{T}}\big|\sqrt{n}\mathbb{E}_n\big\{\xi_i U_i[\widehat{\delta}(t,W_i) - \delta_*(t,W_i)]\big\}\big| \\
&\quad + \sup_{t\in\mathcal{T}}\big|\sqrt{n}\mathbb{E}_n\big\{\xi_i\widehat{\delta}(t,W_i)[\widehat{h}(W_i) - h_*(W_i)]\big\}\big| =: \|\mathrm{IV}_{a,n}\|_{\mathcal{T}} + \|\mathrm{IV}_{b,n}\|_{\mathcal{T}}.
\end{aligned}$$

$\|\mathrm{IV}_{a,n}\|_{\mathcal{T}}$

Recalling that $\delta_k(t,w) = p^k(w)^\top Q_k^{-1}\psi_k(t)$, by T

$$\begin{aligned}
\|\mathrm{IV}_{a,n}\|_{\mathcal{T}} = \sup_{t\in\mathcal{T}}\Big|\sqrt{n}\mathbb{E}_n\Big\{\xi_i U_i\Big[&p^{k_n}(W_i)^\top \widehat{Q}_{k_n}^{-1}\widehat{\psi}_{k_n}(t) \pm p^{k_n}(W_i)^\top Q_{k_n}^{-1}\widehat{\psi}_{k_n}(t) \\
&\pm \delta_{k_n}(t,W_i) - \delta_*(t,W_i)\Big]\Big\}\Big| \\
\leqslant \sup_{t\in\mathcal{T}}&\Big|\widehat{\psi}_{k_n}(t)^\top(\widehat{Q}_{k_n}^{-1} - Q_{k_n}^{-1})\sqrt{n}\mathbb{E}_n\left[p^{k_n}(W_i)\xi_i U_i\right]\Big| \\
+ \sup_{t\in\mathcal{T}}&\Big|[\widehat{\psi}_{k_n}(t) - \psi_{k_n}(t)]^\top Q_{k_n}^{-1}\sqrt{n}\mathbb{E}_n\left[p^{k_n}(W_i)\xi_i U_i\right]\Big| \\
+ \sup_{t\in\mathcal{T}}&\big|\sqrt{n}\mathbb{E}_n\left[\xi_i U_i\{\delta_{k_n}(t,W_i) - \delta_*(t,W_i)\}\right]\big| \\
=: \|\mathrm{IV}_{a,1,n}\|_{\mathcal{T}} &+ \|\mathrm{IV}_{a,2,n}\|_{\mathcal{T}} + \|\mathrm{IV}_{a,3,n}\|_{\mathcal{T}}.
\end{aligned}$$

$\|\mathrm{IV}_{a,1,n}\|_{\mathcal{T}}$ By the ξ_i's being i.i.d., zero-mean, unit variance and independent of the data,

$$\begin{aligned}\mathrm{E}\Big[\Big\|Q_{k_n}^{-1/2}\sqrt{n}\mathbb{E}_n\left[p^{k_n}\left(W_i\right)\xi_i U_i\right]\Big\|^2\Big] &= \mathrm{E}[\xi^2 U^2 p^{k_n}\left(W\right)^\top Q_{k_n}^{-1}p^{k_n}\left(W\right)]\\ &= \mathrm{E}[U^2 p^{k_n}\left(W\right)^\top Q_{k_n}^{-1}p^{k_n}\left(W\right)]\\ &\lesssim \mathrm{E}[p^{k_n}\left(W\right)^\top Q_{k_n}^{-1}p^{k_n}\left(W\right)] = k_n,\end{aligned}$$

so by M we have

$$\Big\|Q_{k_n}^{-1/2}\sqrt{n}\mathbb{E}_n\left[p^{k_n}\left(W_i\right)\xi_i U_i\right]\Big\| \lesssim_{\mathrm{P}} \sqrt{k_n}. \tag{1.I.9}$$

Given that $\sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}\left(t\right)^\top \widehat{Q}_{k_n}^{-1}\|$ (Section 1.I.1), by CS, the Min-Max theorem, Lemma 1.21, and the previous display,

$$\begin{aligned}\|\mathrm{IV}_{a,1,n}\|_{\mathcal{T}} &= \sup_{t\in\mathcal{T}}\Big|\widehat{\psi}_{k_n}\left(t\right)^\top \widehat{Q}_{k_n}^{-1}(Q_{k_n}-\widehat{Q}_{k_n})Q_{k_n}^{-1}\sqrt{n}\mathbb{E}_n\left[p^{k_n}\left(W_i\right)\xi_i U_i\right]\Big|\\ &\leqslant \Big\|Q_{k_n}^{-1}\sqrt{n}\mathbb{E}_n\left[p^{k_n}\left(W_i\right)\xi_i U_i\right]\Big\|\sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}\left(t\right)^\top \widehat{Q}_{k_n}^{-1}(Q_{k_n}-\widehat{Q}_{k_n})\|\\ &\leqslant \Big\|Q_{k_n}^{-1}\sqrt{n}\mathbb{E}_n\left[p^{k_n}\left(W_i\right)\xi_i U_i\right]\Big\|\|\widehat{Q}_{k_n}-Q_{k_n}\|_{\mathrm{op}}\sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}\left(t\right)^\top \widehat{Q}_{k_n}^{-1}\|\\ &\lesssim \Big\|Q_{k_n}^{-1/2}\sqrt{n}\mathbb{E}_n\left[p^{k_n}\left(W_i\right)\xi_i U_i\right]\Big\|\|\widehat{Q}_{k_n}-Q_{k_n}\|_{\mathrm{op}}\sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}\left(t\right)^\top \widehat{Q}_{k_n}^{-1}\|\\ &\lesssim_{\mathrm{P}} \sqrt{k_n}\cdot[\zeta_{k_n}^2\ln\left(k_n\right)/n]^{1/2} = [\zeta_{k_n}^2 k_n\ln\left(k_n\right)/n]^{1/2} \to 0.\end{aligned}$$

$\|\mathrm{IV}_{a,2,n}\|_{\mathcal{T}}$

By CS, the Min-Max theorem, (1.I.9), the results of Section 1.I.1, and Assumption 1.8,

$$\begin{aligned}\|\mathrm{IV}_{a,2,n}\|_{\mathcal{T}} &\leqslant \Big\|Q_{k_n}^{-1}\sqrt{n}\mathbb{E}_n\left[p^{k_n}\left(W_i\right)\xi_i U_i\right]\Big\|\sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}\left(t\right)-\psi_{k_n}\left(t\right)\|\\ &\lesssim \Big\|Q_{k_n}^{-1/2}\sqrt{n}\mathbb{E}_n\left[p^{k_n}\left(W_i\right)\xi_i U_i\right]\Big\|\sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}\left(t\right)-\psi_{k_n}\left(t\right)\|\\ &\lesssim_{\mathrm{P}} \sqrt{k_n}\Big[\zeta_{k_n}(\sqrt{k_n/n}+k_n^{-\alpha})+\Big(\sum_{j=1}^{k_n}\|p_{jk_n}\|_{\mathcal{W}}^2\Big)^{1/2}/\sqrt{n}\Big]\\ &= \zeta_{k_n}\sqrt{k_n}(\sqrt{k_n/n}+k_n^{-\alpha})+\Big(\sum_{j=1}^{k_n}\|p_{jk_n}\|_{\mathcal{W}}^2\Big)^{1/2}\sqrt{k_n/n}\to 0.\end{aligned}$$

$\|\mathrm{IV}_{a,3,n}\|_{\mathcal{T}}$

Fix k and let $\mathcal{F}_k' \coloneqq \{(s,z)\mapsto s\left[y-h_*\left(w\right)\right]\left[\delta_k\left(t,w\right)-\delta_*\left(t,w\right)\right]|\, t\in\mathcal{T}\}$. Given that each $\mathrm{E}[f(\xi,Z)]=0$ for each $f\in\mathcal{F}_k'$, the stochastic process IV_n may be viewed as an empirical process $\mathbb{G}_n$ indexed by the changing classes $\mathcal{F}'_{k_n}$. For $f = f_t, f_1 = f_{t_1}, f_2 = f_{t_2} \in \mathcal{F}'_{k_n}$ arbitrary, by the arguments of Section 1.I.1 there exists a function $z \mapsto F_k(z)$ such that

$$\begin{aligned} |f(s,z)| &\leqslant |s| F_k(z), \\ |f_1(s,z) - f_2(s,z)| &\leqslant |s| F_k(z) \|t_1 - t_2\|, \\ \text{and } \|F_k\|_{P,2} &\lesssim \sqrt{k} \quad (\text{as } k \to \infty). \end{aligned}$$

The ξ_i's being zero mean, unit variance and independent of the data implies that $F'_k : (s,z) \mapsto |s| F_k(z)$ is an envelope for $\mathcal{F}'_k$ with $(\mathrm{E}[F'_k(\xi, Z)^2])^{1/2} = \|F_k\|_{P,2} \lesssim \sqrt{k}$ as $k \to \infty$, satisfying

$$|f_1(s,z) - f_2(s,z)| \leqslant F'_k(s,z) \|t_1 - t_2\|.$$

Using $\mathcal{T}$ compact and the previous display, by van der Vaart and Wellner (1996, Theorem 2.7.11) we see that

$$N_{[\,]}(\varepsilon (\mathrm{E}[F'_k(\xi, Z)^2])^{1/2}, \mathcal{F}'_k, L^2(\xi, Z)) \leqslant (C/\varepsilon)^{d_t}, \quad \varepsilon \in (0,1].$$

and thus

$$J_{[\,]}\left(\delta, \mathcal{F}'_k, L^2(\xi, Z)\right) \leqslant \int_0^\delta \sqrt{1 + d_t \ln(C/\varepsilon)} \mathrm{d}\varepsilon, \quad \delta > 0.$$

where the right-hand side does not depend on k. In particular, $J_{[\,]}\left(1, \mathcal{F}'_{k_n}, L^2(\xi, Z)\right) \lesssim 1$. Defining

$$\sigma_n^2 \coloneqq \sup_{f \in \mathcal{F}'_{k_n}} \mathbb{E}_n[f(\xi_i, Z_i)^2]$$

we see that

$$\sigma_n^2 = \sup_{t \in \mathcal{T}} \mathbb{E}_n \{\xi_i^2 U_i^2 [\delta_{k_n}(t, W_i) - \delta_*(t, W_i)]^2\} \leqslant \mathbb{E}_n \{\xi_i^2 U_i^2 \|\delta_{k_n}(\cdot, W_i) - \delta_*(\cdot, W_i)\|_{\mathcal{T}}^2\}$$

such that

$$\mathrm{E}(\sigma_n^2) \leqslant \mathrm{E}\{\xi^2 U^2 \|\delta_{k_n}(\cdot, W) - \delta_*(\cdot, W)\|_{\mathcal{T}}^2\} \lesssim \mathrm{E}\{\|\delta_{k_n}(\cdot, W) - \delta_*(\cdot, W)\|_{\mathcal{T}}^2\} = R_{\delta, k_n}^2,$$

where the $\lesssim$ follows from the ξ_i's being zero mean, unit variance, and independent of the data and Assumption 1.4, and the last equality follow from the definitions of δ_k and $R_{\delta,k}$.

Consider the two cases: (1) $R_{\delta,k_n} / \|F_{k_n}\|_{P,2} \to 0$ and (2) $R_{\delta,k_n} / \|F_{k_n}\|_{P,2} \nrightarrow 0$.

Case 1: $R_{\delta,k_n}/\|F_{k_n}\|_{P,2} \to 0$. Given that $\sqrt{\mathrm{E}\left(\sigma_n^2\right)} \leqslant C_1 R_{\delta,k_n}$, by the change of variables $\varepsilon' := \varepsilon/C_1$ we have

$$\begin{aligned} J_{[\,]}\left(\sqrt{\mathrm{E}\left(\sigma_n^2\right)}/\|F_{k_n}\|_{P,2}, \mathcal{F}'_k, L^2\left(\xi, Z\right)\right) &\leqslant J_{[\,]}\left(C_1 R_{\delta,k_n}/\|F_{k_n}\|_{P,2}, \mathcal{F}'_k, L^2\left(\xi, Z\right)\right) \\ &= C_1 \int_0^{R_{\delta,k_n}/\|F_{k_n}\|_{P,2}} \sqrt{1 + d_t \ln\left(C_3/\varepsilon'\right)}\mathrm{d}\varepsilon' \\ &=: C_1 \overline{J}_{[\,]}\left(R_{\delta,k_n}/\|F_{k_n}\|_{P,2}\right) \end{aligned} \tag{1.I.10}$$

By van der Vaart and Wellner (2011, p. 196) we have the maximal inequality

$$\begin{aligned} \mathrm{E}(\|\mathbb{G}_n\|_{\mathcal{F}'_{k_n}}) &\lesssim J_{[\,]}\left(\sqrt{\mathrm{E}\left(\sigma_n^2\right)}/\|F_{k_n}\|_{P,2}, \mathcal{F}'_{k_n}, L^2\left(\xi, Z\right)\right)\|F_{k_n}\|_{P,2} \\ &\lesssim \overline{J}_{[\,]}\left(R_{\delta,k_n}/\|F_{k_n}\|_{P,2}\right)\|F_{k_n}\|_{P,2}, \end{aligned}$$

and from van der Vaart and Wellner (1996, p. 239) we know that an entropy integral (bound) of the form (1.I.10) satisfies $\overline{J}_{[\,]}\left(\delta\right) \lesssim \delta\sqrt{\ln(1/\delta)}$ as $\delta \downarrow 0$. Since $R_{\delta,k_n}/\|F_{k_n}\|_{P,2} \to 0$ holds by hypothesis, the previous display combined with $\|F_{k_n}\|_{P,2} \lesssim \sqrt{k_n}$ yields

$$\begin{aligned} \mathrm{E}(\|\mathbb{G}_n\|_{\mathcal{F}'_{k_n}}) &\lesssim \left(R_{\delta,k_n}/\|F_{k_n}\|_{P,2}\right)\sqrt{\ln\left(\|F_{k_n}\|_{P,2}/R_{\delta,k_n}\right)}\|F_{k_n}\|_{P,2} = R_{\delta,k_n}\sqrt{\ln\left(\|F_{k_n}\|_{P,2}/R_{\delta,k_n}\right)} \\ &\lesssim \Delta_{k_n}\sqrt{\ln\left(k_n/R_{\delta,k_n}\right)}. \end{aligned}$$

Case 2. Suppose that $R_{\delta,k_n}/\|F_{k_n}\|_{P,2} \nrightarrow 0$. Given that $R_{\delta,k_n} \to 0$ (Assumption 1.7), we must have $\|F_{k_n}\|_{P,2} \lesssim R_k$. van der Vaart and Wellner (1996, Theorem 2.14.2) and $J_{[\,]}\left(1, \mathcal{F}'_{k_n}, L^2\left(\xi, Z\right)\right) \lesssim 1$ yield

$$\mathrm{E}(\|\mathbb{G}_n\|_{\mathcal{F}_{k'_n}}) \lesssim J_{[\,]}\left(1, \mathcal{F}_{k'_n}, L^2\left(\xi, Z\right)\right)\|F_{k_n}\|_{P,2} \lesssim \|F_{k_n}\|_{P,2} \lesssim R_{\delta,k_n} \lesssim R_{\delta,k_n}\sqrt{\ln\left(k_n/R_{\delta,k_n}\right)}$$

as in Case 1. The claim $\|\mathrm{IV}_{a,3,n}\|_{\mathcal{T}} \to_{\mathrm{P}} 0$ now follows from M and $R_{\delta,k_n}\sqrt{\ln\left(k_n/R_{\delta,k_n}\right)} \to 0$ (Assumption 1.7).

$\|\mathrm{IV}_{b,n}\|_{\mathcal{T}}$

Given that $\sup_{\mathcal{T}}\|\widehat{\psi}_{k_n}\left(t\right)^\top \widehat{Q}_{k_n}^{-1}\|$ (Section 1.I.1), by CS it follows that

$$\begin{aligned} \|\mathrm{IV}_{b,n}\|_{\mathcal{T}} &= \sup_{t\in\mathcal{T}}\left|\widehat{\psi}_{k_n}\left(t\right)^\top \widehat{Q}_{k_n}^{-1}\sqrt{n}\mathbb{E}_n\left[p^{k_n}\left(W_i\right)\xi_i[\widehat{h}\left(W_i\right) - h_*\left(W_i\right)]\right]\right| \\ &\leqslant \left\|\sqrt{n}\mathbb{E}_n\left[p^{k_n}\left(W_i\right)\xi_i[\widehat{h}\left(W_i\right) - h_*\left(W_i\right)]\right]\right\|\sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}\left(t\right)^\top \widehat{Q}_{k_n}^{-1}\| \\ &\lesssim_{\mathrm{P}} \left\|\sqrt{n}\mathbb{E}_n\left[p^{k_n}\left(W_i\right)\xi_i[\widehat{h}\left(W_i\right) - h_*\left(W_i\right)]\right]\right\|. \end{aligned}$$

To show that the right-hand side $\to_{\mathrm{P}} 0$, note that by the ξ_i's being i.i.d., zero-mean, unit variance and independent of $\{Z_i\}_1^n$, and $\widehat{h}_n$ being $\{Z_i\}_1^n$-measurable,

$$\begin{aligned}
&\mathrm{E}\Big[\Big\|\sqrt{n}\mathbb{E}_n\Big[p^{k_n}(W_i)\,\xi_i[\widehat{h}(W_i)-h_*(W_i)]\Big]\Big\|^2\Big|\{Z_i\}_1^n\Big]\\
&=\mathbb{E}_n\big\{\|p^{k_n}(W_i)\|^2[\widehat{h}(W_i)-h_*(W_i)]^2\big\}\leqslant\Big(\sum_{j=1}^{k_n}\|p_{jk_n}\|_{\mathcal{W}}^2\Big)\|\widehat{h}-h_*\|_{\mathbb{P}_n,2}^2\\
&\lesssim_{\mathrm{P}}\Big[\Big(\sum_{j=1}^{k_n}\|p_{jk_n}\|_{\mathcal{W}}^2\Big)^{1/2}(\sqrt{k_n/n}+k_n^{-\alpha})\Big]^2\to 0,
\end{aligned}$$

where the $\lesssim_{\mathrm{P}}$ follows from Lemma 1.22 and the $\to 0$ from Assumption 1.7. It follows by conditional CS that $\mathrm{E}(\|\sqrt{n}\mathbb{E}_n\{p^{k_n}(W_i)\,\xi_i[\widehat{h}(W_i)-h_*(W_i)]\}\||\{Z_i\}_1^n)\to_{\mathrm{P}} 0$, so $\|\sqrt{n}\mathbb{E}_n\{p^{k_n}(W_i)\,\xi_i$ $[\widehat{h}(W_i)-h_*(W_i)]\}\|\to_{\mathrm{P}} 0$ by Lemma 1.23. This $\to_{\mathrm{P}} 0$ finishes the proof of $\|\mathrm{IV}_{b,n}\|_{\mathcal{T}}\to_{\mathrm{P}} 0$, and therefore the proof of $\|\mathrm{IV}_n\|_{\mathcal{T}}\to_{\mathrm{P}} 0$. □

Lemma 1.16. *If Assumptions 1.1–1.8 hold, then*

$$\|\mathbb{E}_n[\widehat{f}(\cdot,Z_i)-f_*(\cdot,Z_i)]\|_{\mathcal{T}}\xrightarrow{\mathrm{P}}0.$$

Proof of Lemma 1.16. The proof proceeds in steps.

Main

For fixed $t\in\mathcal{T}$ we may write

$$\begin{aligned}
\mathbb{E}_n[\widehat{f}(t,Z_i)-f_*(t,Z_i)]&=\mathbb{E}_n\big\{\omega(t,X_i)\,[\rho(Z_i,\widehat{\beta},\widehat{h}(W_i))-\rho(Z_i,\beta_*,h_*(W_i))]\big\}\\
&\quad-[\widehat{b}(t)-b_*(t)^\top]^\top\mathbb{E}_n[s_*(Z_i)]-\widehat{b}(t)^\top\mathbb{E}_n[\widehat{s}(Z_i)-s_*(Z_i)]\\
&\quad+\mathbb{E}_n\{\widehat{\delta}(t,W_i)\,[Y_i-\widehat{h}(W_i)]-\delta_*(t,W_i)\,U_i\}\\
&=:\mathrm{I}_n(t)+\mathrm{II}_n(t)+\mathrm{III}_n(t)+\mathrm{IV}_n(t).
\end{aligned}$$

The following steps show that the four remainder terms $\to_{\mathrm{P}} 0$ uniformly over $\mathcal{T}$. The claim therefore follows from T.

$\|\mathrm{I}_n\|_{\mathcal{T}}$

Assumption 1.1 and M implies that $\|\widehat{\beta}-\beta_*\|\lesssim_{\mathrm{P}} n^{-1/2}\to 0$. Let $\mathcal{N}_*$ be the open neighborhood provided by Assumption 1.3. Then $\widehat{\beta}\in\mathcal{N}_*$ wp $\to 1$. To simplify notation and ensure that objects are globally well defined, in what follows I will—without loss of generality—assume that $\widehat{\beta} \in \mathcal{N}_*$ with probability equal to one for all n. A mean value expansion of $\beta \mapsto \rho(Z_i, \beta, \widehat{h}(W_i))$ at $\widehat{\beta}$ around β_* and CS show that

$$\begin{aligned}\|\mathrm{I}_n\|_{\mathcal{T}} \leqslant & \sup_{t\in\mathcal{T}}|\mathbb{E}_n\{\omega(t, X_i)[\rho(Z_i, \beta_*, \widehat{h}(W_i)) - \rho(Z_i, \beta_*, h_*(W_i))]\}| \\ & + \|\widehat{\beta} - \beta_*\| \sup_{t\in\mathcal{T}}\|\mathbb{E}_n[\omega(t, X_i)\, \partial_\beta \rho < (Z_i, \overline{\beta}, \widehat{h}(W_i))]\| =: \|\mathrm{I}_{a,n}\|_{\mathcal{T}} + \|\widehat{\beta} - \beta_*\| \|\mathrm{I}_{b,n}\|_{\mathcal{T}},\end{aligned}$$

where $\overline{\beta}$ satisfies $\|\overline{\beta} - \beta_*\| \leqslant \|\widehat{\beta} - \beta_*\|$ such that $\widehat{\beta} \in \mathcal{N}_*$ for n sufficiently large. Since $\|\widehat{\beta} - \beta_*\| \to_{\mathrm{P}} 0$ it suffices to show that $\|\mathrm{I}_{a,n}\|_{\mathcal{T}} \to_{\mathrm{P}} 0$ and $\|\mathrm{I}_{b,n}\|_{\mathcal{T}} \lesssim_{\mathrm{P}} 1$. The arguments of Section 1.I.1 show that

$$\begin{aligned}&\sup_{t\in\mathcal{T}}\|\mathrm{I}_{b,n}(t) - \mathrm{E}_Z[\omega(t, X)\, \partial_\beta \rho(Z, \beta_*, h_*(W))]\| \xrightarrow{\mathrm{P}} 0, \\ &\qquad \text{and} \sup_{t\in\mathcal{T}}\|\mathrm{E}_Z[\omega(t, X)\, \partial_\beta \rho(Z, \beta_*, h_*(W))]\| < \infty.\end{aligned}$$

which together imply $\|\mathrm{I}_{b,n}\|_{\mathcal{T}} \lesssim_{\mathrm{P}} 1$.

$\|\mathrm{I}_{a,n}\|_{\mathcal{T}}$ Abbreviate $(z, v) \mapsto \rho(z, \beta_*, v)$ by ρ_*. By a mean value expansion of $v \mapsto \rho_*(Z_i, v)$ at $\widehat{h}(W_i)$ around $h_*(W_i)$ and T we may be bound $\|\mathrm{I}_{a,n}\|_{\mathcal{T}}$ by

$$\begin{aligned}&\sup_{t\in\mathcal{T}}|\mathbb{E}_n\{\omega(t, X_i)[\partial_v \rho_*(Z_i, \overline{h}(W_i)) - \partial_v \rho_*(Z_i, h_*(W_i))][\widehat{h}(W_i) - h_*(W_i)]\}| \\ &\quad + \sup_{t\in\mathcal{T}}|\mathbb{E}_n\{\omega(t, X_i)\, \partial_v \rho_*(Z_i, h_*(W_i))[\widehat{h}(W_i) - h_*(W_i)]\}| =: \|\mathrm{I}_{a,1,n}\|_{\mathcal{T}} + \|\mathrm{I}_{a,2,n}\|_{\mathcal{T}},\end{aligned}$$

where $|\overline{h}(W_i) - h_*(W_i)| \leqslant |\widehat{h}(W_i) - h_*(W_i)|$. By T and Assumptions 1.2 and 1.3

$$\|\mathrm{I}_{a,1,n}\|_{\mathcal{T}} \lesssim \mathbb{E}_n\{L_1(Z_i)[\widehat{h}(W_i) - h_*(W_i)]^2\} \leqslant \mathbb{E}_n[L_1(Z_i)]\, \|\widehat{h} - h_*\|_{\mathcal{W}}^2 \lesssim_{\mathrm{P}} \|\widehat{h} - h_*\|_{\mathcal{W}}^2 \xrightarrow{\mathrm{P}} 0,$$

where the $\lesssim_{\mathrm{P}}$ follows from M and the $\to_{\mathrm{P}} 0$ from Lemma 1.22. Similarly,

$$\begin{aligned}\|\mathrm{I}_{a,2,n}\|_{\mathcal{T}} &\lesssim \mathbb{E}_n\{|\partial_v \rho_*(Z_i, h_*(W_i))||\widehat{h}(W_i) - h_*(W_i)|\} \\ &\leqslant \mathbb{E}_n\{|\partial_v \rho_*(Z_i, h_*(W_i))|\}\|\widehat{h} - h_*\|_{\mathcal{W}} \xrightarrow{\mathrm{P}} 0.\end{aligned}$$

$\|\mathrm{II}_n\|_{\mathcal{T}}$ Section 1.I.1 shows that $\sup_{t\in\mathcal{T}}\|\widehat{b}(t) - b_*(t)\| \to_{\mathrm{P}} 0$, so by CS, Assumption 1.1, and M

$$\|\mathrm{II}_n\|_{\mathcal{T}} \leqslant \|\mathbb{E}_n[s_*(Z_i)]\| \sup_{t\in\mathcal{T}}\|\widehat{b}(t) - b_*(t)\| \xrightarrow{\mathrm{P}} 0.$$

Section 1.I.1 also shows that $\sup_{t\in\mathcal{T}}\|\widehat{b}\,(t)\| \lesssim_{\mathrm{P}} 1$, so by CS and Assumption 1.8,

$$\|\mathrm{III}_n\|_{\mathcal{T}} \leqslant \|\mathbb{E}_n[\widehat{s}\,(Z_i) - s_*\,(Z_i)]\| \sup_{t\in\mathcal{T}}\|\widehat{b}\,(t)\| \leqslant \|\widehat{s} - s_*\|_{\mathbb{P}_n,2} \sup_{t\in\mathcal{T}}\|\widehat{b}\,(t)\| \xrightarrow{\mathrm{P}} 0.$$

$\|\mathrm{IV}_n\|_{\mathcal{T}}$

For fixed $t \in \mathcal{T}$, adding and subtracting $p^{k_n}\,(W_i)^\top\, Q_{k_n}^{-1}\widehat{\psi}_{k_n}\,(t)\,U_i$ and recalling that $\delta_k\,(t,w) = p^k\,(w)^\top\, Q_k^{-1}\psi_k\,(t)$ we may write

$$\begin{aligned}
\mathrm{IV}_n\,(t) &= \mathbb{E}_n\{U_i[\widehat{\delta}\,(t,W_i) - \delta_*\,(t,W_i)]\} - \mathbb{E}_n\{\widehat{\delta}\,(t,W_i)\,[\widehat{h}\,(W_i) - h_*\,(W_i)]\} \\
&= \widehat{\psi}_{k_n}\,(t)^\top\,(\widehat{Q}_{k_n}^{-1} - Q_{k_n}^{-1})\mathbb{E}_n[p^{k_n}\,(W_i)\,U_i] + [\widehat{\psi}_{k_n}\,(t) - \psi_{k_n}\,(t)]^\top Q_{k_n}^{-1}\mathbb{E}_n[p^{k_n}\,(W_i)\,U_i] \\
&\quad + \mathbb{E}_n\{U_i[\delta_{k_n}\,(t,W_i) - \delta_*\,(t,W_i)]\} - \widehat{\psi}_{k_n}\,(t)^\top\,\widehat{Q}_{k_n}^{-1}\mathbb{E}_n\{p^{k_n}\,(W_i)\,[\widehat{h}\,(W_i) - h_*\,(W_i)]\} \\
&=: \mathrm{IV}_{a,n}\,(t) + \mathrm{IV}_{b,n}\,(t) + \mathrm{IV}_{c,n}\,(t) + \mathrm{IV}_{d,n}\,(t)\,.
\end{aligned}$$

The desired $\|\mathrm{IV}_n\|_{\mathcal{T}} \to_{\mathrm{P}} 0$ will follow by T if we can show that the four remainder terms $\to_{\mathrm{P}} 0$. To this end, note that by the Min-Max theorem,

$$\begin{aligned}
\mathrm{E}(\|Q_k^{-1}\mathbb{E}_n[p^k\,(W_i)\,U_i]\|^2) \lesssim \mathrm{E}(\|Q_k^{-1/2}\mathbb{E}_n[p^k\,(W_i)\,U_i]\|^2) &= \mathrm{E}[U^2 p^k\,(W)^\top\, Q_k^{-1} p^k\,(W)]/n \\
&\lesssim \mathrm{E}[p^k\,(W)^\top\, Q_k^{-1} p^k\,(W)] = k/n,
\end{aligned}$$

so by CS, M we have

$$\|Q_{k_n}^{-1}\mathbb{E}_n[p^{k_n}\,(W_i)\,U_i]\| \lesssim_{\mathrm{P}} \sqrt{k/n} \to 0.$$

Section 1.I.1 shows that $\sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}\,(t)^\top\,\widehat{Q}_{k_n}^{-1}\| \lesssim_{\mathrm{P}} 1$. Moreover, Lemma 1.21 show that $\|\widehat{Q}_{k_n} - Q_{k_n}\|_{\mathrm{op}} \lesssim_{\mathrm{P}} [\zeta_{k_n}^2 \ln(k_n)/n] \to 0$, so by the previous display and CS,

$$\begin{aligned}
\|\mathrm{IV}_{a,n}\|_{\mathcal{T}} &= \|\widehat{\psi}_{k_n}\,(t)^\top\,\widehat{Q}_{k_n}^{-1}(Q_{k_n} - \widehat{Q}_{k_n})Q_{k_n}^{-1}\mathbb{E}_n[p^{k_n}\,(W_i)\,U_i]\|_{\mathcal{T}} \\
&\leqslant \|Q_{k_n}^{-1}\mathbb{E}_n[p^{k_n}\,(W_i)\,U_i]\|\|\widehat{Q}_{k_n} - Q_{k_n}\|_{\mathrm{op}} \sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}\,(t)^\top\,\widehat{Q}_{k_n}^{-1}\| \xrightarrow{\mathrm{P}} 0.
\end{aligned}$$

Section 1.I.1 also shows that

$$\sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}\,(t) - \psi_{k_n}\,(t)\| \lesssim_{\mathrm{P}} \zeta_{k_n}\left(\sqrt{k_n/n} + k_n^{-\alpha}\right) + \left(\sum_{j=1}^{k_n}\|p_{jk_n}\|_{\mathcal{W}}^2\right)^{1/2}/\sqrt{n} \to 0,$$

so by CS

$$\|\mathrm{IV}_{b,n}\|_{\mathcal{T}} \leqslant \|Q_{k_n}^{-1}\mathbb{E}_n[p^{k_n}(W_i)\,U_i]\| \sup_{T}\|\widehat{\psi}_{k_n}(t) - \psi_{k_n}(t)\| \xrightarrow{\mathrm{P}} 0.$$

Section 1.I.1 shows that

$$\|\mathrm{IV}_{c,n}\|_{t\in\mathcal{T}} = \sup_{t\in\mathcal{T}}|\mathbb{E}_n\{U_i[\delta_{k_n}(t,W_i) - \delta_*(t,W_i)]\}| \lesssim_{\mathrm{P}} R_{\delta,k_n}\sqrt{\ln(k_n/R_{\delta,k_n})} \to 0.$$

Lastly, by CS, Lemma 1.22 and $\sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}(t)^{\top}\widehat{Q}_{k_n}^{-1}\| \lesssim_{\mathrm{P}} 1$ we get

$$\begin{aligned}
\|\mathrm{IV}_{d,n}\|_{\mathcal{T}} &\leqslant \|\mathbb{E}_n\{p^{k_n}(W_i)\,[\widehat{h}(W_i) - h_*(W_i)]\}\| \sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}(t)^{\top}\widehat{Q}_{k_n}^{-1}\| \\
&\leqslant \|\mathbb{E}_n\{p^{k_n}(W_i)\,[\widehat{h}(W_i) - h_*(W_i)]\}\| \sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}(t)^{\top}\widehat{Q}_{k_n}^{-1}\| \\
&\leqslant \Big(\sum_{j=1}^{k_n}\|p_{jk_n}\|_{\mathcal{W}}^2\Big)^{1/2}\|\widehat{h} - h_*\|_{\mathbb{P}_n,2} \sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}(t)^{\top}\widehat{Q}_{k_n}^{-1}\| \\
&\lesssim_{\mathrm{P}} \Big(\sum_{j=1}^{k_n}\|p_{jk_n}\|_{\mathcal{W}}^2\Big)^{1/2}\Big(\sqrt{k_n/n} + k_n^{-\alpha}\Big) \to 0.
\end{aligned}$$

This finishes the proof of $\|\mathbb{E}_n[\widehat{f}(\cdot,Z_i) - f_*(\cdot,Z_i)]\|_{\mathcal{T}} \to_{\mathrm{P}} 0$. □

Proof of Theorem 1.3. By a rearrangement and T

$$\begin{aligned}
\|\widehat{G} - G_n^*\|_{\mathcal{T}} &= \Big\|\sqrt{n}\mathbb{E}_n\Big[\big(\xi_i - \overline{\xi}\big)\,\widehat{f}(\cdot,Z_i)\Big] - \sqrt{n}\mathbb{E}_n\left[\big(\xi_i - \overline{\xi}\big)\,f_*(\cdot,Z_i)\right]\Big\|_{\mathcal{T}} \\
&= \Big\|\sqrt{n}\mathbb{E}_n\Big(\xi_i\{\widehat{f}(\cdot,Z_i) - \mathbb{E}_n[\widehat{f}(\cdot,Z_i)]\}\Big) - \sqrt{n}\mathbb{E}_n\left(\xi_i\{f_*(\cdot,Z_i) - \mathbb{E}_n\left[f_*(\cdot,Z_i)\right]\}\right)\Big\|_{\mathcal{T}} \\
&\leqslant \Big\|\sqrt{n}\mathbb{E}_n\Big\{\xi_i[\widehat{f}(\cdot,Z_i) - f_*(\cdot,Z_i)]\Big\}\Big\|_{\mathcal{T}} + |\sqrt{n}\overline{\xi}|\|\mathbb{E}_n[\widehat{f}(\cdot,Z_i) - f_*(\cdot,Z_i)]\|_{\mathcal{T}} \\
&= \|\widehat{G}^u - G_n^{*u}\|_{\mathcal{T}} + |\sqrt{n}\overline{\xi}|\|\mathbb{E}_n[\widehat{f}(\cdot,Z_i) - f_*(\cdot,Z_i)]\|_{\mathcal{T}}.
\end{aligned}$$

The first term on the right $\to_{\mathrm{P}} 0$ by Lemma 1.15. Given that $\sqrt{n}\overline{\xi} \sim \mathrm{N}(0,1), |\sqrt{n}\overline{\xi}| \lesssim_{\mathrm{P}} 1$. The second term therefore goes to zero by Lemma 1.16. □

Proof of Corollary 1.2. Given that $\mathcal{F}$ is Donsker (Lemma 1.2), Kosorok (2008, Theorem 10.4(iv)) implies that $\mathbb{G}_n'' \rightsquigarrow_{\mathrm{P},\xi} \mathbb{G}$ in $\ell^{\infty}(\mathcal{F})$, where $\mathbb{G}_n''(f) \coloneqq n^{-1/2}\sum_{i=1}^{n}\xi_i\left\{f(Z_i) - \mathrm{E}\left[f(Z)\right]\right\}$ and $\mathbb{G}_*$ is a zero-mean Gaussian process with covariance kernel $\mathrm{E}\left[\mathbb{G}(f_1)\,\mathbb{G}(f_2)\right] = \mathrm{E}\left[f_1(Z)\,f_2(Z)\right] - \mathrm{E}\left[f_1(Z)\right]\mathrm{E}\left[f_2(Z)\right]$. Since we may identify $\mathcal{F}$ with $\mathcal{T}$ through $f(\cdot) = f_*(t,\cdot)$, this result is equivalent to $G_n^* \rightsquigarrow_{\mathrm{P},\xi} G_*$ in $\ell^{\infty}(\mathcal{T})$, where G_* is a zero-mean Gaussian process with covariance kernel $\mathrm{E}\left[G_*(t_1)\,G_*(t_2)\right] = \mathrm{E}\left[f_*(t_1,Z)\,f_*(t_2,Z)\right] - \mathrm{E}\left[f_*(t_1,Z)\right]\mathrm{E}\left[f_*(t_2,Z)\right]$. The previous display and Lemma 1.24 now show that $\widehat{G}_n \rightsquigarrow_{\mathrm{P},\xi} G_*$ in $\ell^\infty(\mathcal{T})$.

Assumptions 1.1–1.4 imply that σ_*^2 is continuous, and Assumption 1.9 shows that σ_*^2 is nondegenerate. Lemma 1.25 now shows that the cdf F_* of the random variable $\|G_*\|_{\mu,2}^2$ is everywhere continuous and strictly increasing on $[0,\infty)$. Hence F_* is invertible on $(0,\infty)$ with inverse $(1-\alpha) \mapsto (F_*)^{-1}(1-\alpha) = c_*(\alpha)$, and each $(1-\alpha)$-quantile $c_*(\alpha) \in (0,\infty)$ for $\alpha \in (0,1)$. The convergence $\widehat{G} \rightsquigarrow_{\mathrm{P},\xi} G_*$ in $\ell^\infty(\mathcal{T})$ and Kosorok (2008, Lemma 10.11) imply that the cdf $\widehat{F}$ of $\|\widehat{G}\|_{\mu,2}^2$ converges in probability to F_* pointwise on $[0,\infty)$. Fix $\varepsilon > 0$ and $\alpha \in (0,1)$. Let $r_1 \in \mathbf{R}$ be such that $c_*(\alpha) - \varepsilon < r_1 < c_*(\alpha)$ and $F_*(r_1) < 1-\alpha$. Then $\widehat{F}(r_1) < 1-\alpha$ wp $\to 1$, which implies $c_*(\alpha) - \varepsilon < r_1 \leqslant \widehat{c}(\alpha)$ wp $\to 1$. In particular, $\mathrm{P}(\widehat{c}(\alpha) \geqslant c_*(\alpha) - \varepsilon) \to 1$. Let $r_2 \in \mathbf{R}$ be such that $c_*(\alpha) < r_2 < c_*(\alpha) + \varepsilon$ and $1-\alpha < F_*(r_2)$. Then $1-\alpha < \widehat{F}(r_2)$ wp $\to 1$, which implies $\widehat{c}(\alpha) \leqslant r_2 < c_*(\alpha) + \varepsilon$ wp $\to 1$. In particular, $\mathrm{P}(\widehat{c}(\alpha) < c_*(\alpha) + \varepsilon) \to 1$. It follows that

$$\varlimsup_{n\to\infty} \mathrm{P}(|\widehat{c}(\alpha) - c_*(\alpha)| \geqslant \varepsilon) \leqslant \varlimsup_{n\to\infty} \mathrm{P}(\widehat{c}(\alpha) \geqslant c_*(\alpha) + \varepsilon) + \varlimsup_{n\to\infty} \mathrm{P}(\widehat{c}(\alpha) \leqslant c_*(\alpha) - \varepsilon) = 0.$$

Since $\varepsilon > 0$ was arbitrary, the corollary follows.

1.I.3 Proofs for Section 1.4.6

[Proof of Theorem 1.3] Fix $\alpha \in (0,1)$. By Theorem 1.1, under the null, $T_n \to_d \|G_0\|_{\mu,2}^2$. Let F_0 denote the cdf of $\|G_0\|_{\mu,2}^2$. Given Assumption 1.9, F_0 is continuous on $\mathbf{R}$ and strictly increasing on $[0,\infty)$ (cf. Lemma 1.25). By Theorem 1.2 $\widehat{c}(\alpha) \to_{\mathrm{P}} c_*(\alpha) \in (0,\infty)$, and under the null, $c_*(\alpha) = c_0(\alpha)$, the $(1-\alpha)$-quantile of $\|G_0\|_{\mathcal{T}}$. Fix $\varepsilon > 0$. Then

$$\mathrm{P}(T_n > \widehat{c}(\alpha) \cap \widehat{c}(\alpha) < c_0(\alpha) - \varepsilon; \mathrm{H}_0) \leqslant \mathrm{P}(|\widehat{c}(\alpha) - c_0(\alpha)| > \varepsilon; \mathrm{H}_0) \to 0.$$

It follows by the portmanteau theorem and $[c_0(\alpha) - \varepsilon, \infty)$ closed

$$\begin{aligned}
\varlimsup_{n\to\infty} \mathrm{P}(T_n > \widehat{c}(\alpha); \mathrm{H}_0) &= \varlimsup_{n\to\infty} \mathrm{P}(T_n > \widehat{c}(\alpha) \cap \widehat{c}(\alpha) \geqslant c_0(\alpha) - \varepsilon; \mathrm{H}_0) \\
&\leqslant \varlimsup_{n\to\infty} \mathrm{P}(T_n \geqslant c_0(\alpha) - \varepsilon; \mathrm{H}_0) = \varlimsup_{n\to\infty} \mathrm{P}(T_n \in [c_0(\alpha) - \varepsilon, \infty); \mathrm{H}_0) \\
&\leqslant \mathrm{P}\left(\|G_0\|_{\mathcal{T}} \in [c_0(\alpha) - \varepsilon, \infty)\right).
\end{aligned}$$

Using continuity of F_0, the right-hand side equals $1 - F_0(c_0(\alpha) - \varepsilon)$, so letting $\varepsilon \downarrow 0$ and again using continuity of F_0, we see that $\varlimsup_{n\to\infty} \mathrm{P}(T_n > \widehat{c}(\alpha); \mathrm{H}_0) \leqslant \alpha$. For the other direction, again fix $\varepsilon > 0$. Then

$$\mathrm{P}(T_n > \widehat{c}(\alpha) \cap \widehat{c}(\alpha) > c_0(\alpha) + \varepsilon; \mathrm{H}_0) \leqslant \mathrm{P}(|\widehat{c}(\alpha) - c_0(\alpha)| > \varepsilon; \mathrm{H}_0) \to 0,$$

so by the portmanteau theorem and now $(c_0(\alpha)+\varepsilon,\infty)$ open,

$$\begin{aligned}
\varliminf_{n\to\infty} \mathrm{P}(T_n > \widehat{c}(\alpha);\mathrm{H}_0) &= \varliminf_{n\to\infty} \mathrm{P}(T_n > \widehat{c}(\alpha) \cap \widehat{c}(\alpha) \leqslant c_0(\alpha)+\varepsilon;\mathrm{H}_0) \\
&\geqslant \varliminf_{n\to\infty} \mathrm{P}(T_n > c_0(\alpha)+\varepsilon;\mathrm{H}_0) = \varliminf_{n\to\infty} \mathrm{P}(T_n \in (c_0(\alpha)+\varepsilon,\infty);\mathrm{H}_0) \\
&\geqslant \mathrm{P}\left(\|G_0\|_{\mu,2}^2 \in (c_0(\alpha)+\varepsilon,\infty)\right).
\end{aligned}$$

The right-hand side equals $1-F_0(c_0(\alpha)+\varepsilon)$, so letting $\varepsilon \downarrow 0$, we see that $\varliminf_{n\to\infty}\mathrm{P}(T_n > \widehat{c}(\alpha);\mathrm{H}_0) \geqslant \alpha$. □

Proof of Lemma 1.4. By T and Lemma 1.1

$$\begin{aligned}
\left|T_n^{1/2} - \|\sqrt{n}\mathbb{E}_n[f_*(\cdot,Z_i)]\|_{\mu,2}\right| &\leqslant \|\sqrt{n}\mathbb{E}_n[\rho(Z_i,\widehat{\beta},\widehat{h}(W_i)\,\omega(\cdot,X_i))] - \sqrt{n}\mathbb{E}_n[f_*(\cdot,Z_i)]\|_{\mu,2} \\
&\leqslant \|\sqrt{n}\mathbb{E}_n[\rho(Z_i,\widehat{\beta},\widehat{h}(W_i)\,\omega(\cdot,X_i))] - \sqrt{n}\mathbb{E}_n[f_*(\cdot,Z_i)]\|_{\mathcal{T}} \xrightarrow{\mathrm{P}} 0,
\end{aligned}$$

which implies $|T_n/n - \|\mathbb{E}_n[f_*(\cdot,Z_i)]\|_{\mu,2}^2| \to_{\mathrm{P}} 0$. I may therefore focus on $\|\mathbb{E}_n[f_*(\cdot,Z_i)]\|_{\mu,2}^2$. For $t\in\mathcal{T}$ arbitrary ,

$$\mathbb{E}_n[f_*(\cdot,Z_i)] = \mathbb{E}_n[\omega(t,X)\,\rho(Z_i,\beta_*,h_*(W_i))\omega(t,X_i) + b_*(t)^\top s_*(Z_i) + \delta_*(t,W_i)\,U_i].$$

By T and CS we therefore get

$$\begin{aligned}
&\sup_{t\in\mathcal{T}}|\mathbb{E}_n[f_*(\cdot,Z_i)] - \mathrm{E}[\rho(Z,\beta_*,h_*(W))\,\omega(t,X)]| \\
&\leqslant \sup_{t\in\mathcal{T}}|(\mathbb{E}_n - \mathrm{E})[\rho(Z,\beta_*,h_*(W_i))\omega(t,X_i)]| + \|\mathbb{E}_n[s_*(Z_i)]\|\sup_{t\in\mathcal{T}}\|b_*(t)\| \\
&\quad + \sup_{t\in\mathcal{T}}|\mathbb{E}_n[\delta_*(t,W_i)\,U_i]| =: \mathrm{I}_n + \mathrm{II}_n + \mathrm{III}_n.
\end{aligned}$$

Consider first I_n. Given i.i.d. data, $\mathcal{T}$ compact (Assumption 1.2) $t\mapsto\rho(Z,\beta_*,h_*(W))\omega(t,X)$ continuous on $\mathcal{T}$ (Assumption 1.2), and $\sup_{t\in\mathcal{T}}|\rho(Z,\beta_*,h_*(W))\omega(t,X)| \lesssim |\rho(Z,\beta_*,h_*(W))|$ integrable (Assumptions 1.2 and 1.3), a uniform law of law numbers such as Newey and McFadden (1994, Lemma 2.4) shows that $\mathrm{I}_n \to_{\mathrm{P}} 0$.

Consider next II_n. Step 1.I.1 in the proof of Lemma 1.14 shows that $\sup_{t\in\mathcal{T}}\|b_*(t)\| < \infty$. Hence, by Assumption 1.1 and a weak law of large numbers for i.i.d. data,

$$\mathrm{II}_n \lesssim \|\mathbb{E}_n[s_*(Z_i)]\| \xrightarrow{\mathrm{P}} \|\mathrm{E}[s_*(Z)]\| = 0.$$

Lastly, consider III_n. Let $t\in\mathcal{T}$ and a sequence $t_m \in T$ converging to t be arbitrary. Assumptions 1.2 and 1.3 and the dominated convergence theorem show that $t\mapsto\delta_*(t,w)$

is continuous for any $w \in \mathcal{W}$. Also, by Assumptions 1.2, 1.3 and 1.4 and the fact that conditional expectations are $L^2(P)$-contractions,

$$\begin{aligned}
\mathrm{E}\Big\{\sup_{t\in\mathcal{T}}|\delta_*(t,W)\,U|\Big\} &\lesssim \mathrm{E}\left\{|U|\,\mathrm{E}\left[|\partial_v\rho(Z,\beta_*,h_*(W))|\,|\,W\right]\right\}\\
&\leqslant \left[\mathrm{E}\left(U^2\right)\right]^{1/2}\left[\mathrm{E}\left(\left\{\mathrm{E}\left[|\partial_v\rho(Z,\beta_*,h_*(W))|\,W\right]\right\}^2\right)\right]^{1/2}\\
&\leqslant \left[\mathrm{E}\left(U^2\right)\right]^{1/2}\left\{\mathrm{E}\left[\partial_v\rho(Z,\beta_*,h_*(W))^2\right]\right\}^{1/2} < \infty.
\end{aligned}$$

Hence, $\sup_{t\in\mathcal{T}}|\delta_*(t,W)\,U|$ is integrable. Given i.i.d. data, $\mathcal{T}$ compact, $t \mapsto \delta_*(t,W)\,U$ continuous on $\mathcal{T}$ and $\sup_{t\in\mathcal{T}}|\delta_*(t,W)\,U|$ integrable, Newey and McFadden (1994, Lemma 2.4) shows that

$$\mathrm{III}_n = \sup_{t\in T}|\mathbb{E}_n[\delta_*(t,W_i)\,U_i]| = \sup_{t\in\mathcal{T}}|(\mathbb{E}_n-\mathrm{E})\,[\delta_*(t,W_i)\,U_i]| \xrightarrow{\mathrm{P}} 0.$$

Given that $\mathrm{I}_n, \mathrm{II}_n, \mathrm{III}_n \to_{\mathrm{P}} 0$, we must have

$$\begin{aligned}
&\|\mathbb{E}_n[f_*(\cdot,Z_i)] - \mathrm{E}[\rho(Z,\beta_*,h_*(W))\,\omega(\cdot,X)]\|_{\mu,2}\\
&\leqslant \|\mathbb{E}_n[f_*(\cdot,Z_i)] - \mathrm{E}[\rho(Z,\beta_*,h_*(W))\,\omega(\cdot,X)]\|_{\mathcal{T}} \xrightarrow{\mathrm{P}} 0,
\end{aligned}$$

so by T and CMT it follows that $\|\mathbb{E}_n[f_*(\cdot,Z_i)]\|_{\mu,2}^2 \to_{\mathrm{P}} \|\mathrm{E}[\rho(Z,\beta_*,h_*(W))\,\omega(\cdot,X)]\|_{\mu,2}^2$. The probability limit is positive under the alternative by property (1.4.1) of the weight function and the choice of cdf μ. The conclusion now follows from Lemma 1.26. □

1.I.4 Supporting Lemmas for Section 1.4

Let A_n and B_n be symmetric but otherwise arbitrary random matrices of possibly growing dimension.

Lemma 1.17. *If $\lambda_{\min}(A_n) \geqslant c$ wp $\to 1$ and $\|B_n - A_n\|_{\mathrm{op}} \to_{\mathrm{P}} 0$, then (i) B_n is invertible wp $\to 1$ and (ii) $\lambda_{\min}(B_n)^{-1} \lesssim_{\mathrm{P}} 1$.*

Proof. To establish the first claim, I follow the argument in the proof of Newey (1995, Lemma A.4). For conformable vectors v, given that the maximal eigenvalue of a square matrix is bounded by its maximal singular value, it follows that

$$\begin{aligned}
\lambda_{\min}(B_n) &= \min_{\|v\|=1}\left\{v^\top B_n v\right\} = \min_{\|v\|=1}\left\{v^\top A_n v + v^\top(B_n - A_n)\,v\right\}\\
&\geqslant \min_{\|v\|=1}\left\{v^\top A_n v\right\} - \max_{\|v\|=1} v^\top(B_n - A_n)\,v\\
&= \lambda_{\min}(A_n) - \lambda_{\max}(B_n - A_n) \geqslant \lambda_{\min}(A_n) - \|B_n - A_n\|_{\mathrm{op}}.
\end{aligned}$$

Hence

$$\begin{aligned}
\mathrm{P}\left(\lambda_{\min}(B_n) < c/2\right) &\leqslant \mathrm{P}\left(\lambda_{\min}\left(A_n\right) - \|B_n - A_n\|_{\mathrm{op}} < c/2\right) \\
&= \mathrm{P}\left(\lambda_{\min}\left(A_n\right) - \|B_n - A_n\|_{\mathrm{op}} < c/2 \cap \lambda_{\min}\left(A_n\right) \geqslant c\right) \\
&\quad + \mathrm{P}\left(\lambda_{\min}\left(A_n\right) - \|B_n - A_n\|_{\mathrm{op}} < c/2 \cap \lambda_{\min}\left(A_n\right) < c\right) \\
&\leqslant \mathrm{P}\left(\|B_n - A_n\|_{\mathrm{op}} \geqslant c/2\right) + \mathrm{P}\left(\lambda_{\min}\left(A_n\right) < c\right),
\end{aligned}$$

which implies $\overline{\lim}_{n\to\infty} \mathrm{P}(\lambda_{\min}(B_n) < c/2) \leqslant 0$. It follows that $\mathrm{P}(\lambda_{\min}(B_n) < c/2) \to 0$, i.e. $\mathrm{P}(\lambda_{\min}(B_n) \geqslant c/2) \to 1$. Hence, B_n is invertible wp $\to 1$.

The proof of the first claim shows that $\lambda_{\min}(B_n) \geqslant c/2$ wp $\to 1$. Hence, for any $C > 2/c$ we have $\overline{\lim}_{n\to\infty} \mathrm{P}\left(\lambda_{\min}\left(B_n\right)^{-1} > C\right) \leqslant \overline{\lim}_{n\to\infty} \mathrm{P}\left(\lambda_{\min}\left(B_n\right) < c/2\right) = 0$. In particular, $\lim_{C\to\infty} \overline{\lim}_{n\to\infty} \mathrm{P}(\lambda_{\min}\left(B_n\right)^{-1} > C) = 0$. □

Let $\mathbf{Y}_n$ and H_n denote arbitrary random $n \times 1$ vectors of possibly growing dimension. Set $\mathbf{U}_n \coloneqq \mathbf{Y}_n - H_n$, and, for an arbitrary random matrix B_n with n rows and possibly growing column dimension, let $\breve{\pi}_n \coloneqq \left(B_n^\top B_n\right)^- B_n^\top Y_n$, and $\breve{H}_n \coloneqq B_n \breve{\pi}_n$.

Lemma 1.18. *If $\mathbf{U}_n^\top B_n \left(B_n^\top B_n\right)^- B_n^\top \mathbf{U}_n/n \lesssim_{\mathrm{P}} \varepsilon_n^2$, then for any conformable sequence of vectors $\overline{\pi}_n$,*

$$\|\breve{H}_n - H_n\|^2/n \lesssim_{\mathrm{P}} \varepsilon_n^2 + \|H_n - B_n\overline{\pi}_n\|^2/n.$$

Proof. On the event $\{\lambda_{\min}\left(B_n^\top B_n\right) \geqslant c\}$, $B_n^\top B_n$ is invertible $W_n \coloneqq B_n(B_n^\top B_n)^- B_n^\top$. Because $B_n^\top B_n$ is symmetric, so is its generalized inverse. It follows that W_n is symmetric and idempotent, so its eigenvalues are are bounded by one. Given that W_n also satisfies $W_n B_n = B_n$ Rao (1973, 1b.5(vi)(a)),

$$\begin{aligned}
\|\breve{H}_n - H_n\|^2/n &= \left((W_n Y_n - H_n)^2 = H_n^\top W H_n + Y_n^\top W_n Y_n - 2H_n^\top W_n Y_n\right)/n \\
&= \left[\mathbf{U}_n^\top W_n \mathbf{U}_n + H_n^\top \left(I_n - W_n\right) H_n\right]/n \\
&= \Big[\mathbf{U}_n^\top W_n \mathbf{U}_n + \left(H_n - B_n\overline{\pi}_n\right)^\top \left(I_n - W_n\right)\left(H_n - B_n\overline{\pi}_n\right)\Big]/n \\
&\leqslant \left[\mathbf{U}_n^\top W_n \mathbf{U}_n + \|H_n - B_n\overline{\pi}_n\|^2\right]/n \lesssim_{\mathrm{P}} \varepsilon_n^2 + \|H_n - B_n\overline{\pi}_n\|^2/n,
\end{aligned}$$

where the inequality follows from the Min-Max theorem. □

Lemma 1.19. *If $\lambda_{\min}(A_n) \geqslant c$, $\|B_n^\top B_n/n - A_n\|_{\mathrm{op}} \to_{\mathrm{P}} 0$ and $\mathbf{U}_n^\top B_n \left(B_n^\top B_n\right)^- B_n^\top \mathbf{U}_n/n \lesssim_{\mathrm{P}} \varepsilon_n^2$, then for any conformable sequence of vectors $\overline{\pi}_n$,*

$$\|\breve{\pi}_n - \overline{\pi}_n\|^2 \lesssim_{\mathrm{P}} \varepsilon_n^2 + \|H_n - B_n\overline{\pi}_n\|^2/n,$$

$$\|\check{H} - B_n\overline{\pi}_n\|^2/n \lesssim_{\mathrm{P}} \varepsilon_n^2 + \|H_n - B_n\overline{\pi}_n\|^2/n.$$

Proof. Denote $\overline{H}_n \coloneqq B_n\overline{\pi}_n$. I follow the argument in the proof of Newey (1995, Lemma A.8). By Lemma 1.17, $\lambda_{\min}\left(B_n^\top B_n/n\right) \geqslant c$ wp $\to 1$ and $\lambda_{\min}\left(B_n^\top B_n/n\right)^{-1} \lesssim_{\mathrm{P}} 1$, so by the Min-Max theorem,

$$\begin{aligned}\|\check{\pi}_n - \overline{\pi}_n\|^2 &\leqslant \lambda_{\min}\left(B_n^\top B_n/n\right)^{-1}(\check{\pi}_n - \overline{\pi}_n)^\top\left(B_n^\top B_n\right)(\check{\pi}_n - \overline{\pi}_n)/n. \\ &= \lambda_{\min}\left(B_n^\top B_n/n\right)^{-1}\|\check{H}_n - \overline{H}_n\|^2/n \lesssim_{\mathrm{P}} \|\check{H}_n - \overline{H}_n\|^2/n.\end{aligned}$$

It remains to prove the second claim. Let $W_n \coloneqq B_n(B_n^\top B_n)^- B_n^\top$, which is Given that $B_n^\top B_n$ is symmetric, so is its generalized inverse. It follows that W_n is symmetric and idempotent and therefore positive semidefinite. Hence

$$\begin{aligned}&\|\check{H}_n - \overline{H}_n\|^2 = \|W_n\mathbf{Y}_n - \overline{H}_n\|^2 = \mathbf{Y}_n^\top W_n\mathbf{Y}_n - 2Y_n^\top W_n\overline{H}_n + \overline{H}_n^\top\overline{H}_n \\ &= \mathbf{U}_n^\top W_n\mathbf{U}_n + H_n^\top W_n H_n + 2\mathbf{U}_n^\top W_n H_n - 2H_n^\top W_n\overline{H}_n - 2\mathbf{U}_n^\top W_n\overline{H}_n + \overline{H}_n^\top W_n\overline{H}_n \\ &= (\mathbf{U}_n + H_n - \overline{H}_n)^\top W_n(\mathbf{U}_n + H_n - \overline{H}_n) \leqslant 2\mathbf{U}_n^\top W_n\mathbf{U}_n + 2(H_n - \overline{H}_n)^\top W_n(H_n - \overline{H}_n) \\ &\leqslant 2\mathbf{U}_n^\top W_n\mathbf{U}_n + 2\|H_n - \overline{H}_n\|^2\end{aligned}$$

where $\overline{H}_n^\top\overline{H}_n = \overline{\pi}_n^\top B_n^\top B_n\overline{\pi}_n = \overline{\pi}_n^\top B_n^\top B_n(B_n^\top B_n)^- B_n^\top B_n\overline{\pi}_n\overline{H}_n^\top W_n\overline{H}_n$ follows from definition of a generalized inverse, the first inequality follows from $(v+w)^\top M(v+w) \leqslant 2v^\top Mv + 2w^\top Mw$ for M positive semidefinite, and the second inequality from the Min-Max theorem and the fact that idempotent matrices only have eigenvalues equal to zero or one. Dividing through by n, we get

$$\|\check{H}_n - \overline{H}_n\|^2 \leqslant 2\mathbf{U}_n^\top W_n\mathbf{U}_n/n + 2\|H_n - \overline{H}_n\|^2/n \lesssim_{\mathrm{P}} \varepsilon_n^2 + \|H_n - \overline{H}_n\|^2/n.$$

□

Now let $U_i = Y_i - h^*(W_i)$ as in the main text, write $\mathbf{U}_n$ for the $n \times 1$ vector of U_i's, and define P_k to be the $n \times k$ matrix arising from stacking the $p^k(W_i)$'s.

Lemma 1.20. *If Assumption 1.4 holds and $k_n/n \to 0$, then*

$$\mathbf{U}_n^\top P_{k_n}(P_{k_n}^\top P_{k_n})^- P_{k_n}^\top\mathbf{U}_n/n \lesssim_{\mathrm{P}} k_n/n.$$

Proof. Let $\mathbf{W}_n$ denote the collection $\{W_i | i = 1, \ldots, n\}$. By the law of iterated expectations in combination with i.i.d. data and $\mathrm{E}(U|W) = 0$ a.s., for any $k \in \mathbf{N}$,

$$\begin{aligned}
\mathrm{E}[\mathbf{U}_n^\top P_k (P_k^\top P_k)^- P_k^\top \mathbf{U}_n] &= \mathrm{E}[\mathrm{tr}(\mathbf{U}_n^\top P_k \left(P_k^\top P_k\right)^- P_k^\top \mathbf{U}_n)] = \mathrm{E}[\mathrm{tr}(\left(P_k^\top P_k\right)^- P_k^\top \mathbf{U}_n \mathbf{U}_n^\top P_k)] \\
&= \mathrm{E}[\mathrm{tr}(\widehat{Q}_k^- n^{-1} \mathrm{E}[P_k^\top \mathbf{U}_n \mathbf{U}_n^\top P_k | \mathbf{W}_n])] \\
&= \mathrm{E}\Big[\mathrm{tr}\Big(\widehat{Q}_k^- \frac{1}{n} \mathrm{E}\Big[\sum_{i=1}^n \sum_{j=1}^n U_i U_j p^k(W_i) p^k(W_j)^\top \Big| \mathbf{W}_n\Big]\Big)\Big] \\
&= \mathrm{E}\Big[\mathrm{tr}\Big(\widehat{Q}_k^- \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \mathrm{E}\left(U_i U_j | \mathbf{W}_n\right) p^k(W_i) p^k(W_j)^\top \Big)\Big] \\
&= \mathrm{E}\Big[\mathrm{tr}\Big(\widehat{Q}_k^- \frac{1}{n} \sum_{i=1}^n \mathrm{E}\left(U_i^2 | W_i\right) p^k(W_i) p^k(W_i)^\top \Big)\Big] \\
&\lesssim \mathrm{E}[\mathrm{tr}(\widehat{Q}_k^- \mathbb{E}_n[p^k(W_i) p^k(W_i)^\top])] = \mathrm{E}[\mathrm{tr}(\widehat{Q}_k^- \widehat{Q}_k)] \leqslant k,
\end{aligned}$$

where the last inequality follows from $\widehat{Q}_k^- \widehat{Q}_k$ having eigenvalues equal to zero or one. Dividing by n, the claim now follows from M and $k_n/n \to 0$. □

Lemma 1.21. *If $\zeta_{k_n}^2 \ln(k_n)/n \to 0$ and the eigenvalues of Q_k are bounded from above uniformly in k, then $\|\widehat{Q}_{k_n} - Q_{k_n}\|_{\mathrm{op}} \lesssim_{\mathrm{P}} [\zeta_{k_n}^2 \ln(k_n)/n]^{1/2}$.*

Proof. The matrix $\widehat{Q}_k = \mathbb{E}_n[p^k(W_i)p^k(W_i)]$ is the average of the n independent, symmetric, nonnegative $k \times k$-matrix valued random variables $p^k(W_i) p^k(W_i)^\top$, and the matrix Q_k is is their common mean. Given that the operator norm $\|\cdot\|_{\mathrm{op}}$ is always dominated by the Frobenius norm $\|\cdot\|_F$, and

$$\begin{aligned}
\|p^k(W_i) p^k(W_i)^\top\|_F &= [\mathrm{tr}(p^k(W_i) p^k(W_i)^\top p^k(W_i) p^k(W_i)^\top)]^{1/2} \\
&= [\mathrm{tr}(p^k(W_i)^\top p^k(W_i) p^k(W_i)^\top p^k(W_i))]^{1/2} = \|p^k(W_i)\|^2 \leqslant \zeta_k^2,
\end{aligned}$$

each of these n random matrices satisfy $\|p^k(W_i) p^k(W_i)^\top\|_{\mathrm{op}} \leqslant \zeta_k^2$. By hypothesis, $\|Q_k\|_{\mathrm{op}} = [\lambda_{\max}\left(Q_k^\top Q_k\right)]^{1/2} = \lambda_{\max}(Q_k) \lesssim 1$. Belloni, Chernozhukov, Chetverikov, and Kato (2015, Lemma 6.2), which builds on a fundamental result obtained by Rudelson (1999), therefore implies

$$\mathrm{E}\left[\|\widehat{Q}_{k_n} - Q_{k_n}\|_{\mathrm{op}}\right] \lesssim \frac{\zeta_{k_n}^2 \ln k_n}{n} + \sqrt{\frac{\zeta_{k_n}^2 \ln k_n}{n}}.$$

Since $\zeta_{k_n}^2 \ln(k_n)/n \to 0$, the claim now follows from M. □

Lemma 1.22. *If Assumptions 1.4, 1.5 and 1.6 hold, $k_n/n \to 0$ and $\zeta_{k_n}^2 \ln(k_n)/n \to 0$, then for $\widetilde{\pi}_k$ provided by Assumption 1.6 and $\widetilde{h}_k \coloneqq p^{k\top} \widetilde{\pi}_k$ we have (1) $\|\widehat{\pi} - \widetilde{\pi}_{k_n}\| \lesssim_{\mathrm{P}} \sqrt{k_n/n} + k_n^{-\alpha}$;*

(2) $\|\widehat{h}-\widetilde{h}_{k_n}\|_{\mathbb{P}_n,2} \lesssim_{\mathrm{P}} \sqrt{k_n/n}+k_n^{-\alpha}$*; (3)* $\|\widehat{h}-h_*\|_{\mathbb{P}_n,2} \lesssim_{\mathrm{P}} \sqrt{k_n/n}+k_n^{-\alpha}$*; and, (4)* $\|\widehat{h}-h_*\|_{\mathcal{W}} \lesssim_{\mathrm{P}} \zeta_{k_n}(\sqrt{k_n/n}+k_n^{-\alpha})$

Proof. Assumption 1.5 and Lemma 1.21 imply that $\|\widehat{Q}_{k_n} - Q_{k_n}\|_{\mathrm{op}} \lesssim_{\mathrm{P}} [\zeta_{k_n}^2 \ln(k_n)/n]^{1/2}$, so $\|\widehat{Q}_{k_n} - Q_{k_n}\|_{\mathrm{op}} \to_{\mathrm{P}} 0$. Assumption 1.4 implies that $\mathbf{U}_n^\top P_{k_n}(P_{k_n}^\top P_{k_n})^- P_{k_n}^\top \mathbf{U}_n/n \lesssim_{\mathrm{P}} k_n/n$. Setting up for an application of Lemma 1.19, let $A_n \coloneqq Q_{k_n}$, $B_n \coloneqq P_{k_n}$ the $n \times k_n$ matrix arising from stacking the $p^{k_n}(W_i)^\top$'s, $\mathbf{Y}_n$ the $n \times 1$ vector of Y_i's, H_n the $n \times 1$ vector of $h_*(W_i)$'s, and set $\overline{\pi}_n \coloneqq \widetilde{\pi}_{k_n}$. Then $\breve{\pi}_n = (B_n^\top B_n)^- B_n^\top \mathbf{Y}_n = \widehat{Q}_{k_n}^- \mathbb{E}_n[p^{k_n}(W_i)\,Y_i] = \widehat{\pi}$, and an application of Lemma 1.19 with $\varepsilon_n^2 \coloneqq k_n/n$ yields

$$
\begin{aligned}
\|\widehat{\pi} - \widetilde{\pi}_{k_n}\| &\lesssim_{\mathrm{P}} \sqrt{k_n/n} + \|\widetilde{h}_{k_n} - h_*\|_{\mathbb{P}_n,2},\\
\|\widehat{h} - \widetilde{h}_{k_n}\|_{\mathbb{P}_n,2} &\lesssim_{\mathrm{P}} \sqrt{k_n/n} + \|\widetilde{h}_{k_n} - h_*\|_{\mathbb{P}_n,2}.
\end{aligned}
$$

Similarly, an application of Lemma 1.18 shows that

$$
\|\widehat{h} - h_*\|_{\mathbb{P}_n,2} \lesssim_{\mathrm{P}} \sqrt{k_n/n} + \|\widetilde{h}_{k_n} - h_*\|_{\mathbb{P}_n,2}.
$$

Claims 1, 2 and now all follow from Assumption 1.6 and $\|\widetilde{h}_{k_n} - h_*\|_{\mathbb{P}_n,2} \leqslant \|\widetilde{h}_{k_n} - h_*\|_{\mathcal{W}}$. By T, CS, Claim 1 and Assumption 1.6,

$$
\begin{aligned}
\|\widehat{h} - h_*\|_{\mathcal{W}} &\leqslant \|\widehat{h} - \widetilde{h}_{k_n}\|_{\mathcal{W}} + \|\widetilde{h}_{k_n} - h_*\|_{\mathcal{W}} = \|p^{k_n\top}(\widehat{\pi} - \widetilde{\pi}_{k_n})\|_{\mathcal{W}} + \|\widetilde{h}_{k_n} - h_*\|_{\mathcal{W}}\\
&\leqslant \|\widehat{\pi}_n - \widetilde{\pi}_{k_n}\| \sup_{w \in \mathcal{W}} \|p^{k_n}(w)\| + \|\widetilde{h}_{k_n} - h_*\|_{\mathcal{W}} \leqslant \zeta_{k_n}\|\widehat{\pi}_n - \widetilde{\pi}_{k_n}\| + \|\widetilde{h}_{k_n} - h_*\|_{\mathcal{W}}\\
&\lesssim_{\mathrm{P}} \zeta_{k_n}(\sqrt{k_n/n} + k_n^{-\alpha}) + k_n^{-\alpha} \lesssim \zeta_{k_n}(\sqrt{k_n/n} + k_n^{-\alpha}),
\end{aligned}
$$

where the $\lesssim$ follows from $\zeta_k \nrightarrow 0$ as $k \to \infty$. □

Lemma 1.23. *If* X_n *is a sequence of nonnegative random variables defined on a common probability space* $(\Omega, \mathcal{F}, \mathrm{P})$*,* $\mathcal{F}_n$ *is a sequence of sub-*σ*-algebras, and* $\mathrm{E}(X_n|\,\mathcal{F}_n) \to_{\mathrm{P}} 0$*, then* $X_n \to_{\mathrm{P}} 0$*.*

Proof. Fix $n \in \mathbf{N}$, let $Y_n \coloneqq \mathrm{E}(X_n|\,\mathcal{F}_n)$ and let $A_n \coloneqq \{Y_n = 0\}$. Then $X_n = 0$ almost everywhere on A_n. Indeed, if X_n is not zero almost everywhere on A_n, then there exists $C \in (0, \infty)$ such that $B_{n,C} \coloneqq \{\omega \in A_n|\, X_n(\omega) > 1/C\}$ satisfies $\mathrm{P}(B_{n,C}) > 0$. By definition of (a version of) the conditional expectation of X_n given $\mathcal{F}_n$, we must have $\int_A X_n \mathrm{d}P = \int_A Y_n \mathrm{d}P$ for every $A \in \mathcal{F}_n$ and, in particular, for A_n. Since $Y_n = 0$ on A_n and $B_{n,C} \subset A_n$, it follows that

$$
0 = \int_{A_n} Y_n \mathrm{d}P = \int_{A_n} X_n \mathrm{d}P \geqslant \int_{B_{n,C}} X_n \mathrm{d}P \geqslant \mathrm{P}(B_{n,C})/C,
$$

which contradicts $\mathrm{P}(B_{n,C}) > 0$. Since $n \in \mathbf{N}$ was arbitrary, we have shown that $X_n = 0$ on A_n for each $n \in \mathbf{N}$. Now, fix $\varepsilon, \delta > 0$. Then $\mathrm{P}(X_n > \varepsilon \cap Y_n = 0) = 0$ by the previous claim, and it follows that

$$\begin{aligned}\mathrm{P}(X_n > \varepsilon) &= \mathrm{P}(X_n > \varepsilon \cap Y_n = 0) + \mathrm{P}(X_n > \varepsilon \cap 0 < Y_n \leqslant \delta\varepsilon) + \mathrm{P}(X_n > \varepsilon \cap Y_n > \delta\varepsilon)\\ &\leqslant \mathrm{P}\left(X_n > \delta^{-1} Y_n > 0\right) + \mathrm{P}(Y_n > \delta\varepsilon).\end{aligned}$$

Given that Y_n is $\mathcal{F}_n$ measurable, by conditional M we have

$$\begin{aligned}\mathrm{P}\left(X_n > \delta^{-1} Y_n > 0\right) &= \mathrm{E}\left[\mathbf{1}_{Y_n > 0} \mathrm{P}\left(X_n > \delta^{-1} Y_n \middle| \mathcal{F}_n\right)\right] \leqslant \mathrm{E}\left[\mathbf{1}_{Y_n > 0} \delta \mathrm{E}(X_n | \mathcal{F}_n) / Y_n\right]\\ &= \delta \mathrm{P}(Y_n > 0) \leqslant \delta.\end{aligned}$$

By $Y_n \to_{\mathrm{P}} 0$ and the previous two displays we see that for any $\varepsilon, \delta > 0, \overline{\lim}\, \mathrm{P}(X_n > \varepsilon) \leqslant \delta$, so the claim follows from letting $\delta \to 0$. □

Lemma 1.24. *Let X_n and Y_n be sequences of stochastic processes defined on a common probability space $(\Omega, \mathcal{F}, P)$ and taking values in a separable metric space $(\mathbb{D}, d)$, and let $\mathcal{F}_n$ be a sequence of sub-σ-algebras. If $X_n \rightsquigarrow_{\mathrm{P},\mathcal{F}} X$ in $\mathbb{D}$ and $d(X_n, Y_n) \to_{\mathrm{P}} 0$, then $Y_n \rightsquigarrow_{\mathrm{P},\mathcal{F}} X$ in $\mathbb{D}$.*

Proof. By T

$$\begin{aligned}&\sup_{h \in \mathrm{BL}_1(\mathbb{D})} \left|\mathrm{E}[h(Y_n)| \mathcal{F}_n] - \mathrm{E}[h(X)]\right|\\ &\leqslant \sup_{h \in \mathrm{BL}_1(\mathbb{D})} \left|\mathrm{E}[h(Y_n) - h(X_n)| \mathcal{F}_n]\right| + \sup_{h \in \mathrm{BL}_1(\mathbb{D})} \left|\mathrm{E}[h(X_n)| \mathcal{F}_n] - \mathrm{E}[h(X)]\right|\\ &\leqslant d(X_n, Y_n) \wedge 2 + o_{\mathrm{P}}(1) \overset{\mathrm{P}}{\to} 0.\end{aligned}$$

□

Let μ be the cdf from the main text, which is absolutely continuous and bounded away from zero on $\mathcal{T}$ given by Assumption 1.2.

Lemma 1.25. *Let $X = \{X_t | t \in \mathcal{T}\}$ be a zero-mean Gaussian process indexed by $\mathcal{T} \subset \mathbf{R}^{d_t}$ compact with almost every sample path $t \mapsto X_t(\omega)$ continuous, μ an absolutely continuous, everywhere strictly positive probability measure on $\mathcal{T}$, and $\int_{\mathcal{T}} X_t^2 \mathrm{d}\mu(t) < \infty$ a.s. Moreover, let $t \mapsto \sigma_t^2 \coloneqq \mathrm{E}(X_t^2)$ be continuous and nondegenerate on $\mathcal{T}$. Then the cdf F of $\|X\|_{\mu,2}^2 \coloneqq \int X_t^2 \mathrm{d}\mu(t)$ is everywhere continuous and strictly increasing on $[0, \infty)$.*

Proof of Lemma 1.25. Given that the law of X is Gaussian and the functional $\|\cdot\|_{\mu,2} : \ell^\infty(\mathcal{T}) \to \mathbf{R}_+$ is Lipschitz (hence lower semi-continuous), Davydov, Lifshits, and Smorodina,

N. V. (1998, Theorem 11.1) implies that (i) $F_{\|X\|_{\mu,2}}$ is everywhere continuous except possibly at the separation point zero; (ii) $F_{\|X\|_{\mu,2}}$ is absolutely continuous on $(0,\infty)$, (iii) $F_{\|X\|_{\mu,2}}$ is differentiable on $(0,\infty)$ except on an at most countable exceptional set $\Delta \subset (0,\infty)$; (iv) the derivative $F'_{\|X\|_{\mu,2}}$ is positive on $(0,\infty)\setminus\Delta$ and $\mathrm{P}\left(\|X\|_{\mu,2} \in A\right) = \int_A F'_{\|X\|_{\mu,2}}(r)\,\mathrm{d}r$ for any $A \subset \mathbf{R}$ Borel [where $F'_{\|X\|_{\mu,2}}(r)$ is understood to be zero for $r \notin (0,\infty)\setminus\Delta$]. Letting $r, s \in \mathbf{R}$ be such that $0 \leqslant r < s$, we have $\mathbf{1}_{(-\infty,r]} \leqslant \mathbf{1}_{(-\infty,s]}$, which holds with strict inequality for any $t \in (r,s]$. The claim that $F_{\|X\|_{\mu,2}}$ is strictly increasing on $[0,\infty)$ now follows from (iv) by integrating with respect to $F'_{\|X\|_{\mu,2}}(t)\,\mathrm{d}t$. To show that $F_{\|X\|_{\mu,2}}$ is continuous at zero, note that $t \mapsto \sigma_t^2$ being continuous and nondegenerate on the compact $\mathcal{T}$ imply that $\sup_{t\in\mathcal{T}} \sigma_t^2$ is attained and strictly positive. Hence, there exists $t_{\max} \in \mathcal{T}$ such that $\sigma^2_{\max} \coloneqq \sigma^2_{t_{\max}} > 0$. By Gaussianity, the associated marginal satisfies $\mathrm{P}(X_{t_{\max}} \neq 0) = 1$, i.e., there exists $A \subset \Omega$ such that $\mathrm{P}(A) = 1$ and $X_{t_{\max}}(\omega) \neq 0$ for every $\omega \in A$. By assumption there exists $B \subset \Omega$ such that $\mathrm{P}(B) = 1$ and the sample path $t \mapsto X_t(\omega)$ is continuous for every $\omega \in B$. Thus, for every $\omega \in A \cap B$, there exists a neighborhood $C(\omega)$ of $t_{\max}$ in $\mathcal{T}$ such that $X_t(\omega) \neq 0$ for all $t \in C(\omega)$. Given that μ is an absolutely continuous, everywhere strictly positive probability measure on $\mathcal{T}$, $\mu(C(\omega)) > 0$ for all $\omega \in A \cap B$. Consequently, for each $\omega \in A \cap B$, $\int_{\mathcal{T}} X_t(\omega)^2\,\mathrm{d}\mu(t) \geqslant \int_{C(\omega)} X_t(\omega)^2\,\mathrm{d}\mu(t) > 0$. Given that $\mathrm{P}(A \cap B) = 1$, we have shown that $\mathrm{P}(\int_{\mathcal{T}} X_t(\omega)^2\,\mathrm{d}\mu(t) > 0) = 1$, which is equivalent to $F_{\|X\|_{\mu,2}}(0) = \mathrm{P}\left(\|X\|_{\mu,2} = 0\right) = 0$, as desired. The conclusion now follows from $F(u) = F_{\|X\|_{\mu,2}}(\sqrt{u})$ on $u \in [0,\infty)$. □

Lemma 1.26. *If $X_n \to_{\mathrm{P}} c > 0$ and $Y_n \to_{\mathrm{P}} 0$, then $\mathrm{P}(X_n > Y_n) \to 1$.*

Proof. The union bound implies

$$\begin{aligned}
\mathrm{P}\left(X_n \leqslant Y_n\right) &\leqslant \mathrm{P}\left(X_n \leqslant Y_n \text{ and } Y_n \leqslant c/2\right) + \mathrm{P}\left(Y_n > c/2\right) \\
&\leqslant \mathrm{P}\left(X_n \leqslant c/2\right) + \mathrm{P}\left(|Y_n| > c/2\right) \\
&\leqslant \mathrm{P}\left(|X_n - c| \geqslant c/2\right) + \mathrm{P}\left(|Y_n| > c/2\right).
\end{aligned}$$

Both terms on the right-hand side go to zero. Taking the lim sup shows that $\mathrm{P}\left(X_n \leqslant Y_n\right) \to 0$. □

1.J Proofs for Section 1.5

1.J.1 Proofs for Section 1.5.4

Proof of Lemma 1.4. The proof of the claim follows from two applications of Theorem 1.7. I first verify the conditions of Theorem 1.7 for estimation of the single best linear predictor L_*. Using the conclusion from Theorem 1.7, I then verify these conditions once more for estimation of the very many best linear predictors $\{L_{k*}\}_1^q$. The result will then follow from the union bound.

For the *first* application, where $q = 1$, $\beta_{1*} = h_*$ and $\beta_{10} = h_0$, observe that Assumption 1.16 follows from Assumption 1.11, Assumption 1.17 holds trivially since no outcomes are estimated, Assumption 1.18 follows from Assumption 1.12, and Assumption 1.19 is implied by Assumption 1.14. Theorem 1.7 and the hypotheses of the lemma now shows that there exists c, C, C' and n_0 depending only on $c_0, c_0', c_1, C_1, c_2, C_2$ and c_2' such that for all $n \geqslant n_0$, $\|\widehat{L} - L_*\|_{\mathbb{P}_n,2} \leqslant C'\sqrt{s\ln(pn)/n}$.

For the *second* application, where $q = q, \beta_{k*} = \mu_{k*}$ and $\beta_{k0} = \mu_{k0}$, observe that Assumption 1.16 follows from Assumption 1.11, and Assumption 1.18 is implied by Assumptions 1.12 and 1.13. To verify the remaining Assumption 1.17, note that $e_{ik} = X_k\partial_v\rho(Z_i, \widehat{\beta}, \widehat{L}(W_i)) - X_k\partial_v\rho(Z_i, \beta_*, L_*(W_i))$ by Assumptions 1.12 and 1.13 implies the bound

$$|e_{ik}| \leqslant C_1|\partial_v\rho(Z_i, \widehat{\beta}, \widehat{L}(W_i)) - \partial_v\rho(Z_i, \beta_*, L_*(W_i))| \leqslant C'[\|\widehat{\beta} - \beta_*\| + |\widehat{L}(W_i) - L_*(W_i)|].$$

The previous bound implies the following bound on the outcome estimation error:

$$\Delta \leqslant C'(\|\widehat{\beta} - \beta_*\| + \|\widehat{L} - L_*\|_{\mathbb{P}_n,2}).$$

Using Assumptions 1.10 and 1.14 (the latter to get $a_n \leqslant C_2$), Lemma 1.27 implies $\|\widehat{\beta} - \beta_*\| \leqslant C'\sqrt{\ln(n)/n}$ wp $\geqslant 1 - Cn^{-c}$ for constants c, C and C' depending only on C_1, C_2 and c_2. Assumption 1.17 now follows from the first application of Theorem 1.7 and the union bound. □

Proof of Lemma 1.5. Under the assumptions of the lemma and the (maintained) assumption that $\widehat{L}$ and the $\widehat{L}_k$'s are the Lasso estimates of L_* and the L_{k*}'s, respectively, resulting from using conservatively or truly polynomially penalty loadings (such as the penalty loadings resulting from Algorithms 1.2 and 1.3), Lemma 1.4 shows that there exists c, C, C' and n_0 depending only on $c_0, c_0', c_1, C_1, c_2, C_2$ and c_2' such that for all $n \geqslant n_0$

$$\|\widehat{L} - L_*\|_{\mathbb{P}_n,2} + \max_{1\leqslant k\leqslant q}\|\widehat{L}_k - L_{k*}\|_{\mathbb{P}_n,2} \leqslant C'\sqrt{\frac{s\ln(pqn)}{n}} \quad \text{wp} \geqslant 1 - Cn^{-c}. \tag{1.J.1}$$

The remainder of the proof is divided into steps.

Main

Let $s_i \equiv s\left(Z_i\right)$. By T,

$$\begin{aligned}
&\left|T - T_*\right| \\
&\leqslant \max_{1 \leqslant k \leqslant q} \Big| \sqrt{n}\mathbb{E}_n[\psi_k(Z_i, \widehat{\beta}, \widehat{L}\left(W_i\right), \widehat{L}_k\left(W_i\right))] \\
&\quad - \sqrt{n}\mathbb{E}_n[\psi_k(Z_i, \beta_*, L_*\left(W_i\right), L_{k*}\left(W_i\right)) + b_k^\top s_i] \Big| \\
&\leqslant \max_{1 \leqslant k \leqslant q} \left| \sqrt{n}\mathbb{E}_n[\psi_k(Z_i, \beta_*, \widehat{L}\left(W_i\right), \widehat{L}_k\left(W_i\right)) - \psi_k(Z_i, L_*\left(W_i\right), L_{k*}\left(W_i\right))] \right| \\
&\quad + \max_{1 \leqslant k \leqslant q} \left| \sqrt{n}\mathbb{E}_n[\psi_k(Z_i, \widehat{\beta}, \widehat{L}\left(W_i\right), \widehat{L}_k\left(W_i\right)) - \psi_k(Z_i, \beta_*, \widehat{L}\left(W_i\right), \widehat{L}_k\left(W_i\right))] - b_k^\top \sqrt{n}\mathbb{E}_n\left(s_i\right) \right| \\
&=: \mathrm{I} + \mathrm{II}.
\end{aligned}$$

Step I below shows that for some C'

$$\mathrm{P}\left(\mathrm{I} > C'\sqrt{s^2 \ln^3(pqn)/n}\right) \leqslant 7n^{-1},$$

while Step II below shows that for some c, C and C',

$$\mathrm{P}\left(\mathrm{II} > C' \max\left\{\sqrt{s \ln^2(pqn)/n}, n^{-c_2/4}/\sqrt{\ln\left(pqn\right)}, a_n\right\}\right) \leqslant Cn^{-c}.$$

The claim now follows from the three previous displays in combination with the union bound.

I

This step shows that for some C',

$$\mathrm{P}\left(\mathrm{I} > C'\sqrt{s^2 \ln^3(pqn)/n}\right) \leqslant 7n^{-1}. \tag{1.J.2}$$

Let $\rho_i\left(v\right) \coloneqq \rho(Z_i, \beta_*, v)$. Decompose the kth summand and use MVT, to get

$$\begin{aligned}
&\sqrt{n}\mathbb{E}_n[\psi_k(Z_i, \beta_*, \widehat{L}\left(W_i\right), \widehat{L}_k\left(W_i\right)) - \psi_k(Z_i, \beta_*, L_*\left(W_i\right), L_{k*}\left(W_i\right))] \\
&= \sqrt{n}\mathbb{E}_n[X_{ik}\partial_v\rho_i(W_i^\top \overline{h}_i^{(k)}) - W_i^\top \overline{\mu}_i^{(k)}]W_i^\top(\widehat{h} - h_*) + \sqrt{n}\mathbb{E}_n[(Y_i - W_i^\top \overline{h}_i^{(k)})W_i^\top(\widehat{\mu}_k - \mu_{k*})] \\
&= \sqrt{n}\mathbb{E}_n\{[X_{ik}\partial_v\rho_i(W_i^\top h_*) - W_i^\top \mu_{k*}]W_i^\top(\widehat{h} - h_*)\} + \sqrt{n}\mathbb{E}_n[(Y_i - W_i^\top h_*)W_i^\top(\widehat{\mu}_k - \mu_{k*})] \\
&\quad + \sqrt{n}\mathbb{E}_n\{X_{ik}[\partial_v\rho_i(W_i^\top \overline{h}_i^{(k)}) - \partial_v\rho_i(W_i^\top h_*)]W_i^\top(\widehat{h} - h_*)\} \\
&\quad + \sqrt{n}\mathbb{E}_n[W_i^\top(\overline{\mu}_i^{(k)} - \mu_{*k})W_i^\top(h_* - \widehat{h})] + \sqrt{n}\mathbb{E}_n[W_i^\top(\mu_{k*} - \widehat{\mu}_k)W_i^\top(\overline{h}_i^{(k)} - h_*)],
\end{aligned}$$

where $(W_i^\top \overline{h}_i^{(k)}, W_i^\top \overline{\mu}_i^{(k)})$ lies on the line segment connecting $(W_i^\top \widehat{h}, W_i^\top \widehat{\mu}_k)$ and $(W_i^\top h_*, W_i^\top \mu_{*k})$. The MVE in the previous display implies

$$\begin{aligned}
\mathrm{I} &= \max_{1\leqslant k\leqslant q} |\sqrt{n}\mathbb{E}_n[\psi_k(Z_i,\beta_*,\widehat{L}(W_i),\widehat{L}_k(W_i)) - \sqrt{n}\mathbb{E}_n\psi_j(Z_i,\beta_*,L_*(W_i),L_{k*}(W_i))]| \\
&\leqslant \max_{1\leqslant k\leqslant q} |\sqrt{n}\mathbb{E}_n\{[\partial_v\rho_i(W_i^\top h_*)X_{ik} - W_i^\top \mu_{k*}]W_i^\top(\widehat{h}-h_*)\}| \\
&\quad + \max_{1\leqslant k\leqslant q} |\sqrt{n}\mathbb{E}_n[(Y_i - W_i^\top h_*)W_i^\top(\widehat{\mu}_k - \mu_{k*})]| \\
&\quad + \max_{1\leqslant k\leqslant q} |\sqrt{n}\mathbb{E}_n\{X_{ik}[\partial_v\rho_i(W_i^\top \overline{h}_i^{(k)}) - \partial_v\rho_i(W_i^\top h_*)]W_i^\top(\widehat{h}-h_*)\}| \\
&\quad + \max_{1\leqslant k\leqslant q} |\sqrt{n}\mathbb{E}_n[W_i^\top(\overline{\mu}_i^{(k)} - \mu_{k*})W_i^\top(\widehat{h}-h_*)]| \\
&\quad + \max_{1\leqslant k\leqslant q} |\sqrt{n}\mathbb{E}_n[W_i^\top(\widehat{\mu}_k - \mu_{k*})W_i^\top(\overline{h}_i^{(k)}-h_*)]| =: \mathrm{I}_a + \mathrm{I}_b + \mathrm{I}_c + \mathrm{I}_d + \mathrm{I}_e. \qquad (1.J.3)
\end{aligned}$$

Equation (1.J.2) now follows from Steps I_a–I_e below.

I_a

This step shows that for some C',

$$\mathrm{P}\left(\mathrm{I}_a > C'\sqrt{s^2\ln^3(pqn)/n}\right) \leqslant Cn^{-c}. \qquad (1.J.4)$$

To establish the claim, note that by Hölder's inequality,

$$\mathrm{I}_a \leqslant \|\widehat{h}-h_*\|_1 \max_{(j,k)\in[p]\times[q]} \left|\sqrt{n}\mathbb{E}_n\{[X_{ik}\partial_v\rho_i(W_i^\top h_*) - W_i^\top \mu_{k*}]W_{ij}\}\right| =: \mathrm{I}_{a,1}\times \mathrm{I}_{a,2}.$$

Equation (1.J.1) there exists a constant C' such that

$$\mathrm{P}\left(\mathrm{I}_{a,1} > C'\sqrt{s^2\ln(pqn)/n}\right) = \mathrm{P}\left(\|\widehat{h}-h_*\|_1 > C'\sqrt{s^2\ln(pqn)/n}\right) \leqslant Cn^{-c}.$$

By Assumptions 1.12 and 1.13, the summands appearing in I_{a2} are i.i.d. and bounded. By definition on μ_{k*}, $\mathrm{E}\{[X_k\partial_v\rho(Z,\beta_*,W^\top h_*) - W^\top\mu_{k*}]W_j\} = 0$ for all $k \in \{1,\dots,q\}$, so the summands in I_{a1} are also mean-zero. Lemma 1.36 therefore implies that for some constant C' depending only on C_1,

$$\mathrm{P}\left(\mathrm{I}_{a,2} > C'\ln(pqn)\right) \leqslant n^{-1}.$$

Eq. (1.J.4) now follows from the two previous displays and the union bound.

I_b

This step shows that for some C',

$$\mathrm{P}\left(\mathrm{I}_b > C'\sqrt{s^2 \ln^3 (qn)/n}\right) \leqslant Cn^{-c}. \tag{1.J.5}$$

To establish the claim, note that by Hölder's inequality,

$$\begin{aligned}\mathrm{I}_b &\equiv \max_{1\leqslant k\leqslant q} |\sqrt{n}\mathbb{E}_n[(Y_i - W_i^\top h_*)W_i^\top(\widehat{\mu}_k - \mu_{k*})]| \\ &\leqslant \max_{1\leqslant j\leqslant q} \|\widehat{\mu}_k - \mu_{k*}\|_1 \max_{1\leqslant j\leqslant p} |\sqrt{n}\mathbb{E}_n[(Y_i - W_i^\top h_*)W_{ik}]| \\ &=: \mathrm{I}_{b,1} \times \mathrm{I}_{b,2}.\end{aligned}$$

Equation (1.J.1) shows that there exists a constant C' such that

$$\mathrm{P}\left(\mathrm{I}_{b1} > C'\sqrt{s^2 \ln (pqn)/n}\right) = \mathrm{P}\left(\max_{1\leqslant k\leqslant q} \|\widehat{\mu}_k - \mu_{k*}\|_1 > C'\sqrt{s^2 \ln (pqn)/n}\right) \leqslant Cn^{-c}.$$

Assumptions 1.12 and 1.13 and the definition of h_* show that the summands appearing in I_{b2} are i.i.d., mean-zero and bounded. Lemma 1.36 therefore implies that for some constant C' depending only on C_1,

$$\mathrm{P}\left(\mathrm{I}_{b,2} > C' \ln (pqn)\right) \leqslant n^{-1}.$$

Equation (1.J.5) now follows from the two previous displays and the union bound.

I_c This step shows that for some C',

$$\mathrm{P}\left(\mathrm{I}_c > C'\sqrt{s^2 \ln^2 (pqn)/n}\right) \leqslant Cn^{-c}. \tag{1.J.6}$$

To establish the claim, note that by Assumptions 1.12 and 1.13 and MVT

$$\begin{aligned}\mathrm{I}_c &\equiv \sqrt{n} \max_{1\leqslant k\leqslant q} |\mathbb{E}_n\{X_{ik}[\partial_v\rho_i(W_i^\top \overline{h}_i^{(k)}) - \partial_v\rho_i(W_i^\top h_*)]W_i^\top(\widehat{h} - h_*)\}| \\ &\leqslant \sqrt{n}\mathbb{E}_n[|W_i^\top(\widehat{h} - h_*)| \max_j |X_{ik}||\partial_v\rho_i(W_i^\top \overline{h}_i^{(k)}) - \partial_v\rho_i(W_i^\top h_*)|] \\ &\leqslant C_1\sqrt{n}\mathbb{E}_n[\max_{1\leqslant k\leqslant q} |\partial_v\rho_i(W_i^\top \overline{h}_i^{(k)}) - \partial_v\rho_i(W_i^\top h_*)||W_i^\top(\widehat{h} - h_*)|] \\ &\leqslant C'\sqrt{n}\mathbb{E}_n[\max_{1\leqslant k\leqslant q} |W_i^\top(\overline{h}_i^{(k)} - h_*)||W_i^\top(\widehat{h} - h_*)|] \\ &\leqslant C'\sqrt{n}\mathbb{E}_n\{[W_i^\top(\widehat{h} - h_*)]^2\} = C'\sqrt{n}\|\widehat{L} - L_*\|^2_{\mathbb{P}_n,2}.\end{aligned}$$

Equation (1.J.1) implies that wp $\geqslant 1-Cn^{-c}$, $\|\widehat{L}-L_*\|_{\mathbb{P}_n,2}^2 \leqslant C's\ln(pqn)/n$, (1.J.6) follows from the previous display.

I_e and I_e

These steps show that for some C',

$$\mathrm{P}\left(\mathrm{I}_d+\mathrm{I}_e > C'\sqrt{s^2\ln^2(pqn)/n}\right) \leqslant Cn^{-c}. \tag{1.J.7}$$

To establish the claim, note that by CS,

$$\begin{aligned}
\mathrm{I}_d &\equiv \sqrt{n}\max_{1\leqslant k\leqslant q}|\mathbb{E}_n[W_i^\top(\overline{\mu}_k^{(k)}-\mu_{k*})W_i^\top(\widehat{h}-h_*)]| \\
&\leqslant \sqrt{n}\max_{1\leqslant k\leqslant q}\mathbb{E}_n[|W_i^\top(\overline{\mu}_k^{(k)}-\mu_{*j})W_i^\top(\widehat{h}-h_*)|] \\
&\leqslant \sqrt{n}(\mathbb{E}_n\{[W_i^\top(\widehat{h}-h_*)]^2\})^{1/2}\max_{1\leqslant k\leqslant q}(\mathbb{E}_n\{[W_i^\top(\overline{\mu}_k^{(k)}-\mu_{k*})]^2\})^{1/2} \\
&\leqslant \sqrt{n}(\mathbb{E}_n\{[W_i^\top(\widehat{h}-h_*)]^2\})^{1/2}\max_{1\leqslant k\leqslant q}(\mathbb{E}_n\{[W_i^\top(\widehat{\mu}_k-\mu_{k*})]^2\})^{1/2} && \text{(MVT)} \\
&= \sqrt{n}\|\widehat{L}-L_*\|_{\mathbb{P}_n,2}\max_{1\leqslant k\leqslant q}\|\widehat{L}_k-L_{k*}\|_{\mathbb{P}_n,2}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathrm{I}_e &\equiv \sqrt{n}\max_{1\leqslant k\leqslant q}|\mathbb{E}_n[W_i^\top(\widehat{\mu}_k-\mu_{k*})W_i^\top(\overline{h}^{(k)}-h_*)]| \\
&\leqslant \sqrt{n}\max_{1\leqslant k\leqslant q}\mathbb{E}_n[|W_i^\top(\widehat{\mu}_j-\mu_{*j})W_i^\top(\overline{h}^{(k)}-h_*)|] \\
&\leqslant \sqrt{n}\max_{1\leqslant k\leqslant q}(\mathbb{E}_n\{[W_i^\top(\widehat{\mu}_k-\mu_{k*})]^2\})^{1/2}(\mathbb{E}_n\{[W_i^\top(\overline{h}^{(k)}-h_*)]^2\})^{1/2} \\
&\leqslant \sqrt{n}(\mathbb{E}_n\{[W_i^\top(\widehat{h}-h_*)]^2\})^{1/2}\max_{1\leqslant k\leqslant q}(\mathbb{E}_n\{[W_i^\top(\widehat{\mu}_k-\mu_{k*})]^2\})^{1/2} && \text{(MVT)} \\
&= \sqrt{n}\|\widehat{L}-L_*\|_{\mathbb{P}_n,2}\max_{1\leqslant k\leqslant q}\|\widehat{L}_k-L_{k*}\|_{\mathbb{P}_n,2},
\end{aligned}$$

so combining we get

$$\mathrm{I}_d+\mathrm{I}_e \leqslant 2\sqrt{n}\|\widehat{L}-L_*\|_{\mathbb{P}_n,2}\max_{1\leqslant k\leqslant q}\|\widehat{L}_k-L_{k*}\|_{\mathbb{P}_n,2}.$$

The claim now follows from (1.J.1) and the union bound.

II

This claim shows that

$$\mathrm{P}\left(\mathrm{II} > C' \max\left\{\sqrt{s\ln^2(pqn)/n}, n^{-c_2/4}/\sqrt{\ln(pqn)}, a_n\right\}\right) \leqslant Cn^{-c}. \tag{1.J.8}$$

Noting that $\partial_\beta \psi_k(z, \beta, w^\top h, w^\top \mu_k) = x_j \partial_\beta \rho(z, \beta, w^\top h)$, a mean-value expansion yields

$$\begin{aligned}
\mathrm{II} &\equiv \max_{1\leqslant k\leqslant q} |\sqrt{n}\mathbb{E}_n[\psi_k(Z_i, \widehat{\beta}, \widehat{L}(W_i), \widehat{L}_k(W_i)) - \psi_j(Z_i, \beta_*, \widehat{L}(W_i), \widehat{L}_k(W_i))] - b_k^\top \sqrt{n}\mathbb{E}_n(s_i)| \\
&= \max_{1\leqslant k\leqslant q} |\mathbb{E}_n[X_{ik}\partial_{\beta^\top}\rho(Z_i, \overline{\beta}^{(k)}, \widehat{L}(W_i))]\sqrt{n}(\widehat{\beta} - \beta_*) - b_k^\top \sqrt{n}\mathbb{E}_n(s_i)| \\
&\leqslant \max_{1\leqslant k\leqslant q} |\mathbb{E}_n\{X_{ik}[\partial_{\beta^\top}\rho(Z_i, \overline{\beta}^{(k)}, \widehat{L}(W_i)) - \partial_{\beta^\top}\rho(Z_i, \beta_*, L_*(W_i))]\}\sqrt{n}(\widehat{\beta} - \beta_*)| \\
&\quad + \max_{1\leqslant k\leqslant q} |\{(\mathbb{E}_n - \mathrm{E})[X_{ik}\partial_\beta \rho(Z_i, \beta_*, L_*(W_i))]\}^\top \sqrt{n}(\widehat{\beta} - \beta_*)| \\
&\quad + \max_{1\leqslant j\leqslant q} |b_k^\top[\sqrt{n}(\widehat{\beta} - \beta_*) - \sqrt{n}\mathbb{E}_n(s_i)]| =: \mathrm{II}_a + \mathrm{II}_b + \mathrm{II}_c.
\end{aligned}$$

Eq. (1.J.8) now follows from Steps II_a–II_c below and the union bound.

II_a

This step shows that for some c, C and C',

$$\mathrm{P}\left(\mathrm{II}_a > C'\sqrt{s\ln^2(pqn)/n}\right) \leqslant Cn^{-c}. \tag{1.J.9}$$

To establish the claim, note that by CS,

$$\begin{aligned}
\mathrm{II}_a &\leqslant \|\sqrt{n}(\widehat{\beta} - \beta_*)\| \max_{1\leqslant j\leqslant q} \|\mathbb{E}_n\{X_{ik}[\partial_{\beta^\top}\rho(Z_i, \overline{\beta}^{(k)}, W_i^\top \widehat{h}) - \partial_{\beta^\top}\rho(Z_i, \beta_*, W_i^\top h_*)]\}\| \\
&=: \mathrm{II}_{a,1} \times \mathrm{II}_{a,2}.
\end{aligned} \tag{1.J.10}$$

Lemma 1.27 implies that for some c, C and C',

$$\mathrm{P}\left(\mathrm{II}_{a,1} > C'\sqrt{\ln n}\right) = \mathrm{P}\left(\|\sqrt{n}(\widehat{\beta} - \beta_*)\| > C'\sqrt{\ln n}\right) \leqslant Cn^{-c}. \tag{1.J.11}$$

By Assumptions 1.10, 1.12 and 1.13, MVT and CS,

$$\begin{aligned}
\mathrm{II}_{a,2} &\leqslant \mathbb{E}_n\{\max_{1\leqslant k\leqslant q} |X_{ij}| \, \|\partial_\beta \rho(Z_i, \overline{\beta}^{(k)}, W_i^\top \widehat{h}) - \partial_\beta \rho(Z_i, \beta_*, W_i^\top h_*)\|\} \\
&\leqslant C'(\max_k \|\overline{\beta}^{(k)} - \beta_*\| + \mathbb{E}_n[|W_i(\widehat{h} - h_*)|]) \leqslant C'(\|\widehat{\beta} - \beta_*\| + \|\widehat{L} - L_*\|_{\mathbb{P}_n,2})
\end{aligned}$$

From Lemma 1.27 we know that wp $\geqslant 1 - Cn^{-c}$, $\|\sqrt{n}(\widehat{\beta} - \beta_*)\| \leqslant C'\sqrt{\ln n}$ for some c, C and C'. Equation (1.J.1) implies that wp $\geqslant 1 - n^{-1}$, $\|\widehat{h} - h_*\|_{2,n} \leqslant C'[s\ln(pqn)/n]^{1/2}$ for some C'. The union bound therefore shows that wp $\geqslant 1 - Cn^{-c}$, $\|\widehat{\beta} - \beta_*\| + \|\widehat{h} - h_*\|_{\mathbb{P}_n,2} \leqslant C'[s\ln(pqn)/n]^{1/2}$ for some c, C and C'. The previous display therefore implies that, for some c, C and C',

$$\mathrm{P}\left(\mathrm{II}_b > C'\sqrt{s\ln(pqn)/n}\right) \leqslant Cn^{-c}. \tag{1.J.12}$$

Eq. (1.J.9) now follows from (1.J.10), (1.J.11) and (1.J.12) and the union bound.

II_b

This step shows that for some c, C and C',

$$\mathrm{P}\left(\mathrm{II}_b > C'n^{-c_2/4}/\sqrt{\ln(pqn)}\right) \leqslant Cn^{-c}. \tag{1.J.13}$$

To establish the claim, note that by Assumption 1.10,

$$\begin{aligned}
\mathrm{II}_b &\equiv \max_{1\leqslant k\leqslant q}|\{(\mathbb{E}_n - \mathrm{E})\left[X_{ik}\partial_\beta\rho(Z_i, \beta_*, W_i^\top h_*)\right]\}^\top\sqrt{n}(\widehat{\beta} - \beta_*)| \\
&\leqslant \mathrm{II}_{a,1} \times \max_{1\leqslant k\leqslant q}\|(\mathbb{E}_n - \mathrm{E})\left[X_{ik}\partial_\beta\rho(Z_i, \beta_*, W_i^\top h_*)\right]\| \\
&\leqslant \sqrt{C_1}\mathrm{II}_{a,1} \times \max_{(j,k)\in[d]\times[q]}|(\mathbb{E}_n - \mathrm{E})\left[X_{ik}\partial_{\beta_j}\rho(Z_i, \beta_*, W_i^\top h_*)\right]| =: \sqrt{C_1}\mathrm{II}_{a,1} \times \mathrm{II}_{b,1}. \quad (1.J.14)
\end{aligned}$$

Assumptions 1.12 and 1.13 imply that the summands appearing in $\mathrm{II}_{b,1}$ are bounded. Lemma 1.34 therefore implies that

$$\mathrm{E}\left[\max_{(j,k)\in[d]\times[q]}|(\mathbb{E}_n - \mathrm{E})\left[X_{ik}\partial_{\beta_j}\rho(Z_i, \beta_*, W_i^\top h_*)\right]|\right] \leqslant C'\frac{\ln(dq)}{\sqrt{n}} \leqslant C''\frac{\ln(pqn)}{\sqrt{n}}.$$

By M, for some C and C', and the $c_2 > 0$ provided by Assumption 1.14,

$$\mathrm{P}\left(\mathrm{II}_{b,1} > C'n^{-c_2/4}/\ln(pqn)\right) \leqslant C''n^{c:2/4}\sqrt{\frac{\ln^4(pqn)}{n}} \leqslant Cn^{c_2/4-c_2/2} = Cn^{-c_2/4}. \tag{1.J.15}$$

Eq. (1.J.13) now follows from (1.J.11), (1.J.14), and (1.J.15) and the union bound.

II_c

This step shows that for some c, C and C',

$$\mathrm{P}\left(\mathrm{II}_c > C' a_n\right) \leqslant C n^{-c}. \tag{1.J.16}$$

To establish the claim, note that

$$\mathrm{II}_c \equiv \max_{1 \leqslant k \leqslant q} |b_k^\top [\sqrt{n}(\widehat{\beta} - \beta_*) - \sqrt{n}\mathbb{E}_n\left(s_i\right)]| \leqslant \|\sqrt{n}(\widehat{\beta} - \beta_*) - \sqrt{n}\mathbb{E}_n\left(s_i\right)\| \max_{1 \leqslant k \leqslant q} \|b_k\|$$

Assumption 1.10 implies that for some c, C and positive sequence a_n

$$\mathrm{P}\left(\|\sqrt{n}(\widehat{\beta} - \beta_*) - \sqrt{n}\mathbb{E}_n\left(s_i\right)\| > a_n\right) \leqslant C n^{-c}.$$

Moreover, by Assumptions 1.12 and 1.13,

$$\max_{1 \leqslant k \leqslant q} \|b_j\| \leqslant \mathrm{E}\Big[\max_{1 \leqslant k \leqslant q} |X_k| \, \|\partial_\beta \rho(Z, \beta_*, W^\top h_*)\|\Big] \leqslant C'.$$

Eq. (1.J.16) now follows from the three previous displays and the union bound. □

1.J.2 Proofs for Section 1.5.5

Proof of Lemma 1.6. The proof is divided into steps. Throughout the proof I let P_ξ abbreviate the conditional law $\mathrm{P}_\xi\left(\cdot\right) \coloneqq \mathrm{P}\left(\cdot \middle| \left\{Z_i\right\}_1^n\right)$.

Main

Recall that

$$\mathcal{W}_* \equiv \max_{1 \leqslant k \leqslant q} |\sqrt{n}\mathbb{E}_n\{[\psi_k(Z_i, \beta_*, L_*\left(W\right), L_{k*}\left(W\right)) + b_j^\top s_i]\xi_i\}|,$$
$$\mathcal{W} \equiv \max_{1 \leqslant k \leqslant q} |\sqrt{n}\mathbb{E}_n\{[\psi_j(Z_i, \widehat{\beta}, \widehat{L}\left(W_i\right), \widehat{L}_k\left(W_i\right)) + \widehat{b}_k^\top \widehat{s}_i]\xi_i\}|.$$

where

$$\widehat{b}_j \coloneqq \mathbb{E}_n[X_{ij} \partial_\beta \rho(Z_i, \widehat{\beta}, W_i^\top \widehat{h})]$$

and $\widehat{s}$ is an estimator of s provided by Assumption 1.15. Adding and subtracting $\widehat{b}_k^\top \sqrt{n}\mathbb{E}_n\left(s_i\xi_i\right)$ and $\psi_k(Z_i,\beta_*,\widehat{L}\left(W_i\right),\widehat{L}_k\left(W_i\right))\xi_i$, by T and a MVE we get

$$\begin{aligned}
|\mathcal{W}-\mathcal{W}_*| \leqslant{} & \max_{1\leqslant k\leqslant q}\big|\sqrt{n}\mathbb{E}_n\big\{[\psi_k(Z_i,\widehat{\beta},\widehat{L}\left(W_i\right),\widehat{L}_k\left(W_i\right))+\widehat{b}_j^\top\widehat{s}_i]\xi_i \\
& \quad -[\psi_k(Z_i,\beta_*,L_*\left(W_i\right),L_{k*}\left(W_i\right))+b_{k*}^\top s_i]\xi_i\big\}\big| \\
={} & \max_{1\leqslant k\leqslant q}\big|\sqrt{n}\mathbb{E}_n\{[\psi_k(Z_i,\beta_*,\widehat{L}\left(W_i\right),\widehat{L}_k\left(W_i\right))-\psi_k(Z_i,\beta_*,L_*\left(W_i\right),L_{k*}\left(W_i\right))]\xi_i\} \\
& \qquad +\mathbb{E}_n\{\xi_iX_{ik}\partial_{\beta^\top}\rho(Z_i,\overline{\beta}^{(k)},\widehat{L}\left(W_i\right))\}\sqrt{n}(\widehat{\beta}-\beta_*) \\
& \qquad +(\widehat{b}_k-b_{k*})^\top\sqrt{n}\mathbb{E}_n\left(s_i\xi_i\right)+\widehat{b}_k^\top\sqrt{n}\mathbb{E}_n[(\widehat{s}_i-s_i)\xi_i]\big| \\
\leqslant{} & \max_{1\leqslant k\leqslant q}|\sqrt{n}\mathbb{E}_n\{[\psi_j(Z_i,\beta_*,\widehat{L}\left(W_i\right),\widehat{L}_k\left(W_i\right))-\psi_j(Z_i,\beta_*,L_*\left(W_i\right),L_{k*}\left(W_i\right))]\xi_i\}| \\
& +\max_{1\leqslant k\leqslant q}|\mathbb{E}_n\{\xi_iX_{ik}\partial_{\beta^\top}\rho(Z_i,\overline{\beta}^{(k)},\widehat{L}\left(W_i\right))\}\sqrt{n}(\widehat{\beta}-\beta_*)| \\
& +\max_{1\leqslant k\leqslant q}|(\widehat{b}_k-b_{k*})^\top\sqrt{n}\mathbb{E}_n\left(s_i\xi_i\right)|+\max_{1\leqslant k\leqslant q}|\widehat{b}_k^\top\sqrt{n}\mathbb{E}_n[(\widehat{s}_i-s_i)\xi_i]| \\
=:{} & \mathrm{I}+\mathrm{II}+\mathrm{III}+\mathrm{IV}. \qquad (1.J.17)
\end{aligned}$$

The claim of the lemma now follows from Steps I–IV below and the union bound.

I

This step shows that for some C',

$$\mathrm{P}\left(\mathrm{P}_\xi\Big(\mathrm{I}>C'\Big[\sqrt{s\ln^2\left(pqn\right)/n}\vee s^{3/2}\ln^{3/2}\left(pqn\right)/n\Big]\Big)>Cn^{-c}\right)\leqslant Cn^{-c}. \qquad (1.J.18)$$

A MVE and T shows that

$$\begin{aligned}
\mathrm{I}\equiv{} & \max_{1\leqslant k\leqslant q}|\sqrt{n}\mathbb{E}_n\{[\psi_k(Z_i,\beta_*,\widehat{L}\left(W_i\right),\widehat{L}_k\left(W_i\right))-\psi_j(Z_i,\beta_*,L_*\left(W_i\right),L_{k*}\left(W_i\right))]\xi_i\}| \\
\leqslant{} & \max_{1\leqslant k\leqslant q}|\sqrt{n}\mathbb{E}_n\{[X_{ik}\partial_v\rho_i(W_i^\top h_*)X_{ik}-W_i^\top\mu_{k*}]W_i^\top(\widehat{h}-h_*)\xi_i\}| \\
& +\max_{1\leqslant k\leqslant q}|\sqrt{n}\mathbb{E}_n[(Y_i-W_i^\top h_*)W_i^\top(\widehat{\mu}_k-\mu_{k*})\xi_i]| \\
& +\max_{1\leqslant k\leqslant q}|\sqrt{n}\mathbb{E}_n\{[\partial_v\rho_i(W_i^\top\overline{h}^{(k)})-\partial_v\rho_i(W_i^\top h_*)]X_{ik}W_i^\top(\widehat{h}-h_*)\xi_i\}| \\
& +\max_{1\leqslant k\leqslant q}|\sqrt{n}\mathbb{E}_n[W_i^\top(\overline{\mu}^{(k)}-\mu_{k*})W_i^\top(\widehat{h}-h_*)\xi_i]| \\
& +\max_{1\leqslant k\leqslant q}|\sqrt{n}\mathbb{E}_n[W_i^\top(\widehat{\mu}_k-\mu_{k*})W_i^\top(\overline{h}^{(k)}-h_*)\xi_i]|. \\
=:{} & \mathrm{I}_a+\mathrm{I}_b+\mathrm{I}_c+\mathrm{I}_d+\mathrm{I}_e, \qquad (1.J.19)
\end{aligned}$$

where each $W_i^\top \overline{h}^{(k)}$ lies on the line segment connecting $W_i^\top \widehat{h}$ and $W_i^\top h_*$, and each $W_i^\top \overline{\mu}_k^{(k)}$ lies on the line segment connecting $W_i^\top \widehat{\mu}_k$ and $W_i^\top \mu_{k*}$. Equation (1.J.18) now follows from Steps I_a–I_e below and the union bound.

I_a

This step shows that for some C',

$$\mathrm{P}\left(\mathrm{P}_\xi\left(\mathrm{I}_a > C'\sqrt{s\ln^2(pqn)/n}\right) > Cn^{-c}\right) \leqslant Cn^{-c}. \tag{1.J.20}$$

Recall that

$$\mathrm{I}_a \equiv \max_{1\leqslant k\leqslant q}|\sqrt{n}\mathbb{E}_n\{[X_{ik}\partial_v\rho_i(W_i^\top h_*) - W_i^\top \mu_{k*}]W_i^\top(\widehat{h}-h_*)\xi_i\}|$$

Conditional on the data, $\{\sqrt{n}\mathbb{E}_n\{[X_{ik}\partial_v\rho_i(W_i^\top h_*) - W_i^\top \mu_{k*}]W_i^\top(\widehat{h}-h_*)\xi_i\}\}_{k=1}^q$ is centered Gaussian with maximal variance given by

$$\sigma^2 = \max_{1\leqslant k\leqslant q}\mathbb{E}_n\{[X_{ik}\partial_v\rho_i(W_i^\top h_*) - W_i^\top \mu_{k*}]^2[W_i^\top(\widehat{h}-h_*)]^2\} \leqslant C_1^2\|\widehat{L}-L_*\|_{\mathbb{P}_n,2}^2,$$

where I have used Assumption 1.13. Lemma 1.41 now implies that that for some C'

$$\mathrm{P}_\xi\left(\mathrm{I}_a > C'\|\widehat{L}-L_*\|_{\mathbb{P}_n,2}\sqrt{\ln(qn)}\right) \leqslant n^{-1}.$$

Thus, on the event $\mathcal{E} \coloneqq \{\|\widehat{L}-L_*\|_{\mathbb{P}_n,2} \leqslant C'[s\ln(pqn)/n]^{1/2}\}$, for some C',

$$\mathrm{P}_\xi\left(\mathrm{I}_a > C'\sqrt{s\ln^2(pqn)/n}\right) \leqslant n^{-1}.$$

Eq. (1.J.20) now follows from (1.J.1) and the previous display.

I_b

This step shows that for some C',

$$\mathrm{P}\left(\mathrm{P}_\xi\left(\mathrm{I}_b > C'\sqrt{s\ln^2(pqn)/n}\right) > n^{-1}\right) \leqslant n^{-1}. \tag{1.J.21}$$

Recall that

$$\mathrm{I}_b \equiv \max_{1\leqslant k\leqslant q}|\sqrt{n}\mathbb{E}_n[(Y_i - W_i^\top h_*)W_i^\top(\widehat{\mu}_k - \mu_{k*})\xi_i]|.$$

Conditional on the data, $\{\sqrt{n}\mathbb{E}_n[(Y_i - W_i^\top h_*)W_i^\top(\widehat{\mu}_k - \mu_{k*})\xi_i]\}_{k=1}^q$ is centered Gaussian with maximal variance given by

$$\sigma^2 = \max_{1\leqslant k\leqslant q} \mathbb{E}_n\{[Y_i - W_i^\top h_*]^2[W_i^\top(\widehat{\mu}_k - \mu_{k*})]^2\} \leqslant C' \max_{1\leqslant k\leqslant q}\|\widehat{L}_k - L_{k*}\|_{\mathbb{P}_n,2}^2,$$

where I have used Assumption 1.12. Lemma 1.41 therefore implies that for some C',

$$\mathrm{P}_\xi\left(\mathrm{I}_b > C' \max_{1\leqslant k\leqslant q}\|\widehat{L}_k - L_{k*}\|_{\mathbb{P}_n,2}\sqrt{\ln(pqn)}\right) \leqslant n^{-1}.$$

Hence, on the event $\mathcal{E} \coloneqq \{\max_{1\leqslant k\leqslant q}\|\widehat{L}_k - L_{k*}\|_{\mathbb{P}_n,2} \leqslant C'[s\ln(pqn)/n]^{1/2}\}$, it follows that for some C',

$$\mathrm{P}_\xi\left(\mathrm{I}_b > C'\sqrt{s\ln^2(pqn)/n}\right) \leqslant n^{-1}.$$

Eq. (1.J.21) now follows from (1.J.1) and the previous display.

I_c

This step shows that for some C',

$$\mathrm{P}\left(\mathrm{P}_\xi\left(\mathrm{I}_c > C's^{3/2}\ln(qn)/n\right) > Cn^{-c}\right) \leqslant Cn^{-c}. \tag{1.J.22}$$

To establish the claim, first recall

$$\mathrm{I}_c \equiv \max_{1\leqslant k\leqslant q}|\sqrt{n}\mathbb{E}_n\{X_{ik}[\partial_v\rho_i(W_i^\top\overline{h}^{(k)}) - \partial_v\rho_i(W_i^\top h_*)]W_i^\top(\widehat{h} - h_*)\xi_i\}|.$$

Conditional on the data, $\{\sqrt{n}\mathbb{E}_n\{X_{ik}[\partial_v\rho_i(W_i^\top\overline{h}^{(k)}) - \partial_v\rho_i(W_i^\top h_*)]W_i^\top(\widehat{h} - h_*)\xi_i\}\}_{k=1}^q$ is centered Gaussian with maximal variance given by

$$\begin{aligned}\sigma^2 &= \max_{1\leqslant k\leqslant q}\mathbb{E}_n\{X_{ik}^2[\partial_v\rho_i(W_i^\top\overline{h}^{(k)}) - \partial_v\rho_i(W_i^\top h_*)][W_i^\top(\widehat{h} - h_*)]^2\}\\ &= \mathbb{E}_n\{[W_i^\top(\widehat{h} - h_*)]^2\max_{1\leqslant k\leqslant q}X_{ik}^2[\partial_v\rho_i(W_i^\top\overline{h}^{(k)}) - \partial_v\rho_i(W_i^\top h_*)]\}\\ &\leqslant C'\mathbb{E}_n\{[W_i^\top(\widehat{h} - h_*)]^2[W_i^\top(\overline{h}^{(k)} - h_*)]^2\} \leqslant C'\mathbb{E}_n\{[W_i^\top(\widehat{h} - h_*)]^4\}\\ &\leqslant C'\|\widehat{h} - h_*\|_1^2\mathbb{E}_n\{\|W_i\|_\infty^2[W_i^\top(\widehat{h} - h_*)]^2\} \leqslant C''\|\widehat{h} - h_*\|_1^2\|\widehat{L} - L_*\|_{\mathbb{P}_n,2}^2,\end{aligned}$$

where I have used H, MVT and Assumptions 1.12 and 1.13. Lemma 1.41 now implies that for some C',

$$\mathrm{P}_{\xi}\left(\mathrm{I}_c > C' \|\widehat{h} - h_*\|_1 \|\widehat{L} - L_*\|_{\mathbb{P}_n,2} \sqrt{\ln{(qn)}}\right) \leqslant n^{-1}.$$

Hence, on the event $\mathcal{E} \coloneqq \{\|\widehat{h} - h_*\|_1 \|\widehat{L} - L_*\|_{\mathbb{P}_n,2} \leqslant C_1^2 s^{3/2} \ln{(pqn)}/n\}$, it follows that for some C',

$$\mathrm{P}_{\xi}\left(\mathrm{I}_c > C' s^{3/2} \ln{(pqn)}/n\right) \leqslant n^{-1}.$$

Eq. (1.J.22) now follows from (1.J.1) and the previous display.

I_d and I_e

This step shows that

$$\mathrm{P}\left(\mathrm{P}_{\xi}\left(\mathrm{I}_d + \mathrm{I}_e > C' s^{3/2} \ln^{3/2}{(pqn)}/n\right) > Cn^{-c}\right) \leqslant Cn^{-c}. \tag{1.J.23}$$

Recall that

$$\begin{aligned}
\mathrm{I}_d &\equiv \max_{1\leqslant k\leqslant q} |\sqrt{n}\mathbb{E}_n[W_i^\top(\overline{\mu}^{(k)} - \mu_{k*})W_i^\top(\widehat{h} - h_*)\xi_i]|, \\
\mathrm{I}_e &\equiv \max_{1\leqslant j\leqslant q} |\sqrt{n}\mathbb{E}_n[W_i^\top(\widehat{\mu}_k - \mu_{k*})W_i^\top(\overline{h}^{(k)} - h_*)\xi_i]|.
\end{aligned}$$

Consider first I_d. Conditional on the data, $\{\sqrt{n}\mathbb{E}_n[W_i^\top(\overline{\mu}^{(k)} - \mu_{k*})W_i^\top(\widehat{h} - h_*)\xi_i]\}_{k=1}^q$ is centered Gaussian with maximal variance given by

$$\begin{aligned}
\sigma^2 &= \max_{1\leqslant k\leqslant q} \mathbb{E}_n\{[W_i^\top(\widehat{h} - h_*)]^2[W_i^\top(\overline{\mu}^{(k)} - \mu_{k*})]^2\} \leqslant \max_{1\leqslant k\leqslant q} \mathbb{E}_n\{[W_i^\top(\widehat{h} - h_*)]^2[W_i^\top(\widehat{\mu}_k - \mu_{k*})]^2\} \\
&\leqslant \|\widehat{h} - h_*\|_1^2 \max_{1\leqslant k\leqslant q} \mathbb{E}_n\{\|W_i\|_\infty^2[W_i^\top(\widehat{\mu}_k - \mu_{k*})]^2\} \leqslant C'\|\widehat{h} - h_*\|_1^2 \max_{1\leqslant k\leqslant q} \|\widehat{L}_k - L_{k*}\|_{\mathbb{P}_n,2}^2,
\end{aligned}$$

where I have used MVT, H and Assumption 1.12. Lemma 1.41 implies that for some C',

$$\mathrm{P}_{\xi}\left(\mathrm{I}_d > C'\|\widehat{h} - h_*\|_1 \max_{1\leqslant k\leqslant q} \|\widehat{L}_k - L_{k*}\|_{\mathbb{P}_n,2} \sqrt{\ln{(qn)}}\right) \leqslant n^{-1}. \tag{1.J.24}$$

Consider next I_e. Conditional on the data, $\{\sqrt{n}\mathbb{E}_n[W_i^\top(\widehat{\mu}_k - \mu_{k*})W_i^\top(\overline{h}^{(k)} - h_*)\xi_i]\}_{k=1}^q$ is centered Gaussian with maximal variance given by

$$\sigma^2 = \max_{1\leqslant k\leqslant q} \mathbb{E}_n\{[W_i^\top(\overline{h}_i^{(j)} - h_*)]^2[W_i^\top(\widehat{\mu}_j - \mu_{*j})]^2\} \leqslant \max_{1\leqslant k\leqslant q} \mathbb{E}_n\{[W_i^\top(\widehat{h} - h_*)]^2[W_i^\top(\widehat{\mu}_j - \mu_{*j})]^2\}$$

$$\leqslant \|\widehat{h}-h_*\|_1^2 \max_{1\leqslant k\leqslant q} \mathbb{E}_n\{\|W_i\|_\infty^2[W_i^\top(\widehat{\mu}_j-\mu_{*j})]^2\} \leqslant C'\|\widehat{h}-h_*\|_1^2 \max_{1\leqslant k\leqslant q}\|\widehat{L}_k-L_{k*}\|_{\mathbb{P}_n,2}^2,$$

where I have used H, MVT and Assumption 1.12. Lemma 1.41 implies that for some C',

$$\mathrm{P}_\xi\left(\mathrm{I}_e > C'\|\widehat{h}-h_*\|_1 \max_{1\leqslant k\leqslant q}\|\widehat{L}_k-L_{k*}\|_{\mathbb{P}_n,2}\sqrt{\ln(qn)}\right) \leqslant n^{-1}. \tag{1.J.25}$$

Eqs. (1.J.24) and (1.J.25) imply that, on the event $\mathcal{E} \coloneqq \{\|\widehat{h}-h_*\|_1 \max_{1\leqslant k\leqslant q}\|\widehat{L}_k-L_{k*}\|_{\mathbb{P}_n,2} \leqslant C's^{3/2}\ln(pqn)/n\}$, for some C',

$$\mathrm{P}_e\left(\mathrm{I}_d+\mathrm{I}_e > C's^{3/2}\ln^{3/2}(pqn)/n\right) \leqslant n^{-1}.$$

Eq. (1.J.23) now follows from (1.J.1) in combination with the union bound.

II

This step shows that for some c, C and C',

$$\mathrm{P}\left(\mathrm{P}_\xi\left(\mathrm{II} > C'\sqrt{\ln^2(pqn)/n}\right) > Cn^{-c}\right) \leqslant Cn^{-c}. \tag{1.J.26}$$

To establish the claim, note that

$$\begin{aligned}\mathrm{II} &\equiv \max_{1\leqslant k\leqslant q}|\mathbb{E}_n\{\xi_i X_{ik}\partial_{\beta^\top}\rho(Z_i,\overline{\beta}^{(k)},\widehat{L}(W_i))\}\sqrt{n}(\widehat{\beta}-\beta_*)| \\ &\leqslant \|\widehat{\beta}-\beta_*\| \max_{1\leqslant k\leqslant q}\|\sqrt{n}\mathbb{E}_n\{e_i X_{ij}\partial_\beta\rho(Z_i,\overline{\beta}^{(j)},\widehat{L}(W_i))\}\| =: \mathrm{II}_a \times \mathrm{II}_b.\end{aligned} \tag{1.J.27}$$

By Lemma 1.27 we know that, for some c, C and C',

$$\mathrm{P}\left(\mathrm{II}_a > C'\sqrt{\ln(n)/n}\right) = \mathrm{P}\left(\|\widehat{\beta}-\beta_*\| > C'\sqrt{\ln(n)/n}\right) \leqslant Cn^{-c}. \tag{1.J.28}$$

By H and Assumption 1.12,

$$\begin{aligned}\mathrm{II}_b &\equiv \max_{1\leqslant k\leqslant q}\|\sqrt{n}\mathbb{E}_n\{\xi_i X_{ik}\partial_\beta\rho(Z_i,\overline{\beta}^{(k)},\widehat{L}(W_i))\}\| \\ &\leqslant \sqrt{C_1}\max_{(j,k)\in[d]\times q}|\sqrt{n}\mathbb{E}_n\{\xi_i X_{ik}\partial_{\beta_j}\rho(Z_i,\overline{\beta}^{(k)},\widehat{L}(W_i))\}|.\end{aligned} \tag{A1.10}$$

Conditional on the data, $\{\sqrt{n}\mathbb{E}_n\{e_i X_{ij}\partial_{\beta_k}\rho(Z_i,\overline{\beta}^{(j)},W_i^\top \widehat{h})\}\}_{(j,k)\in[d]\times[q]}$ is a centered Gaussian process with maximal variance given by

$$\begin{aligned}\sigma^2 &= \max_{(j,k)\in[d]\times[q]}\mathbb{E}_n\{X_{ik}^2[\partial_{\beta_j}\rho(Z_i,\overline{\beta}^{(k)},\widehat{L}\,(W_i))]^2\}\\ &\leqslant C_1^2\max_{(j,k)\in[d]\times[q]}\mathbb{E}_n\{[\partial_{\beta_j}\rho(Z_i,\overline{\beta}^{(k)},\widehat{L}\,(W_i))]^2\}\\ &\leqslant C_1^2\Big(\max_{(j,k)\in[d]\times[q]}\mathbb{E}_n\{[\partial_{\beta_j}\rho(Z_i,\beta_*,L_*\,(W_i))]^2\}\\ &\qquad+\max_{(j,k)\in[d]\times[q]}\mathbb{E}_n\{[\partial_{\beta_j}\rho(Z_i,\overline{\beta}^{(j)},\widehat{L}\,(W_i))-\partial_{\beta_j}\rho(Z_i,\beta_*,L_*\,(W_i))]^2\}\Big)\\ &\leqslant C'[1+(\max_j\|\overline{\beta}^{(k)}-\beta_*\|+\|\widehat{L}-L_*\|_{\mathbb{P}_n,2})^2]\leqslant C'[1+(\|\widehat{\beta}-\beta_*\|+\|\widehat{L}-L_*\|_{\mathbb{P}_n,2})^2],\end{aligned}$$

where I have used Assumptions 1.12 and 1.13 and MVT. The two previous displays, Lemma 1.41, $d\leqslant C_1$ imply that, for some C',

$$\mathrm{P}_\xi\left(\mathrm{II}_b>C'(1+\|\widehat{\beta}-\beta_*\|+\|\widehat{L}-L_*\|_{\mathbb{P}_n,2})\sqrt{\ln{(qn)}}\right)\leqslant n^{-1}.$$

Let $\mathcal{E}\,(t)\coloneqq\{\|\widehat{\beta}-\beta_*\|+\|\widehat{h}-h_*\|_{2,n}\leqslant t\}$. Then Lemma 1.27, (1.J.1), and the union bound show that for some c,C and C',

$$\mathrm{P}\left(\|\widehat{\beta}-\beta_*\|+\|\widehat{L}-L_*\|_{\mathbb{P}_n,2}>C'\right)\leqslant Cn^{-c}.$$

The two previous displays therefore show that, for some c,C and C',

$$\mathrm{P}\left(\mathrm{P}_\xi\left(\mathrm{II}_b>C'\sqrt{\ln{(qn)}}\right)>n^{-1}\right)\leqslant\mathrm{P}\left(\mathcal{E}\,(C')^c\right)\leqslant Cn^{-c}.\tag{1.J.29}$$

Eq. (1.J.26) now follows from (1.J.27), (1.J.28), (1.J.29) and the union bound.

III

This step shows that for some c,C and C',

$$\mathrm{P}\left(\mathrm{P}_\xi\Big(\mathrm{III}>C'\Big[\sqrt{\frac{s\ln^2{(pqn)}}{n}}\vee\frac{n^{-c_2/4}}{\sqrt{\ln{(pqn)}}}\Big]\Big)>Cn^{-c}\right)\leqslant Cn^{-c}.\tag{1.J.30}$$

To establish the claim, note that

$$\mathrm{III}\equiv\max_{1\leqslant k\leqslant q}|(\widehat{b}_k-b_{k*}*)^\top\sqrt{n}\mathbb{E}_n\,(s_i\xi_i)|\leqslant\|\sqrt{n}\mathbb{E}_n\,(s_i\xi_i)\|\max_{1\leqslant k\leqslant q}\|\widehat{b}_k-b_{k*}\|=:\mathrm{III}_a\times\mathrm{III}_b.\tag{1.J.31}$$

Conditional on the data, $\{\sqrt{n}\mathbb{E}_n\left(s_{ij}\xi_i\right)\}_{j=1}^d$ is centered Gaussian with maximal variance given by

$$\sigma^2 \coloneqq \max_{1\leqslant j\leqslant d} \mathrm{var}_\xi\left(\sqrt{n}\mathbb{E}_n\left(s_{ij}\xi_i\right)\right) = \max_{1\leqslant j\leqslant d} \mathbb{E}_n\left(s_{ij}^2\right) \leqslant C',$$

where I have used Assumption (1.10). Lemma 1.41 and $d \leqslant C_1$ now show that for some C', $\mathrm{P}_\xi(\mathrm{III}_a > C'\sqrt{\ln n}) \leqslant n^{-1}$, so

$$\mathrm{P}_\xi\left(\mathrm{III}_a > C'\sqrt{\ln\left(qn\right)}\right) \leqslant n^{-1}. \tag{1.J.32}$$

Hence, for some C',

$$\mathrm{P}\left(\mathrm{P}_\xi\left(\mathrm{III}_a > C'\sqrt{\ln\left(qn\right)}\right) > n^{-1}\right) = 0. \tag{1.J.33}$$

Note also that,

$$\begin{aligned}\mathrm{III}_b \leqslant \max_{1\leqslant k\leqslant q} &\|\mathbb{E}_n\{X_{ik}[\partial_\beta\rho(Z_i,\widehat{\beta},\widehat{L}\left(W_i\right)) - \partial_\beta\rho(Z_i,\beta_*,L_*\left(W_i\right))]\}\| \\ &+ \max_{1\leqslant k\leqslant q}\|\left(\mathbb{E}_n - \mathrm{E}\right)\left[X_{ik}\partial_\beta\rho(Z_i,\beta_*,L_*\left(W_i\right))\right]\| \eqqcolon \mathrm{III}_{b,1} + \mathrm{II}_c.\end{aligned} \tag{1.J.34}$$

For $t > 0$, let $\mathcal{E}\left(t\right)$ denote the event

$$\mathcal{E}\left(t\right) \coloneqq \left\{\|\widehat{\beta} - \beta\| + \|\widehat{h} - h_*\|_{2,n} \leqslant t\sqrt{s\ln\left(pqn\right)/n}\right\}.$$

Then on $\mathcal{E}\left(t\right)$, by Assumptions 1.12 and 1.13,

$$\begin{aligned}\mathrm{III}_{b,1} &\leqslant \mathbb{E}_n\left[\max_{1\leqslant k\leqslant q}\left|X_{ik}\right|\|\partial_\beta\rho(Z_i,\widehat{\beta},\widehat{L}\left(W_i\right)) - \partial_\beta\rho(Z_i,\beta_*,L_*\left(W_i\right))\|\right] \\ &\leqslant C'\left(\|\widehat{\beta} - \beta\| + \|\widehat{L} - L_*\|_{2,n}\right) \leqslant C't\sqrt{s\ln\left(qn\right)/n}.\end{aligned}$$

Lemma 1.27, (1.J.1), and the union bound guarantee that for sufficiently large t as well as some c and C, $\mathrm{P}\left(\mathcal{E}\left(t\right)^c\right) \leqslant Cn^{-c}$. Choose such a t. Then for some c, C and C',

$$\mathrm{P}\left(\mathrm{III}_{b,1} > C'\sqrt{s\ln\left(pqn\right)/n}\right) \leqslant Cn^{-c}. \tag{1.J.35}$$

Eqs. (1.J.34), (1.J.35) and (1.J.16) in combination with the union bound show that for some c, C and C',

$$\mathrm{P}\left(\mathrm{III}_b > C'\Big[\sqrt{\frac{s\ln(pqn)}{n}} \vee \frac{n^{-c_2/4}}{\ln(pqn)}\Big]\right) \leqslant Cn^{-c}. \tag{1.J.36}$$

Eq. (1.J.30) now follows from (1.J.31), (1.J.32), (1.J.36), and the union bound.

IV

This step shows that for some c, C and C',

$$\mathrm{P}\left(\mathrm{P}_\xi\left(\mathrm{IV} > C'b_n\sqrt{\ln n}\right) > n^{-1}\right) \leqslant Cn^{-c}. \tag{1.J.37}$$

To establish the claim, note that by CS,

$$\mathrm{IV} \equiv \max_{1\leqslant k\leqslant q}|\widehat{b}_k^\top\sqrt{n}\mathbb{E}_n[(\widehat{s}_i - s_i)\xi_i]| \leqslant \|\sqrt{n}\mathbb{E}_n[(\widehat{s}_i - s_i)\xi_i]\| \max_{1\leqslant k\leqslant q}\|\widehat{b}_k\| =: \mathrm{IV}_a \times \mathrm{IV}_b.$$

Given that $d \leqslant C_1$,

$$\mathrm{IV}_a \leqslant \sqrt{C_1}\max_{1\leqslant j\leqslant d}|\sqrt{n}\mathbb{E}_n[(\widehat{s}_{ij} - s_{ij})\xi_i]|.$$

Conditional on the data, $\{\sqrt{n}\mathbb{E}_n[(\widehat{s}_{ij} - s_{ij})\xi_i]\}_{j=1}^d$ is centered Gaussian with maximal variance

$$\sigma^2 = \max_{1\leqslant j\leqslant d}\mathbb{E}_n[(\widehat{s}_{ij} - s_{ij})^2] \leqslant \mathbb{E}_n[\|\widehat{s}_i - s_i\|^2] = \|\widehat{s} - s\|_{\mathbb{P}_n,2}^2.$$

Lemma 1.41 and $d \leqslant C_1$ and the two previous displays imply that for some C'

$$\mathrm{P}_\xi\left(\mathrm{IV}_a > C'\|\widehat{s} - s\|_{\mathbb{P}_n,2}\sqrt{\ln n}\right) \leqslant n^{-1}.$$

On the event $\mathcal{E} := \{\|\widehat{s} - s\|_{\mathbb{P}_n,2} \leqslant b_n\}$, with b_n provided by Assumption 1.10, we therefore have that for some C'

$$\mathrm{P}_\xi\left(\mathrm{IV}_a > C'b_n\sqrt{\ln n}\right) \leqslant n^{-1}.$$

so for some c, C and C',

$$\mathrm{P}\left(\mathrm{P}_\xi\left(\mathrm{IV}_a > C'b_n\sqrt{\ln n}\right) > n^{-1}\right) \leqslant \mathrm{P}\left(\mathcal{E}^c\right) \leqslant Cn^{-c}. \tag{1.J.38}$$

Note also that

$$\mathrm{IV}_b \equiv \max_{1\leqslant k\leqslant q} \|\widehat{b}_k\| \leqslant \max_{1\leqslant k\leqslant q} \|b_{k*}\| + \max_{1\leqslant k\leqslant q} \|\widehat{b}_k - b_{k*}\| = \max_{1\leqslant k\leqslant q} \|b_{k*}\| + \mathrm{III}_b.$$

Assumptions 1.12 and 1.13 and J show that

$$\max_{1\leqslant k\leqslant q} \|b_{k*}\| = \max_{1\leqslant k\leqslant q} \|\mathrm{E}[X_k \partial_\beta \rho(Z, \beta_*, L_* (W))]\| \leqslant \max_{1\leqslant k\leqslant q} \mathrm{E}[|X_k| \, \|\partial_\beta \rho(Z, \beta_*, L_* (W))\|] \leqslant C' \tag{1.J.39}$$

By (1.J.36), for some c, C and C',

$$\mathrm{P}\left(\mathrm{III}_b > C'[\sqrt{s \ln (pqn)/n} + n^{-c_2/4}/\ln (pqn)]\right) \leqslant Cn^{-c}.$$

Assumption 1.14 implies that $\sqrt{s \ln (pqn)/n} + n^{-c_2/4}/\ln (pqn)$ is a bounded sequence. The previous display therefore implies that for some c, C and C'

$$\mathrm{P}\left(\mathrm{III}_b > C'\right) \leqslant Cn^{-c}. \tag{1.J.40}$$

Eqs. (1.J.39) and (1.J.40) yield that for some c, C and C',

$$\mathrm{P}\left(\mathrm{IV}_b > C'\right) \leqslant Cn^{-c}. \tag{1.J.41}$$

Eq. (1.J.37) now follows from (1.J.38), (1.J.41) and the union bound. □

1.J.3 Proofs for Section 1.5.6

Proof of Theorem 1.5. The theorem will follow from an application of Chernozhukov, Chetverikov, and Kato (2013, Corollary 3.1(ii)). Assumptions 1.10, 1.12 and 1.13 and the hypotheses of the theorem suffice for their Condition (E.2) with B_n depending only on C_1 (i.e., constant in n). Given that the growth condition $\ln^7 (qn) \leqslant C_2 n^{1-c_2}$ is assumed, it remains to verify their conditions (14) and (15) and their ζ-condition. To this end, first label the maximum of the lower bounds appearing in the (inner) probability statements of Lemmas 1.5 and 1.6 by

$$\zeta_1'' \coloneqq \zeta_1 \vee \zeta_1' = C' \max\left\{\sqrt{\frac{s^2 \ln^3 (pqn)}{n}}, \frac{s^{3/2} \ln^{3/2} (pqn)}{n}, \frac{n^{-c_2/4}}{\sqrt{\ln (pqn)}}, a_n, b_n \sqrt{\ln n}\right\}.$$

Similarly, label the maximum of the probability bounds of Lemmas 1.5 and 1.6 by $\zeta_2'' \coloneqq Cn^{-c}$. Given that $(\zeta_{1,}'', \zeta_2'')$ provide relax the bounds in Lemmas 1.5 and 1.6, by the same lemmas

$$\begin{aligned} \mathrm{P}\left(|T - T_*| > \zeta_1''\right) &\leqslant \zeta_2'', \\ \mathrm{P}\left(\mathrm{P}_\xi\left(|\mathcal{W} - \mathcal{W}_*| > \zeta_1''\right) > \zeta_2''\right) &\leqslant \zeta_2''. \end{aligned}$$

Conditions (14) and (15) of Chernozhukov et al. (2013) therefore holds simultaneously for $n \geqslant n_0$, where n_0 depends only on $c_0, c_0', c_1, C_1, c_2, C_2$ and C_2'. The growth conditions of Assumptions 1.14 and 1.15 imply that ζ_1'' thus defined satisfies $\zeta_1''\sqrt{\ln q} \leqslant C'n^{-c'}$ for c' and C' depending only on $c_0, c_0', c_1, C_1, c_2, C_2$ and C_2', so $\zeta_1''\sqrt{\ln q} + \zeta_2'' \leqslant C'n^{-c'}$ for some c' and C' depending only on $c_0, c_0', c_1, C_1, c_2, C_2$ and C_2'. □

Proof of Theorem 1.6. The proof of Theorem 1.5 shows that $\max_{1\leqslant k\leqslant q} \mathrm{E}[f_{k*}(Z)^2] \leqslant C'^2$ for some C' depending only on C_1. Under the hypothesis $\max_{1\leqslant k\leqslant q} \mathrm{E}[f_{k*}(Z)^2] \geqslant c_1^2$ of the theorem, from arguments similar to the ones used in proving Lemma 1.5 it follows that there exist constants c' and C' depending only on c_1 and C_1 such that $c'/2 \leqslant \max_{1\leqslant k\leqslant q}\{\mathbb{E}_n[\widehat{f}_k(Z_i)^2]\}^{1/2} \leqslant 2C'$ wp $\to$ 1. Define $\widehat{\sigma}_{(q)}^2 \coloneqq \max_{1\leqslant k\leqslant q} \mathbb{E}_n[\widehat{f}_k(Z_i)^2]$ and let $\mathcal{E} \coloneqq \{c'/2 \leqslant \widehat{\sigma}_{(q)} \leqslant 2C'\}$. Conditional on the data, $\{\sqrt{n}\mathbb{E}_n[\widehat{f}_k(Z_i)\xi_i]\}_{k=1}^q$ is a Gaussian process, so Borell's inequality (van der Vaart and Wellner, 1996, Proposition A.2.1) shows that for any $t > 0$,

$$\mathrm{P}_\xi\left(\mathcal{W} > t\right) \leqslant 2\exp\left\{-\frac{t^2}{8\left[\mathrm{E}_\xi\left(\mathcal{W}\right)\right]^2}\right\}.$$

Fix $\alpha \in (0,1)$. Setting $t = c_{\mathcal{W}}(\alpha)$ and rearranging, we get

$$c_{\mathcal{W}}(\alpha) \leqslant \mathrm{E}_\xi\left(\mathcal{W}\right)\sqrt{8\ln\left(2/\alpha\right)}.$$

A Gaussian maximal inequality (Lemma 1.39) shows that $\mathrm{E}_\xi(\mathcal{W}) \lesssim \widehat{\sigma}_{(q)}\sqrt{\ln q}$. Combining this inequality with the previous display, on $\mathcal{E}$, for some A absolute,

$$c_{\mathcal{W}}(\alpha) \leqslant A\widehat{\sigma}_{(q)}\sqrt{\ln q}\sqrt{8\ln\left(2/\alpha\right)} \leqslant \left(2C'\right)\sqrt{8\ln\left(2/\alpha\right)}\sqrt{\ln q} \eqqcolon C_\alpha'\sqrt{\ln q}, \tag{1.J.42}$$

with C_α' depending only on C_1 and α. Given that $\{\sqrt{n}\mathbb{E}_n[\widehat{f}_k(Z_i)\xi_i]\}_{k=1}^q$ is a Gaussian process conditional on the data, for any $t > 0$,

$$\begin{aligned} \mathrm{P}_\xi\left(\mathcal{W} > t\right) &\geqslant \max_{1\leqslant k\leqslant q} \mathrm{P}_\xi\left(\left|\sqrt{n}\mathbb{E}_n[\widehat{f}_k(Z_i)\xi_i]\right| > t\right) = \max_{1\leqslant k\leqslant q} \mathrm{P}_\xi\left(\left|\mathrm{N}\left(0, \widehat{\sigma}_k^2\right)\right| > t\right) \\ &= \mathrm{P}_\xi\left(\left|\mathrm{N}(0, \widehat{\sigma}_{(q)}^2)\right| > t\right) = 2\left[1 - \Phi\left(t/\widehat{\sigma}_{(q)}\right)\right] \end{aligned}$$

Setting $t = c_{\mathcal{W}}(\alpha)$ and rearranging we see that, on $\mathcal{E}$,

$$c_{\mathcal{W}}(\alpha) \geqslant \Phi^{-1}(1-\alpha/2)\,\widehat{\sigma}_{(q)} \geqslant \Phi^{-1}(1-\alpha/2)\,(c'/2) =: c'_{\alpha}, \tag{1.J.43}$$

where c'_{α} depends only on c_1 and α. The proof of Theorem 1.5 also shows that $\mathrm{E}[\max_{1\leqslant k\leqslant q} f_{k*}(Z)^2] \leqslant C'^2$ for some C' depending only on C_1. Now

$$\begin{aligned}
\mathrm{P}(T \leqslant c_{\mathcal{W}}(\alpha)) &= \mathrm{P}(T \leqslant c_{\mathcal{W}}(\alpha) \cap T_* - T \leqslant c_{\mathcal{W}}(\alpha)) + \mathrm{P}(T \leqslant c_{\mathcal{W}}(\alpha) \cap T_* - T > c_{\mathcal{W}}(\alpha)) \\
&\leqslant \mathrm{P}(T_* \leqslant 2c_{\mathcal{W}}(\alpha)) + \mathrm{P}(T_* - T > c_{\mathcal{W}}(\alpha)) \\
&\leqslant \mathrm{P}(T_* \leqslant 2c_{\mathcal{W}}(\alpha)) + \mathrm{P}(|T - T_*| > c_{\mathcal{W}}(\alpha)) =: P_1 + P_2.
\end{aligned}$$

The probability P_1 satisfies

$$\begin{aligned}
P_1 &= \mathrm{P}\left(T_* \leqslant 2c_{\mathcal{W}}(\alpha) \cap c_{\mathcal{W}}(\alpha) \leqslant v_q\sqrt{n}/4\right) + \mathrm{P}\left(T_* \leqslant 2c_{\mathcal{W}}(\alpha) \cap c_{\mathcal{W}}(\alpha) > v_q\sqrt{n}/4\right) \\
&\leqslant \mathrm{P}\left(T_*/\sqrt{n} \leqslant v_q/2\right) + \mathrm{P}\left(c_{\mathcal{W}}(\alpha) > v_q\sqrt{n}/4\right) \\
&\leqslant \mathrm{P}\left(\left|T_*/\sqrt{n} - v_q\right| \geqslant v_q/2\right) + \mathrm{P}\left(c_{\mathcal{W}}(\alpha) > v_q\sqrt{n}/4\right) =: P_{1a} + P_{1b}.
\end{aligned}$$

The proof of Lemma 1.5 in fact shows that $\mathrm{E}[\max_{1\leqslant k\leqslant q} f_{k*}(Z)^2] \leqslant C'^2$ for some C' depending only on C_1. The first part of the maximal inequality in Lemma 1.34 and $\mathrm{E}[\max_{1\leqslant k\leqslant q} f_{k*}(Z)^2] \leqslant C'^2$ for some C' depending only on C_1 imply that

$$\mathrm{E}\left\{\max_{1\leqslant k\leqslant q} |(\mathbb{E}_n - \mathrm{E})[f_{k*}(Z_i)]|\right\} \leqslant \frac{C' \ln q}{\sqrt{n}}$$

for some C' depending only on C_1. From the previous display, the definition of v_q, and M we therefore see that

$$\begin{aligned}
P_{1a} &= \mathrm{P}\left(\left|\max_{1\leqslant k\leqslant q} |\mathbb{E}_n[f_{k*}(Z)]| - \max_{1\leqslant k\leqslant q} |\mathbb{E}_n[\rho_*(Z, \beta_*, L_*(W))X_k]|\right| \geqslant v_q/2\right) \\
&\leqslant \mathrm{P}\left(\max_{1\leqslant k\leqslant q} |(\mathbb{E}_n - \mathrm{E})[f_{k*}(Z)]| \geqslant v_q/2\right) \leqslant 2v_q^{-1}\mathrm{E}\left\{\max_{1\leqslant k\leqslant q} |(\mathbb{E}_n - \mathrm{E})[f_{k*}(Z_i)]|\right\} \\
&\leqslant \frac{C' v_q^{-1} \ln q}{\sqrt{n}} \to 0.
\end{aligned}$$

Using (1.J.42), we see that

$$\begin{aligned}
P_{1b} &\leqslant \mathrm{P}\left(c_{\mathcal{W}}(\alpha) > v_q\sqrt{n}/4 \cap \mathcal{E}\right) + \mathrm{P}(\mathcal{E}^c) \leqslant \mathbf{1}\left(C'_{\alpha}\sqrt{\ln q} > v_q\sqrt{n}/4\right) + \mathrm{P}(\mathcal{E}^c) \\
&\leqslant \mathbf{1}\left[v_q^{-1}\ln q/\sqrt{n} > 1/(4C'_{\alpha})\right] + \mathrm{P}(\mathcal{E}^c) \to 0.
\end{aligned}$$

Letting $\zeta_1 = \zeta_{1n} > 0$ be the constant defined in Lemma (1.5), the probability P_2 satisfies

$$P_2 = \mathrm{P}\left(\left|T - T_*\right| \zeta_1 > \zeta_1 c_{\mathcal{W}}\left(\alpha\right)\right) \leqslant \mathrm{P}\left(\left|T - T_*\right| > \zeta_1\right) + \mathrm{P}\left(\zeta_1 > c_{\mathcal{W}}\left(\alpha\right)\right) =: P_{2a} + P_{2b}.$$

Now, $P_{2a} \to 0$ by Lemma (1.5). Lastly, using (1.J.43) and $\zeta_1 \to 0$ it follows that

$$P_{2b} \leqslant \mathrm{P}\left(\zeta_1 > c_{\mathcal{W}}\left(\alpha\right) \cap \mathcal{E}\right) + \mathrm{P}\left(\mathcal{E}^c\right) \leqslant \mathbf{1}\left(\zeta_1 > c_\alpha'\right) + \mathrm{P}\left(\mathcal{E}^c\right) \to 0.$$

□

1.J.4 Supporting Lemmas for Section 1.5

Lemma 1.27. *If Assumption 1.10 holds, then there exists c, C and C' depending only on C_1, c_2 and C_2 such that*

$$\mathrm{P}(\|\sqrt{n}(\widehat{\beta} - \beta_*)\| > C'\sqrt{\ln n}) \leqslant Cn^{-c}.$$

Proof. Hoeffding's inequality for bounded random variables, the union bound, and $\|s\left(Z\right)\| \leqslant C_1$ show that for any $t > 0, \mathrm{P}(\|\sqrt{n}\mathbb{E}_n\left[s_*\left(Z_i\right)\right]\| > \sqrt{2d}C_1 t) \leqslant 2d\mathrm{e}^{-t^2}$. Since $d \leqslant C_1$, setting $t \coloneqq \sqrt{\ln n}$ implies $\mathrm{P}(\|\sqrt{n}\mathbb{E}_n\left[s\left(Z_i\right)\right]\| > \sqrt{2}C_1^{3/2}\sqrt{\ln n}) \leqslant 2C_1 n^{-1}$. Assumption 1.10 yields

$$\mathrm{P}(\|\sqrt{n}(\widehat{\beta} - \beta_*) - \mathbb{E}_n\left[s_*\left(Z_i\right)\right]\| > a_n) \leqslant C_2 n^{-c_2},$$

so by T and the union bound,

$$\begin{aligned}
&\mathrm{P}(\|\sqrt{n}(\widehat{\beta} - \beta_*)\| > a_n + \sqrt{2}C_1^{3/2}\sqrt{\ln n}) \\
&\leqslant \mathrm{P}(\|\sqrt{n}(\widehat{\beta} - \beta_*) - \mathbb{E}_n\left[s_*\left(Z_i\right)\right]\| > a_n) + \mathrm{P}(\|\sqrt{n}\mathbb{E}_n\left[s_*\left(Z_i\right)\right]\| > \sqrt{2}C_1^{3/2}\sqrt{\ln n}) \\
&\leqslant C_2 n^{-c_2} + 2C_1 n^{-1} \leqslant Cn^{-c}.
\end{aligned}$$

Given that $a_n \to 0, a_n + \sqrt{2}C_1^{3/2}\sqrt{\ln n} \leqslant C'\sqrt{\ln n}$, □

Lemma 1.28 (Vector Hoeffding's inequality). *If $\{X_i\}_1^n$ are independent centered $\mathbf{R}^d$-valued random variables satisfying $\|X_i\| \leqslant M$ for all $1 \leqslant i \leqslant n$, then for each $t > 0$,*

$$\mathrm{P}\left(\|\mathbb{E}_n\left(X_i\right)\| > t\right) \leqslant 2d\exp\left(-\frac{nt^2}{2dM^2}\right).$$

Hence, for each $t > 0$,

$$\mathrm{P}(\left\|\sqrt{n}\mathbb{E}_n\left(X_i\right)\right\| > \sqrt{2d}Mt) \leqslant 2d\mathrm{e}^{-t^2}.$$

Proof. By the union bound and Hoeffding's inequality for bounded random variables

$$\begin{aligned}\mathrm{P}\left(\|\mathbb{E}_n\left(X_i\right)\| > t\right) &\leqslant \mathrm{P}(\max_{1\leqslant k\leqslant d}|\mathbb{E}_n\left(X_{ik}\right)| > t/\sqrt{d}) \leqslant \sum_{j=1}^{d}\mathrm{P}(|\mathbb{E}_n\left(X_{ik}\right)| > t/\sqrt{d}) \\ &\leqslant 2d\exp\left(-\frac{nt^2}{2dM^2}\right).\end{aligned}$$

□

1.K Sparse Eigenvalues and Compatibility Constants

The following result is essentially Rudelson and Vershynin (2008, Lemma 3.8).

Lemma 1.29. *Let $\{W_i\}_1^n, n \geqslant 2$, be independent random variables taking values in $\mathbf{R}^p, p \geqslant 2$. Define $M \coloneqq \max_{1\leqslant i\leqslant n}\|W_i\|_\infty$. Then for any $m \in \{1,\ldots,p\}$ we have*

$$\mathrm{E}\Big\{\sup_{\|\delta\|_0\leqslant m,\|\delta\|=1}|\left(\mathbb{E}_n - \overline{\mathrm{E}}\right)[(W_i^\top\delta)^2]|\Big\} \leqslant \delta_n^2\left(m\right) + \delta_n\left(m\right)\sup_{\|\delta\|_0\leqslant m,\|\delta\|=1}\sqrt{\overline{\mathrm{E}}[(W_i^\top\delta)^2]},$$

where C' is universal and

$$\delta_n\left(m\right) \coloneqq C'[\mathrm{E}(M^2)]^{1/2}\sqrt{\frac{m\ln p}{n}}\left[1 + (\ln m)\sqrt{\ln n}\right].$$

Rudelson and Vershynin (2008, Lemma 3.8), a lemma which is stated conditionally on the data, leaves the constant appearing in their δ_n as $C\left(M\right)$. In order to clarify the exact dependence of δ_n on M, I include a complete proof.

Let $\kappa\left(a, A\right)$ denote the *compatibility constant* associated with a real symmetric matrix $A \in \mathbf{R}^{q\times q}$,

$$\kappa\left(a, A\right) \coloneqq \min_{\substack{T\subset\{1,\ldots,q\}\\|T|\leqslant s}}\inf_{\substack{\delta\in\mathbf{R}^q\setminus\{\mathbf{0}\}\\\|\delta_{T^c}\|_1\leqslant a\|\delta_T\|_1}}\frac{(s\delta^\top A\delta)^{1/2}}{\|\delta_T\|_1},$$

where $a > 0$. Note that any such compatibility constant depends on the sparsity level s and dimension q, although such dependencies are suppressed throughout.

Define the *minimal sparse eigenvalue* $\phi_{\min}\left(m, A\right)$,

$$\phi_{\min}\left(m, A\right) \coloneqq \inf_{1\leqslant\|\delta\|_0\leqslant m}\frac{\delta^\top A\delta}{\|\delta\|^2},$$

and *maximal sparse eigenvalue* $\phi_{\max}(m, A)$,

$$\phi_{\max}(m, A) \coloneqq \sup_{1 \leqslant \|\delta\|_0 \leqslant m} \frac{\delta^\top A \delta}{\|\delta\|^2},$$

which are defined for $m \in \{1, \ldots, p\}$. The following lemma is implicit in the proof of Bickel, Ritov, and Tsybakov (2009, Lemma 4.1(ii)). I include the proof for easy reference.

Lemma 1.30. *Then for any real symmetric matrix A. Then for any $a > 0$,*

$$\kappa(a, A) \geqslant \max_{1 \leqslant m \leqslant p} \left\{ \sqrt{\phi_{\min}(s + m, A)} - a\sqrt{\frac{\phi_{\max}(m, A)\, s}{m}} \right\}.$$

Lemma 1.31. *Suppose that $\{W_i\}_1^n, n \geqslant 2$, are independent random variables taking values in $\mathbf{R}^p, p \geqslant 2$, which are uniformly bounded by $B < \infty$, $\max_{1 \leqslant i \leqslant n} \|W_i\|_\infty \leqslant B$, and whose average population matrix $\overline{\mathrm{E}}(W_i W_i^\top)$ has m-sparse eigenvalues, $m \geqslant 1$, bounded above by $\phi_H(m) < \infty$ and away from zero by $\phi_L(m) > 0$. Then*

$$\mathrm{P}\left(\phi_{\min}\left(m, \mathbb{E}_n(W_i W_i^\top)\right) < \phi_L(m)/2\right) \leqslant 2\phi_L(m)^{-1}\left[\delta_n^2(m) + \delta_n(m)\sqrt{\phi_H(m)}\right],$$

$$\textit{and} \quad \mathrm{P}\left(2\phi_H(m) < \phi_{\max}\left(m, \mathbb{E}_n(W_i W_i^\top)\right)\right) \leqslant \phi_H(m)^{-1}\left[\delta_n^2(m) + \delta_n(m)\sqrt{\phi_H(m)}\right],$$

where C' is a universal constant and

$$\delta_n(m) \coloneqq C' B \sqrt{\frac{m \ln p}{n}}\left[1 + (\ln m)\sqrt{\ln n}\right].$$

Lemma 1.32. *Let $\{W_i\}_1^n$ be independent $\mathbf{R}^p$-valued random variables with $\max_{1 \leqslant i \leqslant n} \|W_i\|_\infty \leqslant C_1$, $c_1^2 \leqslant \lambda_{\min}(\overline{\mathrm{E}}(W_i W_i^\top)) \leqslant \lambda_{\max}(\overline{\mathrm{E}}(W_i W_i^\top)) \leqslant C_1^2$ and $p \geqslant 2$, and suppose that $s \ln^5(pn)/n \leqslant C_2 n^{-c_2}$. Then there exists constants c and C depending only on c_1, C_1, c_2 and C_2 such that*

$$\mathrm{P}\left(\{\phi_{\min}(s\ln(n) + s) < c_1/2, \mathbb{E}_n(W_i W_i^\top)\} \cup \{2C_1 < \phi_{\max}(s \ln n), \mathbb{E}_n(W_i W_i^\top)\}\right) \leqslant C n^{-c}.$$

Lemma 1.33. *Let $\{W_i\}_1^n$ be independent $\mathbf{R}^p$-valued random variables with $\max_{1 \leqslant i \leqslant n} \|W_i\|_\infty \leqslant C_1$, $c_1^2 \leqslant \lambda_{\min}(\overline{\mathrm{E}}(W_i W_i^\top)) \leqslant \lambda_{\max}(\overline{\mathrm{E}}(W_i W_i^\top)) \leqslant C_1^2$ and $p \geqslant 2$, and suppose that $s \ln^5(pn)/n \leqslant C_2 n^{-c_2}$. Then there exists constants c and C depending only on c_1, C_1, c_2 and C_2 such that for $n \geqslant \exp(16a^2 C_1/c_1)$,*

$$\mathrm{P}\left(\kappa(a) \geqslant \sqrt{c_1/8}\right) \geqslant 1 - C n^{-c}.$$

1.L Inequalities

The following lemma specializes Chernozhukov, Chetverikov, and Kato (2015, Lemma 8) to i.i.d. and bounded data, and is a useful variation of standard maximal inequalities.

Lemma 1.34. *Let $\{X_i\}_1^n$ be i.i.d., centered, $\mathbf{R}^p$-valued random variables with $p \geqslant 2$. Then*

$$\mathrm{E}\left[\max_{1\leqslant j\leqslant p} |\mathbb{E}_n\left(X_{ij}\right)|\right] \lesssim \Big(\max_{1\leqslant j\leqslant p}[\mathrm{E}\left(X_j^2\right)]^{1/2} + [\mathrm{E}(\max_j X_j^2)]^{1/2}\Big)\frac{\ln p}{\sqrt{n}},$$

where the constant is universal. Consequently, if, in addition, $\|X\|_\infty \leqslant M$ constant, then

$$\mathrm{E}\left[\max_{1\leqslant j\leqslant p} |\mathbb{E}_n\left(X_{ij}\right)|\right] \leqslant C\left(M\right)\frac{\ln p}{\sqrt{n}},$$

where $C\left(M\right)$ depends only on M.

The following lemma is a special case of Massart (2000, Theorem 4), a version of Talagrand's inequality.

Lemma 1.35 (Talagrand's inequality). *Let $\{X_i\}_1^n$ be independent, centered random variables with values in $[-M, M]^p$ for some $M > 0$, and $\sigma^2 \coloneqq \max_{1\leqslant j\leqslant N} \overline{\mathrm{E}}\left(X_{ij}^2\right)$. Then for any $\varepsilon, t > 0$,*

$$\mathrm{P}\left(\max_{1\leqslant j\leqslant p}|\mathbb{E}_n\left(X_{ij}\right)| > \left(1+\varepsilon\right)\mathrm{E}\Big[\max_{1\leqslant j\leqslant p}|\mathbb{E}_n\left(X_{ij}\right)|\Big] + \sigma\sqrt{\frac{2\kappa t}{n}} + \frac{\kappa\left(\varepsilon\right)Mt}{n}\right) \leqslant \mathrm{e}^{-t},$$

wgere κ and $\kappa\left(\varepsilon\right)$ may be taken equal to $\kappa = 4$ and $\kappa\left(\varepsilon\right) = 2.5 + 32\varepsilon^{-1}$. Consequently, there exists constants C and C' depending only on M such that for any $t > 0$,

$$\mathrm{P}\left(\max_{1\leqslant j\leqslant p}|\mathbb{E}_n\left(X_{ij}\right)| > 2\mathrm{E}\Big[\max_{1\leqslant j\leqslant p}|\mathbb{E}_n\left(X_{ij}\right)|\Big] + C'\sqrt{\frac{t}{n}} + C''\frac{t}{n}\right) \leqslant \mathrm{e}^{-t}.$$

The next lemma combines Lemmas 1.34 and 1.35.

Lemma 1.36. *Let $\{X_i\}_1^n$ be i.i.d., zero-mean random variables taking values in $[-M, M]^p$ for some $M > 0$. Then there exists a constant C' depending only on M such that*

$$\mathrm{P}\left(\max_{1\leqslant j\leqslant p}|\mathbb{E}_n\left(X_{ij}\right)| > C'\frac{\ln\left(pn\right)}{\sqrt{n}}\right) \leqslant n^{-1}.$$

Let $\psi_p(t) \coloneqq \mathrm{e}^{t^p} - 1$. The ψ_p-*Orlisz norm* $\|X\|_{\psi_p}$ of a random variable is

$$\|X\|_{\psi_p} \coloneqq \inf\{C > 0 | \,\mathrm{E}[\psi_p\left(|X|/C\right)] \leqslant 1\}. \tag{1.L.1}$$

Lemma 1.37. *Let X be a zero-mean, Gaussian random variable with variance σ^2. Then its ψ_2-Orlisz norm $\|X\|_{\psi_2} = \sqrt{8/3}\sigma$.*

Lemma 1.38. *Let the function ψ be nonnegative, convex function, and strictly increasing on $\mathbf{R}_+$. Then for any $C > 0$ satisfying $\max_{1\leqslant j\leqslant p} \mathrm{E}\left[\psi\left(|X_j|/C\right)\right] \leqslant 1$, we have $\mathrm{E}\left[\max_{1\leqslant j\leqslant p}|X_j|\right] \leqslant C\psi^{-1}(p)$.*

Lemma 1.39 (Gaussian Maximal Inequality). *Let $\{X_j\}_1^p$ be centered Gaussian with maximal variance $\sigma^2 \coloneqq \max_{1\leqslant j\leqslant p} \mathrm{E}\left(X_j^2\right)$. Then*

$$\mathrm{E}\Big[\max_{1\leqslant j\leqslant p}|X_j|\Big] \leqslant \sigma\sqrt{8/3}\sqrt{\ln(1+p)}.$$

Consequently, if, in addition, $p \geqslant 2$,

$$\mathrm{E}\Big[\max_{1\leqslant j\leqslant p}|X_j|\Big] \leqslant \sigma\sqrt{16/3}\sqrt{\ln p}.$$

The next lemma is a finite-dimensional version of Borell's inequality, The proof of the lemma follows from the proof of van der Vaart and Wellner (1996, Proposition A.2.1).

Lemma 1.40 (Borell's inequality). *Let $X \sim \mathrm{N}\left(\mathbf{0}_{p\times 1}, \Sigma\right), \|X\|_\infty \coloneqq \max_{1\leqslant j\leqslant p}|X_j|$, and $\sigma^2 \coloneqq \max_{1\leqslant j\leqslant p}\mathrm{E}\left(X_j^2\right)$. For every $t > 0$,*

$$\mathrm{P}\left(\|X\|_\infty > \mathrm{E}\left(\|X\|_\infty\right) + \sqrt{2}\sigma t\right) \leqslant \mathrm{e}^{-t^2}.$$

Lemma 1.41 (Gaussian Deviation Inequality). *Let $X \sim \mathrm{N}\left(\mathbf{0}_{p\times 1}, \Sigma\right), p \geqslant 2, \|X\|_\infty \coloneqq \max_{1\leqslant j\leqslant p}|X_j|$, and $\sigma^2 \coloneqq \max_{1\leqslant j\leqslant p}\mathrm{E}\left(X_j^2\right)$. Then there exists a universal constant K such that for every $n \geqslant 1$,*

$$\mathrm{P}\left(\|X\|_\infty > K\sigma\sqrt{\ln(pn)}\right) \leqslant n^{-1}.$$

A random variable X with mean $\mu \coloneqq \mathrm{E}(X)$ is said to be *subgaussian* if there exists $\sigma \in \mathbf{R}_+$ such that $\mathrm{E}\left[\mathrm{e}^{t(X-\mu)}\right] \leqslant \mathrm{e}^{t^2\sigma^2/2}$ for all $t \in \mathbf{R}$. We say that the *subgaussianity parameter of X* is (at most) σ and call the smallests of such σ's the *optimal subgaussianity parameter.* The following lemma is taken from Marchal and Arbel (2017, Theorem 1).

Lemma 1.42 (Optimal Subgaussianity Parameter for Beta Distribution). *For a* Beta(α,β) *distributed random variable, $\alpha, \beta \in (0,\infty)$, a simple and explicit upper bound of the optimal subgaussianity parameter is $1/4\,(\alpha+\beta+1)$.*

Subgaussian random variables yield a simple maximal inequality.

Lemma 1.43 (Subgaussian Maximal Inequality). *Let $\{X_j\}_1^p, p \geqslant 2$, be subgaussian random variables with common subgaussianity parameter σ. Then*

$$\mathrm{E}\Big(\max_{1\leqslant j\leqslant p} X_j\Big) \leqslant \sigma\sqrt{2\ln p} \quad \textit{and} \quad \mathrm{E}\Big[\max_{1\leqslant j\leqslant p} |X_j|\Big] \leqslant \sigma\sqrt{2\ln(2p)}.$$

Lemma 1.44. *Suppose that $c_1 \leqslant [\overline{\mathrm{E}}(X_{ijk}^2)]^{1/2} \leqslant [\overline{\mathrm{E}}(|X_{ijk}|^{2+\delta})]^{1/(2+\delta)} \leqslant C_1$ for some $0 < \delta \leqslant 1$, and all $j \in \{1,\ldots,p\}$, where c_1 and C_1 depend only on δ. If $\ln(pqn) \leqslant C_2 n^{1-c_2}$ for some $\frac{8}{8+4\delta} < c_2 < 1$, then for every $0 \leqslant c_3 \leqslant 1$ and every*

$$n \geqslant (2C_1\sqrt{C_2}/c_1)^{2/[c_2-8/(8+4\delta)]}$$

we have

$$\mathrm{P}\left(\max_{(j,k)\in[p]\times[q]} |\mathcal{S}_{njk}| > \Phi^{-1}\left(1 - n^{-c_3}/(2pq)\right)\right) \leqslant \left[1 + A\left(1 + C_1/c_1\right)^{2+\delta}\right] n^{-c_3}.$$

1.M Proofs for Section 1.G

Proof of Theorem 1.7. Suppose first that the penalty loadings are *conservatively* polynomially valid. Following the outline of the proof of BCCH (Theorem 1), the proof has five steps.

Step 1.

For $a > 0$, consider the compatibility constants

$$\kappa_k^*(a) := \min_{|T|\leqslant s} \min_{\substack{\delta\neq\mathbf{0} \\ \|\widehat{\Upsilon}_k^*\delta_{T^c}\|_1 \leqslant a\|\widehat{\Upsilon}_k^*\delta_T\|_1}} \frac{\sqrt{s}\|\delta\|_{2,n}}{\|\widehat{\Upsilon}_k^*\delta_T\|_1}, \quad k \in \{1,\ldots,q\}, \tag{1.M.1}$$

where the T's are understood to be nonempty subsets of $\{1,\ldots,p\}$. The κ_k^*'s control the modulus of continuity between the prediction norm $\|\delta\|_{2,n}$ and the $\widehat{\Upsilon}_k^*$-weighted ℓ_1 norms. The main result of this step is the following lemma, which contains the determistic part of the argument in establishing the rate of the Lasso estimators.

Lemma 1.45. *Suppose that Assumption 1.16 holds, that $\lambda \geqslant c_0 n \max_{1\leqslant k\leqslant q}\|S_k^*\|_\infty$ for $c_0 > 1$ and $c_0' \in (0,1]$ and the penalty loadings satisfy (1.G.15) with $\ell > 1/c_0$. Then for each $k \in \{1,\ldots,q\}$,*

$$\|\widehat{\beta}_k - \beta_{k0}\|_{2,n} \leqslant \left(u + \frac{1}{c_0}\right)\frac{\lambda\sqrt{s}}{\kappa_k^*(\overline{c})\,n} + 2\left(c_s + \Delta\right),$$

where $\bar{c} \coloneqq (uc_0 + 1) \,/\, (\ell c_0 - 1)$.

Proof. The proof follows the argument in Bickel, Ritov, and Tsybakov (2009). Fix $k \in \{1, \ldots, q\}$. Denote $\delta_k \coloneqq \widehat{\beta}_k - \beta_{k0}$ and recall $T_{k0} = \operatorname{supp}(\beta_{k0})$. Note that $|T_{k0}| \leqslant s$. Expand the squares to arrive at

$$\begin{aligned}
&\mathbb{E}_n[(\widehat{Y}_{ik} - W_i^\top \widehat{\beta}_k)^2] - \mathbb{E}_n[(\widehat{Y}_{ik} - W_i^\top \beta_{k0})^2] \\
&= \|\delta_k\|_{2,n}^2 - 2\mathbb{E}_n(\varepsilon_{ik} W_i^\top \delta_k) - 2\mathbb{E}_n(r_{ik} W_i^\top \delta_k) - 2\mathbb{E}_n(e_{ik} W_i^\top \delta_k) \\
&\geqslant \|\delta_k\|_{2,n}^2 - \|S_k^*\|_\infty \|\widehat{\mathit{\Upsilon}}_k^* \delta_k\|_1 - 2\,(c_s + \Delta)\, \|\delta_k\|_{2,n},
\end{aligned}$$

where the inequality follows from the Hölder and Cauchy-Schwarz inequalities. By definition of $\widehat{\beta}_k$ as a minimizer

$$\mathbb{E}_n[(\widehat{W}_{ik} - W_i^\top \widehat{\beta}_k)^2] + \frac{\lambda}{n}\|\widehat{\mathit{\Upsilon}}_k \widehat{\beta}_k\|_1 \leqslant \mathbb{E}_n[(\widehat{Y}_{ik} - W_i^\top \beta_{k0})^2] + \frac{\lambda}{n}\|\widehat{\mathit{\Upsilon}}_k \beta_{k0}\|_1,$$

which combined with the previous display implies

$$\begin{aligned}
\|\delta_k\|_{2,n}^2 &\leqslant \frac{\lambda}{n}\left(\|\widehat{\mathit{\Upsilon}}_k \beta_{k0}\|_1 - \|\widehat{\mathit{\Upsilon}}_k \widehat{\beta}_k\|_1\right) + \|S_k^*\|_\infty \|\widehat{\mathit{\Upsilon}}_k^* \delta_k\|_1 + 2\,(c_s + \Delta)\,\|\delta_k\|_{2,n} \\
&\leqslant \frac{\lambda}{n}\left(\|\widehat{\mathit{\Upsilon}}_k \delta_{kT_{k0}}\|_1 - \|\widehat{\mathit{\Upsilon}}_k \delta_{kT_{k0}^c}\|_1\right) + \|S_k^*\|_\infty \|\widehat{\mathit{\Upsilon}}_k^* \delta_k\|_1 + 2\,(c_s + \Delta)\,\|\delta_k\|_{2,n} \\
&\leqslant \frac{\lambda}{n}\left(u\|\widehat{\mathit{\Upsilon}}_k^* \delta_{kT_{k0}}\|_1 - \ell\|\widehat{\mathit{\Upsilon}}_k^* \delta_{kT_{k0}^c}\|_1\right) + \frac{\lambda}{c_0 n}\left(\|\widehat{\mathit{\Upsilon}}_k^* \delta_{kT_{k0}}\|_1 + \|\widehat{\mathit{\Upsilon}}_k^* \delta_{kT_{k0}^c}\|_1\right) \\
&\quad + 2\,(c_s + \Delta)\,\|\delta_k\|_{2,n} \\
&= \left(u + \frac{1}{c_0}\right)\frac{\lambda}{n}\|\widehat{\mathit{\Upsilon}}_k^* \delta_{T_{k0}}\|_1 - \left(\ell - \frac{1}{c_0}\right)\frac{\lambda}{n}\|\widehat{\mathit{\Upsilon}}_k^* \delta_{kT_{k0}^c}\|_1 + 2\,(c_s + \Delta)\,\|\delta_k\|_{2,n}. \qquad (1.\mathrm{M}.2)
\end{aligned}$$

If $\|\delta_k\|_{2,n} \leqslant 2\,(c_s + \Delta)$, then the claim follows. Suppose therefore that $\|\delta_k\|_{2,n} > 2\,(c_s + \Delta)$. Then the previous display implies

$$\|\widehat{\mathit{\Upsilon}}_k^* \delta_{kT_{k0}^c}\|_1 \leqslant \left[\left(u + \frac{1}{c_0}\right) \Big/ ofT \left(\ell - \frac{1}{c_0}\right)\right] \|\widehat{\mathit{\Upsilon}}_k^* \delta_{kT_{k0}}\|_1 = \bar{c}\|\widehat{\mathit{\Upsilon}}_k^* \delta_{kT_{k0}}\|_1,$$

from which it follows that

$$\|\widehat{\mathit{\Upsilon}}_k^* \delta_{kT_{k0}}\|_1 \leqslant \frac{\sqrt{s}\|\delta_k\|_{2,n}}{\kappa_k^*(\bar{c})}.$$

Combining this inequality with (1.M.2) we see that

$$\|\delta_k\|_{2,n}^2 \leqslant \left(u + \frac{1}{c_0}\right)\frac{\lambda}{n}\|\widehat{\mathit{\Upsilon}}_k^* \delta_{kT_{k0}}\|_1 + 2\,(c_s + \Delta)\,\|\delta_k\|_{2,n}$$

$$\leqslant \left(u + \frac{1}{c_0}\right) \frac{\lambda\sqrt{s}\|\delta_k\|_{2,n}}{\kappa_k^*\left(\overline{c}\right)n} + 2\left(c_s + \Delta\right)\|\delta_k\|_{2,n},$$

which leads to the desired bound. □

Step 2

In this step I prove a lemma about the quantiles of the maximum of the "noise" $S_k^* = 2(\widehat{\Upsilon}_k^*)^{-1}\mathbb{E}_n(\varepsilon_{ik}W_i)$, which suggests the level of the penalty.

Lemma 1.46. *Suppose that Assumptions 1.16, 1.18 and 1.19 hold and the penalty level λ is specified as in (1.G.10). Then there exists C and n_0 depending only on c_1, C_1, C_2 and c_2' such that for all $n \geqslant n_0$,*

$$\mathrm{P}\left(c_0 \max_{1\leqslant k\leqslant q}\|S_k^*\|_\infty > \lambda/n\right) \leqslant Cn^{-c_0'}.$$

Proof. Observe that

$$\mathrm{P}\left(c_0 \max_{1\leqslant k\leqslant q}\|S_k^*\|_\infty > \lambda/n\right) = \mathrm{P}\left(\max_{1\leqslant k\leqslant q}\|\sqrt{n}S_k^*/2\|_\infty > \Phi^{-1}\left(1 - n^{-c_0'}/\left(2pq\right)\right)\right)$$

Each $\sqrt{n}S_{kj}^*/2$ is bounded in absolute value by the self-normalized sum $\sqrt{n}S_{kj}^{**}/2 = \mathbb{E}_n\left(\varepsilon_{ik}W_{ij}\right)/[\mathbb{E}_n\left(\varepsilon_{ik}^2W_{ij}^2\right)]^{1/2}$, so the claim follows Lemma 1.44 with $c_3 = c_0', \delta = 3$ and $X_{ijk} = \varepsilon_{ik}W_{ij}$, such that $\mathcal{S}_{njk} = \sqrt{n}S_{kj}^{**}/2$. □

Step 3

Lemma 1.47. *If Assumptions 1.18 and 1.19 hold, then there exists c, C, c', C' and n_0 depending only on c_1, C_1, c_2, C_2 and c_2' such that for all $n \geqslant n_0$,*

$$\mathrm{P}\left(c' \leqslant \widehat{\gamma}_{kj}^* \leqslant C' \textit{ for all } (j,k) \in [p]\times[q]\right) \geqslant 1 - Cn^{-c}.$$

Proof. Assumption 1.18, Markov's inequality and a maximal inequality for bounded random variables (Lemma 1.34) show

$$\mathrm{P}\left(\max_{1\leqslant k\leqslant q}|\left(\mathbb{E}_n - \mathrm{E}\right)\left(\varepsilon_{ik}^2\right)| > n^{-1/4}\right) \leqslant n^{1/4}\mathrm{E}\left[\max_{1\leqslant k\leqslant q}|\left(\mathbb{E}_n - \mathrm{E}\right)\left(\varepsilon_{ik}^2\right)|\right] \leqslant C'\left(\frac{\ln^4\left(q\right)}{n}\right)^{1/4}.$$

Assumption 1.19 implies that there exists c and C depending only on c_1, C_1, c_2, C_2 and c_2'

such that with probability $\geqslant 1 - Cn^{-c}$,

$$\max_{1 \leqslant k \leqslant q} |\left(\mathbb{E}_n - \mathrm{E}\right)(\varepsilon_{ik}^2)| \leqslant n^{-1/4}$$

Using Assumption 1.18 and the previous diplay, it follows that there exists c, C, c' and C' depending only on c_1, C_1, c_2, C_2 and c_2' such that with probability $\geqslant 1 - Cn^{-c}$,

$$\begin{aligned} \max_{(j,k) \in [p] \times [q]} |\widehat{\gamma}_{kj}^{*2} - \gamma_{kj}^{*2}| &\leqslant \max_{1 \leqslant k \leqslant q} |\left(\mathbb{E}_n - \mathrm{E}\right)(\varepsilon_{ik}^2)| \max_{1 \leqslant i \leqslant n} W_{ij}^2 + \max_{1 \leqslant k \leqslant q} \mathrm{E}(\varepsilon_k^2) \max_{1 \leqslant j \leqslant p} \Big| \max_{1 \leqslant i \leqslant n} W_{ij}^2 - M_j^2 \Big| \\ &\leqslant C_1^2 \max_{1 \leqslant k \leqslant q} |\left(\mathbb{E}_n - \mathrm{E}\right)(\varepsilon_{ik}^2)| + 2C_1^3 \max_{1 \leqslant j \leqslant p} \Big| \max_{1 \leqslant i \leqslant n} |W_{ij}| - M_j \Big| \leqslant C' n^{-c'}, \end{aligned}$$

where $\gamma_{kj}^{*2} = \mathrm{E}(\varepsilon_k^2) M_j^2$. By Assumption 1.18 the γ_{kj}^{*2}'s are bounded from above and away from zero. The previous diplay now shows that for sufficiently large n and with probability approaching one polynomially fast, so are the $\widehat{\gamma}_{kj}^{*2}$'s. These upper and lower bounds depend only on c_1, C_1, c_2, C_2 and c_2'. □

Step 4

Lemma 1.48. *If Assumption 1.18 holds, then there exists c, C, c' and n_0 depending only on $c_0, c_0', c_1, C_1, c_2, C_2$ and c_2' such that for all $n \geqslant n_0$,*

$$\mathrm{P}\left(\min_{1 \leqslant k \leqslant q} \kappa_k^* \left(\bar{c}\right) \geqslant c' \right) \geqslant 1 - Cn^{-c}.$$

Proof. Let $\widehat{\gamma}_{\min}^* := \min_{(j,k) \in [p] \times [q]} \widehat{\gamma}_{kj}^*$ and $\widehat{\gamma}_{\max}^* := \max_{(j,k) \in [p] \times [q]} \widehat{\gamma}_{kj}^*$. By Lemma 1.47 there exists c, C, c', C' and n_0 depending only on c_1, C_1, c_2, C_2 and c_2' such that for all $n \geqslant n_0$, $c' \leqslant \widehat{\gamma}_{\min}^* \leqslant \widehat{\gamma}_{\max}^* \leqslant C'$ with probability $\geqslant 1 - Cn^{-c}$. To avoid division by zero or multiplication by infinity, take $n \geqslant n_0$ and work on the event $\{c' \leqslant \widehat{\gamma}_{\min}^* \leqslant \widehat{\gamma}_{\max}^* \leqslant C'\}$.

Fix $k \in \{1, \ldots, q\}$. By construction of $\widehat{\gamma}_{\min}^*$ and $\widehat{\gamma}_{\max}^*$, $\|\widehat{\Upsilon}_k^* \delta\|_1 \geqslant \widehat{\gamma}_{\min}^* \|\delta\|_1$ and $\|\widehat{\Upsilon}_k^* \delta\|_1 \leqslant \widehat{\gamma}_{\max}^* \|\delta\|_1$ for any $\delta \in \mathbf{R}^p$. Let T be such that $|T| \leqslant s$ and let $\delta \neq \mathbf{0}$ satisfy $\|\widehat{\Upsilon}_k^* \delta_{T^c}\|_1 \leqslant C \|\widehat{\Upsilon}_k^* \delta_T\|_1$. Then $\|\delta_{T^c}\|_1 \leqslant \left(\widehat{\gamma}_{\max}^* C / \widehat{\gamma}_{\min}^*\right) \|\delta_T\|_1$, and it follows that

$$\kappa_k^* \left(a\right) \geqslant \frac{1}{\widehat{\gamma}_{\max}^*} \kappa \left(\widehat{\gamma}_{\max}^* a / \widehat{\gamma}_{\min}^*\right) \geqslant \frac{1}{C'} \kappa \left(C' a / c'\right),$$

where the second inequality follows from the event $\{c' \leqslant \widehat{\gamma}_{\min}^* \leqslant \widehat{\gamma}_{\max}^* \leqslant C'\}$ and κ being a nonincreasing function. Taking the minimum over $k \in \{1, \ldots, q\}$ and setting $a = \bar{c}$ we see that

$$\min_{1 \leqslant k \leqslant q} \kappa_k^* \left(\bar{c}\right) \geqslant \frac{1}{C'} \kappa \left(C' \bar{c} / c'\right),$$

Using Assumptions 1.18 and 1.19 Lemma 1.33 implies that there exists c, C and n_0 depending only on $c_0, c_0', c_1, C_1, c_2, C_2$ and c_2' such that $\kappa\left(C'\bar{c}/c'\right) \geqslant \sqrt{c_1^2/8}$ with probability $\geqslant 1 - Cn^{-c}$ at least for $n \geqslant n_0$. The claim now follows from the previous display and the union bound. □

Step 5

Combine the results of all the previous steps: Given that $\lambda = 2c_0\sqrt{n}\Phi^{-1}(1 - n^{-c_0'}/(2pq)) \lesssim \sqrt{n\ln(pqn)}$, that the penalty loadings $\{\widehat{\Upsilon}_k\}_1^q$ are conservatively polynomially valid, that $\min_{1\leqslant k\leqslant q}\kappa_k^*\left(\bar{c}\right)$ is bounded away from zero (Step 4), $c_s \leqslant C_1\sqrt{s/n}$ (Assumption proof 1.11) and $\Delta \leqslant C_1\sqrt{s\ln(pqn)/n}$ (Assumption 1.17) with probability approaching one polynomially fast, it follows that there exists c, C, C' and n_0 depending only on $c_0, c_0', c_1, C_1, c_2, C_2$ and c_2' such that for all $n \geqslant n_0$ with probability $\geqslant 1 - Cn^{-c}$ we have

$$\max_{1\leqslant k\leqslant q}\|\widehat{\beta}_k - \beta_{k0}\|_{2,n} \leqslant \left(u + \frac{1}{c_0}\right)\frac{\lambda\sqrt{s}}{\min_{1\leqslant k\leqslant q}\kappa_k^*\left(\bar{c}\right)n} + 2\left(c_s + \Delta\right) \leqslant C'\sqrt{\frac{s\ln\left(pqn\right)}{n}}.$$

By the triangle inequality and Assumption 1.11, for at least the same n's and with at least the same probability (but possibly different C'),

$$\max_{1\leqslant k\leqslant q}\|\widehat{\beta}_k - \beta_{k*}\|_{2,n} \leqslant \max_{1\leqslant k\leqslant p}\|\widehat{\beta}_k - \beta_{k0}\|_{2,n} + c_s \leqslant C'\sqrt{\frac{s\ln\left(pqn\right)}{n}}.$$

Suppose next that the penalty loadings are *truly* polynomially valid.

Step 1'

Define compatibility constants $\{\kappa_k^{**}\}_1^q$ as in (1.M.1) using $\widehat{\Upsilon}_k^{**}$-weighted ℓ_1 norms instead. Using an argument virtually identical to the one used in proving Lemma 1.45, we may show that if $\lambda \geqslant c_0 n\max_{1\leqslant k\leqslant q}\|S_k^{**}\|_\infty$, $c_0 > 1$, $c_0' \in (0, 1]$, and the penalty loadings satisfy (1.G.16) with $\ell > 1/c_0$, then for each $k \in \{1, \dots, q\}$,

$$\|\widehat{\beta}_k - \beta_{k0}\|_{2,n} \leqslant \left(u + \frac{1}{c_0}\right)\frac{\lambda\sqrt{s}}{\kappa_k^{**}\left(\bar{c}\right)n} + 2\left(c_s + \Delta\right),$$

where $\bar{c} \coloneqq \left(uc_0 + 1\right)/\left(\ell c_0 - 1\right)$.

Step 2'

Given that each $\sqrt{n}S_k^{**}/2$ is equal to a vector of self-normalized sums, the argument used to prove Lemma 1.46 shows that there exists C and n_0 depending only on c_1, C_1, C_2 and c_2'

such that for all $n \geqslant n_0$,

$$\mathrm{P}\left(c_0 \max_{1 \leqslant k \leqslant q} \|S_k^{**}\|_\infty > \lambda/n\right) \leqslant Cn^{-c_0'}.$$

Step 3'

Lemma 1.49. *If Assumptions 1.18 and 1.19 hold, then there exists c, C, c', C' and n_0 depending only on c_1, C_1, c_2, C_2 and c_2' such that for all $n \geqslant n_0$,*

$$\mathrm{P}\left(c' \leqslant \widehat{\gamma}_{kj}^{**} \leqslant C' \textit{ for all } (j,k) \in [p] \times [q]\right) \geqslant 1 - Cn^{-c}.$$

Proof. Assumption 1.18, Markov inequality and a maximal inequality for bounded random variables (Lemma 1.34) imply that

$$\begin{aligned}\mathrm{P}\left(\max_{(j,k)\in[p]\times[q]} |\widehat{\gamma}_{kj}^{**2} - \gamma_{kj}^{**2}| > n^{-1/4}\right) &\leqslant n^{1/4}\mathrm{E}\left[\max_{(j,k)\in[p]\times[q]} |\left(\mathbb{E}_n - \mathrm{E}\right)\left(\varepsilon_{il}^2 x_{ij}^2\right)|\right] \\ &\leqslant n^{1/4} C' \frac{\ln(pq)}{\sqrt{n}} = C' \left(\frac{\ln^4(pq)}{n}\right)^{1/4},\end{aligned}$$

where $\gamma_{kj}^{**2} = \mathrm{E}\left(\varepsilon_k^2 W_j^2\right)$. Assumption 1.19 implies that there exist c and C depending only on c_1, C_1, c_2, C_2 and c_2' such that with probability $\geqslant 1 - Cn^{-c}$,

$$\max_{(j,k)\in[p]\times[q]} |\widehat{\gamma}_{kj}^{**2} - \gamma_{kj}^{**2}| \leqslant n^{-1/4}.$$

Assumption 1.18 implies that the γ_{kj}^{**}'s are bounded from above and away from zero. By the previous diplay, for sufficiently large n and with probability approaching one polynomially fast, so are the $\widehat{\gamma}_{kj}^{**}$'s. These bounds depend only on c_1, C_1, c_2, C_2 and c_2'. □

Step 4'

Using Lemma 1.49 and an argument virtually identical to the one used in Step 4, we may derive a probability bound we find that there exists c, C, c' and n_0 depending only on $c_0, c_0', c_1, C_1, c_2, C_2$ and c_2' such that for all $n \geqslant n_0$,

$$\mathrm{P}\left(\min_{1 \leqslant k \leqslant q} \kappa_k^{**}(\overline{c}) \geqslant c'\right) \geqslant 1 - Cn^{-c}.$$

Step 5'

By an argument virtually identical to the one in Step 5, Steps 1'–4' show that there exists c, C and C' such that with probability $\geqslant 1 - Cn^{-c}$, $\max_{1\leqslant k\leqslant q}\|\widehat{\beta}_k - \beta_{k*}\|_{2,n} \leqslant C'\sqrt{s\ln(pqn)/n}$. This conclusion completes the proof of Theorem 1.7. □

1.N Proofs for Section 1.H

1.N.1 Proofs for Section 1.H.1

Proof of Lemma 1.7. To establish conservative polynomial validity of the initial penalty loadings we need to show that there exists ℓ, u, c, C and n_0 depending only on c_1, C_1, c_2, C_2, c_2' and C_2' such that for all $n \geqslant n_0$,

$$\mathrm{P}\left(\ell\widehat{\gamma}^*_{kj} \leqslant \widehat{\gamma}_{kj} \leqslant u\widehat{\gamma}^*_{kj} \text{ for all } (j,k) \in [p]\times[q]\right) \leqslant Cn^{-c},$$

with $0 < \ell \leqslant 1 \leqslant u$ and both $\ell \to 1$ and $u \to u' \geqslant 1$ polynomially fast. For this purpose, define (infeasible) penalty loadings

$$\begin{aligned}
\check{\gamma}^2_{kj} &\coloneqq \mathbb{E}_n[(Y_{ik} - \overline{Y}_k)^2] \max_{1\leqslant i\leqslant n} W^2_{ij},\\
\widetilde{\gamma}^2_{kj} &\coloneqq \mathbb{E}_n(\widetilde{Y}^2_{ik}) \max_{1\leqslant i\leqslant n} W^2_{ij},\\
\gamma^2_{kj} &\coloneqq \mathrm{E}(\widetilde{Y}^2_k)M^2_j,\\
\gamma^2_{kj} &\coloneqq \mathrm{E}(\varepsilon^2_k)M^2_j,
\end{aligned}$$

where $\overline{Y}_k \coloneqq \mathbb{E}_n(Y_{ik})$ and $\widetilde{Y}_{ik} \coloneqq Y_{ik} - \mathrm{E}(Y_k)$. Under Assumptions 1.17 and 1.18, the initial penalty loadings,

$$\widehat{\gamma}^2_{kj} = \mathbb{E}_n\{[\widehat{Y}_{ik} - \mathbb{E}_n(\widehat{Y}_{ik})]^2\} \max_{1\leqslant i\leqslant n} W^2_{ij},$$

must satisfy

$$\begin{aligned}
\max_{(j,k)\in[p]\times[q]} |\widehat{\gamma}^2_{kj} - \check{\gamma}^2_{kj}| &= \max_{(i,j)\in[n]\times[p]} W^2_{ij} \max_{1\leqslant k\leqslant q} |\mathbb{E}_n\{[(\widehat{Y}_{ik} - \mathbb{E}_n(\widehat{Y}_{ik})]^2\} - \mathbb{E}_n[(Y_{ik} - \overline{Y}_k)^2]|\\
&= \max_{(i,j)\in[n]\times[p]} W^2_{ij} \max_{1\leqslant k\leqslant q} |\mathbb{E}_n(\widehat{Y}^2_{ik}) - [\mathbb{E}_n(\widehat{Y}_{ik})]^2 - \mathbb{E}_n(Y^2_{ik}) + [\mathbb{E}_n(Y_{ik})]^2|\\
&\leqslant C^2_1\big(\max_{1\leqslant k\leqslant q} |\mathbb{E}_n(\widehat{Y}^2_{ik} - Y^2_{ik})| + \max_{1\leqslant k\leqslant q} |[\mathbb{E}_n(\widehat{Y}_{ik})]^2 - [\mathbb{E}_n(Y_{ik})]^2|\big)\\
&= C^2_1\big(\max_{1\leqslant k\leqslant q} |\mathbb{E}_n[(\widehat{Y}_{ik} + Y_{ik})e_{ik}| + \max_{1\leqslant k\leqslant q} |\mathbb{E}_n(\widehat{Y}_{ik} + Y_{ik})\mathbb{E}_n(e_{ik})|\big)
\end{aligned}$$

$$
\begin{aligned}
&\leqslant 6C_1^3 \max_{1\leqslant k\leqslant q} \mathbb{E}_n(|e_{ik}|) \\
&\leqslant 6C_1^3 \Delta && \text{(CS)} \\
&\leqslant 6C_1^4 \sqrt{s \ln(qn)/n} \text{ with probability } \geqslant 1 - C_2 n^{-c_2}.
\end{aligned}
$$

Assumption 1.19 therefore implies that

$$
\max_{(j,k)\in[p]\times[q]} |\widehat{\gamma}^2_{kj} - \check{\gamma}^2_{kj}| \leqslant 6C_1^4 \sqrt{C_2} n^{-c_2} \text{ with probability } \geqslant 1 - C_2 n^{-c_2}.
$$

Noting that $Y_{ik} - \overline{Y}_k = \widetilde{Y}_{ik} - \mathbb{E}_n(\widetilde{Y}_{ik})$, we must have

$$
\begin{aligned}
\max_{(j,k)\in[p]\times[q]} |\check{\gamma}^2_{kj} - \widetilde{\gamma}^2_{kj}| &= \max_{(i,j)\in[n]\times[p]} W_{ij}^2 \max_{1\leqslant k\leqslant q} |\{\mathbb{E}_n[(Y_{ik} - \overline{Y}_k)^2] - \mathbb{E}_n(\widetilde{Y}^2_{ik})]\} \\
&\leqslant C_1^2 \max_{1\leqslant k\leqslant q} |\mathbb{E}_n\{[\widetilde{Y}_{ik} - \mathbb{E}_n(\widetilde{Y}_{ik})]^2\} - \mathbb{E}_n(\widetilde{Y}^2_{ik})| \\
&= C_1^2 \Big[\max_{1\leqslant k\leqslant q} |\mathbb{E}_n(\widetilde{Y}_{ik})| \Big]^2.
\end{aligned}
$$

By a maximal inequality for bounded random variables (Lemma 1.34), with probability $\geqslant 1 - n^{-1/4}$, for some C' depending only on C_1 we have $\max_{1\leqslant l\leqslant p} |\mathbb{E}_n(\widetilde{Y}_{il})| \leqslant C' \ln(q)/n^{1/4}$, which by the growth condition $\ln^4(q) \leqslant C_2 n^{1-c_2}$, in turn, is $\leqslant C' n^{-c'}$ for c' and (potentially different) C' depending only on C_1, c_2 and C_2. Hence, with the same probability,

$$
\max_{(j,k)\in[p]\times[q]} |\widehat{\gamma}^2_{kj} - \widetilde{\gamma}^2_{kj}| \leqslant C' n^{-c'}, \tag{1.N.1}
$$

for potentially different c' and C'. A maximal inequality for bounded random variables (Lemma 1.34) also shows that, with probability $\geqslant 1 - n^{-1/4}$, we have $\max_{1\leqslant k\leqslant q} |(\mathbb{E}_n - \mathrm{E})(\widetilde{Y}^2_{il})| \leqslant C' \ln(q)/n^{1/4}$, which by the growth condition $\ln^4(q) \leqslant C_2 n^{1-c_2}$, in turn, is $\leqslant C' n^{-c'}$ for c' and (potentially different) C' depending only on C_1, c_2 and C_2. Adding and subtracting $\mathrm{E}(\widetilde{Y}^2_k) \max_{1\leqslant i\leqslant n} W^2_{ij}$ for given (j,k) and using Assumption 1.18, we get

$$
\begin{aligned}
\max_{(j,k)\in[p]\times[q]} |\widetilde{\gamma}^2_{kj} - \gamma^2_{kj}| &= \max_{(j,k)\in[p]\times[q]} |\mathbb{E}_n[(Y_{ik} - \overline{Y}_k)^2] \max_{1\leqslant i\leqslant n} W^2_{ij} - \mathrm{E}(\widetilde{Y}^2_k) M^2_j| \\
&\leqslant \max_{1\leqslant k\leqslant q} |(\mathbb{E}_n - \mathrm{E})(\widetilde{Y}^2_{ik})| \max_{(i,j)\in[n]\times[p]} W^2_{ij} + \max_{1\leqslant k\leqslant q} \mathrm{E}(\widetilde{Y}^2_k) \max_{1\leqslant j\leqslant p} \Big| \max_{1\leqslant i\leqslant n} W^2_{ij} - M^2_j \Big| \\
&\leqslant C_1^2 \max_{1\leqslant k\leqslant q} |(\mathbb{E}_n - \mathrm{E})(\widetilde{Y}^2_{ik})| + C_1^2 \max_{1\leqslant j\leqslant p} \Big| \max_{1\leqslant i\leqslant n} W^2_{ij} - M^2_j \Big| \\
&\leqslant C_1^2 \max_{1\leqslant k\leqslant q} |(\mathbb{E}_n - \mathrm{E})(\widetilde{Y}^2_{ik})| + 2C_1^3 \max_{1\leqslant j\leqslant p} \Big| \max_{1\leqslant i\leqslant n} |W_{ij}| - M_j \Big| \\
&\leqslant C' n^{-c'} \quad \text{wp} \geqslant 1 - Cn^{-c}
\end{aligned} \tag{1.N.2}
$$

for potentially different c, C, c' and C'. Given that $\gamma_{kj}^2 = \mathrm{E}(\widetilde{Y}_k^2)M_j^2 \geqslant \mathrm{E}(\varepsilon_k^2)M_j^2 = \gamma_{kj}^{*2}$, the previous two displays show that for potentially different c, C, c' and C', with probability $\geqslant 1 - Cn^{-c}$ we have

$$\widehat{\gamma}_{kj}^2 \geqslant \gamma_{kj}^2 - C'n^{-c'} \geqslant \gamma_{kj}^{*2} - C'n^{-c'} \text{ for all } (j,k) \in [p] \times [q] \,.$$

A maximal inequality for bounded random variables (Lemma 1.34) shows that, with probability $\geqslant 1 - n^{-1/4}$, we have $\max_{1 \leqslant k \leqslant q} |(\mathbb{E}_n - \mathrm{E})(\varepsilon_{ik}^2)| \leqslant C' \ln(q)/n^{1/4}$, which by the growth condition $\ln^4(q) \leqslant C_2 n^{1-c_2}$, in turn, is $\leqslant C'n^{-c'}$ for c' and (potentially different) C' depending only on C_1, c_2 and C_2. Adding and subtracting $\mathrm{E}(\varepsilon_k^2) \max_{1 \leqslant i \leqslant n} W_{ij}^2$ for given j and using Assumption 1.18, we get

$$\begin{aligned}
\max_{(j,k)\in[p]\times[q]} |\widehat{\gamma}_{kj}^{*2} - \gamma_{kj}^{*2}| &= \max_{(j,k)\in[p]\times[q]} |\mathbb{E}_n(\varepsilon_{ik}^2) \max_{1\leqslant i\leqslant n} W_{ij}^2 - \mathrm{E}(\varepsilon_k^2)M_j^2| \\
&\leqslant \max_{1\leqslant k\leqslant q} |\left(\mathbb{E}_n - \mathrm{E}\right)\left(\varepsilon_{ik}^2\right)| \max_{(i,j)\in[n]\times[p]} W_{ij}^2 + \max_{1\leqslant k\leqslant q} \mathrm{E}(\varepsilon_k^2) \max_{1\leqslant j\leqslant p} \Big| \max_{1\leqslant i\leqslant n} W_{ij}^2 - M_j^2 \Big| \\
&\leqslant C_1^2 \max_{1\leqslant k\leqslant q} |\left(\mathbb{E}_n - \mathrm{E}\right)\left(\varepsilon_{ik}^2\right)| + 2C_1^3 \max_{1\leqslant j\leqslant p} \Big| \max_{1\leqslant i\leqslant n} |W_{ij}| - M_j \Big| \\
&\leqslant C'n^{-c'} \quad \text{wp} \geqslant 1 - Cn^{-c}
\end{aligned}$$

for potentially different c, C, c' and C'. The previous two displays now show that for potentially different c, C, c' and C', with probability $\geqslant 1 - Cn^{-c}$ we have

$$\widehat{\gamma}_{kj}^2 \geqslant \widehat{\gamma}_{kj}^{*2} - C'n^{-c'} \text{ for all } (j,k) \in [p] \times [q] \,.$$

By $\gamma_{kj}^{*2} = \mathrm{E}(\varepsilon_k^2)M_j^2 \geqslant \mathrm{E}(\varepsilon_k^2 W_j^2)$ and Assumption 1.18, the γ_{kj}^*'s are bounded away from zero. The previous diplay shows that for sufficiently large n and with probability approaching one polynomially fast, so are the $\widehat{\gamma}_{kj}^*$'s. This observation combined with the previous display shows that there exists $\ell = \ell_n \in (0, 1]$ nonrandom such that for sufficiently large n

$$\ell\widehat{\gamma}_{kj}^* \leqslant \widehat{\gamma}_{kj} \text{ for all } (j,k) \in [p] \times [q]$$

with probability $\geqslant 1 - Cn^{-c}$, where $\ell \to 1$ polynomially fast. On the other hand, (1.N.1) and (1.N.2) combined with the $\widehat{\gamma}_{kj}^{*2}$'s being bounded away from zero with probability approaching one polynomially fast and the γ_{kj}^2's being bounded from above (Assumption 1.18) show that there exists a $u = u_n \geqslant 1$ nonrandom such that for sufficiently large n

$$\widehat{\gamma}_{kj} \leqslant u\widehat{\gamma}_{kj}^* \text{ for all } (j,k) \in [p] \times [q]$$

with probability $\geqslant 1 - Cn^{-c}$, where $u \to u' \geqslant 1$ polynomially fast. The statement of the lemma now follows from the previous two displays and the union bound. □

Proof of Lemma 1.8. To establish true polynomial validity of the refined penalty loadings we need to show that there exists ℓ, u, c, C and n_0 depending only on $c_0, c_0', c_1, C_1, c_2, C_2$ and c_2' such that for $n \geqslant n_0$,

$$\mathrm{P}\left(\ell\widehat{\gamma}_{kj}^{**} \leqslant \widehat{\gamma}_{kj} \leqslant u\widehat{\gamma}_{kj}^{**} \text{ for all } (j,k) \in [p] \times [q]\right) \leqslant Cn^{-c},$$

with $0 < \ell \leqslant 1 \leqslant u$ and both $\ell \to 1$ and $u \to 1$ polynomially fast. For this purpose, define $\gamma_{kj}^{**2} \coloneqq \mathrm{E}(\varepsilon_k^2 W_j^2)$, which by Assumption 1.18 are bounded from above and away from zero by constants depending only on c_1 and C_1. The claim therefore follows if we can show that there exists c, C, c', C' and n_0 depending only on $c_0, c_0', c_1, C_1, c_2, C_2$ and c_2' such that for all $n \geqslant n_0$,

$$\max_{(j,k)\in[p]\times[q]} |\widehat{\gamma}_{kj}^2 - \widehat{\gamma}_{kj}^{**2}| \leqslant C'n^{-c'} \text{ with probability} \geqslant 1 - Cn^{-c},$$

and that there exists (possibly different) c, C, c', C' and n_0 depending only on $c_0, c_0', c_1, C_1, c_2, C_2$ and c_2' such that for all $n \geqslant n_0$,

$$\max_{(j,k)\in[p]\times[q]} |\widehat{\gamma}_{kj}^{**2} - \gamma_{kj}^{**2}| \leqslant C'n^{-c'} \text{ with probability} \geqslant 1 - Cn^{-c}.$$

Note that $\widehat{\varepsilon}_{ik} = \widehat{Y}_{ik} - W_i^\top \widehat{\beta}_k = e_{ik} + \varepsilon_{ik} - W_i^\top \delta_k$, where $\delta_k \coloneqq \widehat{\beta}_k - \beta_{k*}$. It follows that

$$\begin{aligned}
\max_{(j,k)\in[p]\times[q]} |\widehat{\gamma}_{kj}^2 - \widehat{\gamma}_{kj}^{**2}| &= \max_{(j,k)\in[p]\times[q]} |\mathbb{E}_n[(\widehat{\varepsilon}_{ik}^2 - \varepsilon_{ik}^2)W_{ij}^2]| \\
&= \max_{(j,k)\in[p]\times[q]} |\mathbb{E}_n\{[e_{ik}^2 + 2\varepsilon_{ik}e_{ik} + (W_i^\top\delta_k)^2 - 2(\varepsilon_{ik} + e_{ik})W_i^\top\delta_k]W_{ij}^2\}| \\
&\leqslant \max_{(j,k)\in[p]\times[q]} \mathbb{E}_n\{[e_{ik}^2 + 2|\varepsilon_{ik}||e_{ik}| + (W_i^\top\delta_k)^2 + 2(|\varepsilon_{ik}| + |e_{ik}|)||W_i^\top\delta_k|]W_{ij}^2\} \\
&\leqslant C_1^2 \max_{1\leqslant k\leqslant q} \mathbb{E}_n\{[e_{ik}^2 + 2C_1|e_{ik}| + (W_i^\top\delta_k)^2 + 4C_1|W_i^\top\delta_k|] \\
&\leqslant C_1^2(\Delta^2 + 2C_1\Delta + \max_{1\leqslant k\leqslant q}\|\delta_k\|_{2,n}^2 + 4C_1 \max_{1\leqslant k\leqslant q}\|\delta_k\|_{2,n}),
\end{aligned}$$

where the last inequality follows from Cauchy-Schwarz inequality. The first requirement now follows Assumption 1.17, the hypothesis on the $\widehat{\beta}_k$'s and $s\ln(pqn)/n \leqslant C_2 n^{-c_2}$. The second requirement follows from a maximal inequality for bounded random variables (Lemma 1.34) and $\ln^4(pq) \leqslant C_2 n^{1-c_2}$:

$$\begin{aligned}\mathrm{P}\left(\max_{(j,k)\in[p]\times[q]}|\widehat{\gamma}_{kj}^{**2}-\gamma_{kj}^{**2}|>n^{-1/4}\right) &\leqslant n^{1/4}\mathrm{E}\left(\max_{(j,k)\in[p]\times[q]}|\widehat{\gamma}_{kj}^{**2}-\gamma_{kj}^{**2}|\right)\\ &= n^{1/4}\mathrm{E}\left[\max_{(j,k)\in[p]\times[q]}|\left(\mathbb{E}_n-\mathrm{E}\right)\left(\varepsilon_{ik}^2W_{ij}^2\right)\right]\\ &\leqslant n^{1/4}C'\frac{\ln(pq)}{\sqrt{n}}=C'\left(\frac{\ln^4(pq)}{n}\right)^{1/4}\leqslant Cn^{-c}.\end{aligned}$$

□

1.N.2 Proofs for Section 1.H.2

Proof of Lemma 1.9. The claim will follow from an application of Lemma 1.7 with $q \coloneqq 1$. Assumption 1.17 holds trivially since Y is observed. Assumption 1.18 is implied by Assumption 1.12. Lastly, the growth condition $\ln^4(q) \leqslant C_2 n^{1-c_2}$ is trivial since $q = 1$. □

Proof of Lemma 1.10. The claim will follow from an application of Lemma 1.8 with $q \coloneqq 1$. The performance bound (1.H.4) (with $\widehat{\beta}_1 \coloneqq \widehat{h}$ and $\beta_{1*} \coloneqq h_*$) is implied by Lemma 1.4 and the conservative polynomial validity of the initial penalty loadings (Lemma 1.9). Assumption 1.18 is implied by Assumption 1.12. Lastly, the growth condition $s\ln(pn) \vee \ln^4(p) \leqslant C_2 n^{1-c_2}$ is implied by Assumption 1.14. □

Proof of Lemma 1.11. The claim will follow from an application of Lemma 1.7 with $q \coloneqq q$. Assumption 1.17 is implied by Assumptions 1.10, 1.12, 1.13 and 1.14 (cf. the proof of Lemma 1.4). Assumption 1.18 is implied by Assumptions 1.12 and 1.13. The growth condition $\ln^4(q) \leqslant C_2 n^{1-c_2}$ is implied by Assumption 1.14. □

Proof of Lemma 1.12. The claim will follow from an application of Lemma 1.8 with $q \coloneqq q$. The performance bound (1.H.4) (with $\widehat{\beta}_k \coloneqq \widehat{\mu}_k$ and $\beta_{k*} \coloneqq \mu_{k*}, k \in \{1,\dots,q\}$) is implied by Lemma 1.4 and the conservative polynomial validity of the penalty loadings for both h_* and $\{\mu_{k*}\}_1^q$ estimation (Lemmas 1.9 and 1.11). Assumption 1.18 is implied by Assumptions 1.12 and 1.13. The growth requirement $s\ln(pqn) \vee \ln^4(qn) \leqslant C_2 n^{1-c_2}$ is implied by Assumption 1.14. □

1.O Proofs for Section 1.K

PROOF OF LEMMA 1.29. For $T \subset \{1, \ldots, p\}$, let $B_q^T \coloneqq \{\theta \in \mathbf{R}^p | \, \|\delta\|_q \leqslant 1, \operatorname{supp}(\delta) \subset T\}$ and $D_q^m \coloneqq \cup_{|T| \leqslant m} B_q^T$. Then

$$I \coloneqq \mathrm{E}\Big\{ \sup_{\|\delta\|_0 \leqslant m, \|\delta\|=1} \big| \big(\mathbb{E}_n - \overline{\mathrm{E}}\big) \left[(W_i^\top \delta)^2\right] \big| \Big\} \leqslant \mathrm{E}\Big\{ \sup_{\delta \in D_2^m} \big| \big(\mathbb{E}_n - \overline{\mathrm{E}}\big) \left[(W_i^\top \delta)^2\right] \big| \Big\}.$$

By a symmetrization (Ledoux and Talagrand, 2013, Lemma 6.3),

$$n\mathrm{E}\Big\{ \sup_{\delta \in D_2^m} \big| \big(\mathbb{E}_n - \overline{\mathrm{E}}\big) \left[(W_i^\top \delta)^2\right] \big| \Big\} \leqslant 2\mathrm{E}\Big(\mathrm{E}\Big\{ \sup_{\delta \in D_2^m} \Big| \sum_{i=1}^n [\varepsilon_i (W_i^\top \delta)^2] \Big| \,\Big|\, \{W_i\}_1^n \Big\}\Big), \tag{1.O.1}$$

where $\{\varepsilon_i\}_1^n$ denotes independent Rademacher random variables, independent of $\{W_i\}_1^n$. Let $\{g_i\}_1^n$ denote independent standard Gaussian random variables, independent of both $\{\varepsilon_i\}_1^n$ and $\{W_i\}_1^n$. Given that $\mathrm{E}(|g_i|) = \sqrt{2/\pi}$, $\varepsilon_i |g_i|$ is distributed as g_i, and $\{\varepsilon_i\}_1^n, \{g_i\}_1^n$ and $\{W_i\}_1^n$ are independent,

$$\begin{aligned}
&\mathrm{E}\Big\{ \sup_{\delta \in D_2^m} \Big| \sum_{i=1}^n [\varepsilon_i (W_i^\top \delta)^2] \Big| \,\Big|\, \{W_i\}_1^n \Big\} \\
&= \sqrt{\pi/2}\mathrm{E}\Big\{ \sup_{\delta \in D_2^m} \Big| \sum_{i=1}^n [\varepsilon_i \mathrm{E}\left(|g_i| \,|\, \{\varepsilon_i\}_1^n, \{W_i\}_1^n\right) (W_i^\top \delta)^2] \Big| \,\Big|\, \{W_i\}_1^n \Big\} \\
&\leqslant \sqrt{\pi/2}\mathrm{E}\Big\{ \sup_{\delta \in D_2^m} \Big| \sum_{i=1}^n [\varepsilon_i |g_i| (W_i^\top \delta)^2] \Big| \,\Big|\, \{W_i\}_1^n \Big\} && \text{(Jensen)} \\
&= \sqrt{\pi/2}\mathrm{E}\Big\{ \sup_{\delta \in D_2^m} \Big| \sum_{i=1}^n [g_i (W_i^\top \delta)^2] \Big| \,\Big|\, \{W_i\}_1^n \Big\}. && (1.O.2)
\end{aligned}$$

By Dudley's inequality (Ledoux and Talagrand, 2013, Theorem 11.17)

$$\mathrm{E}\Big\{ \sup_{\delta \in D_2^m} \Big| \sum_{i=1}^n [g_i (W_i^\top \delta)^2] \Big| \,\Big|\, \{W_i\}_1^n \Big\} \lesssim \int_0^{\mathrm{diam}(D_2^m)} \sqrt{\ln N\left(\epsilon, D_2^m, d\right)} \mathrm{d}\epsilon, \tag{1.O.3}$$

where $\mathrm{diam}(D_2^m) = \sup_{\delta, \delta' \in D_2^m} d\left(\theta, \theta'\right)$ for d being the intrinsic L^2 metric (conditional on $\{W_i\}_1^n$). This metric satisfies

$$\begin{aligned}
d\left(\delta, \delta'\right) &\equiv \Big(\mathrm{E}\Big\{ \Big[\sum_{i=1}^n [g_i (W_i^\top \delta)^2] - \sum_{i=1}^n [g_i (W_i^\top \delta')^2 \Big]^2 \,\Big|\, \{W_i\}_1^n] \Big\}\Big)^{1/2} \\
&= \Big\{ \sum_{i=1}^n [(W_i^\top \delta)^2 - (W_i^\top \delta')^2]^2 \Big\}^{1/2}
\end{aligned}$$

$$
\begin{aligned}
&\leqslant \Big[\sum_{i=1}^{n}(W_i^\top\delta + W_i^\top\delta')^2\Big]^{1/2}\max_{1\leqslant i\leqslant n}|W_i^\top\delta - W_i^\top\delta'|\\
&\leqslant 2\sup_{\delta\in D_2^m}\Big[\sum_{i=1}^{n}(W_i^\top\delta)^2\Big]^{1/2}\\
&=: 2R\|\delta-\delta'\|_W\\
&= 2\sqrt{m}R\|\delta/\sqrt{m}-\delta'/\sqrt{m}\|_W
\end{aligned}
$$

where I have defined $R \coloneqq \sup_{\delta\in D_2^m}[\sum_{i=1}^n (W_i^\top\delta)^2]^{1/2}$ and the norm $\|\delta\|_W \coloneqq \max_{1\leqslant i\leqslant n}|W_i^\top\delta|$. The previous display shows that

$$
N\left(\epsilon, D_2^m, d\right) \leqslant N(\epsilon/(2\sqrt{m}R), D_2^m/\sqrt{m}, \|\cdot\|_W).
$$

Note that

$$
\begin{aligned}
\operatorname{diam}(D_2^m) &\leqslant 2\sup_{\delta\in D_2^m}\Big[\sum_{i=1}^{n}(W_i^\top\delta)^4\Big]^{1/2}\\
&\leqslant 2\sup_{\delta\in D_2^m}\Big[\sum_{i=1}^{n}(W_i^\top\delta)^2\|W_i\|_\infty^2\|\delta\|_1^2\Big]^{1/2} && \text{(Hölder)}\\
&\leqslant 2\sup_{\delta\in D_2^m}\Big[\sum_{i=1}^{n}(W_i^\top\delta)^2M^2(\sqrt{m})^2\Big]^{1/2} && (\delta\in D_2^m)\\
&= 2\sqrt{m}MR.
\end{aligned}
$$

The previous two displays combine to show that

$$
\begin{aligned}
\int_0^{\operatorname{diam}(D_2^m)}\sqrt{\ln N\left(\epsilon, D_2^m, d\right)}\mathrm{d}\epsilon &\leqslant \int_0^{2\sqrt{m}MR}\sqrt{\ln N(\epsilon/(2\sqrt{m}R), D_2^m/\sqrt{m}, \|\cdot\|_W)}\mathrm{d}\epsilon\\
&= 2\sqrt{m}R\int_0^M\sqrt{\ln N(\epsilon, D_2^m/\sqrt{m}, \|\cdot\|_W)}\mathrm{d}\epsilon \quad (\epsilon' \coloneqq \epsilon/(2\sqrt{m}R))\\
&=: 2\sqrt{m}R\int_0^M H\left(\epsilon\right)\mathrm{d}\epsilon. \qquad (1.\text{O}.4)
\end{aligned}
$$

If $\delta\in B_2^T/\sqrt{m}$, then $\operatorname{supp}(\delta)\subset T$ and $\|\delta\|\leqslant 1/\sqrt{m}$. It follows that $\|\theta\|_1\leqslant 1$, so $\delta\in B_1^T$. It follows that

$$
B_2^T/\sqrt{m}\subset B_1^T. \qquad (1.\text{O}.5)
$$

Similarly,

$$D_2^m/\sqrt{m} \subset D_1^m \subset B_1, \tag{1.O.6}$$

where $B_1 \coloneqq \{\delta \in \mathbf{R}^p | \, \|\delta\|_1 \leqslant 1\}$. The discussion following the proof of Rudelson and Vershynin (2008, Lemma 3.9) shows that $N(\epsilon, D_1^m, \|\cdot\|_W) \leqslant \sum_{j=1}^m \binom{p}{j}(1+2M/\epsilon)^m$ for all $\epsilon > 0$. Stirling's approximation shows that $\sum_{j=1}^m \binom{p}{j} \leqslant (A_2 p/m)^m \leqslant (A_2 p)^m$ for A_2 universal. This bound and the first containment in (1.O.6) therefore show that

$$\begin{aligned} H(\epsilon) &\leqslant \sqrt{m}[\sqrt{\ln(A_2 p)} + \sqrt{\ln(1+2M/\epsilon)}] \\ &\lesssim \sqrt{m}[\sqrt{\ln p} + \sqrt{\ln(1+2M/\epsilon)}] \eqqcolon H_1(\epsilon) \,. \qquad (p \geqslant 2) \end{aligned}$$

Rudelson and Vershynin (2008, Lemma 3.9) shows that $N(\epsilon, B_1, \|\cdot\|_X) \leqslant (2p)^{A_1 \varepsilon^{-2} M^2 \ln n}$ for all $\epsilon > 0$ and A_1 universal. This bound and the second containment (1.O.6) show that for any $\epsilon > 0$,

$$H(\epsilon) \leqslant \sqrt{\ln[(2p)^{A_1 \epsilon^{-2} M^2 \ln n}]} \lesssim M\sqrt{\ln p}\sqrt{\ln n}\epsilon^{-1} \eqqcolon H_2(\epsilon) \,. \qquad (p \geqslant 2)$$

The previous two displays imply that for any $a \in (0, M]$,

$$\begin{aligned} J(m, M) &\coloneqq 2\sqrt{m}R\int_0^M H(\epsilon)\,\mathrm{d}\epsilon \\ &= 2\sqrt{m}R\Big[\int_0^a H(\epsilon)\,\mathrm{d}\epsilon + \int_a^M H(\epsilon)\,\mathrm{d}\epsilon\Big] \\ &\leqslant 2\sqrt{m}R\Big[\int_0^a H_1(\epsilon)\,\mathrm{d}\epsilon + \int_a^M H_2(\epsilon)\,\mathrm{d}\epsilon\Big]. \end{aligned} \tag{1.O.7}$$

The first integral in (1.O.7) satisfies

$$\int_0^a H_1(\epsilon)\,\mathrm{d}\epsilon \lesssim \sqrt{m}\Big[a\sqrt{\ln p} + \int_0^a \sqrt{\ln(1+2M/\epsilon)}\mathrm{d}\epsilon\Big].$$

Integrating by parts, using $u(\epsilon) \coloneqq \sqrt{\ln(1+2M/\epsilon)}$ and $v'(\epsilon) \coloneqq 1$, the right-hand side integral becomes

$$\begin{aligned} &\int_0^a \sqrt{\ln(1+2M/\epsilon)}\mathrm{d}\epsilon \\ &= \epsilon\sqrt{\ln(1+2M/\epsilon)}\Big|_0^a - \int_0^a \epsilon\frac{1}{2\sqrt{\ln(1+2M/\epsilon)}} \cdot \frac{1}{1+2M/\epsilon} \cdot \frac{(-2M)}{\epsilon^2}\mathrm{d}\epsilon \end{aligned}$$

$$= a\sqrt{\ln(1+2M/a)} + \int_0^a \frac{1}{2\sqrt{\ln(1+2M/\epsilon)}} \cdot \frac{1}{1+2M/\epsilon} \cdot \frac{2M}{\epsilon} \mathrm{d}\epsilon.$$

Substituting $u \coloneqq 2M/\epsilon$, the remaining integral may be written as

$$\begin{aligned}
&\int_0^a \frac{1}{2\sqrt{\ln(1+2M/\epsilon)}} \cdot \frac{1}{1+2M/\epsilon} \cdot \frac{2M}{\epsilon} \mathrm{d}\epsilon \\
&= \int_{+\infty}^{2M/a} \frac{1}{2\sqrt{\ln\left(1+u\right)}} \cdot \frac{1}{1+u} \cdot u \cdot \frac{(-2M)}{u^2} \mathrm{d}u \\
&= M \int_{2M/a}^{+\infty} \frac{1}{\sqrt{\ln\left(1+u\right)}} \cdot \frac{1}{u^2+u} \mathrm{d}u.
\end{aligned}$$

Since $a \in (0, M]$, $2M/a \geqslant 2$, and thus $\ln\left(1 + 2M/a\right) \geqslant 1$. It follows that

$$M \int_{2M/a}^{+\infty} \frac{1}{\sqrt{\ln\left(1+u\right)}} \cdot \frac{1}{u^2+u} \mathrm{d}u \leqslant M \int_{2M/a}^{+\infty} \frac{1}{u^2} \mathrm{d}u = \frac{a}{2}.$$

The four previous displays combine to show that

$$\begin{aligned}
\int_0^a H_1\left(\epsilon\right) \mathrm{d}\epsilon &\lesssim \sqrt{m}\left[a\sqrt{\ln p} + a\sqrt{\ln(1+2M/a)} + \frac{a}{2}\right] \\
&\lesssim a\sqrt{m}\left[\sqrt{\ln p} + \sqrt{\ln(1+2M/a)}\right]. \qquad (p \geqslant 2)
\end{aligned}$$

The second integral in (1.O.7) satisfies

$$\begin{aligned}
\int_a^M H_2\left(\epsilon\right) \mathrm{d}\epsilon &= \sqrt{A_1} M \sqrt{\ln\left(2p\right)} \sqrt{\ln n} \int_a^M \epsilon^{-1} \\
&\lesssim M\sqrt{\ln p}\sqrt{\ln n}\left(\ln M - \ln a\right). \qquad (p \geqslant 2)
\end{aligned}$$

Applying the previous two bounds to (1.O.7), we get

$$\begin{aligned}
J\left(m, M\right) &\lesssim amR\left[\sqrt{\ln p} + \sqrt{\ln(1+2M/a)}\right] \\
&\quad + \sqrt{m}MR\sqrt{\ln p}\sqrt{\ln n}\left(\ln M - \ln a\right).
\end{aligned}$$

Choosing $a \coloneqq M/\sqrt{m}$, which is allowed by $m \geqslant 1$, we arrive at

$$\begin{aligned}
J\left(m, M\right) &\lesssim \sqrt{m}MR\left[\sqrt{\ln p} + \sqrt{\ln(1+2\sqrt{m})} + \left(\ln m\right)\sqrt{\ln p}\sqrt{\ln n}\right] \\
&\lesssim \sqrt{m \ln p} MR\left[1 + \left(\ln m\right)\sqrt{\ln n}\right]. \qquad (1.O.8)
\end{aligned}$$

The sequence of inequalities (1.O.1), (1.O.2), (1.O.3), (1.O.4) and (1.O.8) now yield that for some A universal and $\delta_n(m) \coloneqq A\sqrt{m \ln p}[\mathrm{E}(M^2)]^{1/2}[1 + (\ln m)\sqrt{\ln n}]$,

$$\begin{aligned}
I &\leqslant A\sqrt{m \ln p}\mathrm{E}(MR)\left[1 + (\ln m)\sqrt{\ln n}\right]/n \\
&\leqslant A\sqrt{m \ln p}\left[\mathrm{E}\left(M^2\right)\right]^{1/2}\left[1 + (\ln m)\sqrt{\ln n}\right]\left[\mathrm{E}\left(R^2/n\right)\right]^{1/2} && \text{(Cauchy-Schwarz)} \\
&= \delta_n(m)\left[\mathrm{E}\left(R^2/n\right)\right]^{1/2} \\
&\leqslant \delta_n(m)\Big(I + \sup_{\|\delta\|_0 \leqslant m, \|\delta\|=1} \overline{\mathrm{E}}[(W_i^\top \delta)^2]\Big)^{1/2} && \text{(Triangle)} \\
&\eqqcolon a\sqrt{I+b}.
\end{aligned}$$

Now, for I, a and b nonnegative, $I \leqslant a\sqrt{I+b}$ is equivalent to $I^2 - a^2 I - a^2 b \geqslant 0$. The largest root of this U-shaped quadratic function is $\frac{1}{2}[a^2 + (a^4 + 4a^2 b)^{1/2}] \leqslant \frac{1}{2}(2a^2 + 2a\sqrt{b}) = a^2 + a\sqrt{b}$. Hence

$$I \leqslant \delta_n^2(m) + \delta_n(m) \sup_{\|\delta\|_0 \leqslant m, \|\delta\|=1} \sqrt{\overline{\mathrm{E}}[(W_i^\top \delta)^2]}.$$

□

Proof of Lemma 1.30. Fix A real symmetric and $a > 0$. Suppress the dependence on A such that $\kappa(a) \coloneqq \kappa(a, A), \phi_{\min}(a) \coloneqq \phi_{\min}(a, A)$ and $\phi_{\max}(a) \coloneqq \phi_{\max}(a, A)$. Let T be an arbitrary subset of $\{1, \ldots, p\}$ of cardinality at most s. Without loss of generality we may assume that T is nonempty such that $|T| \geqslant 1$. (If the minimum in the definition of $\kappa(a)$ is attained at $T = \emptyset$, then $\kappa(a) = +\infty$, and there is nothing to prove.) Let $m \in \{1, \ldots, p\}$ and $\delta \neq \mathbf{0}$ such that $\|\delta_{T^c}\|_1 \leqslant a\|\delta_T\|_1$ be arbitrary. Following the proof of Bickel, Ritov, and Tsybakov (2009, Lemma 4.1), partition $T^c = \cup_{k=1}^K T_k$ into $K = \lceil (p - |T|)/m \rceil$ subsets, $\{T_k\}_1^K$, where $|T_k| = m, k = 1, \ldots, K-1$, and $|T_K| \leqslant m$ such that T_k is contains the indices corresponding to the m (in absolute value) largest coordinates of δ outside $T \cup (\cup_{j=1}^{k-1} T_j)$ for $k = 1, \ldots, K-1$, and T_K is the remaining subset. Then using $\delta = \delta_S + \delta_{S^c}$ for any $S \subset \{1, \ldots, p\}$, by the triangle inequality,

$$\|\delta\|_{2,n} \geqslant \|\delta_{T \cup T_1}\|_{2,n} - \|\delta_{(T \cup T_1)^c}\|_{2,n}.$$

Given that $|T \cup T_1| = |T| + |T_1| \leqslant s + m$,

$$\|\delta_{T \cup T_1}\|_{2,n} \geqslant \sqrt{\phi_{\min}(s+m)}\|\delta_{T \cup T_1}\| \geqslant \sqrt{\phi_{\min}(s+m)}\|\delta_T\|.$$

Since $\{T_k\}_1^K$ partition T^c, $(T \cup T_1)^c = T^c \cap T_1^c = \cup_{k=2}^K T_k$, and $|T_k| \leqslant m$,

$$\|\delta_{(T\cup T_1)^c}\|_{2,n} = \Big\| \sum_{k=2}^K \delta_{T_k} \Big\|_{2,n} \leqslant \sum_{k=2}^K \|\delta_{T_k}\|_{2,n} \leqslant \sqrt{\phi_{\max}(m)} \sum_{k=2}^K \|\delta_{T_k}\|,$$

where I have used that $\phi_{\max}$ is a nondecreasing function. By construction of the T_k's, $\max_{j\in T_{k+1}} |\delta_j| \leqslant \min_{j\in T_k} |\delta_j| \leqslant \|\delta_{T_k}\|_1/m$, and thus $\|\delta_{T_{k+1}}\| \leqslant \|\delta_{T_k}\|_1/\sqrt{m}$ for each $k \in \{1, \dots, K-1\}$. Hence

$$\sum_{k=2}^K \|\delta_{T_k}\| \leqslant \frac{1}{\sqrt{m}} \sum_{k=1}^{K-1} \|\delta_{T_k}\|_1 \leqslant \frac{\|\delta_{T^c}\|_1}{\sqrt{m}} \leqslant \frac{c_0\|\delta_T\|_1}{\sqrt{m}}.$$

The previous four displays show that

$$\|\delta\|_{2,n} \geqslant \sqrt{\phi_{\min}(s+m)} \frac{\|\delta_T\|_1}{\sqrt{s}} - a\sqrt{\phi_{\max}(m)} \frac{\|\delta_T\|_1}{\sqrt{m}}.$$

Rearranging we get

$$\frac{\sqrt{s}\|\delta\|_{2,n}}{\|\delta_T\|_1} \geqslant \sqrt{\phi_{\min}(s+m)} - a\sqrt{\frac{\phi_{\max}(m)\, s}{m}}.$$

□

Proof of Lemma 1.31. Denote

$$\begin{aligned} \phi_{\min}(m) &\coloneqq \phi_{\min}\left(m, \mathbb{E}_n(W_i W_i^\top)\right), \\ \phi_{\max}(m) &\coloneqq \phi_{\max}(m, \mathbb{E}_n(W_i W_i^\top)), \\ \text{and } V_n(m) &\coloneqq \sup_{\|\delta\|_0 \leqslant m, \|\delta\|=1} |\left(\mathbb{E}_n - \overline{\mathrm{E}}\right)[(W_i^\top \delta)^2]|. \end{aligned}$$

Given that

$$\phi_{\min}(m) \equiv \inf_{\|\delta\|_0 \leqslant m, \|\delta\|=1} \mathbb{E}_n[(W_i^\top \delta)^2] \geqslant \inf_{\|\delta\|_0 \leqslant m, \|\delta\|=1} \overline{\mathrm{E}}[(W_i^\top \delta)^2] - V_n(m) \geqslant \phi_L - V_n(m),$$

we must have

$$\begin{aligned} \mathrm{P}(\phi_{\min}(m) < \phi_L(m)/2) &\leqslant \mathrm{P}(V_n(m) > \phi_L(m)/2) \leqslant 2\phi_L(m)^{-1}\,\mathrm{E}[V_n(m)] \\ &\leqslant 2\phi_L(m)^{-1}\left[\delta_n^2(m) + \delta_n(m) \sup_{\|\delta\|_0 \leqslant m, \|\delta\|=1} \sqrt{\overline{\mathrm{E}}[(W_i^\top \delta)^2]}\right] \\ &\qquad \text{(Lemma 1.29)} \end{aligned}$$

$$\leqslant 2\phi_L(m)^{-1}\left[\delta_n^2(m) + \delta_n(m)\sqrt{\phi_H(m)}\right].$$

Given that

$$\phi_{\max}(m) \equiv \sup_{\|\delta\|_0 \leqslant m, \|\delta\|=1} \mathbb{E}_n[(W_i^\top \delta)^2] \leqslant \sup_{\|\delta\|_0 \leqslant m, \|\delta\|=1} \overline{\mathrm{E}}[(W_i^\top \delta)^2] + V_n(m) \leqslant \phi_H(m) + V_n(m),$$

we must have

$$\begin{aligned}
\mathrm{P}\left(2\phi_H(m) < \phi_{\max}(m)\right) &\leqslant \mathrm{P}\left(V_n(m) > \phi_H(m)\right) \leqslant \phi_H(m)^{-1}\,\mathrm{E}\left[V_n(m)\right] \\
&\leqslant \phi_H(m)^{-1}\left[\delta_n^2(m) + \delta_n(m) \sup_{\|\delta\|_0 \leqslant m, \|\delta\|=1} \sqrt{\overline{\mathrm{E}}[(W_i^\top \delta)^2]}\right] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Lemma 1.29)} \\
&\leqslant \phi_H(m)^{-1}\left[\delta_n^2(m) + \delta_n(m)\sqrt{\phi_H(m)}\right].
\end{aligned}$$

□

Proof of Lemma 1.32. Denote

$$\begin{aligned}
\phi_{\min}(m) &\coloneqq \phi_{\min}\left(m, \mathbb{E}_n(W_i W_i^\top)\right), \\
\phi_{\max}(m) &\coloneqq \phi_{\max}\left(m, \mathbb{E}_n(W_i W_i^\top)\right), \\
\phi_L(m) &\coloneqq \phi_{\min}\left(m, \overline{\mathrm{E}}(W_i W_i^\top)\right), \\
\text{and } \phi_H(m) &\coloneqq \phi_{\max}\left(m, \overline{\mathrm{E}}(W_i W_i^\top)\right).
\end{aligned}$$

Given that the eigenvalues of $\overline{\mathrm{E}}(W_i W_i^\top)$ are bounded away from zero,

$$\begin{aligned}
\phi_L(s\ln(n) + s) &= \inf_{\|\delta\|_0 \leqslant s\ln(n)+s, \|\delta\|=1} \overline{\mathrm{E}}[(W_i^\top \delta)^2] \\
&\geqslant \min_{\|\delta\|=1} \overline{\mathrm{E}}[(W_i^\top \delta)^2] = \lambda_{\min}\left(\overline{\mathrm{E}}(W_i W_i^\top)\right) \geqslant c_1^2.
\end{aligned}$$

Lemma 1.31 therefore implies that

$$\mathrm{P}\left(\phi_{\min}(s\ln(n) + s) < c_1^2/2\right) \leqslant 2c_1^{-2}\left[\delta_n^2(s\ln(n) + s) + \sqrt{C_1}\delta_n(s\ln(n) + s)\right]$$

for some A universal and

$$\delta_n(m) \coloneqq AC_1\sqrt{\frac{m\ln p}{n}}\left[1 + (\ln m)\sqrt{\ln n}\right], \quad m \geqslant 1.$$

Given that

$$\begin{aligned}\delta_n\left(s\ln(n)+s\right) &\lesssim C_1\sqrt{\frac{s\ln(n)\ln(p)+s\ln(p)}{n}}\left[1+\left(\ln\left(s\ln(n)+s\right)\right)\sqrt{\ln n}\right]\\ &\lesssim C_1\sqrt{\frac{s\ln^5(pn)}{n}},\end{aligned}$$

there exists constants $c_{(1)}$ and $C_{(1)}$ depending only on c_1, C_1, c_2 and C_2 such that

$$\mathrm{P}\left(\phi_{\min}\left(s\ln(n)+s\right)<c_1^2/2\right)\leqslant C_{(1)}n^{-c_{(1)}}.$$

Given that the eigenvalues of $\overline{\mathrm{E}}(W_iW_i^\top)$ are bounded from above,

$$\begin{aligned}\phi_H\left(s\ln n\right) &= \sup_{\|\delta\|_0\leqslant s\ln n,\|\delta\|=1}\overline{\mathrm{E}}[(W_i^\top\delta)^2]\\ &\leqslant \max_{\|\delta\|=1}\overline{\mathrm{E}}[(W_i^\top\delta)^2]=\lambda_{\max}\left(\overline{\mathrm{E}}(W_iW_i^\top)\right)\leqslant C_1^2.\end{aligned}$$

Lemma 1.31 therefore implies that

$$\mathrm{P}\left(2C_1^2<\phi_{\max}\left(s\ln n\right)\right)\leqslant C_1^{-2}\Big[\delta_n^2\left(s\ln n\right)+\sqrt{C_1}\delta_n\left(s\ln n\right)\Big].$$

Given that

$$\begin{aligned}\delta_n\left(s\ln(n)\right) &\lesssim C_1\sqrt{\frac{s\ln(n)\ln(p)}{n}}\left[1+\left(\ln\left(s\ln(n)\right)\right)\sqrt{\ln n}\right]\\ &\lesssim C_1\sqrt{\frac{s\ln^5(pn)}{n}},\end{aligned}$$

there exists constants $c_{(2)}$ and $C_{(2)}$ depending only on c_1, C_1, c_2 and C_2 such that

$$\mathrm{P}\left(2C_1^2<\phi_{\max}\left(s\ln n\right)\right)\leqslant C_{(2)}n^{-c_{(2)}}.$$

The claim now follows from the union bound. □

Proof of Lemma 1.33. By Lemma 1.30, on the event $\mathcal{E}\coloneqq\{c_1^2/2\leqslant\phi_{\min}\left(s\ln(n)+s\right)\leqslant\phi_{\max}\left(s\ln n\right)\leqslant 2C_1^2\}$ we have

$$\kappa(a)\geqslant\max_{1\leqslant m\leqslant p}\left\{\sqrt{\phi_{\min}(s+m)}-a\sqrt{\frac{\phi_{\max}(m)\,s}{m}}\right\}$$

$$\begin{aligned}&\geqslant \sqrt{\phi_{\min}\left(s\ln\left(n\right)+s\right)}-a\sqrt{\frac{\phi_{\max}\left(s\ln n\right)}{\ln n}}\\&\geqslant\sqrt{\frac{c_1^2}{2}}-a\sqrt{\frac{2C_1^2}{\ln n}},\end{aligned}$$

which is $\geqslant \sqrt{c_1/8}$ if and only if $n \geqslant \exp(16a^2C_1^2/c_1^2)$. The claim now follows from Lemma 1.32, which shows that $\mathrm{P}\left(\mathcal{E}^c\right) \leqslant Cn^{-c}$ for c and C depending only on c_1, C_1, c_2 and C_2. □

1.P Proofs for Section 1.L

Proof of Lemma 1.34. Chernozhukov, Chetverikov, and Kato (2015, Lemma 8) shows that

$$\mathrm{E}\left[\max_{1\leqslant j\leqslant p}\left|\mathbb{E}_n\left(X_{ij}\right)\right|\right]\lesssim \max_{1\leqslant j\leqslant p}[\mathrm{E}\left(X_j^2\right)]^{1/2}\sqrt{\frac{\ln p}{n}}+\left[\mathrm{E}\left(\max_{(i,j)\in[n]\times[p]}X_{ij}^2\right)\right]^{1/2}\frac{\ln p}{n}.$$

The claim therefore follows from

$$\mathrm{E}(\max_{(i,j)\in[n]\times[p]}X_{ij}^2)\leqslant\sum_{i=1}^n\mathrm{E}(\max_{1\leqslant j\leqslant p}X_{ij}^2)=n\mathrm{E}(\max_{1\leqslant j\leqslant p}X_j^2)\leqslant nM^2.$$

□

Proof of Lemma 1.36. The maximal inequality (Lemma 1.34) implies

$$\mathrm{E}[\max_{1\leqslant j\leqslant p}\left|\mathbb{E}_n\left(X_{ij}\right)\right|]\leqslant C'\frac{\ln p}{\sqrt{n}},$$

where C' depends only on M. Taking $t = \ln n$, the second part of Talagrand's inequality (Lemma 1.35) combined with the previous display imply that for some C', C'' and C''' depending only on M,

$$\mathrm{P}\left(\max_{1\leqslant j\leqslant p}\left|\mathbb{E}_n\left(X_{ij}\right)\right|>C'\frac{\ln p}{\sqrt{n}}+C''\sqrt{\frac{\ln n}{n}}+C'''\frac{\ln n}{n}\right)\leqslant n^{-1}.$$

Noting that

$$\frac{\ln\left(pn\right)}{\sqrt{n}}\geqslant\max\left\{\frac{\ln p}{\sqrt{n}},\sqrt{\frac{\ln n}{n}},\frac{\ln n}{n}\right\},$$

the claim now follows from recasting C' as the maximum of the three constants. □

Proof of Lemma 1.37. If $\sigma^2 = 0$ then $X = 0$ a.s., and $\mathrm{E}[\psi_2(|X|/C)] = 0$ for any $C > 0$. The infimum of such C's is zero, so $\|X\|_{\psi_2} = 0$ as claimed. Assume therefore $\sigma^2 > 0$. Since $(X/\sigma)^2 \sim \chi_1^2$, $\mathrm{E}\{\exp[(X/C)^2]\} = \mathrm{E}\{\exp[(\sigma/C)^2(X/\sigma)^2]\}$ is the moment-generating function (MGF) of a chi-square random variable with one degree of freedom evaluated at $(\sigma/C)^2$. Guess and verify that $(\sigma/C)^2 < \frac{1}{2}$, such that this MGF is well defined at $(\sigma/C)^2$. Then $\mathrm{E}\{\exp[(\sigma/C)^2(X/\sigma)^2]\} = [1 - 2(\sigma/C)^2]^{-1/2}$. Now $\mathrm{E}[\psi_2(|X|/C)] \leqslant 1$ if and only if $\mathrm{E}\{\exp[(X/C)^2]\} \leqslant 2$, which rearranges to $C \geqslant \sqrt{8/3}\sigma$. Given that $(\sigma/C)^2 < \frac{1}{2}$ rearranges to $C > \sqrt{2}\sigma$, and Since $8/3 > 8/4 = 2$, the earlier guess was indeed correct. □

Proof of Lemma 1.38. By convexity, increasingness and property of C,

$$\psi\left(\mathrm{E}\Big[\max_{1\leqslant j\leqslant p}|X_j|/C\Big]\right) \leqslant \mathrm{E}\Big[\psi\Big(\max_{1\leqslant j\leqslant p}|X_j|/C\Big)\Big] = \mathrm{E}\Big[\max_{1\leqslant j\leqslant p}\psi(|X_j|/C)\Big]$$
$$\leqslant \sum_{j=1}^{p}\mathrm{E}\left[\psi\left(|X_j|/C\right)\right] \leqslant p\max_{1\leqslant j\leqslant p}\mathrm{E}\left[\psi\left(|X_j|/C\right)\right] \leqslant p.$$

so by strict increasingness,

$$\mathrm{E}\Big[\max_{1\leqslant j\leqslant p}|X_j|\Big] \leqslant C\psi^{-1}(p).$$

□

Proof of Lemma 1.39. The function $\psi_2(t) = \mathrm{e}^{t^2} - 1$ is nonnegative, convex, and strictly increasing on $\mathbf{R}_+$. Lemma 1.37 shows that $\|X_j\|_{\psi_2} = \sqrt{8/3}\sigma_j$, so taking $C \coloneqq \max_j\|X_j\|_{\psi_2} = \sqrt{8/3}\max_j\sigma_j$, we must have $\max_j\mathrm{E}\left[\psi_2\left(|X_j|/C\right)\right] \leqslant 1$. Lemma 1.38 therefore yields

$$\mathrm{E}\Big[\max_{1\leqslant j\leqslant p}|X_j|\Big] \leqslant \sqrt{8/3}\sqrt{\ln(1+p)}\max_{1\leqslant j\leqslant p}\sigma_j.$$

Then second inequality follows from $\ln(1+p) < 2\ln p$ for $p \geqslant 2$. □

Proof of Lemma 1.41. The Gaussian maximal inequality (Lemma 1.39) implies that

$$\mathrm{E}\Big[\max_{1\leqslant j\leqslant p}|X_j|\Big] \leqslant K\sigma\sqrt{\ln p},$$

for K universal. Given that

$$K\sigma\sqrt{\ln p} + \sqrt{2}\sigma\sqrt{\ln n} \leqslant (K+\sqrt{2})\sigma\sqrt{\ln(pn)},$$

Borell's inequality with $t \coloneqq \sqrt{\ln n}$, shows that

$$\begin{aligned}\mathrm{P}\left(\|X\|_\infty > (K+\sqrt{2})\sigma\sqrt{\ln{(pn)}}\right) &\leqslant \mathrm{P}\left(\|X\|_\infty > K\sigma\sqrt{\ln p} + \sqrt{2}\sigma\sqrt{\ln n}\right) \\ &\leqslant \mathrm{P}\left(\|X\|_\infty > \mathrm{E}\Big[\max_{1\leqslant j\leqslant p}|X_j|\Big] + \sqrt{2}\sigma\sqrt{\ln n}\right) \leqslant n^{-1}.\end{aligned}$$

Recast the universal constant as $K+\sqrt{2}$. □

Proof of Lemma 1.43. For any $\lambda > 0$, using J followed by subgaussianity

$$\begin{aligned}\mathrm{E}\Big(\max_{1\leqslant j\leqslant p} X_j\Big) &= \frac{1}{\lambda}\mathrm{E}\Big\{\ln\Big[\exp\big(\lambda\max_{1\leqslant j\leqslant p} X_j\big)\Big]\Big\} \leqslant \frac{1}{\lambda}\ln\Big\{\mathrm{E}\Big[\exp\big(\lambda\max_{1\leqslant j\leqslant p} X_j\big)\Big]\Big\} \\ &= \frac{1}{\lambda}\ln\Big[\mathrm{E}\Big(\max_{1\leqslant j\leqslant p}\mathrm{e}^{\lambda X_j}\Big)\Big] \leqslant \frac{1}{\lambda}\ln\Big[\sum_{j=1}^p \mathrm{E}\left(\mathrm{e}^{\lambda X_j}\right)\Big] \leqslant \frac{1}{\lambda}\ln\Big(\sum_{j=1}^p \mathrm{e}^{\lambda^2\sigma^2/2}\Big) \\ &= \frac{1}{\lambda}\ln\left(p\mathrm{e}^{\lambda^2\sigma^2/2}\right) = \frac{\ln p}{\lambda} + \frac{\lambda\sigma^2}{2}.\end{aligned}$$

If $\sigma^2 = 0$, then we may let $\lambda \to \infty$ to obtain a zero upper bound. If $\sigma^2 > 0$, then the right-hand side has minimum $\sigma\sqrt{2\ln p}$, attained at $\lambda = \sqrt{2\ln p}/\sigma$, which establishes the first claim. The second claim follows from the first by noting that

$$\max_{1\leqslant j\leqslant p}|X_j| = \max_{1\leqslant j\leqslant 2p} X_j,$$

where $X_{p+j} \coloneqq -X_j$ for $j \in \{1,\ldots,p\}$. □

1.P.1 Moderate Deviation Inequalities for Self-Normalized Sums

Let $\{X_i\}_1^n$ be independent, centered random variables with $0 < \mathrm{E}(X_i^2) < \infty$. Define

$$\begin{aligned}\mathcal{S}_n &\coloneqq \frac{\sum_{i=1}^n X_i}{(\sum_{i=1}^n X_i^2)^{1/2}} \\ d_{n,\delta} &\coloneqq \frac{[\sum_{i=1}^n \mathrm{E}(X_i^2)]^{1/2}}{[\sum_{i=1}^n \mathrm{E}(|X_i|^{2+\delta})]^{1/(2+\delta)}}, \quad 0 < \delta \leqslant 1.\end{aligned}$$

Theorem 1.8 (de la Pena, Lai, and Shao, 2009, Theorem 7.4). *For any* $0 \leqslant x \leqslant d_{n,\delta}$,

$$\begin{aligned}\frac{\mathrm{P}\left(\mathcal{S}_n \geqslant x\right)}{1-\Phi\left(x\right)} &= 1 + O\left(1\right)\left(\frac{1+x}{d_{n,\delta}}\right)^{2+\delta}, \\ \frac{\mathrm{P}\left(\mathcal{S}_n \leqslant -x\right)}{\Phi\left(-x\right)} &= 1 + O\left(1\right)\left(\frac{1+x}{d_{n,\delta}}\right)^{2+\delta},\end{aligned}$$

where $|O(1)| \leqslant A$ *for* A *absolute.*

Corollary 1.1 (of Theorem 1.8). *For any* $0 \leqslant x \leqslant d_{n,\delta}$,

$$\left|\frac{\mathrm{P}\left(|\mathcal{S}_n| \geqslant x\right)}{2\left[1-\Phi\left(x\right)\right]} - 1\right| \leqslant A\left(\frac{1+x}{d_{n,\delta}}\right)^{2+\delta},$$

for the same absolute constant A *as in Theorem 1.8.*

Proof. Fix $0 \leqslant x \leqslant d_{n,\delta}$. By Theorem 1.8,

$$1 - A\left(\frac{1+x}{d_{n,\delta}}\right)^{2+\delta} \leqslant \frac{\mathrm{P}\left(\mathcal{S}_n \geqslant x\right)}{1-\Phi\left(x\right)} \leqslant 1 + A\left(\frac{1+x}{d_{n,\delta}}\right)^{2+\delta}$$

and

$$1 - A\left(\frac{1+x}{d_{n,\delta}}\right)^{2+\delta} \leqslant \frac{\mathrm{P}\left(\mathcal{S}_n \leqslant -x\right)}{\Phi\left(-x\right)} \leqslant 1 + A\left(\frac{1+x}{d_{n,\delta}}\right)^{2+\delta}.$$

We know $\Phi(-x) = 1 - \Phi(x)$. Now,

$$\mathrm{P}\left(|\mathcal{S}_n| \geqslant x\right) = \mathrm{P}\left(\mathcal{S}_n \geqslant x\right) + \mathrm{P}\left(\mathcal{S}_n \leqslant -x\right)$$

implies both

$$\mathrm{P}\left(|\mathcal{S}_n| \geqslant x\right) \leqslant 2\left[1-\Phi\left(x\right)\right]\left[1 + A\left(\frac{1+x}{d_{n,\delta}}\right)^{2+\delta}\right]$$

and

$$\mathrm{P}\left(|\mathcal{S}_n| \geqslant x\right) \geqslant 2\left[1-\Phi\left(x\right)\right]\left[1 - A\left(\frac{1+x}{d_{n,\delta}}\right)^{2+\delta}\right].$$

□

Lemma 1.50. *Suppose that* $c_1 \leqslant [\overline{\mathrm{E}}(X_i^2)]^{1/2} \leqslant [\overline{\mathrm{E}}(|X_i|^{2+\delta})]^{1/(2+\delta)} \leqslant C_1$ *for some* $0 < \delta \leqslant 1$, *where* c_1 *and* C_1 *depend only on* δ. *If* $0 \leqslant \Phi^{-1}\left(1-\alpha\right) \leqslant d_{n,\delta}$, *then*

$$\mathrm{P}\left(|\mathcal{S}_n| > \Phi^{-1}\left(1-\alpha\right)\right) \leqslant 2\alpha\left[1 + A\left(1 + C_1/c_1\right)^{2+\delta}\right],$$

for the same absolute constant A *as in Theorem 1.8.*

Proof. Given that $0 \leqslant \Phi^{-1}\left(1-\alpha\right) \leqslant d_{n,\delta}$, Corollary 1.1 applies with $x = \Phi^{-1}\left(1-\alpha\right)$. Since we also have

$$d_{n,\delta} = \frac{n^{\delta/(4+2\delta)}[\overline{\mathrm{E}}\,(X_i^2)]^{1/2}}{[\overline{\mathrm{E}}(|X_i|^{2+\delta})]^{1/(2+\delta)}} \geqslant \frac{c_1 n^{\delta/(4+2\delta)}}{C_1} \geqslant \frac{c_1}{C_1},$$

it follows that

$$\begin{aligned}
&\mathrm{P}\left(|\mathcal{S}_n| > \Phi^{-1}\left(1-\alpha\right)\right) \\
&\leqslant 2[1-\Phi(\Phi^{-1}\left(1-\alpha\right))]\left[1 + A\left(\frac{1+\Phi^{-1}\left(1-\alpha\right)}{d_{n,\delta}}\right)^{2+\delta}\right] \\
&\leqslant 2\alpha\left[1 + A\left(C_1/c_1+1\right)^{2+\delta}\right].
\end{aligned}$$

□

Now let $\{X_i\}_1^n$ denote independent, centered $\mathbf{R}^{p\times q}$-valued random variables, where $pq \geqslant 2$. Define

$$\begin{aligned}
\mathcal{S}_{njk} &\coloneqq \frac{\sum_{i=1}^n X_{ijk}}{(\sum_{i=1}^n X_{ijk}^2)^{1/2}}, \\
M_{njk,\delta} &\coloneqq \frac{[\overline{\mathrm{E}}\left(X_{ijk}^2\right)]^{1/2}}{[\overline{\mathrm{E}}(|X_{ijk}|^{2+\delta})]^{1/(2+\delta)}}.
\end{aligned}$$

Proof of Lemma 1.44. Since $\Phi^{-1}\left(1-z\right) \leqslant \sqrt{2\ln\left(1/z\right)}$ for $0 < z < 1$ and $0 \leqslant c_3 \leqslant 1$, we must have

$$\Phi^{-1}\left(1-n^{-c_3}/\left(2pq\right)\right) \leqslant \sqrt{2\ln(2pqn^{c_3})} \leqslant 2\sqrt{\ln(pqn)} \leqslant 2\sqrt{C_2}n^{(1-c_2)/2}.$$

The moment conditions imply $n^{\delta/(4+2\delta)}\min_{1\leqslant j\leqslant p} M_{njk,\delta} \geqslant c_1 n^{\delta/(4+2\delta)}/C_1$. The condition $2\sqrt{C_2}n^{(1-c_2)/2} \leqslant c_1 n^{\delta/(4+2\delta)}/C_1$ rearranges to the condition on $n \geqslant (2C_1\sqrt{C_2}/c_1)^{2/[c_2-8/(8+4\delta)]}$. It follows that $\Phi^{-1}\left(1-n^{-c_3}/\left(2pq\right)\right) \leqslant n^{\delta/(4+2\delta)}\min_{1\leqslant j\leqslant p} M_{njk,\delta}$, so by the union bound and Lemma 1.50 with $\alpha = n^{-c_3}/\left(2pq\right)$,

$$\begin{aligned}
\mathrm{P}\left(\max_{(j,k)\in[p]\times[q]} |\mathcal{S}_{njk}| > \Phi^{-1}\left(1-n^{-c_3}/\left(2pq\right)\right)\right) &\leqslant \sum_{j=1}^p\sum_{k=1}^q \mathrm{P}\left(|\mathcal{S}_{njk}| > \Phi^{-1}\left(1-n^{-c_3}/\left(2pq\right)\right)\right) \\
&\leqslant \sum_{j=1}^p\sum_{k=1}^q 2\frac{n^{-c_3}}{2pq}\left[1 + A\left(1+C_1/c_1\right)^{2+\delta}\right] \\
&= \left[1 + A\left(1+C_1/c_1\right)^{2+\delta}\right] n^{-c_3}.
\end{aligned}$$

□

Chapter 2

Identification and Estimation of a Generalized Panel Regression Model

2.1 Introduction

Many microeconometric panel data models involve limited dependent variables. In limited dependent variable models, the observable dependent variable Y is generated by a latent, unobserved dependent variable Y^* through a map D, which cannot be reversed. Hence, only the "limited" version $Y = D(Y^*)$ of the latent dependent variable Y^* is observed by the econometrician.

Examples of *econometric* models involving limited dependent variables are the panel data binary threshold crossing, censored, truncated and discrete choice models. One example of an underlying *economic* limited dependent variable model is a labor force participation model of married females. Here, the observable dependent variable is whether a married female is participating in the labor market or not, and the unobservable dependent variable is the married female's willingness to participate. Another example is interval coding of wealth or, more generally, data censoring in survey data. In the interval–coding example, the unobservable dependent variable is family wealth, while the observable dependent variable is family wealth recorded up to a wealth bracket of, say, \$100,000–\$125,000. That is, the only thing recorded in the data is that wealth is within the bracket—not the value itself.

Often, the latent dependent variable is assumed to depend on a function h_o of the observable regressors $\mathbf{X}_t$ and unobservable random terms. In the panel setting, the unobservables are divided into a time–invariant unobservable component, or "fixed effect," α and an unobservable random term, or "disturbance," ϵ_t. Hence, the latent dependent variable is typically modeled as $Y_t^* = F(h_o(\mathbf{X}_t), \alpha, \epsilon_t)$ for some map F, which allows us to write the observable dependent variable as

$$Y_t = D \circ F(h_o(\mathbf{X}_t), \alpha, \epsilon_t), \tag{2.1.1}$$

where $D \circ F$ denotes the composite function.

The traditional approach to estimation of limited dependent variable panel data models is to specify a parametric form for the function h_o and for the conditional distribution of the unobservable random components (α, ϵ_t) given the regressors. Furthermore, the composite function $D \circ F$ is typically not only specified up to an unknown parameter but *completely* specified. As a result, such panel data models are vulnerable to the choice of parametric families in which h_o and the conditional distribution of the unobservable random components are assumed to belong, as well as the particular choice of $D \circ F$. Specifically, misspecification of either of these components may lead standard estimators to be inconsistent.

In this paper, I analyze identification and estimation of a class of nonseparable panel data models of the form in (2.1.1), which I refer to jointly as the *generalized panel regression model* (GPRM). I show that it is possible to identify the GPRM without imposing any parametric structure on the function h_o of the regressors, the mapping $D \circ F$ through which the function of regressors, fixed effect, and disturbance term generate the dependent variable, or the conditional distribution of unobservables (α, ϵ_t).

Building on my identification result, I develop a *series maximum rank correlation estimator* (SMRCE) of the function of h_o of the regressors and provide conditions under which consistency in an L^2 sense is achieved. I also provide conditions under which both L^2 and uniform convergence rates of the SMRCE may be derived.

The GPRM assumes that the dependent variable is monotonically related to the function h_o of the regressors, but the form of this relationship is left unspecified. The monotonicity property is a feature shared by the limited dependent variable models mentioned above as well as the panel regression model and some duration, and transformation models. As such, identification and estimation of all of these panel models may be analyzed within a unified framework provided by this paper.

The method of MRC minimizes misspecification biases stemming from an incorrectly imposed form of $D \circ F$ by leaving this composite mapping unspecified. The method of series, or, more generally, sieves, minimizes the possibility of misspecification stemming from an incorrectly imposed parametric form by instead allowing h_o to be nonparametric. Taken together, the method of SMRC estimation thus limits the scope of misspecification and may be used to investigate the validity of estimates obtained under parametric assumptions.

The GPRM admits *arbitrary* dependence between the explanatory variables and the fixed effect and permits a fixed effect of *any* finite dimension. Dependence between the explanatory variables and the fixed effect is often predicted by economic theory, in particular when the explanatory variable is itself an outcome of an agent's decision problem, where unobservable (to the econometrician) random components enter. One example of such a model is the human capital–earnings model, where individuals choose their level of education at least partially based on their own innate ability. Allowing for a *multidimensional* fixed effect to enter in a flexible way permits one to proceed with estimation even in cases where the fixed effect cannot be credibly summarized by a scalar unobservable component which may only enter additively.

Besides leaving $D \circ F$ unspecified, MRC estimation also allows one to estimate the function h_o of the regressors without specifying the conditional distribution of the unobservables. Because the MRCE avoids specifying the conditional distribution of the unobservables, it is said to be *distribution–free*. Distribution–free estimation was introduced to econometrics by Manski (1975), who showed that the parameters of h_o (assumed linear) in a cross–sectional multinomial choice model could be consistently estimated without specifying the distribution of the disturbance terms. Many other papers have developed semiparametric distribution–free methods in the case of the cross–sectional binary choice model (see, for example, Cosslett 1983 and Manski 1985). Manski (1987) showed how to identify and consistently estimate a linear h_o in a binary threshold crossing model involving panel data.

The estimators from the previous papers were robust to misspecification of the (conditional) distribution of the unobserved random component(s), but required h_o to be parametric and, specifically, linear. Matzkin (1992) developed a nonparametric and distribution–free estimator of the cross–sectional binary threshold crossing and binary choice models, thus avoiding misspecification bias stemming from an incorrect functional form for h_o and the distribution of the disturbance term (here assumed independent of the regressors).

The previous papers all assume the map $D \circ F$ to be known, i.e. fully specified. To avoid this potential source of misspecification, Han (1987) introduced the *generalized regression model* in a cross–sectional framework. Han showed that, subject to mild monotonicity requirements, one could consistently estimate the parameters of h_o (assumed linear), while leaving the map $D \circ F$ and the distribution of the disturbance unspecified. Matzkin (1991) extended the analysis of Han by developing a consistent estimator for a nonparametric h_o assumed to belong to a set of increasing, concave, and linearly homogeneous functions. Maintaining the linear h_o assumption, Abrevaya (2000) extended Han in another direction by adapting the model to panel data.

I contribute to two literatures within econometrics, the first being the literature on generalized regression models and, hence, distribution–free methods. My *first* contribution to this literature is to provide conditions under which the function h_o of the regressors is *nonparametrically* identified within a rich class of functions in a generalized *panel* regression model. In order to obtain identification, I assume that the regressors may be partitioned such that h_o is additively separable in each element of the partition, and that the function through which one of the partitioning sets enters is linearly homogeneous and strictly increasing in one argument. Special cases satisfying this requirement include the assumption of (full) linear homogeneity in Matzkin (1991), where the partition is trivial, and the assumption of additive separability into the value of one regressor, where the partitioning separates a single, hence "special," regressor from the remaining variables (see Manski 1985, 1987; Abrevaya 2000 for examples of the use of a special regressor when h_o is assumed linear).[1] *Second,* I provide conditions under which the nonparametric estimator suggested by my identification result is consistent, and *third,* I derive convergence rates in both an L^2–type and uniform distance. The present paper is the first to obtain consistency in the panel GRM, and the first to derive rates of convergence in any nonparametric cross–sectional or panel GRM.[2]

I also contribute to the literature on rates of convergence of series, or, more generally, sieve estimators by establishing convergence rates for a series estimator based on maximization of a *nonsmooth* objective function constructed from *unconventional* conditional moment restrictions. The latter moment conditions are similar to the moment restrictions used to construct the maximum score estimator (MSE) of Manski (1985), which is known to possess nonstandard asymptotic properties.

The fact that the objective function is based on unconventional conditional moment restrictions similar to the MSE implies that the sufficient conditions for derivation of the rates of convergence provided by e.g. Chen and Pouzo (2012) for conventional conditional moment restriction estimators do not apply. (See also Section 2.3.1.) The lack of smoothness implies that general results for convergence rates of series or sieve estimators, such as the

[1] Matzkin modified the arguments of Abrevaya (2000) to obtain nonparametric identification of h_o within a class of increasing, concave, and linearly homogeneous functions (see Matzkin, 2007, Section 4.5.3). Matzkin's identifying assumptions are different from the assumptions invoked in this paper, and neither set of assumptions is nested in the other due to the different treatments of the fixed effect. Hence, the two identification results should be viewed as complementary.

[2] Recently, Berry and Haile (2009) introduced a heterogeneous generalized regression model with group effects and provided sufficient conditions for identification using multiple "special regressor" assumptions. Souza-Rodrigues (2014) developed an estimator of the model of Berry and Haile (2009) and established consistency and convergence rates of the estimator. However, Berry and Haile (2009), and therefore Souza-Rodrigues (2014), conduct their asymptotic analysis as both the number of groups (n) *and* the number of members of each group (T_i) grow without bound. Their asymptotic analysis is therefore more demanding in terms of the observational framework than the typical "large n–small T" panel setting studied in this paper. While the large n–large T_i assumption may appear natural in a study of, say, many and large markets, it is typically unnatural in the context of panel data.

ones provided by Chen (2007), Chen and Shen (1998), Chen and White (1999), or Shen and Wong (1994), do not apply. (See also Remark 2.5.) These comments highlight that the SMRCE must be treated differently than conventional moment restriction estimators.

In the next section I introduce the GPRM and discuss maximum rank correlation in more detail. I provide sufficient conditions for identifying the function h_o of the regressors in Section 2.3 and employ this identification result to develop a series estimator for h_o in Section 2.4. I establish its consistency in Section 2.5, and derive convergence rates of the proposed estimator in Section 2.6. Section 1.6 summarizes. Formal proofs of theorems are in Section 2.A and supporting results in Section 2.B of the appendix.

Notation

Throughout this paper, $\|\mathbf{x}\|$ denotes the Euclidean norm $\|\mathbf{x}\| := \|\mathbf{x}\|_e = (\sum_{i=1}^d x_i^2)^{1/2}$ when applied to a finite–dimensional vector $\mathbf{x} \in \mathbf{R}^d, d \in \mathbf{N}$, $\|\mathbf{A}\|$ the Frobenius norm $\|\mathbf{A}\| := \|\mathbf{A}\|_F = (\sum_{i,j=1}^d a_{ij}^2)^{1/2}$ when applied to a $d \times d$ matrix $\mathbf{A} = (a_{ij})$, and $\|f\|_{\mathcal{X}}$ the supremum norm of a real–valued function f with domain $\mathcal{X} \subset \mathbf{R}^d$, $\|f\|_{\mathcal{X}} := \sup_{\mathbf{x}\in\mathcal{X}} |f(\mathbf{x})|$. For a $d \times d$ matrix $\mathbf{A}$, $\lambda_j(\mathbf{A}), j = 1, \ldots, d$, denotes its eigenvalues, and, in particular, $\underline{\lambda}(\mathbf{A})$ and $\overline{\lambda}(\mathbf{A})$ its smallest and largest eigenvalue, respectively. Given positive numbers a_n and b_n for $n \geqslant 1$, $a_n \lesssim b_n$ and $a_n = O(b_n)$ both mean that a_n/b_n is bounded, and $a_n \asymp b_n$ that both $a_n \lesssim b_n$ and $b_n \lesssim a_n$. The underlying probability space is $(\Omega, \Sigma, \mathbb{P})$. Given $\mathbf{R}^d$–valued random variables $\mathbf{V}_n, n \in \mathbf{N}$, $\mathbf{V}_n \lesssim_P b_n$ and $\mathbf{V}_n = O_P(b_n)$ both denote that $\mathbf{V}_n/b_n$ is bounded in probability, i.e. $\lim_{c\to\infty} \limsup_{n\to\infty} \mathbb{P}(\|\mathbf{V}\| \geqslant cb_n) = 0$ and $\mathbf{V}_n = o_P(b_n)$ denotes that $\mathbf{V}_n/b_n$ converges to zero in probability, i.e. $\lim_{n\to\infty} \mathbb{P}(\|\mathbf{V}_n\| \geqslant cb_n) = 0$ for all $c > 0$. I use E to denote the expectation taking with respect to the distribution ν (the law of $\mathbf{V}$), i.e. $\mathrm{E}(\mathbf{V}) = \int v\nu(\mathrm{d}v)$, and $\mathbb{E}_n$ the expectation relative to the empirical distribution, i.e. $\mathbb{E}_n(\mathbf{V}) = (1/n)\sum_{i=1}^n \mathbf{V}_i$ where $\mathbf{V}_i, i = 1, \ldots, n$, is a random sample from ν. For a function f of a reference random variable $\mathbf{V}$ clear from context, I use $\mathrm{E}(f)$ and $\mathbb{E}_n(f)$ as short for $\mathrm{E}(f(\mathbf{V}))$ and $\mathbb{E}_n(f(\mathbf{V}))$, respectively.

2.2 Model

Let $\mathbf{Z} := (\mathbf{Z}_t : t = 1, \ldots, T) := ((Y_t, \mathbf{X}_t) : t = 1, \ldots, T)$ be a T–period observation from a distribution P with support equal to a subset $\mathcal{S}$ of $\mathbf{R}^{T(1+d_x)}$. Here Y_t is a scalar response variable, and $\mathbf{X}_t$ a vector of regressors of dimension d_x varying over time $t = 1, \ldots, T$. Throughout the paper I consider a sequence of nonparametric generalized regression models for panel data, or *generalized panel regression models* (GPRMs), indexed by the cross–sectional dimension, or "sample size," n,

$$Y_{it} = D \circ F\left(h_o\left(\mathbf{X}_{it}\right), \alpha_i, \epsilon_{it}\right), \quad t = 1, \ldots, T, \quad i = 1, \ldots, n, \tag{2.2.1}$$

where h_o is a real–valued function of d_x variables, α_i is a d_α–dimensional time–invariant unobserved component, or "fixed effect,"[3] ϵ_{it} is a scalar unobserved component, F is a real–valued function strictly increasing in its first and last argument, and D is a real–valued increasing function of its argument. The object of interest is the function h_o or functionals thereof.[4]

As noted in Section 2.1, in the context of limited dependent variable models, one may think of $Y_t^* := F\left(h_o\left(\mathbf{X}_t\right), \alpha, \epsilon_i\right)$ as the latent, unobservable dependent variable and $Y_t = D(Y_t^*)$ as the observable dependent variable. Hence, D may be thought of as the *censoring rule*, or, more generally, *observational rule*, although it need not be given such an interpretation.

Although the assumptions provided in this paper readily extend to the case an arbitrary number of periods T, allowing for a general T introduces unnecessarily complicated notation without much additional insight. I therefore consider the case where two periods are available for each person, and for the remainder of this paper $T = 2$.[5]

As mentioned in Section 2.1, Han (1987) introduced the generalized regression model in a cross–sectional framework, and Abrevaya (2000) extended Han's generalized regression model to allow for panel data.[6] Han and Abrevaya showed that several cross–sectional and panel regression models were nested in their respective frameworks. For example, let the fixed effect α be scalar, h_o linear, such that $h_o(\mathbf{x}) = \mathbf{x}^\top \beta_o$ for β_o in $\mathbf{R}^{d_x}$ unknown, and $F(u_1, u_2, u_3) = u_1 + u_2 + u_3$. If $D(v) = v$, then the model (2.2.1) reduces to the linear panel

[3] Although α_i as a whole is unobserved, time–invariant regressors, such as gender, are absorbed into α_i, which may therefore contain observable elements. In what follows, I need not distinguish between unobservable and observable time–invariant variables. I will therefore refer to them jointly as the "fixed effect."

[4] Examples of such functionals of h_o are 1. the ℓth partial derivative $h_o \mapsto (\partial h_o / \partial x_\ell)(\mathbf{x})$, 2. the ℓth average partial derivative $h_o \mapsto \int (\partial h_o / \partial x_\ell)(\mathbf{x}) \mathrm{d}\mu\left(\mathbf{x}\right)$; and, 3. the conditional average partial derivative $h_o \mapsto \int (\partial h_o / \partial x_\ell) \mathrm{d}\mu(\mathbf{x}|\tilde{\mathbf{x}})$, where, in each case, μ is some known or estimable measure.

[5] Following the arguments made in Charlier, Melenberg, and Soest (1995), I may also extend the analysis to accommodate an unbalanced panel.

[6] Here I treat the model as a model for panel data. However, it may be viewed in the more general context as a model for *group–level data*, In the latter setting, each i corresponds to a group, and a t in $\mathcal{T}_i := \{t : t \text{ is in group } i\}$ a member of group i. Note that the size T_i of $\mathcal{T}_i$ need not be equal for all groups i. Depending on the economic setting one has in mind, in a group–level model one may consider asymptotics as the *number of groups* (n) increases without bound, the *number of members* (T_i) *of each group* increases without bound, or both. Here I consider asymptotics as $n \to \infty$ for $T_i = T$ fixed and small. I therefore find it natural to view the model as a model for panel data as in Abrevaya (2000).

regression model

$$Y_{it} = \mathbf{X}_{it}^{\top}\beta_o + \alpha_i + \epsilon_{it}, \quad t = 1, 2, \quad i = 1, \ldots, n.$$

If instead $D(v) = \mathbf{1}\,(v \geqslant 0)$, where $\mathbf{1}(A)$ is the indicator function taking on the value one if the statement A is true and zero otherwise, we arrive at the linear panel binary threshold crossing model (see, for example, Manski, 1987), and for $D(v) = v\mathbf{1}\,(v \geqslant 0)$ we get the linear panel censored regression model (see, for example, Honoré, 1992). Other notable special cases are transformation and duration models (see Han, 1987; Abrevaya, 2000, for more examples).

Han and Abrevaya both worked under the assumption that h_o takes on a linear form, i.e. $h_o(\mathbf{x}) = \mathbf{x}^{\top}\beta_o$ for β_o in $\mathbf{R}^{d_x}$ unknown. In this paper, I assume that the function h_o satisfies mild regularity conditions but does not belong to a known, finite–dimensional parametric family.

Let $\mathbf{Z}_i, i = 1, \ldots, n$, be a sample of independent observations from P. When $h_o(\mathbf{x}) = \mathbf{x}^{\top}\beta_o$, for suitable choices of "ranking functions" $(y_1, y_2) \mapsto H(y_1, y_2)$ Abrevaya (2000) proposed estimating β_o by a maximizer $\widehat{\beta}_n^H$ of $\widetilde{Q}_n\,(\beta; H)$ on $\mathbf{R}^{d_x}$, where

$$\widetilde{Q}_n\,(\beta; H) := \frac{1}{n}\sum_{i=1}^{n}\left[H\,(Y_{i1}, Y_{i2})\,\mathbf{1}(\mathbf{X}_{i1}^{\top}\beta > \mathbf{X}_{i2}^{\top}\beta) + H\,(Y_{i2}, Y_{i1})\,\mathbf{1}(\mathbf{X}_{i1}^{\top}\beta < \mathbf{X}_{i2}^{\top}\beta)\right]. \quad (2.2.2)$$

Abrevaya called the class of estimators $\widehat{\beta}_n^H$ *rank estimators.*

Rank estimators may be motivated by a simple principle. For concreteness, consider the ranking function $H(y_1, y_2) = \mathbf{1}\,(y_1 > y_2)$. In this case, the objective function $\widetilde{Q}_n(\beta; H)$ provides a measure of the within–individual (positive) association between the outcome Y_{it} and index of regressors $\mathbf{X}_{it}^{\top}\beta$ at the coefficient vector β. For a given i, assuming that the disturbances ϵ_{it} are i.i.d. over time conditional on $\mathbf{X}_{it}, t = 1, 2$, and α_i, the monotonicity of $D \circ F$ guarantees that

$$\begin{gathered}\mathbb{P}\,(Y_{it} \geqslant Y_{is}|\,\mathbf{X}_{i1}, \mathbf{X}_{i2}) \geqslant \mathbb{P}\,(Y_{it} \leqslant Y_{is}|\,\mathbf{X}_{i1}, \mathbf{X}_{i2}) \\ \text{whenever} \quad \mathbf{X}_{it}^{\top}\beta_o \geqslant \mathbf{X}_{is}^{\top}\beta_o, \quad t, s = 1, 2, \quad t \neq s.\end{gathered}$$

In words, it is more likely than not that $Y_{it} \geqslant Y_{is}$ whenever $\mathbf{X}_{it}^{\top}\beta_o \geqslant \mathbf{X}_{is}^{\top}\beta_o$. As ties turn out to play no role, they can be dropped when constructing the criterion function in (2.2.2). Maximizing $\beta \mapsto \widetilde{Q}_n(\beta; H)$ for $H(y_1, y_2) = \mathbf{1}\,(y_1 > y_2)$ yields the *maximum rank correlation estimator* of β_o, but the same principle of estimation carries over to a variety of suitable ranking functions H.[7] The idea of maximum rank correlation and, more generally, rank estimation are closely related to *Kendall's coefficient of concordance*, or *Kendall's Tau*, which is a measure of association between two random variables.[8]

In this paper, I treat h_o as *nonparametric* and estimate the function by maximizing the *sample rank correlation*

$$Q_n(h) = \frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{1}\left(Y_{i1} > Y_{i2}\right)\mathbf{1}(h(\mathbf{X}_{i1}) > h(\mathbf{X}_{i2})) + \mathbf{1}\left(Y_{i1} < Y_{i2}\right)\mathbf{1}(h(\mathbf{X}_{i1}) < h(\mathbf{X}_{i2}))\right] \tag{2.2.3}$$

over a *sieve space* $\mathcal{H}_k$, where $k = k_n$ is a positive integer that increases with n.

It can be computationally difficult to maximize Q_n over an infinite–dimensional parameter space $\mathcal{H}$, and, even if maximization is computationally feasible, the resulting estimator may suffer from inconsistency and/or a slow rate of convergence. These problems arise because maximization over an infinite–dimensional, noncompact space need no longer be a well–posed problem (see, for example, Chen 2007). The method of sieves offers a solution to the issue of ill–posedness by maximizing the objective Q_n over a sequence of less complex parameter spaces $\mathcal{H}_k$ called *sieves* (Grenander 1981*).* To ensure consistency, the complexity of the sieves is required to increase with the sample size n.

2.3 Identification

From (2.2.3) it is clear that $Q_n(h) = Q_n(c_1 h + c_2)$ for constants c_1 and c_2 with $c_1 > 0$. Hence, if h maximizes Q_n, then so does any positive, affine transformation $c_1 h + c_2$ of h. Hence, in order to achieve identification of h_o, at a minimum one needs to fix the scale c_1 and location c_2.

In the case of $h_o(\mathbf{x}) = \mathbf{x}^\top \beta_o$ with β_o in $\mathbf{R}^{d_x}$ unknown, this identification issue is typically overcome by restricting the parameter space to be a $d_x - 1$ dimensional subset of $\mathbf{R}^{d_x}$. Han (1987) fixes the norm of β_o to be one, thus restricting the parameter space to the d_x–dimensional unit sphere. Sherman (1993) uses $\{\beta \in \mathbf{R}^{d_x} : \beta_{d_x} = 1\}$, thus fixing the

[7] See Abrevaya (2000) for examples. The idea of rank estimation goes back to Cavanagh and Sherman (1998) who dealth with estimation of cross–sectional monotonic index models in which Han's (1987) GRM is nested.

[8] Two points are said to be *concordant* if the line joining them has positive slope, and *discordant* if the slope is negative. View $P_1(\beta) := (Y_{i1}, \mathbf{X}_{i1}^\top\beta)$ and $P_2(\beta) := (Y_{i2}, \mathbf{X}_{i2}^\top\beta)$ as two points in the plane. For a given β and the choice of $(y_1, y_2) \mapsto H(y_1, y_2) = \mathbf{1}(y_1 > y_2)$, $\widetilde{Q}_n(\beta; H)$ yields the fraction of the n cross–sectional units whose two data points $(Y_{i1}, \mathbf{X}_{i1}^\top\beta)$ and $(Y_{i2}, \mathbf{X}_{i2}^\top\beta)$ are concordant at β. As such, the criterion resembles *Kendall's coefficient of concordance* as a function of β, although here I do not penalize discordant points.

coefficient on the last regressor to one, which turns out to be a more convenient normalization for the purpose of deriving asymptotics.

In the nonparametric case, the identification problem may be overcome in a number of different ways. Typical restrictions are shape and/or separability conditions.[9] Here I proceed by assuming that the set of basic regressors may be partitioned into two parts such that 1. h_o is additively separable in each collection of variables, and 2. for one part of the partition the associated mapping is linearly homogeneous and strictly increasing in one argument. This assumption nests two special cases: 1. h_o is additively separable into the value of a single regressor (the "special regressor"); and, 2. h_o is (fully) linearly homogeneous and strictly increasing in one of its arguments.

For identification purposes I make the following assumptions.

A. 1 (Mappings D and F). *The functions D and F in (2.2.1) satisfy: 1. D is nonconstant; 2. D is increasing; 3. F is strictly increasing in its first and last argument.*

A. 2 (Disturbances). *Conditional on $(\mathbf{X}_{i1}, \mathbf{X}_{i2}, \alpha_i)$, ϵ_{i1} and ϵ_{i2} are independently and identically distributed (i.i.d.).*

For any subset π' of the indices $\{1, \ldots, d_x\}$, let $\mathbf{x}_{\pi'}$ denote the vector of regressors associated with $\pi', \mathbf{x}_{\pi'} := (x_j)_{j \in \pi'}$, and $\mathbf{x}_{-\pi'}$ the—possibly empty—vector of all other regressors, $\mathbf{x}_{-\pi'} := (x_j)_{j \notin \pi'}$. Let $|\pi'|$ denote the cardinality of the selection π'. The following assumption implicitly defines the parameter space.

A. 3 (Parameter Space). *1. $\Psi_{\pi'}$ is the space of real–valued, continuous, linearly homogeneous functions ψ with domain $\mathcal{D}_{\pi'} = \mathbf{R}^{|\pi'|}$ that are strictly increasing in the first coordinate and satisfy $\psi(\overline{\mathbf{x}}_{\pi'}) = \overline{c}_{\pi'}$ for some $\overline{\mathbf{x}}_{\pi'}$ in $\mathcal{D}_{\pi'}$ and some constant $\overline{c}_{\pi'}$. 2. $\Phi_{\pi'}$ is the space of real–valued, continuous functions φ with domain $\mathcal{D}_{-\pi'} \subset \mathbf{R}^{|-\pi'|}$ satisfying $\varphi(\overline{\mathbf{x}}_{-\pi'}) = \overline{c}_{-\pi'}$ for some $\overline{\mathbf{x}}_{-\pi'}$ in $\mathcal{D}_{-\pi'}$ and some constant $\overline{c}_{-\pi'}$. 3. $\mathcal{H}_{\pi'}$ is the set of functions h that may be written as $(\mathbf{x}_{\pi'}, \mathbf{x}_{-\pi'}) \mapsto h(\mathbf{x}_{\pi'}, \mathbf{x}_{-\pi'}) = \psi(\mathbf{x}_{\pi'}) + \varphi(\mathbf{x}_{-\pi'})$ for some ψ in $\Psi_{\pi'}$ and some φ in $\Phi_{\pi'}$. 4. For some known, nonempty subset π of $\{1, \ldots, d_x\}$, h_o is in $\mathcal{H}_{\pi}$.*

Let $\mathbf{X}_\pi := \{\mathbf{X}_{\pi,t}\}_{t=1}^2$ and $\mathbf{X}_{-\pi} := \{\mathbf{X}_{-\pi,t}\}_{t=1}^2$ denote the π–regressors and non–π–regressors in both periods, $\mathbf{X}_t := (\mathbf{X}_{\pi,t}, \mathbf{X}_{-\pi,t})$ all regressors in period t, and $\mathbf{X} := \{\mathbf{X}_t\}_{t=1}^2$ all regressors in both periods.

A. 4 (Observables). *Let π be as in A.3. 1. Conditional on $\mathbf{X}_{-\pi}$, $\mathbf{X}_\pi$ possesses a Lebesgue density on $\mathbf{R}^{2|\pi|}$. 2. For every $s \neq t$ and all $\mathbf{x} = (\mathbf{x}_s, \mathbf{x}_t)$ satisfying $h_o(\mathbf{x}_s) < h_o(\mathbf{x}_t)$ there exists $y^* := y^*(\mathbf{x})$ such that $\mathbb{P}\left(Y_s \leqslant y^* | \mathbf{X}_s = \mathbf{x}_s\right) > \mathbb{P}\left(Y_t \leqslant y^* | \mathbf{X}_t = \mathbf{x}_t\right)$.*

[9] For an excellent overview of methods for achieving identification in the class of nonseparable, single–index models in which (2.2.1) belongs, see Matzkin (2007).

Remark 2.1. 1. A.1.1 rules out a trivial model where Y_{it} is constant, while A.1.2 and A.1.3 restate the monotonicity assumptions maintained in the GPRM.[10]

2. A.2 states that for every cross–sectional unit, conditional on the regressors in all periods and the fixed effect, the disturbances are i.i.d. over time. This assumption may be viewed as the GPRM analog of the assumption of strict exogeneity conditional on the fixed effect often invoked in the analysis of the linear panel regression model.[11]

3. A.3 states that the basic regressors may be partitioned into two parts such that h_o is additively separable in each part of the partition, and one part of the partition enters a linearly homogeneous function, which is strictly increasing in at least one element. Given this assumption, I define the parameter space $\mathcal{H}$ as $\mathcal{H} := \mathcal{H}_\pi$.

4. It is possible that h_o has multiple representations satisfying A.3. If multiple selections satisfy A.3.4, then I choose one such selection π and consider identification of h_o within that particular space $\mathcal{H}_\pi$.

5. Given the selection π, for notational convenience I assume that $\mathbf{X}_{\pi,t}$ constitutes the *first* $|\pi|$ basic regressors, and $\mathbf{X}_{-\pi,t}$ the *last* $|-\pi| = d_x - |\pi|$ basic regressors. Recasting π as the index corresponding to the "cutoff" regressor determining the partition, I may then write the regressors as $\mathbf{X}_t = (\mathbf{X}_\pi, \mathbf{X}_{-\pi}) = (X_1, \ldots, X_\pi, X_{\pi+1}, \ldots, X_{d_x})$. After this relabeling, I may write any h in $\mathcal{H}$ as

$$h(\mathbf{x}_\pi, \mathbf{x}_{-\pi}) = \psi\left(\mathbf{x}_\pi\right) + \varphi\left(\mathbf{x}_{-\pi}\right) = \psi\left(x_1, \ldots, x_\pi\right) + \varphi\left(x_{\pi+1}, \ldots, x_{d_x}\right).$$

6. Given the relabeling of the regressors provided by the preceding remark, A.4.1 states that the first π regressors in both time periods are jointly continuously distributed with full support regardless of the values of the remaining regressors. This assumption is employed to show that the regressor values allowing one to distinguish h_o from other functions in $\mathcal{H}$ occur with positive probability. Similar continuity assumptions are frequently invoked in the semiparametric literature, when the index function is assumed to be linear.[12]

7. A.4.2 is a (high–level) requirement on the support of the outcomes. A.1 and A.2

[10] The assumption of increasingness, as opposed to decreasingness, is immaterial. As long as D is monotone and F strictly monotone in the relevant arguments, we can always recast the model in a way such that A.1.2 and A.1.3 hold.

[11] The regressors $\mathbf{X}_{it}, t = 1, \ldots, T$, are said to be *strictly exogenous conditional on the fixed effect* if $\mathrm{E}(\epsilon_{it} | \mathbf{X}_{i1}, \ldots, \mathbf{X}_{iT}, \alpha_i) = 0$ for all $t = 1, \ldots, T$.

[12] See, for example, Abrevaya (2000) Assumption 3(b), in the context of GPRM, or Manski (1987) Assumption 2(b), in the context of a panel binary threshold crossing model. Both papers impose a linear form on the index function and may therefore specialize the continuity assumption to the *differences* of, say, the first regressor. As linear homogeneity does not necessarily imply linearity, I cannot make a similar reduction. Instead I place the continuity assumption directly on the joint distribution.

together yield that if $h_o(\mathbf{X}_1) < h_o(\mathbf{X}_2)$ then $\mathrm{P}(Y_1 \leqslant y | \mathbf{X}_1) \geqslant \mathrm{P}(Y_2 \leqslant y | \mathbf{X}_2)$ for all y in $\mathbf{R}$. A.4.2 guarantees that the inequality is strict for some y^*, possibly depending on the regressors. This assumption is used to show that h_o maximizes a certain population objective function uniquely.

8. To simplify notation further, for the remainder of the paper I assume that the point of normalization $\overline{\mathbf{x}}_{-\pi}$ and the value at said normalization point $\overline{c}_{-\pi}$ are zeros. In other words, all φ pass through their respective origin. As a consequence of the linear homogeneity of each ψ in Ψ_π, so does the h for which $\mathbf{x} \mapsto h(\mathbf{x}) = \psi(\mathbf{x}_\pi) + \varphi(\mathbf{x}_{-\pi})$. Although this additional assumption is irrelevant for the purpose of establishing identification, it facilitates the construction of a sieve in Section 2.4.

The following theorem constitutes the main implication provided by the GPRM.

Theorem 2.1 (Main Implication). *If A.1 and A.2 hold, then*

$$\mathbb{P}(Y_1 > Y_2 | \mathbf{X}) \left\{ \begin{matrix} \geqslant \\ \leqslant \end{matrix} \right\} \mathbb{P}(Y_1 < Y_2 | \mathbf{X}) \text{ whenever } h_o(\mathbf{X}_1) \left\{ \begin{matrix} \geqslant \\ \leqslant \end{matrix} \right\} h_o(\mathbf{X}_2)$$

The two probability statements in Theorem 2.1 may be viewed as conditional moment inequalities provided by the true h_o. Such moment inequalities may be exploited to (point) identify h_o. For this purpose, define the *population objective function* Q by

$$\begin{aligned} Q(h) &:= \mathrm{E}\left(\mathbf{1}(Y_1 > Y_2)\,\mathbf{1}(h(\mathbf{X}_1) > h(\mathbf{X}_2)) + \mathbf{1}(Y_1 < Y_2)\,\mathbf{1}(h(\mathbf{X}_1) < h(\mathbf{X}_2))\right) \\ &= \mathbb{P}(Y_1 > Y_2, h(\mathbf{X}_1) > h(\mathbf{X}_2)) + \mathbb{P}(Y_1 < Y_2, h(\mathbf{X}_1) < h(\mathbf{X}_2)) \end{aligned}$$

for h in $\mathcal{H}$. Note that, disregarding equalities, the probability statements in the definition of Q mimick the probability statements provided by Theorem 2.1, except that h_o is replaced by an arbitrary h in $\mathcal{H}$. Note also that Q equals the expectation of the sample objective, i.e. $Q(h) = \mathrm{E}(\mathbb{E}_n(f))$.

The next theorem is the first main result of this paper.

Theorem 2.2 (Identification). *If A.1–A.4 hold, then h_o is the unique maximizer of Q on $\mathcal{H}$.*

Theorem 2.2 shows both 1. h_o is a maximizer of Q on $\mathcal{H}$— thus establishing *existence* of a solution—and 2. the solution is *unique.* Hence, under the assumptions stated in Theorem 2.2, h_o maximizes the probability of *rank correlation,* or *positive concordance*, in the population. The theorem provides a way of identifying h_o within $\mathcal{H}$ in a constructive manner. In Section

2.4 I build on this identification result to construct an estimator of h_o as a solution to the associated sample problem.

2.3.1 Comparison with Conventional Moment Equality and Inequality Models

Chen and Pouzo (2012) considers estimation of *conditional moment restriction models* (CMRMs) that may be written in the form

$$\mathrm{E}\left(r\left(\mathbf{Y}, \mathbf{X}; \theta\right) \middle| \mathbf{W}\right) = 0,$$

where r is a potentially nonsmooth "generalized residual" function known up to the potentially infinite–dimensional parameter θ, and $\mathbf{W}$ a collection of instrumental variables. Given that the main implication of the GPRM is a collection of CMRs, one may wonder whether the GPRM is simply a CMRM (in the Chen and Pouzo 2012 sense) in disguise.

To see that the GPRM is *not* a conventional CMRM, note that if the CMRs provided by Theorem 2.1 hold with strict inequality, then they may be summarized as

$$\operatorname{sgn}\left(\mathbb{P}\left(Y_1 > Y_2 \middle| \mathbf{X}\right) - \mathbb{P}\left(Y_1 < Y_2 \middle| \mathbf{X}\right)\right) = \operatorname{sgn}\left(h_o\left(\mathbf{X}_1\right) - h_o\left(\mathbf{X}_2\right)\right), \tag{2.3.1}$$

where $\operatorname{sgn}(u)$ denotes the sign function

$$\operatorname{sgn}\left(u\right) := \begin{cases} 1, & u > 0 \\ 0, & u = 0 \\ -1, & u < 0. \end{cases}$$

Due to the nonlinearity of the sign function, one cannot simply "pull out" the expectation in (2.3.1) to write the main implication as a conventional CMR. As the "generalized residual" function implicit in (2.3.1) is not known up to h_o (it depends on all the parameters of the model in an unknown way), the GPRM does not fall within the class of models treated by Chen and Pouzo (2012).

Computing expectations over $\mathbf{X}$, we arrive at the *unconditional* moment restriction

$$\mathrm{E}_{\mathbf{X}}\left(\operatorname{sgn}\left(\mathbb{P}\left(Y_1 > Y_2 \middle| \mathbf{X}\right) - \mathbb{P}\left(Y_1 < Y_2 \middle| \mathbf{X}\right)\right)\right) = \mathrm{E}_{\mathbf{X}}\left(\operatorname{sgn}\left(h_o\left(\mathbf{X}_1\right) - h_o\left(\mathbf{X}_2\right)\right)\right). \tag{2.3.2}$$

As in the case of the conditional moment restriction, (2.3.2) is *not* a conventional unconditional moment restriction in the sense of Hansen (1982) and the generalized method of moments (GMM) framework.[13] Hence, the GPRM does not fit into the conventional moment restriction estimation framework, be it conditional or unconditional.

One may, however, summarize the GPRM as an unconditional moment *inequality* model, albeit at a loss of information. To see this, write the Theorem 2.1 inequalities as

$$[\mathbb{P}(Y_1 > Y_2|\, \mathbf{X}) - \mathbb{P}(Y_1 < Y_2|\, \mathbf{X})]\,[h_o(\mathbf{X}_1) - h_o(\mathbf{X}_2)] \geqslant 0 \text{ for all } (\mathbf{x}_1, \mathbf{x}_2) \in \mathcal{X} \times \mathcal{X}.$$

Declaring the left–hand side as the generalized residual does not suffice to define a conventional conditional moment inequality model, as the resulting generalized residual is still not known up to h. Define therefore instead $r(\mathbf{Y}, \mathbf{X}; h) := [\mathbf{1}(Y_1 > Y_2) - \mathbf{1}(Y_1 < Y_2)][h(\mathbf{X}_1) - h(\mathbf{X}_2)]$, which is a function known up to h. Then taking the expectation with respect to $\mathbf{X}$ and using the law of iterated expectations, we may convert the set of inequalities given by Theorem 2.1 into the single unconditional moment inequality $\mathrm{E}(r(\mathbf{Y}, \mathbf{X}; h)) \geqslant 0$. Note that the integrand r is a smooth function of the parameter h. However, if we were to use this single moment inequality to narrow down the infinite–dimensional parameter space $\mathcal{H}$, the resulting identified set is likely to remain large, and hence uninformative. The above discussion illustrates that there is likely to be a trade-off between smoothness of the integrand in the parameter and the level of identification. If we convert the *unconventional conditional* moment *equalities* provided by the main implication of the GPRM into a *conventional unconditional* moment *inequality*, then we will likely have to give up point identification.

2.4 Estimation

The parameter of interest is h_o, the index function in (2.2.1). The parameter space is $\mathcal{H} = \mathcal{H}_\pi$, where $\mathcal{H}_\pi$ is defined in A.3. I approximate h_o by *linear forms* $\mathbf{x} \mapsto p^k(\mathbf{x})^\top \beta$, a *series*, where

$$\mathbf{x} \mapsto p^k(\mathbf{x}) := (p_1(\mathbf{x}), \ldots, p_k(\mathbf{x}))^\top,$$

is a k–vector of *approximating functions*, or, *basis functions*. These basis function may change with k, although I do not express such possible dependence in the notation. Throughout this paper I maintain the assumption that the number of series terms $k = k_n$ is chosen such that $\ln(k) \lesssim \ln(n)$.

Given that h_o is in $\mathcal{H}_\pi$, it is of the form $h_o(\mathbf{x}_\pi, \mathbf{x}_{-\pi}) = \psi_o(\mathbf{x}_\pi) + \varphi_o(\mathbf{x}_{-\pi})$. To achieve the approximation above I therefore approximate both ψ_o and φ_o by linear forms $\mathbf{x}_\pi \mapsto$

[13] The unconditional moment condition is similar to the moment condition on which the maximum score estimator is based (see Manski 1985, p. 315).

$q^{k_\pi}(\mathbf{x}_\pi)^\top \gamma$ and $\mathbf{x}_{-\pi} \mapsto r^{k_{-\pi}}(\mathbf{x}_{-\pi})^\top \delta$, where

$$\begin{aligned}\mathbf{x}_\pi \mapsto & q^{k_\pi}(\mathbf{x}_\pi) := (q_1(\mathbf{x}_\pi), \ldots, q_{k_\pi}(\mathbf{x}_\pi))^\top, \\ \mathbf{x}_{-\pi} \mapsto & r^{k_{-\pi}}(\mathbf{x}_{-\pi}) := \left(r_1(\mathbf{x}_{-\pi}), \ldots, r_{k_{-\pi}}(\mathbf{x}_{-\pi})\right)^\top,\end{aligned}$$

are $k_\pi-$ and $k_{-\pi}$–vectors of approximating functions $\{q_j\}$ and $\{r_j\}$, respectively. To construct the p–basis, I set $k := k_\pi + k_{-\pi}$ and collect the $q-$ and r–bases and their coefficients as

$$\mathbf{x} \mapsto p^k(\mathbf{x}) := (q^{k_\pi}(\mathbf{x}_\pi)^\top, r^{k_{-\pi}}(\mathbf{x}_{-\pi})^\top)^\top, \quad \beta := (\gamma^\top, \delta^\top)^\top. \tag{2.4.1}$$

For each ℓ in $\mathbf{N}$, define the function spaces

$$\begin{aligned}\Psi_\ell := & \big\{\psi_\gamma = q^{\ell\top}\gamma : \psi(a\mathbf{x}_\pi) = a\psi(\mathbf{x}_\pi) \text{ for all } a \in \mathbf{R}, \psi(\overline{\mathbf{x}}_\pi) = \overline{c}_\pi \\ & x_1 \mapsto \psi(x_1, x_2, \ldots, x_\pi) \text{ strictly increasing}, \gamma \in \mathbf{R}^\ell\big\}, \\ \Phi_\ell := & \big\{\varphi_\delta = r^{\ell\top}\delta : \varphi(\mathbf{0}) = 0, \delta \in \mathbf{R}^\ell\big\},\end{aligned}$$

where $\mathbf{0}$ is the $(d_x - \pi)$–dimensional zero. Then, under mild conditions on the choice of the bases, Ψ_ℓ is a subset of $\Psi := \Psi_\pi$ defined as in A.3.1, and Φ_ℓ is a subset of $\Phi := \Phi_\pi$ defined as in A.3.2 (restricting Φ_ℓ to the domain of Φ).

For each k_π and $k_{-\pi}$ in $\mathbf{N}$, let Γ_{k_π} and $\Delta_{k_{-\pi}}$ be compact subsets of $\mathbf{R}^{k_\pi}$ and $\mathbf{R}^{k_{-\pi}}$, respectively. For $k = k_\pi + k_{-\pi}$, define $\mathcal{B}_k$ as $\mathcal{B}_k := \mathcal{B}_{k_\pi, k_{-\pi}} := \Gamma_{k_\pi} \times \Delta_{k_{-\pi}}$. Then $\mathcal{B}_k$ is a compact subset a $\mathbf{R}^k$. Define the sieve space $\mathcal{H}_k$ as the space of functions spanned by p^k subject to the coefficients in $\mathcal{B}_k$:

$$\mathcal{H}_k := \mathcal{H}_k(\mathcal{B}_k) := \big\{\psi_\gamma(\mathbf{x}_\pi) + \varphi_\delta(\mathbf{x}_{-\pi}) : \psi_\gamma \in \Psi_{k_\pi}, \varphi_\delta \in \Phi_{k_{-\pi}}, (\gamma^\top, \delta^\top) \in \mathcal{B}_k\big\}. \tag{2.4.2}$$

Then $\mathcal{H}_k$ is a subset of $\mathcal{H} = \mathcal{H}_\pi$ as defined in A.3.3. In what follows, I will interchangeably write an element of $\mathcal{H}_k$ as h_β or $p^{k\top}\beta$ with the understanding that p^k and β are constructed as in (2.4.1).

For any h in $\mathcal{H}_k$ and $\mathbf{z} = (y_1, \mathbf{x}_1, y_2, \mathbf{x}_2)$ in the support, we may now define

$$f_h(\mathbf{z}) := \mathbf{1}(y_1 > y_2)\mathbf{1}(h(\mathbf{x}_1) > h(\mathbf{x}_2)) + \mathbf{1}(y_1 < y_2)\mathbf{1}(h(\mathbf{x}_1) < h(\mathbf{x}_2)).$$

Given a random sample $\mathbf{Z}_i, i = 1, \ldots, n$, let Q_n be the empirical average of f_h on the sieve space $\mathcal{H}_{k_n}$:

$$Q_n(h) := \mathbb{E}_n(f_h) = \frac{1}{n}\sum_{i=1}^{n} f_h(\mathbf{Z}_i), \quad h \in \mathcal{H}_{k_n}. \tag{2.4.3}$$

I define a *series maximum rank correlation estimator* (SMRCE) as any maximizer $\widehat{h}_n$ of Q_n on $\mathcal{H}_{k_n}$.

Remark 2.2. 1. Since Q_n is a step function with range of finite cardinality, a maximizer on $\mathcal{H}_{k_n}$ always exists. To see this, note that for each k in $\mathbf{N}$, the hyperplanes $\{\beta \in \mathbf{R}^k : [p^k(\mathbf{X}_{i1}) - p^k(\mathbf{X}_{i2})]^\top \beta = 0\}, i = 1, \ldots, n$, divide $\mathbf{R}^k$ into (at most) $\sum_{j=0}^{k} \binom{n}{j}$ regions (see Schläfli 1901, as quoted on p. 921 of Dudley 1978). On each region the function $\beta \mapsto Q_n(h_\beta)$ is necessarily constant. As a consequence, Q_n takes at most $\sum_{j=0}^{k} \binom{n}{j}$ steps over $\mathcal{H}_{k_n}$. The constancy over each region also implies that there will generally be a continuum of maximizers.

2. The preceding remark shows that the compactness of $\mathcal{B}_k$ is not needed to establish existence of a maximizer. Indeed, even if $\mathcal{B}_k = \mathbf{R}^k$, the sample objective function Q_n will have a maximizer on $\mathcal{H}_k(\mathbf{R}^k)$. However, the *population* objective Q need not have range of finite cardinality on $\mathcal{H}_k$, and the compactness of $\mathcal{B}_k$ is used to establish existence of a maximizer of Q on $\mathcal{H}_k$ or a closed subset thereof.

2.5 Consistency

Given that h_o belongs to an infinite–dimensional space, to address the question of consistency of the series maximum rank correlation estimator (SMRCE), I must choose the metric in which convergence is defined. For this purpose, define the map ρ on $\mathcal{H} \times \mathcal{H}$ by

$$\rho(h_1, h_2) := (1/2)(\|h_1 - h_2\|_{1,2} + \|h_1 - h_2\|_{2,2}),$$

where $\|h_1 - h_2\|_{t,2} := (\int_{\mathcal{X}} |h_1 - h_2|^2 \, \mathrm{d}\nu_t)^{1/2}$ defines the L^2–distance between h_1 and h_2 using the distribution ν_t of $\mathbf{X}_t, t = 1, 2$. Because ρ is the average of $L^2(\nu_t)$ metrics over t, it is itself a metric. Hence, $\mathcal{H}$ endowed with ρ is a metric space.[14,15]

Define the positive–semidefinite $k \times k$ matrices $\Gamma_{t,k}, t = 1, 2$, by

$$\Gamma_{t,k} := \mathrm{E}\big(p^k(\mathbf{X}_t) p^k(\mathbf{X}_t)^\top\big) = \int_{\mathcal{X}} p^k p^{k\top} \mathrm{d}\nu_t,$$

[14] Strictly speaking, ρ is not a proper metric. To see this, note that of two distinct functions h_1 and h_2 in $\mathcal{H}$ differ only on a set A for which $\nu_t(A) = 0, t = 1, 2$, then ρ assigns $\rho(h_1, h_2) = 0$ even though $h_1 \neq h_2$. Hence ρ fails the identity of indiscernibles axiom in the definition of a metric. Since ρ satisties the axioms of positivity and symmetry as well as the triangle inequality, it is, however, a *pseudo–metric* on $\mathcal{H}$. If I define two functions h_1 and h_2 to be *equivalent*, denoted $h_1 \sim h_2$, if they differ only on a set of ν_t–measure zero, $t = 1, 2$, then ρ is a proper metric on the set of *equivalence classes* $\mathcal{H}/\!\sim$ of $\mathcal{H}$. In what follows, I will ignore this distinction and simply refer to ρ as a "metric" on $\mathcal{H}$, and $(\mathcal{H}, \rho)$ as a "metric space."

[15] The metric $\rho := \rho_T$ extends naturally to the case of $T \geqslant 2$ by defining $\rho_T(h_1, h_2) := (1/T) \sum_{t=1}^{T} \|h_1 - h_2\|_{t,2}$, where $\|h_1 - h_2\|_{t,2}$ denotes the $L^2(\nu_t)$ distance, $t = 1, \ldots, T$.

let $\widetilde{Q}$ be the function defined by $\beta \mapsto Q\left(h_\beta\right)$ on $\mathcal{B}_k$, and let the $k \times k$ matrix $\mathbf{A}_k$ denote its second–derivative (when well–defined), $\beta \mapsto \mathbf{A}_k\left(\beta\right) := (\partial^2 \widetilde{Q}/\partial\beta\partial\beta^\top)(\beta)$. The following assumptions are used to establish consistency of the SMRCE with respect to ρ.

A. 5 (**Basis functions**). *The (positive) eigenvalues of* $\Gamma_{t,k}, t = 1, 2,$ *are bounded away from zero and from above uniformly in* $k \in \mathbf{N}$.

A. 6 (**Sieve spaces**). *The sieve spaces* $\mathcal{H}_k = \mathcal{H}_k(\mathcal{B}_k)$ *and their generating sets* $\mathcal{B}_k$ *are such that: 1.* $\mathcal{B}_k$ *is compact in* $\mathbf{R}^k$ *for* $k \in \mathbf{N}$. *2.* $\mathcal{H}_k \subset \mathcal{H}_{k+1} \subset \mathcal{H}$ *for* $k \in \mathbf{N}$. *3. There exists a sequence* $\{h_k\}_{k\geqslant 1}$ *such that* $h_k \in \mathcal{H}_k, k \in \mathbf{N}$, *and* $\|h_k - h_o\|_{\mathcal{X}} \to 0$ *as* $k \to \infty$.

A. 7 (**Objective function**). *1.* $\widetilde{Q}$ *is twice continuously differentiable on* $\mathcal{B}_k$ *for* $k \in \mathbf{N}$. *2. The (negative) eigenvalues of* $\mathbf{A}_k\left(\beta\right)$, *are bounded from below and away from zero uniformly in* $\beta \in \mathcal{B}_k$ *and* $k \in \mathbf{N}$. *3. The maximizer* $\beta_k^* := \operatorname*{argmax}_{\beta\in\mathcal{B}_k} \widetilde{Q}\left(\beta\right)$ *belongs to the interior of* $\mathcal{B}_k$ *for all* $k \in \mathbf{N}$.

Remark 2.3. 1. A.5 imposes some regularity on the regressors $p^k\left(\mathbf{X}_t\right), t = 1, 2$. This assumption provides a link between the distance between functions $h_1 = h_{\beta_1}$ and $h_2 = h_{\beta_2}$ in $\mathcal{H}_k$ and the distance between their coefficient vectors β_1 and β_2. Roughly speaking, this assumption ensures that the properties of $\mathcal{B}_k$ are inherited by $\mathcal{H}_k$. The assumption also plays a role in showing that the estimation problem is well–posed.[16]

2. By the preceding remark, A.6.1 is used to establish compactness of each $\mathcal{H}_k$. Combined with the continuity of $\widetilde{Q}$, the assumption also guarantees that a maximizer on closed subsets of $\mathcal{H}_k$ always exists (via Weierstrass's Extreme Value Theorem). Although not needed to ensure the existence of the SMRCE itself (see Remark 2.2), given the computational aspect of finding the SMRCE in practice, restricting attention to a (large) subset of $\mathbf{R}^k$ when conducting maximization appears natural. A.6.2 states that the sieve spaces are growing but contained in the parameter space. A.6.3

A.8 says that h_o can be approximated *uniformly* by a sequence in the sieve. Note that this is stronger than assuming that h_o can be approximated with respect to the ρ metric. The stronger assumption is needed because the population objective function Q is continuous with respect to the supremum metric on $\mathcal{H}$, but not necessarily with respect to ρ on $\mathcal{H}$.

3. A.7.1 imposes some regularity on the objective $\beta \mapsto \widetilde{Q}(\beta)$, and this regularity is inherited by $h \mapsto Q(h)$. Although the *sample* objective function Q_n is a step function, and

[16] A maximization problem is said to be *well posed* (with respect to a metric ρ), if for all sequences $\{h_k\}_{k\geqslant 1}$ in $\mathcal{H}$ such that $Q(h_o) - Q(h_k) \to 0$ as $k \to \infty$, we have $\rho(h_k, h_o) \to 0$; and *ill posed* if there exists a sequence $\{h_k\}_{k\geqslant 1}$ in $\mathcal{H}$ such that $Q(h_o) - Q(h_k) \to 0$ as $k \to \infty$ but $\rho(h_k, h_o) \nrightarrow 0$. See Chen (2007) for a discussion of ill–posedness in the context of sieve extremum estimation and Carrasco, Florens, and Renault (2007) for a survey of linear inverse problems within structural estimation.

thus discontinuous, for an increasing sample size the averaging acts as a "smoother," and under mild conditions the *population* objective is continuous. A.7.1 makes this conclusion somewhat stronger by requiring that $\widetilde{Q}$ is in fact twice continuously differentiable. A.7.2 implies that $\widetilde{Q}$ is strictly concave on $\mathcal{B}_k$ for all k. The strict concavity guarantees a unique maximizer on $\mathcal{B}_k$, which justifies the wording of A.7.3. The concavity also prevents the objective from "turning up" again when one moves away from the maximizer. For example, $\widetilde{Q}$ cannot drift along an asymptote to the value of the maximum as k increases without bound. That the maximum of $\widetilde{Q}$ on $\mathcal{B}_k$ is well separated plays an important part in showing that the population maximization problem is well–posed.

A.7.3 guarantees that the first–order necessary condition for a maximum is satisfied with equality in all arguments, which simplifies later expansions.

The above assumptions lead us to the second main result of this paper.

Theorem 2.3 (Consistency). *If A.1–A.7 hold, and $k_n/n \to 0$ as $n \to \infty$, then any maximizer $\widehat{h}_n$ of Q_n on $\mathcal{H}_{k_n}$ satisfies $\rho(\widehat{h}_n, h_o) \to_P 0$ as $n \to \infty$.*

2.6 Convergence Rates

To derive the rate of convergence of the series maximum rank correlation estimator (SMRCE), I need to invoke stronger assumptions than needed for consistency alone.

A. 8 (Approximation). *There exists a constant $\alpha > 0$ and a sequence $\{h_k\}_{k \geqslant 1}$ such that: 1. $h_k \in \mathcal{H}_k$ for all $k \in \mathbf{N}$; and 2. $\|h_k - h_o\|_{\mathcal{X}} \lesssim k^{-\alpha}$ as $k \to \infty$.*

For any $\delta \geqslant 0$, let $B_\rho(h, \delta)$ denote the open (ρ) δ–ball in $\mathcal{H}$ centered at h in $\mathcal{H}$, $B_\rho(h, \delta) := \{h' \in \mathcal{H} : \rho(h', h) < \delta\}$.

A. 9 (Locally Lipschitz Objective). *There exists a constant $\delta^* > 0$ such that for all h in $B_\rho\left(h_o, \delta^*\right) : Q(h_o) - Q(h) \lesssim \rho(h, h_o)$.*

Remark 2.4. 1. A.8, which replaces A.6.3 says that h_o can be approximated *uniformly* by a sequence in the sieve, and the uniform approximation is of a particular order. The uniform approximation error is assumed to have (at least) polynomial decay in terms of the series truncation k. The constant α is typically a function of the smoothness of h_o and dimensionality of $\mathbf{X}_t$. (See also Remark 2.5.3).

2. A.9 is a local smoothness condition on the population objective function Q. The assumption states that for all h in $\mathcal{H}$ sufficiently close to h_o, Q behaves as a Lipschitz function. Because A.9 involves h_o, and h_o is nonparametric, this assumption cannot be phrased in terms of the coefficient vectors. However, A.9 may still be be viewed as a strengthening of A.7.

Let ζ_k be defined by

$$\zeta_k := \sup_{\mathbf{x}\in\mathcal{X}} \|p^k(\mathbf{x})\|$$

With these additional conditions and the definition of ζ_k we arrive at the third, and last, main result of this paper.

Theorem 2.4 (**Rates**). *1. If A.1–A.9 hold, and $k_n/n \to 0$ as $n \to \infty$, then for any maximizer $\widehat{h}_n$ of Q_n on $\mathcal{H}_{k_n}$ we have*

$$\rho(\widehat{h}_n, h_o) \lesssim_P (k_n/n)^{1/4} + k_n^{-\alpha}. \tag{2.6.1}$$

2. If, in addition, $\zeta_{k_n}^4 k_n/n \to 0$ as $n \to \infty$, then

$$\|\widehat{h}_n - h_o\|_{\mathcal{X}} \lesssim_P \zeta_{k_n}[(k_n/n)^{1/4} + k_n^{-\alpha}]. \tag{2.6.2}$$

3. If k_n is chosen such that $k_n \asymp n^{1/(4\alpha+1)}$, then

$$\rho(\widehat{h}_n, h_o) \lesssim_P n^{-\alpha/(4\alpha+1)},$$
$$\|\widehat{h}_n - h_o\|_{\mathcal{X}} \lesssim_P \zeta_{k_n} n^{-\alpha/(4\alpha+1)},$$

provided $k_n/n \to 0$ and $\zeta_{k_n}^4 k_n/n \to 0$, respectively, as $n \to \infty$.

Remark 2.5. 1. Parts 1 and 2 of Theorem 2.4 provide upper bounds on the rate of convergence of the SMRCE $\widehat{h}_n$ in terms of the ρ and uniform metrics for general choices of the sequence of series truncation terms $\{k_n\}_{n\geqslant 1}$ (subject to the conditions $k_n/n \to 0$ and $\zeta_{k_n}^4 k_n/n \to 0$, respectively). The *first* part of the right–hand side bound in (2.6.1) may be interpreted as the rate of convergence of the *standard error* of the estimator, where the "error" is measured in relation to the maximizer $h_{k_n}^*$ of the population objective Q, when maximization is restricted to the sieve space $\mathcal{H}_{k_n}$. This maximizer is, in some sense, the "best," or "risk–minimizing" element in $\mathcal{H}_{k_n}$. The *second* part of the bound in (2.6.1) may be interpreted as the rate of convergence of the *bias* measured as the distance between h_o and its best approximation in the sieve space $\mathcal{H}_{k_n}$. The proof of Theorem 2.4 involves deriving convergence rates for these two components in turn.

2. The interpretation of the right–hand side bound in (2.6.2) is similar. However, as the notion of convergence in Part 2 of Theorem 2.4 is *uniform* convergence, one must have some control over the behavior of (the norm of) the basis terms. The term ζ_k captures this behavior and will generally depend on the approximating properties of the basis used to

construct the estimation series. For extensive reviews of the approximating properties of different series, see Huang (1998) and Chen (2007).

3. In Part 3 of Theorem 2.4 I maximize the rate of convergence provided by Parts 1 and 2 of Theorem 2.4 by choosing the series truncation sequence to such that the standard error and bias terms convergence to zero at the same rate. For typical smoothness classes we have $\alpha = p/d_x$, where p denotes the degree of smoothness, often expressed in terms of the number of times h_o is continuously differentiable. (For Hölder classes $p = p' + \gamma$, where p' is the number of continuous derivatives and γ is the Hölder exponent.) For such smoothness classes, the convergence rate with respect to ρ becomes $\rho(\widehat{h}_n, h_o) \lesssim_P n^{-p/(4p+d_x)}$. From this expression, we see that the smoother the h_o, the faster the convergence. The expression also illustrates that, like many other nonparametric estimators, the SMRCE suffers from the curse of dimensionality; the higher the d_x, the slower the rate of convergence. One can likely... employ the additive separability in A.3 to reduce d_x to the maximum number $\pi \vee (d_x - \pi)$ of the arguments of ψ_o and φ_o by considering estimation of each of the two components of h_o separately.

4. Stone (1980; 1982) shows that the optimal (global) rates of convergence for nonparametric regression in the L_2 norm is $\|\widehat{g}_n - g_o\|_2 \lesssim_P n^{-p/(2p+d)}$, where p again denotes the smoothness of g_o and d the number of arguments. To the best of my knowledge, the optimal rate of convergence of nonparametric estimators of h_o in the parameter space $\mathcal{H}$ remains unknown. In fact, it remains to be proven (or disproven) that such an optimal rate exists. However, *if* Stone's result carries over to the generalized (panel) regression, then—by the preceding remark—the rate provided in Part 3 of Theorem 2.4 is suboptimal. I conjecture that one may refine the argument used to derive the bound on the standard error in Part 1 of Theorem 2.4 to achieve the tigher bound of $(k_n/n)^{1/2}$. With this tighter bound, the rate provided by Part 3 of Theorem 2.4 coincides with Stone's optimal rate.

5. Theorem 3.2 of Chen (2007), Theorem 1 of Chen and Shen (1998), Theorem 3.1 of Chen and White (1999), and Theorems 1 and 2 of Shen and Wong (1994) provide sufficient conditions for deriving rates of convergence of (approximate) sieve M–estimators, a class in which the SMRCE belongs. One might suspect that we can simply verify the conditions of said papers to arrive at the rate of convergence. However, the theorems of Chen (2007), Chen and Shen (1998), and Chen and White (1999) all rely on a smoothness requirement of the objective integrand corresponding to a local continuity requirement. The MRC objective function Q_n is an average of indicator functions, which need be even locally smooth in the parameter.

Theorem 1 of Shen and Wong (1994) requires a bound on the L^∞–metric entropy of a certain space of objective integrand differences (see their Condition C3). This assumption may also be viewed as a (high–level) smoothness requirement on the objective integrands. It is not clear how one should construct such an entropy bound in the case of the objective integrands enter the MCR objective. In fact, it is not clear that such a bound even exists in the case of the SMRC estimation problem, where the objective integrands have points of discontinuity.[17] Hence, none of these general rate of convergence results cover the case of SMRC estimation.

2.7 Conclusion

In this paper I show that that it is possible to identify a generalized panel regression models (GPRM) without imposing any parametric structure on 1. the function of observable explanatory variables, 2. the systematic function through which the function of observable explanatory variables, fixed effect, and disturbance term generate the outcome variable, or 3. the distribution of unobservables.

The GPRM nests frameworks such as panel regression, binary threshold crossing, censored regression, duration, and transformation models. The GPRM admits arbitrary dependence between the explanatory variables and the fixed effects, and permits a fixed effect of any finite dimension.

I develop a series maximum rank correlation estimator (SMRCE) of the function of observable explanatory variables, and provide conditions under which L_2–consistency is achieved. I also provide conditions under which both L_2–type and uniform convergence rates of the SMRCE may be derived.

To the best of my knowledge, this paper is the first to state conditions under which identification and consistency is achieved and convergence rates derived with a typical panel (i.e. “large” n–small T”) observational framework, while allowing all three above–mentioned elements of the model to be nonparametric.

[17] In the proof of Theorem 2.4 I derive a bound on the $L^p(\mathcal{Q})$–metric entropy of the space of objective integrands that holds for any $1 \leqslant p < \infty$ and any probability measure $\mathcal{Q}$. [See (2.A.8).] However, attempting to extend the result to the case of $p = \infty$ as required by Shen and Wong (1994), the entropy bound becomes trivial.

Chapter 2 Appendices

2.A Proofs

2.A.1 Proofs for Section 2.3

The proof of Theorem 2.1 relies on the following lemma.

Lemma 2.1. *If A.1 and A.2 hold, then*

$$\mathbb{P}\left(Y_1 > Y_2 \middle| \mathbf{X}, \alpha\right) \left\{ \begin{array}{c} \geqslant \\ \leqslant \end{array} \right\} \mathbb{P}\left(Y_1 < Y_2 \middle| \mathbf{X}, \alpha\right) \text{ whenever } h_o\left(\mathbf{X}_1\right) \left\{ \begin{array}{c} \geqslant \\ \leqslant \end{array} \right\} h_o\left(\mathbf{X}_2\right).$$

Proof. Let $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ be such that $h_o\left(\mathbf{x}_1\right) \leqslant h_o\left(\mathbf{x}_2\right)$ and $\mathbf{a}$ a value in the support of α. Since F is increasing in its first and last argument,

$$F\left(h_o\left(\mathbf{x}_1\right), \mathbf{a}, e\right) \leqslant F\left(h_o\left(\mathbf{x}_2\right), \mathbf{a}, e\right), \tag{2.A.1}$$

for all e in the support of ϵ_t given $(\mathbf{X}_1, \mathbf{X}_2, \alpha) = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{a})$. It follows that

$$\begin{aligned}
&\mathbb{P}\left(Y_1 > Y_2 \middle| \mathbf{X} = \mathbf{x}, \alpha = \mathbf{a}\right) \\
=&\int_{\mathbf{R}^2} \mathbf{1}\left(D \circ F\left(h_o\left(\mathbf{x}_1\right), \mathbf{a}, e_1\right) > D \circ F\left(h_o\left(\mathbf{x}_2\right), \mathbf{a}, e_2\right)\right) \mathrm{d}F_{\epsilon_1, \epsilon_2 | \mathbf{X}, \alpha}\left(e_1, e_2 | \mathbf{x}, \mathbf{a}\right) \\
=&\int_{\mathbf{R}^2} \mathbf{1}\left(D \circ F\left(h_o\left(\mathbf{x}_1\right), \mathbf{a}, e_1\right) > D \circ F\left(h_o\left(\mathbf{x}_2\right), \mathbf{a}, e_2\right)\right) \mathrm{d}F_{\epsilon | \mathbf{X}, \alpha}\left(e_1 | \mathbf{x}, \mathbf{a}\right) \mathrm{d}F_{\epsilon | \mathbf{X}, \alpha}\left(e_2 | \mathbf{x}, \mathbf{a}\right) \\
=&\int_{\mathbf{R}^2} \mathbf{1}\left(D \circ F\left(h_o\left(\mathbf{x}_1\right), \mathbf{a}, e_2\right) > D \circ F\left(h_o\left(\mathbf{x}_2\right), \mathbf{a}, e_1\right)\right) \mathrm{d}F_{\epsilon_1, \epsilon_2 | \mathbf{X}, \alpha}\left(e_2, e_1 | \mathbf{x}, \mathbf{a}\right) \\
\leqslant&\int_{\mathbf{R}^2} \mathbf{1}\left(D \circ F\left(h_o\left(\mathbf{x}_2\right), \mathbf{a}, e_2\right) > D \circ F\left(h_o\left(\mathbf{x}_1\right), \mathbf{a}, e_1\right)\right) \mathrm{d}F_{\epsilon_1, \epsilon_2 | \mathbf{X}, \alpha}\left(e_2, e_1 | \mathbf{x}, \mathbf{a}\right) \\
=&\mathbb{P}\left(Y_1 < Y_2 \middle| \mathbf{X} = \mathbf{x}, \alpha = \mathbf{a}\right),
\end{aligned}$$

where $F_{\epsilon_1, \epsilon_2 | \mathbf{X}, \alpha}$ denotes the joint CDF of (ϵ_1, ϵ_2) conditional on $(\mathbf{X}, \alpha)$, and I have used that ϵ_1 and ϵ_2 are i.i.d. conditional on $(\mathbf{X}_1, \mathbf{X}_2, \alpha)$, (2.A.1), and D increasing. An analogous argument establishes the reverse probability inequality in the event of $h_o(\mathbf{x}_1) \geqslant h_o(\mathbf{x}_2)$. Since $\mathbf{x}_1$ and $\mathbf{x}_2$ were arbitrary, these inequalities establish the lemma. □

Proof of Theorem 2.1. Let $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ be such that $h_o(\mathbf{x}_1) \leqslant h_o(\mathbf{x}_2)$ and $\mathbf{a}$ a value in the support of α. Lemma 2.1 implies that $\mathbb{P}(Y_1 > Y_2|\mathbf{X} = \mathbf{x}, \alpha = a) \leqslant \mathbb{P}(Y_1 < Y_2|\mathbf{X} = \mathbf{x}, \alpha = a)$. Take expectations over α given $\mathbf{X} = \mathbf{x}$ to arrive at the desired conclusion. □

The proof of Theorem 2.2 relies on Lemmas 2.2–2.4.

Lemma 2.2. *If A.3 and A.4 hold, then for any $h \in \mathcal{H}$, $\mathbf{1}(h(\mathbf{X}_1) = h(\mathbf{X}_2)) = 0$ a.s. $[\nu_{\mathbf{X}}]$.*

Proof. Let h be in $\mathcal{H}$. Then $h(\mathbf{x}) = \psi(\mathbf{x}_\pi) + \varphi(\mathbf{x}_{-\pi})$ for some $\psi \in \Psi$ and $\varphi \in \Phi$. Write $\mathbb{P}(h(\mathbf{X}_1) = h(\mathbf{X}_2))$ as the integral

$$\begin{aligned}&\int \mathbf{1}(h(\mathbf{X}_1) = h(\mathbf{X}_2))\,\mathrm{d}\nu_{\mathbf{X}}\\ =&\int \Big[\int \mathbf{1}(\psi(\mathbf{x}_{\pi,1}) + \varphi(\mathbf{x}_{-\pi,1}) = \psi(\mathbf{x}_{\pi,2}) + \varphi(\mathbf{x}_{-\pi,2}))\, f(\mathbf{x}_\pi|\,\mathbf{x}_{-\pi})\,\mathrm{d}\mathbf{x}_\pi\Big]\mathrm{d}\nu_{\mathbf{X}_{-\pi}}(\mathbf{x}_{-\pi}),\end{aligned}$$

where the inner integration is with respect to Lebesgue measure, and $f := f_{\mathbf{X}_\pi|\mathbf{X}_{-\pi}}$ denotes the joint PDF of $\mathbf{X}_\pi$ conditional on $\mathbf{X}_{-\pi}$. By continuity and linear homogeneity of ψ, for any fixed $\mathbf{x}_{-\pi}$ the set $\{\mathbf{x}_\pi : \psi(\mathbf{x}_{\pi,1}) + \varphi(\mathbf{x}_{-\pi,1}) = \psi(\mathbf{x}_{\pi,2}) + \varphi(\mathbf{x}_{-\pi,2})\}$ has zero Lebesgue measure. A.4.4 therefore implies that the inner integral is zero a.s. $[\nu_{\mathbf{X}_{-\pi}}]$. □

Lemma 2.3. *If A.1, A.2, and A.4.2 hold, then $\mathbb{P}(Y_1 > Y_2|\,\mathbf{X}) \gtrless \mathbb{P}(Y_1 < Y_2|\,\mathbf{X})$ whenever $h(\mathbf{X}_1) \gtrless h_o(\mathbf{X}_2)$ a.s. $[\nu_{\mathbf{X}}]$.*

Proof. Let $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ be such that $h_o(\mathbf{x}_1) < h_o(\mathbf{x}_2)$ and $\mathbf{a}$ a value in the support of α. For $y \in \mathbf{R}$ we have

$$\begin{aligned}&\mathbb{P}(Y_1 \leqslant y|\,\mathbf{X} = \mathbf{x}, \alpha = \mathbf{a})\\ &= \int_{\mathbf{R}} \mathbf{1}(D \circ F(h_o(\mathbf{x}_1), \mathbf{a}, e) \leqslant y)\,\mathrm{d}F_{\epsilon_1|\mathbf{X},\alpha}(e|\mathbf{x}, \mathbf{a})\\ &= \int_{\mathbf{R}} \mathbf{1}(D \circ F(h_o(\mathbf{x}_1), \mathbf{a}, e) \leqslant y)\,\mathrm{d}F_{\epsilon|\mathbf{X},\alpha}(e|\mathbf{x}, \mathbf{a})\\ &\geqslant \int_{\mathbf{R}} \mathbf{1}(D \circ F(h_o(\mathbf{x}_2), \mathbf{a}, e) \leqslant y)\,\mathrm{d}F_{\epsilon_2|\mathbf{X},\alpha}(e|\mathbf{x}, \mathbf{a})\\ &= \mathbb{P}(Y_2 \leqslant y|\,\mathbf{X} = \mathbf{x}, \alpha = \mathbf{a}),\end{aligned}$$

where I have used that ϵ_1 and ϵ_2 are i.i.d. conditional on $(\mathbf{X}, \alpha)$ (with marginal CDF $F_{\epsilon|\mathbf{X},\alpha}$), (2.A.1), and D increasing. Since $y \in \mathbf{R}$ was arbitrary, we have

$$\mathbb{P}(Y_1 \leqslant y|\,\mathbf{X} = \mathbf{x}, \alpha = \mathbf{a}) \geqslant \mathbb{P}(Y_2 \leqslant y|\,\mathbf{X} = \mathbf{x}, \alpha = \mathbf{a}),\ y \in \mathbf{R}. \qquad (2.A.2)$$

Now rewrite $\mathbb{P}\left(Y_1 \leqslant Y_2 \middle| \mathbf{X}=\mathbf{x}, \alpha=\mathbf{a}\right)$ as the integral

$$\begin{aligned}
&\int_{\mathbf{R}} \mathbb{P}\left(Y_1 \leqslant D \circ F\left(h_o\left(\mathbf{x}_2\right), \mathbf{a}, e\right) \middle| \mathbf{X}=\mathbf{x}, \alpha=\mathbf{a}\right) \mathrm{d}F_{\epsilon_2|\mathbf{X},\alpha}\left(e|\mathbf{x}, \mathbf{a}\right) \\
&= \int_{\mathbf{R}} \mathbb{P}\left(Y_1 \leqslant D \circ F\left(h_o\left(\mathbf{x}_2\right), \mathbf{a}, e\right) \middle| \mathbf{X}=\mathbf{x}, \alpha=\mathbf{a}\right) \mathrm{d}F_{\epsilon|\mathbf{X},\alpha}\left(e|\mathbf{x}, \mathbf{a}\right) \\
&> \int_{\mathbf{R}} \mathbb{P}\left(Y_2 \leqslant D \circ F\left(h_o\left(\mathbf{x}_2\right), \mathbf{a}, e\right) \middle| \mathbf{X}=\mathbf{x}, \alpha=\mathbf{a}\right) \mathrm{d}F_{\epsilon|\mathbf{X},\alpha}\left(e|\mathbf{x}, \mathbf{a}\right) \\
&\geqslant \int_{\mathbf{R}} \mathbb{P}\left(Y_2 \leqslant D \circ F\left(h_o\left(\mathbf{x}_1\right), \mathbf{a}, e\right) \middle| \mathbf{X}=\mathbf{x}, \alpha=\mathbf{a}\right) \mathrm{d}F_{\epsilon_1|\mathbf{X},\alpha}\left(e|\mathbf{x}, \mathbf{a}\right) \\
&= \mathbb{P}\left(Y_1 \geqslant Y_2 \middle| \mathbf{X}=\mathbf{x}, \alpha=\mathbf{a}\right),
\end{aligned}$$

where the strict inequality comes from A.4.2 and (2.A.2). It follows that

$$\mathbb{P}\left(Y_1 > Y_2 \middle| \mathbf{X}=\mathbf{x}, \alpha=\mathbf{a}\right) < \mathbb{P}\left(Y_1 < Y_2 \middle| \mathbf{X}=\mathbf{x}, \alpha=\mathbf{a}\right).$$

An analogous argument establishes the reverse probability inequality in the event of $h_o\left(\mathbf{x}_1\right) > h_o\left(\mathbf{x}_2\right)$. Compute expectations over α given $\mathbf{X}=\mathbf{x}$ to arrive at the desired conclusion. Since $\mathbf{x}_1$ and $\mathbf{x}_2$ were arbitrary, this establishes the lemma. □

Lemma 2.4. *If A.3 and A.4 hold and $h \in \mathcal{H}$ is such that $h \neq h_o$, then there exists a subset $N_1 \times N_2$ of $\mathcal{X}^2$ such that 1. $\nu_{\mathbf{X}}(N_1 \times N_2) > 0$, and 2. for all $(\mathbf{x}_1, \mathbf{x}_2) \in N_1 \times N_2$: $[h\left(\mathbf{x}_1\right) - h\left(\mathbf{x}_2\right)][h_o\left(\mathbf{x}_1\right) - h_o\left(\mathbf{x}_2\right)] < 0$.*

Proof. Let h be in $\mathcal{H}$, so $h(\mathbf{x}) = \psi(\mathbf{x}_\pi) + \varphi(\mathbf{x}_{-\pi})$ for some $\psi \in \Psi$ and some $\varphi \in \Phi$. If $h \neq h_o$, then there exists $\widehat{\mathbf{x}} =\in \mathcal{X}$ such that $h(\widehat{\mathbf{x}}) \neq h_o(\widehat{\mathbf{x}})$. Assume that $h(\widehat{\mathbf{x}}) < h_o(\widehat{\mathbf{x}})$; the reverse case is analogous. By A.(3), $\psi(\overline{\mathbf{x}}_\pi) = \psi_o(\overline{\mathbf{x}}_\pi)$, $\varphi(\overline{\mathbf{x}}_{-\pi}) = \varphi_o(\overline{\mathbf{x}}_{-\pi})$, and ψ and ψ_o are both linearly homogeneous and increasing it their first argument. It follows that there exists $a \in \mathbf{R}$ such that

$$\begin{aligned}
h\left(\widehat{\mathbf{x}}\right) = \psi\left(\widehat{\mathbf{x}}_\pi\right) + \varphi\left(\widehat{\mathbf{x}}_{-\pi}\right) &< \psi\left(a\overline{\mathbf{x}}_\pi\right) + \varphi\left(\overline{\mathbf{x}}_{-\pi}\right) = a\psi\left(\overline{\mathbf{x}}_\pi\right) + \varphi\left(\overline{\mathbf{x}}_{-\pi}\right) \\
&= a\psi_o\left(\overline{\mathbf{x}}_\pi\right) + \varphi_o\left(\overline{\mathbf{x}}_{-\pi}\right) < \psi_o\left(a\overline{\mathbf{x}}_\pi\right) + \varphi_o\left(\overline{\mathbf{x}}_{-\pi}\right) < h\left(\widehat{\mathbf{x}}\right).
\end{aligned}$$

A.(3) implies that ψ, φ, ψ_o, and φ are continuous. Hence there exists a neighborhood N_1 of $\widehat{\mathbf{x}}$ and a neighborhood N_2 of $(\gamma\overline{\mathbf{x}}_\pi, \overline{\mathbf{x}}_{-\pi})$ such that for all $(\mathbf{x}_1, \mathbf{x}_2) \in N_1 \times N_2$ we have $h\left(\mathbf{x}_1\right) < h\left(\mathbf{x}_2\right)$ and $h_o\left(\mathbf{x}_1\right) > h_o\left(\mathbf{x}_2\right)$. As a consequence, for all $(\mathbf{x}_1, \mathbf{x}_2) \in N_1 \times N_2$: $[h\left(\mathbf{x}_1\right) - h\left(\mathbf{x}_2\right)][h_o\left(\mathbf{x}_1\right) - h_o\left(\mathbf{x}_2\right)] < 0$. Write $\nu_{\mathbf{X}}\left(N_1 \times N_2\right)$ as

$$\int \Big[\int \mathbf{1}_{N_1 \times N_2}\left(\left(\mathbf{x}_{\pi,1}, \mathbf{x}_{-\pi,1}\right), \left(\mathbf{x}_{\pi,2}, \mathbf{x}_{-\pi,2}\right)\right) f_{\mathbf{X}_\pi|\mathbf{X}_{-\pi}}\left(\mathbf{x}_\pi|\mathbf{x}_{-\pi}\right) \mathrm{d}\mathbf{x}_\pi \Big] \mathrm{d}\nu_{\mathbf{X}_{-\pi}}\left(\mathbf{x}_{-\pi}\right).$$

Since $\mathbf{X}_\pi$ is continuously distributed with full support conditional on $\mathbf{X}_{-\pi}$, the inner integral is positive for a.s. $[\nu_{\mathbf{X}_{-\pi}}]$. □

Proof of Theorem 2.2. By Lemma 2.2,

$$\mathbf{1}\left(h\left(\mathbf{x}_1\right) < h\left(\mathbf{x}_2\right)\right) = 1 - \mathbf{1}\left(h\left(\mathbf{x}_1\right) > h\left(\mathbf{x}_2\right)\right) \text{ a.s. } [\nu_{\mathbf{X}}].$$

Consider $(\mathbf{x}_1, \mathbf{x}_2) \in \mathcal{X}^2$ and $h \in \mathcal{H}$. The integrand of the objective function may be (a.s.) written as

$$\begin{aligned} h \mapsto & \left[\mathbb{P}\left(Y_1 > Y_2 \middle| \mathbf{X} = \mathbf{x}\right) - \mathbb{P}\left(Y_1 < Y_2 \middle| \mathbf{X} = \mathbf{x}\right)\right] \mathbf{1}\left(h\left(\mathbf{x}_1\right) > h\left(\mathbf{x}_2\right)\right) \\ & + \mathbb{P}\left(Y_1 < Y_2 \middle| \mathbf{X} = \mathbf{x}\right), \end{aligned} \tag{2.A.3}$$

Theorem 2.1 implies for h_o maximizes this integrand for all $\mathbf{x} \in \mathcal{X}^2$. By the law of iterated expectations we may deduce that h_o maximizes Q on $\mathcal{H}$, establishing existence of a solution. The objective difference is

$$\begin{aligned} Q\left(h_o\right) - Q\left(h\right) = & \mathrm{E}_{\mathbf{X}}\left(\left[\mathbf{1}\left(h\left(\mathbf{X}_1\right) > h\left(\mathbf{X}_2\right)\right) - \mathbf{1}\left(h\left(\mathbf{X}_1\right) > h\left(\mathbf{X}_2\right)\right)\right]\right. \\ & \times \left.\left[\mathbb{P}\left(Y_1 > Y_2 \middle| \mathbf{X}\right) - \mathbb{P}\left(Y_1 < Y_2 \middle| \mathbf{X}\right)\right]\right) \\ \geqslant & \mathrm{E}_{\mathbf{X}}\left(\mathbf{1}_{N_1 \times N_2}(\mathbf{X}_1, \mathbf{X}_2)\left[\mathbf{1}\left(h\left(\mathbf{X}_1\right) > h\left(\mathbf{X}_2\right)\right) - \mathbf{1}\left(h\left(\mathbf{X}_1\right) > h\left(\mathbf{X}_2\right)\right)\right]\right. \\ & \times \left.\left[\mathbb{P}\left(Y_1 > Y_2 \middle| \mathbf{X}\right) - \mathbb{P}\left(Y_1 < Y_2 \middle| \mathbf{X}\right)\right]\right) > 0, \end{aligned}$$

where the first inequality holds because the integrand is positive a.s. $[\nu_{\mathbf{X}}]$, cf. (2.A.3), and the second follows from Lemma 2.3 and 2.4. The strict inequality establishes uniqueness. □

2.A.2 Proofs for Section 2.5

The proof of Theorem 1.4 relies on Lemmas 2.5–2.11.

Lemma 2.5 (Metric equivalence). *If A.5 holds, then for any $k \in \mathbf{N}, h_1 = h_{\beta_1}, h_2 = h_{\beta_2} \in \mathcal{H}_k$, we have $\rho\left(h_1, h_2\right) \asymp \|\beta_1 - \beta_2\|$.*

Proof. Given that the eigenvalues of $\Gamma_{k,t}$ are bounded from above by some constant $0 < \bar{c}_\Gamma < \infty$, we have

$$\begin{aligned} 2\rho\left(h_1, h_2\right) &= \Big(\int_{\mathcal{X}} |p^{k\top}(\beta_1 - \beta_2)|^2 \mathrm{d}\nu_1\Big)^{1/2} + \Big(\int_{\mathcal{X}} |p^{k\top}(\beta_1 - \beta_2)|^2 \mathrm{d}\nu_2\Big)^{1/2} \\ &= [(\beta_1 - \beta_2)^\top \Gamma_{k,1}(\beta_1 - \beta_2)]^{1/2} + [(\beta_1 - \beta_2)^\top \Gamma_{k,2}(\beta_1 - \beta_2)]^{1/2} \\ &\leqslant [\overline{\lambda}(\Gamma_{k,1})^{1/2} + \overline{\lambda}(\Gamma_{k,2})^{1/2}]\|\beta_1 - \beta_2\| \end{aligned} \tag{2.A.4}$$

$$\leqslant 2\overline{c}_{\Gamma}^{1/2}\|\beta_1 - \beta_2\|.$$

Since the eigenvalues of $\Gamma_{k,t}$ are bounded away from zero by some constant $0 < \underline{c}_{\Gamma} < \infty$, a similar calculation yields $\rho\,(h_1, h_2) \geqslant \underline{c}_{\Gamma}^{1/2}\|\beta_1 - \beta_2\|$. □

For a metric space $(\mathcal{F}, \rho_{\mathcal{F}})$ and $\epsilon > 0$, define the *covering number* $N(\epsilon, \mathcal{F}, \rho_{\mathcal{F}})$ *of* $\mathcal{F}$ as the smallest number of ϵ–balls in the metric $\rho_{\mathcal{F}}$ on $\mathcal{F}$ needed to cover the space $\mathcal{F}$. (If no such smallest number exist, $N(\epsilon, \mathcal{F}, \rho_{\mathcal{F}})$ is defined to be $+\infty$.) If $\rho_{\mathcal{F}}$ is induced by a norm $\|\cdot\|_{\mathcal{F}}$ on $\mathcal{F}$, then $N(\epsilon, \mathcal{F}, \|\cdot\|_{\mathcal{F}})$ is understood as $N(\epsilon, \mathcal{F}, \rho_{\mathcal{F}})$. Also, define the *diameter* $\operatorname{diam}(\mathcal{F})$ *of* $\mathcal{F}$ by $\operatorname{diam}(\mathcal{F}) := \sup_{f_1, f_2 \in \mathcal{F}} \rho_{\mathcal{F}}(f_1, f_2)$, i.e., the largest $\rho_{\mathcal{F}}$–distance between two elements f_1 and f_2 of $\mathcal{F}$.

Lemma 2.6 (**Compact sieve space**). *If A.5 and A.6.1 hold, then $\mathcal{H}_k$ is compact (ρ) for all $k \in \mathbf{N}$.*

Proof. Fix $k \in \mathbf{N}$. I show that $\mathcal{H}_k$ is closed and totally bounded (ρ), which is equivalent to compactness. 1. *Closed.* Let $\{h_m\}_{m \geqslant 1} \subset \mathcal{H}_k$ be a sequence converging (ρ) to $h \in \mathcal{H}$. Then there exist a sequence $\{\beta_m\}_{m \geqslant 1} \subset \mathcal{B}_k$ such that $h_m = h_{\beta_m}, m \in \mathbf{N}$. Since $\mathcal{B}_k$ is compact in $\mathbf{R}^k$, $\{\beta_m\}$ has a convergent subsequence $\{\beta_{m_\ell}\}_{\ell \geqslant 1}$. Let $\beta := \lim_{\ell \to \infty} \beta_{m_\ell}$ be its limit. Since $\mathcal{B}_k$ is closed in $\mathbf{R}^k$ (Heine–Borel), β is in $\mathcal{B}_k$ and hence h_β in $\mathcal{H}_k$. By Lemma 2.5, $\rho\,(h_{m_\ell}, h_\beta) \lesssim \|\beta_{m_\ell} - \beta\|$, so $h_{m_\ell} \to h_\beta$ in ρ as $\ell \to \infty$. Given that a sequence can have only one limit with respect to ρ, it follows that $h = h_\beta \in \mathcal{H}_k$, and $\mathcal{H}_k$ is closed (ρ). 2. *Totally bounded.* Given that $\mathcal{B}_k$ is compact in $\mathbf{R}^k$, it is bounded (Heine–Borel), so $\operatorname{diam}(\mathcal{B}_k) < \infty, k \in \mathbf{N}$. Hence, for all $\epsilon > 0$ we have $N(\epsilon, \mathcal{B}_k, \|\cdot\|) \leqslant \operatorname{diam}(\mathcal{B}_k)/\epsilon^k$. From (2.A.4) it follows that

$$N(\epsilon, \mathcal{H}_k, \rho) \leqslant \frac{\overline{c}_{\Gamma}^{k/2}\operatorname{diam}(\mathcal{B}_k)}{\epsilon^k} < \infty \text{ for any } \epsilon > 0,$$

so $\mathcal{H}_k$ is totally bounded. □

Lemma 2.7 (ρ**–continuity on** $\mathcal{H}_k$). *If A.5, A.6.1, and A.7.1 hold, then Q is continuous with respect to ρ on $\mathcal{H}_k$ for each $k \in \mathbf{N}$.*

Proof. Let $k \in \mathbf{N}$, and let $\{h_m\}_{m \geqslant 1} \subset \mathcal{H}_k$ be such that $h_m \to h$ in ρ. By Lemma 2.6, $\mathcal{H}_k$ is closed, so $h \in \mathcal{H}_k$, and there is a $\beta \in \mathcal{B}_k$ such that $h = h_\beta$. By Lemma 2.5 the sequence $\{\beta_m\}_{m \geqslant 1} \subset \mathcal{B}_k$ satisfying $h_m = h_{\beta_m}, m \in \mathbf{N}$, converges (ρ) and has $\beta \in \mathcal{B}_k$ as its limit. By the continuity of $\widetilde{Q}$ on $\mathcal{B}_k$ in A.7.1 we have

$$\lim_{m \to \infty} Q(h_m) = \lim_{m \to \infty} \widetilde{Q}(\beta_m) = \widetilde{Q}(\beta) = Q(h).$$

□

Lemma 2.8 **($\|\cdot\|_{\mathcal{X}}$–continuity on $\mathcal{H}$).** *If A.3 and A.4 hold, then Q is continuous with respect to the uniform metric on $\mathcal{H}$.*

Proof. Let $h \in \mathcal{H}$ and $\{h_m\}_{m\geqslant 1} \subset \mathcal{H}$ be such that $\|h_m - h\|_{\mathcal{X}} \to 0$ as $m \to \infty$. Let $\mathbf{x}$ be a point in the support of $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ such that $h(\mathbf{x}_1) > h(\mathbf{x}_2)$. Since $h_m \to h$ uniformly on $\mathcal{X}$, $h_m \to h$ pointwise, and it must be that $h_m(\mathbf{x}_1) > h_m(\mathbf{x}_2)$ for all m sufficiently large. From this observation and the definition of $f(\mathbf{z}, h)$ as

$$f(\mathbf{z}, h) = \mathbf{1}_{\{y_1>y_2\}}\mathbf{1}_{\{h(\mathbf{x}_1)>h(\mathbf{x}_2)\}} + \mathbf{1}_{\{y_1<y_2\}}\mathbf{1}_{\{h(\mathbf{x}_1)<h(\mathbf{x}_2)\}},$$

it follows that

$$\{\mathbf{z} : f(\mathbf{z}, h) \text{ is discontinuous at } h\} \subset \{\mathbf{z} : h(\mathbf{x}_1) = h(\mathbf{x}_2)\}.$$

From the inclusion in the preceding display and Lemma 2.2 we get

$$\mathbb{P}\left(f(\mathbf{Z}, h) \text{ is discontinuous at } h\right) \leqslant \mathrm{P}\left(h(\mathbf{X}_1) = h(\mathbf{X}_2)\right) = 0.$$

Hence, $h \mapsto f(\mathbf{Z}, h)$ is continuous with respect to $\|\cdot\|_{\mathcal{X}}$ a.s. Since the constant function equal to one constitutes a $\mathbb{P}$–integrable majorant for each of the maps $\mathbf{z} \mapsto f(\mathbf{z}, h_m), m \in \mathbf{N}$, Lebesgue's Dominated Convergence Theorem (DCT) implies

$$\lim_{m\to\infty} Q\left(h_m\right) = \lim_{m\to\infty} \int f(\cdot, h_m)\mathrm{d}P = \int \lim_{m\to\infty} f(\cdot, h_m)\mathrm{d}P = \int f(\cdot, h)\mathrm{d}P = Q(h).$$

Hence, $Q(h_m) \to Q(h)$ whenever $h_m \to h$ in $\|\cdot\|_{\mathcal{X}}$. □

Lemma 2.9 **(Well–posedness).** *If A.1—A.7 hold, then for all $\epsilon > 0, k \in \mathbf{N}$:*

$$Q\left(h_o\right) - \sup_{\{h\in\mathcal{H}_k:\rho(h,h_o)\geqslant\epsilon\}} Q\left(h\right) \gtrsim \epsilon^2.$$

Proof. Let $h_k \equiv p^{k\top}\beta_k, \beta_k \in \mathcal{B}_k, k \geqslant 1$, denote the sequence from A.6.3, which by assumption converges uniformly to h_o on $\mathcal{X}$. Lemma 2.8 implies that $Q(h_o) - Q(h_k) \to 0$. Twice continuous differentiability (A.7.1) and uniqueness of the (interior) maximizer (A.7.3) in combination with a second–order mean value expansion of $\widetilde{Q}$ around β_k^* imply

$$\begin{aligned} Q(h_k) - Q(h_k^*) = \widetilde{Q}(\beta_k) - \widetilde{Q}(\beta_k^*) &= (1/2)(\beta_k - \beta_k^*)^\top \mathbf{A}_k(\overline{\beta}_k)(\beta_k - \beta_k^*) \\ &\geqslant (-\overline{c}_A/2)\|\beta_k - \beta_k^*\|^2. \end{aligned}$$

Here $\overline{\beta}_k$ is between β_k and β_k^*, and the inequality follows from A.7.2 using $-\infty < -\overline{c}_A < 0$ as the lower bound on the (negative) eigenvalues. Rearranging the inequality in the preceding display and invoking Theorem 2.2 (identification) we get

$$\|\beta_k^* - \beta_k\|^2 \leqslant (2/\overline{c}_A)[Q(h_k^*) - Q(h_k)] \leqslant (2/\overline{c}_A)[Q(h_o) - Q(h_k)].$$

The left–hand side is equivalent to $\rho(h_k^*, h_k)^2$ (Lemma 2.5) and the right–hand side goes to zero. Deduce that $\rho(h_k^*, h_k) \to 0$. Given that $h_k \to h_o$ uniformly on $\mathcal{X}$ (A.6.3 again), $h_k \to h_o$ in ρ, so by the triangle inequality

$$\rho(h_k^*, h_o) \leqslant \rho(h_k^*, h_k) + \rho(h_k, h_o) \to 0,$$

and $h_k^* \to h_o$ in ρ.

Let $\epsilon > 0, k \in \mathbf{N}$ and $h \in \mathcal{H}_k \backslash B_\rho(h_o, \epsilon)$ be arbitrary, where $B_\rho(h_o, \epsilon) := \{h \in \mathcal{H} : \rho(h, h_o) < \epsilon\}$. The preceding paragraph shows that the sequence $\{h_j^*\}_{j \geqslant 1}$ will eventually belong to $B_\rho(h_o, \epsilon)$. We may therefore choose $k' := k'(\epsilon, h, k) \geqslant k$ large enough such that $h_{k'}^* \in B_\rho(h_o, \epsilon/2)$ for all $j \geqslant k'$. By A.6.2, $\mathcal{H}_k \subset \mathcal{H}_{k+1}, k \in \mathbf{N}$, and $[\mathcal{H}_k \backslash B_\rho(\epsilon, h_o)] \subset \mathcal{H}_{k'}$. The triangle inequality yields

$$\rho\left(h, h_{k'}^*\right) \geqslant \rho\left(h, h_o\right) - \rho\left(h_{k'}^*, h_o\right) \geqslant \epsilon/2.$$

Using that the (negative) eigenvalues of the Hessian are away from zero by some constant $-\infty < -\underline{c}_A < 0$ (A.7), a mean–value expansion yields

$$Q(h) - Q(h_{k'}^*) \leqslant (-\underline{c}_A/2)\|\beta - \beta_{k'}^*\|^2,$$

or, rearranging,

$$Q(h_{k'}^*) - Q(h) \geqslant (\underline{c}_A/2)\|\beta - \beta_{k'}^*\|^2. \tag{2.A.5}$$

From Lemma 2.5 we know that

$$\rho(h, h_{k'}^*) \leqslant \overline{c}_\Gamma \|\beta - \beta_{k'}^*\|^2. \tag{2.A.6}$$

Combining (2.A.5), (2.A.6) and Theorem 2.2, we arrive at

$$Q\left(h_o\right) - Q\left(h\right) \geqslant Q\left(h_{k'}^*\right) - Q\left(h\right) \geqslant (\underline{c}_A/2\overline{c}_\Gamma)\rho\left(h, h_{k'}^*\right)^2 \geqslant (\underline{c}_A/8\overline{c}_\Gamma)\epsilon^2.$$

Given that $h \in \mathcal{H}_k \backslash B_\rho(h_o, \epsilon)$ was arbitrary, $Q\left(h_o\right) - \sup_{h \in \mathcal{H}_k \backslash B_\rho(h_o,\epsilon)} Q\left(h\right) \geqslant (\underline{c}_A / 8\overline{c}_\Gamma)\epsilon^2$. □

To avoid the double subscript, let $f(\cdot, \beta)$ denote f_{h_β}, where for any $\mathbf{z} \in \mathcal{S}$ and any $\gamma \in \mathbf{R}^k$ we have

$$
\begin{aligned}
f\left(\mathbf{z}, \beta\right) = & \mathbf{1}\left(y_1 > y_2\right) \mathbf{1}(p^k\left(\mathbf{x}_1\right)^\top \beta > p^k\left(\mathbf{x}_2\right)^\top \beta) \\
& + \mathbf{1}\left(y_1 < y_2\right) \mathbf{1}(p^k\left(\mathbf{x}_1\right)^\top \beta < p^k\left(\mathbf{x}_2\right)^\top \beta).
\end{aligned} \tag{2.A.7}
$$

Let $\widetilde{\mathcal{F}}_k := \{f(\cdot, \gamma) : \beta \in \mathbf{R}^k\}$, and note that $\mathcal{F}_k$ is a subset of $\widetilde{\mathcal{F}}_k$ because $\mathcal{B}_k \subseteq \mathbf{R}^k$.

Lemma 2.10 (VC Properties). *For each $k \in \mathbf{N}$, $\widetilde{\mathcal{F}}_k$ is VC–subgraph with VC index $V(\widetilde{\mathcal{F}}_k) \lesssim k$.*

Proof. Let s, γ_0, γ_1, and γ_2 be real numbers, and let δ_1 and δ_2 be vectors in $\mathbf{R}^k$. For each $\mathbf{z} \in \mathcal{S}$, define

$$
g\left(\mathbf{z}, s; \gamma_0, \gamma_1, \gamma_2, \delta_1, \delta_2\right) := \gamma_0 s + \gamma_1 y_1 + \gamma_2 y_2 + p^k\left(\mathbf{x}_1\right)^\top \delta_1 + p^k\left(\mathbf{x}_2\right)^\top \delta_2,
$$

and define the function space

$$
\mathcal{G}_k := \{g\left(\cdot, \cdot; \gamma_0, \gamma_1, \gamma_2, \delta_1, \delta_2\right) : \gamma_0, \gamma_1, \gamma_2 \in \mathbf{R}, \delta_1, \delta_2 \in \mathbf{R}^k\}.
$$

Then $\mathcal{G}_k$ forms a vector space over $\mathbf{R}$ on $\mathcal{S} \times \mathbf{R}$ of dimension $2k + 3$. Lemma 2.6.15 in van der Vaart and Wellner (1996) therefore implies that $\mathcal{G}_k$ is VC–subgraph with VC–index $V(\mathcal{G}_k) \leqslant \dim\left(\mathcal{G}_k\right) + 2 = 2k + 5$. In particular, $V(\mathcal{G}_k) \lesssim k$.

Let $f_\beta \in \widetilde{\mathcal{F}}_k$. The *subgraph* of $f\left(\cdot, \beta\right)$ is defined as

$$
\operatorname{subgraph}\left(f\left(\cdot, \beta\right)\right) := \left\{(\mathbf{z}, s) \in \mathcal{S} \times \mathbf{R} : s < f\left(\mathbf{z}, \beta\right)\right\}.
$$

Using (2.A.7) the subgraph of any $f \in \widetilde{\mathcal{F}}_k$ may be written as

$$
\begin{aligned}
& \big(\{y_1 - y_2 > 0\} \cap \{p^k\left(\mathbf{x}_1\right)^\top \beta - p^k\left(\mathbf{x}_2\right)^\top \beta > 0\} \cap \{s \geqslant 1\}^c \cap \{s > 0\}\big) \\
& \cup \big(\{y_2 - y_1 > 0\} \cap \{p^k\left(\mathbf{x}_2\right)^\top \beta - p^k\left(\mathbf{x}_1\right)^\top \beta > 0\} \cap \{s \geqslant 1\}^c \cap \{s > 0\}\big) \\
=: & \left(\{g_1 > 0\} \cap \{g_2 > 0\} \cap \{g_3 \geqslant 1\}^c \cap \{g_4 > 0\}\right) \\
& \cup \left(\{g_5 > 0\} \cap \{g_6 > 0\} \cap \{g_7 \geqslant 1\}^c \cap \{g_8 > 0\}\right).
\end{aligned}
$$

The first set in the above union is the intersection of four sets, three of which belongs to a VC class and the fourth being the complement of a set belonging to a VC class. Lemma

2.6.17 of van der Vaart and Wellner (1996) implies that the class of sets formed by such intersections is also a VC class. By the same argument, the second set in the above union belongs to a VC class. Another application of Lemma 2.6.17 of van der Vaart and Wellner (1996) yields that the subgraphs of $\widetilde{\mathcal{F}}_k$ form a VC class, i.e. $\widetilde{\mathcal{F}}_k$ is VC–subgraph. Lemma 2.5 yields that the order of the VC–index is preserved by taking finite unions and intersections of VC classes, so $V(\widetilde{\mathcal{F}}_k) \lesssim k$. □

Lemma 2.11 (Uniform Law of Large Numbers). *If $k_n/n \to 0$ as $n \to \infty$ then*

$$\sup_{f \in \mathcal{F}_{k_n}} \left|\mathbb{E}_n\left(f\right) - \mathrm{E}\left(f\right)\right| \to_P 0 \textit{ as } n \to \infty.$$

Proof. Consider $\widetilde{\mathcal{F}}_k$ from Lemma 2.10. Since $\widetilde{\mathcal{F}}_k$ is VC–subgraph, Theorem 2.6.4 of van der Vaart and Wellner (1996) implies that for any probability measure $\mathcal{Q}$ and $0 < \epsilon < 1$, the $L^p(\mathcal{Q})$–covering number of $\widetilde{\mathcal{F}}_k$ satisfies

$$N(\epsilon, \widetilde{\mathcal{F}}_k, L^p\left(\mathcal{Q}\right)) \lesssim V(\mathcal{F}_k)(4e)^{V(\mathcal{F}_k)}(1/\epsilon)^{p[V(\mathcal{F}_k)-1]},\ 1 \leqslant p < \infty. \tag{2.A.8}$$

Define the $L^p(\mathcal{Q})$–entropy of $\widetilde{\mathcal{F}}_k$ as $H(\epsilon, \widetilde{\mathcal{F}}_k, L^p(\mathcal{Q})) := \ln N(\epsilon, \widetilde{\mathcal{F}}_k, L^p\left(\mathcal{Q}\right))$. Using that $V(\widetilde{\mathcal{F}}_k) \lesssim k$ (Lemma 2.10), we get that

$$H(\epsilon, \widetilde{\mathcal{F}}_k, L^p\left(\mathcal{Q}\right)) \lesssim k \ln(1/\epsilon).$$

As the bound holds for every $\mathcal{Q}$, it holds for the empirical measure P_n, so

$$H(\epsilon, \widetilde{\mathcal{F}}_k, L^p\left(P_n\right)) \lesssim k \ln(1/\epsilon).$$

For every $k \in \mathbf{N}$, $\mathcal{F}_k = \mathcal{F}_k(\mathcal{B}_k)$ is contained in $\widetilde{\mathcal{F}}_k = \mathcal{F}_k(\mathbf{R}^k)$, so this $L^p(P_n)$–entropy bound applies to $\mathcal{F}_k$ as well. Let $p = 2$. For each fixed $0 < \epsilon < 1$, $\ln(1/\epsilon)$ may be absorbed into the constant and we see that

$$(1/n)H(\epsilon, \mathcal{F}_{k_n}, L^2\left(P_n\right)) \lesssim k_n/n \to 0 \text{ as } n \to \infty.$$

Hence $(1/n)H(\epsilon, \mathcal{F}_{k_n}, L^2\left(\mathbb{P}_n\right)) \to 0$ in probability. Given that the constant function equal to one constitutes a bounded envelope for each $\mathcal{F}_k, k \in \mathbf{N}$, Lemma 3.6 of van de Geer (2000) yields the desired conclusion. □

Proof of Theorem 1.4. Given the preceding lemmas, the proof of consistency given here is now implicit in the proof of Theorem 3.1 in Chen (2007). For the sake of completeness I will provide the remaining steps. By Remark 2.2, a maximizer $\widehat{h}_n$ exists. Let $B_\rho\left(h_o, \epsilon\right) :=$

$\{h \in \mathcal{H} : \rho(h, h_o) < \epsilon\}$ denote the open ϵ–ball centered at h_o relative to the metric ρ. Since $\mathcal{H}_{k_n} \cap B(h_o, \epsilon)^c$ is a closed subset of the compact $\mathcal{H}_{k_n}$, it is compact. Since Q is a continuous function defined on the compact $\mathcal{H}_{k_n}$ (Lemma 2.7), Weierstrass's Extreme Value Theorem implies that $\sup_{h \in \mathcal{H}_{k_n} \backslash B(h_o,\epsilon)} Q(h)$ exists.

Let $\{h_k\}_{k \geqslant 1}$ denote the sequence from A.6.3. It follows that

$$\begin{aligned}
&\mathbb{P}\left(\rho(\widehat{h}_n, h_o) \geqslant \epsilon\right) = \mathbb{P}\left(\widehat{h}_n \in \mathcal{H}_{k_n} \backslash B(h_o, \epsilon)\right) \qquad (2.A.9)\\
&\leqslant \mathbb{P}\left(Q(h_o) - Q(\widehat{h}_n) \gtrsim \epsilon^2\right)\\
&= \mathbb{P}\left(Q(h_o) - Q_n(\widehat{h}_n) + Q_n(\widehat{h}_n) - Q(\widehat{h}_n) \gtrsim \epsilon^2\right)\\
&\leqslant \mathbb{P}\left(Q(h_o) - Q_n(h_{k_n}) + Q_n(\widehat{h}_n) - Q(\widehat{h}_n) \gtrsim \epsilon^2\right)\\
&= \mathbb{P}\left(Q(h_o) - Q(h_{k_n}) + Q(h_{k_n}) - Q_n(h_{k_n}) + Q_n(\widehat{h}_n) - Q(\widehat{h}_n) \gtrsim \epsilon^2\right)\\
&\leqslant \mathbb{P}\Big(|Q(h_o) - Q(h_{k_n})| + 2 \sup_{h \in \mathcal{H}_{k_n}} |Q_n(h) - Q(h)| \gtrsim \epsilon^2\Big),
\end{aligned}$$

where the first inequality follows from Lemma 2.9 and the second by the fact that $\widehat{h}_n$ maximizes Q_n on $\mathcal{H}_{k_n}$. By Lemma 2.11, $\{\mathcal{F}_{k_n}\}_{n \geqslant 1}$ satisfies the Uniform Law of Large Numbers, so

$$\sup_{h \in \mathcal{H}_{k_n}} |Q_n(h) - Q(h)| = \sup_{f \in \mathcal{F}_{k_n}} |\mathbb{E}_n(f) - \mathrm{E}(f)| \to_P 0 \text{ as } n \to \infty.$$

The definition of h_{k_n} and Lemma 2.8 imply that $|Q(h_o) - Q(h_{k_n})| \to 0$ as $n \to \infty$. Hence, the right–hand side probability in (2.A.9) goes to zero as $n \to \infty$. □

2.A.3 Proofs for Section 2.6

The derivation of the convergence rates relies on two additional lemmas. Define $\widehat{f}_n := f_{\widehat{h}_n}$, which is well defined given that the SMRCE is well defined (see also Remark 2.2). Since $\widehat{h}_n$ maximizes $Q_n(h) = \mathbb{E}_n(f_h)$ on $\mathcal{H}_k$ and $\mathcal{F}_k = \mathcal{F}_k(\mathcal{H}_k)$, we must have $\widehat{f}_n \in \underset{f \in \mathcal{F}_k}{\operatorname{argmax}} \mathbb{E}_n(f)$. For notational convenience, I will suppress the dependence on n in k throughout this section. For any $\delta > 0$, define the *δ–restriction of* $\mathcal{F}_k$ as

$$\mathcal{F}_k(\delta) := \Big\{f \in \mathcal{F}_k : \sup_{f' \in \mathcal{F}_k} \mathrm{E}(f') - \mathrm{E}(f) \leqslant \delta\Big\}.$$

Define the *excess risk* as the difference $\widehat{\delta}_n := \sup_{f' \in \mathcal{F}_k} \mathrm{E}(f') - \mathrm{E}(\widehat{f}_n)$, where $\mathrm{E}(\widehat{f}_n)$ should be read as $\mathrm{E}_{\mathbf{Z}}(\widehat{f}_n(\mathbf{Z})) = \int_{\mathcal{S}} \widehat{f}_n(\mathbf{z}) \mathrm{d}P(\mathbf{z})$. The rate of convergence of the excess risk plays a key role in deriving the rate of convergence of the SMRCE.

Lemma 2.12. *If A.1–A.9 hold, and $k_n/n \to 0$, then the excess risk satisfies $\widehat{\delta}_n \lesssim_P \sqrt{k_n/n}$ as $n \to \infty$.*

Proof. By Lemma 2.6, $\mathcal{H}_k$ is compact with respect to ρ, and by Lemma 2.7, Q is continuous on $\mathcal{H}_k$. Hence, by Weierstrass's Extreme Value Theorem, Q has a maximizer on $\mathcal{H}_k$. As a consequence the map $f \mapsto \mathrm{E}\,(f)$ has a maximizer on $\mathcal{F}_k$. Denote such a maximizer by f_k^*. As the maximum is attained, the *excess risk* may be expressed as $\widehat{\delta}_n = \mathrm{E}_Z(f_k^* - \widehat{f}_n)$. Note that $\mathbb{E}_n(\widehat{f}_n - f_k^*) \geqslant 0$ as f_k^* need not solve the sample problem. It follows that

$$\begin{aligned}
\widehat{\delta}_n &= \mathrm{E}_Z(f_k^* - \widehat{f}_n) \\
&= \mathrm{E}_Z(f_k^* - \widehat{f}_n) + \mathbb{E}_n(f_k^* - \widehat{f}_n) - \mathbb{E}_n(f_k^* - \widehat{f}_n) \\
&\leqslant \left(\mathrm{E}_Z - \mathbb{E}_n\right)(f_k^* - \widehat{f}_n) \\
&= |\left(\mathbb{E}_n - \mathrm{E}_Z\right)(\widehat{f}_n - f_k^*)|.
\end{aligned}$$

Since $\widehat{f}_n \in \mathcal{F}_K(\widehat{\delta}_n) \subset \mathcal{F}_k$, we arrive at the (crude) bound on the excess risk:

$$\widehat{\delta}_n \leqslant \sup_{f \in \mathcal{F}_K} \left|\left(\mathbb{E}_n - \mathrm{E}\right)\left(f - f_k^*\right)\right|. \tag{2.A.10}$$

From Lemma 2.10 we know that $\mathcal{F}_k$ is VC–subgraph with VC index $V(\mathcal{F}_k) \lesssim k$, and from (2.A.8) it follows that

$$\sup_{\mathcal{Q}} N\left(\epsilon, \mathcal{F}_k, L_2\left(\mathcal{Q}\right)\right) \lesssim \left(\frac{\left[V\left(\mathcal{F}_k\right)\left(4e\right)^{V(\mathcal{F}_k)}\right]^{1/2[V(\mathcal{F}_k)-1]}}{\epsilon}\right)^{2[V(\mathcal{F}_k)-1]}, \tag{2.A.11}$$

where the supremum is over all probability measures $\mathcal{Q}$. In the language of Chernozhukov, Chetverikov, and Kato (2014a), $\mathcal{F}_k$ is *VC*(b_k, a_k, v_k)*–type* for the choices $b_k := 1, a_k := [V\left(\mathcal{F}_k\right)(4e)^{V(\mathcal{F}_k)}]^{-2[V(\mathcal{F}_k)-1]}$, and $v_k := 2\left[V\left(\mathcal{F}_k\right) - 1\right]$. In what follows I set up for an application of the (Talagrand–type) inequality taken from Chernozhukov, Chetverikov, and Kato (2014a) and, for convenience, restated in Theorem 2.6.

The function class $\overline{\mathcal{F}}_k := \left\{f - f_k^* : f \in \mathcal{F}_k\right\}$ is nothing more than $\mathcal{F}_k$ recentered as f_k^*, so the bound on the uniform covering number in (2.A.11) applies to $\overline{\mathcal{F}}_k$ as well. As the elements of $\overline{\mathcal{F}}_k$ are differences of indicator functions, $b_k = 1$ constitutes as constant envelope for $\overline{\mathcal{F}}_k$. Set $\sigma_k^2 := 1$. Since $\mathrm{E}\left(f - f_k^*\right)^2 \leqslant 1$ for all $f \in \mathcal{F}_k$, we have $\sup_{f \in \overline{\mathcal{F}}_k} \mathrm{var}\,(f) \leqslant \sigma_k^2 = b_k^2$. Given that $b_k = \sigma_k$, the requirement that $b_k^2 v_k \ln\left(a_k b_k/\sigma_k\right) \leqslant n\sigma_k^2$ for application of the CCK inequality is satisfied provided $v_k \ln(a_k) \leqslant n$. Since $a_k = [V(\mathcal{F}_k)(4e)^{V(\mathcal{F}_k)}]^{1/v_k}$, we have

$$v_k \ln(a_k) = \ln V(\mathcal{F}_k) + V(\mathcal{F}_k)\ln(4e) \lesssim k,$$

and the requirement of the CCK inequality boils down to $k \lesssim n$, which holds because $k/n \to 0$. The inequality allows one to pick $t_k \leqslant n\sigma_k^2/b_k^2$. Here I choose $t_n = \ln(n)$, which increases without bound when $n \to \infty$. Guess that $t_n \leqslant v_k \ln(a_k b_k/\sigma_k)$, which may be verified later. Then the CCK inequality in Theorem 2.6 implies that

$$\mathbb{P}\Big(\sup_{f\in\overline{\mathcal{F}}_k} |(\mathbb{E}_n - \mathrm{E})(f)| \gtrsim U_n(\widehat{\delta}_n)\Big) \leqslant e^{-t_n} = \frac{1}{n}, \tag{2.A.12}$$

where $U_n(\widehat{\delta}_n) = \sqrt{\sigma_k^2[t_n \vee v_k \ln(a_k b_k/\sigma_k)]/n} = \sqrt{k/n}$. From (2.A.12), it follows that

$$\limsup_{n\to\infty} \mathbb{P}\Big(\sup_{f\in\overline{\mathcal{F}}_k} |(\mathbb{E}_n - \mathrm{E})(f)| \gtrsim U_n(\widehat{\delta}_n)\Big) = 0,$$

which, in turn, implies that

$$\sup_{f\in\overline{\mathcal{F}}_k} |(\mathbb{E}_n - \mathrm{E})(f)| \lesssim_P U_n(\widehat{\delta}_n) = \sqrt{k/n}.$$

Combining (2.A.10) with the preceding display yields the desired result. □

Let $h_k = h_{\beta_k}, k \in \mathbf{N}$, be a sequence of functions in the sieve satisfying A.8, and let $h_k^* = h_{\beta_k^*}, k \in \mathbf{N}$, where β_k^* is the maximizer of $\widetilde{Q}$ on $\mathcal{B}_k$ from A.7.3.

Lemma 2.13. *If A.1–A.9 hold, then $\rho(h_{k_n}^*, h_{k_n}) \lesssim k_n^{-\alpha}$ as $n \to \infty$.*

Proof. The proof involves two steps. First, I establish a crude bound on the rate of convergence of $\rho(h_k^*, h_k)$, where the dependence on n has been suppressed. Second, I iterate on a set of inequalities to use the crude bound to speed up the rate of convergence to the desired rate.

Step 1. A.5 and A.7 imply

$$\begin{aligned}\rho(h_k^*, h_k)^2 &\lesssim \|\beta_k^* - \beta_k\|_e^2 \lesssim \widetilde{Q}(\beta_k^*) - \widetilde{Q}(\beta_k) \\ &= Q(h_k^*) - Q(h_k) \leqslant Q(h_o) - Q(h_k).\end{aligned} \tag{2.A.13}$$

By A.8, $h_k \to h_o$, so for sufficiently large k, A.9 applies, and $Q(h_o) - Q(h_k) \lesssim \rho(h_k, h_o)$. Assume without loss of generality that this bound holds for $k \in \mathbf{N}$. Now A.8 and (2.A.13) implies that

$$
\begin{aligned}
Q\left(h_o\right)-Q\left(h_k\right) &\lesssim k^{-\alpha}. \qquad (2.A.14)\\
\rho(h_k^{*}, h_k) &\lesssim k^{-\alpha/2}, \qquad (2.A.15)
\end{aligned}
$$

Step 2. Decompose as follows

$$
Q\left(h_o\right)-Q\left(h_k\right)=\left[Q\left(h_o\right)-Q\left(h_{k^2}\right)\right]+\left[Q\left(h_{k^2}\right)-Q\left(h_k\right)\right], \tag{2.A.16}
$$

where $\{k^2\}$ denotes the subsequence $\{k_n^2\}_{n\in\mathbf{N}}$ of $\{k_n\}_{n\in\mathbf{N}}$. I now bound the second term in (2.A.16).

Write $\mathbf{0}$ for the (k^2-k)–dimensional zero vector. Then

$$
\begin{aligned}
Q\left(h_{k^2}\right)-Q\left(h_k\right) \leqslant Q\left(h_{k^2}^{*}\right)-Q\left(h_k\right) &= \widetilde{Q}\left(\beta_{k^2}^{*}\right)-\widetilde{Q}\left(\beta_k\right) \qquad (2.A.17)\\
&\lesssim \|\beta_{k^2}^{*}-(\beta_k^{\top}, \mathbf{0}^{\top})^{\top}\|_e^2 \lesssim \rho(h_{k^2}^{*}, h_k)^2.
\end{aligned}
$$

Eqs. (2.A.14)–(2.A.15) yield $Q\left(h_0\right)-Q\left(h_{k^2}\right) \lesssim k^{-2\alpha}$ and $\rho(h_{k^2}^{*}, h_{k^2}) \lesssim k^{-\alpha}$. Hence, by the triangle inequality and A.8,

$$
\rho(h_{k^2}^{*}, h_o) \leqslant \rho(h_{k^2}^{*}, h_{k^2})+\rho(h_{k^2}, h_o) \lesssim k^{-\alpha}+k^{-2\alpha} \lesssim k^{-\alpha}.
$$

Using this result in combination with the triangle inequality and A.8, we get

$$
\rho(h_{k^2}^{*}, h_k) \leqslant \rho(h_{k^2}^{*}, h_o)+\rho(h_k, h_o) \lesssim k^{-\alpha}+k^{-\alpha} \lesssim k^{-\alpha}.
$$

Plugging this bound into (2.A.17) yields

$$
Q\left(h_{k^2}\right)-Q\left(h_k\right) \lesssim (k^2)^{-\alpha} \lesssim k^{-2\alpha}
$$

Another application of A.8 implies that $Q\left(h_o\right)-Q\left(h_{k^2}\right) \lesssim k^{-2\alpha}$. Gathering these two results in the decomposition (2.A.16) produces

$$
Q\left(h_o\right)-Q\left(h_k\right) \lesssim k^{-2\alpha}+k^{-2\alpha} \lesssim k^{-2\alpha},
$$

and the bound in (2.A.13) from Step 1 yields the desired result. □

Proof of Theorem 2.4. A.5, A.7, and Lemma (2.12) imply that, with probability approaching one as $n \to \infty$,

$$\begin{aligned}\rho(\widehat{h}_n, h_k^*)^2 &\lesssim \|\widehat{\beta}_n - \beta_k^*\|_e^2 \lesssim \widetilde{Q}(\beta_k^*) - \widetilde{Q}(\widehat{\beta}_n) = Q\left(h_k^*\right) - Q(\widehat{h}_n) \\ &= \mathrm{E}_Z(f_k^* - \widehat{f}_n) = \widehat{\delta}_n \lesssim_P (k/n)^{1/2}\,.\end{aligned}$$

Hence $\rho(\widehat{h}_n, h_k^*) \lesssim_P (k/n)^{1/4}$ with probability approaching one. Combining this result with Lemma 2.13, A.8, and the triangle inequality,

$$\rho(\widehat{h}_n, h_o) \leqslant \rho(\widehat{h}_n, h_k^*) + \rho(h_k^*, h_k) + \rho(h_k, h_o) \tag{2.A.18}$$

$$\lesssim_P (k/n)^{1/4} + k^{-\alpha} + k^{-\alpha} \lesssim_P (k/n)^{1/4} + k^{-\alpha}, \tag{2.A.19}$$

with probability approaching one as $n \to \infty$. This establishes Part 1.

To establish Part 2, let $\mathbf{x} \in \mathcal{X}$ be arbitrary. By the Bunyakovsky–Cauchy–Schwarz (BCS) inequality,

$$|\widehat{h}_n\left(\mathbf{x}\right) - h_k^*\left(\mathbf{x}\right)| \leqslant \|p^k\left(\mathbf{x}\right)\|_e \|\widehat{\beta}_n - \beta_k^*\|_e \leqslant \zeta_k \|\widehat{\beta}_n - \beta_k^*\|_e.$$

Since $\|\widehat{\beta}_n - \beta_k^*\|_e \lesssim \rho(\widehat{h}_n, h_k^*) \lesssim_P (k/n)^{1/4}$, we have

$$\|\widehat{h}_n - h_k\|_{\mathcal{X}} \lesssim_P \zeta_k (k/n)^{1/4}.$$

Again by the BCS inequality,

$$|h_k^*(\mathbf{x}) - h_k(\mathbf{x})| \leqslant \zeta_k \|\beta_k^* - \beta_k\|_e.$$

Since $\|\beta_k^* - \beta_k\|_e \lesssim \rho(h_k^*, h_k) \lesssim k^{-\alpha}$, we have

$$\|h_k^* - h_k\|_{\mathcal{X}} \lesssim \zeta_k k^{-\alpha}.$$

Now, by the triangle inequality and A.8,

$$\begin{aligned}\|\widehat{h}_n - h_o\|_{\mathcal{X}} &\leqslant \|\widehat{h}_n - h_k^*\|_{\mathcal{X}} + \|h_k^* - h_k\|_{\mathcal{X}} + \|h_k - h_o\|_{\mathcal{X}} \\ &\lesssim_P \zeta_k \left(k/n\right)^{1/4} + \zeta_k k^{-\alpha} + k^{-\alpha}.\end{aligned}$$

The third term is negligible compared to the second and may therefore be ignored. □

2.B Technical Appendix

The following theorem is van der Vaart and Wellner (2009) Theorem 1.1.

Theorem 2.5 (VC Index Bounds). *Let $V := \sum_{j=1}^{m} V_j$ be the sum of VC indices V_j from m VC–classes $\mathcal{C}_j$. Then the following bounds hold:* $\left\{ \begin{array}{c} V(\sqcup_{j=1}^{m}\mathcal{C}_j) \\ V(\sqcap_{j=1}^{m}\mathcal{C}_j) \\ V(\boxtimes_{j=1}^{m}\mathcal{C}_j) \end{array} \right\} \leqslant c_1 V \ln\left(\frac{c_2 m}{e^{Ent(\underline{V})/\overline{V}}}\right) \leqslant c_1 V \ln(c_2 m)$*, where* $\underline{V} := (V_1, \ldots, V_m), c_1 := e/[(e-1)\ln(2)] \doteq 2.28231\ldots, c_2 := e/\ln(2) \doteq 3.92165\ldots,$ *and*

$$Ent(\underline{V}) := \frac{1}{m}\sum_{j=1}^{m} V_j \ln(V_j) - \overline{V}\ln(\overline{V})$$

is the "entropy" of the V_j's under the discrete uniform distribution with weights $1/m$ and $\overline{V} := (1/m)\sum_{j=1}^{m} V_j$.

The following (Talagrand–type) inequality is Chernozhukov, Chetverikov, and Kato (2014a) Theorem B.1.

Theorem 2.6 (CCK Inequality). *Let $\mathbf{V}_1, \ldots, \mathbf{V}_n$ be i.i.d. random variables taking values in a measurable space $(S, \mathcal{S})$. Suppose that $\mathcal{G}$ is a nonempty, pointwise measurable class of functions on S uniformly bounded by a constant b such that there exists constants $a \geqslant e$ and $v > 1$ with $\sup_Q N(b\epsilon, \mathcal{G}, L_2(Q)) \leqslant (a/\epsilon)^v$ for $0 < \epsilon \leqslant 1$. Let σ^2 be a constant such that $\sup_{g\in\mathcal{G}} \mathrm{var}(g(\mathbf{V})) \leqslant \sigma^2 \leqslant b^2$. If $b^2 v \ln(ab/\sigma) \leqslant n\sigma^2$, then for all $t \leqslant n\sigma^2/b^2$,*

$$\mathrm{P}\Big(\sup_{g\in\mathcal{G}}\Big|\sum_{i=1}^{n} g(\mathbf{V}_i) - \mathrm{E}(g(\mathbf{V}))\Big| > A[n\sigma^2 \max(t, v\ln(ab/\sigma))]^{1/2}\Big) \leqslant e^{-t},$$

where $A > 0$ is an absolute constant.

Bibliography

Abrevaya, J. (2000). Rank estimation of a generalized fixed-effects regression model. *Journal of Econometrics 95*(1), 1–23.

Ai, C. and X. Chen (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica 71*(6), 1795–1843.

Aliprantis, C. D. and K. Border (2006). *Infinite dimensional analysis: A Hitchhiker's Guide.* Springer Science & Business Media.

Andrews, D. W. and W. Ploberger (1994). Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica: Journal of the Econometric Society*, 1383–1414.

Bajari, P., H. Hong, J. Krainer, and D. Nekipelov (2010). Estimating static models of strategic interactions. *Journal of Business & Economic Statistics 28*(4), 469–482.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica 80*(6), 2369–2429.

Belloni, A. and V. Chernozhukov (2011). *High dimensional sparse econometric models: An introduction.* Springer.

Belloni, A. and V. Chernozhukov (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli 19*(2), 521–547.

Belloni, A., V. Chernozhukov, D. Chetverikov, and K. Kato (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics 186*(2), 345–366.

Belloni, A., V. Chernozhukov, and C. Hansen (2014a). High-dimensional methods and inference on structural and treatment effects. *The Journal of Economic Perspectives 28*(2), 29–50.

Belloni, A., V. Chernozhukov, and C. Hansen (2014b). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies 81*(2), 608–650.

Belloni, A., V. Chernozhukov, and C. B. Hansen (2013). Inference for High-Dimensional Sparse Econometric Models. In *Advances in Economics and Econometrics*, Volume 3 of *Econometric Society Monographs*. Cambridge University Press.

Bera, A. K., G. Montes-Rojas, and W. Sosa-Escudero (2010). General specification testing with locally misspecified models. *Econometric Theory 26*(6), 1838–1845.

Berry, S. T. and P. A. Haile (2009). *Identification of a heterogeneous generalized regression model with group effects*. Yale University, Cowles Foundation for Research in Economics.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 1705–1732.

Bierens, H. J. (1982). Consistent model specification tests. *Journal of Econometrics 20*(1), 105–134.

Bierens, H. J. (1984). Model specification testing of time series regressions. *Journal of Econometrics 26*(3), 323–353.

Bierens, H. J. (1990). A consistent conditional moment test of functional form. *Econometrica: Journal of the Econometric Society*, 1443–1458.

Bierens, H. J. (2016). *Econometric Model Specification: Consistent Model Specification Tests and Semi-nonparametric... Modeling and Inference*. World Scientific.

Bierens, H. J. and W. Ploberger (1997). Asymptotic theory of integrated conditional moment tests. *Econometrica: Journal of the Econometric Society*, 1129–1151.

Boning, W. B. and F. Sowell (1999). Optimality for the integrated conditional moment test. *Econometric Theory 15*(5), 710–718.

Carrasco, M., J.-P. Florens, and E. Renault (2007). Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. *Handbook of econometrics 6*, 5633–5751.

Cavanagh, C. and R. P. Sherman (1998). Rank estimators for monotonic index models. *Journal of Econometrics 84*(2), 351–381.

Charlier, E., B. Melenberg, and A. v. Soest (1995). A smoothed maximum score estimator for the binary choice panel data model with an application to labour force participation. *Statistica Neerlandica 49*(3), 324–342.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics 6*, 5549–5632.

Chen, X., O. Linton, and I. Van Keilegom (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, 1591–1608.

Chen, X. and D. Pouzo (2009). Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *Journal of Econometrics 152*(1), 46–60.

Chen, X. and D. Pouzo (2012). Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica 80*(1), 277–321.

Chen, X. and X. Shen (1998). Sieve extremum estimates for weakly dependent data. *Econometrica*, 289–314.

Chen, X. and H. White (1999). Improved rates and asymptotic normality for nonparametric neural network estimators. *Information Theory, IEEE Transactions on 45*(2), 682–691.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey (2016). Double Machine Learning for Treatment and Causal Parameters. *arXiv:1608.00060 [stat]*. arXiv: 1608.00060.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2017). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*.

Chernozhukov, V., D. Chetverikov, and K. Kato (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics 41*(6), 2786–2819.

Chernozhukov, V., D. Chetverikov, and K. Kato (2014a). Anti-concentration and honest, adaptive confidence bands. *The Annals of Statistics 42*(5), 1787–1818.

Chernozhukov, V., D. Chetverikov, and K. Kato (2014b). Gaussian approximation of suprema of empirical processes. *The Annals of Statistics 42*(4), 1564–1597.

Chernozhukov, V., D. Chetverikov, and K. Kato (2015). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Probability Theory and Related Fields 162*(1-2), 47–70.

Chernozhukov, V., J. C. Escanciano, H. Ichimura, and W. K. Newey (2016). Locally robust semiparametric estimation. *arXiv preprint arXiv:1608.00033*.

Chernozhukov, V., C. Hansen, and M. Spindler (2015a). Post-selection and post-regularization inference in linear models with many controls and instruments. *The American Economic Review 105*(5), 486–490.

Chernozhukov, V., C. Hansen, and M. Spindler (2015b). Valid post-selection and post-regularization inference: An elementary, general approach. *Annual Review of Economics* (7), 649–688.

Cosslett, S. R. (1983). Distribution-free maximum likelihood estimator of the binary choice model. *Econometrica: Journal of the Econometric Society*, 765–782.

Davydov, Y. A., M. A. Lifshits, and Smorodina, N. V. (1998). *Local properties of distributions of stochastic functionals.* American Mathematical Soc.

de Jong, R. M. (1996). The Bierens test under data dependence. *Journal of Econometrics 72*(1), 1–32.

de la Pena, V. H., T. L. Lai, and Q.-M. Shao (2009). *Self-normalized processes. Probability and its Applications.* New York. Springer-Verlag, Berlin.

DeVore, R. A. and G. G. Lorentz (1993). *Constructive approximation*, Volume 303. Springer Science & Business Media.

Donald, S. G., G. W. Imbens, and W. K. Newey (2003). Empirical likelihood estimation and consistent tests with conditional moment restrictions. *Journal of Econometrics 117*(1), 55–93.

Donald, S. G. and W. K. Newey (1994). Series estimation of semilinear models. *Journal of Multivariate Analysis 50*(1), 30–40.

Dudley, R. M. (1978). Central limit theorems for empirical measures. *The Annals of Probability*, 899–929.

Escanciano, J. C. (2006). A consistent diagnostic test for regression models using projections. *Econometric Theory 22*(6), 1030–1051.

Fan, Y. and Q. Li (2000). Consistent model specification tests: Kernel-based tests versus Bierens' ICM tests. *Econometric Theory 16*(6), 1016–1041.

Fuller, W. C., C. F. Manski, and D. A. Wise (1982). New evidence on the economic determinants of postsecondary schooling choices. *Journal of Human Resources*, 477–498.

Gozalo, P. L. (1993). A consistent model specification test for nonparametric estimation of regression function models. *Econometric Theory 9*(3), 451–477.

Grenander, U. (1981). *Abstract inference*. Wiley New York.

Han, A. K. (1987). Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. *Journal of Econometrics 35*(2), 303–316.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, 1029–1054.

Hardle, W. and E. Mammen (1993). Comparing nonparametric versus parametric regression fits. *The Annals of Statistics*, 1926–1947.

Hong, Y. and H. White (1995). Consistent specification testing via nonparametric series regression. *Econometrica: Journal of the Econometric Society*, 1133–1159.

Honoré, B. E. (1992). Trimmed LAD and least squares estimation of truncated and censored regression models with fixed effects. *Econometrica: journal of the Econometric Society*, 533–565.

Horowitz, J. L. and W. Härdle (1994). Testing a parametric model against a semiparametric alternative. *Econometric theory 10*(5), 821–848.

Huang, J. Z. (1998). Projection estimation in multiple regression with application to functional ANOVA models. *The annals of statistics 26*(1), 242–272.

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference*. Springer Science & Business Media.

Lavergne, P. and Q. Vuong (2000). Nonparametric significance testing. *Econometric Theory 16*(4), 576–601.

Ledoux, M. and M. Talagrand (2013). *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media.

Lee, L.-f. (2005). A C(alpha)-type gradient test in the GMM approach. *Working Paper, Department of Economics, Ohio State University.*

Leeb, H. and B. M. Pötscher (2008). Sparse estimators and the oracle property, or the return of Hodges' estimator. *Journal of Econometrics 142*(1), 201–211.

Li, Q. and S. Wang (1998). A simple consistent bootstrap test for a parametric regression function. *Journal of Econometrics 87*(1), 145–165.

Lorentz, G. G. (1966). *Approximation of functions.* Holt, Rinehart and Winston.

Manski, C. (1991). Nonparametric estimation of expectations in the analysis of discrete choice under uncertainty. In *Nonparametric and semiparametric methods in econometrics and statistics*, pp. 259–275.

Manski, C. F. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics 3*(3), 205–228.

Manski, C. F. (1985). Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of econometrics 27*(3), 313–333.

Manski, C. F. (1987). Semiparametric analysis of random effects linear models from binary panel data. *Econometrica: Journal of the Econometric Society*, 357–362.

Marchal, O. and J. Arbel (2017, April). On the sub-Gaussianity of the Beta and Dirichlet distributions. *arXiv:1705.00048 [math, stat]*. arXiv: 1705.00048.

Massart, P. (2000). About the constants in Talagrand's concentration inequalities for empirical processes. *Annals of Probability*, 863–884.

Matzkin, R. L. (1991). A nonparametric maximum rank correlation estimator. In *Nonparametric and Semiparametric Methods in Econometrics and Statistics: Proceedings of the Fifth International Symposium in Economic Theory and Econometrics*, Volume 5, pp. 277. Cambridge University Press.

Matzkin, R. L. (1992). Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models. *Econometrica: Journal of the Econometric Society*, 239–270.

Matzkin, R. L. (2007). Nonparametric identification. *Handbook of Econometrics 6*, 5307–5368.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of applied econometrics 5*(2), 99–135.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, 1349–1382.

Newey, W. K. (1995). Convergence rates for series estimators. In G. Maddala, P. C. Phillips, and T. Srinivasan (Eds.), *Advances in Econometrics and Quantitative Economics*. Blackwell.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics 79*(1), 147–168.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics 4*, 2111–2245.

Neyman, J. (1959). Optimal asymptotic tests of composite statistical hypotheses. *Probability and statistics 57*, 213.

Ossiander, M. (1987). A central limit theorem under metric entropy with L2 bracketing. *The Annals of Probability*, 897–919.

Pakes, A. and D. Pollard (1989). Simulation and the asymptotics of optimization estimators. *Econometrica: Journal of the Econometric Society*, 1027–1057.

Pollard, D. (1990). Empirical processes: theory and applications. In *NSF-CBMS regional conference series in probability and statistics*, pp. i–86. JSTOR.

Poterba, J. M., S. F. Venti, and D. A. Wise (1994). 401 (k) plans and tax-deferred saving. In *Studies in the Economics of Aging*, pp. 105–142. University of Chicago Press.

Poterba, J. M., S. F. Venti, and D. A. Wise (1995). Do 401 (k) contributions crowd out other personal saving? *Journal of Public Economics 58*(1), 1–32.

Rao, C. R. (1973). *Linear statistical inference and its applications*, Volume 2. Wiley New York.

Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, 931–954.

Rudelson, M. (1999). Random vectors in the isotropic position. *Journal of Functional Analysis 164*(1), 60–72.

Rudelson, M. and R. Vershynin (2008). On sparse reconstruction from Fourier and Gaussian measurements. *Communications on Pure and Applied Mathematics 61*(8), 1025–1045.

Rudin, W. (1976). Principles of Mathematical Analysis (International Series in Pure & Applied Mathematics).

Santos, A. (2012). Inference in nonparametric instrumental variables with partial identification. *Econometrica 80*(1), 213–275.

Schläfli, L. (1901). *Theorie der vielfachen Kontinuität*, Volume 38. Zürcher & Furrer.

Schumaker, L. (2007). *Spline functions: basic theory.* Cambridge University Press.

Shen, X. and W. H. Wong (1994). Convergence rate of sieve estimates. *The Annals of Statistics*, 580–615.

Sherman, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica: Journal of the Econometric Society*, 123–137.

Souza-Rodrigues, E. (2014). Nonparametric estimation of a generalized regression model with group effects. *University of Toronto Department of Economics Working Paper*.

Stinchcombe, M. B. and H. White (1998). Consistent specification testing with nuisance parameters present only under the alternative. *Econometric theory 14*(03), 295–325.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The annals of Statistics*, 1348–1360.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 1040–1053.

Stone, C. J. (1985). Additive regression and other nonparametric models. *The annals of Statistics*, 689–705.

Stute, W. (1997). Nonparametric model checks for regression. *The Annals of Statistics*, 613–641.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Timan, A. F. (1963). *Theory of approximation of functions of a real variable*, Volume 34. Courier Corporation.

van de Geer, S. (2000). *Empirical Processes in M-estimation*, Volume 6. Cambridge university press Cambridge.

van der Vaart, A. and J. A. Wellner (2011). A local maximal inequality under uniform entropy. *Electronic Journal of Statistics 5*(2011), 192.

van der Vaart, A. W. (2000). *Asymptotic statistics*, Volume 3. Cambridge university press.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes*. Springer.

van der Vaart, A. W. and J. A. Wellner (2009). A note on bounds for VC dimensions. *Institute of Mathematical Statistics collections 5*, 103.

Whang, Y.-J. (2001). Consistent specification testing for conditional moment restrictions. *Economics Letters 71*(3), 299–306.

Willis, R. J. and S. Rosen (1979). Education and self-selection. *Journal of political Economy 87*(5, Part 2), S7–S36.

Wooldridge, J. M. (1991). On the application of robust, regression-based diagnostics to models of conditional means and conditional variances. *Journal of econometrics 47*(1), 5–46.

Wooldridge, J. M. (1992). A test for functional form against nonparametric alternatives. *Econometric Theory 8*(4), 452–475.

Yatchew, A. J. (1992). Nonparametric regression tests based on least squares. *Econometric Theory 8*(4), 435–451.

Zheng, J. X. (1996). A consistent test of functional form via nonparametric estimation techniques. *Journal of Econometrics 75*(2), 263–289.