Three Essays in Microeconometrics

by

Max H. Farrell

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Economics)
in The University of Michigan
2014

Doctoral Committee:

Associate Professor Matias D. Cattaneo, Chair
Professor Xuming He
Professor Lutz Kilian
Professor Jeffrey A. Smith

For my wife Mindy, my son Henry, and my daughter Rosalind

ACKNOWLEDGEMENTS

First and foremost, I thank my loving wife Mindy. Without her love, support, and encouragement, this dissertation would not exist at all. She is the foundation upon which I stand and the pinnacle to which I aspire. I am deeply thankful and grateful for our children, Henry and Rosalind. Although they present the greatest challenges in life, they also give the greatest rewards. Their joy has made this work possible. This dissertation sometimes came at the expense of my family, and I hope they agree that it was worth the sacrifice.

I can not adequately express my gratitude to my advisor, Matias Cattaneo. Matias and I started together in the Department of Economics six years ago, and he first inspired me to study econometrics. In that time, I have continuously benefited from his guidance and friendship. I thank him for his encouragement, insight, generosity with time, and seemingly inexhaustible patience. Matias is responsible for turning me into the scholar I am today (for better or worse).

I have been lucky enough to be surrounded by excellent faculty at the University of Michigan, and though I have benefited from the wisdom and experience of many, I must single out the efforts of Xuming He, Lutz Kilian, and Jeff Smith. Xuming kept me on sound statistical ground, and his insight has been invaluable. Lutz broadened my horizons, and could always be counted upon to separate the wheat from the chaff. Jeff has always helped keep my work tied to the real world of economics.

I wish to thank the many friends in my life, whose support kept me going. I am particularly grateful to Eric Ohrn for making the countless hours in Lorch 119 (and the basement of Hatcher and the 2^{nd} floor of Shapiro) bearable, and even fun.

Finally, I must thank Jay Shelton and Jonathan Gruber for putting me on this path, and for keeping me there.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

CHAPTER I

Introduction

This dissertation explores the theoretical finite sample and asymptotic properties of several econometric estimators. Two central themes are robustness and pragmatism. This dissertation develops theory that is focused on real-world problems faced in applied econometrics. To provide increased robustness and reliability in applications, theories and tools must accurately capture the true behavior and construction of estimators. This often means taking explicit account of procedures (e.g. variable selection in Chapter II) that traditional, nonrobust theories ignore. By capturing and studying these steps, complex and technical though they may be, it is often possible to deliver highly robust results that are also accessible and easily implemented. Although this dissertation spans a broad range of microeconometrics, this common foundation links the chapters together. That is, the goal of the theory develop is to deliver practicable methods for empirical research or put already-common practices on sound theoretical footing.

Chapter II concerns robust inference on average treatment effects following model selection. In the selection on observables framework, this chapter shows how to construct confidence intervals based on a doubly-robust estimator that are robust to model selection errors and prove that they are valid uniformly over a large class of treatment effect models. The class allows for multivalued treatments with heterogeneous effects (in observables), general heteroskedasticity, and selection amongst (possibly) more covariates than observations. The estimator attains the semiparametric efficiency bound under appropriate conditions. Precise conditions are given for any model selector to yield these results, and it is shown how to combine data-driven selection with economic theory. For implementation, the group lasso is proposed and new technical results are derived for high-dimensional, sparse multinomial logistic regression. A simulation study shows that the estimator performs very well in finite samples

over a wide range of models. Revisiting the National Supported Work demonstration data, the method yields accurate estimates and tight confidence intervals.

Chapter III, joint with Matias Cattaneo, studies the asymptotic properties of partitioning estimators of the conditional expectation function and its derivatives. Mean-square and uniform convergence rates are established and shown to be optimal under simple and intuitive conditions. The uniform rate explicitly accounts for the effect of moment assumptions, which is useful in semiparametric inference. A general asymptotic integrated mean-square error approximation is obtained and used to derive an optimal plug-in tuning parameter selector. A uniform Bahadur representation is developed for linear functionals of the estimator. Using this representation, asymptotic normality is established, along with consistency of a standard-error estimator. The finite-sample performance of the partitioning estimator is examined and compared to other nonparametric techniques in an extensive simulation study.

Chapter IV, also joint with Matias Cattaneo, studies the large sample properties of a subclassification-based estimator of the Dose-Response Function under Ignorability. This work relies on the theory developed in the prior chapter, in addition to newly developed results. Employing standard regularity conditions, it is shown that the estimator is root-n consistent, asymptotically linear, and semiparametric efficient in large samples. A consistent estimator of the standard-error is also developed under the same assumptions. In a Monte Carlo experiment the finite sample performance of this simple and intuitive estimator is compared to others commonly employed in the literature.

CHAPTER II

Robust Inference on Average Treatment Effects with Possibly More Covariates than Observations

2.1 Introduction

Model selection has always had a place in empirical economics, whether or not it is formally acknowledged. A key problem in modern empirical work is that researchers face datasets with large numbers of variables, sometimes more than observations. A complementary problem is that economic theory and prior knowledge may mandate controlling for certain variables, but are generally silent regarding functional form. These two problems force researchers to search for a model that is simultaneously parsimonious and adequately flexible. Many formal methods are computationally infeasible with a large number of variables. A typical response to this challenge is to iteratively search over a small set of alternative specifications, guided only by the researcher's taste and intuition. No matter the approach used, inference almost never takes into account this "specification search" and the resulting confidence intervals are not robust to model selection mistakes, and hence are unreliable in empirical work.

This problem is particularly important in estimating average treatment effects under selection on observables, because in this framework using the right covariates is crucial for identification and correct inference. In this context, we provide an easy-to-implement and objective method for covariate selection and post-selection inference on average treatment effects.[1] We establish four main results for multivalued treatments effects with arbitrary heterogeneity in observables and heteroskedasticity. First

[1] Treatment effects, missing data, measurement error, and data combination models are equivalent under selection on observables. Thus, all our results immediately apply to those contexts. For reviews of these literatures, see Tsiatis (2006), Heckman and Vytlacil (2007), Imbens and Wooldridge (2009), and Wooldridge (2010).

and foremost, we show that a doubly-robust estimator is robust to model selection errors, a newly-discovered virtue of this class of estimators.[2] By taking explicit account of the model selection stage and its inherent selection errors, we derive precise conditions required for any model selector to deliver confidence intervals for average treatment effects that are uniformly valid over a large class of data-generating processes. Second, we show that a simple refitting procedure allows researchers to augment variables chosen according economic theory with data-driven selection to deliver flexible inference that remains uniformly valid. Third, we prove that our proposed estimator is asymptotically linear and attains the semiparametric efficiency bound, under standard conditions imposed in the program evaluation literature. Fourth, we derive new technical results for multinomial (and binary) logistic regression, the most widely used model for treatment assignment.

Inference following model selection is notoriously difficult. In a sequence of papers, Leeb and Pötscher (2005, 2008a, 2008b, 2009, 2009) have shown that inference relying too heavily on model selection can not be made uniformly valid. Loosely speaking, uniform validity of a confidence interval captures the idea that the interval should have the same quality (coverage) for many data-generating processes. This theoretical property is practically important because it implies greater reliability in applications. Our proposed methods for post model selection inference build upon the path-breaking recent work of Belloni, Chernozhukov, and Hansen (2013). We circumvent, without contradicting, the impossibility results of Leeb and Pötscher by not insisting on perfect selection, but rather explicitly accounting for inevitable model selection errors in the asymptotic approximations.[3]

Our approach, based on the doubly-robust estimator, has several key features. The name “doubly-robust” reflects that it is robust to misspecification of either the treatment equation (propensity score) or the outcome equation, a property obtained by combining inverse probability weighting and regression imputation. First, we show that this robustness extends to model selection, enabling us to allow for selection errors in both equations without impacting inference. Second, we capture arbitrary treatment effect heterogeneity (dependence of the effect on an individual’s observed characteristics), which is crucial in empirical work. With such heterogeneity, the average treatment effect and the treatment on the treated differ, and hence we present

[2] Doubly-robust estimation and its role in program evaluation is discussed by Robins and Rotnitzky (1995), Kang and Schafer (2007, with discussion), van der Laan and Robins (2003), Tan (2010), and references therein.

[3] Efron (2013) and Berk, Brown, Buja, Zhang, and Zhao (2013) also propose methods for post-selection inference, both quite distinct from our method.

results for both. Third, the doubly-robust estimator also stems from the semiparametric efficient moment conditions, and hence we obtain the semiparametric efficiency bound, even under heteroskedasticity, under standard additional conditions. Taking all these features together enables us to obtain uniform inference over such a large class of treatment effects models.

In recent independent work, Belloni, Chernozhukov, and Hansen (2013, draft dated July 19), propose a similar approach. Their main focus is a partially linear model, in which the coefficient of a treatment indicator will recover the average effect of a binary treatment only if the effect is constant across observables, but Section 5 of their most recent draft, developed independently from our work, considers heterogeneous effects. There are two broad differences in our approaches. First, we allow for multivalued treatments, which offers a larger set of estimands and can thus enhance the understanding of program impacts.[4] We show how to improve model selection in this context by pooling information across treatment levels. Second, although in both cases the doubly-robust estimator is used for average treatment effects following a (quite different) model selection step,[5] we exploit certain features of the estimator to produce two benefits: (i) our procedure requires demonstrably weaker conditions on the model selection stage (see Assumption II.5); and (ii) none of our results require using variables selected for the treatment equation in the outcome model estimation, and vice versa (their "post double selection" method), and indeed, we show doing so requires stronger assumptions (see Assumption II.6).

Our analysis is conducted under selection on observables, which has a long tradition and remains quite popular in empirical economics.[6] Covariate selection has three crucial roles to play in this framework. First, using more observed covariates, and more flexibly, may help proxy for unobserved confounding and hence increase the plausibility of unconfoundedness. Second, it is natural that some variables are not part of the causal mechanism under study, and therefore should be excluded. Third, the efficient conditioning set must contain those variables that drive the outcome, which are not necessarily those important for treatment assignment. This reasoning mandates contradicting goals for practitioners: a large, rich set of controls on the one hand, and parsimony on the other. Our approach is a formal, theory-driven attempt

[4] Discussion and applications may be found in, for example Imbens (2000), Lechner (2001), Imai and van Dyk (2004), Abadie (2005), Cattaneo (2010), and Cattaneo and Farrell (2011b).

[5] They use different asymptotic variance estimators, and for treatment effects on the treated they do not exploit the simplification discussed in Remark 1.

[6] For other approaches and reviews of the literature, see, e.g., Holland (1986), Hahn (1998), Horowitz and Manski (2000), Chen, Hong, and Tarozzi (2004, 2008), Bang and Robins (2005), Abadie and Imbens (2006), Wooldridge (2007), and references therein.

to reconcile this contradiction.

A special feature of our analysis is that we match the empirical realities of large data sets by considering selection from amongst (possibly) more covariates than observations, so-called *high-dimensional* data. The goal of variable selection is to find a small model that is nonetheless sufficiently flexible to capture unknown features of the data-generating process required for inference. If a small model can perfectly capture the unknown feature it is said to be *exactly sparse*. A far more realistic scenario is *approximate sparsity*, when the bias from using a small model is well-controlled, but nonzero. Sparsity is a natural framework for thinking about model selection. Indeed, any time only a few of the available variables are used, a sparsity assumption has effectively been made. It is common empirical practice to report results from several small models, but for these results to be valid one must assume these specifications give high-quality, sparse representations of the unknown features. The alternative we provide involves selecting a sparse, yet flexible, model from among a large set of variables. Results may then be compared with more traditional methods used in practice.

With the aim of mimicking common empirical practice we estimate the propensity score with multinomial logistic regression. To handle this nonlinear model under approximate sparsity we employ the group lasso (Yuan and Lin 2006) coupled with a novel penalty that controls both the noise and bias simultaneously. In our view, the group lasso is particularly well-suited to multivalued treatments because it pools information across all treatment levels to aid selection. Our results are stated in the language of treatment effects, but apply to general data structures and are of independent interest.[7] To the best of our knowledge this is the first detailed study of an approximately sparse, nonlinear model in the high-dimensional literature. Much of the literature has focused on linear models (see Buhlmann and van de Geer (2011) for a survey), while prior studies of nonlinear models often assume exact sparsity, or present limited results.[8] Furthermore, these studies often use high-level conditions that can be hard to verify. In contrast, we obtain sharp results for logistic regression

[7] Our techniques build on prior studies, in particular Bickel, Ritov, and Tsybakov (2009), Lounici, Pontil, van de Geer, and Tsybakov (2009, 2011), Obozinski, Wainwright, and Jordan (2011), Belloni and Chernozhukov (2011b), Belloni, Chen, Chernozhukov, and Hansen (2012).

[8] Examples include van de Geer (2008), Belloni and Chernozhukov (2011a), or Negahban, Ravikumar, Wainwright, and Yu (2012). Bach (2010) only gives an error bound on coefficients in exactly sparse logistic regression, which can not yield our results; and does not consider prediction error or post-selection estimation. In independent work, Belloni, Chernozhukov, and Wei (2013) study exactly sparse logistic regression, also using Bach's (2010) tools, but are focused on a different inference goal.

under the same simple and intuitive conditions used for linear modeling by exploiting mathematical techniques of self-concordant functions put forth by Bach (2010). We also provide extensions to prior work on linear models needed to apply them in treatment effect estimation.

Finally, we offer numerical evidence on the finite sample performance of our procedure. In a small simulation study we find that our procedure delivers very accurate coverage of confidence intervals even for models where covariate selection is difficult, either because of a low signal-to-noise ratio or lack of sparsity, thus highlighting the uniform validity of inference. We also apply our method to the widely-used National Supported Work Demonstration data (LaLonde 1986) and find very accurate estimates and tight confidence intervals (see Table 2.1).

The remainder of the paper proceeds as follows. In Section 2.2, we give a short, self-contained overview of the main results. For ease of reference, Section 2.2.3 collects all notation. Section 2.3 describes the treatment effect model. Sparse models are discussed in Section 2.4, which shows how several models that are commonly used in empirical work fit in this framework. Section 2.5 presents our estimation method and gives complete results on treatment effect inference. The proposed group lasso approach to sparse modeling is detailed in Section 2.6, including theoretical results on model selection and estimation. Section 2.7 presents the numerical evidence on finite sample behavior of our procedure. Section 2.8 concludes. The main proofs are presented in the Appendix, while the remainder are available in a supplement.

2.2 Overview of Results and Notation

Here we give an overview of the main contributions of the paper. We first discuss treatment effect inference with a general model selector. Then in Section 2.2.2 we discuss our new results for the group lasso, our proposed model selector. Section 2.2.3 collects notation to be used throughout.

2.2.1 Treatment Effects and Results on Post-Selection Inference

We consider a multivalued treatment, with status indicated by $D \in \{0, 1, \ldots, \mathcal{T}\}$. Interest lies in mean effects of the treatment on a scalar outcome Y. Let $\{Y(t)\}_{t=0}^{\mathcal{T}}$ be the (latent) potential outcomes: $Y(t)$ is the outcome a unit would have under $D = t$. $Y(t)$ is only observed for units with $D = t$; that is, $Y = \sum_{t=0}^{\mathcal{T}} \mathbb{1}\{D = t\}Y(t)$. Many interesting parameters combine means of potential outcomes, and having multivalued treatments allows for a wider range of estimands. Define the mean of one potential outcome as

$$\mu_t = \mathbb{E}[Y(t)].$$

To fix ideas, $\mu_1 - \mu_0$ is the average treatment effect in the binary case ($D \in \{0, 1\}$). Sections 2.3 and 2.5 consider more general average effects, including effects on treated groups. For simplicity, in this section we focus a single μ_t.

We use the selection on observables framework to identify μ_t. For a vector of covariates X, define the generalized propensity score and conditional outcome regressions as

$$p_t(x) = \mathbb{P}[D = t | X = x] \qquad \text{and} \qquad \mu_t(x) = \mathbb{E}[Y | D = t, X = x].$$

For identification it is sufficient to assume that $\mathbb{E}[Y(t)|D, X] = \mathbb{E}[Y(t)|X]$ (mean independence) and $p_t(X)$ is bounded away from zero (overlap) for all treatment levels. Broadly, these two assumptions imply that units from one treatment group are good proxies for other treatments and that there are always such proxies available (see Section 2.3).

Suppose we have an i.i.d. sample $\{(y_i, d_i, x_i')\}_{i=1}^n$ from (Y, D, X'). Then, for model-selection-based estimators $\hat{p}_t(x_i)$ and $\hat{\mu}_t(x_i)$, we estimate μ_t with

$$\hat{\mu}_t = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{\mathbb{1}\{d_i = t\}(y_i - \hat{\mu}_t(x_i))}{\hat{p}_t(x_i)} + \hat{\mu}_t(x_i) \right\}.$$

This doubly-robust estimator combines regression imputation and inverse probability weighting, and remains consistent if either $p_t(x)$ or $\mu_t(x)$ is misspecified. Following widespread empirical practice, we estimate $\hat{p}_t(x_i)$ with multinomial logistic regression and $\hat{\mu}_t(x_i)$ linearly (see Section 2.6). The choice of covariates in $\hat{p}_t(x_i)$ and $\hat{\mu}_t(x_i)$ is crucial, impacting consistency, efficiency, and finite sample performance. Covariate selection based on ad hoc, iterative searches is common in empirical evaluations, but is informal, not objective, and not replicable. Balancing tests are also commonly used in this context, but have the additional drawback of assuming the same covariates are important for outcomes and treatment assignment, and more generally do not weight the covariates by their importance for bias.

On the other hand, our proposed procedure gives practitioners an easy to implement, fully objective tool to perform data-driven covariate selection and treatment effect inference, with replicable results.[9] Importantly, we do not preclude the addition

[9] For the final estimation step, the doubly-robust estimator is available in STATA 13 and the package of Cattaneo, Drukker, and Holland (forthcoming) (as the default). The covariate selection

of variables known to be important from economic theory or prior knowledge. Our procedure is intended to supplement these variables with a flexible set of controls, guarding against misspecification or overfitting.

The following theorem is an example of the more general results presented in Section 2.5.2, wherein we also define V_t and $\hat{V}_t = \hat{V}^W_{\boldsymbol{\mu}}(t) + \hat{V}^B_{\boldsymbol{\mu}}(t,t)$.

Theorem II.1. *Consider a sequence $\{P_n\}$ of data-generating processes that obey, for each n, Assumptions II.3, and II.4 below. We require two conditions on the model selector:*

(i) $\sum_{i=1}^n (\hat{p}_t(x_i) - p_t(x_i))^2/n = o_{P_n}(1)$ *and* $\sum_{i=1}^n (\hat{\mu}_t(x_i) - \mu_t(x_i))^2/n = o_{P_n}(1)$*;*

(ii) $\left(\sum_{i=1}^n \mathbb{1}\{d_i = t\}(\hat{p}_t(x_i) - p_t(x_i))^2/n\right)\left(\sum_{i=1}^n \mathbb{1}\{d_i = t\}(\hat{\mu}_t(x_i) - \mu_t(x_i))^2/n\right) = o_{P_n}(n^{-1})$.

Under these conditions, $\sqrt{n}(\hat{\mu}_t - \mu_t) \to_d N(0, V_t)$ and $\hat{V}_t/V_t \to_{P_n} 1$. For each n, let $\boldsymbol{P}_n$ be the set of data-generating processes satisfying Assumption II.3, and II.4 and conditions (i) and (ii). Then

$$\sup_{P \in \boldsymbol{P}_n} \left| \mathbb{P}_P \left[\mu_t \in \left\{ \hat{\mu}_t \pm c_\alpha \sqrt{\hat{V}_t/n} \right\} \right] - (1-\alpha) \right| \to 0,$$

where $c_\alpha = \Phi^{-1}(1 - \alpha/2)$.

This result establishes the uniform validity of an asymptotic confidence interval for μ_t, overcoming all the post model selection inference challenges: robustness to model selection errors, selecting a model that is small but flexible enough to capture the features of the underlying data generating process, and still retaining efficiency under additional, standard conditions (see Section 2.5.3). Intuitively, this is similar to (but distinct from) overcoming pretesting bias in other contexts.

Two general conditions are placed on the model selector. The first is a mild consistency requirement. The second is analogous to the commonly-used, high-level requirement in semiparametrics that first-stage components converge faster than $n^{-1/4}$. However, because we use the doubly-robust moment condition we only have the product of the two estimation errors; this requirement can be easier to satisfy if one or the other function is easier to estimate (e.g. if one function is very smooth or very sparse). In high-dimensional models the rates for the first stage depend on the sample size, the number of covariates considered, and the sparsity level. We propose to use

stage is easily implemented in R; code is available upon request. A self-contained STATA package is under development.

the group lasso and prove that these estimators satisfy (i) and (ii). Importantly, the rate will depend on the total number of covariates only logarithmically, allowing for a large number.

2.2.2 Model Selection Stage

We propose refitting following group lasso selection, and show that it meets all requirements on the model selector. The group lasso is well-suited to program evaluation applications because covariates are penalized according to their overall contribution in all treatment groups. This has two consequences. First, information from all treatments is pooled when doing selection, and hence a weaker signal may be extracted, which improves the selection properties. Second, the selected variables are common to all treatment levels. From a practical point of view this is desirable, as interest rarely lies in a single μ_t, but rather a collection, and substantial commonality is expected in the variables important for different treatment levels. Formally, the group lasso gives the union of all supports. The group lasso is easy to implement, as discussed in Remark 4.

We consider high-dimensional, sparse models for $p_t(x)$ and $\mu_t(x)$. These are defined by a p-dimensional vector X^* based on the original variables X, with $p > n$ allowed. The X^* may consist of any combination of the original variables, interactions, flexible parametric transformations, and/or nonparametric series terms (such as splines or polynomials). A model is approximately sparse if there are $s < n$ of these terms that yield a good approximation ($s \to \infty$ is allowed). To build intuition, suppose that $p_t(x)$ and $\mu_t(x)$ follow p-dimensional parametric models. Then the sparsity assumption is that there is an s-dimensional model that has sufficiently small specification bias. In the nonparametric case, sparsity is weaker than (but analogous to) the familiar assumption that a small set of basis functions can approximate the unknown objects well. In practice researchers employ a hybrid of these approaches, which is covered by our results. Section 2.4 gives more detail and examples.

We form $\hat{p}_t(x)$ and $\hat{\mu}_t(x)$ in two steps (complete details in Section 2.6). First, the group lasso is applied separately to multinomial logistic and least squares regression to select covariates from X^*. We then estimate $p_t(x)$ and $\mu_t(x)$ by refitting unpenalized models using the selected variables, possibly augmented with controls that are known to be important from prior work or economic theory. It is not desirable for a model selector to discard theory and prior work, and our procedure explicitly avoids this. We also allow for using logistic-selected variables in the linear model refitting and vice versa, but this is not necessary for uniform inference nor efficiency (and requires stronger assumptions).

Our main results give precise bounds for the number of covariates selected and the estimation error, both for the penalized and unpenalized estimates. These bounds, given in Section 2.6.3, are nonasymptotic: exact constants are given for each bound and these bounds are valid for any given n, p, and s, provided our assumptions are met. Such results are complex and so we give the following intuitive, asymptotic result (The notation O_{P_n} is defined in Section 2.2.3).

Corollary II.2. *Suppose the biases from the best s_d- and s_y-term approximations to $p_t(x)$ and $\mu_t(x)$ are bounded by b_s^d and b_s^y, respectively. Then under the assumptions in Section 2.6.3, and $\delta > 0$ described therein, with high probability we have:*

1. $\sum_{i=1}^n (\hat{p}_t(x_i) - p_t(x_i))^2/n = O_{P_n}\left(n^{-1} s_d \log(p \vee n)^{3/2+\delta} + (b_s^d)^2 s_d\right)$ *and*

2. $\sum_{i=1}^n (\hat{\mu}_t(x_i) - \mu_t(x_i))^2/n = O_{P_n}\left(n^{-1} s_y \log(p \vee n)^{3/2+\delta} + (b_s^y)^2\right)$.

These two results for our proposed group lasso estimators can be directly used to verify the high-level conditions in Theorem II.1 above. The product of these rates makes explicit the advantage discussed above of using condition (ii) in Theorem II.1, by showing exactly how the errors depend on the total number of variables, the sparsity, and the bias. Section 2.6.3 also shows that the number of variables selected is the same order as the sparsity level, and provides bounds on the logistic and linear coefficients directly. Both these results are important for certain steps in treatment effect estimation that aren't reflected in the simple statement of Theorem II.1. These results appear to be entirely new for the multinomial logistic regression, for any version of the lasso. From a practical point of view, these results provide formal justification for using multinomial logistic regression, coupled with group lasso selection and post-selection refitting.

2.2.3 Notation

We collect here notation to be used for the rest of the paper. The population data generating process (DGP) is denoted by P_n and is defined by the joint law of the random variables $(Y, D, X')'$. For a given n, $\{(y_i, d_i, x_i')'\}_{i=1}^n$ constitute n i.i.d. draws from P_n. In general, the DGP may vary with n, along with features such as parameters, distributions, an so forth, as discussed in Section 2.4.2. This is generally suppressed for notational clarity.

We further adopt the following conventions.

Treatments. Define the treatment sets $\overline{\mathbb{N}}_{\mathcal{T}} = \{0, 1, 2, \ldots, \mathcal{T}\}$ and $\mathbb{N}_{\mathcal{T}} = \{1, 2, \ldots, \mathcal{T}\}$. No order is assumed in the treatments. For each unit i, d_i indicates treatment assignment, and define $d_i^t = \mathbb{1}\{d_i = t\}$. Let $n_t = \sum_{i=1}^n d_i^t$ be the number of individuals with treatment t and define $\underline{n} = \min_{t \in \overline{\mathbb{N}}_{\mathcal{T}}} n_t$ and $\overline{n} = \max_{t \in \overline{\mathbb{N}}_{\mathcal{T}}} n_t$. Further define $\overline{\mathcal{T}} = \mathcal{T} + 1$.

Vectors. Define $\mathbb{N}_p = \{1, 2, \ldots, p\}$. For a doubly-indexed collection of scalars $\{\delta_{t,j} : t \in \overline{\mathbb{N}}_{\mathcal{T}}, j \in \mathbb{N}_p\}$, define $\delta_{\cdot,j} \in \mathbb{R}^{\overline{\mathcal{T}}}$ as the vector that collects over all t for fixed j; $\delta_{t,\cdot} \in \mathbb{R}^p$ collects over $j \in \mathbb{N}_p$ for fixed t; and $\delta_{\cdot,\cdot} \in \mathbb{R}^{p \times \overline{\mathcal{T}}}$ the concatenation of all $\delta_{t,\cdot}$. For simplicity, we write δ_t for $\delta_{t,\cdot}$. When considering the multinomial logistic model, t will vary only over $\mathbb{N}_{\mathcal{T}}$ but the notation will be maintained (or, equivalently, normalize $\delta_{0,\cdot} = 0$). For a set $S \subset \mathbb{N}_p$, let $\delta_{t,S} \in \mathbb{R}^{\text{card}(S)}$ be the vector of $\{\delta_{t,j} : j \in S\}$ for fixed t and similarly let $\delta_{\cdot,S} \in \mathbb{R}^{|S| \times \overline{\mathcal{T}}} = \{\delta_{t,j} : t \in \overline{\mathbb{N}}_{\mathcal{T}}, j \in S\}$.

Norms. Single bars will be either absolute value or cardinality of a set, and will be clear from the context. For a vector v, let $\|v\|_1$ and $\|v\|_2$ denote the ℓ_1 and ℓ_2 norms, respectively. For the group lasso, define the mixed ℓ_2/ℓ_1 norm as $|||\delta_{\cdot,\cdot}|||_{2,1} = \sum_{j \in \mathbb{N}_p} \|\delta_{\cdot,j}\|_2$. It will always be the case that the ("outer") ℓ_1 norm is over the covariates and the ("inner") ℓ_2 norm is over the treatments (in our application). When discussing the multinomial logistic model, treatments will be restricted to $\mathbb{N}_{\mathcal{T}}$ with no change in notation.

Data-Generating Processes. The DGP for a fixed n will be denoted by P_n. The set of all such P_n that we allow for will be $\boldsymbol{P}_n$. As shorthand for a sequence we will use $\{P_n\} = \{P_n : n \geq 1, P_n \in \boldsymbol{P}_n\}$. Expectations and probabilities will be understood to be taken against P_n, though notationally suppressed: $\mathbb{E}[W] = \mathbb{E}_{P_n}[W]$ denotes the population expectation for a random variable W and $\mathbb{P}[A] = \mathbb{P}_{P_n}[A]$ the probability of event A. For asymptotic arguments dependence on n is explicit, so that $O_{P_n}(\cdot)$ and $o_{P_n}(\cdot)$ have their usual meaning with the understanding that the measure P_n is used for each n.

The empirical expectation will be denoted $\mathbb{E}_n[w_i] = \sum_{i=1}^n w_i/n$. Also, define $\mathbb{E}_{n,t}[w_i] = \sum_{i \in \mathbb{I}_t} w_i/n_t = \sum_{i=1}^n d_i^t w_i/n_t$ for observations with treatment t.

2.3 Treatment Effects Model

In this section we formally define the treatment effects model and the parameters of interest. Recall that $D \in \{0, 1, \ldots, \mathcal{T}\}$ indicates treatment status, $\{Y(t)\}_{t \in \overline{\mathbb{N}}_{\mathcal{T}}}$ are the (latent) potential outcomes, and $Y(t)$ is only observed for units with $D = t$; that is, $Y = \sum_{t \in \overline{\mathbb{N}}_\mathcal{T}} Y(t)$. The building blocks of many general estimands are the averages

$$\mu_t = \mathbb{E}[Y(t)], \qquad t \in \overline{\mathbb{N}}_\mathcal{T}, \tag{2.1}$$

and

$$\mu_{t,t'} = \mathbb{E}[Y(t)|D = t'], \qquad t, t' \in \overline{\mathbb{N}}_\mathcal{T} \times \overline{\mathbb{N}}_\mathcal{T}, \tag{2.2}$$

In the binary case, the average treatment effect is given by $\mu_1 - \mu_0$, whereas the treatment on the treated is $\mu_{1,1} - \mu_{0,1}$. Having a multivalued treatment allows for a much larger range of interesting estimands. To fix ideas, we keep as running examples two leading cases from the literature. First, the so-called dose-response function: the $(\mathcal{T} + 1)$-vector $\boldsymbol{\mu} = (\mu_0, \mu_1, \ldots, \mu_\mathcal{T})'$. Second, define $\boldsymbol{\tau}$ as the $\mathcal{T}$-vector with element t given by $\mu_{t,t} - \mu_{0,t}$. This gives the effect of each treatment relative to the baseline $t = 0$, only for those who received that treatment. These vectors are by no means the only interesting estimands constructed from μ_t and $\mu_{t,t'}$; many others are discussed by Lechner (2001), Heckman and Vytlacil (2007), and others.

The following two conditions are sufficient to identify μ_t and $\mu_{t,t'}$.

Assumption II.3 (Identification)**.** *For all $t \in \overline{\mathbb{N}}_\mathcal{T}$ and almost surely X, P_n obeys:*

(a) (Mean independence) $\mathbb{E}[Y(t)|D, X = x] = \mathbb{E}[Y(t)|X = x]$, and

(b) (Overlap) $\mathbb{P}[D = t|X = x]) \geq p_{\min} > 0$ for all $t \in \overline{\mathbb{N}}_\mathcal{T}$.

This assumption is a form of "ignorability" coined by Rosenbaum and Rubin (1983). This model allows arbitrary treatment effect heterogeneity in observables, but not unobservables. This assumption is standard in the program evaluation literature, and its plausibility has been discussed at length, so we omit a general discussion (see, e.g., Imbens (2004), Wooldridge (2010, Chapter 21), and references therein). However, in the context of model selection, two remarks on II.3(a) are warranted.

First, in place of Assumption II.3(a), it is more common to instead assume full conditional independence: $Y \perp\!\!\!\perp D|X$. However, as observed by Heckman, Ichimura, and Todd (1997), the weaker mean independence is sufficient. For our purposes, the "gap" between the two assumptions is important. Suppose full independence holds only conditional on a set of variables strictly larger than the variables entering the mean functions (e.g. the excess variables affect higher moments). In this case, because mean independence is still sufficient, we need not aim to select the larger set of covariates. Our results of course hold under full independence, which is important for the efficiency discussed in Section 2.5.3 below.

Second, the main drawback of Assumption II.3(a) is that it does not give identification of average effects on transformations of $Y(t)$. However, we are expressly interested in model selection on the mean function of the level of $Y(t)$, and hence Assumption II.3(a) is more natural. To operationalize model selection, structure must be placed on $\mathbb{E}[Y(t)|X=x]$, and hence functional form conditions tied to mean independence are not limiting per se. Indeed, if the parameter of interest is changed, for example to $\mathbb{E}[\log(Y(t))]$, and a sparsity assumption is made for $\mathbb{E}[\log(Y(t))|X=x]$, then our method applies.

Assumption II.3 yields identification of μ_t and $\mu_{t,t'}$ using either inverse weighting or regression, and double robustness follows from combining the two strategies. Recall the notation $p_t(x) = \mathbb{P}[D=t|X=x]$ and $\mu_t(x) = \mathbb{E}[Y|D=t, X=x]$. Applying Assumption II.3 we find that

$$\mathbb{E}\big[\psi_t\big(Y, D, \mu_t(X), p_t(X), \mu_t\big)\big] = \mathbb{E}\left[\frac{\mathbb{1}\{D=t\}Y}{p_t(X)} + \mu_t(X) - \frac{\mathbb{1}\{D=t\}\mu_t(X)}{p_t(X)} - \mu_t\right] = 0 \quad (2.3)$$

and

$$\mathbb{E}\big[\psi_{t,t'}\big(Y, D, \mu_t(X), p_t(X), p_{t'}(X), \mu_{t,t'}\big)\big] = \mathbb{E}\left[\frac{\mathbb{1}\{D=t'\}\mu_t(X)}{p_{t'}} + \frac{p_{t'}(X)}{p_{t'}}\frac{\mathbb{1}\{D=t\}(Y-\mu_t(X))}{p_t(X)} - \mu_{t,t'}\right] = 0, \quad (2.4)$$

where $p_t = \mathbb{P}[D=t]$. The moment condition (2.3) holds if either $p_t(x)$ or $\mu_t(x)$ is misspecified. For $\mu_{t,t'}$, if $\mu_t(x)$ is misspecified, both $p_t(X)$ and $p_{t'}(X)$ must be correctly specified, while if $\mu_t(x)$ is correct, both propensity scores may be misspecified. It is important to note that the forms of $\psi_t(\cdot)$ and $\psi_{t,t'}(\cdot)$ are fixed, so the function itself does not depend on the sample size even if its arguments do. Our estimator will be a plug-in version of this moment condition.

Remark 1 (Simplifications for $\mu_{t,t}$)**.** Identification $\mu_{t,t}$ does not require Assumption II.3. $Y(t)$ is fully observed for the sub-population of interest and so a simple average will deliver $\mu_{t,t} = \mathbb{E}[\mathbb{1}\{D=t\}Y]/p_t$. Note that (2.4) reduces to this when $t=t'$. For $\boldsymbol{\tau}$ this means we must only estimate the function $\mu_t(x_i)$ for $t=0$. Intuitively, we must use control group observations to proxy for treated units, but not the other way around.

Thus, for certain parameters of interest, Assumption II.3 can be weakened to hold only for the comparison group. However, we cover generic estimands, without necessarily specifying a control group, and so we maintain Assumption II.3 for simplicity, rather than keeping track of hosts of special cases. ■

Remark 2 (Efficient Influence Functions)**.** The functions $\psi_t(\cdot)$ and $\psi_{t,t'}(\cdot)$ are the efficient influence functions. Thus, our estimators have the interpretation of being plug-in versions of these influence functions. Indeed, as discussed Section 2.5.3, our estimators will be asymptotically linear with this influence function. ■

2.4 Sparse Models

We now formalize approximate sparsity. It is convenient to work with the linear log-odds ratio form of the multinomial model; the outcome model already being linear. Let X_Y^* and X_D^* be p-dimensional transformations of the covariates X, with $p > n$ allowed. These transformations are specific to the outcome and treatment models, but may overlap. They do not vary with t, nor depend on the DGP. Some examples are given below in Section 2.4.1. For the multinomial logistic model, we take $p_0(x) = 1 - \sum_{t \in \mathbb{N}_\mathcal{T}} p_t(x)$ and write

$$\log\left(\frac{p_t(x)}{p_0(x)}\right) = x_D^{*\prime}\gamma_t^* + B_t^D, \qquad t \in \mathbb{N}_\mathcal{T}. \tag{2.5}$$

Similarly, write the outcome regressions as

$$\mu_t(x) = x_Y^{*\prime}\beta_t^* + B_t^Y, \qquad t \in \overline{\mathbb{N}}_\mathcal{T}, \tag{2.6}$$

The terms $B_t^D = B_t^D(x)$ and $B_t^Y = B_t^Y(x)$ are bias terms arising from the parametric specification. As discussed below, these encompass the usual nonparametric bias as well. When it is clear from the context we often abbreviate both X_D^* and X_Y^* by X^* (and their realizations by x_i^*) and refer to them generically as "covariates". Much discussion applies to both. We assume $\mathbb{E}_n[(x_i^*)^2] = 1$ without loss of generality (see Remark 5).

Approximate sparsity requires that only a small number of the X^* are needed to make the bias small. Define $S_*^D = \bigcup_{\overline{\mathbb{N}}_\mathcal{T}} \text{supp}(\gamma_t^*)$ and $S_*^Y = \bigcup_{\mathbb{N}_\mathcal{T}} \text{supp}(\beta_t^*)$, so that these sets capture all variables important for treatment and outcomes, respectively. We assume that there is some $s_d < n$ such that for $|S_*^D| = s_d$, and similarly $|S_*^Y| = s_y < n$, and B_t^D and B_t^Y are sufficiently small. While a great deal of overlap is expected, in practice it is likely that a few covariates will be more or less important for different treatments, and so we do not require that the supports of $\gamma_t^*, t \in \mathbb{N}_\mathcal{T}$ or

$\beta^*_t, t \in \overline{\mathbb{N}}_{\mathcal{T}}$ are constant over t, nor that S^D_* overlaps with S^Y_*. Instead, it may be better to think of $\mathbb{N}_p \setminus S^D_*$ and $\mathbb{N}_p \setminus S^Y_*$ as the "common nonsupports" of the treatment and outcome equations. When there is no confusion, we will write s for either s_d or s_y.

2.4.1 Parametric and Nonparametric Examples

To concretize the sparse model idea, we now discuss how several models commonly used in practice fit into this framework. These include parametric and nonparametric models for $p_t(x)$ and $\mu_t(x)$, and hybrids of these. A common theme to all examples will be comparison to the *oracle* model: the model that knows the true support in advance. Our uniform inference results include all these examples as special cases because, loosely speaking, we obtain uniformity over DGPs where $p_t(x)$ and $\mu_t(x)$ have sparse representations. We aim for an accessible discussion of each model, and defer technicalities to the literature (Raskutti, Wainwright, and Yu 2010, Rudelson and Zhou 2011, Belloni, Chernozhukov, and Hansen 2013).

Example 1 (Oracle parametric model)**.** Assume models (2.5) and (2.6) hold with $B^D_t = B^Y_t = 0$ and $X^*_D = X^*_Y = X$. Let $p = s = \dim(X)$. All covariates are used in all modeling. If dimension is fixed this is the textbook parametric model, see for example Wooldridge (2010). Alternatively, the dimension can be diverging, but more slowly than n. We are not aware of any work which covers this case explicitly, though for the first stage, He and Shao (2000) cover linear and logistic regression, and their results easily extend to multinomial logistic models.[10]

The vast majority of treatment effect studies adopt this model (with dimension fixed), taking the set of covariates as given. In our framework, this is equivalent to the researcher having access to prior knowledge of which covariates are important and which are not. Such knowledge no doubt plays an important role, but it can not cover all situations or all variables in a data set. Furthermore, as more data become available, the researcher does not increase the complexity of their model. ■

Example 2 (Exactly sparse parametric model)**.** Retain the exact parametric structure of the prior example, but let $\dim(X) = p$ be possibly larger than n, and assume that S^Y_* and S^D_* are unknown sets of cardinality less than n. Model selection must be performed. Often, researchers (implicitly) rely on the *oracle property*, that S^Y_* and

[10] Even with diverging dimensions, the parametric multinomial logistic model relies on the independence of irrelevant alternatives.

S_*^D can be found with probability approaching one, and conduct inference conditioning on this event. This approach can never be made uniformly valid, and is known to have poor finite sample properties, as shown by Leeb and Pötscher. ■

Example 3 (Approximately sparse parametric model)**.** Again suppose a purely parametric model, so that $X_D^* = X_Y^* = X$ and $\dim(X) = p$, possibly greater than n. Suppose that there exist coefficients $\gamma_{\cdot,\cdot}^0$ and $\beta_{\cdot,\cdot}^0$ such that $\log[p_t(x)/p_0(x)] = x_D^*{}'\gamma_t^0$ and $\mu_t(x) = x'\beta_t^0$ exactly, but instead of any coefficients being precisely zero, suppose they may be ordered such that $|\gamma_{t,j}^0| \propto j^{-\alpha_\gamma}$ and $|\beta_{t,j}^0| \propto j^{-\alpha_\beta}$, with α_γ and α_γ at least 2. With this rapid decay, there exist s_d and s_y that are $o(n)$ such that Equations (2.5) and (2.6), and other conditions needed, are satisfied for $\gamma_{t,j}^* = \gamma_{t,j}^0$ for $j \leq s_d$ and $\beta_{t,j}^* = \beta_{t,j}^0$ for $j \leq s_y$ and the rest truncated to zero. That is S_*^D and S_*^Y collect the largest coefficients and $B_t^D = \sum_{\mathbb{N}_p \setminus S_*^D} x_j \gamma_{t,j}^0$, and similarly for B_t^Y. ■

Example 4 (Semiparametric model)**.** Assume $p_t(x)$ and $\mu_t(x)$ are unknown functions that can be well-approximated by a linear combination of s_d and s_y basis functions, respectively (e.g. are sufficiently smooth). In (2.5) and (2.6), $\gamma_{\cdot,\cdot}^*$ and $\beta_{\cdot,\cdot}^*$ are the coefficients of these approximations, while B_t^D and B_t^Y are the usual nonparametric biases. $X_D^* = R_D(X)$ and $X_Y^* = R_Y(X)$ are series terms used in the approximation. Standard semiparametric analyses, such as Hirano, Imbens, and Ridder (2003), Imbens, Newey, and Ridder (2007), or Cattaneo (2010), can be viewed in this context as oracle models that know in advance which terms yield the best approximation, typically assumed to be the first terms. Instead, we only require that some s_d (or s_y) of a set of p series terms give good approximations. This allows for greater flexibility in applications, where there is no knowledge of which series terms to use, and the researcher may want to mix terms from different bases. ■

Example 5 (Mixed parametric and semiparametric model)**.** Partition $X = (X_1, X_2)$. Suppose that the true log-odds function satisfies $\log[p_t(x)/p_0(x)] = x_1'\gamma_t^1 + h_t(x_2) + B_t^1(x)$, where $B_t^1(x)$ is a specification bias and $h_t(\cdot)$ is a smooth unknown function. For a set of basis functions $R_D(x_2)$, there will exist coefficients γ_t^2 such that $h_t(x_2) = R_D(x_2)'\gamma_t^2 + B_t^2(x_2)$ and so

$$\log\left(\frac{p_t(x)}{p_0(x)}\right) = x_D^*{}'\gamma_t^* + B_t^D, \quad x_D^* = (x_1', R_D(x_2)')', \quad \gamma_t^* = (\gamma_t^{1\prime}, \gamma_t^{2\prime})',$$

$$\text{and} \quad B_t^D = B_t^1 + B_t^2.$$

We require that some collection of variables and series terms give a good, sparse approximation, without placing explicit conditions on how many of either. Implicitly, one will restrict the other. For example, if the dimension of the parametric part is large, then we require that $h_t(\cdot)$ can be more easily approximated. We treat $\mu_t(x)$ the same. This example is closest to actual practice, where some variables (e.g. dummies) enter in a known way and should not be considered part of a nonparametric object, while other covariates must be considered flexibly. ■

2.4.2 Conceptual considerations in n-varying DGPs

We close this section with a discussion of how the DGP may vary with sample size. Much of the DGP, including parameters and distributions, is allowed to depend on n. Perhaps the most salient features that do not depend on n are the set of treatments and the functions ψ_t and $\psi_{t,t'}$. It is likely that our results can be extended to accommodate a growing number of treatments, but that is beyond the scope of our study. In the models (2.5) and (2.6), X^*, $\gamma^*_{\cdot,\cdot}$, and $\beta^*_{\cdot,\cdot}$ must depend on n by construction. Our results on estimation of these models are nonasymptotic: exact constants are provided that are defined for a fixed n. For treatment effect inference, we use triangular array asymptotics to retain the dependence on n of the DGP. The interpretation of the results does, and should, change depending on what is assumed about the DGP. To illustrate, let us return to Examples 2 and 4.

First, consider the simple parametric models of Example 2. We may now define $\mu_t = \mathbb{E}[\mathbb{E}[Y(t)|X]] = \mathbb{E}[X']\beta^*_t$, which depends on n by construction. That is, given an exact parametric specification for $\mathbb{E}[Y(t)|X]$ with a diverging number of covariates, the parameter to be estimated, μ_t, must depend on n. This may seem unnatural, as we typically think of the "true" parameters being features of a (large) fixed study population. However, with a diverging number of covariates, the idea of a fixed DGP is not clear. Indeed, if we estimate $\mu_t = \mu_t^{(n_1)}$ based upon n_1 observations, and then proceed to gather n_2 *more* observations, when we re-estimate our target is now $\mu_t^{(n_1+n_2)} \neq \mu_t^{(n_1)}$. One possible resolution is as follows. First, the parameter of interest is $\mu_t^{(\infty)} = \mathbb{E}[Y(t)]$, which is defined without reference to covariates. We can view each successive n-dependent μ_t as an approximation of $\mu_t^{(\infty)}$ based upon $p = p_n$ covariates. Note well that in our thought experiment, $p_{n_1} \neq p_{n_1+n_2}$, and so additional variables should have been collected for all $n_1 + n_2$ samples.

Contrast this with the semiparametric model in Example 4. It is common to assume the population DGP is fixed over n. The treatment effects may be constructed in terms of the underlying variables, e.g. $\mu_t^{(\infty)} = \mathbb{E}[Y(t)] = \mathbb{E}[\mathbb{E}[Y(t)|X]]$, with X^* serving only the purpose of aiding in approximating the regression functions. Model selection is performed on series terms, not underlying variables, to estimate the coefficients $\gamma^*_{\cdot,\cdot}$ and $\beta^*_{\cdot,\cdot}$. If $\mu_t = \mathbb{E}[X_Y^{*\prime}]\beta_t^* + \mathbb{E}[B_t^Y]$ does not depend on n, the bias term, by definition, exactly compensates for the n-dependence in $\mathbb{E}[X_Y^{*\prime}]\beta_t^*$. We emphasize that our inference results allow for general n-dependence in the DGP, and interpretation by the econometrician must take careful account of any conceptual assumptions.

2.5 Main Results on Treatment Effect Estimation and Inference

In this section we present results on uniformly valid treatment effect inference. We first present the estimators and conditions required for a generic model selector to yield uniform inference. We then give theoretical results, and close with a short discussion of efficiency.

2.5.1 Estimation Procedure with a Generic Model Selector

The moment functions $\psi_t(\cdot)$ and $\psi_{t,t'}(\cdot)$ of Equations (2.3) and (2.4) have fixed and known form, and so for (model selection based) estimators $\hat{p}_t(x)$ and $\hat{\mu}_t(x)$, we can define

$$\hat{\mu}_t = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{d_i^t(y_i - \hat{\mu}_t(x_i))}{\hat{p}_t(x_i)} + \hat{\mu}_t(x_i)\right\} \tag{2.7}$$

and

$$\hat{\mu}_{t,t'} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{d_i^{t'}\hat{\mu}_t(x_i)}{\hat{p}_{t'}} + \frac{\hat{p}_{t'}(x_i)}{\hat{p}_{t'}}\frac{d_i^t(y_i - \hat{\mu}_t(x_i))}{\hat{p}_t(x_i)}\right\}, \tag{2.8}$$

where $\hat{p}_t = n_t/n$. By combining these estimators appropriately we can construct estimators $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\tau}}$ for the dose-response function $\boldsymbol{\mu}$ and the vector $\boldsymbol{\tau}$, respectively, and any other estimand. Notice that when $t = t'$ $\hat{\mu}_{t,t}$ is an average over the appropriate subpopulation: $\hat{\mu}_{t,t} = \mathbb{E}_{n,t}[y_i]$.

Although in this section we allow for generic model selection based estimates $\hat{p}_t(x)$ and $\hat{\mu}_t(x)$, it is important to distinguish between estimates based upon selected sets that have no "additional randomness" and those that do. Model selection based estimation will naturally have two steps: first data-driven selection and then refitting to ameliorate the shrinkage bias and allow the researcher to augment the selected variables. Let $\tilde{S}^D$ and $\tilde{S}^Y$ be the selected sets and $\hat{S}^D$ and $\hat{S}^Y$ be the final sets of variables used in the refitting. We will say that these contain no "additional randomness" if the added variables (i.e. $\hat{S} \setminus \tilde{S}$, for Y or D) are nonrandomly selected, such as from economic theory or prior knowledge. On the other hand, the added variables may be selected from a random process beyond that included in $\tilde{S}$. The leading example would be using logistic-selected variables in the regressions or vice versa. Then the variables used in $\hat{\mu}_t(x_i)$ depend not only on the randomness of $\tilde{S}^Y$, but also on that of $\tilde{S}^D$, and hence on $\{d_i\}_{i=1}^n$. Stronger conditions are required for the estimators with additional randomness.

The choice of method is in part dependent on the assumptions of the underlying model. To illustrate, first, return to Example 2, where we have a purely parametric model with $X = X_D^* = X_Y^*$. The researcher may want to set $\hat{S}^D \supset \tilde{S}^D \cup \tilde{S}^Y$, in order to have a better chance that $S_*^Y \subset \hat{S}^D$. The set $\hat{S}^D$ now contains additional randomness due to $\tilde{S}^Y$. Conversely, consider Example 4. It is natural to include "low-order" basis functions for each underlying covariate, say linear and quadratic polynomials. Thus, the researcher may want to include these in $\hat{S}$, whether or not selected by the group lasso. However, there is no reason that the series terms useful for approximating the functions $\mu_t(x)$ would be useful for $p_t(x)$, or vice versa, and no additional randomness is injected.

We now state the sufficient conditions used for treatment effect estimation and inference. For exposition, we present these in three groups: those concerning the underlying DGP, requirements of $\hat{p}_t(x)$ and $\hat{\mu}_t(x)$ in the "no additional randomness" case, and finally the stronger conditions to allow for "additionally random" selected sets. Begin with conditions on the DGP. Let $U \equiv Y(t) - \mu_t(X)$ and impose the following condition.

Assumption II.4 (Data Generating Process). *For each n, the following are true for the DGP P_n.*

(a) (y_i, d_i, x_i) *is an i.i.d. sample from* (Y, D, X)*, where the data generating process obeys Equations* (2.5) *and* (2.6) *such that* $|S_*^Y| = s_d$ *and* $|S_*^D| = s_y$.

(b) The covariates X^* *have bounded support, with* $\max_{j \in \mathbb{N}_p} X_j^* \leq \mathcal{X} < \infty$*, uniformly in n. Transformations may depend on n but not the underlying data generating process.*

(c) $\mathbb{E}[|U|^4 \mid X] \leq \mathcal{U}^4$*, uniformly in n.*

(d) $\min_{j \in \mathbb{N}_p,\ t \in \bar{\mathbb{N}}_\mathcal{T}} \mathbb{E}[X_j^{*2} U^2] \wedge \mathbb{E}[X_j^{*2}(\mathbb{1}\{D = t\} - p_t(X))^2]$ *is bounded away from zero, uniformly in n.*

(e) For some $r > 0$: $\mathbb{E}[|\mu_t(x_i)\mu_{t'}(x_i)|^{1+r}]$ *and* $\mathbb{E}[|u_i|^{4+r}]$ *are bounded, uniformly in n.*

The conditions of Assumption II.4 are mild and intuitive. Assumption II.4(a) restricts attention to cross-sectional applications and codifies the requirement that the underlying functions have sparse representations. The condition of bounded covariates is unlikely to be a limitation in practice. Any X^* that are underlying variables will naturally be bounded in applications. This condition is automatically satisfied for most common choices of basis functions employed in nonparametric estimation. Finally, Assumptions II.4(c), II.4(d), and II.4(e) are weak moment conditions on the potential outcome models, including allowing the errors to be heteroskedastic and non-Gaussian. Excepting the support requirements of II.4(a) these conditions are not unique to high-dimensional models or model selection. Formalizing the requirement of uniform bounds in n is needed when doing array asymptotics.

We now give precisely the conditions on the model selector for uniformly valid inference.

Assumption II.5 (Model Selector Restrictions). *The model selection based estimators $\hat{p}_t(x)$ and $\hat{\mu}_t(x)$ obey the following for a sequence $\{P_n\}$, uniformly in $t \in \overline{\mathbb{N}}_{\mathcal{T}}$.*

(a) $\mathbb{E}_n[(\hat{p}_t(x_i) - p_t(x_i))^2] = o_{P_n}(1)$ *and* $\mathbb{E}_n\left[(\hat{\mu}_t(x_i) - \mu_t(x_i))^2\right] = o_{P_n}(1)$*,*

(b) $\mathbb{E}_{n,t}[(\hat{\mu}_t(x_i) - \mu_t(x_i))^2]^{1/2}\mathbb{E}_{n,t}[(\hat{p}_t(x_i) - p_t(x_i))^2]^{1/2} = o_{P_n}(n^{-1/2})$.

The first is a mild consistency requirement. The second is more interesting. It is analogous to the commonly-used, high-level requirement in semiparametrics that each first-step component converge at $n^{-1/4}$ at least.[11] Belloni, Chernozhukov, and Hansen (2013) use just such a condition. However, by making use of the doubly-robust property we have the weaker condition shown, involving the product. If one function is relatively easy to estimate Assumption II.5(b) can be satisfied even if the other does not converge at $n^{-1/4}$. In high-dimensional, sparse models the rates for the first stage depend on the sample size, the number of covariates considered, and the sparsity level. Thus, if one function requires fewer covariates to estimate, i.e. smaller p or s, then greater complexity can be allowed for in the other (capturing, in particular, their relative smoothness).

For our proposed group lasso selectors, recalling the results of Corollary II.2, Assumption II.5(a) will be satisfied if $(\sqrt{n^{-1}s_d \log(p \vee \underline{n})^{3/2+\delta_D}} + b_s^d\sqrt{s_d}) \to 0$ and $(\sqrt{n^{-1}s_y \log(p \vee \underline{n})^{3/2+\delta_Y}} + b_s^y) \to 0$. Further, II.5(b) is satisfied if their product is $o(n^{-1/2})$, which clearly shows how the relative sparsities and smoothnesses may interact.

[11] See, for example, Newey (1994a), Newey and McFadden (1994), and Chen (2007), and references therein.

When considering the additional-randomness estimators, we need a stronger bound on the regression errors U and more conditions on the first stage.

Assumption II.6 (Regularity conditions for union estimators)**.** *The model selection based estimators $p_t(x)$ and $\hat{\mu}_t(x)$ obey the following for a sequence $\{P_n\}$, uniformly $t \in \overline{\mathbb{N}}_{\mathcal{T}}$:*

$$\big(\max_{i \in \mathbb{I}_t} |u_i|\big) \left|\mathbb{E}_{n,t}[(\hat{p}_t(x_i) - p_t(x_i))^2]\right| = o_{P_n}(n^{-1/2})$$

and

$$\|\hat{\gamma}_t - \gamma_t^*\|_1 \vee \|\hat{\beta}_t - \beta_t^*\|_1 = o_{P_n}(\log(p)^{-1}).$$

These stronger conditions are needed because we must apply bounds for self-normalized sums (de la Peña, Lai, and Shao 2009). Belloni, Chen, Chernozhukov, and Hansen (2012) were the first to use these techniques in high-dimensional, sparse models. The first condition is a high-level condition that can be verified with conditions on the errors and a bound for estimation. For example, if we follow Belloni, Chen, Chernozhukov, and Hansen (2012) and assume that $\max_{i \in \mathbb{N}_n} |u_i| = O_{P_n}(n^{1/q})$ for some $q > 2$, then Assumption II.6 is met under our group lasso results if both $\big(n^{1/2+1/q}\left[n^{-1} s_d \log(p \vee \underline{n})^{3/2+\delta_D} + (b_s^d)^2 s_d\right]\big)$ and

$$\left(\log(p)\left[\sqrt{n^{-1} s_d^2 \log(p \vee \underline{n})^{3/2+\delta_D}} + b_s^d s_d \vee b_s^y\right]\right)$$

converge to zero. Note that as q increases, the stringency of the rate restriction decreases. For example, if the u_i are Gaussian, q can be taken to be any (large) positive number.

Remark 3 (Linear Probability Models)**.** Some authors advocate a linear probability model for the function $p_t(x)$, instead of the multinomial logistic form. Our results cover this case as well. Note that all we require are sufficiently high quality approximations to the underlying objects. If Assumptions II.5, and II.6 if appropriate,[12] are met then uniform inference is possible using a linear probability model. Our group lasso results (Theorems II.14 and II.15) can be used directly to verify these conditions. In the same vein, multinomial logistic regression can be used to estimate $\mu_t(x)$ if the outcome Y is discretely valued. ■

[12] Assumption II.6 can be slightly weakened in this case due to the linear link function.

2.5.2 Theoretical Results

We now come to our main results on inference on average treatment effects. Most of our discussion will concern μ_t and $\boldsymbol{\mu}$; similar points apply to results for $\mu_{t,t'}$ and $\boldsymbol{\tau}$. Our first result concerns consistency of our estimates under misspecification.

Theorem II.7 (Double Robustness)**.** *Consider a sequence $\{P_n\}$ of data-generating processes. Suppose that for some $p_t^0(x)$ and $\mu_t^0(x)$, $\mathbb{E}_n[(\hat{p}_t(x_i) - p_t^0(x_i))^2] = o_{P_n}(1)$ and $\mathbb{E}_n[(\hat{\mu}_t(x_i) - \mu_t^0(x_i))^2] = o_{P_n}(1)$. Let Assumptions II.3, and II.4 hold for each n, with the regularity conditions also holding for $p_t^0(x)$ and $\mu_t^0(x)$. If $p_t^0(x) = p_t(x)$ or $\mu_t^0(x) = \mu_t(x)$, then $|\hat{\mu}_t - \mu_t| = o_{P_n}(1)$.*

This theorem formalizes the double-robustness property of our estimators: the propensity score or regression may be misspecified if the limiting objects must be well-behaved. Compare to Assumption II.5(a). The nearly identical result for $\mu_{t,t'}$ is omitted to save space.

We now turn to our main inference results. First we demonstrate a Bahadur representation of a generic $\hat{\mu}_t$ or $\hat{\mu}_{t,t'}$. These are shown to be equivalent to a sample average of the moment functions $\psi_t(\cdot)$ and $\psi_{t,t'}(\cdot)$, respectively, after proper centering and scaling, evaluated at the true $p_t(x_i)$ and $\mu_t(x_i)$. Using these results, asymptotic normality can be obtained for general estimands. We state explicit results for the leading examples $\boldsymbol{\mu}$ and $\boldsymbol{\tau}$.

Before giving the results for $\boldsymbol{\mu}$, we need an asymptotic variance formula. Let the conditional variance of the potential outcomes be $\sigma_t^2(x) = \mathbb{E}[U^2|D = t, X = x]$. Define the $\overline{\mathcal{T}}$-square matrix $V_{\boldsymbol{\mu}}$ with elements

$$V_{\boldsymbol{\mu}}[t,t'] = \mathbb{1}\{t = t'\}\mathbb{E}\left[\frac{\sigma_t^2(X)}{p_t(X)}\right] + \mathbb{E}\left[(\mu_t(X) - \mu_t)(\mu_{t'}(X) - \mu_{t'})\right] \equiv V_{\boldsymbol{\mu}}^W(t) + V_{\boldsymbol{\mu}}^B(t,t').$$

Straightforward plug-in estimators for these two components are given by

$$\hat{V}_{\boldsymbol{\mu}}^W(t) = \mathbb{E}_n\left[\frac{d_i^t(y_i - \hat{\mu}_t(x_i))^2}{\hat{p}_t(x_i)^2}\right]$$

and

$$\hat{V}_{\boldsymbol{\mu}}^B(t,t') = \mathbb{E}_n\left[(\hat{\mu}_t(x_i) - \hat{\mu}_t)(\hat{\mu}_{t'}(x_i) - \hat{\mu}_{t'})\right].$$

Our first result gives the asymptotic behavior of $\hat{\mu}_t$ and $\hat{\boldsymbol{\mu}}$ for a sequence of DGPs.

Theorem II.8 (Estimation of Average Treatment Effects)**.** *Consider a sequence $\{P_n\}$ of data-generating processes that obey Assumptions II.3, II.4, and II.5 for each n. If*

$\hat{\mu}_t(x_i)$ and $\hat{p}_t(x_i)$ do not have additional randomness in the estimated supports, we have:

1. $\sqrt{n}(\hat{\mu}_t - \mu_t) = \sum_{i=1}^n \psi_t(y_i, d_i^t, \mu_t(x_i), p_t(x_i), \mu_t)/\sqrt{n} + o_{P_n}(1)$*;*
2. $V_{\boldsymbol{\mu}}^{-1/2}\sqrt{n}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \to_d \mathcal{N}(0, I_{\overline{\mathcal{T}}})$*; and*
3. $\hat{V}_{\boldsymbol{\mu}}^W(t) - V_{\boldsymbol{\mu}}^W(t) = o_{P_n}(1)$ *and* $\hat{V}_{\boldsymbol{\mu}}^B(t,t') - V_{\boldsymbol{\mu}}^B(t,t') = o_{P_n}(1)$*.*

If, in addition, Assumption II.6 holds, then the same is true when the supports contain additional randomness.

Theorem II.8 itself may appear standard, but what is nonstandard is that the model selection step of the estimation has been explicitly accounted for. This immediately gives the following uniform inference results.

Corollary II.9 (Uniformly Valid Inference)**.** *Let $\boldsymbol{P}_n$ be the set of data-generating processes satisfying the conditions of Theorem II.8 for a given n. For a fixed, twice uniformly continuously differentiable function $G : \mathbb{R}^{\overline{\mathcal{T}}} \to \mathbb{R}$ with gradient ∇_G such that* $\liminf_{n\to\infty} \|\nabla_G(\boldsymbol{\mu})\|_2$ *is bounded away from zero, we have:*

$$\sup_{P \in \boldsymbol{P}_n} \left| \mathbb{P}_P \left[G(\boldsymbol{\mu}) \in \left\{ G(\hat{\boldsymbol{\mu}}) \pm c_\alpha \sqrt{\nabla_G(\hat{\boldsymbol{\mu}})' \hat{V}_{\boldsymbol{\mu}} \nabla_G(\hat{\boldsymbol{\mu}})/n} \right\} \right] - (1-\alpha) \right| \to 0,$$

where $c_\alpha = \Phi^{-1}(1-\alpha/2)$.

Corollary II.9 shows that these procedures are uniformly valid over the class of DGPs we consider, and hence will be reliable in applications. This method of proving uniformity follows Belloni, Chernozhukov, and Hansen (2013) and Romano (2004), and is distinct from the approach of Andrews and Guggenberger (2009). By not relying on an oracle property, we avoid the uniformity problems demonstrated by Leeb and Pötscher, as discussed before.

Our results for the treatment effects on the treated, $\mu_{t,t'}$, are conceptually similar. The variance formula for $\boldsymbol{\tau}$ is slightly more cumbersome notationally. Define the $\mathcal{T}$-square matrix $V_{\boldsymbol{\tau}}$ with elements

$$\begin{aligned} V_{\boldsymbol{\tau}}[t,t'] &= \mathbb{1}\{t=t'\}\mathbb{E}\left[\frac{p_t(X)}{p_t^2}\left[\sigma_t^2(X) + (\mu_t(X) - \mu_0(X) - \mu_{t,t} + \mu_{0,t})^2\right]\right] \\ &\qquad + \mathbb{E}\left[\frac{p_t(X)p_{t'}(X)}{p_t p_{t'} p_0(X)}\sigma_0^2(X)\right] \\ &\equiv V_{\boldsymbol{\tau}}^W(t) + V_{\boldsymbol{\tau}}^B(t,t'). \end{aligned}$$

Straightforward plug-in estimators for these two components are given by

$$\hat{V}_{\boldsymbol{\tau}}^{W}(t) = \mathbb{E}_n \left[\frac{d_i^t}{\hat{p}_t^2} \left[(y_i - \hat{\mu}_0(x_i) - \hat{\mu}_{t,t} + \hat{\mu}_{0,t})^2 \right] \right]$$

and

$$\hat{V}_{\boldsymbol{\tau}}^{B}(t,t') = \mathbb{E}_n \left[\frac{\hat{p}_t(x_i)\hat{p}_{t'}(x_i)}{\hat{p}_t \hat{p}_{t'} \hat{p}_0(x_i)^2} d_i^0 (y_i - \hat{\mu}_0(x_i))^2 \right].$$

Note that we needn't estimate $\mu_t(x)$ and $\sigma_t^2(x)$, again due to the simplification discussed in Remark 1. With this notation, we have the following results.

Theorem II.10 (Estimation of Treatment Effects on Treated Groups)**.** *Consider a sequence $\{P_n\}$ of data-generating processes that obey Assumptions II.3, II.4, and II.5 for each n. Then under P_n, as $n \to \infty$, if $\hat{\mu}_t(x_i)$ and $\hat{p}_t(x_i)$ do not have additional randomness in the estimated supports:*

1. $\sqrt{n}(\hat{\mu}_{t,t'} - \mu_{t,t'}) = \sum_{i=1}^{n} \psi_{t,t'}(y_i, d_i^t, \mu_t(x_i), p_t(x_i), p_{t'}(x_i), \mu_{t,t'})/\sqrt{n} + o_{P_n}(1)$*;*
2. $V_{\boldsymbol{\tau}}^{-1/2}\sqrt{n}(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}) \to_d \mathcal{N}(0, I_{\mathcal{T}})$*; and*
3. $\hat{V}_{\boldsymbol{\tau}}^{W}(t) - V_{\boldsymbol{\tau}}^{W}(t) = o_{P_n}(1)$ *and* $\hat{V}_{\boldsymbol{\tau}}^{B}(t,t') - V_{\boldsymbol{\tau}}^{B}(t,t') = o_{P_n}(1)$*.*

If, in addition, Assumption II.6 holds, then the same is true when the supports contain additional randomness.

Corollary II.11 (Uniformly Valid Inference)**.** *Let $\boldsymbol{P}_n$ be the set of data-generating processes satisfying the conditions of Theorem II.10 for a given n. For a fixed, twice uniformly continuously differentiable function $G : \mathbb{R}^{\mathcal{T}} \to \mathbb{R}$ with gradient ∇_G such that* $\liminf_{n\to\infty} \|\nabla_G(\boldsymbol{\tau})\|_2$ *is bounded away from zero, we have:*

$$\sup_{P \in \boldsymbol{P}_n} \left| \mathbb{P}_P \left[G(\boldsymbol{\tau}) \in \left\{ G(\hat{\boldsymbol{\tau}}) \pm c_\alpha \sqrt{\nabla_G(\hat{\boldsymbol{\tau}})' \hat{V}_{\boldsymbol{\tau}} \nabla_G(\hat{\boldsymbol{\tau}})/n} \right\} \right] - (1-\alpha) \right| \to 0,$$

where $c_\alpha = \Phi^{-1}(1-\alpha/2)$.

2.5.3 Efficiency Considerations

The prior theoretical results are aimed at delivering robust inference. In this section, we briefly discuss the efficiency of our estimator. We consider two efficiency criteria: semiparametric efficiency and oracle efficiency. The former deals with the variance of the final estimator, whereas the latter is directly about the efficacy of the model selection. To put each criterion on sound conceptual footing, we separate discussion and restrict each to the most appropriate set of models.

For semiparametric efficiency, assume that $p_t(x)$ and $\mu_t(x)$ are nonparametric objects, as in Example 4. Recall that X are fixed-dimension variables and the DGP does not vary with n. If we "upgrade" the mean independence of Assumption II.3(a) to full independence, namely $\{Y(t)\}_{\overline{\mathbb{N}}_\mathcal{T}} \perp\!\!\!\perp D|X$, then Theorems II.8 and Theorem II.10 immediately yield asymptotically linearity and semiparametric efficiency, attaining Hahn's (1998) and Cattaneo's (2010) bounds.

Let us turn to oracle efficiency. An alternative to our approach is to prove that the true support can be found with probability approaching one (the oracle property), then conduct inference conditioning on this event. This approach cannot be made uniformly valid, but may be of interest in the causal setting when restricted to exactly sparse models (there is no "true" support in approximately sparse models), because discovering the true support is equivalent to finding the variables in the causal mechanism (White and Lu 2011). This may be interesting in its own right, or for future applications by way of hypothesis generation. Further, efficiency can be improved because only variables appearing in $\mu_t(x_i) = \mathbb{E}[Y|D = t, x_i]$ should be used, hence $S_*^D \setminus S_*^Y$ are not needed and $S_*^Y \setminus S_*^D$ can be ignored for propensity score estimation.

Perfect selection requires two strong conditions: (i) an orthogonality condition on the Gram matrixes that restricts the correlation between the variables in and out of the true support (Zhao and Yu 2006, Bach 2008), and (ii) a *beta-min* condition bounding the nonzero coefficients away from zero. Intuitively, highly correlated variables can not be distinguished, nor can coefficients sufficiently close to zero be found with certainty. Both bounds may depend on n. Under such conditions, it is possible to show that S_*^Y and S_*^D can be found with probability approaching one.

2.6 Group Lasso Selection and Estimation

We now give details for group lasso model selection and estimation. This section is quite technical. Our main theorems are given in Section 2.6.3. To set up these results, we first make precise how selection and refitting are implemented. Section 2.6.1 develops our (apparently) novel penalty choice for multinomial logistic regression. Restricted and sparse eigenvalues, key quantities in our bounds, are discussed in Section 2.6.2. Discussion will be model-specific so we use the general notation X^* and s.

We first select covariates by applying the group lasso penalty to the multinomial logistic loss (for the propensity scores) and to least squares loss (to estimate the outcome regression). The loss functions are defined as

$$\mathcal{M}(\gamma_{\cdot,\cdot}) = \sum_{t \in \mathbb{N}_\mathcal{T}} \mathbb{E}_n \left[-d_i^t \log \left(\hat{p}_t(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}}) \right) \right] \qquad \text{and} \qquad \mathcal{E}(\beta_{\cdot,\cdot}) = \sum_{t \in \overline{\mathbb{N}}_\mathcal{T}} \mathbb{E}_{n,t}[(y_i - x_i^{*\prime}\beta_t)^2],$$

where we denote the multinomial logit function as

$$\hat{p}_t(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}}) = \exp(x_i^{*\prime}\gamma_t) \Big/ \left(1 + \sum_{t \in \mathbb{N}_\mathcal{T}} \exp(x_i^{*\prime}\gamma_t) \right)$$

. Then, the group lasso estimates for the propensity score coefficients, denoted $\tilde{\gamma}_{\cdot,\cdot}$, solve

$$\tilde{\gamma}_{\cdot,\cdot} = \operatorname*{arg\,min}_{\gamma_{\cdot,\cdot} \in \mathbb{R}^{p\mathcal{T}}} \left\{ \mathcal{M}(\gamma_{\cdot,\cdot}) + \lambda_D |||\gamma_{\cdot,\cdot}|||_{2,1} \right\}, \tag{2.9}$$

where λ_D is a penalty parameter discussed in detail below and $|||\gamma_{\cdot,\cdot}|||_{2,1}$ is the mixed ℓ_2/ℓ_1 norm defined above. Similarly, the regression estimates solve

$$\tilde{\beta}_{\cdot,\cdot} = \operatorname*{arg\,min}_{\beta_{\cdot,\cdot} \in \mathbb{R}^{p\overline{\mathcal{T}}}} \left\{ \mathcal{E}(\beta_{\cdot,\cdot}) + \lambda_Y |||\beta_{\cdot,\cdot}|||_{2,1} \right\}. \tag{2.10}$$

To ameliorate the downward bias induced by the penalty and to allow for researcher-added variables, we refit unpenalized models.[13] Let $\tilde{S}^D = \{j : \|\tilde{\gamma}_{\cdot,j}\|_2 > 0\}$ and $\tilde{S}^Y = \{j : \|\tilde{\beta}_{\cdot,j}\|_2 > 0\}$ be the selected covariates and $\hat{S}^D$ and $\hat{S}^Y$ those used in refitting.[14] We require $\hat{S} \supset \tilde{S}$ and $|\hat{S}| \leq s$ for D and Y (we will prove that $|\tilde{S}| \leq s$ in both cases). The refitting estimators solve

$$\hat{\gamma}_{\cdot,\cdot} = \operatorname*{arg\,min}_{\gamma_{\cdot,\cdot},\ \mathrm{supp}(\gamma_t) = \hat{S}^D} \{\mathcal{M}(\gamma_{\cdot,\cdot})\} \tag{2.11}$$

and

$$\hat{\beta}_{\cdot,\cdot} = \operatorname*{arg\,min}_{\beta_{\cdot,\cdot},\ \mathrm{supp}(\beta_t) = \hat{S}^Y} \{\mathcal{E}(\beta_{\cdot,\cdot})\}. \tag{2.12}$$

[13] The bias is away from the pseudo-true coefficients of the sparse parametric representation, $\gamma^*_{\cdot,\cdot}$ and $\beta^*_{\cdot,\cdot}$. There is no relation to specification biases B_t^D and B_t^Y.

[14] When $\mathrm{supp}(\gamma_t^*)$ and $\mathrm{supp}(\beta_t^*)$ will not vary much over t, the group lasso is known to have better properties than the ordinary lasso in terms of selection and convergence. Obozinski, Wainwright, and Jordan (2011) give a sharp bound on the overlap necessary to yield improvements, while Huang and Zhang (2010), Kolar, Lafferty, and Wasserman (2011), and Lounici, Pontil, van de Geer, and Tsybakov (2011) also demonstrate advantages of the group lasso approach. These works show, among other things, that the group lasso advantage increases as $\mathcal{T}$ increases, and with the group structure, may perform better with smaller samples. We defer to the works cited for a formal discussion.

2.6.1 Choice of Penalty

We now turn to choice of the penalty parameters λ_D and λ_Y. These must be chosen so that, with high probability, the penalty dominates the noise, which is captured by the magnitude of the score in the dual of the $|||\cdot|||_{2,1}$ norm.

For linear regression, we set

$$\lambda_Y = \frac{4\mathcal{X}\mathcal{U}\sqrt{\mathcal{T}}}{\sqrt{\underline{n}}}\left(1 + \frac{\log(p \vee \underline{n})^{3/2+\delta_Y}}{\sqrt{\mathcal{T}}}\right)^{1/2}, \tag{2.13}$$

for some $\delta_Y > 0$, so that

$$\lambda_Y > 4 \max_{j \in \mathbb{N}_p} \|\mathbb{E}_{n,t}[u_i x_{i,j}^*]\|_2,$$

with probability $1 - \mathcal{P}$ for small (and shrinking) $\mathcal{P}$, following Lounici, Pontil, van de Geer, and Tsybakov (2011). This penalty is of the form $\lambda_Y \propto \Lambda(1 + r_n)$, where Λ is an upper bound on the true score. The rate r_n balances the rate of convergence against the concentration effect: larger r_n slows the rate of convergence, but makes the probability of concentration of the group lasso estimate higher, by shrinking $\mathcal{P}$.

For the multinomial logistic regression, we instead find λ_D such that

$$\lambda_D > 2 \max_{j \in \mathbb{N}_p} \|\mathbb{E}_n[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}) - d_i^t)x_{i,j}^*]\|_2$$

with probability $1 - \mathcal{P}$. Note that $\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})$ appears instead of $p_t(x_i)$,[15] which implies that the bias and noise are simultaneously dominated. To achieve this, we set

$$\lambda_D = 2\mathcal{X}\sqrt{\mathcal{T}}\left[b_s^d + \frac{1}{\sqrt{\underline{n}}}\left(1 + \frac{\log(p \vee \underline{n})^{3/2+\delta_D}}{\sqrt{\mathcal{T}}}\right)^{1/2}\right], \tag{2.14}$$

for some $\delta_D > 0$. The form of λ_D is $\Gamma + \Lambda(1 + r_n)^{1/2}$, where the added Γ bounds the bias contribution. To the best of our knowledge, choosing the penalty in this way to handle an approximately sparse, nonlinear model is new in the high-dimensional, sparse literature, and may be useful in future research.

In the Appendix we show that, for $\delta = \delta_Y$ or δ_D, the concentration probability is given by

$$\mathcal{P} = \frac{4\sqrt{\log(2p)(1 + 64\log(12p)^2)}}{\log(p \vee \underline{n})^{3/2+\delta}}, \tag{2.15}$$

Remark 4. For given δ_Y and δ_D, the only unknown quantities in λ_Y and λ_D are $\mathcal{X}$, $\mathcal{U}$, and b_s^d. In practice, we set $b_s^d = 0$ for two reasons: first, bias estimation may be

[15] The multiple of 2 instead of 4 is a related technicality.

difficult; and second, the only consequence of a smaller penalty is (perhaps) a slight reduction in efficiency. We use $\max_{i \leq n} \max_{j \in \mathbb{N}_p} |x^*_{i,j}|$ to estimate $\mathcal{X}$, after scaling (see Remark 5). We estimate $\mathcal{U}$ by iteration. Given an initial estimate $\hat{\mu}_t^{(0)}(x)$, $\hat{\mathcal{U}}^{(k)} = \mathbb{E}_n[(y_i - \hat{\mu}_t^{(k-1)}(x_i))^4]^{1/4}$, where $\hat{\mu}_t^{(k)}(x_i)$, $k > 0$, is based on Eqn. (2.12). The initial estimate can be least squares on a few variables, a regularized method tuned by cross validation, or other options. ■

Remark 5 (Weighted Penalties)**.** Two final remarks are in order regarding weighting the group lasso penalty. First, one may weight the ℓ_2 portion of the penalty, as in

$$\lambda_D \sum_{j \in \mathbb{N}_p} \| \boldsymbol{X}_j \gamma_{\cdot,j} \|_2,$$

where $\boldsymbol{X}_j$ is the design matrix for covariate j, across all the treatments. (Other weight matrixes are possible.) With this choice, the estimate is invariant to within group (treatment) reparameterizations, and is thus scale invariant for each covariate. We therefore assume throughout that $\mathbb{E}_n[(x_i^*)^2] = 1$ without loss of generality.

Second, the ℓ_1 norm can be weighted to give a penalty of the form

$$\lambda_D \sum_{j \in \mathbb{N}_p} w_j \| \gamma_{\cdot,j} \|_2$$

. Two common choices for w_j are the number of variables in group j or an adaptive penalty from a pilot estimate. Our groups are equally sized, and although adaptive procedures may improve oracle properties (Zou 2006, Wei and Huang 2010), our goal is not perfect selection. ■

2.6.2 Restricted Eigenvalues

The local behavior of optimizations (2.9), (2.10), (2.11), and (2.12) is captured by their respective Hessians, which involve the second moment matrix of the covariates. The eigenvalues of such matrixes will be explicit in our bounds. We are interested in finite sample bounds, and so we will only discuss the empirical Gram matrixes (see Remark 6). Define

$$Q = \mathbb{E}_n[x_i^* x_i^{*\prime}] \qquad \text{and} \qquad Q_t = \mathbb{E}_{n,t}[x_i^* x_i^{*\prime}]. \tag{2.16}$$

In high-dimensional data, both are singular, and so we use restricted eigenvalues and sparse eigenvalues (Bickel, Ritov, and Tsybakov 2009).

For the multinomial logistic regression, the minimal restricted eigenvalue is defined by

$$\kappa_D^2 \leq \min_{\delta} \left\{ \frac{\sum_{t \in \mathbb{N}_{\mathcal{T}}} \delta_t' Q \delta_t}{\|\delta_{\cdot, S_D^*}\|_2^2} : \delta \in \mathbb{R}^{p\mathcal{T}} \setminus \{0\}, \left|\left|\left| \delta_{\cdot, \{S_*^D\}^c} \right|\right|\right|_{2,1} \leq 3 \left|\left|\left| \delta_{\cdot, S_*^D} \right|\right|\right|_{2,1} \right\}. \qquad (2.17)$$

For least squares estimation we instead use

$$\kappa_Y^2 \leq \min_{\delta} \left\{ \frac{\sum_{t \in \overline{\mathbb{N}}_{\mathcal{T}}} \delta_t' Q_t \delta_t}{\|\delta_{\cdot, S_Y^*}\|_2^2} : \delta \in \mathbb{R}^{p\overline{\mathcal{T}}} \setminus \{0\}, \left|\left|\left| \delta_{\cdot, \{S_*^Y\}^c} \right|\right|\right|_{2,1} \leq 3 \left|\left|\left| \delta_{\cdot, S_*^Y} \right|\right|\right|_{2,1} \right\}. \qquad (2.18)$$

The only difference is that Q appears for κ_D, whereas Q_t are used in κ_Y. The restricted set, or cone constraint, requires the magnitude of $\delta_{\cdot,\cdot}$ off the true support be small relative to the true support, measured in the group lasso norm. We will show that $(\tilde{\gamma}_{\cdot,\cdot} - \gamma^*_{\cdot,\cdot})$ and $(\tilde{\beta}_{\cdot,\cdot} - \beta^*_{\cdot,\cdot})$ obey the respective constraints.

In contrast, the refitting errors $(\hat{\gamma}_{\cdot,\cdot} - \gamma^*_{\cdot,\cdot})$ and $(\hat{\beta}_{\cdot,\cdot} - \beta^*_{\cdot,\cdot})$ (from (2.11) and (2.12)) may not obey the cone constraint, but are known to be sparse. This motivates the use of sparse eigenvalues. For a set $S \subset \mathbb{N}_p$ and a $p \times p$ matrix $\tilde{Q}$, define

$$\underline{\phi}\{\tilde{Q}, S\}^2 = \min_{\delta \in \mathbb{R}^p, \, \text{supp}(\delta) = S} \frac{\delta' \tilde{Q} \delta}{\|\delta\|_2^2} \qquad \text{and} \qquad \overline{\phi}\{\tilde{Q}, S\}^2 = \max_{\delta \in \mathbb{R}^p, \, \text{supp}(\delta) = S} \frac{\delta' \tilde{Q} \delta}{\|\delta\|_2^2}. \qquad (2.19)$$

Finally, it will be useful to define a bound on $\overline{\phi}\{\tilde{Q}, S\}$ over all subsets of a certain size. To this end, for any integer m, define $\overline{\overline{\phi}}(\tilde{Q}, m) = \max_{S \subset \mathbb{N}_p, \, |S| \leq m} \overline{\phi}\{\tilde{Q}, S\}$.

We take these quantities to be primitive, and defer discussion to the literature. For example, see van de Geer and Buhlmann (2009), Huang and Zhang (2010), Raskutti, Wainwright, and Yu (2010), Rudelson and Zhou (2011), and Belloni, Chernozhukov, and Hansen (2013). In particular, Huang and Zhang (2010) show that the group lasso may need fewer observations to satisfy conditions on sparse eigenvalues.

Remark 6. Often, invertibility of Q and Q_t relies on their convergence to nonsingular population counterparts.[16] Some of the papers cited above verify conditions on the restricted and sparse eigenvalues by just this approach. Our theorems can be restated in this way by conditioning on the event that Q and Q_t are close to their counterparts in the appropriate sense, and adjusting the probability with which the conclusions hold. We instead take bounds to be infinite if the minimum eigenvalues are zero. ■

[16] This is standard in fixed-dimension models, and has been used for diverging-dimensions parametric models (He and Shao 2000) and nonparametrics (Newey 1997, Huang 2003, Belloni, Chen, Chernozhukov, and Kato 2012, Cattaneo and Farrell 2013, Chen and Christensen 2013).

2.6.3 Theoretical Results

We now have the necessary notation and assumptions to state our theoretical results on group lasso estimation, beginning with multinomial logistic regression, followed by a terse treatment of linear models. Corollary II.2 is a special case of the results in this section, see Remarks 7 and 8.

Our first result is a nonasymptotic bound on the group lasso estimates from (2.9).

Theorem II.12 (Group Lasso Estimation of Multinomial Logistic Models)**.** *Suppose Assumptions II.3(b), II.4(a), II.4(b), and II.4(c) hold and that* $\max_{i \leq n} b^d_{t,i} \leq b^d_s$*. Define* $A_p = 0 \vee (p_{\min}/(p_{\min} - b^d_s))$ *and*

$$R_{\mathcal{M}} = \left(\frac{A_p}{p_{\min}}\right)^{\bar{\mathcal{T}}} \frac{3\mathcal{T} A_K \lambda_D \sqrt{s}}{\kappa_D}, \qquad \textit{for} \qquad A_K > 2\frac{\kappa_D^2}{\kappa_D^2 - 8\mathcal{X}\sqrt{\mathcal{T}}\lambda_D s}.$$

Then with probability $1 - \mathcal{P}$

$$\max_{t \in \mathbb{N}_{\mathcal{T}}} \mathbb{E}_n[(\hat{p}_t(\{x_i^{*\prime}\tilde{\gamma}_t\}_{\mathbb{N}_{\mathcal{T}}}) - p_t(x_i))^2]^{1/2} \leq R_{\mathcal{M}} + b^d_s,$$

$$\max_{t \in \mathbb{N}_{\mathcal{T}}} \|\tilde{\gamma}_t - \gamma_t^*\|_1 \leq \left(\frac{|\tilde{S}^D \cup S_D^*|}{\underline{\phi}\{Q, \tilde{S}^D \cup S_D^*\}}\right)^{1/2} R_{\mathcal{M}},$$

and

$$|\tilde{S}^D| \leq 8 s L_n \left\{\min_{m \in \mathbb{N}_Q^D} \overline{\overline{\phi}}(Q, m)\right\},$$

where

$$\mathbb{N}_Q^D = \left\{m \in \{1, 2, \ldots n\} : m > 8 s L_n \overline{\overline{\phi}}(Q, m)\right\}, \qquad \textit{and} \qquad L_n = \left(\frac{R_{\mathcal{M}}}{\lambda_D \sqrt{s}}\right)^2.$$

This theorem is new to the literature, to the best of our knowledge. Much of the detail involves capturing the finite sample behavior of the Hessian and Gram matrixes. We discuss the features of this result in the following remarks.

- The Hessian of $\mathcal{M}(\gamma_{\cdot,\cdot})$ is $\mathbb{E}_n[\mathcal{H}_i \otimes x_i^* x_i^{*\prime}]$ for a $\mathcal{T}$-square matrix $\mathcal{H}_i$ that depends the coefficients and x_i^* through the estimated probabilities $\hat{p}_t(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_{\mathcal{T}}})$. The error $R_{\mathcal{M}}$ depends on how well-controlled is this matrix. The factors $p_{\min}$, A_p, and A_K capture the behavior of $\mathcal{H}_i$ and κ_D^{-1} accounts for the rest. Under overlap, the true probabilities are bounded above $p_{\min}$, and hence $p_{\min}^{-\bar{\mathcal{T}}}$ captures the nonsingularity of the population version of $\mathcal{H}_i$. To get to this point requires two steps. First, the sparse parametric representations $\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})$ must also be bounded away from zero, leading to the factor of A_p. This is essentially a bias condition, which in the asymptotic case holds trivially: A_p may be chosen arbitrarily close to one as $b_s^d \to 0$. Second, A_K controls the neighborhood in which $\hat{p}_t(\{x_i^{*\prime}\tilde{\gamma}_t\}_{\mathbb{N}_\mathcal{T}})$ is also bounded away from zero. Intuitively (and asymptotically), the estimate will be in a small (shrinking) neighborhood of the $\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})$. In asymptotics A_K may be chosen arbitrarily close to 2, which stems from the factor of 1/2 in a quadratic expansion of $\mathcal{M}(\cdot)$. A lower bound on A_K is required in finite samples to ensure that $\hat{p}_t(\{x_i^{*\prime}\tilde{\gamma}_t\}_{\mathbb{N}_\mathcal{T}})$ is positive, and hence the two-term expansion is valid. This is analogous to Belloni and Chernozhukov's (2011a) "restricted nonlinear impact coefficient" approach, with a central difference that A_K is captured in our bound directly.

- The maximal sparse eigenvalues are crucial to the bound on $|\tilde{S}^D|$. In many prior results, the latter is bounded using the largest eigenvalue of Q itself, i.e. $\overline{\overline{\phi}}(Q, n)$. Adapting the technique of Belloni and Chernozhukov (2011b) to the present case, we are able to find a tighter bound, which yields sparsity proportional to s under weaker conditions. This is crucial for refitting.

- For the linear model the constants in the group lasso bounds can offset the (logarithmic) suboptimality in rate (Huang and Zhang 2010, Lounici, Pontil, van de Geer, and Tsybakov 2011), and this may be true here as well.

The error bounds for post-selection estimation are more complex and depends in part on the good properties of the initial group lasso fit. The following theorem gives our results.

Theorem II.13 (Post-Selection Multinomial Logistic Regression)**.** *Suppose the conditions of Theorem II.12 hold. For*

$$A_K > 2\left\{\frac{\underline{\phi}\{Q, \hat{S}_D \cup S_D^*\}^2}{\underline{\phi}\{Q, \hat{S}_D \cup S_D^*\}^2 - \mathcal{X}\sqrt{\mathcal{T}}\lambda_D|\hat{S}_D \cup S_D^*|}\right\}$$
$$\vee \left\{\frac{\underline{\phi}\{Q, \hat{S}_D \cup S_D^*\}}{\underline{\phi}\{Q, \hat{S}_D \cup S_D^*\} - 2R_\mathcal{M}\mathcal{X}\sqrt{\mathcal{T}}\sqrt{|\hat{S}_D \cup S_D^*|}}\right\},$$

define

$$R'_\mathcal{M} = \left(\frac{A_p}{p_{\min}}\right)^{\overline{\mathcal{T}}} \frac{\mathcal{T} A_K \lambda_D \sqrt{|\hat{S}_D \cup S_D^*|}}{2\underline{\phi}\{Q, \hat{S}_D \cup S_D^*\}}$$

and

$$R''_{\mathcal{M}} = \{R_{\mathcal{M}}\} \vee \left\{ R'_{\mathcal{M}} + \left[R'_{\mathcal{M}} R_{\mathcal{M}} + \left(\frac{A_p}{p_{\min}} \right)^{\overline{\mathcal{T}}} \mathcal{T} A_K R^2_{\mathcal{M}} \right]^{1/2} \right\}.$$

Then with probability $1 - \mathcal{P}$,

$$\max_{t \in \mathbb{N}_{\mathcal{T}}} \mathbb{E}_n[(\hat{p}_t(\{x_i^{*\prime}\hat{\gamma}_t\}_{\mathbb{N}_{\mathcal{T}}}) - p_t(x_i))^2]^{1/2} \leq R''_{\mathcal{M}} + b^d_s,$$

and

$$\max_{t \in \mathbb{N}_{\mathcal{T}}} \|\hat{\gamma}_t - \gamma^*_t\|_1 \leq \left(\frac{|\tilde{S}^D \cup S^*_D|}{\underline{\phi}\{Q, \tilde{S}^D \cup S^*_D\}} \right)^{1/2} R''_{\mathcal{M}}.$$

This result is the first study of post-selection estimation in approximately sparse logistic models, for any $\mathcal{T} \geq 1$. We explicitly capture the dependence on the loss function $\mathcal{M}(\gamma_{\cdot,\cdot})$ and the impact of the initial group lasso fit. It is not readily discernible if these bounds improve upon the group lasso estimates. This in part depends on the DGP and the selection success of the initial fit. It would be interesting to have an explicit characterization of the improvements offered by refitting. In this result, further lower bounds on A_K are required to handle the sparse eigenvalues, compared to the restricted version in Theorem II.12. The role played by A_K is the same in both cases, as with the other factors.

It is worth noting that, despite the complexity of multinomial logistic regression, the conditions for Theorems II.12 and II.13 are simple and intuitive, and match those used for linear models.

Remark 7 (Asymptotics for Multinomial Logistic Regression)**.** It is relatively straightforward to state asymptotic rates of convergence, as done in Corollary II.2. The first conclusion there is immediate from Theorem II.13 if in addition to the conditions required, we also impose that $\lambda_D \sqrt{s} = o(1)$, κ_D and $\min_{S:|S|=O(s)} \underline{\phi}\{Q, S\}$ are bounded away from zero, and $\overline{\overline{\phi}}(Q, \cdot)$ is bounded, uniformly in the set $\mathbb{N}^D_Q$. This also implies that $|\tilde{S}^D| = O_{P_n}(s)$ and $\|\gamma_t - \gamma^*_t\|_1 = O_{P_n}(\sqrt{n^{-1}s^2 \log(p \vee \underline{n})^{3/2+\delta}} + b^d_s s)$, which is important for verifying Assumption II.6.

The rates of convergence for the propensity score estimates and the ℓ_1 error of the coefficients are minimax optimal up to a factor of $\log(p \vee \underline{n})^{1/2+\delta}$. A tighter bias condition (by $\sqrt{s}$) is required than in the linear model case, due to the bias in estimating the Hessian.[17] Inspection of the proof shows that this condition can be

[17] The more stringent requirement may be an artifact of the proof. However, it is worth noting that using a different proof method and considering only an oracle series estimator in a semiparametric model, Cattaneo (2010, Theorem B-1) found the same bias requirement.

dropped in the binary case. ■

We now give our results for group lasso estimation of the conditional outcome regressions. In computing $\mu_t(x_i)$ for $d_i^t \neq 1$ we are performing out of sample prediction, which slightly complicates the bounds. Our first result is on the initial group lasso fit.

Theorem II.14 (Group Lasso Estimation of Linear Models)**.** *Suppose Assumption II.4(a), II.4(b), II.4(c) hold and that* $\max_{i \leq n} b_{t,i}^y \leq b_s^y$. *Define*

$$R_{\mathcal{E}} = \left(\frac{3\lambda_Y \sqrt{s}}{\kappa_Y} + 2b_s^y \right).$$

Then with probability $1 - \mathcal{P}$

$$\max_{t \in \overline{\mathbb{N}}_{\mathcal{T}}} \mathbb{E}_n[(x_i^{*\prime}\tilde{\beta}_t - \mu_t(x_i))^2]^{1/2} \leq \left(\frac{\overline{\phi}\{Q, \tilde{S}^Y \cup S_Y^*\}}{\underline{\phi}\{Q_t, \tilde{S}^Y \cup S_Y^*\}} \right)^{1/2} R_{\mathcal{E}} + b_s^y,$$

$$\max_{t \in \overline{\mathbb{N}}_{\mathcal{T}}} \left\| \tilde{\beta}_t - \beta_t^* \right\|_1 \leq \left(\frac{|\tilde{S}^Y \cup S_Y^*|}{\underline{\phi}\{Q, \tilde{S}^Y \cup S_Y^*\}} \right)^{1/2} \left(\frac{\overline{\phi}\{Q, \tilde{S}^Y \cup S_Y^*\}}{\underline{\phi}\{Q_t, \tilde{S}^Y \cup S_Y^*\}} \right)^{1/2} R_{\mathcal{E}},$$

and

$$|\tilde{S}^Y| \leq 32 s L_n \left\{ \min_{m \in \mathbb{N}_Q^Y} \sum_{t \in \overline{\mathbb{N}}_{\mathcal{T}}} \overline{\overline{\phi}}(Q_t, m) \right\},$$

where

$$\mathbb{N}_Q^Y = \left\{ m \in \{1, 2, \ldots, \overline{n}\} : m > 32 s L_n \sum_{t \in \overline{\mathbb{N}}_{\mathcal{T}}} \overline{\overline{\phi}}(Q_t, m) \right\}, \qquad \textit{and} \qquad L_n = \left(\frac{R_{\mathcal{E}} + b_s^y}{\lambda_Y \sqrt{s}} \right)^2.$$

This theorem generalizes Lounici, Pontil, van de Geer, and Tsybakov (2011) to the nonparametric, approximately sparse case, improves the sparsity bound, and gives out of sample prediction (imputation) results. The analogous generalization for within sample prediction loss (e.g. multi-task learning), $\mathbb{E}_{n,t}[(x_i^{*\prime}\tilde{\beta}_t - \mu_t(x_i))^2]^{1/2}$, may be found in the Appendix.

For refitting, we are predicting for the entire sample and so we utilize the general results given by Belloni, Chen, Chernozhukov, and Hansen (2012) for post-selection estimation of least squares. The following result is a direct implication of their Lemma 7 and our Theorem II.14.

Theorem II.15 (Post-Selection Linear Regression)**.** *Suppose* $\log(p) = o(n^{1/3})$ *and* $\min_{j \in \mathbb{N}_p} \mathbb{E}[X_j^{*2} U^2] > 0$ *hold in addition to the conditions of Theorem II.14. Then*

$$\mathbb{E}_n[(x_i'\hat{\beta}_t - \mu_t(x_i))^2]^{1/2} \leq A_1 \sqrt{\frac{s(\mathcal{T} \wedge \log(s\mathcal{T}))}{n\underline{\phi}\{Q, S_Y^*\}}} + A_2 \sqrt{\frac{|\hat{S}_Y \setminus S_Y^*| \log(p\mathcal{T})}{n\underline{\phi}\{Q, S_Y^{FP}\}}} + A_3 \sqrt{\mathbb{E}_n[(x_i^{*\prime}\tilde{\beta}_t - \mu_t(x_i))^2]}$$

and

$$\max_{t \in \bar{\mathbb{N}}_{\mathcal{T}}} \left\| \hat{\beta}_t - \beta_t^* \right\|_1 \leq A_4 \left(\frac{|\hat{S}_Y \cup S_Y^*|}{\underline{\phi}\{Q, \hat{S}_Y \cup S_Y^*\}} \mathbb{E}_n[(x_i'\hat{\beta}_t - \mu_t(x_i))^2] \right)^{1/2},$$

where for absolute constants A_k*, k=1, 2, 3, 4 that do not depend on* n *nor the DGP.*

As above, the performance of the refitting procedure depends in part on the success of the initial group lasso fit. Indeed, the middle term is dropped if the true support union is found. The constants A_k, k=1, 2, 3, 4 are not given explicitly but are known to be absolute bounds (de la Peña, Lai, and Shao 2009). This result is less precise than Theorems II.12 and II.13, but sufficient to verify Assumptions II.5 and II.6.

Remark 8 (Asymptotics for Multinomial Logistic Regression)**.** As in Remark 7, we can now recover Corollary II.2 by imposing that, uniformly in $\bar{\mathbb{N}}_{\mathcal{T}}$: κ_Y and $\min_{S:|S|=O(s)} \underline{\phi}\{Q_t, S\} \wedge \underline{\phi}\{Q, S\}$ are bounded away from zero, and $\bar{\bar{\phi}}(Q, \cdot) \vee \bar{\bar{\phi}}(Q_t, \cdot)$ is bounded, also uniformly in the set $\mathbb{N}_Q^Y$. This also yields $|\tilde{S}^Y| = O_{P_n}(s)$ and $\|\tilde{\beta}_t - \beta_t^*\|_1 = O_{P_n}(\sqrt{n^{-1}s^2 \log(p \vee \underline{n})^{3/2+\delta}} + b_s^y \sqrt{s})$. ■

2.7 Numerical and Empirical Evidence

2.7.1 Simulation Study

To illustrate the uniform validity of our inference procedure we conducted a small-scale Monte Carlo exercise to study how our estimator behaves as the propensity score and regression functions change, and the model selection problem becomes more or less difficult. For simplicity we focus on the average effect of a binary treatment with . We generate 1000 observations $(y_i, d_i, x_i')'$, with $p = 500$, from the models in Example 3. The covariates include an intercept, with the remainder drawn from $N(0, \Sigma)$, with covariance $\Sigma[j_1, j_2] = 2^{-|j_1 - j_2|}, 2 \leq j_1, j_2 \leq 500$. Errors are standard Normal. The crucial aspect of the DGP are the coefficient vectors β_0^0, β_1^0, and γ^0. We consider a range of models, defined by positive scalars ρ_β, ρ_γ, α_β, and α_γ, as follows:

$$\beta_0^0 = \rho_\beta(-1, 1, -1, 2^{-\alpha_\beta}, -3^{-\alpha_\beta}, \ldots, j^{-\alpha_\beta}, \ldots, p^{-\alpha_\beta})', \qquad \beta_1^0 = -\beta_0^0, \qquad \text{and}$$
$$\gamma^0 = \rho_\gamma(1, -1, 1, -2^{-\alpha_\gamma}, 3^{-\alpha_\gamma}, \ldots, j^{-\alpha_\gamma}, \ldots, -p^{-\alpha_\gamma})'.$$

The multipliers ρ_β and ρ_γ affect the signal-to-noise ratio (the variance is fixed), but not the sparsity. For very small values distinguishing the large and small coefficients is difficult for a given sample size. The exponents α_β and α_γ control the sparsity, where for small values a sparse representation is not possible.

Figure 2.1 shows the empirical coverage rate of 95% confidence interval for $\mu_1 - \mu_0$ for different DGPs. Panel (a) shows the multipliers ρ_β and ρ_γ ranging over 0.01 (weak signal) to 1 (strong), with $\alpha_\beta = \alpha_\gamma = 2$. Panel (b) varies the sparsity exponents α_β and α_γ range over 1/8 (not sparse) to 4 (very sparse), with $\rho_\beta = \rho_\gamma = 1$. Of 1000 observations total, the (mean) size of the comparison group declines from 497 to 302 as ρ_γ increases and 444 to 303 as α_γ increases, over their given ranges. Coverage is exceedingly accurate over all signal strengths, and breaks down only when neither $\mu_t(x_i)$ nor $p_t(x_i)$ is sparse, which is exactly when Assumption II.5(b) (or condition (ii) of Theorem II.1) can not be satisfied. Note that coverage accuracy is retained when only one function is sparse, showcasing the double-robustness property.

2.7.2 Empirical Application

To illustrate the role that model selection can play in a real-world application, we revisit the National Supported Work (NSW) demonstration. The NSW has been analyzed numerous times since LaLonde (1986). Our aim is a simple study of model selection, not a comprehensive or conclusion evaluation of the NSW. As such, we focus on the subsample used by Dehejia and Wahba (1999) and the Panel Study of Income Dynamics (PSID) comparison sample, taking as given their data definitions, sample selection, and trimming rules. Detailed discussion of these choices, and the NSW program may be found in Dehejia and Wahba (1999, 2002) (hereafter DW99 and DW02) and Smith and Todd (2005), and references therein. Briefly, the outcome of interest is earnings following a job training program. The dataset includes a treatment indicator, post-treatment earnings (1978), two years of pre-treatment earnings (1974[18] and 1975), as well as age, education, a marital status, and indicators for Black and Hispanic. Thus, X consists of seven variables.

[18] This naming follows DW99, though the variable itself may be measured outside 1974, see discussion in the works cited.

Our goal is to highlight the role model selection has in inference, and hence we will be interested in comparing specifications. We will keep the estimator fixed: all estimates will be based on the doubly-robust estimator (2.8) (with standard errors from Section 2.5.2). We will consider the following three:

1. **No Selection**: X, (earn1974)^2, (earn1975)^2, (age)^2, and (educ)^2;

2. **Informally Selected:** The above, plus $\mathbb{1}\{$educ<HS$\}$, $\mathbb{1}\{$earn1974=0$\}$, $\mathbb{1}\{$earn1975=0$\}$, and ($\mathbb{1}\{$earn1974=0$\}\times$Hispanic). This specification was selected by DW02 using an informal balance test.

3. **Group Lasso Selection:** X, $\mathbb{1}\{$educ<HS$\}$, $\mathbb{1}\{$earn1974=0$\}$, $\mathbb{1}\{$earn1975=0$\}$, all possible interactions, and all polynomials up to order five of the continuous covariates (age, educ, earn1974, earn1975).

For specifications 1 and 2, we use the same covariates in the outcome and treatment models. In addition, all specifications include an intercept and we include education and pre-treatment income in the refitting step following model selection. We follow DW99 and DW02 and discard controls with estimated propensity score larger (smaller) than the maximum (minimum) in the treated sample.[19]

Table 2.1 presents results from these three specifications, and includes the experimental arm of the NSW. The group lasso based estimate performs very well: the point estimate is accurate and the interval is tight. Selecting from 171 possible covariates allows for a great deal of flexibility, but the sparsity of the estimate keeps the variance well-controlled. The no-selection point estimate is accurate, but fails to yield significance, while the specification of DW02 yields a significant, but overly high estimate and wide confidence intervals. The benefits of explicit model selection are clear.

2.8 Discussion

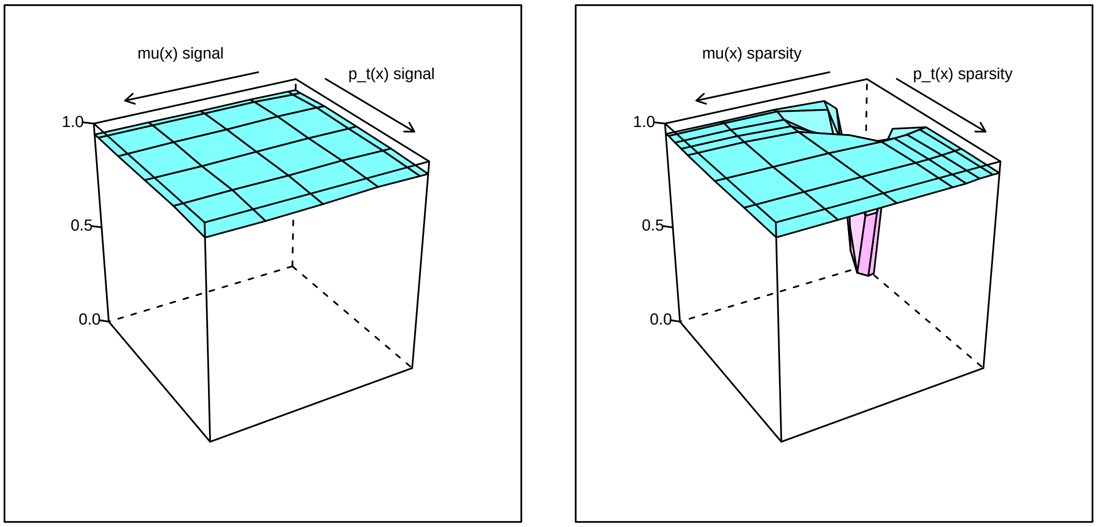
The main results of this paper established a method to achieve uniformly valid inference on average effects of multivalued treatments even after model selection among possibly more covariates than observations. We demonstrated robustness to model selection errors, misspecification, and heterogeneous effects in observables. To accomplish this, we proved new results on group lasso estimation of multinomial logistic

[19] A formal treatment of trimming is beyond the scope of the present study. The goal of our analysis is illustrative, and hence we take DW99's trimming as given. This issue is discussed by DW99, DW02, and Smith and Todd (2005).

regression models. Numerical evidence shows that our method is quite promising for applications.

We handle very general treatment effects models, but restrict attention to studying impacts on the mean. A useful and natural extension would be to consider quantile treatment effects (Firpo 2007) or more generic moment condition based estimands (Cattaneo 2010). Under appropriate regularity conditions, it seems plausible that such an extension can be made. However, the first stage estimation is quite complex in our framework, and this extension would require additional nontrivial technical work. In additional, we plan to develop a formal choice for the penalty parameter that is optimal in some sense, beyond the simple discussion in Section 2.6.1.

Figure 2.1: Empirical Coverage of 95% Confidence Intervals, Varying Signal Strength and Sparsity of $p_t(x)$ and $\mu_t(x)$



(a) Varying Signal Strength

(b) Varying Sparsity

Table 2.1: Analysis of NSW Demonstration: Treatment Effects on the Treated and Confidence Intervals for Various Specifications

	Number of Variables		Sample Sizes[(c)]			
Specifications:	Before selection[(a)]	After selection[(b)]	Control	Treated	ATT	95% CI
Experimental Benchmark	–	–	260	185	1794	[110, 3479]
Doubly-Robust Estimates						
Specification 1 (No Selection)	N/A	11	1211	185	1664	[-276, 3604]
DW02 (Informal Selection)	??	15	1058	185	2528	[149, 4908]
Refitting after Group Lasso Selection	171	20/6	1735	185	1737	[33, 3441]

Notes: All analyses use the DW99 subsample and PSID control group. Specifications vary, but all estimates and standard errors of from the method defined in Section 2.5 with the exception of the partially linear model.

(a) Not counting the intercept. The total set of variables considered by DW02 is not known.

(b) For the group lasso estimators, the two numbers given are for those used in the outcome regressions and propensity score, respectively. For other doubly-robust estimators all variables are used in the propensity score and outcome models.

(c) The full sample begins with 2490 controls and 185 treated units. Control observations outside the range of estimated propensity scores in the treated sample are discarded.

CHAPTER III

Optimal Convergence Rates, Bahadur Representation, and Asymptotic Normality of Partitioning Estimators

3.1 Introduction

Nonparametric estimation of an unknown conditional expectation function and its derivatives is an important problem in econometrics (see, e.g., Ichimura and Todd (2007) and references therein). In many applications the object of interest is a conditional expectation, its derivative, or functional thereof, while in other cases their nonparametric estimators are employed as a first step in a semiparametric procedure. The implementation of nonparametric estimators requires suitable large sample properties, including sufficiently rapid rates of convergence and known asymptotic distributions. Series- and kernel-based methods are examples whose properties are now well understood.

This paper studies the large sample properties of an estimator of the regression function and its derivatives known as *partitioning*. This estimation strategy is alternatively referred to as *blocking*, *subclassification*, or *stratification*. The estimator is constructed by partitioning the support of the conditioning variables into disjoint cells, which become smaller with the sample size, and within each the unknown regression function (and its derivatives) is approximated by linear least-squares using a fixed-order polynomial basis (other bases are possible). Consistent estimation is achieved as the cells become small enough to remove the error of the parametric approximation. For a recent textbook discussion of this estimation strategy see Györfi, Kohler, Krzyżak, and Walk (2002, Chapter 4). After the necessary notation and assumptions are introduced, we provide a detailed comparison between partitioning estimators and other nonparametric estimators in Section 3.2.2 below.

The partitioning estimator, although simple and intuitive, has not received a thorough treatment in the econometrics or statistics literature. The available results typically concern mean-square rates for special cases (see, e.g., Kohler, Krzyżak, and Walk (2006) and references therein). The main goal of this paper is to provide a general asymptotic treatment of partitioning estimators. Our analysis yields the following new insights. First, employing simple and intuitive sufficient conditions, in most cases weaker than those in the existing literature, mean-square and uniform convergence rates of the partitioning estimator are established and shown to be optimal. More generally, the uniform convergence rate explicitly highlights a natural trade-off between moment assumptions and rate restrictions. Second, we characterize the leading terms of a conditional integrated mean-square error expansion and provide an optimal plug-in selector for the tuning parameter. Third, we derive a uniform Bahadur-type representation of linear functionals of the partitioning estimator, which is used to establish asymptotic normality under simple and intuitive conditions, with a suitable standard-error estimator. We cover both regular and irregular estimands. The applicability of the new results is illustrated with three examples: (i) derivative of the regression function at a point, (ii) partial and full means, and (iii) weighted average derivatives. Our results are also useful in other contexts in econometrics, as discussed in Section 3.1.1 below.

The paper proceeds as follows. In the remainder of this section we give the main motivations for our work, discussing in particular the importance of our results for both empirical and theoretical econometrics. Section 3.2 describes the partitioning estimator formally and also provides a comparison to other nonparametric estimators. Rates of convergence and a general integrated mean-square error expansion for the partitioning estimator are given in Section 3.3, while a Bahadur-type representation for linear functionals of the estimator and asymptotic normality with valid standard-error estimators are developed in Section 3.4. The results of a Monte Carlo study are summarized in Section 3.5. Finally, Section 3.6 concludes. Proofs are gathered in the appendix. A supplement is available upon request containing detailed technical proofs and greatly expanded simulation results.

3.1.1 Motivation and Preliminary Discussion

Studying the large-sample properties of partitioning estimators may be interesting and important for a variety of reasons, some theoretical and others methodological. The partitioning estimator has specific features and asymptotic optimality properties that make it a useful addition to the econometrics toolkit: a complement, not a substitute, to the arsenal of nonparametric procedures commonly employed in econometrics. This estimator is attractive because it is very tractable and enjoys useful asymptotic representations leading to intuitive results, as well as other features that may be useful in econometric applications.

In particular, the partitioning estimator is potentially discontinuous in finite samples (just like nearest-neighbor estimators). This specific characteristic may be an advantage from a practical point of view, and could also lead to an estimator with desirable theoretical properties. The "binning" underlying the partitioning estimator arises naturally in many economic problems, where units (people, firms, etc.) in the same bin share similar economic behavior, and therefore partitioning-based inference procedures have been proposed to retain this natural interpretability (see applications below). From a theoretical perspective, we are interested in understanding the asymptotic properties of partitioning given its potential discontinuity in finite samples, and how they compare with results for other nonparametric procedures. We briefly discuss three implications of this discontinuity, which make the partitioning estimator theoretically and practically interesting in our view.

1. *Shape Restrictions: Convergence Rates.* Nonparametric estimation typically assumes the estimand is smooth and most estimators are constructed imposing some of the underlying smoothness assumed. The partitioning estimator does not impose smoothness and therefore allows us to understand what effects imposing this shape restriction may have on asymptotic properties, which arguably is of theoretical interest. For instance, we establish optimal uniform convergence rates for partitioning, showing (by example) that imposing smoothness is not necessary for this result. This finding is not ex-ante obvious in our view, especially given other known results (see Section 3.2.2).

2. *Shape Restrictions: Bias-Variance Trade-Off.* From a more practical perspective, removing the smoothness restriction may be interpreted as "freeing up restrictions". This means that the estimator will have a different bias-variance behavior in finite samples. To fix ideas, consider the linear partitioning and linear regression spline estimators of a univariate regression function. For each sample size, both are (piecewise linear) least squares fits, and differ only in that the spline is required to be continuous (see Section 3.2.2). That is, the linear spline is a restricted least squares problem compared to the partitioning estimate. From linear model results, it follows that the spline has larger bias

than the partitioning estimate, but smaller variance.[1] Neither can be strictly superior or inferior, based on the usual bias-variance trade-off, and in fact the partitioning estimator may have better properties from a theoretical point of view.

3. *Diagnostics*. The potential discontinuity of the partitioning estimator in finite samples makes it a useful complement to existing smooth estimates already available in the literature. Specifically, the partitioning estimate may be used as a diagnostic check on the underlying smoothness assumptions imposed by other procedures, particularly if such assumptions are in question for a certain region of the support. Furthermore, the discontinuous partitioning estimate can be used to characterize the overall variability of the data relative to a smoothed-out estimate (see the regression discontinuity application below for an example).

Further motivation for our work stems from the role of partitioning estimators in empirical economics. Perhaps originating with the regressogram of Tukey (1947), partitioning-based procedures have been suggested in many contexts where "binning" has a natural interpretation, despite their formal properties being unknown in most cases. We close this section by briefly discussing four examples where partitioning estimation arises in econometrics: as an exploratory device, a nonparametric estimator, and two semiparametric cases.

Application: Regression Discontinuity. Partitioning estimators are used heuristically in the regression discontinuity (RD) design for two purposes: (i) to plot a smoothed-out cloud of points along with global polynomial fits of the underlying regression function for control and treatment units, and (ii) to investigate whether the data suggests the presence of other possible discontinuities in the underlying conditional expectation of potential outcomes, as a form of falsification test. Imbens and Lemieux (2008) review the RD literature, and explicitly advocate partitioning (calling it a "histogram-type estimate") to assess the plausibility of the RD design. Our general result in Theorem III.5 is employed in Calonico, Cattaneo, and Titiunik (2012) to derive an optimal choice of partition length in this context, thereby providing a systematic way of plotting RD data. ■

Application: Porfolio Sorting. In understanding anomalous asset returns, a common approach is "portfolio sorts", in which assets are partitioned into homogeneous

[1] This claim assumes that the estimators are misspecified in finite samples, as is the case with nonparametric estimators in general. This remains true when comparing to kernel-based estimators.

groups according to characteristics that may drive anomalies. A number of informal and formal analyses are then performed on the sorted assets, including tests of monotonicity and comparison of extremes. See, e.g., Fama and French (2008). Our results may be used to develop formal nonparametric inference for this type of application. ■

Application: Subclassification on Observables. In many econometric contexts units are divided in groups according to their observed characteristics, and then inference is conducted first within each subclass and then overall. Under an ignorability assumption, for example, subclassification (or partitioning) has been proposed in multiple forms to estimate treatment effects. Imbens and Wooldridge (2009) give a recent survey of the program evaluation literature, which includes several examples of partitioning-based procedures. Despite many such procedures have been proposed and used in empirical work, there is a paucity of rigorous asymptotic theory. The theoretical results presented herein may be used to characterize the large-sample properties of those partitioning-based procedures. For instance, in Cattaneo and Farrell (2011b) we employ these results to formalize the properties of a partitioning-based estimator of the average treatment effect and dose-response function. ■

Application: Average Derivatives. Partitioning also yields simple and intuitive estimators for derivatives of the regression function. Based on this observation, Banerjee (2007) recently proposed a partitioning-based semiparametric average derivative estimator. In Section 4 we discuss an alternative semiparametric estimator for (weighted) average derivatives, and establish its asymptotic properties under general, easy-to-interpret sufficient conditions. ■

3.2 The Partitioning Estimator

3.2.1 Setup and Estimator

Before describing the estimator we introduce some notation. For a scalar, vector, or matrix A we denote $|A| = \sqrt{\operatorname{tr}(A'A)}$. For a multi-index $k = (k_1, k_2, \cdots, k_d) \in \mathbb{Z}_+^d$, we let $[k] = k_1 + \cdots + k_d$, $x^k = x_1^{k_1} \cdots x_d^{k_d}$ for $x = (x_1, \cdots, x_d)' \in \mathbb{R}^d$, and $\partial^k h(x) = \partial^{[k]} h(x)/(\partial^{k_1} x_1 \cdots \partial^{k_d} x_d)$ for smooth enough function $h(x)$.

We impose the following assumption on the data generating process throughout.

Assumption III.1.

(a) $(Y_1, X_1'), \cdots, (Y_n, X_n')$ *is an i.i.d. sample from* (Y, X')*, and* $X \in \mathcal{X}$ *is continuously distributed with Lebesgue density* $f(x)$*.*

(b) $\mathcal{X} \subset \mathbb{R}^d$ *is given by* $\mathcal{X} = \times_{\ell=1}^{d} \mathcal{X}_\ell$*, a Cartesian product of compact, convex intervals.*

(c) $\mathbb{E}[|Y|^{2+\eta} \mid X]$ *is bounded for some* $\eta \geq 0$*.*

(d) $f(x)$ *is bounded and bounded away from zero on* $\mathcal{X}$*.*

(e) $\mu(x) = \mathbb{E}[Y|X = x]$ *is* S*-times continuously differentiable on (an extension of)* $\mathcal{X}$*, and satisfies* $|\partial^m \mu(x) - \partial^m \mu(x')| \leq C\,|x - x'|^\alpha$*, for some constants* $C > 0$ *and* $\alpha \in (0, 1]$*, and all* $x, x' \in \mathcal{X}$ *and* $[m] = S$*.*

We discuss the salient features of this assumption in the following remarks.

- Part (a) restricts attention to cross-sectional contexts with continuous regressors. Our results can be extended to cover some form of time-dependent data, or to include discrete regressors by working conditionally, although we do not consider these extensions here to simplify the discussion and notation.

- Part (b) requires regressors with compact support. The assumed rectangular structure is without loss of generality for most of the results presented here. The compact support assumption has the main advantage of allowing for the density $f(x)$ to be bounded away from zero on the full support of X, but has the potential drawback of introducing bias at the boundary of the support. This assumption is also imposed for nonparametric series estimators (Newey 1997) and nonparametric local polynomials (Fan and Gijbels 1996), but it can be relaxed in semiparametric inference by considering weaker (weighted) norms (Chen 2007). In this paper we only focus on the conventional mean-square and uniform norms.

  This assumption is important because it can affect the attainable convergence rates for nonparametric regression estimators in general. Specifically, in the case of mean-square convergence, Kohler, Krzyżak, and Walk (2009) show that it is possible to attain Stone's (1982) optimal L_2 convergence rate even without compactness as long as certain moment conditions hold, and Kohler, Krzyżak, and Walk (2006) show that a cleverly constructed special partitioning estimator attains this rate. In the case of the uniform convergence rate, it appears to be an open question whether Stone's (1982) bound is achievable without compactness.

- Part (c) allows for the case of $\eta = 0$ (i.e., bounded second conditional moment only), and the generality will be useful in the derivation of the uniform convergence rate.

- Part (d) ensures that all cells in the partition will contain enough observations asymptotically, and appears difficult to relax without affecting the rates of convergence.

- Part (e) is a classical smoothness condition controlling the amount of bias reduction possible, when coupled with an appropriate basis choice employed within each cell.

To describe the nonparametric procedure, we first give a precise description of the partitioning scheme. For a sequence $J_n \to \infty$ as $n \to \infty$, partition each $\mathcal{X}_\ell$ into the J_n disjoint intervals $[p_{\ell,j-1}, p_{\ell,j}), j = 1, \ldots, J_n - 1$, and $[p_{\ell,J_n-1}, p_{\ell,J_n}]$, with $p_{\ell,j-1} < p_{\ell,j}$ for all j. The complete partition of $\mathcal{X}$ consists of the J_n^d cells formed as Cartesian products of all such intervals. Let $P_j \subset \mathbb{R}^d$ denote a generic cell of the partition, $j = 1, \ldots, J_n^d$, and for $x \in \mathbb{R}^d$, let $\mathbb{1}_{P_j}(x)$ be the indicator for $x \in P_j$. Throughout, we suppress the dependence on n for notational convenience: all aspects of the partition implicitly depend on n.

To guarantee that each cell is well defined we require that $|p_{\ell,j} - p_{\ell,j-1}| \asymp J_n^{-1}$ for all $\ell = 1, \ldots, d$ and $j = 1, \ldots, J_n$, where for scalars a and b, $a \asymp b$ denotes that $C_* b \leq a \leq C^* b$ for positive constants C_* and C^* that do not depend on $j = 1, \ldots, J_n$ nor n. Hence, by construction the partition satisfies $\text{vol}(P_j) \asymp J_n^{-d}$, where $\text{vol}(P_j)$ denotes the volume of cell P_j. A simple, natural partitioning scheme meeting this requirement is evenly dividing the support of each covariate, although other possibilities are allowed so long as all intervals decrease proportionally to J_n.

Within each cell the unknown conditional expectation is approximated by solving a least squares problem. For fixed $K \in \mathbb{N}$, let $r(x_\ell) = (1, x_\ell, x_\ell^2, \ldots, x_\ell^{K-1})'$ denote the vector of powers up to degree $K - 1$ on a single covariate $x_\ell \in \mathcal{X}_\ell$. Let $R(x)$ represent a column vector containing the complete polynomial basis of degree $K - 1$ formed as the Kronecker product of the $r(x_\ell)$, discarding terms with degree exceeding $K - 1$. Thus, each element of $R(x)$ is given by $x^k = x_1^{k_1} \cdots x_d^{k_d}$ for a unique $k \in \{k \in \mathbb{Z}_+^d : [k] \leq K - 1\}$. We assume $R(x)$ is ordered ascendingly in $k \in \mathbb{Z}_+^d$ and $\ell = 1, \ldots, d$. For example, if $K = 1$ then $R(x) = (1)$ and sample means are fitted in each cell, while if $K = 2$ then $R(x) = (1, x_1, \ldots, x_d)'$, corresponding to ordinary linear least squares. This construction is explicitly meant to cover the general, unrestricted case, although in applications other bases may be of interest. For example, if $\mu(x)$ additively separable, then the interactions between covariates may be excluded from the basis, leading to a simpler least squares problem. This additional flexibility is useful, for example, in estimation via control functions. The goal of this construction is to ensure that $R(x)$ is flexible enough to remove bias up to the appropriate order (see Lemma B.2).

The choice of K is intimately related to bias reduction. Setting a higher K allows for a more flexible functional form within each cell and hence lower bias, provided the underlying function is sufficiently smooth.[2] In this sense, the partitioning scheme and the choice of K play the same role for the partitioning estimator that the choice of specific higher-order kernel plays in kernel-based estimation, while the choice of J_n is analogous to the choice of bandwidth in a kernel context. The partitioning scheme and (fixed) K represent the smoothing parameter, and $J_n \to \infty$ is the tuning parameter of the nonparametric procedure.

Let $R_j(x) = \mathbb{1}_{P_j}(x)R(x)$ denote basis restricted to the cell containing x. Using this notation the partitioning regression estimator of order K is given by:

$$\begin{aligned} \hat{\mu}(x) &= \sum_{j=1}^{J_n^d} R_j(x)'\hat{\beta}_j, \qquad \hat{\beta}_j = \left(R_j'R_j\right)^- R_j'Y, \\ R_j &= \left(R_j(X_1), \ldots, R_j(X_n)\right)', \qquad Y = (Y_1, \ldots, Y_n)', \end{aligned} \tag{3.1}$$

where A^- denotes any generalized symmetric inverse. Under regularity conditions given below, and with proper scaling, the matrix $R_j'R_j$ will be positive definite uniformly in j with probability approaching one (see Lemma B.4), and the standard inverse will exist. The structure given in Eqn. (3.1) implies that $\hat{\mu}(x)$ is a (random) function that has at most finitely many discontinuities, is almost everywhere differentiable, and is of bounded variation. (Qualifiers such as "almost everywhere" and "for n large enough are usually omitted for simplicity.)

To construct an estimator of the derivatives of $\mu(x)$, let $m \in \mathbb{Z}_+^d$ be a multi-index and $\partial^m\mu(x)$ denote a partial derivative of order $[m]$. An intuitive estimator of $\partial^m\mu(x)$ is

$$\widehat{\partial^m\mu(x)} \equiv \partial^m\hat{\mu}(x) \equiv \sum_{j=1}^{J_n^d} \mathbb{1}_{P_j}(x)\left(\partial^m R(x)\right)'\hat{\beta}_j, \tag{3.2}$$

which we take as the definition throughout. In words, $\partial^m\hat{\mu}(x)$ is defined as the derivative of the estimated polynomial regression function, restricted to a particular cell containing x (as there are no boundary issues in differentiating $R(x)$). Because $\partial^m R(x)$ has zeros in some components, the resulting estimator employs a lower degree

[2] This bias reduction is asymptotic. Ruppert and Wand (1994, Remark 4) provide a very interesting discussion, for local polynomials, that highlights how in finite samples this smoothing-bias reduction could be more than offset by increased variability.

basis but the least squares problem within each cell is unaffected. This intuitively corresponds to estimating the rougher function $\partial^m \mu(x)$. The main results of this paper also cover the estimation of derivatives of the regression function, provided K is large enough.

3.2.2 Related Literature

The partitioning estimator is closely related to, but different from, other nonparametric estimators available in the literature. In this section we describe how it relates to two common estimators: series and local polynomials.

From a series estimation perspective, the partitioning estimator may be recast as a linear sieve estimator. Define $\mathbf{R}_n(x) = (R_1(x)', \ldots, R_{J_n^d}(x)')'$ by collecting the bases over all J_n^d cells, and set $\mathbf{R}_n = [\mathbf{R}_n(X_1), \ldots, \mathbf{R}_n(X_n)]'$. The partition regression estimator can then be written as

$$\hat{\mu}(x) = \mathbf{R}_n(x)'\hat{\mathbf{B}}_n, \qquad \hat{\mathbf{B}}_n = (\mathbf{R}_n'\mathbf{R}_n)^{-}\mathbf{R}_n'Y = (\hat{\beta}_1', \ldots, \hat{\beta}_{J_n^d}')'.$$

This representation implies that results available from the sieve estimation literature are in principle applicable to the partitioning estimator. But by exploiting the specific structure of the partitioning estimator we are able to obtain faster uniform convergence rates and new results such as derivative estimation, an integrated mean-square error expansion, and a Bahadur representation, while improving on rate restrictions and using simple primitive conditions, when compared to the results available in the general series estimation literature (Newey (1997), de Jong (2002), and Belloni, Chen, Chernozhukov, and Kato (2012)).

Regression splines are series estimators for which improved results are available (Huang 2003). Partitioning estimators and polynomial splines are intuitively similar, but fundamentally different smoothing procedures. Both estimators rely on a refining partition of the support with fixed-order basis functions: an order K spline uses $K-1$ degree polynomials (in our notation). The key distinguishing characteristic is that at the cell boundaries (called "knots") the spline estimate is forced to be smooth whenever possible: a spline of order K is $(K-2)$-times differentiable at each knot. For precisely this reason splines are usually regarded as a "global" smoother. In contrast, partitioning estimators place no restriction on the behavior of the polynomials at the boundary of each cell, and hence the basis functions are truly local (and compactly supported). In this paper we show that the partitioning estimators have the same optimal L_2 convergence rate under the same rate restrictions as polynomial splines.

We also show that the partitioning estimator achieves the optimal uniform convergence rate for levels and derivatives. General series estimators are only known to have suboptimal uniform rates (de Jong 2002). (In personal communication, X. Chen shared preliminary work showing that spline least squares regression may achieve the optimal uniform convergence rate under certain conditions (Chen and Huang 2003).)

Kernel-based local polynomials are another class of nonparametric estimators of the regression function and its derivatives. Partitioning estimators are conceptually (and numerically) distinct from the kernel-based local polynomial estimators discussed in Fan and Gijbels (1996) and the local polynomial estimators discussed in Eggermont and LaRiccia (2009, Chapter 16), which are also different from each other. These local polynomial approaches and the partitioning estimators differ in the way that observations are grouped: the local polynomial approaches use observations near the evaluation point, as determined by the choice of kernel and bandwidth, while partitioning estimators use observations within each cell, regardless of the particular evaluation point. This fact implies that partitioning estimators are naturally discontinuous while local polynomials are not. The partitioning estimator can be viewed as a local polynomial estimator with a particular variable bandwidth and a uniform spherical kernel.

To describe how the local polynomials and the partitioning estimators differ, consider the estimation of the regression function (a similar discussion applies to derivative estimation). Both estimation procedures solve the following weighted least-squares problem:

$$\hat{\beta}_n(x) = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^{\dim(B(\cdot))}} \sum_{i=1}^{n} W_n(X_i, x) \left(Y_i - B(X_i, x)'\beta\right)^2 ,$$

where $W_n(X_i, x)$ is a non-negative weighting function and $B(X_i, x)$ is a choice of polynomial basis. Both local polynomials estimators mentioned above employ $W_n(X_i, x) = K((X_i - x)/h_n)/h_n$, for a fixed kernel function $K(\cdot)$ and a bandwidth sequence $h_n \to 0$. Moreover, the local polynomials in Fan and Gijbels (1996) are obtained by choosing $B(X, x) = R(X - x)$ and setting $\hat{\mu}(x) = e_1'\hat{\beta}_n(x)$ with $e_1 = (1, 0, 0, \ldots, 0)'$, while the local polynomial estimator in Eggermont and LaRiccia (2009, Chapter 16) employ $B(X, x) = R(X)$ and set $\hat{\mu}(x) = R(x)'\hat{\beta}_n(x)$. In contrast, the partitioning estimators use $W_n(X_i, x) = \sum_{j=1}^{J_n^d} \mathbb{1}_{P_j}(X_i)\mathbb{1}_{P_j}(x)$ and $B(X, x) = R(X)$, and set $\hat{\mu}(x) = R(x)'\hat{\beta}_n(x)$. Therefore, results from local polynomial methods cannot be applied directly to partitioning estimators.

Finally, as a reviewer pointed out, Stone (1982, Section 3) also suggested another (hybrid) local polynomial procedure which bears some relation to the partitioning estimator studied here. Using our notation, Stone's estimators employ $W_n(X_i, x) = \sum_{j=1}^{J_n^d} \mathbb{1}_{P_j}(X_i)\mathbb{1}\{j : |z - x| \leq h_n, \forall z \in P_j\}/N_j$, where $N_j = \sum_{i=1}^{n} \mathbb{1}_{P_j}(X_i)$ is the number of observations in P_j. This estimator uses all (data in) cells falling *completely* within an h_n-ball around the evaluation point x, in contrast to partitioning which only considers observations in the cell P_j. Moreover, Stone's estimator necessitates the choice of two tuning parameters, J_n and h_n, which are required to satisfy $h_n J_n \to \infty$. The rate restriction that the cells are required to shrink faster than the bandwidth implies that the number of cells in each h_n-ball tends to infinity, and hence asymptotically the weighting is constant in the h_n-ball and symmetric about x, just like a classical local polynomial with a spherical uniform kernel with bandwidth h_n, and not like the partitioning estimators considered here.

In Section 6 we provide further discussion of the potential advantages and disadvantages of the partitioning estimators when compared to series, kernel and nearest-neighbor estimators.

3.3 Convergence Rates and Integrated Mean-Square Expansion

Some further notation is necessary to state the results. Let $a \wedge b = \min\{a, b\}$, $a, b \in \mathbb{R}$. For a function $h(\cdot)$ let $\|h\|_p^p = \int_{\mathcal{X}} |h(x)|^p f(x)dx$ and $\|h\|_\infty = \sup_{x\in\mathcal{X}} |h(x)|$ denote the L_p and L_∞ norms; function arguments are suppressed if there is no confusion.

3.3.1 Rates of Convergence

The following theorem gives the L_2 convergence rate for the partitioning estimate of the regression function and its derivatives.

Theorem III.2. *If Assumption III.1 holds and $J_n^d \log(J_n^d) = o(n)$, then for $s \leq S \wedge (K - 1)$:*

$$\max_{[m]\leq s} \|\partial^m \hat{\mu} - \partial^m \mu\|_2^2 = O_p\left(\frac{J_n^{d+2s}}{n} + J_n^{-2((S+\alpha)\wedge K - s)}\right).$$

This theorem shows that, by setting J_n^d proportional to $n^{d/(2(S+\alpha)+d)}$ and $K \geq S + 1$, the partitioning estimator achieves Stone's (1982) optimal rate, a property shared by other series- and kernel-based estimators. Because the partitioning estimator can be recast as a series estimator, the conclusion in Theorem III.2 for the regression function (i.e., $[m] = 0$) could have been obtained directly from general results in the sieve estimation literature under high-level assumptions. A contribution of this theorem is to obtain such a result under weaker, primitive conditions. In particular, the rate restriction required, $J_n^d \log(J_n^d) = o(n)$, is weaker than the one typically imposed in the general series literature (e.g., Newey (1997) requires the analogue of $J_n^d \max_{[m]\leq s} \|\partial^m \mathbf{R}_n(\cdot)\|_\infty^2 = o(n)$ with $\max_{[m]\leq s} \|\partial^m \mathbf{R}_n(\cdot)\|_\infty^2$ polynomial in J_n^d). This refined rate restriction was also used by Huang (2003) for multivariate regression splines and by Belloni, Chen, Chernozhukov, and Kato (2012) for general series estimation, but employing the operator norm instead of the (stronger) Frobenius norm used herein.

Theorem 1 also contributes to the literature in two additional ways. First, existing results for partitioning estimators of $\mu(\cdot)$ only yield the optimal rate when Y is bounded, and otherwise give suboptimal rates (see, Györfi, Kohler, Krzyżak, and Walk (2002, Corollaries 11.2 and 19.3)). Second, this result shows that the partitioning estimator of derivatives of $\mu(\cdot)$ achieves the optimal rate under the same weak conditions. This result, which appears to be new for the partitioning estimation literature, is often useful in econometric applications (e.g., average marginal effects).

Next, we discuss the L_∞ convergence rate of the partitioning estimator.[3]

Theorem III.3. *Suppose the conditions of Theorem III.2 hold. If, in addition, for some $\xi \in [0, 1 \wedge \eta]$ the partition satisfies $J_n^{d\xi(1+2/\eta)} \log(J_n^d)^{2-(1+2/\eta)\xi} = O(n)$, with $0/0 \equiv 0$, then for $s \leq S \wedge (K-1)$:*

$$\max_{[m]\leq s} \|\partial^m \hat{\mu} - \partial^m \mu\|_\infty^2 = O_p\left(\frac{J_n^{(2-\xi)d+2s} \log(J_n^d)^\xi}{n} + J_n^{-2((S+\alpha)\wedge K - s)} \right).$$

The parameter ξ is a user-defined choice, which depends on the underlying moment condition of Assumption III.1(c). This parameter is not a tuning parameter in the classical nonparametric sense, but rather is explicitly introduced in Theorem III.3 for potential applications. As formalized in Lemma B.5, ξ allows for greater or lesser weight placed on the tails of the (conditional) distribution of the outcome variable, which in turn provides a trade-off between the rate restriction imposed and the (possibly suboptimal) rate of convergence of the estimator. Consider two examples: (i) if $\mathbb{E}[Y^4|X] < \infty$ (i.e., $\eta = 2$), then the additional requirement of Theorem III.3 is $J_n^{2d} = O(n)$ for $\xi = 1$, implying essentially $(S+\alpha) \wedge K \geq d/2$, and (ii) if $\eta = 0$, which implies $\xi = 0$, only bounded conditional variance is assumed, and Theorem

[3] In the appendix we also provide conditions for the result to hold almost surely.

III.3 gives the (suboptimal) rate $J_n^{2(d+s)}/n$ with convergence implying the other rate restrictions. For $0 < \xi < 1$ neither rate restriction in Theorem III.3 implies the other.

With an appropriate choice of J_n, the convergence rate in Theorem III.3 will be optimal if $\eta \geq 1$, allowing for $\xi = 1$, provided the rate restrictions are satisfied. Known results for general series estimation (e.g., Newey (1997) and de Jong (2002)) do not achieve this optimal uniform rate, and impose stronger side restrictions, which implies that the rate-optimality of the partitioning estimator cannot be deduced from those results. Conventional local polynomial estimators, on the other hand, do achieve this optimal uniform rate (e.g., Masry (1996)) but these results do not apply to the partitioning estimator, as discussed above.

In semiparametric contexts it may be neither necessary nor desirable that the nonparametric component attain the optimal rate, if the goal is to minimize the restrictions imposed. Forcing the preliminary nonparametric estimator to achieve the optimal rate may require overly-restrictive conditions on the model (e.g., moments) or tuning/smoothing parameters. Theorem III.3 shows that these conditions may be ameliorated by an appropriate choice of ξ. See Cattaneo and Farrell (2011b) for an application of this result.

Finally, we note that the bias rate in Theorems III.2 and III.3, $J_n^{-2((S+\alpha)\wedge K-s)}$, highlights the fact that the partitioning estimator is not "adaptive" to the underlying smoothness of the regression function: to improve the convergence rate of the estimator, the order K must be chosen "large enough" given the unknown smoothness level, S. This feature is common to many other nonparametric estimators, including local polynomials and regression splines.[4] Although there are estimators that "adapt" to the underlying smoothness, these are usually believed to have poor finite-sample properties.

3.3.2 Integrated Mean-Square Error Expansion

We present a general conditional Integrated Mean-Square Error (IMSE) asymptotic expansion for the partitioning estimator.[5] We focus on evenly split partitions for notational simplicity, but the results may be extended to other partitioning schemes. We briefly discuss how to derive a direct plug-in rule for selecting the value of J_n

[4] Assume $s = 0$ for simplicity. An order p local polynomial estimator (p odd) employing bandwidth $h_n \to 0$ has bias-rate $h_n^{(S+\alpha)\wedge p}$ (e.g., Fan and Gijbels (1996) or Masry (1996)). An order p regression spline estimator employing knots $\kappa_n \to \infty$ has bias-rate $\kappa_n^{-((S+\alpha)\wedge p)}$ (e.g., Huang (2003) or Belloni, Chen, Chernozhukov, and Kato (2012)).

[5] The supplemental appendix also contains an unconditional IMSE expansion for the special case of $K = 1$.

using this expansion, which provides an alternative to the cross-validation procedures discussed in Györfi, Kohler, Krzyżak, and Walk (2002, Chapters 8, 13).

We impose the following additional assumption.

Assumption III.4.

(a) $\sigma^2(x) = \mathbb{V}[Y|X = x]$ *and* $f(x)$ *are continuous on* $\mathcal{X}$.

(b) $\mu(x)$ *is* $(S + 1)$*-times continuously differentiable on (an extension of)* $\mathcal{X}$.

These additional smoothness conditions allow us to characterize the leading constants in the asymptotic IMSE expansion, as opposed to giving bounds in the rates of convergence. Assumption III.4(b) is a slight strengthening of Assumption III.1(e). Let $\text{vol}(\mathcal{X})$ denote the volume of the support, with $|\mathcal{X}_\ell|$ denoting the length of the interval $\mathcal{X}_\ell$ for $\ell = 1, 2, \ldots, d$, and set $\mathcal{X}data = (X_1, \ldots, X_n)'$. To save some notation, we also assume $K = S + 1$.

Theorem III.5. *Suppose the conditions of Theorem III.2 and Assumption III.4 hold. If* $w(x)$ *is continuous on* $\mathcal{X}$*, then:*

$$\int_{\mathcal{X}} \mathbb{E}\left[(\partial^m \hat{\mu}(x) - \partial^m \mu(x))^2 \mid \mathcal{X}data\right] w(x)dx$$

$$= \frac{J_n^{d+2[m]}}{n}[\mathscr{V}_{K,d,m} + o_p(1)] + J_n^{-2(K-[m])}[\mathscr{B}_{K,d,m} + o_p(1)],$$

where $\mathscr{V}_{K,d,m}$ *and* $\mathscr{B}_{K,d,m}$ *are given in Eqns.* (B.4) *and* (B.5) *in the Appendix.*

This result gives a general conditional IMSE expansion valid for any dimension d, any order K, and any derivative m. Under similar conditions, analogous results restricted to $d > 1$ with $[m] = 0$ or $d = 1$ with $[m] = m \geq 0$ are given by Ruppert and Wand (1994) for conventional local polynomial estimators and for $d = 1$ with $[m] = m \geq 0$ by Huang (2003) and Zhou and Wolfe (2000) for regression splines.

We leave the exact expressions of the constants $\mathscr{V}_{K,d,m}$ and $\mathscr{B}_{K,d,m}$ for the general case in the Appendix as they are notationally cumbersome. These expressions simplify considerably for interesting special cases. Specifically, consider estimating $\mu(x)$, i.e., $[m] = 0$. While $\mathscr{B}_{K,d,0}$ remains cumbersome (see Eqn. (B.6)), the variance constant reduces to

$$\mathscr{V}_{K,d,0} = \frac{\dim(R(\cdot))}{\text{vol}(\mathcal{X})} \int_{\mathcal{X}} \frac{\sigma^2(x)}{f(x)} w(x)dx,$$

for any d and K. If, in addition, we restrict attention to the univariate case,

$$\mathscr{B}_{K,1,0} = \frac{\mathrm{vol}(\mathcal{X})^{2K}}{2^{2K+1}(K!)^2}\left(\frac{2}{1+2K} - P_K' Q_K^{-1} P_K\right)\int_{\mathcal{X}} \left(\partial^K \mu(x)\right)^2 w(x)dx,$$

where $P_K = \int_{-1}^{1} R(x)x^K dx$ and $Q_K = \int_{-1}^{1} R(x)R(x)'dx$. Alternatively, for the piecewise constant fit in the multivariate case, we obtain the tidy expression

$$\mathscr{B}_{1,d,0} = \frac{1}{12}\sum_{\ell=1}^{d} |\mathcal{X}_\ell|^2 \int_{\mathcal{X}} \left(\frac{\partial \mu(x)}{\partial x_\ell}\right)^2 w(x)dx.$$

In all possible cases, minimization of the general asymptotic IMSE obtained in Theorem III.5 with respect to J_n gives the optimal choice

$$J_n^* = \left\langle (\mathscr{C}_{K,d,m}\, n)^{\frac{1}{d+2K}} \right\rangle, \qquad \mathscr{C}_{K,d,m} = \frac{2(K-[m])\, \mathscr{B}_{K,d,m}}{(d+2[m])\, \mathscr{V}_{K,d,m}},$$

and $\langle \cdot \rangle$ denotes the nearest integer. A feasible plug-in rule can be easily constructed by using preliminary estimators for the unknown objects in $\mathscr{C}_{K,d,m}$.

3.4 Bahadur Representation and Asymptotic Normality

This section studies the asymptotic behavior of partitioning-based estimators of linear functionals of the regression function. We establish a uniform Bahadur representation, asymptotic normality, and consistency of a suitable standard-error estimator, for both regular and irregular (not root-n estimable) estimands. The estimand of interest is given by $\theta = \theta(\mu)$ and we consider the simple plug-in estimator $\hat{\theta} = \theta(\hat{\mu})$. The following assumption characterizes the class of functionals considered.

Assumption III.6. *$\theta(\tilde{\mu}) \in \mathbb{R}$ is linear, and $|\theta(\tilde{\mu})| \leq C \max_{[m]\leq s} \|\partial^m \tilde{\mu}\|_\infty$, for some $C > 0$.*

This assumption restricts the class of functionals to be linear and bounded (i.e., continuous) in the appropriate uniform norm. It is not difficult to extend the results presented here to cover non-linear functionals, although this extension is omitted to conserve space.[6] Many interesting econometric applications are covered by linear

[6] This extension is achieved by a standard "linearization" argument: first the functional is assumed to be differentiable in the appropriate sense (e.g. Frechet differentiable with respect to an appropriate norm), and then rate restrictions are imposed so that the linearization error is asymptotically negligible.

functionals of the regression function. For concreteness, consider the following three examples.[7]

Example 6. Pointwise Inference $\theta_{1,m}(\mu) = \partial^m \mu(x)$, $m \in \mathbb{Z}_+^d$, $[m] < K$, where differentiation is defined in Eqn. (3.2). This irregular estimand is useful for nonparametric inference for the regression function and its derivatives. ■

Example 7. Partial and Full Means $\theta_{2,\delta}(\mu) = \int_{\times_{\ell=1}^{\delta} \mathcal{X}_\ell} \mu(x) f(x_1, \cdots, x_\delta) dx_1 \cdots dx_\delta$, $\delta \leq d$, where components of $x \in \mathcal{X}$ not integrated over are held fixed at some value (we assume the x_ℓ are ordered such that integration is over the first δ covariates). Estimating partial and full means is an important problem in econometrics (see, e.g., Newey (1994b)). It is well known that $\theta_{2,\delta}(\hat{\mu})$ will not be $\sqrt{n}$-consistent unless $\delta = d$, though the convergence rate increases as more regressors are integrated out. ■

Example 8. Weighted Average Derivative $\theta_{3,m}(\mu) = -\int_{\mathcal{X}} \mu(x)(\partial^m w(x))dx$, $[m] = 1$, where $w(x)$ is a continuously differentiable weighting (trimming) function that vanishes outside a compact subset of $\mathcal{X}$. The functional in this example corresponds to the indirect weighted average derivative (integration by parts gives $\theta_{3,m}(\mu) = \int_{\mathcal{X}} (\partial^m \mu(x)) w(x) dx$), and leads to a simpler estimator based on the regression function directly. Estimating weighted average derivatives is a well-studied problem (see, e.g., Stoker (1986)). The conditions on the weighting function $w(x)$ are essential to eliminate the influence of the boundary of the regressors' support, and achieve $\sqrt{n}$-consistency. ■

The first result in this section establishes a uniform Bahadur-type representation for $\theta(\hat{\mu})$. Specifically, we show that the estimator may be represented as an average of independent, conditionally mean-zero random variables forming a triangular array based on certain smoothing weights, plus a remainder that enjoys a particular rate of convergence. This representation facilitates verification of a variety of properties of semiparametric estimators employing the partitioning estimator as a preliminary step.[8]

To describe the result, define $\varepsilon_i = Y_i - \mu(X_i)$, $i = 1, \ldots, n$, and $q_j = \mathbb{P}[X \in P_j]$, $j = 1, \ldots, J_n^d$. Because $q_j \asymp J_n^{-d}$ by Assumption III.1(d), q_j captures the rate of convergence of each cell (as well as the local behavior of $f(x)$ in each cell). The

[7] For other examples of linear (and non-linear) functionals of interest see, e.g., Andrews (1991), Newey (1997), Chen (2007), Ichimura and Todd (2007), and references therein.

[8] For a recent detailed discussion of the applicability of the Bahadur representation to semiparametric inference, and such a result for kernel-based local polynomials, see Kong, Linton, and Xia (2010).

Bahadur representation of the partitioning-based estimator is then given by:

$$\theta(\hat{\mu}) - \theta(\mu) = \frac{1}{n}\sum_{i=1}^{n} \Psi_n(X_i)\varepsilon_i + \theta(\nu_n), \qquad \Psi_n(z) = \sum_{j=1}^{J_n^d} \Theta_j'\Omega_j^{-1}R_j(z)/q_j, \tag{3.3}$$

with $\Theta_j = (\theta([R_j(\cdot)]_1), \ldots, \theta([R_j(\cdot)]_{\dim(R(\cdot))}))'$, where $[\cdot]_g$ denotes the g^{th} element of a vector, and $\Omega_j = \mathbb{E}\left[R_j(X)R_j(X)'\right]/q_j$.

The smoothing weight $\Psi_n(x)$ is a nonrandom function which varies with n only through the partitioning scheme. By linearity of the functional, the Bahadur representation for $\hat{\mu}$ automatically yields the result for $\theta(\hat{\mu})$ in Eqn. (3.3). To be concrete, in the appendix we first write $\hat{\mu}(x) - \mu(x) = \sum_{i=1}^{n} \psi_n(x, X_i)\varepsilon_i/n + \nu_n(x)$ with $\psi_n(x, z) = \sum_{j=1}^{J_n^d} R_j(x)'\Omega_j^{-1}R_j(z)/q_j$ and a remainder $\nu_n(x)$, and then obtain the smoothing weight and remainder in Eqn. (3.3) by applying the functional $\theta(\cdot)$ to ψ_n and ν_n, respectively: $\Psi_n(z) = \theta(\psi_n(\cdot, z))$ and $\theta(\nu_n)$. The following theorem characterizes the uniform convergence rate of $\theta(\nu_n)$.[9]

Theorem III.7. *Let Assumption III.6 hold with $s \leq S \wedge (K-1)$, and consider the representation in Eqn.* (3.3)*. If the conditions of Theorem III.3 hold, then:*

$$\theta(\nu_n) = O_p\left(\frac{J_n^{(2-\xi/2)d+s}\log(J_n^d)^{1+\xi/2}}{n^{3/2}} + \frac{J_n^{d+s}}{n} + J_n^{-((S+\alpha)\wedge K - s)} \right).$$

Before stating the asymptotic normality result, it is helpful to first discuss an asymptotic variance formula, which also captures the rate of convergence in general. To this end, define

$$V_n = \mathbb{E}\left[\Psi_n(X)^2\sigma^2(X)\right] = \sum_{j=1}^{J_n^d} \Theta_j'\Omega_j^{-1}\Gamma_j\Omega_j^{-1}\Theta_j/q_j, \tag{3.4}$$

with $\Gamma_j = \mathbb{E}\left[R_j(X)R_j(X)'\sigma^2(X)\right]/q_j$. Since a linear least squares estimate is computed within each cell, the asymptotic variance is of the Huber-Eicker-White heteroskedasticity robust form. A plug-in sample analogue of V_n is given by

$$\begin{aligned} \hat{V}_n &= \frac{1}{n}\sum_{i=1}^{n}(\hat{\Psi}_n(X_i)\hat{\varepsilon}_i)^2 = \sum_{j=1}^{J_n^d} \mathbb{1}_{n,j}\Theta_j'\hat{\Omega}_j^{-1}\hat{\Gamma}_j\hat{\Omega}_j^{-1}\Theta_j/q_j, \qquad \hat{\varepsilon}_i = Y_i - \hat{\mu}(X_i) \\ \hat{\Omega}_j &= \frac{1}{n}\sum_{i=1}^{n} R_j(X_i)R_j(X_i)'/q_j, \qquad \hat{\Gamma}_j = \frac{1}{n}\sum_{i=1}^{n} R_j(X_i)R_j(X_i)'\hat{\varepsilon}_i^2/q_j. \end{aligned} \tag{3.5}$$

[9] An almost sure version of this theorem is available in the appendix.

Notice that q_j is artificially introduced to take explicit account for the convergence rate of each sample (and population) average. These quantities are unknown, but they exactly cancel out in the formulation above, leading to a feasible estimator of the large-sample variance.

Theorem III.8. *Suppose the conditions of Theorem III.7 hold with $\eta \geq 0$, that $\sigma^2(x)$ is bounded away from zero on $\mathcal{X}$, and $\theta(\nu_n) = o_p(\sqrt{V_n}/\sqrt{n})$.*

(a) For $\eta > 0$, if $0 < \|\Psi_n\|_2 \to \infty$ and $\|\Psi_n\|_{2+\eta}/\|\Psi_n\|_2 = o(n^{\eta/(4+2\eta)})$, then:

$$\frac{\sqrt{n}(\theta(\hat{\mu}) - \theta(\mu))}{\sqrt{V_n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\Psi_n(X_i)\varepsilon_i}{\sqrt{V_n}} + o_p(1) \to_d \mathcal{N}(0,1),$$

If, in addition, $\|\hat{\mu} - \mu\|_\infty = o_p(1)$, then $\hat{V}_n/V_n \to_p 1$.

(b) If $\|\Psi_n - \Psi\|_2 \to 0$, $0 < \|\Psi\|_2 < \infty$, and $\theta(\mu) = \mathbb{E}[\Psi(X)\mu(X)]$, then:

$$\frac{\sqrt{n}(\theta(\hat{\mu}) - \theta(\mu))}{\sqrt{V_n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\Psi(X_i)\varepsilon_i}{\sqrt{V}} + o_p(1) \to_d \mathcal{N}(0,1),$$

and $V_n \to V = \mathbb{E}[\Psi(X)^2\sigma^2(X)]$. If, in addition, $\|\hat{\mu} - \mu\|_\infty = o_p(1)$, then $\hat{V}_n/V_n \to_p 1$.

This result gives simple and intuitive sufficient conditions for asymptotic normality of a partitioning-based plug-in estimator of $\theta = \theta(\mu)$, and for consistency of a suitable standard-error estimator. The theorem is divided in two parts, which are mutually exclusive, depending on the asymptotic behavior of the smoothing weights in the Bahadur representation. This approach is similar in spirit to the central limit theorems of Newey (1997) for series estimators (compare to his Assumptions 6 and 7), but using the Bahadur representation we put simple sufficient conditions directly on the smoothing weights. These results automatically apply to vector-valued estimands, although we restrict θ to be scalar for simplicity.

The distinctive feature separating the cases is mean-square continuity of the functional $\theta(\cdot)$ and its Riesz representation (see, e.g., van der Vaart (1991)). These conditions are not imposed in Theorem III.8(a), so the estimand is irregular, and the CLT is obtained by directly exploiting the triangular array structure of the Bahadur representation. In contrast, in Theorem III.8(b) these conditions imply that the estimand is $\sqrt{n}$-consistent and asymptotically linear with influence function $\psi_i = \Psi(X_i)\varepsilon_i$, which permits an easy characterization of the asymptotic variance. This case is important because it gives easy-to-verify sufficient conditions for asymptotic linearity.

The high-level conditions in Theorem III.8 need to be verified in each application (i.e. for a particular $\theta(\cdot)$). We demonstrate the applicability of this theorem by returning to the three examples introduced above, and giving simple primitive conditions under which the high-level conditions hold for the partitioning plug-in estimator.

Example 9. Pointwise Inference (continued) Suppose the conditions of Theorem III.3 hold with $\eta > 0$ and the partition satisfies $J_n^{(2-\xi)d} \log(J_n^d)^{1+\xi/2} = o(n)$ and $\sqrt{n} J_n^{-d/2-(S+\alpha)\wedge K} \to 0$. Then, for $[m] < K$ the conditions of Theorem III.8(a) are met, as $\|\Psi_n\|_p^p \asymp J_n^{(p-1)d+p[m]}$ and $V_n \asymp J_n^{d+2[m]}$. Therefore, $\partial^m \hat{\mu}(x) = \partial^m \mu(x) + O_p(J_n^{d/2+[m]}/\sqrt{n})$. The rate restrictions are quite mild in this example. Negligibility of the remainder term requires the "variance" condition $J_n^{(3/2-\xi/2)d} \log(J_n^d)^{1+\xi/2} = o(n)$; standard error estimation necessitates the only slightly stronger restriction above. In the case $\xi = \eta = 1$ (in Theorem III.3), the two coincide, giving $J_n^d \log(J_n^d)^{3/2} = o(n)$, and only three bounded moments are assumed. As a comparison, the central limit theorem of Newey (1997) for regression splines requires the analogue of $J_n^{2d}/n \to 0$ and $\sqrt{n} J_n^{-(S+\alpha)\wedge K} \to 0$, and assumes four bounded moments. These improvements are due to the fact that we are able to exactly characterize the convergence rate of V_n, and to the faster rates of convergence and weaker rate restrictions obtained in the previous section for partitioning estimators. ■

Example 10. Partial and Full Means (continued) Begin with the irregular case ($\delta < d$). Suppose the conditions of Theorem III.3 hold with $\eta > 0$ and the partition satisfies $J_n^{[(3-\xi)d+\delta]/2} \log(J_n^d)^{1+\xi/2} = o(n)$ and $\sqrt{n} J_n^{-(d-\delta)/2-(S+\alpha)\wedge K} \to 0$. The conditions of Theorem III.8(a) are met as $\|\Psi_n\|_p^p \asymp J_n^{(p-1)(d-\delta)}$ and hence $V_n \asymp J_n^{d-\delta}$. For some values of δ and ξ, this may imply $\|\hat{\mu} - \mu\|_\infty \to_p 0$, otherwise the exponent on J_n must be (slightly) increased to $(2-\xi)d + \delta/2$. These rate restrictions are strengthened by $J_n^{\delta/2}$ compared to the pointwise case, exactly the decrease in the order of the variance. As δ increases to d, the rate of the variance decreases, leading to the rate of convergence $\theta_{2,\delta}(\hat{\mu}) = \theta_{2,\delta}(\mu) + O_p(J_n^{(d-\delta)/2}/\sqrt{n})$, which shows that the estimator is $\sqrt{n}$-consistent only in the full mean case. In this case, $\theta_{2,d}(\mu) = \int_{\mathcal{X}} \mu(x) f(x) dx = \mathbb{E}[\Psi(X)\mu(X)]$, with $\Psi(x) = 1$. Moreover, $\Psi_n(x) = \sum_{j=1}^{J_n^d} e_1' R_j(x) = 1$, and hence $\|\Psi_n - \Psi\|_2 = 0$, which verifies the conditions in Theorem III.8(b). ■

Example 11. Weighted Average Derivative (continued) Suppose the conditions of Theorem III.3 hold and the partition satisfies $J_n^{(2-\xi/2)d} \log(J_n^d)^{1+\xi/2} = o(n)$ and $\sqrt{n} J_n^{-(S+\alpha)\wedge K} \to 0$. Then, the conditions of Theorem III.8(b) hold and uniform consistency of $\hat{\mu}(x)$ is implied. Specifically, note that $\theta_{3,m}(\mu) = \int_{\mathcal{X}} \mu(x)(\partial^m w(x)) dx =$

$\mathbb{E}[\Psi(X)\mu(X)]$, with $\Psi(x) = -f(x)^{-1}\partial^m w(x)$, and hence

$$\Psi_n(x) = \sum_{j=1}^{J_n^d} R_j(x)'\Omega_j^{-1}\mathbb{E}[R_j(X)\Psi(X)]/q_j.$$

Under an appropriate smoothness assumption, there will exist $\{\gamma_j^0\}$ such that

$$\max_{1 \leq j \leq J_n^d} \|\mathbb{1}_{P_j}(\cdot)\Psi(\cdot) - R_j(\cdot)'\gamma_j^0\|_\infty = o(1)$$

, yielding the mean-square convergence condition. Hence $\theta_{3,m}$ will be $\sqrt{n}$-consistent. ■

It is important to mention that Theorems III.7 and III.8 (and the examples discussed above) are established using uniform norms, which leads to the simple and general sufficient conditions above. In some examples, however, it is possible to improve on these sufficient conditions by relying on the (weaker) L_2 norm. For instance, if the linear functional is continuous with respect to the L_2 norm (and hence regular), then it is possible to improve on the rate restrictions of Theorem III.8 by relying on sharper rates on the remainder of the Bahadur representation. In the specific case of partitioning estimators, because of the sharp uniform rates obtained in this paper the difference between the mean-square and uniform convergence rates is only a slow-varying function (i.e., $\log(J_n^d)$) under appropriate moment assumptions, and hence using the stronger uniform norm is not too restrictive.

3.5 Monte Carlo Evidence

We report a subset of the results from an extensive Monte Carlo study that we conducted to explore the finite-sample performance of the partitioning estimator in comparison to local polynomials and regression splines. We focused on estimating the regression function $\mu(x)$, and examined two measures of global accuracy, root integrated mean-square error (MSE) and integrated mean absolute error (MAE), as well as root mean-square error (RMSE) at interior and boundary points of $\mathcal{X}$. The full set of results from our simulation study is available in the online supplement, and includes different sample sizes, dimensions and distributions for the covariates, regression functions, and levels of variability. In addition, as a complement to the nonparametric results presented here, Cattaneo and Farrell (2011b) report another extensive simulation study employing the partitioning estimator as a preliminary estimator in semiparametric treatment effect estimation.

We generated 5,000 simulated data sets according to $Y_i = \mu(X_{i,1}, X_{i,2}) + \varepsilon_i,\ i = 1, \ldots, n$, with $\varepsilon_i \sim_{\text{iid}} \mathcal{N}(0, \sigma^2)$. The covariates are independently distributed as truncated Beta(B, B) distributions. We set $\sigma^2 = 1$ and consider both $B = 1$ (uniform) and $B = 1/2$ (which places more mass at the boundary), and truncate to $[0.05, 0.95]$. We discuss only four different specifications for the regression function $\mu(x_1, x_2)$ in this section:

$$
\begin{aligned}
\text{Model 1: } \mu(x_1, x_2) &= 0.7 \exp\left\{-3\left((4x_1 - 2 + 0.8)^2 + 8(x_2 - 1/2)^2\right)\right\} \\
&\qquad\qquad + \exp\left\{-3\left((4x_1 - 2 - 0.8)^2 + 8(x_2 - 1/2)^2\right)\right\}, \\
\text{Model 2: } \mu(x_1, x_2) &= \sin(5x_1)\sin(10x_2), \\
\text{Model 3: } \mu(x_1, x_2) &= \left((1 - (4x_1 - 2)^2)^2\right)(\sin(5x_2)/5), \\
\text{Model 4: } \mu(x_1, x_2) &= \mathbb{1}\{(4x_1 - 2) \in [-2, 1]\}((4x_1 - 2)^7 - 19)/20 \\
&\quad - \mathbb{1}\{(4x_1 - 2) \in (-1, 0]\}(4x_1 - 2)^2 \\
&\quad + \mathbb{1}\{(4x_1 - 2) \in (0, 1/2]\}(4x_1 - 2)^4/2 \\
&\quad + \mathbb{1}\{(4x_1 - 2) \in (1/2, 1]\}(4x_1 - 2)^5 \\
&\quad + \mathbb{1}\{(4x_1 - 2) \in (1, 2]\}(2 - (4x_1 - 2)^3) \\
&\quad + 4.26\left(\exp(-3x_2) - 4\exp(-6x_2) + 3\exp(-9x_2)\right).
\end{aligned}
$$

These bivariate regression functions are graphed in Figure 3.1. These models are adapted from Fan and Gijbels (1996, Chapter 7.5), Braun and Huang (2005) and Eggermont and LaRiccia (2009, Chapter 22.1) to the simulation setup consider here. Model 4 is discontinuous, but we nonetheless include it as another potentially interesting case for comparison.

For all three nonparametric estimators, we use linear and cubic polynomials (i.e. $K = 2$ and $K = 4$ in our notation). We employ a product Epanechnikov kernel with a common bandwidth for local polynomials, and a tensor product of B-splines for regression splines. The tuning parameters are chosen as follows: for local polynomials, the bandwidth is chosen to minimize the asymptotic conditional IMSE derived by Ruppert and Wand (1994); for the partitioning estimator, we use J_n^* defined above; and for regression splines, we use $J_n^* + 1$ knots for each covariate, also placed uniformly in the support. The final choice may not be optimal for regression splines, but is made for two reasons. First, it permits a direct comparison between partitioning and splines, highlighting the role of the inherent discontinuities of the partitioning estimator. Second, direct plug-in rules for splines are available only for special cases.

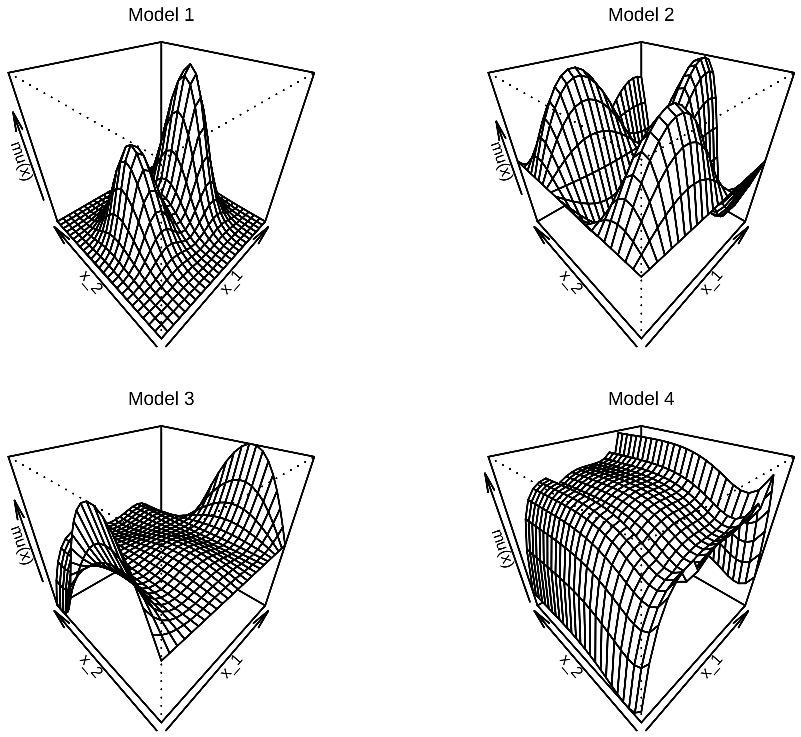


Figure 3.1: Regression functions for simulations.

We use both infeasible and feasible tuning parameters selectors, where the data-driven selectors were implemented by extending the procedure outlined by Fan and Gijbels (1996, Section 4.2) to $d = 2$. The infeasible tuning parameter formula is invalid for Model 4, and thus we set $h_n^* = 1/3$ for local polynomials and $J_n^* = 3$ for partitioning and regression splines. We employed the feasible selectors for all models (e.g., ignoring the lack of continuity in Model 4).

We report here only the case $n = 1,000$, presented in Tables 3.1 and 3.2. The former shows results for $B = 1/2$, followed by the uniform case. The first two columns show the infeasible tuning parameter and the rounded mean feasible choice across simulations. In general, no estimator dominates the others and hence an absolute ranking does not emerge from this simulation study. The partitioning estimator is on par with the other two estimators in many cases, by any of the accuracy measures. In general the global measures are not particularly useful to rank these estimators, and although it appears that local polynomials perform better “on average”, the differences are usually small. The partitioning estimator often outperforms the others at the point $(x_1, x_2) = (0.1, 0.1)$, indicating good boundary performance.

The discontinuities of the partitioning estimator require further discussion. By comparing to B-splines, we observe that according to the global accuracy measures these discontinuities do not appear to have a deleterious effect: the partitioning estimator is often on par with splines, and occasionally more accurate. The discontinuities are measure zero, so this may not be surprising, but it shows that the asymptotics provide a good finite-sample approximation in this case. Local polynomials tend to slightly outperform both B-splines and partitioning estimators in terms of these global measures, but not in all cases. The pointwise results are also mixed, depending on the data generating process and the evaluation point considered. For instance, in Model 2 at the point $(x_1, x_2) = (1/2, 1/2)$ with $B = 1/2$, the partitioning estimator performs poorly (though the feasible local polynomial estimator is hardly better), except for the infeasible linear fit. This suggests that a practitioner interested in inference at a particular point should not place a cell boundary at that point. Nonetheless, in other cases the partitioning estimator outperforms its continuous counterpart, even in point estimation. The partitioning estimator performs very well in the discontinuous Model 4, even though cell boundaries are not placed at the discontinuities.

3.6 Conclusion

This paper aimed to give a thorough asymptotic treatment of partitioning estimators of the regression function, its derivatives, and functionals thereof. We established (optimal) mean-square and uniform rates of convergence, a general conditional IMSE expansion and an optimal plug-in rule to select the number of cells, and finally a Bahadur representation and asymptotic normality for linear functionals of the estimator, with valid standard-error estimates. We also showed how these results apply to a few examples, and performed an extensive simulation study.

This estimation strategy appears to have some advantages and disadvantages when compared to other popular nonparametric procedures. Indeed, one goal of this paper was to provide a comprehensive analysis of the partitioning estimators to permit formal comparison to different nonparametric procedures, as we discussed along the manuscript. While the partitioning estimator is simple and intuitive (and has been proposed in the econometric literature before), it has the perhaps unappealing feature of being discontinuous in finite samples. This property is also shared by the popular nearest-neighbor estimator (e.g., Györfi, Kohler, Krzyżak, and Walk (2002, Chapter 6)), but not by the conventional series- and kernel-based estimators in usual cases. Thus, while we view this estimator as potentially useful in applications (e.g., as a preliminary exploratory device), it is important to highlight that it does ignore the underlying smoothness of the regression function when constructing the estimate. From a theoretical perspective, it is nonetheless interesting that imposing such smoothness is not needed to construct a nonparametric regression estimator that achieves the usual optimal rates of convergence. Moreover, this result shows that the partitioning estimator is not overfitting, even though it enjoys more degrees-of-freedom by not imposing smoothness restrictions as other estimators do (e.g., regression splines).

Table 3.1: Error Comparisons for Local Polynomials, B-Splines, and Partitioning Estimators

	Tuning Parameter		Root Integrated MSE		Ingetrated MAE		Point Estimation RMSE					
							(0.5,0.5)		(0.1,0.5)		(0.1,0.1)	
Degree:	Linear	Cubic	Linear	Cubic	Linear	Cubic	Linear	Cubic	Linear	Cubic	Linear	Cubic
Model 1, $X_{i,\ell} \sim \beta(0.5, 0.5)$												
Infeasible Estimation												
Local Polynomial	0.17	0.24	0.160	0.190	0.123	0.147	0.168	0.188	0.151	0.186	0.178	0.216
B-splines	9	4	0.174	0.172	0.134	0.133	0.215	0.235	0.111	0.162	0.161	0.167
Partitioning	9	4	0.179	0.205	0.138	0.156	0.180	0.662	0.137	0.388	0.151	0.169
Feasible Estimation												
Local Polynomial	0.28	0.27	0.199	0.201	0.146	0.155	0.082	0.148	0.095	0.114	0.083	0.095
B-splines	4	1	0.167	0.176	0.125	0.137	0.416	0.244	0.160	0.124	0.139	0.163
Partitioning	4	1	0.177	0.174	0.133	0.133	0.317	0.256	0.240	0.183	0.136	0.159
Model 2, $X_{i,\ell} \sim \beta(0.5, 0.5)$												
Infeasible Estimation												
Local Polynomial	0.12	0.18	0.204	0.232	0.160	0.180	0.203	0.214	0.190	0.267	0.182	0.321
B-splines	16	4	0.209	0.172	0.165	0.132	0.411	0.151	0.270	0.155	0.176	0.169
Partitioning	16	4	0.252	0.217	0.196	0.165	0.573	0.824	0.383	0.396	0.166	0.171
Feasible Estimation												
Local Polynomial	0.31	0.23	0.487	0.489	0.390	0.387	0.702	0.748	0.602	0.604	0.277	0.284
B-splines	4	4	0.343	0.172	0.275	0.132	0.171	0.151	0.146	0.155	0.226	0.169
Partitioning	4	4	0.373	0.217	0.304	0.165	0.869	0.824	0.298	0.396	0.308	0.171
Model 3, $X_{i,\ell} \sim \beta(0.5, 0.5)$												
Infeasible Estimation												
Local Polynomial	0.15	0.37	0.167	0.156	0.128	0.119	0.142	0.107	0.162	0.132	0.186	0.169
B-splines	16	4	0.176	0.158	0.136	0.121	0.284	0.151	0.220	0.155	0.175	0.167
Partitioning	16	4	0.233	0.204	0.180	0.155	0.433	0.681	0.305	0.388	0.174	0.170
Feasible Estimation												
Local Polynomial	0.33	0.27	0.259	0.265	0.164	0.172	0.113	0.119	0.233	0.242	0.179	0.190
B-splines	4	4	0.187	0.166	0.144	0.128	0.220	0.151	0.142	0.145	0.138	0.166
Partitioning	4	4	0.199	0.195	0.155	0.148	0.214	0.550	0.250	0.329	0.124	0.177
Model 4, $X_{i,\ell} \sim \beta(0.5, 0.5)$												
Infeasible Estimation												
Local Polynomial	0.33	0.33	0.666	0.318	0.478	0.242	0.140	0.176	0.307	0.148	0.200	0.194
B-splines	9	9	0.726	0.340	0.575	0.268	0.223	0.276	0.166	0.181	0.198	0.191
Partitioning	9	9	0.652	0.347	0.511	0.266	0.151	0.234	0.205	0.236	0.177	0.210
Feasible Estimation												
Local Polynomial	0.19	0.26	0.636	0.537	0.523	0.442	0.401	0.409	0.661	0.367	0.595	0.897
B-splines	4	1	0.794	0.627	0.616	0.505	0.692	0.364	0.171	0.180	0.165	0.208
Partitioning	4	1	0.784	0.621	0.603	0.497	0.748	0.436	0.208	0.216	0.161	0.177

Notes. Tuning parameters are local polynomial bandwidth and the number of cells for partitioning estimation and B-splines, as described in the text. Feasible tuning parameters reported are the (rounded) mean of all estimated values. Integrated MSE and MAE are estimated by averaging over the design points in each simulated data set.

Table 3.2: Error Comparisons for Local Polynomials, B-Splines, and Partitioning Estimators

	Tuning Parameter		Root Integrated MSE		Ingetrated MAE		Point Estimation RMSE					
							(0.5,0.5)		(0.1,0.5)		(0.1,0.1)	
Degree:	Linear	Cubic	Linear	Cubic	Linear	Cubic	Linear	Cubic	Linear	Cubic	Linear	Cubic
Model 1, $X_{i,\ell} \sim \beta(1,1)$												
Infeasible Estimation												
Local Polynomial	0.16	0.23	0.166	0.191	0.126	0.144	0.149	0.160	0.175	0.195	0.258	0.275
B-splines	9	4	0.186	0.175	0.146	0.133	0.212	0.203	0.129	0.181	0.216	0.246
Partitioning	9	4	0.183	0.205	0.142	0.156	0.164	0.570	0.150	0.398	0.193	0.236
Feasible Estimation												
Local Polynomial	0.26	0.26	0.215	0.213	0.165	0.170	0.059	0.127	0.104	0.121	0.099	0.108
B-splines	4	1	0.173	0.187	0.133	0.146	0.397	0.239	0.204	0.150	0.173	0.234
Partitioning	4	1	0.181	0.184	0.139	0.143	0.341	0.266	0.287	0.205	0.159	0.208
Model 2, $X_{i,\ell} \sim \beta(1,1)$												
Infeasible Estimation												
Local Polynomial	0.12	0.18	0.204	0.227	0.156	0.171	0.174	0.174	0.208	0.266	0.263	0.400
B-splines	16	4	0.205	0.171	0.159	0.128	0.369	0.129	0.286	0.177	0.249	0.247
Partitioning	16	4	0.252	0.216	0.196	0.164	0.515	0.690	0.412	0.407	0.223	0.238
Feasible Estimation												
Local Polynomial	0.28	0.23	0.505	0.508	0.411	0.411	0.677	0.704	0.568	0.570	0.315	0.324
B-splines	4	4	0.327	0.171	0.257	0.128	0.168	0.129	0.171	0.177	0.329	0.247
Partitioning	4	4	0.362	0.216	0.292	0.164	0.755	0.690	0.318	0.407	0.428	0.238
Model 3, $X_{i,\ell} \sim \beta(1,1)$												
Infeasible Estimation												
Local Polynomial	0.16	0.38	0.161	0.148	0.120	0.110	0.113	0.091	0.175	0.138	0.263	0.206
B-splines	9	4	0.176	0.159	0.136	0.118	0.115	0.130	0.148	0.177	0.220	0.246
Partitioning	9	4	0.194	0.203	0.150	0.154	0.103	0.572	0.160	0.399	0.196	0.239
Feasible Estimation												
Local Polynomial	0.33	0.27	0.211	0.218	0.133	0.140	0.103	0.103	0.216	0.220	0.166	0.171
B-splines	4	4	0.172	0.164	0.128	0.123	0.174	0.129	0.162	0.165	0.170	0.241
Partitioning	4	4	0.182	0.190	0.137	0.142	0.181	0.451	0.238	0.333	0.139	0.231
Model 4, $X_{i,\ell} \sim \beta(1,1)$												
Infeasible Estimation												
Local Polynomial	0.33	0.33	0.613	0.322	0.424	0.246	0.136	0.160	0.194	0.149	0.225	0.228
B-splines	9	9	0.677	0.347	0.508	0.275	0.139	0.256	0.152	0.190	0.296	0.276
Partitioning	9	9	0.619	0.350	0.465	0.269	0.134	0.189	0.182	0.249	0.236	0.281
Feasible Estimation												
Local Polynomial	0.19	0.26	0.603	0.512	0.484	0.408	0.353	0.359	0.554	0.329	0.708	0.942
B-splines	4	1	0.727	0.543	0.527	0.418	0.515	0.304	0.254	0.195	0.253	0.319
Partitioning	4	1	0.723	0.524	0.522	0.393	0.557	0.471	0.304	0.293	0.257	0.259

Notes. Tuning parameters are local polynomial bandwidth and the number of cells for partitioning estimation and B-splines, as described in the text. Feasible tuning parameters reported are the (rounded) mean of all estimated values. Integrated MSE and MAE are estimated by averaging over the design points in each simulated data set.

CHAPTER IV

Efficient Estimation of the Dose-Response Function under Ignorability using Subclassification on the Covariates

4.1 Introduction

Treatment effect models are a prime example of a missing data problem. Units are assumed to have a collection of distinct random potential outcomes but, depending on their treatment status, only one of these outcomes is observed. The population parameters of interest in these models are usually some feature of the marginal distributions of the potential outcomes such as the means or quantiles. These parameters, however, are not identifiable from a random sample of observed outcomes and treatment statuses without further assumptions because of the potential presence of selectivity bias; a non-random missing data problem. A common identifying assumption in these models is called Ignorability, which includes a key restriction on the data generating process known as unconfoundedness or selection on observables. This assumption imposes random missing data after conditioning on a set of predetermined always-observed covariates, and permits the development of flexible inference procedures by first working conditionally on the covariates and then averaging out appropriately.

In the context of finite multi-valued treatment effects, a simple and interesting estimand is the Dose-Response Function (DRF), which describes the mean effect of each treatment level on the outcome of interest.[1] Under Ignorability, many different semiparametric estimators for the DRF may be constructed using flexible approaches,

[1] See, e.g., (Imbens 2000), (Lechner 2001), (Imai and van Dyk 2004), (Cattaneo 2010), and references therein.

including nonparametric regression methods, matching techniques, inverse probability weighting schemes, procedures based on the estimated (generalized) propensity score, and hybrid procedures that combine some of these techniques.[2] These estimators, which include a preliminary nonparametric estimator, are well known to be root-n consistent (where n is the sample size) and asymptotically normal under appropriate regularity conditions, provided certain restrictions on the tuning and smoothing parameters involved in the estimation are satisfied. In most cases these estimators are also asymptotically linear and semiparametric efficient.

This chapter develops a new semiparametric efficient estimator of the DRF based on the idea of subclassification, blocking, or stratification on the observed predetermined covariates. The estimator proceeds by first dividing the support of the observed covariates into disjoint cells, also called blocks or stratums, then carrying on inference using only observations within each cell, and finally averaging out appropriately. Intuitively, for cells "small enough," the potential outcomes within each cell are approximately missing completely at random by virtue of Ignorability, leading to a consistent, asymptotically linear, and semiparametric efficient estimator under conventional regularity conditions. Moreover, using this idea we also develop a simple and intuitive consistent standard-error estimator, leading to asymptotically valid confidence intervals for the population parameter of interest.

The idea behind the semiparametric estimator discussed in this chapter may be traced back to the early work of (Cochran 1968), who informally discusses the idea of subclassification with a univariate continuous covariate in observational studies. In this chapter we formally derive a first-order, asymptotically linear large sample approximation for a class of subsclassification-based semiparametric estimators that allow for an arbitrary number of continuous covariates as well as an arbitrary large polynomial of approximation within each cell. These results are also connected to the work of Rosenbaum and Rubin (1983, 1984), who discuss inference by subclassifying observations based on the estimated propensity score in observational studies. In this chapter, however, subclassification is done directly on the observed covariates rather than on the estimated (generalized) propensity score, thereby avoiding preliminary nonparametric estimation of the propensity score and the related technical issues of generated regressors and random denominators. The ongoing work of Cattaneo, Imbens, Pinto, and Ridder (2009) addresses the delicate issue of subclassification-

[2] For a review on the program evaluation and missing data literatures, see, e.g., Chen, Hong and Tarozzi (2004, 2008), (Heckman and Vytlacil 2007), (Imbens and Wooldridge 2009), (Bang and Robins 2005), (Tsiatis 2006), (Wooldridge 2007), and references therein.

based inference using the estimated propensity score. The results in this chapter can be viewed as a first step toward developing the theoretical properties of such a procedure by considering the "known (generalized) propensity score" case, since a known propensity score may be treated as a univariate observed covariate and the results herein apply immediately.[3]

The subclassification-based estimator studied in this chapter may also be viewed as a two-step semiparametric estimator that depends on a special nonparametric procedure called Partitioning. In this chapter we exploit this idea, together with some of the recently developed asymptotic results presented in (Cattaneo and Farrell 2013) for nonparametric partitioning estimators, to provide sufficient conditions for the efficient semiparametric estimation of the DRF, and to construct simple and easy-to-implement consistent standard-error estimators. We assess the performance of these large sample approximations in a Monte Carlo experiment.

The rest of the chapter is organized as follows. Section 2 introduces the multi-valued treatment effect model, discusses identification, and describes (both intuitively and formally) the subclassification-based semiparametric estimator. Section 3 develops the asymptotic properties of this estimator, while Section 4 reports the main results of a simulation study. Finally, Section 5 summarizes the work presented here and discusses possible extensions. All technical derivations are contained in the Appendix.

4.2 Model, Identification and Estimator

This chapter focuses on the estimation of the Dose-Response Function in the context of a (finite) multi-valued treatment effect model. Suppose that $(Y_i, X_i', T_i)'$, $i = 1, 2, \cdots, n$, is an observed random sample, where Y_i is an outcome variable, $X_i \in \mathcal{X} \subset \mathbb{R}^d$ is a vector of continuous covariates, and $T_i \in \mathcal{T} = \{0, \cdots, \tau\}$ with $\mathcal{T}$ a finite set of treatments or groups. The procedure discussed below may be easily generalized to allow for discrete covariates by computing the estimator for each fixed distinct combination, and then averaging out appropriately, as it is standard in the literature. However, to simplify the discussion (and notation) we consider only continuous predetermined covariates. The outcome variable Y_i is assumed to satisfy $Y_i = D_{0,i}Y_i(0) + \cdots + D_{\tau,i}Y_i(\tau)$, where $D_{t,i} = \mathbf{1}(T_i = t)$, $t = 0, \cdots, \tau$, is a treatment or group indicator, and $Y_i(0), \cdots, Y_i(\tau)$ are $\tau + 1$ random potential outcomes.

[3] For further discussion on estimators combining subclassification and regression see, e.g., (Imbens 2004) and (Imbens and Wooldridge 2009).

($\mathbf{1}(\cdot)$ denotes the indicator function.) For each unit $i = 1, \cdots, n$, only one of the $\tau + 1$ potential outcomes is observed, according to the value of T_i. This leads to the fundamental problem of causal inference in the context of program evaluation (e.g., (Holland 1986)), a classical missing data problem.

The estimand of interest is the DRF given by $\mu = (\mu_0, \cdots, \mu_\tau)'$ with $\mu_t = \mathbb{E}[Y_i(t)]$. More general estimands are briefly discussed in Section 5, which summarizes potential extensions to the work undertaken in this chapter. Because all but one of the potential outcomes are missing for each unit, μ is not identifiable from the data without further assumptions. The following identification assumption is commonly used in the missing data and program evaluation literatures.

Assumption 1. (**Weak Ignorability**) For all $t \in \mathcal{T}$:

(a) $Y_i(t) \perp\!\!\!\perp D_{t,i} \mid X_i$.

(b) $0 < e_{\min} \leq \mathbb{P}[T_i = t | X_i]$.

Assumption 1(a) corresponds to a (weak) version of unconfoundedness or selection on observables, and implies that after conditioning on the observed covariates missing data occurs completely at random. This assumption is strong, but commonly employed in the literature. Assumption 1(b) ensures that the generalized propensity score $e_t(x) = \mathbb{P}[T_i = t | X_i = x]$ is bounded away from zero, an important condition for semiparametric efficient estimation. This assumption, and different variations thereof, has been commonly used in the missing data, measurement error, and treatment effect literatures.

Assumption 1 implies that

$$\mu_t = \mathbb{E}[Y_i(t)] = \mathbb{E}\left[\frac{D_{t,i}Y_i}{e_t(X_i)}\right] = \mathbb{E}\left[\frac{\mathbb{E}[D_{t,i}Y_i|X_i]}{e_t(X_i)}\right] = \mathbb{E}[\mathbb{E}[Y_i|T_i = t, X_i]],$$

which leads to a variety of semiparametric plug-in (feasible) estimation approaches for the DRF. These alternative representations motivate inverse probability weighting, imputation, and projection estimation, among other possibilities. For a discussion of these alternative, well-known approaches see, e.g., Chen, Hong and Tarozzi (2004, 2008), (Bang and Robins 2005), (Imbens, Newey, and Ridder 2007), (Tsiatis 2006), (Heckman and Vytlacil 2007), (Imbens and Wooldridge 2009), and references therein. Regardless of the particular identifying approach employed, in all cases at least one nonparametric estimator is required, unless the researcher is willing to impose strong parametric assumptions. Suitable implementations of flexible, semiparametric estimators are available in the literature when using local polynomials (including kernels) or sieves (including series), and these estimators are known to be asymptotically linear and semiparametric efficient under appropriate regularity conditions. (An important alternative estimator is the matching estimator of (Abadie and Imbens 2006) which is not asymptotically linear.)

To motivate the subclassification estimator considered in this chapter, note that if the potential outcomes are assumed to be missing completely at random, that is, if $Y(t) \perp\!\!\!\perp D_t$, then a simple (possibly inefficient) estimator of μ_t is given by

$$\bar{Y}_t = \frac{1}{\bar{W}_t} \sum_{i=1}^{n} D_{t,i} Y_i, \qquad \bar{W}_t = \sum_{i=1}^{n} D_{t,i},$$

which is a simple weighted average of the observed outcomes. However, if the data are not missing completely at random, $\bar{Y}_t$ will be inconsistent for μ_t in general. Nonetheless, Assumption 1 leads to a similar idea based on subclassification on the observed covariates X_i. Suppose that $\mathcal{X}$ is compact and that $\mathcal{P}_n = \{P_j : j = 1, \cdots, J_n^d\}$ is a disjoint partition covering $\mathcal{X}$ with typical cell P_j (implicit dependence on n through the partitioning scheme is suppressed for notational ease). Within each (small) cell P_j of the the partition $\mathcal{P}_n$, Assumption 1 implies that $Y(t)$ is "approximately" independent of D_t, suggesting the following subclassification-based estimator:

$$\hat{\mu}_t = \sum_{j=1}^{J_n^d} \frac{N_j}{n} \bar{Y}_{j,t}, \qquad \bar{Y}_{j,t} = \frac{1}{N_{j,t}} \sum_{i=1}^{n} \mathbf{1}_{P_j}(X_i) D_{t,i} Y_i,$$

$$N_j = \sum_{i=1}^{n} \mathbf{1}_{P_j}(X_i), \qquad N_{j,t} = \sum_{i=1}^{n} \mathbf{1}_{P_j}(X_i) D_{t,i}, \qquad \mathbf{1}_{P_j}(x) = \mathbf{1}(x \in P_j).$$

The "local" estimate $\bar{Y}_{j,t}$ is only well defined when $N_{j,t} > 0$, which is guaranteed in large samples by Assumption 1(b), provided that the cells are not too small. A proper definition of this estimator needs to account for the potential empty cells in finite samples, as done formally below. From an intuitive point of view $N_{j,t}/N_j \approx \mathbb{P}[D_{t,i} = 1 | X_i \in P_j] \approx e_t(X_i)$. Thus, under appropriate regularity conditions and for a fine enough partition, it is natural to expect that

$$\hat{\mu}_t = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J_n^d} \frac{N_j}{N_{j,t}} \mathbf{1}_{P_j}(X_i) D_{t,i} Y_i \approx \mu_t.$$

If all cells of the partition become small as $J_n^d \to \infty$, this subclassification-based estimator may be viewed as a semiparametric estimator given by

$$\hat{\mu}_t = \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_t(X_i), \qquad \hat{\mu}_t(x) = \sum_{j=1}^{J_n^d}\frac{N_j}{N_{j,t}}\mathbf{1}_{P_j}(x)D_{t,i}Y_i,$$

where J_n^d corresponds to the tuning parameter underlying the nonparametric procedure. In fact, $\hat{\mu}_t(x)$ corresponds to a special case of the nonparametric estimator of a regression function known as Partitioning (see, e.g., (Györfi, Kohler, Krzyżak, and Walk 2002, Chapter 4) and (Cattaneo and Farrell 2013)).

Valid first-order, asymptotically linear, semiparametric inference requires a delicate choice of tuning and smoothing parameters so that the higher-order variance and the higher-order bias of the statistic are asymptotically negligible. For the partitioning estimator, J_n^d is the tuning parameter which "controls" the variance of the estimator: the smaller the cells (i.e., the larger J_n^d), the larger the variance. The bias, on the other hand, is (partially) determined by the "quality" of approximation: within each cell, the approximation is based on the sample mean of $D_{t,i}Y_i$, leading to an approximation error proportional to the inverse of the length of the cell. Thus, if bias is a concern, a natural way to improve the approximation is to use a more flexible polynomial in X_i within each block.

These insights lead to the following subclassification-based estimator, which is the main object of study in this chapter. The following notation is needed to formally describe the estimator. For fixed $K \in \mathbb{N}$, let $R(x)$ represent a column vector containing the complete polynomial basis of order $(K-1)$ based on $x \in \mathbb{R}^d$, that is, for $x = (x_1, \cdots, x_d)'$ and $\alpha = (\alpha_1, \cdots, \alpha_d)' \in \mathbb{Z}_+^d$ (a multi-index), with $|\alpha| = \alpha_1 + \cdots + \alpha_d$ and $x^\alpha = x_1^{\alpha_1}\cdots x_d^{\alpha_d}$, each element of $R(x)$ is given by x^α for $\alpha \in \{\mathsf{a} \in \mathbb{Z}_+^d : |\mathsf{a}| \leq K-1\}$. For example, if $d=1$ then $R(x) = (1, x, x^2, \cdots, x^{K-1})'$. Within each cell P_j, the basis is denoted by $R_j(x) = \mathbf{1}(x \in P_j)R(x)$. Using this notation, a subclassification-based estimator (of order $K-1$) is given by

$$\hat{\mu}_t = \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_t(X_i), \qquad \hat{\mu}_t(x) = \sum_{j=1}^{J_n^d} R_j(x)'\hat{\beta}_j, \qquad \hat{\beta}_j = \mathbf{1}_{n,j}(R_{j,t}'R_{j,t})^{-}R_{j,t}'Y,$$

$$R_{j,t} = (D_{t,1}R_j(X_1), \cdots, D_{t,n}R_j(X_n))', \qquad Y = (Y_1, \cdots, Y_n)',$$

where $\mathbf{1}_{n,j} = \mathbf{1}(\lambda_{\min}(\hat{\Omega}_{j,t}) > c)$, with $\lambda_{\min}(A)$ the minimum eigenvalue of a matrix A, $\hat{\Omega}_{j,t} = R_{j,t}'R_{j,t}/(nq_j)$, $q_j = \mathbb{P}[X_i \in P_j]$, and c a fixed positive constant.

This estimator is quite intuitive: within each cell, the unknown regression function is approximated by a polynomial of order $K-1$ in X_i, which is used to impute missing values for each observation, and then the imputed values are averaged out to obtained the final estimator. As shown in the Appendix, under appropriate regularity conditions, $\mathbf{1}_{n,j}$ takes the value 1 with probability approaching one, so that the least squares problem within each cell of the partition is (asymptotically) well defined.

4.3 Large Sample Results

This section describes the large sample properties of estimator introduced in the previous section. The following assumption imposes a set of simple restrictions on the data generating process.

Assumption 2. (a) X_i has compact support $\mathcal{X} \subset \mathbb{R}^d$, and its (Lebesgue) density is bounded and bounded away from zero.
(b) $\mathbb{E}[|Y_i(t)|^4|X_i]$ is bounded for all $t \in \mathcal{T}$.
(c) $\mu_t(x)$ is S_μ-times continuously differentiable for all $t \in \mathcal{T}$.
(d) $e_t(x)$ is S_e-times continuously differentiable for all $t \in \mathcal{T}$.

Assumption 2(a) is important, and may be relaxed only when certain special partitioning schemes are employed and more stringent moment assumptions are imposed, but is otherwise difficult to weaken. Assumptions 2(b)-(d) implicitly control the rate of convergence in uniform norm of the partitioning nonparametric estimator, as shown in (Cattaneo and Farrell 2013), and are standard in nonparametric and semiparametric estimation.

Regarding the partitioning nonparametric estimator, the following assumption will be imposed throughout. For scalars sequence $\{a_j : j = 1, \cdots, J_n\}$, let $a_j \asymp J_n^{-1}$ denote that $C_* J_n^{-1} \leq a_j \leq C^* J_n^{-1}$ with C_* and C^* universal positive constants not depending on n nor $j = 1, \cdots, J_n$.

Assumption 3. (a) For $\ell = 1, \cdots, d$ and $J_n \in \mathbb{N}$, let the ℓ-dimension of $\mathcal{X}$ be partitioned into the J_n disjoint intervals $[p_{\ell,j-1}, p_{\ell,j})$, $j = 1, \cdots, J_n - 1$, and $[p_{\ell,J_n-1}, p_{\ell,J_n}]$, satisfying $p_{\ell,j-1} < p_{\ell,j}$ for all j, and $|p_{\ell,j} - p_{\ell,j-1}| \asymp J_n^{-1}$. The complete partition of $\mathcal{X}$ consists of the J_n^d sets formed as Cartesian products of all such intervals, with typical cell denoted P_j.
(b) For some $K \in \mathbb{N}$, $R(x)$ represents the complete polynomial basis of order $K-1$ based on $x \in \mathbb{R}^d$.

Assumption 3(a) imposes natural restrictions on the partitioning scheme employed, which guarantee that each cell is well defined. By construction, each cell must satisfy: $\mathrm{vol}(P_j) \asymp J_n^{-d}$, or equivalently, for some positive constants C_* and C^*: $C_* J_n^d \leq \min_{1\leq j\leq J_n^d} \mathrm{vol}(P_j) \leq \max_{1\leq j\leq J_n^d} \mathrm{vol}(P_j) \leq J_n^{-d} \leq C^* J_n^d$, where $\mathrm{vol}(\cdot)$ denotes the volume of cell P_j. The simplest possible scheme is an evenly spaced partition, but Assumption 3(a) allows other possibilities so long as all cells continue to decrease proportionally to J_n^d. For a simple example, one may use a partition twice as fine in a region of abundant data compared to a sparse region (e.g., where the density is low). Assumption 3(b) specifies the degree of the polynomial used in the approximation within each cell. This assumption is meant to cover the general, unrestricted case, although in applications other (restricted) bases may be of interest. For example, if $\mu_t(x)$ is assumed to be additively separable, then the interactions between covariates may not be included in the basis $R(x)$, leading to a simpler least squares problem. The goal of Assumption 3(b) is to ensure that $R(x)$ is flexible enough to remove bias up to "order K", as shown in the Appendix.

The following theorem establishes that $\hat{\mu}_t$ has an asymptotically linear representation, with the well-known efficient influence function for μ_t (see, e.g., (Hahn 1998)).

Theorem IV.1. *Suppose Assumptions 1–3 hold, and*

$$\sqrt{n} J_n^{-K\wedge S_\mu \wedge S_e} \to 0$$

and

$$J_n^{10d/7} \log(J_n)^2/n \to 0$$

. Then, for all $t \in \mathcal{T}$,

$$\sqrt{n}(\hat{\mu}_t - \mu_t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_t(Y_i, X_i, T_i) + o_p(1),$$

where

$$\psi_t(Y_i, X_i, T_i) = \frac{D_{t,i}(Y_i - \mu_t(X_i))}{e_t(X_i)} + \mu_t(X_i) - \mu_t.$$

Theorem 1 shows that there exists a choice of $J_n^d \to \infty$ such that $\hat{\mu}$ is asymptotically linear and semiparametric efficient, provided both $\mu_t(x)$ and $e_t(x)$ are smooth enough and K is large enough.[4] The rate restrictions in Theorem 1 describe the lower and upper bounds on the rate of growth for the nonparametric tuning parameter, as is

[4] The rate restrictions imposed in Theorem 1 are in general not minimal, and may be relaxed in certain cases. It is possible to show by example that a necessary condition is given by $J_n^d/n \to 0$.

common in semiparametric inference. This condition formalizes the intuition above: the first statement requires sufficiently small cells to control bias, while the second ensures the nonparametric variance does not grow too fast.

It follows from this theorem that $\sqrt{n}(\hat{\mu}-\mu) \to_d \mathcal{N}(0,V)$, with V the semiparametric efficiency bound for μ, that is, V has (t,s)-element $(1 \leq t,s \leq \tau)$ given by

$$V_{[t,s]} = \mathbb{E}\left[\mathbf{1}(t=s)\frac{\sigma_t^2(X_i)}{e_t(X_i)} + (\mu_t(X_i)-\mu_t)(\mu_s(X_i)-\mu_s)\right],$$

where $\sigma_t^2(X_i) = \mathbb{V}[Y_i(t)|X_i]$. See, e.g., (Cattaneo 2010) for a discussion on this and related results.

In order to construct feasible, asymptotically valid confidence intervals for μ a consistent estimator of the standard errors is needed. Several alternatives are in principle possible, although a subclassification-based estimator seems most natural in the present context. One such estimator may be justified as follows. The overall asymptotic variance may be decomposed into the sum of the "within" and "between" variance as follows:

$$V_{[t,s]} = \mathbb{E}\left[\mathbf{1}(t=s)\frac{\sigma_t^2(X_i)}{e_t(X_i)}\right] + \mathbb{E}\left[(\mu_t(X_i)-\mu_t)(\mu_s(X_i)-\mu_s)\right] = V_{W,[t,s]} + V_{B,[t,s]}.$$

It is intuitive to separately estimate each component. First, because a least squares estimate is computed within each cell, a natural choice for $\hat{V}_{W,[t,s]}$ is a Huber-Eicker-White heteroskedasticity-robust estimator:

$$\hat{V}_{W,[t,s]} = \mathbf{1}(t=s)\sum_{j=1}^{J_n^d} \mathbf{1}_{n,j}\hat{L}_j'\hat{\Omega}_{j,t}^{-1}\hat{\Sigma}_{j,t}\hat{\Omega}_{j,t}^{-1}\hat{L}_j, \qquad \hat{L}_j = \frac{1}{nq_j}\sum_{i=1}^{n} R_j(X_i),$$

$$\hat{\Sigma}_{j,t} = \frac{1}{n}\sum_{i=1}^{n} R_j(X_i)R_j(X_i)'D_{t,i}(Y_i - \hat{\mu}_t(X_i))^2.$$

This estimator has a simple, intuitive representation when $K=1$, given by

$$\hat{V}_{W,[t,s]} = \mathbf{1}(t=s)\frac{1}{n}\sum_{i=1}^{n}\hat{\sigma}_t^2(X_i), \qquad \hat{\sigma}_t^2(x) = \sum_{j=1}^{J_n^d}\frac{N_j^2}{N_{j,t}^2}\mathbf{1}_{P_j}(x)D_{t,i}(Y_i - \hat{\mu}_t(X_i))^2.$$

Second, for $V_{B,[t,s]}$, a simple partitioning-based plug-in estimator is:

$$\hat{V}_{B,[t,s]} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_t(X_i)-\hat{\mu}_t)(\hat{\mu}_s(X_i)-\hat{\mu}_s).$$

The following theorem verifies that both estimators, $\hat{V}_{W,[t,s]}$ and $\hat{V}_{B,[t,s]}$, are indeed consistent for their population counterparts.

Theorem IV.2. *Suppose the conditions of Theorem 1 hold. Then, for all $t, s \in \mathcal{T}$, $\hat{V}_{W,[t,s]} \to_p V_{W,[t,s]}$ and $\hat{V}_{B,[t,s]} \to_p V_{B,[t,s]}$.*

It follows immediately from Theorem 2 that $\hat{V}$ with typical (t, s)-element $(1 \leq t, s \leq \tau)$ given by $\hat{V}_{[t,s]} = \hat{V}_{W,[t,s]} + \hat{V}_{B,[t,s]}$ is a consistent estimator of V, leading to a consistent estimator of the semiparametric efficiency bound obtained in Theorem 1.

4.4 Simulations

In this section we report the results of a Monte Carlo study of the subclassification-based estimator. We focus on a binary treatment (i.e. $\tau = 2$) and conduct inference on the average treatment effect throughout, both for simplicity and to facilitate comparison with other results in the program evaluation literature.

The data generating process we consider is as follows. $X_{1i} \sim \texttt{Uniform}[-2, 2]$, $Y_i(0) = \mu_0 + X_{1i} + \eta_{0i}$, $Y_i(1) = \mu_1 + \exp\{X_{1i}\} - \mathbb{E}[\exp\{X_{1i}\}] + \eta_{1i}$, and $T_i = \mathbf{1}\{X_{1i}^3/3 - X_{1i} + \eta_{2i} > 0\}$, where the errors $\eta_{ki} \sim N(0, 2)$, $k = 0, 1, 2$, and are mutually independent. We also consider a heteroskedastic variant of this model, in which $\eta_{1i} \sim N(0, 2X_i^2)$. Further, we extend these models to include a second covariate by generating $X_{2i} \sim \texttt{Uniform}[-2, 2]$ independently of X_{1i} then setting $Y_i(0) = \mu_0 + X_{1i} + X_{2i} + \eta_{0i}$, $Y_i(1) = \mu_1 + X_{1i}^3 X_{2i}^2 + \exp\{X_{1i}\} + \exp\{X_{2i}\} - 2\mathbb{E}[\exp\{X_{1i}\}] + \eta_{1i}$, and all else as above. In all cases we set $\mu_0 = 0$ and $\mu_1 = 1$, so that the average treatment effect is one. We conduct simulations of each model with sample sizes of 500 and 1,000, both using 2,000 replications and evenly spaced cells.

The average bias for the univariate model are reported in Figure 1 for a range of J_n. The homoskedastic and heteroskedastic model produce very similar results, and the discussion below applies to both. The figure demonstrates several salient features of the subclassification estimator. First, the increased flexibility of the nonparametric estimator resultant from increasing the number of cells initially decreases the bias. The nonparametric procedure relies on $J_n^d \to \infty$ as $n \to \infty$, and the bias decreases accordingly: see (A-2) in the appendix. The second important feature is the choice of the (fixed) parameter K, giving the order of the fit within each cell. Recall from above that $K = 1$ corresponds to fitting means within each cell, $K = 2$ gives a linear fit, and so forth. As aforementioned, a larger K improves the theoretic bias properties of the estimator and for modest values of J_n this is borne out in Figure 1: For J_n

below 10, fitting means within each cell may not be sufficient to remove bias, but an increase merely to linear fits is often adequate. Much more modest improvements result from a quadratic fit.

However, as J_n increases further, the bias properties decline: the estimator has increased bias compared to fewer cells, and substantially so for the quadratic fit. This is a consequence of the least squares problem being ill-posed in an increasing number of cells. Heuristically, for the (fixed) n chosen, these J_n represent sequences for which the rate restrictions of Theorem 1 do not hold, and hence the distributional approximation is invalid. Recalling the formulation above, for these "empty cells" $\mathbf{1}_{nj} = 0$ and the matrix $\hat{\Omega}_{j,t}$ is singular (or near singular; in practice a numerical cut-off is employed). Hence, these cells are not included in the estimate, leading to bias. It is beyond the scope of this work to study a formal trimming procedure, but one that controls for empty cells in a systematic way may lead to improved performance for certain choices of J_n. These results (and similarly those of the bivariate specification below) may be interpreted as a cautionary tale regarding choice of smoothing and tuning parameters in nonparametric estimation in general. It is also important to note that this phenomenon does not impact the estimator with degree zero (fitting means) as severely, since only one observation per cell is required. Indeed, for bias the piecewise constant version of the subclassification-based estimator is quite robust to the choice of J_n. Finally, note that the increased sample size expands the range of J_n for which the estimator performs well, for any choice of K.

Figure 2 reports coverage rates for 95% confidence intervals for the univariate models. Many of the same conclusions are evident. For modest values of J_n, increased K leads to more accurate coverage, but beyond a certain value, coverage declines more rapidly for higher values of K. Again, the robustness to choice of J_n for $K = 1$ is evident. The coverage is remarkably accurate for even a large number of cells. The variance estimator accounts for heteroskedasticity quite well: only a small loss is evident. In practice, the "empty-cell" issue is likely to be a greater concern.

Figures 3–6 show the Gaussian approximation for the four univariate models. The estimator approximates the semiparametric efficiency bound for several different choices of J_n and K, matching the result of Theorem 1. In all cases, the same conclusions above are evident and the heteroskedasticity makes little difference. For moderate choices, the robustness is again demonstrated. However, for $n = 500$ and a large J_n, the estimator is biased and the variance is inflated: the bottom-right graph in Figures 3 and 5 shows that the approximation can be poor. Again, increasing the sample size ameliorates the issue, as would be expected from the theory in Section 3.

The conclusions from the bivariate model are somewhat different. Figures 7 and 8 report bias and coverage for the bivariate models. Note that the range of J_n is restricted compared to the univariate models. Recall that the theory requires J_n^d cells, so that the points marked as "10" in Figures 7 and 8 actually utilize $J_n^2 = 100$ total cells. Here the empty cell problem has become severe, and both the bias and coverage properties become extremely poor. Also observe that for smaller values of J_n, fitting means is no longer sufficient to remove bias or produce accurate coverage: both are more accurate for the quadratic fit for a larger range of J_n. This illustrates the tension between the bias and variance conditions in Theorem 1 for the sequence J_n. This "curse of dimensionality" is a common problem in nonparametric estimation. Figures 9–12 show the Gaussian approximations, which exhibit the same issues. In some cases, the approximation is extremely poor for these sample sizes. However, observe that for moderate values of J_n (e.g., $J_n = 3, J_n^d = 9$), the semiparametric efficiency bound is approximated well for certain choices of K. To investigate this further, we simulated the bivariate homoskedastic model with a sample size of $n = 2{,}000$. The Gaussian approximation is shown in Figure 13. As would be expected, the estimator performs better for a wider range of J_n and K. The bias and coverage results (not shown) are also substantially improved. When considering additional regressors, researchers should keep this "curse of dimensionality" in mind.

Finally, we compare the partitioning estimator to several others common in the literature: inverse probability weighting (IPW), a series-based imputation estimator, and M-nearest-neighbor matching. The propensity score is estimated using a logistic regression on a power series of X_i up to order four or six, and then the average treatment effect is estimated by inverse weighting as in (Hirano, Imbens, and Ridder 2003). The series estimator uses nonparametric regression to impute missing outcomes in much the same spirit as the partitioning estimator, but the approximation is global, see (Imbens, Newey, and Ridder 2007). Here we use a power series of degree four or six, but to approximate the underlying regression function instead of the propensity score. Finally, we consider nearest-neighbor matching (Abadie and Imbens 2006). We implement this in Stata using the `nnmatch` command of (Abadie, Drukker, Herr, and Imbens 2004). We consider one- and two-neighbor matching, as well as simple and bias-adjusted estimates. For brevity, we consider only the univariate homoskedastic model. Following the above results, we use only 7- and 10-cell partitions, and only up to a linear fit. Table 1 presents mean-square error comparison between the estimators. Gaussian approximations are given in Figures 14 and 15. In the figures the 10-cell subclassification estimator with degree zero is given by the solid line, with the long-dashed line for degree one. Results are comparable with $J_n = 7$, so this is excluded. The comparison estimator is given by the short-dashed line for the "lower" degree (power series of degree four in the case of IPW and Series, or one match) and a dotted line for the "higher" degree (degree six, two matches).

The subclassification estimator performs comparably to these alternatives. Both the IPW and series estimators are known to attain the semiparametric efficiency bound, which is borne out in panels (A) and (B) of Figures 14 and 15. Table 1 shows that these estimators exhibit comparable variance to the subclassification estimator. For a fixed number of matches, nearest-neighbor matching is well-centered but does not attain the bound, and hence it is not surprising that the subclassification estimator is more concentrated, see panels (C) and (D). The MSE of the matching estimator is larger as a result. For a piecewise constant or linear fit, the subclassification estimator appears to be on par with popular choices in the econometrics literature.

4.5 Extensions and Final Remarks

The main result of this chapter (Theorem 1) shows that the subclassification-based estimator of the Dose-Response Function introduced in Section 2 is asymptotically linear and semiparametric efficient under standard regularity conditions. Theorem 2 also demonstrates that a simple, consistent standard errors estimator is easy to construct based on the idea of subclassification. In addition, the simulation study reported in Section 4 suggests that this estimator performs well in finite samples.

The theoretical results presented in this chapter may be easily extended to cover other potential estimands of interest. Perhaps the most natural extension would be to consider estimating the quantiles of the distribution of $Y(t)$, $t \in \mathcal{T}$. (See, e.g., Firpo (2007).) In this case, because the α-th quantile of $Y(t)$, denoted by $q_t(\alpha)$, solves $0 = \mathbb{E}[m(Y(t), q_t(\alpha); \alpha)]$ with $m(y, q; \alpha) = \mathbf{1}(y \leq q) - \alpha$, a natural subclassification-based estimator of $q_t(\alpha)$ would be given by $\hat{q}_t(\alpha) = \arg\min_q |M_n(q; \alpha)|$,

$$M_n(q;\alpha) = \frac{1}{n}\sum_{i=1}^{n} \hat{q}_t(X_i;\alpha), \qquad \hat{q}_t(x;\alpha) = \sum_{j=1}^{J_n^d} R_j(X_i)'\hat{\beta}_{j,\alpha},$$

$$\hat{\beta}_{j,\alpha} = \mathbf{1}_{n,j}(R_{j,t}'R_{j,t})^{-1}R_{j,t}'Y(q;\alpha), \qquad Y(q;\alpha) = (m(Y_1, q;\alpha), \cdots, m(Y_n, q;\alpha))'.$$

Under appropriate regularity conditions, it seems plausible that the resulting estimator $\hat{q}_t(\alpha)$ would also be asymptotically normal and semiparametric efficient. More generally, it is natural to expect that such a result would hold for other estimands as defined by a choice of function $m(\cdot)$ in some appropriate class. For a discussion on related ideas and other potential extensions see, e.g., (Cattaneo 2010) and references therein. These extensions are not considered in this chapter for brevity, and consequently are relegated for future work.

Another useful extension to the present work would be to develop a guide for the choice of J_n in applications. The number of cells is the nonparametric tuning parameter, and its choice is important for the finite sample properties of the estimator, as discussed in Section 4. A natural criterion for choosing J_n would be to consider a mean-square error expansion of the estimator, which could be minimized to find the optimal number of cells. Among other things, this would be a function of K, the smoothing parameter. Following this, a simple "plug-in" estimate could be proposed for the optimal J_n.

Figure 4.1: Empirical Average Bias for Univariate Models

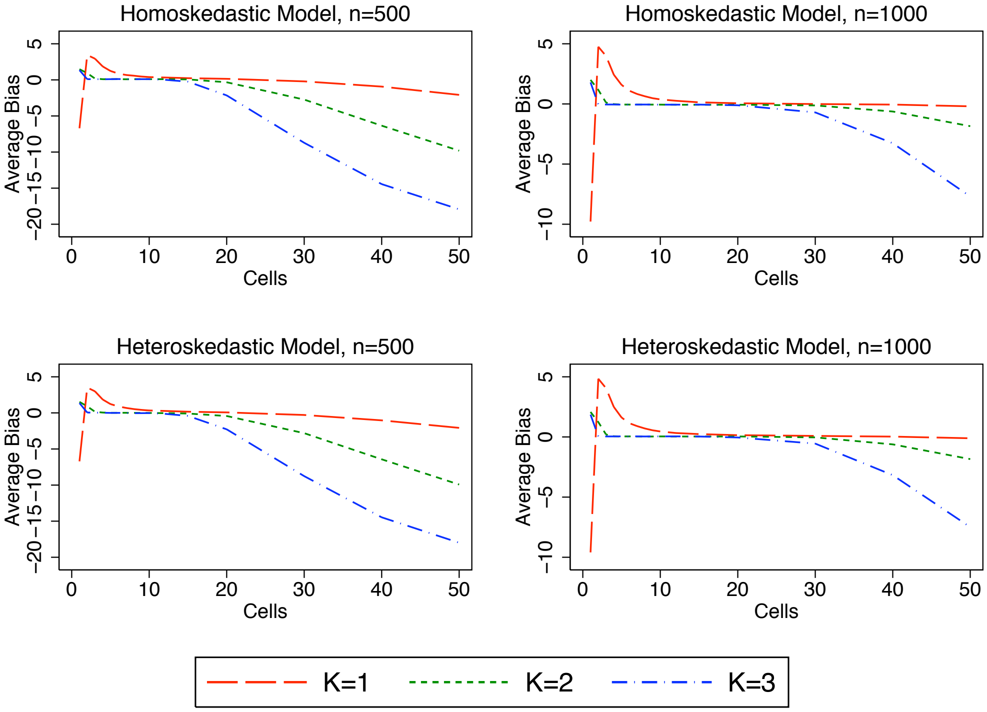


Table 4.1: Mean-Square Error Compared to Alternative Estimators

		$n = 500$			$n = 1{,}000$		
		Bias	Var.	MSE	Bias	Var.	MSE
Subclassification Estimator							
$J_n = 7$	$K = 1$	0.484	10.172	10.407	0.887	9.868	10.655
$J_n = 7$	$K = 2$	-0.14	10.087	10.107	0.015	9.474	9.474
$J_n = 10$	$K = 1$	0.174	10.11	10.14	0.427	9.543	9.726
$J_n = 10$	$K = 2$	-0.138	10.285	10.304	0.021	9.595	9.595
IPW							
Degree 4		-0.137	9.902	9.921	0.016	9.405	9.405
Degree 6		-0.358	10.567	10.695	-0.16	9.738	9.764
Series							
Degree 4		-0.136	9.804	9.822	0.012	9.35	9.35
Degree 6		-0.357	10.446	10.573	-0.169	9.725	9.754
NN-Matching							
Simple	M=1	-0.135	13.24	13.258	-0.003	12.619	12.619
Simple	M=2	-0.113	11.501	11.514	0.021	10.982	10.982
Bias-adj	M=1	-0.138	13.24	13.259	-0.005	12.618	12.618
Bias-adj	M=2	-0.12	11.502	11.516	0.018	10.981	10.981

Figure 4.2: Coverage Rates for 95% Confidence Intervals for Univariate Models

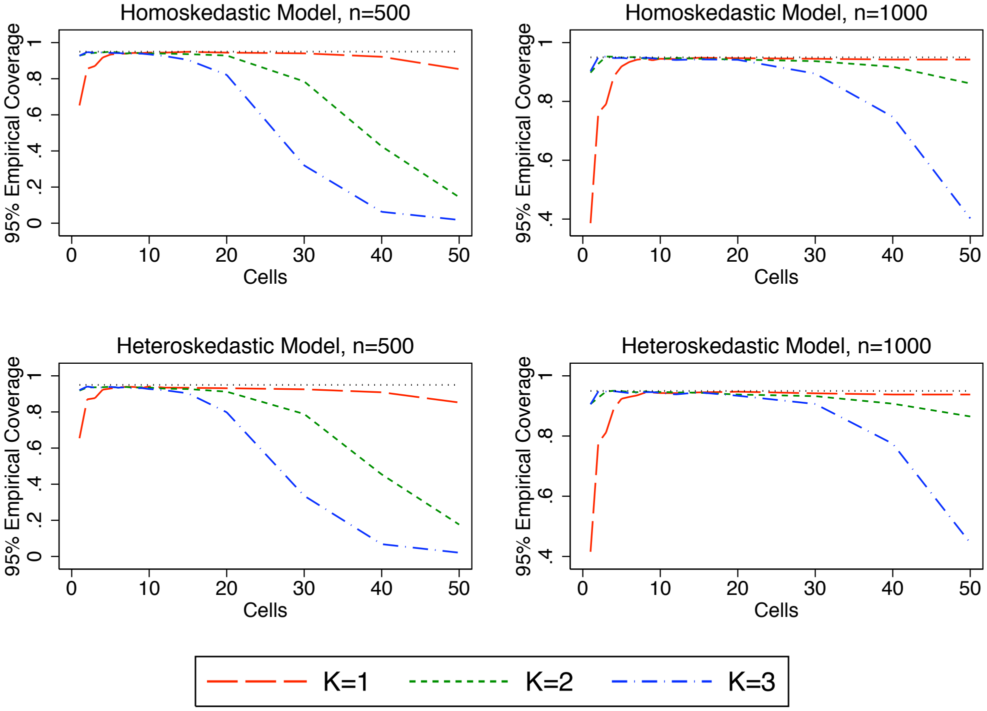


Figure 4.3: Normal Approx. for Univariate Homoskedastic Model, $n = 500$

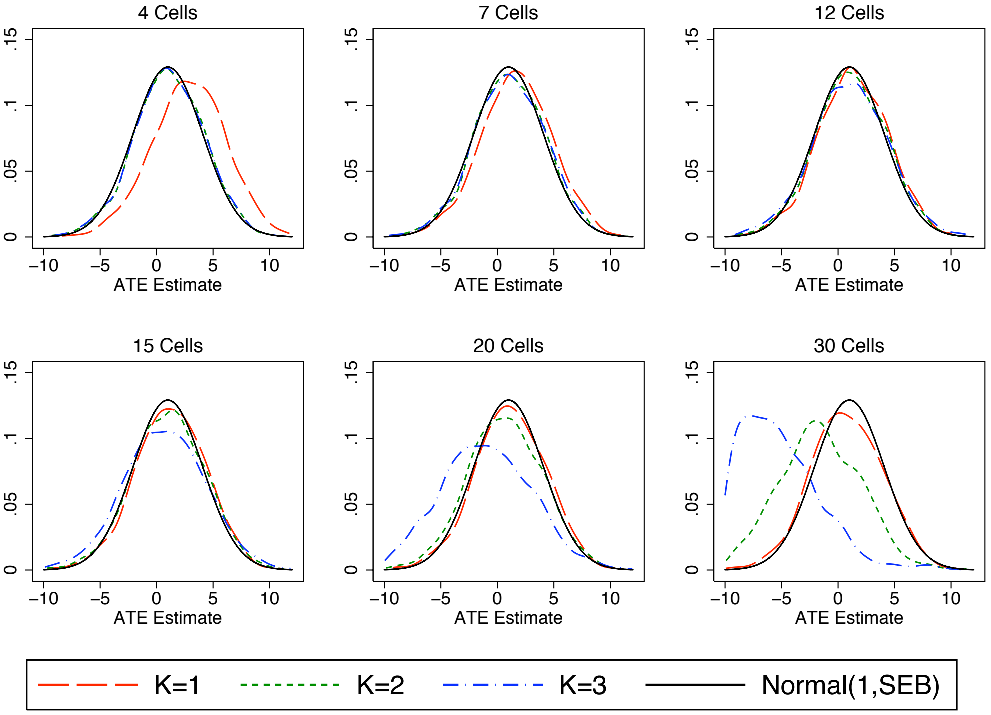


Figure 4.4: Normal Approx. for Univariate Homoskedastic Model, n = 1,000

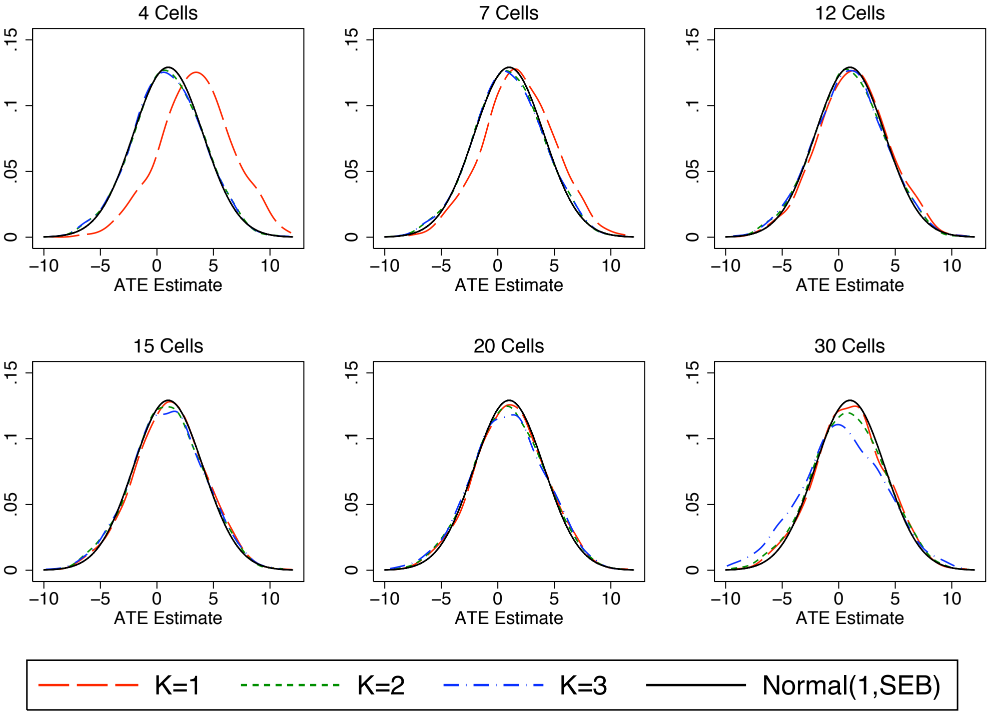


Figure 4.5: Normal Approx. for Univariate Heteroskedastic Model, $n = 500$

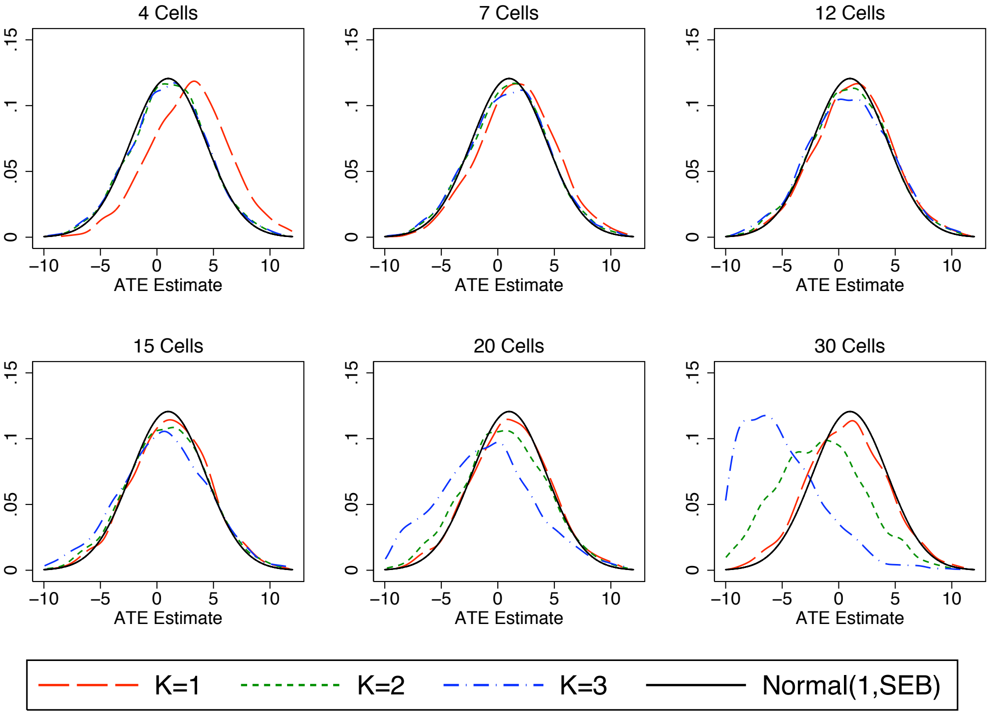


Figure 4.6: Normal Approx. for Univariate Heteroskedastic Model, n = 1,000

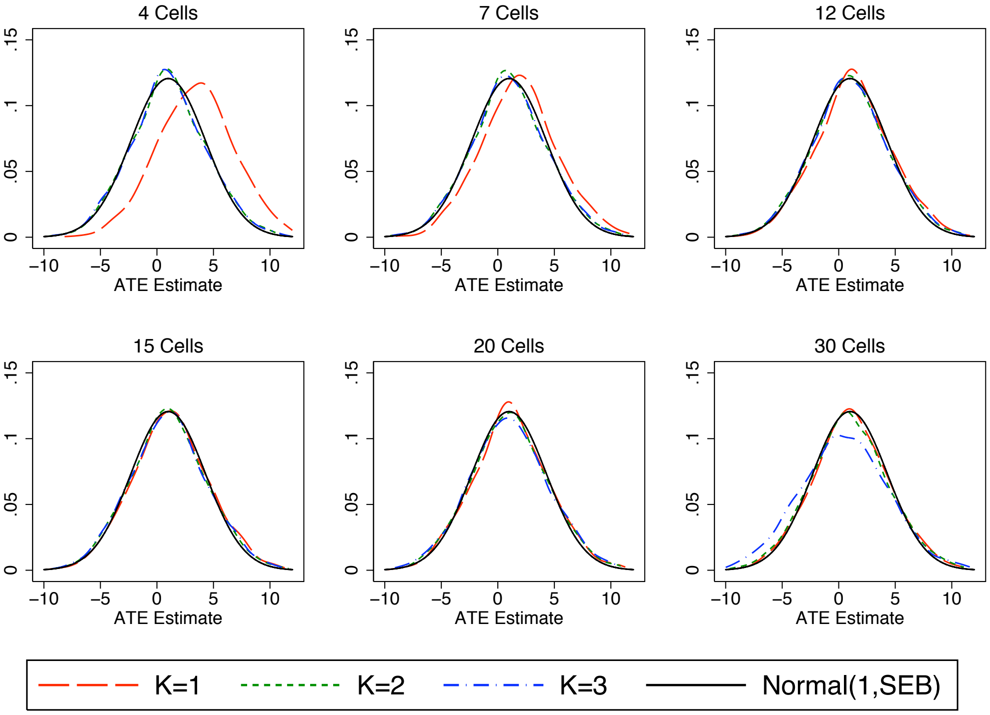
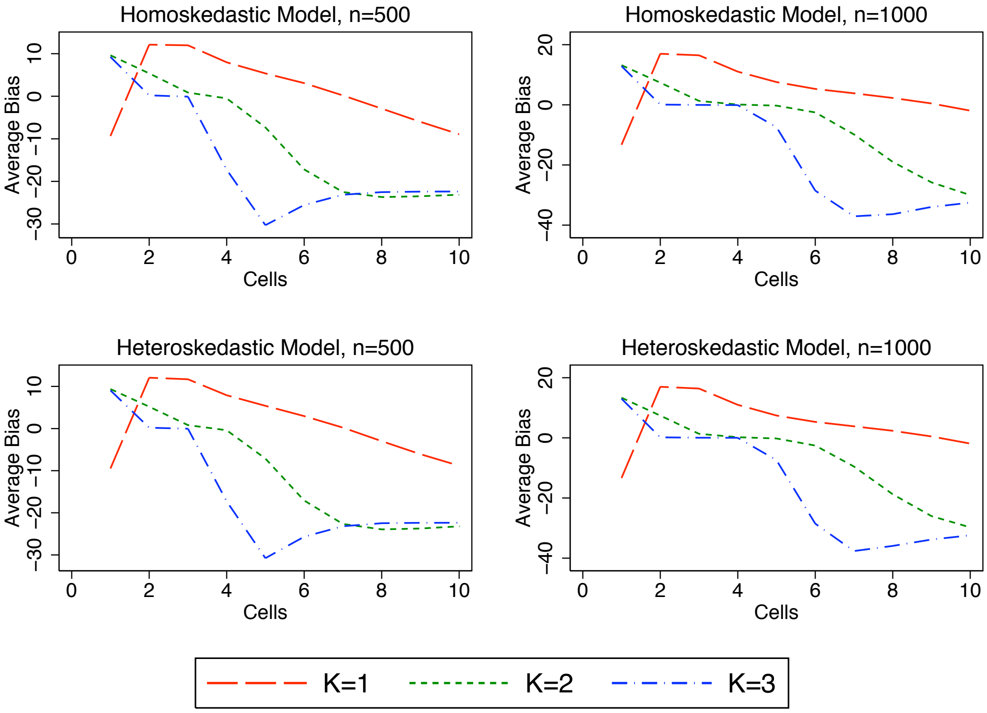


Figure 4.7: Empirical Average Bias for Bivariate Models



Homoskedastic Model, n=500

Homoskedastic Model, n=1000

Heteroskedastic Model, n=500

Heteroskedastic Model, n=1000

Figure 4.8: Coverage Rates for 95% Confidence Intervals for Bivariate Models

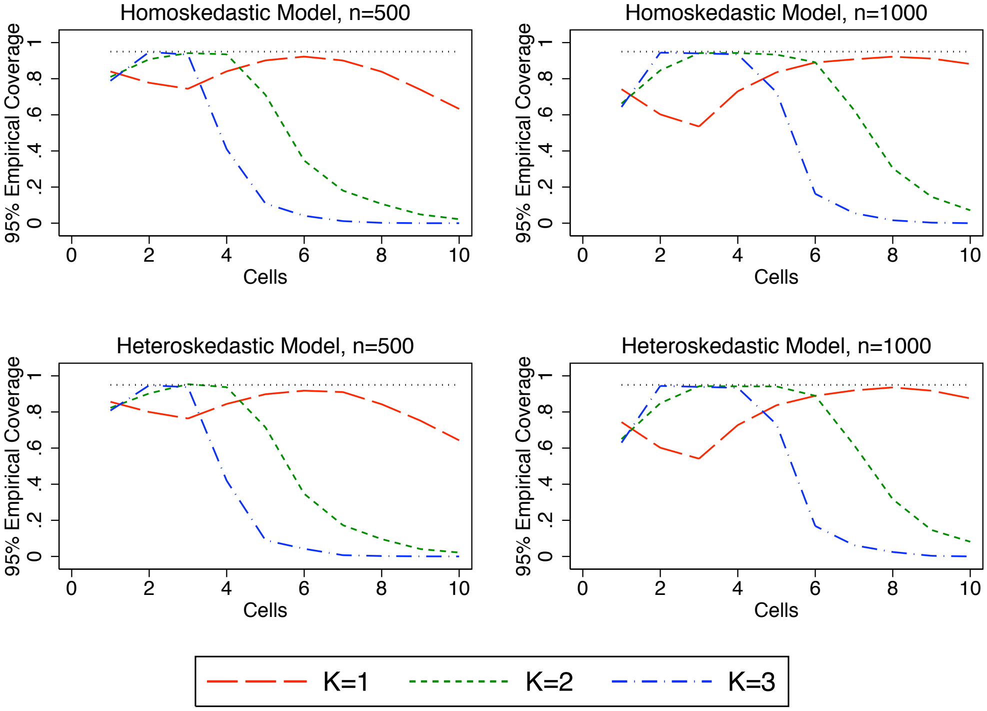


Figure 4.9: Normal Approx. for Bivariate Homoskedastic Model, $n = 500$

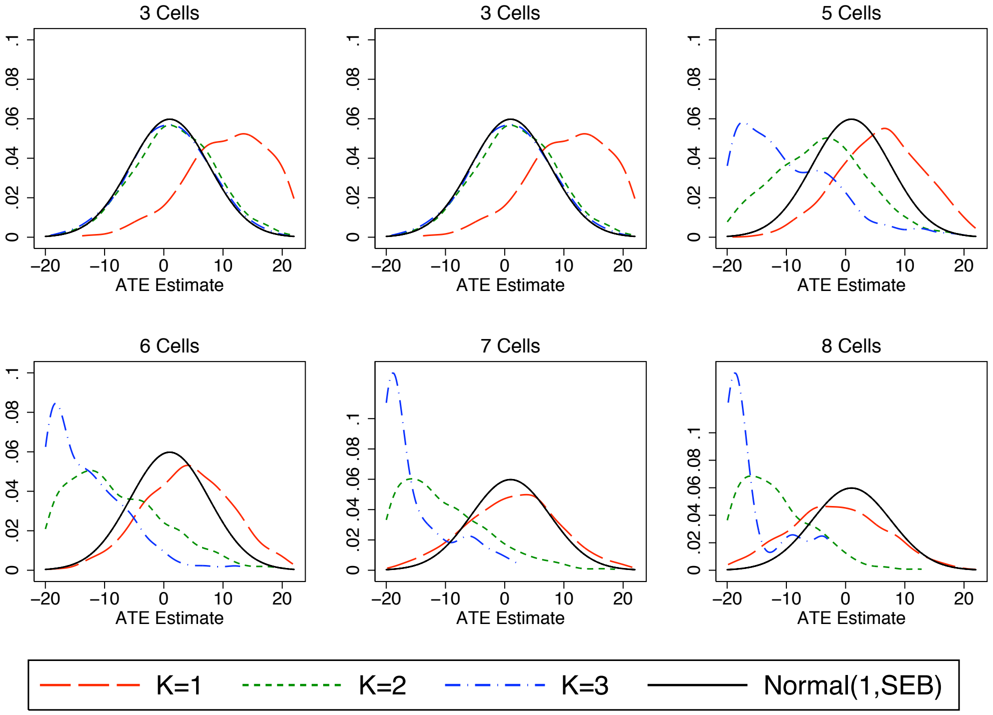


Figure 4.10: Normal Approx. for Bivariate Homoskedastic Model, $n = 1{,}000$

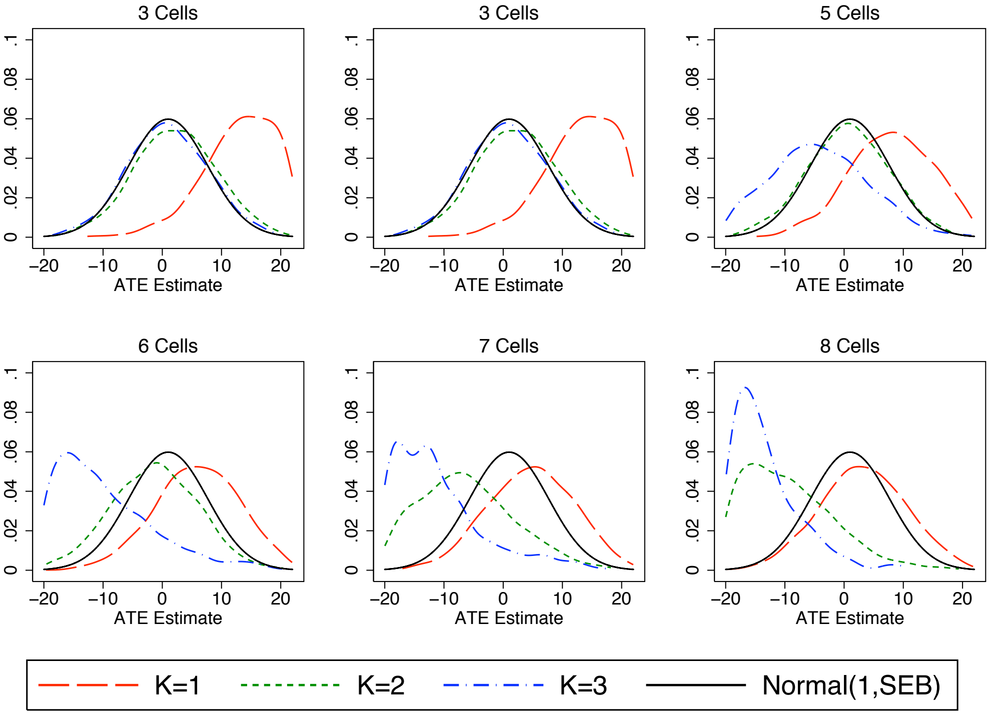


Figure 4.11: Normal Approx. for Bivariate Heteroskedastic Model, $n = 500$

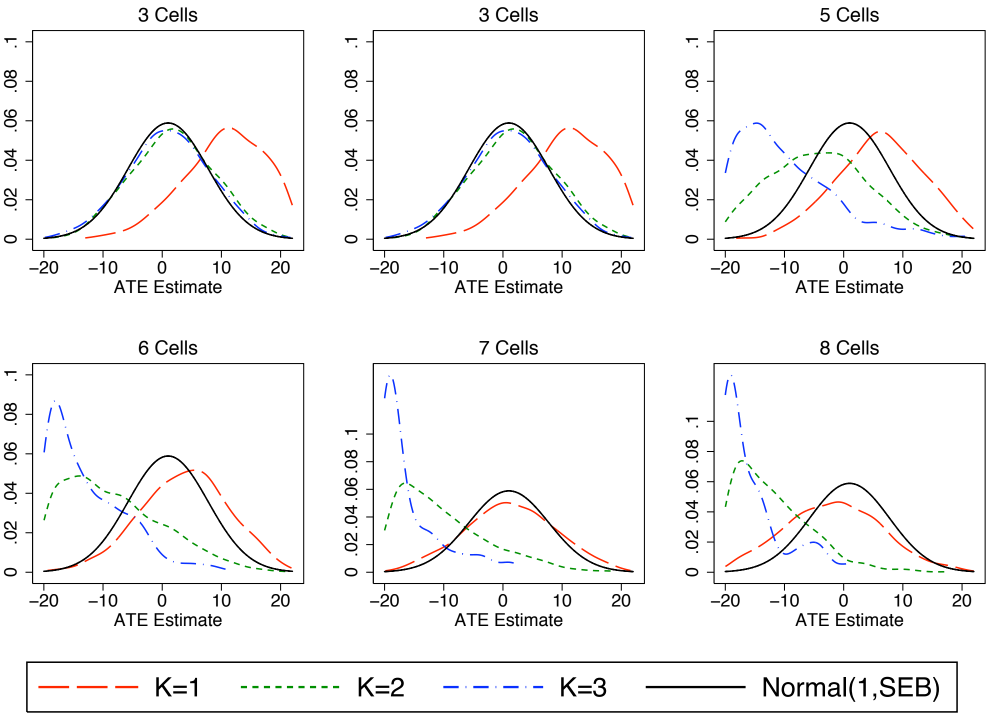


Figure 4.12: Normal Approx. for Bivariate Heteroskedastic Model, $n = 1{,}000$

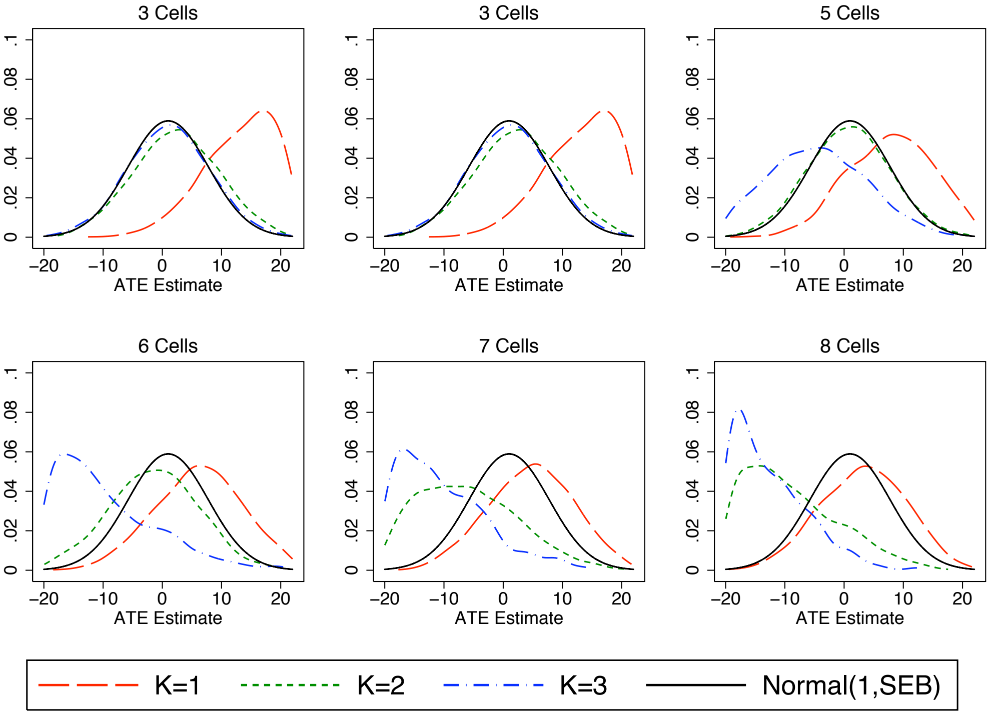


Figure 4.13: Normal Approx. for Bivariate Homoskedastic Model, $n = 2{,}000$

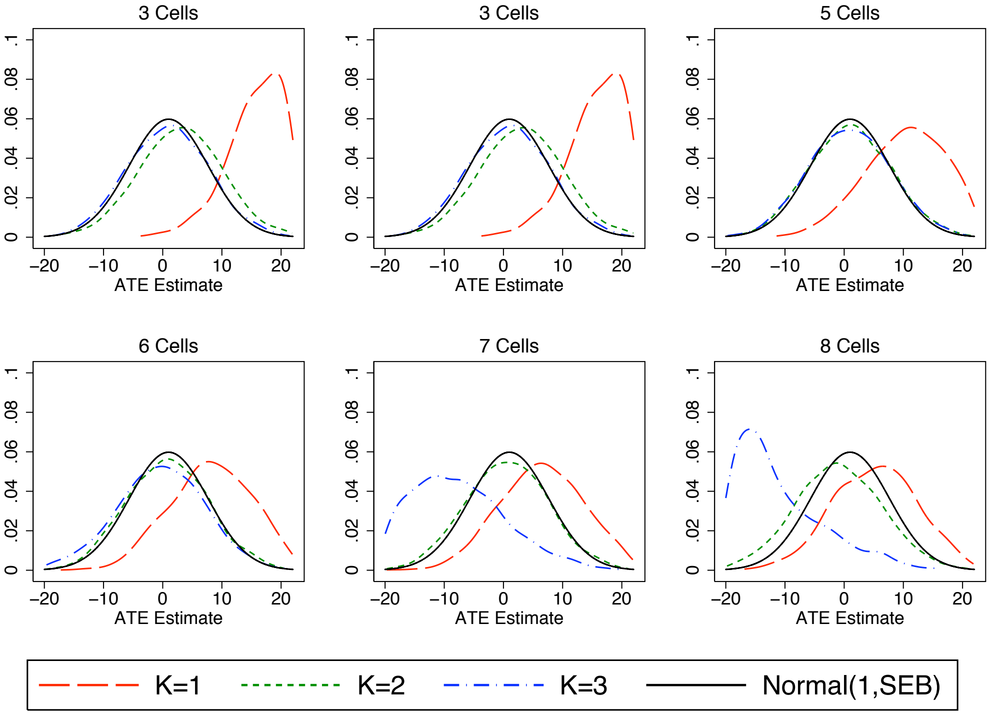


Figure 4.14: Comparison to Alternative Estimators, $n = 500$

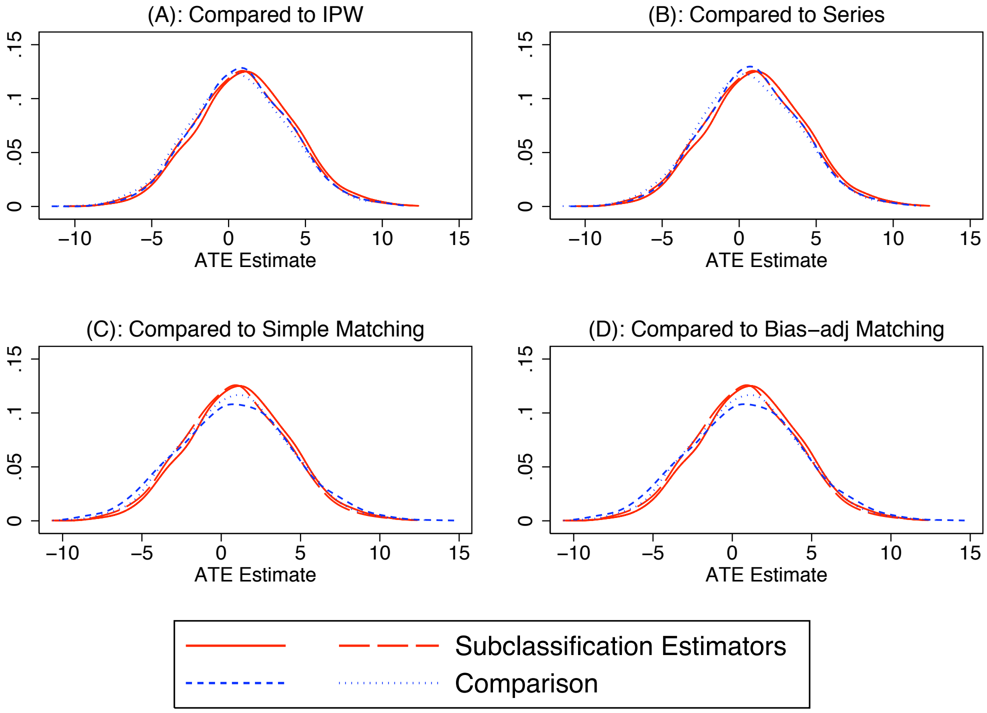
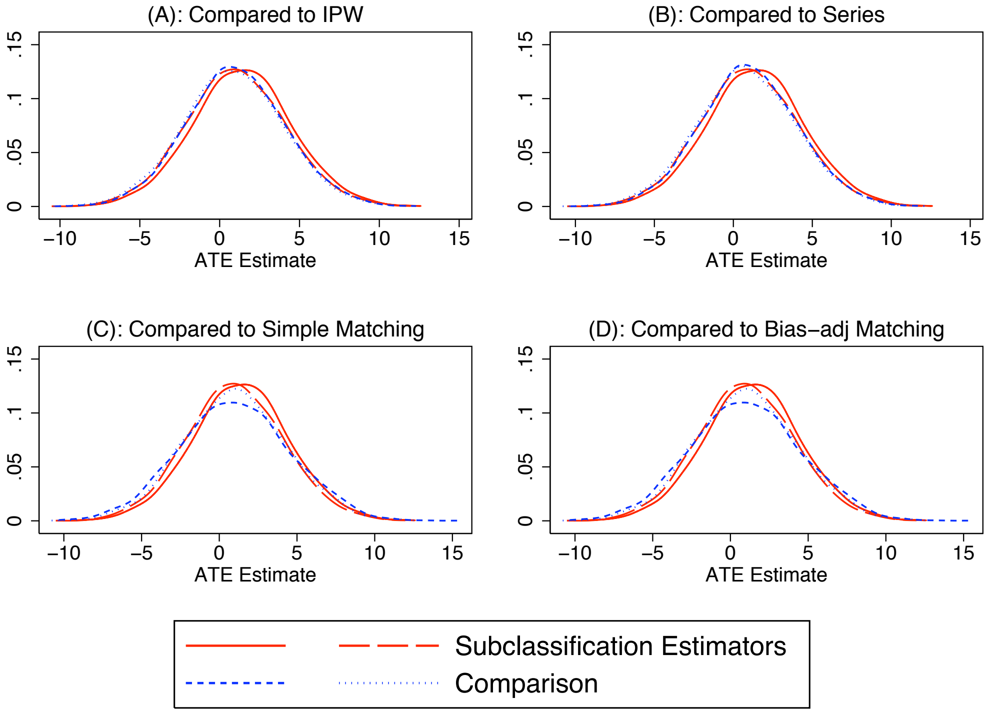


Figure 4.15: Comparison to Alternative Estimators, $n = 1{,}000$



APPENDICES

APPENDIX A

Proofs for Chapter 2

Proofs for Chapter 2: Treatment Effect Inference

The proofs in this section are asymptotic in nature, compared to the nonasymptotic bounds of the next section. It shall be understood that asymptotic order symbols hold for the sequence being considered, as a shorthand for the more formal versions given in the assumptions (e.g. Assumption II.5). C will denote a generic positive constant, which may be a matrix. Define the set of indexes $\mathbb{I}_t = \{i : d_i = t\}$.

Proofs for Average Treatment Effects

Proof of Theorem II.7. SEE SUPPLEMENTAL APPENDIX. □

We first prove Theorem II.8.1 assuming there is no additional randomness injected into the support estimates. Following this, we redo the proof to account for additional randomness. We then turn to the remaining portions of Theorem II.8 and to Corollary II.9, which require shorter arguments.

We make frequent use of the linearization

$$\frac{1}{a} = \frac{1}{b} + \frac{b-a}{ab} = \frac{1}{b} + \frac{b-a}{b^2} + \frac{(b-a)^2}{ab^2}, \tag{A.1}$$

where the first inequality is readily verified, and the second re-applies the first.

Proof of Theorem II.8.1 without Additional Randomness. With $\psi_t(\cdot)$ defined in Eqn.

(2.3), we have $\sqrt{n}(\hat{\mu}_t - \mu_t) = \sqrt{n}\mathbb{E}_n[\psi_t(y_i, d_i^t, \mu_t(x_i), p_t(x_i), \mu_t)] + R_1 + R_2$, where

$$R_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} d_i^t (y_i - \mu_t(x_i)) \left(\frac{1}{\hat{p}_t(x_i)} - \frac{1}{p_t(x_i)} \right)$$

and

$$R_2 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\hat{\mu}_t(x_i) - \mu_t(x_i)) \left(1 - \frac{d_i^t}{\hat{p}_t(x_i)} \right).$$

The proof proceeds by showing that both R_1 and R_2 are $o_{P_n}(1)$.

For R_1, applying the first equality in Eqn. (A.1), we rewrite R_1 as

$$R_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} d_i^t u_i \left(\frac{p_t(x_i) - \hat{p}_t(x_i)}{\hat{p}_t(x_i) p_t(x_i)} \right).$$

Then, applying Assumptions II.3(b) and II.4(c) and the first-stage consistency condition of Assumption II.5(a):

$$\mathbb{E}\left[R_1^2 | \{x_i, d_i\}_{i=1}^n \right] = \mathbb{E}_n \left[\frac{d_i^t \sigma_t(x_i)}{p_t(x_i)^4} \left(p_t(x_i) - \hat{p}_t(x_i) \right)^2 \right] \leq C\mathbb{E}_n[(p_t(x_i) - \hat{p}_t(x_i))^2] = o_{P_n}(1).$$

Next, again using Eqn. (A.1) we have

$$1 - \frac{d_i^t}{\hat{p}_t(x_i)} = \frac{p_t(x_i) - d_i^t}{p_t(x_i)} + \frac{d_i^t(\hat{p}_t(x_i) - p_t(x_i))}{\hat{p}_t(x_i) p_t(x_i)}.$$

We use this to re-write $R_2 = R_{21} + R_{22}$, where

$$R_{21} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\hat{\mu}_t(x_i) - \mu_t(x_i)) \left(\frac{p_t(x_i) - d_i^t}{p_t(x_i)} \right)$$

and

$$R_{22} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\hat{\mu}_t(x_i) - \mu_t(x_i))(\hat{p}_t(x_i) - p_t(x_i)) \left(\frac{d_i^t}{\hat{p}_t(x_i) p_t(x_i)} \right).$$

For the first term, as in R_1, we have

$$\begin{aligned} \mathbb{E}\left[R_{21}^2 | \{y_i, x_i\}_{i=1}^n \right] &= \mathbb{E}_n \left[(\hat{\mu}_t(x_i) - \mu_t(x_i))^2 \left(\frac{p_t(x_i)(1 - p_t(x_i))}{p_t(x_i)^2} \right) \right] \\ &\leq C\mathbb{E}_n \left[(\hat{\mu}_t(x_i) - \mu_t(x_i))^2 \right] = o_{P_n}(1), \end{aligned}$$

by the first-stage consistency condition of Assumption II.5(a). Next, by Hölder's inequality, Assumption II.3(b) and the rate condition of Assumption II.5(b)

$$\begin{aligned}|R_{22}| &\leq \sqrt{n}\left(\max_{i\leq n}\frac{1}{\hat{p}_t(x_i)p_t(x_i)}\right)\sqrt{\mathbb{E}_{n,t}[(\hat{\mu}_t(x_i)-\mu_t(x_i))^2]\mathbb{E}_{n,t}[(\hat{p}_t(x_i)-p_t(x_i))^2]}\\ &= O_{P_n}(1)\sqrt{n}\sqrt{\mathbb{E}_{n,t}[(\hat{\mu}_t(x_i)-\mu_t(x_i))^2]\mathbb{E}_{n,t}[(\hat{p}_t(x_i)-p_t(x_i))^2]} = o_{P_n}(1).\end{aligned}$$

This completes the proof, as $|R_1 + R_2| = o_{P_n}(1)$ by Markov's inequality and the triangle inequality. □

Proof of Theorem II.8.1 with Additional Randomness. We must reconsider the remainders R_1 and R_2. For the former, applying Eqn. (A.1), we find $R_1 = R_{11} + R_{12}$, where

$$R_{11} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{d_i^t u_i}{p_t(x_i)^2}\left(p_t(x_i)-\hat{p}_t(x_i)\right)$$

and

$$R_{12} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{d_i^t u_i}{p_t(x_i)^2\hat{p}_t(x_i)}\left(\hat{p}_t(x_i)-p_t(x_i)\right)^2.$$

For R_{11}, we first add and subtract the parametric representation to get $R_{11} = R_{111} + R_{112}$, where,

$$R_{111} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{d_i^t u_i}{p_t(x_i)^2}\left(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}) - \hat{p}_t(\{x_i^{*\prime}\hat{\gamma}_t\}_{\mathbb{N}_\mathcal{T}})\right)$$

and

$$R_{112} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{d_i^t u_i}{p_t(x_i)^2}\left(p_t(x_i) - \hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})\right).$$

By a two-term mean-value expansion $R_{111} = R_{111a} + R_{111b}$, with

$$R_{111a} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{d_i^t u_i}{p_t(x_i)^2}\sum_{t\in\mathbb{N}_\mathcal{T}}\left\{\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})(1-\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}))(x_i^{*\prime}(\hat{\gamma}_t-\gamma_t^*)\right\}$$

and

$$R_{111b} = \frac{1}{2\sqrt{n}}\sum_{i=1}^{n}\frac{d_i^t u_i}{p_t(x_i)^2}v_i'\bar{\mathcal{H}}v_i,$$

where $v_i = \{x_i^{*\prime}(\hat{\gamma}_t - \gamma_t^*)\}_{\mathbb{N}_\mathcal{T}}$ and $\overline{\mathcal{H}} = \mathcal{H}(\{x_i^{*\prime}\gamma_t^* + m_t x_i^{*\prime}\hat{\gamma}_t\}_{\mathbb{N}_\mathcal{T}})$ for appropriate scalars m_t.

For R_{111a}, consider each term in the sum over $\mathbb{N}_\mathcal{T}$ one at a time; let $R_{111a} = \sum_{t\in\mathbb{N}_\mathcal{T}} R_{111a}(t)$. Let t' denote the original treatment under consideration. Define

$\Sigma_{t,j} = \mathbb{E}\left[(x^*_{i,j})^2\sigma^2_{t'}(x_i)\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})^2(1-\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}))^2/p_{t'}(x_i)^3\right]$. Then proceed as follows

$$\begin{aligned}
R_{111a}(t) &= \frac{1}{\sqrt{n}}\sum_{i=1}^n\left(\frac{d_i^{t'}u_i\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})(1-\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}))}{p_{t'}(x_i)^2}\right)\sum_{j\in\hat{S}_D}x^*_{i,j}(\hat{\gamma}_t-\gamma_t^*)\\
&= \sum_{j\in\hat{S}_D}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^n\left(x^*_{i,j}\frac{d_i^{t'}u_i\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})(1-\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}))}{p_{t'}(x_i)^2\Sigma_{t,j}^{1/2}}\right)\right\}\\
&\qquad\qquad\times\Sigma_{t,j}^{1/2}(\hat{\gamma}_{t,j}-\gamma^*_{t,j})\\
&\le\left(\max_{j\in\mathbb{N}_p}\Sigma_{t,j}^{1/2}\right)\left(\max_{j\in\mathbb{N}_p}\frac{1}{\sqrt{n}}\sum_{i=1}^n x^*_{i,j}\frac{d_i^{t'}u_i\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})(1-\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}))}{p_{t'}(x_i)^2\Sigma_{t,j}^{1/2}}\right)\\
&\qquad\qquad\times\|\hat{\gamma}_t-\gamma_t^*\|_1\\
&= O(1)O_{P_n}(\log(p))\,\|\hat{\gamma}_t-\gamma_t^*\|_1 = o_{P_n}(1).
\end{aligned}$$

Convergence follows under Assumption II.6. For the penultimate equality, it follows from Assumptions II.3(b), II.4(b), and II.4(c) that $\max_{j\in\mathbb{N}_p}\Sigma_{t,j} = O(1)$. Finally, the center factor is shown to be $O_{P_n}(\log(p))$ by applying the moderate deviation theory for self-normalized sums of de la Peña, Lai, and Shao (2009, Theorem 7.4) and in particular Belloni, Chen, Chernozhukov, and Hansen (2012, Lemma 5). To apply this lemma, first note that the summand of the center factor has bounded third moment and second moment bounded away from zero, from Assumptions II.3(b), II.4(b), II.4(c), and the requirements of Assumptions II.5 and II.6. $\Sigma_{t,j}$ normalizes the second moment, and the lemma applies under the first restriction of Assumption II.6.

Again by the results of Tanabe and Sagae (1992) and Assumption II.5, $v_i'\bar{\mathcal{H}}v_i \le C\|v_i\|_2^2$. Thus, using Assumption II.3(b) to bound $\max_{i\le n}p_t(x_i)^{-2} < C$, we find R_{111b} may be bounded as follows:

$$\begin{aligned}
|R_{111b}| &\le C\sum_{t\in\mathbb{N}_\mathcal{T}}\sqrt{n}(\max_{i\in\mathbb{I}_t}|u_i|)\mathbb{E}_{n,t}\left[|x_i^{*\prime}(\hat{\gamma}_t-\gamma_t^*)|^2\right]\\
&\le C\mathcal{T}\max_{t\in\mathbb{N}_\mathcal{T}}\left|\sqrt{n}(\max_{i\in\mathbb{I}_t}|u_i|)\mathbb{E}_{n,t}\left[|\hat{p}_t(\{x_i^{*\prime}\hat{\gamma}_t\}_{\mathbb{N}_\mathcal{T}})-\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})|^2\right]\right| = o_{P_n}(1),
\end{aligned}$$

by the union bound and Assumption II.6, using the Assumptions II.3(b) and II.5(a) to apply Eqn. (A.13) with the inequality reversed.

A variance bound may be applied to R_{112} as in the previous proof, and we have $|R_{112}| = O_{P_n}(b_s) = o_{P_n}(1)$ by Markov's inequality.

Next, R_{12} is simply bounded by

$$
\begin{aligned}
|R_{12}| &\leq \sqrt{n}(\max_{i \in \mathbb{I}_t} |u_i|) \left(\max_{i \in \mathbb{I}_t} \frac{1}{p_t(x_i)^2 \hat{p}_t(x_i)} \right) \mathbb{E}_{n,t} \left[(\hat{p}_t(x_i) - p_t(x_i))^2 \right] \\
&\leq O_{P_n}(1)\sqrt{n}(\max_{i \in \mathbb{I}_t} |u_i|) \mathbb{E}_{n,t} \left[(\hat{p}_t(x_i) - p_t(x_i))^2 \right] = o_{P_n}(1),
\end{aligned}
$$

where the rate follows from Assumptions II.3(b), II.4, and II.5, and this tends to zero by Assumption II.6.

As in the prior proof, write $R_2 = R_{21} + R_{22}$. The same bound is used for R_{22}. However, for R_{21}, add and subtract the pseudotrue values to get $R_{21} = R_{211} + R_{212}$, where

$$
R_{211} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (x_i^{*\prime} \hat{\beta}_t - x_i^* \beta_t^*) \left(\frac{p_t(x_i) - d_i^t}{p_t(x_i)} \right)
$$

and

$$
R_{212} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (x_i^* \beta_t^* - \mu_t(x_i)) \left(\frac{p_t(x_i) - d_i^t}{p_t(x_i)} \right)
$$

For the first term, define $\tilde{\Sigma}_{t,j} = \mathbb{E} \left[(x_{i,j}^*)^2 (d_i^t - p_t(x_i))^2 / p_t(x_i)^2 \right]$ and then proceed as follows:

$$
\begin{aligned}
R_{211} &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{p_t(x_i) - d_i^t}{p_t(x_i)} \right) \sum_{j \in \hat{S}_Y} x_{i,j}^* (\hat{\beta}_{t,j} - \beta_{t,j}^*) \\
&= \sum_{j \in \hat{S}_Y} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{x_{i,j}^* (p_t(x_i) - d_i^t) / p_t(x_i)}{\tilde{\Sigma}_{t,j}^{1/2}} \right\} \tilde{\Sigma}_{t,j}^{1/2} (\hat{\beta}_{t,j} - \beta_{t,j}^*) \\
&\leq \left(\max_{j \in \mathbb{N}_p} \tilde{\Sigma}_{t,j}^{1/2} \right) \left(\max_{j \in \mathbb{N}_p} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{x_{i,j}^* (p_t(x_i) - d_i^t) / p_t(x_i)}{\tilde{\Sigma}_{t,j}^{1/2}} \right) \left\| \hat{\beta}_t - \beta_t^* \right\|_1 \\
&= O(1) O_{P_n}(\log(p)) \left\| \hat{\beta}_t - \beta_t^* \right\|_1 = o_{P_n}(1),
\end{aligned}
$$

where the final line follows exactly as above.

A variance bound may be applied to R_{212} as in the previous proof, and we have $|R_{212}| = O_{P_n}(b_s) = o_{P_n}(1)$ by Markov's inequality. □

Proof of Theorem II.8.2. This claim follows directly from the prior result under the moment conditions of Assumption II.4(e). □

Proof of Theorem II.8.3. We begin with $\hat{V}_W(t)$. Expanding the square and using Eqn.

(A.1), rewrite $\hat{V}^W_{\boldsymbol{\mu}}(t) = \mathbb{E}_n[d_i^t u_i^2 p_t(x_i)^{-2}] + R_{W,1} + R_{W,2} + R_{W,3}$ where

$$R_{W,1} = \mathbb{E}_n\left[\frac{d_i^t u_i^2}{\hat{p}_t(x_i)^2 p_t(x_i)^2}\left(\hat{p}_t(x_i) - p_t(x_i)\right)\left(\hat{p}_t(x_i) + p_t(x_i)\right)\right],$$

$$R_{W,2} = \mathbb{E}_n\left[\frac{d_i^t(\mu_t(x_i) - \hat{\mu}_t(x_i))^2}{\hat{p}_t(x_i)^2}\right], \qquad \text{and} \qquad R_{W,3} = 2\mathbb{E}_n\left[\frac{d_i^t u_i(\mu_t(x_i) - \hat{\mu}_t(x_i))}{\hat{p}_t(x_i)^2}\right].$$

Using Hölder's inequality, Assumptions II.3(b), II.4(e), and II.5(a), we have the following

$$R_{W,1} \leq \left(\max_{i\in\mathbb{I}_t}\frac{\hat{p}_t(x_i) + p_t(x_i)}{\hat{p}_t(x_i)^2 p_t(x_i)^2}\right)\mathbb{E}_n[d_i^t|u_i|^4]^{1/2}\mathbb{E}_n[d_i^t(\hat{p}_t(x_i) - p_t(x_i))^2]^{1/2} = o_{P_n}(1),$$

$$R_{W,2} \leq \left(\max_{i\in\mathbb{I}_t}\frac{1}{\hat{p}_t(x_i)^2}\right)\mathbb{E}_n[d_i^t(\hat{\mu}_t(x_i) - \mu_t(x_i))^2] = o_{P_n}(1),$$

and,

$$R_{W,3} \leq 2\left(\max_{i\in\mathbb{I}_t}\frac{1}{\hat{p}_t(x_i)^2}\right)\mathbb{E}_n[d_i^t|u_i|^2]^{1/2}\mathbb{E}_n[d_i^t(\hat{\mu}_t(x_i) - \mu_t(x_i))^2]^{1/2} = o_{P_n}(1),$$

where $\mathbb{E}_n[|u_i|^4] = O_{P_n}(1)$ from the inequality of von Bahr and Esseen (1965). From the same inequality it follows that $\mathbb{E}_n[d_i^t u_i^2 p_t(x_i)^{-2}] - V^W_{\boldsymbol{\mu}}(t)| = o_{P_n}(1)$, under Assumptions II.3(b) and II.4(c).

Next consider the "between" variance estimator, $\hat{V}^B_{\boldsymbol{\mu}}$. For any $t\overline{\mathbb{N}}_{\mathcal{T}}$ and $t' \in \overline{\mathbb{N}}_{\mathcal{T}}$, define

$$R_{B,1}(t,t') = \mathbb{E}_n\left[(\hat{\mu}_t(x_i) - \mu_t(x_i))(\hat{\mu}_{t'}(x_i) - \mu_{t'}(x_i))\right],$$

$$R_{B,2}(t,t') = \hat{\mu}_t\mathbb{E}_n\left[\hat{\mu}_{t'}(x_i) - \mu_{t'}(x_i)\right], \quad \text{and} \quad R_{B,3}(t,t') = \mathbb{E}_n\left[\mu_t(x_i)(\hat{\mu}_{t'}(x_i) - \mu_{t'}(x_i))\right].$$

From Hölder's inequality, Assumption II.5(a), Theorem II.8.2, the von Bahr and Esseen inequality, and Assumptions II.4(c) and II.4(e) it follows that $R_{B,k}(t,t') = o_{P_n}(1)$ for $k \in \mathbb{N}_3$ and all pairs $(t,t') \in \mathbb{N}_t^2$. With this in mind, we decompose

$$\begin{aligned}\hat{V}^B_{\boldsymbol{\mu}}(t,t') = {} & \mathbb{E}_n\left[\mu_t(x_i)\mu_{t'}(x_i)\right] - \hat{\mu}_t\mathbb{E}_n\left[\mu_{t'}(x_i)\right] - \hat{\mu}_{t'}\mathbb{E}_n\left[\mu_t(x_i)\right] + \hat{\mu}_t\hat{\mu}_{t'} \\ & + R_{B,1}(t,t') + R_{B,2}(t,t') + R_{B,2}(t',t) + R_{B,3}(t,t') + R_{B,3}(t',t).\end{aligned}$$

Consistency of $\hat{V}^B_{\boldsymbol{\mu}}(t,t')$ now follows from the von Bahr and Esseen inequality and Theorem II.8.2. □

Proof of Corollary II.9. Suppose the result did not hold. Then, there would exist a subsequence $P_m \in \mathcal{P}_m$, for each m, such that

$$\lim_{m\to\infty} \left| \mathbb{P}_{P_m} \left[G(\boldsymbol{\mu}) \in \left\{ G(\hat{\boldsymbol{\mu}}) \pm c_\alpha \sqrt{\nabla_G(\hat{\boldsymbol{\mu}}) \hat{V} \nabla_G'(\hat{\boldsymbol{\mu}})/n} \right\} \right] - (1-\alpha) \right| > 0.$$

But this contradicts Theorem II.8, under which $(\nabla_G(\hat{\boldsymbol{\mu}})\hat{V}\nabla_G'(\hat{\boldsymbol{\mu}})/n)^{-1/2}(G(\hat{\boldsymbol{\mu}}) - G(\boldsymbol{\mu}))$ is asymptotically standard normal under the sequence P_m. □

Proofs for Average Treatment Effects on Treated Groups

Proofs are similar to those for Theorem II.8 and Corollary II.9, and hence we omit them to save space.

Proofs for Chapter 2: Group Lasso Selection and Estimation

Unless otherwise noted, all bounds in this section are nonasymptotic. Further, as the proofs are segregated we will use the generic notation X^* and s for the covariates and sparsity level.

Proofs for Multinomial Logistic Models

Lemmas

Lemma A.1 (Score Bound)**.** *For λ_D and $\mathcal{P}$ defined respectively in Eqn.* (2.14) *and Eqn.* (2.15) *we have*

$$\mathbb{P}\left[\max_{j \in \mathbb{N}_p} \|\mathbb{E}_n[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}) - d_i^t)x_{i,j}^*]\|_2 \geq \frac{\lambda_D}{2} \right] \leq \mathcal{P}.$$

Proof. First, by the Cauchy-Schwarz inequality and Assumption II.4(b) and the bias condition, we have

$$\begin{aligned} &\max_{j \in \mathbb{N}_p} \|\mathbb{E}_n[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}) - p_t(x_i))x_{i,j}^*]\|_2 \\ &\qquad\qquad \leq \mathcal{X}\|\mathbb{E}_n[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}) - p_t(x_i))^2]^{1/2}]\|_2 \leq \mathcal{X} b_s^d \sqrt{\mathcal{T}}. \end{aligned}$$

Therefore, by the triangle inequality and the definition of λ_D, with $r_n = \mathcal{T}^{-1/2}\log(p \vee$

$\underline{n})^{3/2+\delta}$,

$$\begin{aligned}
&\mathbb{P}\left[\max_{j\in\mathbb{N}_p}\|\mathbb{E}_n[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})-d_i^t)x_{i,j}^*]\|_2 \geq \frac{\lambda_D}{2}\right]\\
&\quad\leq \mathbb{P}\left[\max_{j\in\mathbb{N}_p}\|\mathbb{E}_n[v_{t,i}x_{i,j}^*]\|_2 + \max_{j\in\mathbb{N}_p}\|\mathbb{E}_n[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})-p_t(x_i))x_{i,j}^*]\|_2 \geq \frac{\lambda_D}{2}\right]\\
&\quad= \mathbb{P}\Bigg[\max_{j\in\mathbb{N}_p}\|\mathbb{E}_n[v_{t,i}x_{i,j}^*]\|_2 + \max_{j\in\mathbb{N}_p}\|\mathbb{E}_n[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})-p_t(x_i))x_{i,j}^*]\|_2\\
&\qquad\qquad\qquad \geq \mathcal{X}\sqrt{\mathcal{T}}\left[b_s^d + \frac{1}{\sqrt{\underline{n}}}\left(1+r_n\right)^{1/2}\right]\Bigg]\\
&\quad= \mathbb{P}\left[\max_{j\in\mathbb{N}_p}\|\mathbb{E}_n[v_{t,i}x_{i,j}^*]\|_2 \geq \frac{\mathcal{X}\sqrt{\mathcal{T}}}{\sqrt{\underline{n}}}\left(1+r_n\right)^{1/2}\right]\\
&\quad= \mathbb{P}\left[\max_{j\in\mathbb{N}_p}\|\mathbb{E}_n[v_{t,i}x_{i,j}^*]\|_2^2 \geq \frac{\mathcal{X}^2\mathcal{T}}{\underline{n}}\left(1+r_n\right)\right],
\end{aligned}$$

canceling the bias terms from each side the squaring.

The residuals $v_{t,i}$ are conditionally mean-zero by definition and satisfy $\mathbb{E}[v_{t,i}^2|x_i] \leq 1$. Using this, Assumption II.4(a) and the definition of $\mathcal{X}$, we find that

$$\mathbb{E}\left[\|\mathbb{E}_n[v_{t,i}x_{i,j}^*]\|_2^2\right] = \sum_{t\in\mathbb{N}_\mathcal{T}}\mathbb{E}\left[\mathbb{E}_n[v_{t,i}x_{i,j}^*]^2\right] = \sum_{t\in\mathbb{N}_\mathcal{T}}\frac{1}{n}\mathbb{E}[v_{t,i}^2(x_{i,j}^*)^2] \leq \frac{\mathcal{X}^2\mathcal{T}}{n}$$

uniformly in $j \in \mathbb{N}_p$. Define the mean-zero random variables $\xi_{t,j}$ as:

$$\xi_{t,j} = (\mathbb{E}_n[v_{t,i}x_{i,j}^*])^2 - \frac{1}{n}\mathbb{E}[V_t^2 X_j^{*2}].$$

Thus, we further bound the probability as follows.

$$\begin{aligned}
\mathbb{P}\left[\max_{j\in\mathbb{N}_p}\|\mathbb{E}_n[v_{t,i}x_{i,j}^*]\|_2 \geq \frac{\mathcal{X}^2\mathcal{T}}{\underline{n}}\left(1+r_n\right)\right] &= \mathbb{P}\left[\max_{j\in\mathbb{N}_p}\|\mathbb{E}_n[v_{t,i}x_{i,j}^*]\|_2^2 - \frac{\mathcal{X}^2\mathcal{T}}{n} \geq \frac{\mathcal{X}^2\mathcal{T}r_n}{n}\right]\\
&\leq \mathbb{P}\left[\max_{j\in\mathbb{N}_p}\sum_{t\in\mathbb{N}_\mathcal{T}}\xi_{t,j} \geq \frac{\mathcal{X}^2\mathcal{T}r_n}{n}\right]\\
&\leq \mathbb{E}\left[\max_{j\in\mathbb{N}_p}\left|\sum_{t\in\mathbb{N}_\mathcal{T}}\xi_{t,j}\right|\right]\frac{n}{\mathcal{X}^2\mathcal{T}r_n}, \qquad \text{(A.2)}
\end{aligned}$$

where final line follows from Markov's inequality.

Next, applying Lemma 9.1 of Lounici, Pontil, van de Geer, and Tsybakov (2011) (with their $m = 1$ and hence $c(m) = 2$) followed by Jensen's inequality and Assumption II.4(c), we find that

$$
\begin{aligned}
\mathbb{E}\left[\max_{j\in\mathbb{N}_p}\left|\sum_{t\in\mathbb{N}_\mathcal{T}}\xi_{t,j}\right|\right] &\le (8\log(2p))^{1/2}\mathbb{E}\left[\left(\sum_{t\in\mathbb{N}_\mathcal{T}}\max_{j\in\mathbb{N}_p}\xi_{t,j}^2\right)^{1/2}\right] \\
&\le (8\log(2p))^{1/2}\left(\mathbb{E}\left[\sum_{t\in\mathbb{N}_\mathcal{T}}\max_{j\in\mathbb{N}_p}\xi_{t,j}^2\right]\right)^{1/2} \\
&\le 4\log(2p)^{1/2}\left(\sum_{t\in\mathbb{N}_\mathcal{T}}\frac{\mathcal{X}^4}{n^2}+\sum_{t\in\mathbb{N}_\mathcal{T}}\mathbb{E}\left[\max_{j\in\mathbb{N}_p}\left|\mathbb{E}_n[v_{t,i}x_{i,j}^*]\right|^4\right]\right)^{1/2}. \quad \text{(A.3)}
\end{aligned}
$$

The leading 4 is $\sqrt{8}\sqrt{2}$, where $\sqrt{2}$ is a byproduct of applying the inequality $(a-b)^2 \le 2(a^2+b^2)$ to $\xi_{t,j}^2$. Again using Lemma 9.1 of Lounici, Pontil, van de Geer, and Tsybakov (2011) (with their $m=4$, and $c(m)=12$ since $c(4) \ge (e^{4-1}-1)/2+2 \approx 11.54$), we bound the expectation in the second term above as follows:

$$
\mathbb{E}\left[\max_{j\in\mathbb{N}_p}\left|\mathbb{E}_n[v_{t,i}x_{i,j}^*]\right|^4\right] \le [8\log(12p)]^{4/2}\mathbb{E}\left[\left(\sum_{i=1}^{n}\max_{j\in\mathbb{N}_p}\left|\frac{v_{t,i}x_{i,j}^*}{n}\right|^2\right)^{4/2}\right] \le \frac{64\log(12p)^2\mathcal{X}^4}{n^2}, \tag{A.4}
$$

using Assumptions II.4(a) and II.4(b).

Now, inserting the results of Eqns. (A.3) and (A.4) into Eqn. (A.2), we have

$$
\begin{aligned}
\mathbb{P}\left[\max_{j\in\mathbb{N}_p}\|\mathbb{E}_n[v_{t,i}x_{i,j}^*]\|_2 \ge \frac{\lambda_D}{4}\right] &\le \frac{4n\log(2p)^{1/2}}{\mathcal{T}\mathcal{X}^2 r_n}\left(\sum_{t\in\mathbb{N}_\mathcal{T}}\frac{\mathcal{X}^4}{n^2}+\sum_{t\in\mathbb{N}_\mathcal{T}}\frac{64\log(12p)^2\mathcal{X}^4}{n^2}\right)^{1/2} \\
&\le \frac{4\log(2p)^{1/2}}{r_n\sqrt{\mathcal{T}}}[1+64\log(12p)^2]^{1/2} = \mathcal{P},
\end{aligned}
$$

using the choice $r_n = \mathcal{T}^{-1/2}\log(p\vee \underline{n})^{3/2+\delta}$. □

Lemma A.2 (Estimate Sparsity)**.** *With probability at least* $1-\mathcal{P}$

$$
|\tilde{S}^D| \le \frac{4}{\lambda_D^2}\overline{\phi}\{Q,\tilde{S}^D\}\sum_{t\in\mathbb{N}_\mathcal{T}}\mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\tilde{\gamma}_t\}_{\mathbb{N}_\mathcal{T}})-\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}))^2\right].
$$

Proof. First, by Karush-Kuhn-Tucker conditions for (2.9), for all $t\in\mathbb{N}_\mathcal{T}$, if $\tilde{\gamma}_{\cdot,j}\neq 0$ it must satisfy

$$
\mathbb{E}_n[x_{i,j}^*(\hat{p}_t(\{x_i^{*\prime}\tilde{\gamma}_t\}_{\mathbb{N}_\mathcal{T}})-d_i^t)] = \lambda_D\frac{\tilde{\gamma}_{t,j}}{\|\tilde{\gamma}_{\cdot,j}\|_2}. \tag{A.5}
$$

Hence, taking the ℓ_2-norm over $t \in \mathbb{N}_{\mathcal{T}}$ for fixed $j \in \tilde{S}^D$, adding and subtracting the true propensity score, using the triangle inequality, and the score bound (A.1), we find that

$$\begin{aligned}\lambda_D &= \left\|\mathbb{E}_n[x^*_{i,j}(\hat{p}_t(\{x^{*\prime}_i\tilde{\gamma}_t\}_{\mathbb{N}_{\mathcal{T}}}) - d^t_i)]\right\|_2 \\ &\le \left\|\mathbb{E}_n[x^*_{i,j}(\hat{p}_t(\{x^{*\prime}_i\gamma^*_t\}_{\mathbb{N}_{\mathcal{T}}}) - d^t_i)]\right\|_2 + \left\|\mathbb{E}_n[x^*_{i,j}(\hat{p}_t(\{x^{*\prime}_i\tilde{\gamma}_t\}_{\mathbb{N}_{\mathcal{T}}}) - \hat{p}_t(\{x^{*\prime}_i\gamma^*_t\}_{\mathbb{N}_{\mathcal{T}}}))]\right\|_2 \\ &\le \lambda_D/2 + \left\|\mathbb{E}_n[x^*_{i,j}(\hat{p}_t(\{x^{*\prime}_i\tilde{\gamma}_t\}_{\mathbb{N}_{\mathcal{T}}}) - \hat{p}_t(\{x^{*\prime}_i\gamma^*_t\}_{\mathbb{N}_{\mathcal{T}}}))]\right\|_2 .\end{aligned}$$

Let $\boldsymbol{P}^*_t$ be the vector of $\{\hat{p}_t(\{x^{*\prime}_i\gamma^*_t\}_{\mathbb{N}_{\mathcal{T}}})\}_{i=1}^n$ and similarly for $\tilde{\boldsymbol{P}}_t$. Collecting terms, then squaring both sides and and summing over $j \in \tilde{S}^D$ (i.e. applying $\|\cdot\|_2^2$ over $j \in \tilde{S}^D$ to both sides) yields

$$\begin{aligned}\sum_{j\in\tilde{S}^D}\lambda_D^2 &\le 4\sum_{j\in\tilde{S}^D}\sum_{t\in\mathbb{N}_{\mathcal{T}}}\mathbb{E}_n[x^*_{i,j}(\hat{p}_t(\{x^{*\prime}_i\tilde{\gamma}_t\}_{\mathbb{N}_{\mathcal{T}}}) - \hat{p}_t(\{x^{*\prime}_i\gamma^*_t\}_{\mathbb{N}_{\mathcal{T}}}))]^2 \\ &= 4\sum_{t\in\mathbb{N}_{\mathcal{T}}}\frac{1}{n^2}\left\|\left[\boldsymbol{X}'(\tilde{\boldsymbol{P}}_t - \boldsymbol{P}^*_t)\right]_{j\in\tilde{S}^D}\right\|_2^2 \\ &\le 4\sum_{t\in\mathbb{N}_{\mathcal{T}}}\frac{\overline{\phi}\{Q,\tilde{S}^D\}}{n}\left\|\tilde{\boldsymbol{P}}_t - \boldsymbol{P}^*_t\right\|_{2,n}^2 \\ &\le 4\overline{\phi}\{Q,\tilde{S}^D\}\sum_{t\in\mathbb{N}_{\mathcal{T}}}\mathbb{E}_n\left[(\hat{p}_t(\{x^{*\prime}_i\tilde{\gamma}_t\}_{\mathbb{N}_{\mathcal{T}}}) - \hat{p}_t(\{x^{*\prime}_i\gamma^*_t\}_{\mathbb{N}_{\mathcal{T}}}))^2\right] .\end{aligned}$$

The result now follows, as the left-hand side is equal to $|\tilde{S}^D|\lambda_D^2$. □

Lemma A.3 (Cone Constraint)**.** *Define $\tilde{\delta}_{\cdot,\cdot} = \tilde{\gamma}_{\cdot,\cdot} - \gamma^*_{\cdot,\cdot}$. With probability $1-\mathcal{P}$, $\tilde{\delta}_{\cdot,\cdot}$ obeys the cone constraint required by the definition of κ_D.*

Proof. By the Cauchy-Schwarz inequality and Lemma A.1,

$$\begin{aligned}&\sum_{t\in\mathbb{N}_{\mathcal{T}}}\mathbb{E}_n\left[(\hat{p}_t(\{x^{*\prime}_i\gamma^*_t\}_{\mathbb{N}_{\mathcal{T}}}) - d^t_i)x^{*\prime}_i\tilde{\delta}_t\right] \\ &\qquad= \sum_{j\in\mathbb{N}_p}\sum_{t\in\mathbb{N}_{\mathcal{T}}}\mathbb{E}_n\left[(\hat{p}_t(\{x^{*\prime}_i\gamma^*_t\}_{\mathbb{N}_{\mathcal{T}}}) - d^t_i)x^*_{i,j}\right]\tilde{\delta}_{t,j} \\ &\qquad\le \sum_{j\in\mathbb{N}_p}\sqrt{\sum_{t\in\mathbb{N}_{\mathcal{T}}}\mathbb{E}_n\left[(\hat{p}_t(\{x^{*\prime}_i\gamma^*_t\}_{\mathbb{N}_{\mathcal{T}}}) - d^t_i)x^*_{i,j}\right]^2}\sqrt{\sum_{t\in\mathbb{N}_{\mathcal{T}}}\tilde{\delta}^2_{t,j}} \\ &\qquad\le \max_{j\in\mathbb{N}_p}\left\{\left\|\mathbb{E}_n\left[(\hat{p}_t(\{x^{*\prime}_i\gamma^*_t\}_{\mathbb{N}_{\mathcal{T}}}) - d^t_i)x^*_{i,j}\right]\right\|_2\right\}\sum_{j\in\mathbb{N}_p}\left\|\tilde{\delta}_{\cdot,j}\right\|_2 \\ &\qquad\le \frac{\lambda_D}{2}\left\|\left|\tilde{\delta}_{\cdot,\cdot}\right|\right\|_{2,1}, \qquad\qquad \text{(A.6)}\end{aligned}$$

with probability at least $1-\mathcal{P}$.

By the optimality of $\tilde{\delta}_{\cdot,\cdot}$, we have

$$\mathcal{M}(\gamma^*_{\cdot,\cdot}+\tilde{\delta}_{\cdot,\cdot})+\lambda_D\left|\!\left|\!\left|\gamma^*_{\cdot,\cdot}+\tilde{\delta}_{\cdot,\cdot}\right|\!\right|\!\right|_{2,1}\leq\mathcal{M}(\gamma^*_{\cdot,\cdot})+\lambda_D\left|\!\left|\!\left|\gamma^*_{\cdot,\cdot}\right|\!\right|\!\right|_{2,1},$$

implying

$$\begin{aligned}\lambda_D\left\{\left|\!\left|\!\left|\gamma^*_{\cdot,\cdot}\right|\!\right|\!\right|_{2,1}-\left|\!\left|\!\left|\gamma^*_{\cdot,\cdot}+\tilde{\delta}_{\cdot,\cdot}\right|\!\right|\!\right|_{2,1}\right\}&\geq\mathcal{M}(\gamma^*_{\cdot,\cdot}+\tilde{\delta}_{\cdot,\cdot})-\mathcal{M}(\gamma^*_{\cdot,\cdot})\\&\geq\sum_{t\in\mathbb{N}_{\mathcal{T}}}\mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_{\mathcal{T}}})-d_i^t)x_i^{*\prime}\tilde{\delta}_t\right],\end{aligned}$$

applying the convexity of $\mathcal{M}$. Using the bound in Eqn. (A.6) and rearranging we find that

$$0\leq\lambda_D\left\{\left|\!\left|\!\left|\gamma^*_{\cdot,\cdot}\right|\!\right|\!\right|_{2,1}-\left|\!\left|\!\left|\gamma^*_{\cdot,\cdot}+\tilde{\delta}_{\cdot,\cdot}\right|\!\right|\!\right|_{2,1}\right\}+\frac{\lambda_D}{2}\left|\!\left|\!\left|\tilde{\delta}_{\cdot,\cdot}\right|\!\right|\!\right|_{2,1}.$$

Canceling λ_D and decomposing the supports, we find that

$$\begin{aligned}0&\leq\frac{1}{2}\left|\!\left|\!\left|\tilde{\delta}_{\cdot,\cdot}\right|\!\right|\!\right|_{2,1}+\left\{\left|\!\left|\!\left|\gamma^*_{\cdot,\cdot}\right|\!\right|\!\right|_{2,1}-\left|\!\left|\!\left|\gamma^*_{\cdot,\cdot}+\tilde{\delta}_{\cdot,\cdot}\right|\!\right|\!\right|_{2,1}\right\}\\&=\frac{1}{2}\left|\!\left|\!\left|\tilde{\delta}_{\cdot,S_*}\right|\!\right|\!\right|_{2,1}+\frac{1}{2}\left|\!\left|\!\left|\tilde{\delta}_{\cdot,S_*^c}\right|\!\right|\!\right|_{2,1}+\left|\!\left|\!\left|\gamma^*_{\cdot,S_*}\right|\!\right|\!\right|_{2,1}-\left|\!\left|\!\left|\gamma^*_{\cdot,S_*}+\tilde{\delta}_{\cdot,S_*}\right|\!\right|\!\right|_{2,1}-\left|\!\left|\!\left|\tilde{\delta}_{\cdot,S_*^c}\right|\!\right|\!\right|_{2,1},\end{aligned}$$

where the second line follows because $\gamma^*_{\cdot,S_*^c}=0$. Collecting terms and applying the triangle inequality yields

$$\begin{aligned}\frac{1}{2}\left|\!\left|\!\left|\tilde{\delta}_{\cdot,S_*^c}\right|\!\right|\!\right|_{2,1}&\leq\frac{1}{2}\left|\!\left|\!\left|\tilde{\delta}_{\cdot,S_*}\right|\!\right|\!\right|_{2,1}+\left|\!\left|\!\left|\gamma^*_{\cdot,S_*}\right|\!\right|\!\right|_{2,1}-\left|\!\left|\!\left|\gamma^*_{\cdot,S_*}+\tilde{\delta}_{\cdot,S_*}\right|\!\right|\!\right|_{2,1}\\&\leq\frac{1}{2}\left|\!\left|\!\left|\tilde{\delta}_{\cdot,S_*}\right|\!\right|\!\right|_{2,1}+\left|\left|\!\left|\!\left|\gamma^*_{\cdot,S_*}\right|\!\right|\!\right|_{2,1}-\left|\!\left|\!\left|\gamma^*_{\cdot,S_*}+\tilde{\delta}_{\cdot,S_*}\right|\!\right|\!\right|_{2,1}\right|\\&\leq\frac{1}{2}\left|\!\left|\!\left|\tilde{\delta}_{\cdot,S_*}\right|\!\right|\!\right|_{2,1}+\left|\!\left|\!\left|\gamma^*_{\cdot,S_*}-\left(\gamma^*_{\cdot,S_*}+\tilde{\delta}_{\cdot,S_*}\right)\right|\!\right|\!\right|_{2,1}\\&=\frac{1}{2}\left|\!\left|\!\left|\tilde{\delta}_{\cdot,S_*}\right|\!\right|\!\right|_{2,1}+\left|\!\left|\!\left|\tilde{\delta}_{\cdot,S_*}\right|\!\right|\!\right|_{2,1}.\end{aligned}$$

Hence $\tilde{\delta}_{\cdot,\cdot}$ belongs to the restricted set of (2.17). □

Proof of Theorem II.12

Define $\tilde{\delta}_{\cdot,\cdot}=\tilde{\gamma}_{\cdot,\cdot}-\gamma^*_{\cdot,\cdot}$. By the optimality of $\tilde{\delta}_{\cdot,\cdot}$, we have

$$+\lambda_D\left|\!\left|\!\left|\gamma^*_{\cdot,\cdot}+\tilde{\delta}_{\cdot,\cdot}\right|\!\right|\!\right|_{2,1}\leq\mathcal{M}(\gamma^*_{\cdot,\cdot})+\lambda_D\left|\!\left|\!\left|\gamma^*_{\cdot,\cdot}\right|\!\right|\!\right|_{2,1}.$$

Rearranging and subtracting the score, we have

$$\begin{aligned}\mathcal{M}(\gamma^*_{\cdot,\cdot} + \tilde{\delta}_{\cdot,\cdot}) - \mathcal{M}(\gamma^*_{\cdot,\cdot}) &- \sum_{t\in\mathbb{N}_\mathcal{T}} \mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}) - d_i^t)x_i^{*\prime}\right]\tilde{\delta}_t \\ &\le \lambda_D\left\{|||\gamma^*_{\cdot,\cdot}|||_{2,1} - \left|\left|\left|\gamma^*_{\cdot,\cdot} + \tilde{\delta}_{\cdot,\cdot}\right|\right|\right|_{2,1}\right\} - \sum_{t\in\mathbb{N}_\mathcal{T}} \mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}) - d_i^t)x_i^{*\prime}\right]\tilde{\delta}_t.\end{aligned} \tag{A.7}$$

The proof proceeds by deriving a further upper bound to the right and a quadratic lower bound of the left. The combination of these will yield a bound on $\mathbb{E}_n[(x_i^{*\prime}\tilde{\delta}_t)^2]^{1/2}$.

Let us begin with the right side of Eqn. (A.7). For the penalized difference of coefficients we have

$$|||\gamma^*_{\cdot,S_*^c}|||_{2,1} - \left|\left|\left|\gamma^*_{\cdot,S_*^c} + \tilde{\delta}_{\cdot,S_*^c}\right|\right|\right|_{2,1} = \left|\left|\left|\tilde{\delta}_{\cdot,S_*^c}\right|\right|\right|_{2,1},$$

because $\gamma^*_{\cdot,S_*^c} = 0$. Therefore,

$$\begin{aligned}\left|\,|||\gamma^*_{\cdot,\cdot}|||_{2,1} - \left|\left|\left|\gamma^*_{\cdot,\cdot} + \tilde{\delta}_{\cdot,\cdot}\right|\right|\right|_{2,1}\right| &= \left|\,|||\gamma^*_{\cdot,S_*}|||_{2,1} - \left|\left|\left|\gamma^*_{\cdot,S_*} + \tilde{\delta}_{\cdot,S_*}\right|\right|\right|_{2,1} - \left|\left|\left|\tilde{\delta}_{\cdot,S_*^c}\right|\right|\right|_{2,1}\right| \\ &\le \left|\,|||\gamma^*_{\cdot,S_*}|||_{2,1} - \left|\left|\left|\gamma^*_{\cdot,S_*} + \tilde{\delta}_{\cdot,S_*}\right|\right|\right|_{2,1}\right| \\ &\le \left|\,\left|\left|\left|\gamma^*_{\cdot,S_*} - \left(\gamma^*_{\cdot,S_*} + \tilde{\delta}_{\cdot,S_*}\right)\right|\right|\right|_{2,1}\right| \\ &= \left|\left|\left|\tilde{\delta}_{\cdot,S_*}\right|\right|\right|_{2,1},\end{aligned}$$

where the second step follows from the triangle inequality and dropping the nonnegative norm, and the third by the triangle inequality again. Thus, using this result for the first term and the bound (A.6) for the second, we find that the right side of Eqn. (A.7) is bounded as follows, using the cone constraint, the Cauchy-Schwarz inequality, and the definition of κ_D from Eqn. (2.17),

$$\begin{aligned}\lambda_D\left|\left|\left|\tilde{\delta}_{\cdot,S_*}\right|\right|\right|_{2,1} + \frac{\lambda_D}{2}\left|\left|\left|\tilde{\delta}_{\cdot,\cdot}\right|\right|\right|_{2,1} &\le \lambda_D\left|\left|\left|\tilde{\delta}_{\cdot,S_*}\right|\right|\right|_{2,1} + \frac{\lambda_D}{2}\left|\left|\left|\tilde{\delta}_{\cdot,S_*}\right|\right|\right|_{2,1} + \frac{\lambda_D}{2}3\left|\left|\left|\tilde{\delta}_{\cdot,S_*}\right|\right|\right|_{2,1} \\ &= 3\lambda_D\left|\left|\left|\tilde{\delta}_{\cdot,S_*}\right|\right|\right|_{2,1} \\ &\le 3\lambda_D\sqrt{|S_*|}\left\|\tilde{\delta}_{\cdot,S_*}\right\|_2 \\ &\le \frac{3\lambda_D\sqrt{|S_*|}}{\kappa_D}\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2}.\end{aligned} \tag{A.8}$$

Note that $\sum_{t\in\mathbb{N}_\mathcal{T}} \tilde{\delta}_t' Q \tilde{\delta}_t = \mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]$.

Now turn to the left side of Eqn. (A.7). Our goal is to show that this is bounded below by a quadratic function. We apply the bounds for Bach's (2010) modified self-concordant functions. To show that $\mathcal{M}(\cdot)$ belongs to this class, we must bound the third derivative in terms of the Hessian. Recall that

$$\hat{p}_t(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}}) = \exp\{x_i^{*\prime}\gamma_t\} / \left(1 + \sum_{\mathbb{N}_\mathcal{T}} \exp\{x_i^{*\prime}\gamma_t\}\right).$$

Define the $\mathcal{T}$-square matrix $\mathcal{H}(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})$ as having the $(t,t') \in \mathbb{N}_\mathcal{T}^2$ entry given by

$$\mathcal{H}(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})_{[t,t']} = \begin{cases} \hat{p}_t(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})(1 - \hat{p}_t(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})) & \text{if } t = t' \\ -\hat{p}_t(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})\hat{p}_{t'}(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}}) & \text{if } t \neq t' \end{cases}$$

First, note that $\mathcal{M}(\gamma_{\cdot,\cdot})$ can be written as

$$\mathcal{M}(\gamma_{\cdot,\cdot}) = \mathbb{E}_n\left[\log\left(1 + \sum_{t\in\mathbb{N}_\mathcal{T}} \exp\{x_i^{*\prime}\gamma_t\}\right) - \sum_{t\in\mathbb{N}_\mathcal{T}} d_i^t(x_i^{*\prime}\gamma_t)\right].$$

Define $F : \mathbb{R}^\mathcal{T} \to \mathbb{R}$ as $F(w) = \log\left(1 + \sum_{t\in\mathbb{N}_\mathcal{T}} \exp(w_t)\right)$, so that

$$\mathcal{M}(\gamma_{\cdot,\cdot}) = \mathbb{E}_n\left[F(w_i) - \sum_{t\in\mathbb{N}_\mathcal{T}} d_i^t w_{i,t}\right]$$

, where $w_{i,t} = x_i^{*\prime}\gamma_t$ and $w_i = \{w_{i,t}\}_{\mathbb{N}_\mathcal{T}}$. Then for any $w \in \mathbb{R}^\mathcal{T}$, $v \in \mathbb{R}^\mathcal{T}$, and scalar α, define $g(\alpha) = F(w + \alpha v) : \mathbb{R} \to \mathbb{R}$. We verify the conditions of Bach (2010, Lemma 1) for this $g(\alpha)$ and $F(w)$. This involves finding the third derivative of $g(\alpha)$, and bounding it in terms of the second (i.e. the Hessian). To this end, note that he multinomial function has the property that $\partial\hat{p}_t(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})/\partial\gamma_t = \hat{p}_t(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})(1 - \hat{p}_t(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}}))x_i^*$ and $\partial\hat{p}_t(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})/\partial\gamma_{t',\cdot} = -\hat{p}_t(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})\hat{p}_{t'}(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})x_i^*$. From these, we find that

$$g'(\alpha) = v'F'(w + \alpha v) = \sum_{t\in\mathbb{N}_\mathcal{T}} v_t \hat{p}_t(w + \alpha v)$$

and

$$g''(\alpha) = v'F''(w + \alpha v)v = v'\mathcal{H}(w + \alpha v)v.$$

To bound $g'''(\alpha)$, we again use the derivatives of $\hat{p}_t(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})$ to find the derivatives of elements $\mathcal{H}(w)$. Routine calculations give, for any $r \neq s \neq t$:

$$\begin{aligned}
\partial\mathcal{H}(w)_{t,t}/\partial w_t &= \hat{p}_t(w)(1-\hat{p}_t(w))(1-2\hat{p}_t(w)) = \mathcal{H}(w)_{t,t}(1-2\hat{p}_t(w)) \\
\partial\mathcal{H}(w)_{t,t}/\partial w_r &= -\hat{p}_t(w)\hat{p}_r(w)(1-\hat{p}_t(w)) + \hat{p}_t(w)^2\hat{p}_r(w) \\
&= \mathcal{H}(w)_{t,t}(\hat{p}_t(w)\hat{p}_r(w)(1-\hat{p}_t(w))^{-1} - \hat{p}_r(w)) \\
\partial\mathcal{H}(w)_{t,s}/\partial w_t &= -\hat{p}_t(w)\hat{p}_s(w)(1-2\hat{p}_t(w)) = \mathcal{H}(w)_{t,s}(1-2\hat{p}_t(w)) \\
\partial\mathcal{H}(w)_{t,s}/\partial w_r &= -\hat{p}_t(w)\hat{p}_s(w)(-2\hat{p}_r(w)) = \mathcal{H}(w)_{t,s}(-2\hat{p}_r(w)).
\end{aligned}$$

Each derivative returns the same Hessian element multiplied by term bounded by 2 in absolute value. Let a_r represent this factor. Then we bound

$$\begin{aligned}
g'''(\alpha) &= \left| \sum_{r\in\mathbb{N}_{\mathcal{T}}} v_r \left.\frac{\partial v'\mathcal{H}(\tilde{w})v}{\partial w_r}\right|_{\tilde{w}=w+\alpha v} \right| = \left| \sum_{r\in\mathbb{N}_{\mathcal{T}}} v_r v'\mathcal{H}(w+\alpha v)v a_r \right| \\
&\leq \sum_{r\in\mathbb{N}_{\mathcal{T}}} v'\mathcal{H}(w+\alpha v)v|v_r||a_r| \leq 2v'\mathcal{H}(w+\alpha v)v \sum_{r\in\mathbb{N}_{\mathcal{T}}} |v_r| = 2\|v\|_1 g''(\alpha) \leq 2\sqrt{\mathcal{T}}\|v\|_2 g''(\alpha).
\end{aligned}$$

Applying Bach's (2010) Lemma 1 with $w_i = \{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_{\mathcal{T}}}$ and $v_i = \{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}$ we get the lower bound

$$\begin{aligned}
M(\gamma^*_{\cdot,\cdot} + \tilde{\delta}_{\cdot,\cdot}) - \mathcal{M}(\gamma^*_{\cdot,\cdot}) - \sum_{t\in\mathbb{N}_{\mathcal{T}}} \mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_{\mathcal{T}}}) - d_i^t)x_i^{*\prime}\right]\tilde{\delta}_t \\
\geq \mathbb{E}_n\left[\frac{v_i'\mathcal{H}(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_{\mathcal{T}}})v_i}{4\mathcal{T}\|v_i\|_2^2}\left(e^{-2\|v_i\|_2} + 2\|v_i\|_2 - 1\right)\right] \\
\geq \mathbb{E}_n\left[\frac{v_i'\mathcal{H}(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_{\mathcal{T}}})v_i}{4\mathcal{T}\|v_i\|_2^2}\left(2\|v_i\|_2^2 - \frac{4}{3}\|v_i\|_2^3\right)\right], \qquad \text{(A.9)}
\end{aligned}$$

where the second inequality follows from Belloni, Chernozhukov, and Wei (2013, Lemma 9).

Tanabe and Sagae (1992, Theorem 1) give $\mathcal{H}(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_{\mathcal{T}}}) \geq \phi_{\min}\{\mathcal{H}(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_{\mathcal{T}}})\}\mathcal{I}_{\mathcal{T}}$, in the positive definite sense, where $\phi_{\min}(A)$ denotes the smallest eigenvalue of A and $\mathcal{I}_T$ is the $\mathcal{T} \times \mathcal{T}$ identity matrix. Then

$$\phi_{\min}\{\mathcal{H}(\{x_i^*\gamma_t^*\}_{\mathbb{N}_{\mathcal{T}}})\} \geq \det\{\mathcal{H}(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_{\mathcal{T}}})\} = \prod_{t\in\overline{\mathbb{N}}_{\mathcal{T}}} \hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_{\mathcal{T}}}) \geq \left(\frac{p_{\min}}{A_p}\right)^{\overline{\mathcal{T}}},$$

where $p_0(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_{\mathcal{T}}}) = 1 - \sum_{t\in\mathbb{N}_{\mathcal{T}}} \hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_{\mathcal{T}}})$ and the first inequality is also due to

Tanabe and Sagae (1992). These results imply that

$$v_i'\mathcal{H}(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})v_i \geq (p_{\min}/A_p)^{\overline{\mathcal{T}}}v_i'\mathcal{I}_\mathcal{T}v_i = (p_{\min}/A_p)^{\overline{\mathcal{T}}}\|v_i\|_2^2$$

and therefore

$$\begin{aligned}
&\mathbb{E}_n\left[\frac{v_i'\mathcal{H}(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})v_i}{4\mathcal{T}\|v_i\|_2^2}\left(2\|v_i\|_2^2-\frac{4}{3}\|v_i\|_2^3\right)\right]\\
&\geq\left(\frac{p_{\min}}{A_p}\right)^{\overline{\mathcal{T}}}\frac{1}{4\mathcal{T}}\mathbb{E}_n\left[2\|v_i\|_2^2-\frac{4}{3}\|v_i\|_2^3\right]\\
&=\left(\frac{p_{\min}}{A_p}\right)^{\overline{\mathcal{T}}}\frac{1}{\mathcal{T}}\frac{\mathbb{E}_n[\|v_i\|_2^2]}{2}\left(1-\frac{2}{3}\frac{\mathbb{E}_n[\|v_i\|_2^3]}{\mathbb{E}_n[\|v_i\|_2^2]}\right). \qquad \text{(A.10)}
\end{aligned}$$

Recall that $v_i=\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}$. To prove a quadratic lower bound, consider two cases, depending on whether

$$\frac{1}{2}\left(1-\frac{2}{3}\frac{\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^3]}{\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]}\right)$$

is above or below $1/A_K$.

In the first case, combining Equations (A.9) and (A.10) gives

$$\begin{aligned}
&M(\gamma_{\cdot,\cdot}^*+\tilde{\delta}_{\cdot,\cdot})-\mathcal{M}(\gamma_{\cdot,\cdot}^*)-\sum_{t\in\mathbb{N}_\mathcal{T}}\mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})-d_i^t)x_i^{*\prime}\right]\tilde{\delta}_t\\
&\qquad\geq\left(\frac{p_{\min}}{A_p}\right)^{\overline{\mathcal{T}}}\frac{1}{\mathcal{T}}\frac{\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]}{A_K}. \qquad \text{(A.11)}
\end{aligned}$$

Now consider the second case, where this bound does not hold. By Lemma A.3, $\tilde{\delta}_{\cdot,\cdot}$ is in the cone defined by (2.17), and therefore

$$\begin{aligned}
&\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_1=\sum_{t\in\mathbb{N}_\mathcal{T}}\sum_{j\in\mathbb{N}_p}\left|x_{i,j}^*\tilde{\delta}_{t,j}\right|\leq\mathcal{X}\left\|\tilde{\delta}_{\cdot,\cdot}\right\|_1\leq\sqrt{\mathcal{T}}\mathcal{X}\left|\left|\left|\tilde{\delta}_{\cdot,\cdot}\right|\right|\right|_{2,1}\\
&=\sqrt{\mathcal{T}}\mathcal{X}4\left|\left|\left|\tilde{\delta}_{\cdot,S_*}\right|\right|\right|_{2,1}\leq\sqrt{\mathcal{T}}\mathcal{X}4\sqrt{|S_*|}\left\|\tilde{\delta}_{\cdot,S_*}\right\|_2\leq\sqrt{\mathcal{T}}\mathcal{X}4\sqrt{|S_*|}\kappa_D^{-1}\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2},
\end{aligned}$$

using Assumption II.4(b), the Cauchy-Schwarz inequality, decomposing the support of $\delta_{\cdot,\cdot}$, and then following the same steps as (A.8). Hence, by subadditivity,

$$\begin{aligned}
&\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^3]\leq\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_1]\\
&\qquad\leq\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{3/2}\sqrt{\mathcal{T}}\mathcal{X}4\sqrt{|S_*|}\kappa_D^{-1}.
\end{aligned}$$

Thus

$$\frac{1}{A_K} > \frac{1}{2}\left(1 - \frac{2}{3}\frac{\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^3]}{\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]}\right) \geq \frac{1}{2}\left(1 - \frac{\sqrt{\mathcal{T}}\mathcal{X}8\sqrt{|S_*|}}{3\kappa_D}\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2}\right),$$

which is equivalent to

$$\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2} > 3\left(1 - \frac{2}{A_K}\right)\frac{\kappa_D}{8\mathcal{X}\sqrt{\mathcal{T}}\sqrt{|S_*|}} \equiv r_n.$$

Because $\mathcal{M}(\gamma^*_{\cdot,\cdot} + \delta_{\cdot,\cdot}) - \mathcal{M}(\gamma_{\cdot,\cdot}) - \sum_{t\in\mathbb{N}_\mathcal{T}} \mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}) - d_i^t)x_i^{*\prime}\right]\delta_t$ is convex in $\delta_{\cdot,\cdot}$, and hence any line segment lies above the function, we have know that

$$\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2} > r_n$$

, we have

$$\begin{aligned}\mathcal{M}(\gamma^*_{\cdot,\cdot} + \tilde{\delta}_{\cdot,\cdot}) - \mathcal{M}(\gamma_{\cdot,\cdot}) - \sum_{t\in\mathbb{N}_\mathcal{T}} \mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}) - d_i^t)x_i^{*\prime}\right]\tilde{\delta}_t &\geq r_n^2 \\ &\geq r_n^2\frac{\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2}}{r_n} = r_n\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2}.\end{aligned}$$

Combining this result with Equations (A.7) and (A.8), we have

$$3\left(1 - \frac{2}{A_K}\right)\frac{\kappa_D}{8\mathcal{X}\sqrt{\mathcal{T}}\sqrt{|S_*|}}\mathbb{E}_n[\|\{x_i^{*\prime}\delta_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2} \leq \frac{3\lambda_D\sqrt{|S_*|}}{\kappa_D}\mathbb{E}_n[\|\{x_i^{*\prime}\delta_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2},$$

which is impossible under the restriction on A_K.

Therefore, Eqn. (A.11) must hold.[1] Combining this with Equations (A.7) and (A.8), we find that

$$\left(\frac{p_{\min}}{A_p}\right)^{\overline{\mathcal{T}}}\frac{1}{\mathcal{T}}\frac{\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]}{A_K} \leq \frac{3\lambda_D\sqrt{|S_*|}}{\kappa_D}\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2}.$$

Thus, dividing through and applying the union bound we find that

$$\max_{t\in\mathbb{N}_\mathcal{T}}\mathbb{E}_n[(x_i^{*\prime}\tilde{\delta}_t)^2]^{1/2} \leq \mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2} \leq \left(\frac{A_p}{p_{\min}}\right)^{\overline{\mathcal{T}}}\frac{3\mathcal{T}A_K\lambda_D\sqrt{|S_*|}}{\kappa_D}. \qquad \text{(A.12)}$$

[1] Intuitively, the deviation $\tilde{\delta}_{\cdot,\cdot}$ is too large for the quadratic bound to hold, and so this analysis is conceptually similar to using Belloni and Chernozhukov's (2011a) restricted nonlinearity impact coefficient, but our characterization is different.

To bound the propensity score error, we apply the mean value theorem and the form of $\partial \hat{p}_t(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})/\partial \gamma_t$. We must linearize with respect to t only (recall that $\hat{p}_t(\{x_i^{*\prime}\tilde{\gamma}_t\}_{\mathbb{N}_\mathcal{T}})$ depends on all of $\tilde{\gamma}_{\cdot,\cdot}$). To this end, define M_t as the $\mathcal{T}$-vector with entry t given by $x_i^{*\prime}\gamma_t^* + \tilde{m}_t x_i^{*\prime}\tilde{\gamma}_t$ for a scalar $\tilde{m}_t \in [0,1]$ and entries $t' \in \mathbb{N}_\mathcal{T} \setminus \{t\}$ equal to $x_i^{*\prime}\gamma_{t'}$. Then we have

$$\left|\hat{p}_t(\{x_i^{*\prime}\tilde{\gamma}_t\}_{\mathbb{N}_\mathcal{T}}) - \hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})\right| = \left|\hat{p}_t(M_t)[1-\hat{p}_t(M_t)]x_i^{*\prime}\tilde{\delta}_t\right| \leq \left|x_i^{*\prime}\tilde{\delta}_t\right|. \qquad \text{(A.13)}$$

Using this result coupled with the triangle inequality, the bias condition, and Eqn. (A.12), we find

$$\begin{aligned}
&\mathbb{E}_n[(\hat{p}_t(\{x_i^{*\prime}\tilde{\gamma}_t\}_{\mathbb{N}_\mathcal{T}}) - p_t(x_i))^2]^{1/2} \\
&\quad\leq \mathbb{E}_n[(\hat{p}_t(\{x_i^{*\prime}\tilde{\gamma}_t\}_{\mathbb{N}_\mathcal{T}}) - \hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}))^2]^{1/2} + \mathbb{E}_n[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}) - p_t(x_i))^2]^{1/2} \\
&\quad\leq \mathbb{E}_n\left[(x_i^{*\prime}\tilde{\delta}_t)^2\right]^{1/2} + b_s^d \\
&\quad\leq \left(\frac{A_p}{p_{\min}}\right)^{\overline{\mathcal{T}}} \frac{3\mathcal{T}A_K\lambda_D\sqrt{|S_*|}}{\kappa_D} + b_s^d.
\end{aligned}$$

The ℓ_1 bound follows from Eqn. (A.12) by the Cauchy-Schwarz inequality and the definition in Eqn. (2.19):

$$\|\tilde{\gamma}_t - \gamma_t^*\|_1 \leq \sqrt{|\tilde{S}^D \cup S_D^*|}\,\|\tilde{\gamma}_t - \gamma_t^*\|_{2,p} \leq \left(\frac{|\tilde{S}^D \cup S_D^*|}{\underline{\phi}\{Q, \tilde{S}^D \cup S_D^*\}}\right)^{1/2} \mathbb{E}_n[(x_i^{*\prime}(\tilde{\gamma}_t - \gamma_t^*))^2]^{1/2}.$$

Finally, we bound the size of the selected set of coefficients. First, note that optimality of $\tilde{\gamma}_{\cdot,\cdot}$ ensures that $|\tilde{S}^D| \leq n$. Then, restating the conclusion Lemma A.2 using the notation of the Theorem and the rate result (A.12), then bounding $\overline{\phi}$ by $\overline{\overline{\phi}}$ we find that

$$|\tilde{S}^D| \leq |S_D^*| 4L_n\overline{\overline{\phi}}\{Q, |\tilde{S}^D|\}.$$

The argument now parallels that used by Belloni and Chernozhukov (2011b), relying on their result on the sublinearity of sparse eigenvalues. Let $\lceil m \rceil$ be the ceiling function and note that $\lceil m \rceil \leq 2m$. For any $m \in \mathbb{N}_Q^D$, suppose that $|\tilde{S}^D| > m$. Then,

$$\begin{aligned}
|\tilde{S}^D| &\leq |S_D^*| 4L_n\overline{\overline{\phi}}\{Q, m(|\tilde{S}^D|/m)\} \\
&\leq \left\lceil |\tilde{S}^D|/m \right\rceil |S_D^*| 4L_n\overline{\overline{\phi}}\{Q, m\} \\
&\leq (|\tilde{S}^D|/m)|S_D^*| 8L_n\overline{\overline{\phi}}\{Q, m\}.
\end{aligned}$$

Rearranging gives

$$m \leq |S_D^*| 8 L_n \overline{\overline{\phi}}\{Q, m\}$$

whence $m \notin \mathbb{N}_Q^D$. Minimizing over $\mathbb{N}_Q^D$ gives the result. □

Proof of Theorem II.13

Define $\hat{\delta}_{\cdot,\cdot} = \hat{\gamma}_{\cdot,\cdot} - \gamma^*_{\cdot,\cdot}$. Many of the arguments parallel those for Theorem II.12. The key differences are that a quadratic lower bound for $\mathcal{M}(\gamma^*_{\cdot,\cdot} + \hat{\delta}_{\cdot,\cdot}) - \mathcal{M}(\gamma^*_{\cdot,\cdot}) - \sum_{t \in \mathbb{N}_{\mathcal{T}}} \mathbb{E}_n \left[(\hat{p}_t(\{x_i^{*\prime} \gamma_t^*\}_{\mathbb{N}_{\mathcal{T}}}) - d_i^t) x_i^{*\prime} \right] \hat{\delta}_t$ may occur, but is not necessary, and $\hat{\delta}_{\cdot,\cdot}$ may not belong to the cone of the restricted eigenvalues, but obeys the sparse eigenvalue constraints.

We first give a suitable upper bound for

$$\mathcal{M}(\gamma^*_{\cdot,\cdot} + \hat{\delta}_{\cdot,\cdot}) - \mathcal{M}(\gamma^*_{\cdot,\cdot}) - \sum_{t \in \mathbb{N}_{\mathcal{T}}} \mathbb{E}_n \left[(\hat{p}_t(\{x_i^{*\prime} \gamma_t^*\}_{\mathbb{N}_{\mathcal{T}}}) - d_i^t) x_i^{*\prime} \right] \hat{\delta}_t$$

. By the Cauchy-Schwarz inequality and the definition of the sparse eigenvalues of Eqn. (2.19),

$$\begin{aligned}
\left\|\left| \hat{\delta}_{\cdot,\cdot} \right|\right\|_{2,1} &= \sum_{j \in \hat{S}_D \cup S_D^*} \left\| \hat{\delta}_{\cdot,j} \right\|_2 \\
&\leq \sqrt{\left| \hat{S}_D \cup S_D^* \right|} \sqrt{\sum_{t \in \mathbb{N}_{\mathcal{T}}} \sum_{j \in \hat{S}_D \cup S_D^*} \hat{\delta}_{t,j}^2} \\
&= \sqrt{\left| \hat{S}_D \cup S_D^* \right|} \sqrt{\sum_{t \in \mathbb{N}_{\mathcal{T}}} \left\| \hat{\delta}_{\cdot,j} \right\|_2^2} \\
&\leq \sqrt{\left| \hat{S}_D \cup S_D^* \right|} \sqrt{\sum_{t \in \mathbb{N}_{\mathcal{T}}} \underline{\phi} \left\{ Q, \hat{S}_D \cup S_D^* \right\}^{-2} \hat{\delta}_t' Q \hat{\delta}_t} \\
&= \sqrt{\left| \hat{S}_D \cup S_D^* \right|} \underline{\phi} \left\{ Q, \hat{S}_D \cup S_D^* \right\}^{-1} \mathbb{E}_n [\| \{x_i^{*\prime} \hat{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}} \|_2^2]^{1/2}. \qquad \text{(A.14)}
\end{aligned}$$

Combining this bound with that of (A.6) yields

$$\begin{aligned}
\left| \sum_{t \in \mathbb{N}_{\mathcal{T}}} \mathbb{E}_n \left[(\hat{p}_t(\{x_i^{*\prime} \gamma_t^*\}_{\mathbb{N}_{\mathcal{T}}}) - d_i^t) x_i^{*\prime} \right] \hat{\delta}_t \right| &\leq \frac{\lambda_D}{2} \left\|\left| \hat{\delta}_{\cdot,\cdot} \right|\right\|_{2,1} \\
&\leq \frac{\lambda_D}{2} \sqrt{\left| \hat{S}_D \cup S_D^* \right|} \underline{\phi} \left\{ Q, \hat{S}_D \cup S_D^* \right\}^{-1} \mathbb{E}_n [\| \{x_i^{*\prime} \hat{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}} \|_2^2]^{1/2}. \quad \text{(A.15)}
\end{aligned}$$

Next we $\mathcal{M}(\gamma^*_{\cdot,\cdot} + \hat{\delta}_{\cdot,\cdot}) - \mathcal{M}(\gamma^*_{\cdot,\cdot})$. By optimality of the post selection estimator $\mathcal{M}(\hat{\gamma}_{\cdot,\cdot}) \leq \mathcal{M}(\tilde{\gamma}_{\cdot,\cdot})$, as $\tilde{S}^D \subset \hat{S}_D$ by construction, and hence the right side of the prior display is bounded by $\mathcal{M}(\tilde{\gamma}_{\cdot,\cdot}) - \mathcal{M}(\gamma^*_{\cdot,\cdot})$. By the mean value theorem, for scalars $\{m_t \in [0,1]\}_{\mathbb{N}_\mathcal{T}}$, $\mathcal{M}(\tilde{\gamma}_{\cdot,\cdot}) - \mathcal{M}(\gamma^*_{\cdot,\cdot})$, the bound of (A.6), the same steps in (A.13), and (A.14) with $\tilde{\delta}_{\cdot,\cdot}$:

$$\begin{aligned}
\mathcal{M}(\gamma^*_{\cdot,\cdot} + \tilde{\delta}_{\cdot,\cdot}) - \mathcal{M}(\gamma^*_{\cdot,\cdot}) &= \sum_{t\in\mathbb{N}_\mathcal{T}} \mathbb{E}_n\left[(d_i^t - \hat{p}_t(\{x_i^{*\prime}\gamma_t^* + m_t x_i^{*\prime}\tilde{\delta}_t\}))x_i^{*\prime}\tilde{\delta}_t\right] \\
&= \sum_{t\in\mathbb{N}_\mathcal{T}} \mathbb{E}_n\left[(d_i^t - \hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}))x_i^{*\prime}\tilde{\delta}_t\right] \\
&\quad + \sum_{t\in\mathbb{N}_\mathcal{T}} \mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}) - \hat{p}_t(\{x_i^{*\prime}\gamma_t^* + m_t x_i^{*\prime}\tilde{\delta}_t\}))x_i^{*\prime}\tilde{\delta}_t\right], \\
&\leq \frac{\lambda_D}{2}\left\|\left|\tilde{\delta}_{\cdot,\cdot}\right|\right\|_{2,1} + \sum_{t\in\mathbb{N}_\mathcal{T}} \mathbb{E}_n\left[m_t(x_i^{*\prime}\tilde{\delta}_t)^2\right]. \\
&\leq \frac{\lambda_D}{2}\frac{\sqrt{|\hat{S}_D \cup S_D^*|}}{\underline{\phi}\{Q, \hat{S}_D \cup S_D^*\}}\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2} + \mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2], \quad \text{(A.16)}
\end{aligned}$$

using that $m_t \leq 1$.

Collecting the bounds of (A.15) and (A.16), and the definition of $R_\mathcal{M}$ gives

$$\begin{aligned}
&\mathcal{M}(\gamma^*_{\cdot,\cdot} + \hat{\delta}_{\cdot,\cdot}) - \mathcal{M}(\gamma^*_{\cdot,\cdot}) - \sum_{t\in\mathbb{N}_\mathcal{T}} \mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}) - d_i^t)x_i^{*\prime}\right]\hat{\delta}_t \\
&\qquad\qquad \leq \frac{\lambda_D}{2}\frac{\sqrt{|\hat{S}_D \cup S_D^*|}}{\underline{\phi}\{Q, \hat{S}_D \cup S_D^*\}}\left(\mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2} + R_\mathcal{M}\right) + R_\mathcal{M}^2.
\end{aligned}$$

Next, we turn to a lower bound. Consider the same two cases as in the proof of Theorem II.12. In the first case, we have the quadratic lower bound:

$$\begin{aligned}
M(\gamma^*_{\cdot,\cdot} + \hat{\delta}_{\cdot,\cdot}) - \mathcal{M}(\gamma^*_{\cdot,\cdot}) - \sum_{t\in\mathbb{N}_\mathcal{T}} \mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}) - d_i^t)x_i^{*\prime}\right]\hat{\delta}_t \\
\geq \left(\frac{p_{\min}}{A_p}\right)^{\overline{\mathcal{T}}}\frac{1}{\mathcal{T}}\frac{\mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]}{A_K}. \quad \text{(A.17)}
\end{aligned}$$

In the other case, this bound may not hold. Arguing as in the proof of Theorem II.12, but applying Eqn. (A.14), we get

$$\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_1 \leq \sqrt{\mathcal{T}}\mathcal{X}\sqrt{|\hat{S}_D \cup S_D^*|}\underline{\phi}\{Q, \hat{S}_D \cup S_D^*\}^{-1}\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2}.$$

Therefore, as above, we find

$$\mathcal{M}(\gamma^*_{\cdot,\cdot} + \hat{\delta}_{\cdot,\cdot}) - \mathcal{M}(\gamma_{\cdot,\cdot}) - \sum_{t \in \mathbb{N}_{\mathcal{T}}} \mathbb{E}_n \left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_{\mathcal{T}}}) - d_i^t) x_i^{*\prime} \right] \hat{\delta}_t \geq r_n \mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]^{1/2}, \tag{A.18}$$

with

$$r_n = \frac{3}{2}\left(1 - \frac{2}{A_K}\right) \frac{\underline{\phi}\{Q, \hat{S}_D \cup S_D^*\}}{\mathcal{X}\sqrt{\mathcal{T}}\sqrt{|\hat{S}_D \cup S_D^*|}}.$$

Collecting the upper bounds of (A.15) and (A.16) and the lower bounds (A.17) and (A.18), and using the definition of $R_{\mathcal{M}}$, we have

$$\begin{aligned} &\left\{ \left(\frac{p_{\min}}{A_p}\right)^{\overline{\mathcal{T}}} \frac{1}{\mathcal{T}} \frac{\mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]}{A_K} \right\} \wedge \left\{ r_n \mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]^{1/2} \right\} \\ &\qquad \leq \frac{\lambda_D}{2} \frac{\sqrt{|\hat{S}_D \cup S_D^*|}}{\underline{\phi}\{Q, \hat{S}_D \cup S_D^*\}} \mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]^{1/2} + \frac{\lambda_D}{2} \frac{\sqrt{|\hat{S}_D \cup S_D^*|}}{\underline{\phi}\{Q, \hat{S}_D \cup S_D^*\}} R_{\mathcal{M}} + R_{\mathcal{M}}^2. \end{aligned}$$

For some $A_1 > 1$, replace the restriction on A_K in the Theorem with the requirement that

$$\begin{aligned} A_K > 2 &\left\{ \frac{\underline{\phi}\{Q, \hat{S}_D \cup S_D^*\}^2}{\underline{\phi}\{Q, \hat{S}_D \cup S_D^*\}^2 - (A_1/3)\mathcal{X}\sqrt{\mathcal{T}}\lambda_D |\hat{S}_D \cup S_D^*|} \right\} \\ &\qquad \vee \left\{ \frac{\underline{\phi}\{Q, \hat{S}_D \cup S_D^*\}}{\underline{\phi}\{Q, \hat{S}_D \cup S_D^*\} - (A_1/3) 2 R_{\mathcal{M}} \mathcal{X}\sqrt{\mathcal{T}}\sqrt{|\hat{S}_D \cup S_D^*|}} \right\}. \end{aligned}$$

Suppose the linear term is the minimum. Then, with the restrictions on A_K (and hence r_n), we have

$$\begin{aligned} r_n \mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]^{1/2} &\leq (r_n/A_1)\left(\mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]^{1/2} + R_{\mathcal{M}}\right) + R_{\mathcal{M}}^2 \\ &\leq (r_n/A_1)\left(\mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]^{1/2} + 2R_{\mathcal{M}}\right). \end{aligned}$$

Therefore

$$\mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]^{1/2} \leq \frac{2R_{\mathcal{M}}}{A_1 - 1}.$$

On the other hand, if the quadratic term is the minimum, define

$$R'_{\mathcal{M}} = \left(\frac{A_p}{p_{\min}}\right)^{\overline{\mathcal{T}}} \frac{\mathcal{T} A_K \lambda_D \sqrt{|\hat{S}_D \cup S_D^*|}}{2\underline{\phi}\{Q, \hat{S}_D \cup S_D^*\}},$$

and we have

$$\mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2] \leq R'_{\mathcal{M}} \mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^1/2 + R'_{\mathcal{M}} R_{\mathcal{M}} + \left(\frac{A_p}{p_{\min}}\right)^{\overline{\mathcal{T}}} \mathcal{T} A_K R_{\mathcal{M}}^2.$$

Then, because $a^2 \leq ab + c$ implies that $a \leq b + \sqrt{c}$, we have the final bound on the log-odds estimates:

$$\mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2} \leq R'_{\mathcal{M}} + \left(R'_{\mathcal{M}} R_{\mathcal{M}} + \left(\frac{A_p}{p_{\min}}\right)^{\overline{\mathcal{T}}} \mathcal{T} A_K R_{\mathcal{M}}^2\right)^{1/2}. \tag{A.19}$$

From this bound on the log-odds estimates, we obtain the bound on the propensity score estimates and the ℓ_1 rate, given by,

$$\begin{aligned} &\max_{t \in \mathbb{N}_\mathcal{T}} \mathbb{E}_n[(\hat{p}_t(\{x_i^{*\prime}\hat{\gamma}_t\}_{\mathbb{N}_\mathcal{T}}) - p_t(x_i))^2]^{1/2} \\ &\qquad \leq \left\{\frac{2R_{\mathcal{M}}}{A_1 - 1}\right\} \vee \left\{R'_{\mathcal{M}} + \left(R'_{\mathcal{M}} R_{\mathcal{M}} + \left(\frac{A_p}{p_{\min}}\right)^{\overline{\mathcal{T}}} \mathcal{T} A_K R_{\mathcal{M}}^2\right)^{1/2}\right\} + b_s^d, \end{aligned}$$

and

$$\begin{aligned} &\max_{t \in \mathbb{N}_\mathcal{T}} \|\hat{\gamma}_t - \gamma_t^*\|_1 \\ &\leq \left(\frac{|\tilde{S}^D \cup S_D^*|}{\underline{\phi}\{Q, \tilde{S}^D \cup S_D^*\}}\right)^{1/2} \left\{\frac{2R_{\mathcal{M}}}{A_1 - 1}\right\} \vee \left\{R'_{\mathcal{M}} + \left(R'_{\mathcal{M}} R_{\mathcal{M}} + \left(\frac{A_p}{p_{\min}}\right)^{\overline{\mathcal{T}}} \mathcal{T} A_K R_{\mathcal{M}}^2\right)^{1/2}\right\}, \end{aligned}$$

by arguments parallel to those used in the proof of Theorem II.12. The results as stated now follow by setting $A_1 = 3$. □

Proofs for Linear Models

SEE SUPPLEMENTAL APPENDIX.

APPENDIX B

Proofs for Chapter 3

Proofs for Chapter 3

Complete technical details may be found in the online supplement. Let C denote a generic positive constant that may take different values in different places. We use $\bigotimes$ for Kronecker products and $\prod$ for usual multiplication. Matrix inequalities are in the positive definite sense. Consecutive uses of the symbol $\asymp$ are interpreted pairwise. For a multi-index k, we define the additional notation: $k! = k_1! \cdots k_d!$, $k \leq \tilde{k} \Leftrightarrow k_1 \leq \tilde{k}_1, \ldots, k_d \leq \tilde{k}_d$, and $\sum_{[k] \leq K} = \sum_{L=0}^{K} \sum_{[k]=L}$ for $K \geq 0$.

Without loss of generality we take the basis to be centered at the midpoint of each cell and scaled by the length of the cell. Observe that centering the polynomial basis around the center of each cell avoids issues of differentiability at the boundary of each cell and the support $\mathcal{X}$. Define the one-to-one function $g(k) : \mathbb{Z}_+^d \to \mathbb{N}$ that gives the index position of $R(x)$ corresponding to entry x^k. Let $g^* = \max_k \{g(k) : k \in \mathbb{Z}_+^d,\ [k] \leq K - 1\}$. For a generic cell P_j, let p_{j*}, $\overline{p}_j$, and p_j^* be the vectors in $\mathbb{R}^d$ giving the start, mid-point, and end of the cell, respectively, and let $\overline{p}_{\ell,j} = (p_{\ell,j} + p_{\ell,j-1})/2 \in \mathbb{R}$ be the midpoint of each interval. Define the matrix functions $D(a)$ to be the $K \times K$ diagonal matrix with entries given by $a^{-(v-1)}$, $v = 1, \ldots, K$ and $L(b)$ to be the $K \times K$ lower triangular matrix with typical element $\binom{u-1}{v-1}(-b)^{u-v}$, $(u, v) \in \{1, \ldots, K : u \geq v\}$. We then take the (rotated) polynomial basis to be given by $\tilde{R}_j(x) \equiv \mathbb{1}_{P_j}(x)\tilde{R}(x) = \mathbb{1}_{P_j}(x) S_K \bigotimes_{\ell=1}^{d} \left\{ D\left(p_{\ell,j} - \overline{p}_{\ell,j}\right) L(\overline{p}_{\ell,j}) r(x_\ell) \right\}$, where S_K is a $g^* \times K^d$ selection matrix which removes terms of degree exceeding $K - 1$. Finally let $\tilde{R}_j = (\tilde{R}_j(X_1), \ldots, \tilde{R}_j(X_n))'$ and (globally) redefine $\Omega_j = \mathbb{E}[\tilde{R}_j(X)\tilde{R}_j(X)']/q_j$ and $\hat{\Omega}_j = \tilde{R}_j'\tilde{R}_j/(nq_j)$, maintaining the same notation for the latter two for simplicity.

Preliminary Lemmas

Several intermediate lemmas are required before proving the main results. These lemmas establish properties of partitioning estimators which may be of independent interest for other applications.

Lemma B.1. *Under Assumption III.1(b), the basis satisfies:*

$$\max_{1\leq j\leq J_n^d}\max_{[m]\leq s}\|\partial^m\tilde{R}_j(\cdot)\|_\infty = O(J_n^s)$$

, for $s\leq K-1$.

Proof. By construction of the partition, for $x\in P_j$, $|x-\overline{p}_j|\leq|p_j^*-\overline{p}_j|\asymp J_n^{-1}$. For fixed $x\in\mathcal{X}$ and a multi-index m such that $[m]\leq K-1$:

$$\begin{aligned}&\left|\partial^m\tilde{R}_j(x)\right|^2\\&\qquad=\frac{1}{(p_j^*-\overline{p}_j)^{2m}}\mathbb{1}_{P_j}(x)\sum_{[k]\leq K-1}\mathbb{1}\{m\leq k\}\left\{\frac{k!}{(k-m)!}\frac{(x-\overline{p}_j)^{k-m}}{(p_j^*-\overline{p}_j)^{k-m}}\right\}^2\\&\qquad=O\left(J_n^{2[m]}\right),\end{aligned}$$

uniformly in $1\leq j\leq J_n^d$, $x\in P_j$, and $\{m:[m]\leq K-1\}$, and in particular for those satisfying $[m]\leq s\leq K-1$, for any such s. □

Lemma B.2. *Define $\mu_j(x)\equiv\mathbb{1}_{P_j}(x)\mu(x)$, and following the definition in Eqn.* (3.2), *$\partial^m\mu_j(x)\equiv\mathbb{1}_{P_j}(x)\partial^m\mu(x)$. Under Assumptions III.1(b) and III.1(e), there is a non-random vector β_j^0, depending only on K and j, such that for $s\leq S\wedge(K-1)$:* $\max_{1\leq j\leq J_n^d}\max_{[m]\leq s}\|\partial^m\mu_j(\cdot)-\partial^m\tilde{R}_j(\cdot)'\beta_j^0\|_\infty=O(J_n^{-((S+\alpha)\wedge K-s)})$.

Proof. Assumption III.1(e) implies that $\partial^m\mu_j(x)$ satisfies the Taylor expansion for $x\in P_j$ given by:

$$\partial^m\mu_j(x)=\sum_{[k]\leq S\wedge(K-1)-[m]}\frac{1}{k!}\left(\partial^{k+m}\mu_j(\overline{p}_j)\right)\left(x-\overline{p}_j\right)^k+O\left(\left|x-\overline{p}_j\right|^{(S+\alpha)\wedge K-[m]}\right),\tag{B.1}$$

with constants which can be made uniform in the multi-index m, s, and j. For $k\in\mathbb{Z}_+^d$ define the function $\beta_j^0(k)=\frac{1}{k!}\left(\partial^k\mu_j(\overline{p}_j)\right)(p_j^*-\overline{p}_j)^k$ and the coefficient vector β_j^0 as the $g^* \times 1$ vector with entry e equal to $\beta_j^0(g^{-1}(e))$. Therefore:

$$\begin{aligned}
&\partial^m \tilde{R}_j(x)'\beta_j^0 \\
&\qquad = \sum_{[k] \leq S \wedge (K-1)} \mathbb{1}\{m \leq k\} \frac{\left(x - \overline{p}_j\right)^{k-m}}{(k-m)!} \partial^k \mu_j(\overline{p}_j) \\
&\qquad = \sum_{[\tilde{k}+m] \leq S \wedge (K-1)} \frac{\left(x - \overline{p}_j\right)^{\tilde{k}}}{\tilde{k}!} \partial^{\tilde{k}+m} \mu_j(\overline{p}_j).
\end{aligned}$$

This matches the Taylor series, hence subtracting from Eqn. (B.1) completes the proof. □

Lemma B.3. *Under Assumption III.1, $\Omega_j \asymp I_{g^*}$, the identity matrix, uniformly in j.*

Proof. By Assumption III.1(d) and the construction of the partition, $q_j \asymp J_n^{-d}$. Applying this result and Assumption III.1(d) again, we have: $\Omega_j \asymp J_n^d \int_{\mathcal{X}} \tilde{R}_j(x)\tilde{R}_j(x)'dx$. Now, by Assumption III.1(b), properties of the Kronecker product, and the construction of the transformed basis,

$$\Omega_j \asymp J_n^d S_K \bigotimes_{\ell=1}^{d} \left\{ \int_{p_{\ell,j-1}}^{p_{\ell,j}} r\left(\frac{x_\ell - \overline{p}_{\ell,j}}{p_{\ell,j} - \overline{p}_{\ell,j}}\right) r\left(\frac{x_\ell - \overline{p}_{\ell,j}}{p_{\ell,j} - \overline{p}_{\ell,j}}\right)' dx_\ell \right\} S_K'.$$

Let H denote the Hilbert matrix of order K, which is positive definite. Changing variables $z = (x_\ell - \overline{p}_{\ell,j})/(p_{\ell,j} - \overline{p}_{\ell,j})$, applying $|p_{\ell,j} - p_{\ell,j-1}| \asymp J_n^{-1}$, changing variables $t = (z+1)/2$, gives

$$\begin{aligned}
\Omega_j &\asymp S_K \left\{ \bigotimes_{\ell=1}^{d} \int_{-1}^{1} r(z)r(z)'dz \right\} S_K' \\
&\qquad \asymp S_K \left\{ \bigotimes_{\ell=1}^{d} \left[D(2)L(-1)\right]^{-1} H \left[L(-1)D(2)\right]^{-1} \right\} S_K' \asymp I_{g^*}. \quad \square
\end{aligned}$$

Lemma B.4. *Let $a_n = n^{-1}J_n^d \log(J_n^d)$. Under the conditions of Theorem III.2: $\max_{1 \leq j \leq J_n^d} |\hat{\Omega}_j - \Omega_j|^2 = O_p(a_n)$. If, in addition, $J_n^d \asymp (n/\log(n))^\gamma$, $\gamma \in (0,1)$, the same is true almost surely.*

Proof. For $k, \tilde{k} \in k \in \mathbb{Z}_+^d : [k] \leq K-1$, let the $(g(k), g(\tilde{k}))$ element of $(\hat{\Omega}_j - \Omega_j)$ be denoted $\sum_{i=1}^{n} W_{ij}(k,\tilde{k})/(nq_j)$, where

$$W_{ij}(k,\tilde{k}) = [\tilde{R}_j(X_i)\tilde{R}_j(X_i)']_{g(k),g(\tilde{k})} - [\mathbb{E}[\tilde{R}_j(X_i)\tilde{R}_j(X_i)']]_{g(k),g(\tilde{k})}$$

. By Lemma B.1 and the triangle inequality, $|W_{ij}(k,\tilde{k})| \leq C$ and $\mathbb{E}[W_{ij}(k,\tilde{k})^2] \leq Cq_j$. Thus by Boole's inequality, K being fixed, Bernstein's inequality, and $q_j \asymp J_n^{-d}$:

$$\begin{aligned}
&\mathbb{P}\left[\max_{1\leq j\leq J_n^d}\left|\hat{\Omega}_j - \Omega_j\right| > (a_n)^{1/2}\varepsilon\right] \\
&\leq CJ_n^d \max_{1\leq j\leq J_n^d}\max_{[k],[\tilde{k}]\leq K-1}\mathbb{P}\left[\left|\sum_{i=1}^{n} W_{ij}(k,\tilde{k})\right| > q_j\sqrt{nJ_n^d\log(J_n^d)}\varepsilon\right] \\
&\leq CJ_n^d \max_{1\leq j\leq J_n^d}\max_{[k],[\tilde{k}]\leq K-1}\exp\left\{-C\frac{q_j^2 nJ_n^d\log(J_n^d)\varepsilon^2}{nq_j + q_j\sqrt{nJ_n^d\log(J_n^d)}\varepsilon}\right\},
\end{aligned}$$

which is arbitrarily small for ε large enough by the rate restriction of Theorem III.2. When $J_n^d \asymp (n/\log(n))^\gamma$, the conclusion holds with probability one by the Borel-Cantelli Lemma. □

Lemma B.5. *Let the conditions of Theorem III.3 hold, and for ξ therein let $r_n^2 = n^{-1}J_n^{d(2-\xi)}\log(J_n^d)^\xi$. Then for $G = (\mu(X_1),\ldots,\mu(X_n))'$, we have $\max_{1\leq j\leq J_n^d}|\tilde{R}_j'(Y-G)/(nq_j)| = O_p(r_n)$. If, in addition, $J_n^d \asymp (n/\log(n))^\gamma$, $\gamma\in(0,1)$, and $\eta > 2(1+\xi\gamma)/(1-\xi\gamma)$, the same is true almost surely.*

Proof. With the convention $0/0 = 0$, define $t_n = J_n^{d\xi/\eta}\log(J_n^d)^{-\xi/\eta}$. Following the same notation as in Lemma B.4, let $H_{ij}(k) = \mathbb{1}_{P_j}(X_i)[\tilde{R}_j(X_i)]_{g(k)}(Y_i\mathbb{1}\{Y_i \leq t_n\} - \mathbb{E}[Y_i\mathbb{1}\{Y_i \leq t_n\}|X_i])$ and $T_{ij}(k) = \mathbb{1}_{P_j}(X_i)[\tilde{R}_j(X_i)]_{g(k)}(Y_i\mathbb{1}\{Y_i > t_n\} - \mathbb{E}[Y_i\mathbb{1}\{Y_i > t_n\}|X_i])$. For the truncated term, since $|H_{ij}(k)| \leq t_n$ and $\mathbb{E}[H_{ij}(k)^2] \leq Cq_j$, Bernstein's inequality and $q_j \asymp J_n^{-d}$ give, for fixed $k\in\mathbb{Z}_+^d$:

$$J_n^d\max_{1\leq j\leq J_n^d}\mathbb{P}\left[\left|\sum_{i=1}^{n} H_{ij}(k)\right| > nq_jr_n\varepsilon\right] \leq C\exp\left\{\log(J_n^d)\left[1 - C\frac{nr_n^2(J_n^d\log(J_n^d))^{-1}\varepsilon^2}{1+t_nr_n\varepsilon}\right]\right\}.$$

For the tails, by Markov's inequality, $\mathbb{E}\left[T_{ij}(k)\right] = 0$, Lemma B.1, Assumption III.1(c), and $q_j \asymp J_n^{-d}$:

$$\begin{aligned}
&J_n^d\max_{1\leq j\leq J_n^d}\mathbb{P}\left[\left|\sum_{i=1}^{n} T_{ij}(k)\right| > nq_jr_n\varepsilon\right] \\
&\qquad\qquad \leq C\frac{J_n^d}{nr_n^2t_n^\eta\varepsilon^2}\max_{1\leq j\leq J_n^d}\frac{1}{q_j^2}\mathbb{E}\left[\mathbb{1}_{P_j}(X_i)\mathbb{E}\left[|Y_i|^{2+\eta}\middle| X_i\right]\right]
\end{aligned}$$

$$\leq C\frac{J_n^{2d}}{nr_n^2 t_n^\eta \varepsilon^2}.$$

These two bounds do not depend on k, and hence by Boole's inequality and K constant,

$$\begin{aligned}\mathbb{P}\left[\max_{1\leq j\leq J_n^d}\left|\tilde{R}_j'(Y-G)/(nq_j)\right| > r_n\varepsilon\right] &\leq CJ_n^d \max_{1\leq j\leq J_n^d}\max_{[k]\leq K-1}\mathbb{P}\left[\left|\sum_{i=1}^n H_{ij}(k)\right| > nq_jr_n\varepsilon\right]\\ &+ CJ_n^d \max_{1\leq j\leq J_n^d}\max_{[k]\leq K-1}\mathbb{P}\left[\left|\sum_{i=1}^n T_{ij}(k)\right| > nq_jr_n\varepsilon\right],\end{aligned}$$

which is arbitrarily small for ε large enough by $\xi \in [0,1]$, the rate restriction of the Theorem, and the definition of t_n. The conclusion holds with probability one by the Borel-Cantelli Lemma if $J_n^d \asymp (n/\log(n))^\gamma$ and $t_n = n^\tau$ for $(1+\xi\gamma)/\eta < \tau < (1-\xi\gamma)/2$. □

Convergence Rates

Proof of Theorem III.2. Define $\mathbb{1}_{n,j} = \mathbb{1}\{\lambda_{\min}(\hat{\Omega}_j) \geq C\}$ for some positive constant C, where $\lambda_{\min}(\hat{\Omega}_j)$ is the smallest eigenvalue, and take $\hat{\mu}(x) = \sum_{j=1}^{J_n^d}\mathbb{1}_{n,j}\tilde{R}_j(x)'\hat{\beta}_j$ (cf. Eqn. (3.1)). As $\min_{1\leq j\leq J_n^d}\mathbb{1}_{n,j} = 1$ w.p.a. 1, this distinction vanishes asymptotically. First:

$$\begin{aligned}\max_{[m]\leq s}\left\|\partial^m\hat{\mu} - \partial^m\sum_{j=1}^{J_n^d}\mathbb{1}_{n,j}\mu_j\right\|_2^2 &\leq \max_{[m]\leq s} 3\sum_{j=1}^{J_n^d}\left\|\mathbb{1}_{n,j}(\partial^m\tilde{R}_j(\cdot))'\hat{\Omega}_j^{-1}\tilde{R}_j'(Y-G)/(nq_j)\right\|_2^2 \qquad (T_{n1})\\ &+ \max_{[m]\leq s} 3\sum_{j=1}^{J_n^d}\left\|\mathbb{1}_{n,j}(\partial^m\tilde{R}_j(\cdot))'\hat{\Omega}_j^{-1}\tilde{R}_j'(G-\tilde{R}_j\beta_j^0)/(nq_j)\right\|_2^2 \qquad (T_{n2})\\ &+ \max_{[m]\leq s} 3\sum_{j=1}^{J_n^d}\left\|\mathbb{1}_{n,j}\left[(\partial^m\tilde{R}_j(\cdot))'\beta_j^0 - \partial^m\mu_j(\cdot)\right]\right\|_2^2. \qquad (T_{n3})\end{aligned}$$

By properties of the trace, Assumption III.1(c), $\tilde{R}_j(\tilde{R}_j'\tilde{R}_j)^{-1}\tilde{R}_j'$ idempotent, K fixed, and $q_j \asymp J_n^{-d}$,

$$\mathbb{E}\left[|\mathbb{1}_{n,j}\hat{\Omega}_j^{-1/2}\tilde{R}_j'(Y-G)/(nq_j)|^2\middle|\{X_i\}\right]$$

$$
\begin{aligned}
&= \frac{\mathbb{1}_{n,j}}{nq_j} \operatorname{tr}\left\{\tilde{R}_j(\tilde{R}_j'\tilde{R}_j)^{-1}\tilde{R}_j'\mathbb{E}\left[(Y-G)(Y-G)'|\,\{X_i\}\right]\right\} \\
&\leq C\frac{\mathbb{1}_{n,j}}{nq_j} \operatorname{tr}\left\{\tilde{R}_j\left(\tilde{R}_j'\tilde{R}_j\right)^{-1}\tilde{R}_j'\right\} \leq \frac{C}{nq_j} \leq \frac{CJ_n^d}{n}.
\end{aligned}
$$

Hence, $T_{n1} \leq O_p(J_n^{2s})\sum_{j=1}^{J_n^d}\mathbb{1}_{n,j}\left|\hat{\Omega}_j^{-1/2}\tilde{R}_j'(Y-G)/nq_j\right|^2\int_{P_j}f(x)dx = O_p\left(J_n^{d+2s}/n\right)$, by Markov's inequality and Lemmas B.1 and B.4.

By Boole's and Bernstein's inequality and the condition of Theorem III.2:

$$
\begin{aligned}
&\mathbb{P}\left[\max_{1\leq j\leq J_n^d}\sum_{i=1}^{n}(\mathbb{1}_{P_j}(X_i)-q_j) > nq_j\varepsilon\right] \\
&\qquad\qquad \leq C\exp\left\{\log(J_n^d)\left[1-C\frac{n}{J_n^d\log(J_n^d)}\frac{\varepsilon^2}{1+\varepsilon}\right]\right\} \to 0. \quad \text{(B.2)}
\end{aligned}
$$

Therefore, by $\tilde{R}_j(\tilde{R}_j'\tilde{R}_j)^{-1}\tilde{R}_j'$ idempotent and Lemma B.2:

$$
\begin{aligned}
&\max_{1\leq j\leq J_n^d}\left|\mathbb{1}_{n,j}\hat{\Omega}_j^{-1/2}\tilde{R}_j'(G-\tilde{R}_j\beta_j^0)/(nq_j)\right|^2 \leq \max_{1\leq j\leq J_n^d}\left|(G-\tilde{R}_j\beta_j^0)'(G-\tilde{R}_j\beta_j^0)/(nq_j)\right| \\
&\quad \leq \max_{1\leq j\leq J_n^d}\left\|\mathbb{1}_{P_j}(\cdot)(\mu(\cdot)-\tilde{R}_j(\cdot)'\beta_j^0)\right\|_\infty^2 \max_{1\leq j\leq J_n^d}\frac{1}{nq_j}\sum_{i=1}^{n}\mathbb{1}_{P_j}(X_i) = O_p\left(J_n^{-2((S+\alpha)\wedge K)}\right). \\
&\hfill \text{(B.3)}
\end{aligned}
$$

Applying Lemmas B.1 and B.4, and $\sum_{j=1}^{J_n^d}\int_{P_j}f(x)dx = 1$, we have

$$
T_{n2} = O_p(J_n^{-2((S+\alpha)\wedge K-s)}).
$$

Finally, Lemma B.2 immediately gives:

$$
T_{n3} = O(J_n^{-2((S+\alpha)\wedge K-s)})
$$

. □

Proof of Theorem III.3. First:

$$
\begin{aligned}
&\max_{[m]\leq s}\left\|\partial^m\hat{\mu}-\partial^m\sum_{j=1}^{J_n^d}\mathbb{1}_{n,j}\mu_j\right\|_\infty^2 \\
&\qquad \leq \max_{1\leq j\leq J_n^d}\max_{[m]\leq s}3\left\|\mathbb{1}_{n,j}(\partial^m\tilde{R}_j(\cdot))'\hat{\Omega}_j^{-1}\tilde{R}_j'(Y-G)/(nq_j)\right\|_\infty^2
\end{aligned}
$$

$$
\begin{aligned}
&+ \max_{1\le j\le J_n^d}\max_{[m]\le s} 3\left\|\mathbb{1}_{n,j}(\partial^m \tilde{R}_j(\cdot))'\hat{\Omega}_j^{-1}\tilde{R}_j'(G-\tilde{R}_j\beta_j^0)/(nq_j)\right\|_\infty^2 \\
&+ \max_{1\le j\le J_n^d}\max_{[m]\le s} 3\left\|\mathbb{1}_{n,j}(\partial^m \tilde{R}_j(\cdot))'\beta_j^0-\partial^m\mu_j(\cdot)\right\|_\infty^2 \\
&= O\left(J_n^{2s}\right) O_p\left(\frac{J_n^{d(2-\xi)}\log(J_n^d)^\xi}{n}\right) + O_p\left(J_n^{-2((S+1)\wedge K-s)}\right),
\end{aligned}
$$

where we apply Lemmas B.1, B.4, and B.5 for the first term; Lemmas B.1 and B.4 and Eqn. (B.3) for the second; and Lemma B.2 for the third. The result follows as $\min_{1\le j\le J_n^d}\mathbb{1}_{n,j}=1$ w.p.a. 1. □

We now demonstrate a version of Theorem III.3 that holds with probability one.

Theorem B.6. *Suppose the conditions of Theorem III.2 hold. If, in addition, for some $\xi\in[0,1\wedge\eta]$ the partition satisfies $J_n^d \asymp (n/\log(n))^\gamma$, $\gamma\in(0,1)$ and $\eta>2(1+\xi\gamma)/(1-\xi\gamma)$, then for $s\le S\wedge(K-1)$:*

$$
\max_{[m]\le s}\|\partial^m\hat{\mu}-\partial^m\mu\|_\infty^2 = O_{as}\left(\frac{J_n^{(2-\xi)d+2s}\log(J_n^d)^\xi}{n}+J_n^{-2((S+\alpha)\wedge K-s)}\right).
$$

Proof of Theorem B.6. The rate restriction on J_n implies that of Theorem III.3, whose proof may thus be strengthened to hold with probability one using Eqn. (B.2). □

Asymptotic Mean-Square Error

We first give three lemmas necessary for results. The first two are straightforward, and Lemma B.9 follows identically to Lemma B.4. Proofs may be found in the supplemental appendix.

Lemma B.7. *Let the conditions of Theorem III.2 hold and $g(\cdot)$ be continuous on $\mathcal{X}$. Then for $h_j(x)=\mathbb{1}_{P_j}(x)h(x)$, with remainder uniform in $1\le j\le J_n^d$: $\int_{P_j}h(z)g(z)dz = g(\overline{p}_j)\int_{P_j}h(z)dz+\max_{1\le j\le J_n^d}\|h_j(\cdot)\|_\infty(o(J_n^{-d}))$.*

Lemma B.8. *Let the conditions of Theorem III.2 hold. If $g(\cdot)$ is continuous on $\mathcal{X}$, then: $\sum_{j=1}^{J_n^d} g(\overline{p}_j)\operatorname{vol}(P_j)=\int_{\mathcal{X}}g(z)dz+o(1)$.*

Lemma B.9. *Under the conditions of Theorem III.5, for Γ_j defined Eqn.* (3.4) *and* *any $k \in \mathbb{Z}_+^d$:*

$$(a) \quad \max_{1\le j\le J_n^d} \left| \frac{1}{nq_j} \sum_{i=1}^n \tilde{R}_j(X_i)\tilde{R}_j(X_i)'\sigma^2(X_i) - \Gamma_j \right|^2 = O_p\left(\frac{J_n^d \log(J_n^d)}{n}\right);$$

$$(b) \quad \max_{1\le j\le J_n^d} \left| \frac{1}{nq_j} \sum_{i=1}^n \tilde{R}_j(X_i)\frac{(X_i-\overline{p}_j)^k}{(p_j^*-\overline{p}_j)^k} - \frac{1}{q_j}\mathbb{E}\left[\tilde{R}_j(X)\frac{(X_i-\overline{p}_j)^k}{(p_j^*-\overline{p}_j)^k}\right] \right|^2 = O_p\left(\frac{J_n^d \log(J_n^d)}{n}\right).$$

Proof of Theorem III.5. We first give some notation and facts used repeatedly throughout. With a slight abuse notation, let $|\mathcal{X}|^k = \prod_{\ell=1}^d |\mathcal{X}_\ell|^{k_\ell}$. Let $\mathcal{U} = \times_{\ell=1}^d[-1,1]$. We frequently use the change of variables $z_\ell = (x_\ell - \overline{p}_{\ell,j})/(p_{\ell,j}-\overline{p}_{\ell,j})$, $\ell = 1,\dots,d$, the Jacobian of which is $\prod_{\ell=1}^d(p_{\ell,j}-\overline{p}_{\ell,j}) = 2^{-d}\operatorname{vol}(P_j) = (2J_n)^{-d}\operatorname{vol}(\mathcal{X})$. For any $k\in\mathbb{Z}_+^d$: $(p_j^*-\overline{p}_j)^k = (2J_n)^{-[k]}|\mathcal{X}|^k$.

Using Lemmas B.1 and B.7 and the change of variables above, we get the following results, which also hold for $w(x) = f(x)$ or $m = 0$:

$$\text{(a)} \int\limits_{\mathcal{X}} (\partial^m \tilde{R}_j(x))(x-\overline{p}_j)^{k-m}w(x)dx$$
$$= 2^{-d}w(\overline{p}_j)(p_j^*-\overline{p}_j)^{k-2m}\operatorname{vol}(P_j)\int\limits_{\mathcal{U}}(\partial^m R(z))z^k dz + o(J_n^{-d-K});$$

$$\text{(b) } \Omega_j = \frac{2^{-d}}{q_j}f(\overline{p}_j)\operatorname{vol}(P_j)\int\limits_{\mathcal{U}} R(z)R(z)'dz + o(J_n^{-d});$$

$$\text{(c)} \int\limits_{X}(\partial^m\tilde{R}_j(x))(\partial^m\tilde{R}_j(x))'w(x)dx$$
$$= \frac{2^{-d}w(\overline{p}_j)\operatorname{vol}(P_j)}{(p_j^*-\overline{p}_j)^{2m}}\int\limits_{\mathcal{U}}(\partial^m R(z))\,(\partial^m R(z))'\,dz + o(J_n^{-d-2[m]}).$$

First consider the conditional variance term: $\int_{\mathcal{X}}\mathbb{V}[\partial^m\hat{\mu}(x) \mid \mathcal{X}data]w(x)dx$. By Lemma B.7, $\Gamma_j = \sigma^2(\overline{p}_j)\Omega_j + o(J_n^{-d})$. Applying this result and Lemmas B.1, B.4, and B.9(a), we have:

$$\mathbb{V}\left[\sum_{j=1}^{J_n^d}(\partial^m\tilde{R}_j(x))'\mathbb{1}_{n,j}\hat{\Omega}_j^{-1}\tilde{R}_jY/(nq_j)|\mathcal{X}data\right]$$
$$= \sum_{j=1}^{J_n^d}\frac{1}{nq_j}(\partial^m\tilde{R}_j(x))'\Omega_j^{-1}\Gamma_j\Omega_j^{-1}(\partial^m\tilde{R}_j(x)) + o_p\left(\frac{J_n^{d+2[m]}}{n}\right)$$

$$= \sum_{j=1}^{J_n^d} \frac{1}{nq_j} \sigma^2(\overline{p}_j) \operatorname{tr} \left\{ \Omega_j^{-1} (\partial^m \tilde{R}_j(x))(\partial^m \tilde{R}_j(x))' \right\} + o_p \left(\frac{J_n^{d+2[m]}}{n} \right).$$

Integrating the above expression, applying Lemma B.7, the above facts and change of variables, and Lemma B.8 (under Assumption III.4(a)), we have:

$$\begin{aligned}
&\sum_{j=1}^{J_n^d} \frac{1}{nq_j} \sigma^2(\overline{p}_j) \operatorname{tr} \left\{ \Omega_j^{-1} \int\limits_X \left(\partial^m \tilde{R}_j(x) \right) \left(\partial^m \tilde{R}_j(x) \right)' w(x) dx \right\} + o_p \left(\frac{J_n^{d+2[m]}}{n} \right) \\
&\quad = \frac{J_n^{d+2[m]}}{n} \frac{2^{2[m]}}{|\mathcal{X}|^{2m} \operatorname{vol}(\mathcal{X})} \left(\int\limits_{\mathcal{X}} \frac{\sigma^2(x)}{f(x)} w(x) dx \right) \\
&\qquad\qquad \times \operatorname{tr} \left\{ \left(\int\limits_{\mathcal{U}} R(z) R(z)' dz \right)^{-1} \int\limits_{\mathcal{U}} \left(\partial^m R(z) \right) \left(\partial^m R(z) \right)' dz \right\} \\
&\qquad\qquad \times [1 + o(1)] + o_p \left(J_n^{d+2[m]}/n \right).
\end{aligned}$$

Next consider the bias portion of the expansion: $\int_{\mathcal{X}} (\mathbb{E}[\hat{\mu}(x) | \mathcal{X} data] - \mu(x))^2 w(x) dx$. Define $T_{K,j,m}(x) = \sum_{k:[k]=K} \left(\partial^k \mu_j(\overline{p}_j) \right) (x - \overline{p}_j)^{k-m}/(k-m)!$, so that under Assumption III.4(b), $\partial^m \mu_j(x) = T_{K,j,m}(x) + o(J_n^{-(K-[m])})$ uniformly in $1 \leq j \leq J_n^d$. Then by Lemmas B.4 and B.9,

$$\begin{aligned}
&\sum_{j=1}^{J_n^d} \partial^m \tilde{R}_j(x)' \mathbb{1}_{n,j} (\tilde{R}_j' \tilde{R}_j)^{-1} \sum_{i=1}^{n} \tilde{R}_j(X_i) \mu(X_i) - \sum_{j=1}^{J_n^d} \partial^m \mu_j(x)) \\
&= \sum_{j=1}^{J_n^d} \left(\partial^m \tilde{R}_j(x)' \mathbb{1}_{n,j} (\tilde{R}_j' \tilde{R}_j)^{-1} \left(\sum_{i=1}^{n} \tilde{R}_j(X_i) \tilde{R}_j(x_i)' \right) \beta_j^0 - \partial^m \mu_j(x) \right) \\
&\qquad + \sum_{j=1}^{J_n^d} \partial^m \tilde{R}_j(x)' \mathbb{1}_{n,j} (\tilde{R}_j' \tilde{R}_j)^{-1} \sum_{i=1}^{n} \tilde{R}_j(X_i) \left(T_{K,j,0}(X_i) + o(J_n^{-K}) \right) \\
&= -\sum_{j=1}^{J_n^d} \mathbb{1}_{n,j} \mathbb{1}_{P_j}(x) T_{k,j,m}(x) + \sum_{j=1}^{J_n^d} \partial^m \tilde{R}_j(x)' \mathbb{1}_{n,j} (\tilde{R}_j' \tilde{R}_j)^{-1} \sum_{i=1}^{n} \tilde{R}_j(X_i) T_{K,j,0}(X_i) \\
&\qquad\quad + o_p \left(J_n^{-(K-[m])} \right) \\
&= \sum_{[k]=K} \sum_{j=1}^{J_n^d} \mathbb{1}_{P_j}(x) \left(\partial^k \mu_j(\overline{p}_j) \right) \left(\frac{\partial^m \tilde{R}_j(x)'}{k! q_j} \Omega_j^{-1} \mathbb{E} \left[\tilde{R}_j(X)(X - \overline{p}_j)^k \right] - \frac{(x - \overline{p}_j)^{k-m}}{(k-m)!} \right) \\
&\qquad\quad + o_p \left(J_n^{-(K-[m])} \right).
\end{aligned}$$

Then since $\min_{1\leq j\leq J_n^d} \mathbb{1}_{n,j} = 1$ w.p.a. 1 by Lemma B.4, the integrated, squared bias becomes:

$$\begin{aligned}
&\sum_{j=1}^{J_n^d} \sum_{\substack{k,\ \tilde{k} \\ [k]=[\tilde{k}]=K}} \left(\partial^k \mu_j(\overline{p}_j)\right)\left(\partial^{\tilde{k}} \mu_j(\overline{p}_j)\right)\Bigg\{ \frac{1}{(k-m)!(\tilde{k}-m)!}\int\limits_{P_j} (x-\overline{p}_j)^{k+\tilde{k}-2m} w(x)dx \\
&\quad + \frac{1}{k!\tilde{k}!}\frac{1}{q_j^2}\int\limits_{P_j} \partial^m \tilde{R}_j(x)'\Omega_j^{-1}\mathbb{E}\left[\tilde{R}_j(X)(X-\overline{p}_j)^k\right] \\
&\quad \times \mathbb{E}\left[(X-\overline{p}_j)^{\tilde{k}}\tilde{R}_j(X)'\right]\Omega_j^{-1}\partial^m \tilde{R}_j(x)w(x)dx \\
&\quad - \frac{1}{k!(\tilde{k}-m)!}\frac{1}{q_j}\int\limits_{P_j} (x-\overline{p}_j)^{\tilde{k}-m}\partial^m \tilde{R}_j(x)'w(x)dx\Omega_j^{-1}\mathbb{E}\left[\tilde{R}_j(X)(X-\overline{p}_j)^k\right] \\
&\quad - \frac{1}{\tilde{k}!(k-m)!}\frac{1}{q_j}\int\limits_{P_j} (x-\overline{p}_j)^{k-m}\partial^m \tilde{R}_j(x)'w(x)dx\Omega_j^{-1}\mathbb{E}\left[\tilde{R}_j(X)(X-\overline{p}_j)^{\tilde{k}}\right]\Bigg\} \\
&\quad + o_p\left(J_n^{-2(K-[m])}\right) \\
&= \sum_{j=1}^{J_n^d} \sum_{\substack{k,\ \tilde{k} \\ [k]=[\tilde{k}]=K}} \left(\partial^k \mu_j(\overline{p}_j)\right)\left(\partial^{\tilde{k}} \mu_j(\overline{p}_j)\right)\{B_1 + B_2 - B_3 - B_4\} + o_p\left(J_n^{-2(K-[m])}\right),
\end{aligned}$$

where the final equality defines the terms B_1–B_4. Applying Lemma B.7 and the change of variables above, and discarding a remainder of order $o(J_n^{-d})O(J_n^{-2(K-[m])})$, these terms are:

$$\begin{aligned}
B_1 &= \frac{w(\overline{p}_j)}{(k-m)!(\tilde{k}-m)!}\int\limits_{P_j} (x-\overline{p}_j)^{k+\tilde{k}-2m}dx \\
&= \frac{(p_j^* - \overline{p}_j)^{k+\tilde{k}-2m} w(\overline{p}_j)\operatorname{vol}(P_j)}{2^d(k-m)!(\tilde{k}-m)!}\int\limits_{\mathcal{U}} z^{k+\tilde{k}-2m}dz; \\
B_2 &= \frac{1}{k!\tilde{k}!}\frac{1}{q_j^2}\int\limits_{P_j} \operatorname{tr}\Big\{(\partial^m \tilde{R}_j(x))'\Omega_j^{-1}\mathbb{E}\left[\tilde{R}_j(X)(X-\overline{p}_j)^k\right] \\
&\qquad \times \mathbb{E}\left[(X-\overline{p}_j)^{\tilde{k}}\tilde{R}_j(X)'\right]\Omega_j^{-1}(\partial^m \tilde{R}_j(x))\Big\}w(x)dx \\
&= \frac{(p_j^* - \overline{p}_j)^{k+\tilde{k}-2m} w(\overline{p}_j)\operatorname{vol}(P_j)}{2^d k!\tilde{k}!}\operatorname{tr}\Bigg\{\left(\int\limits_{\mathcal{U}} R(x)R(x)'dz\right)^{-1}\int\limits_{\mathcal{U}} R(z)z^k dz
\end{aligned}$$

$$\times \int\limits_{\mathcal{U}} R(z)'z^{\tilde{k}}dz \left(\int\limits_{\mathcal{U}} R(z)R(z)'dz\right)^{-1} \int\limits_{\mathcal{U}} \left(\partial^m R(z)\right)\left(\partial^m R(z)\right)' dz\Bigg\};$$

$$B_3 = \frac{(p_j^* - \overline{p}_j)^{k+\tilde{k}-2m} w(\overline{p}_j)\operatorname{vol}(P_j)}{2^d k!(\tilde{k}-m)!} \int\limits_{\mathcal{U}} (\partial^m R(z))' z^{\tilde{k}-m} dz$$
$$\times \left(\int\limits_{\mathcal{U}} R(z)R(z)'dz\right)^{-1} \int\limits_{\mathcal{U}} R(z)z^k dz;$$

and finally B_4 is identical to B_3 except with k and $\tilde{k}$ reversed.

All four terms have the common factor $(p_j^* - \overline{p}_j)^{k+\tilde{k}-2m} w(\overline{p}_j)\operatorname{vol}(P_j)$, which contains all dependence on the partition. By Lemma B.8, the facts at the outset, and that $[k] = [\tilde{k}] = K$: $\sum_{j=1}^{J_n^d} (\partial^k \mu_j(\overline{p}_j))(\partial^{\tilde{k}} \mu_j(\overline{p}_j))(p_j^* - \overline{p}_j)^{k+\tilde{k}-2m} w(\overline{p}_j)\operatorname{vol}(P_j) = (2J_n)^{-2(K-[m])}|\mathcal{X}|^{k+\tilde{k}-2m}$ $\times$ $\int_{\mathcal{X}} (\partial^k \mu_j(x))(\partial^{\tilde{k}} \mu_j(x)) w(x)dx[1+o(1)]$.

Define:

$$\mathscr{V}_{K,d,m} = \frac{2^{2[m]}}{\operatorname{vol}(\mathcal{X})}\left(\prod_{\ell=1}^{d} |\mathcal{X}_\ell|^{-2m_\ell}\right)\left(\int\limits_{\mathcal{X}} \frac{\sigma^2(x)}{f(x)} w(x)dx\right)$$
$$\times \operatorname{tr}\left\{\left(\int\limits_{\mathcal{U}} R(z)R(z)'dz\right)^{-1} \int\limits_{\mathcal{U}} \left(\partial^m R(z)\right)\left(\partial^m R(z)\right)' dz\right\}; \tag{B.4}$$

$$\begin{aligned}
\mathscr{B}_{K,d,m} = \frac{1}{2^{2(K+d-[m])}} & \sum_{\substack{k,\ \tilde{k}\\ [k]=[\tilde{k}]=K}} \left(\prod_{\ell=1}^{d} |\mathcal{X}_\ell|^{k_\ell+\tilde{k}_\ell-2m_\ell}\right) \left(\int_{\mathcal{X}} \left(\partial^k\mu(x)\right)\left(\partial^{\tilde{k}}\mu(x)\right) w(x)dx\right) \\
& \times \left\{ \frac{1}{(k-m)!(\tilde{k}-m)!} \int_{\mathcal{U}} z^{k+\tilde{k}-2m} dz \right. \\
& \qquad \times + \frac{1}{k!\tilde{k}!} \operatorname{tr}\Bigg[\left(\int_{\mathcal{U}} R(z)R(z)' dz\right)^{-1} \int_{\mathcal{U}} R(z) z^{k} dz \\
& \qquad \times \int_{\mathcal{U}} R(z)' z^{\tilde{k}} dz \left(\int_{\mathcal{U}} R(z)R(z)' dz\right)^{-1} \int_{\mathcal{U}} \left(\partial^m R(z)\right)\left(\partial^m R(z)\right)' dz \Bigg] \\
& - \frac{1}{k!(\tilde{k}-m)!} \int_{\mathcal{U}} (\partial^m R(z))' z^{\tilde{k}-m} dz \left(\int_{\mathcal{U}} R(z)R(z)' dz\right)^{-1} \int_{\mathcal{U}} R(z) z^{k} dz \\
& \left. - \frac{1}{\tilde{k}!(k-m)!} \int_{\mathcal{U}} (\partial^m R(z))' z^{k-m} dz \left(\int_{\mathcal{U}} R(z)R(z)' dz\right)^{-1} \int_{\mathcal{U}} R(z) z^{\tilde{k}} dz \right\}.
\end{aligned} \tag{B.5}$$

Combining all the above steps we obtain the final result, with $\min_{1\leq j\leq J_n^d} \mathbb{1}_{n,j} = 1$ w.p.a. 1. □

Finally, we note that for $[m] = 0$:

$$\begin{aligned}
\mathscr{B}_{K,d,0} = \frac{1}{2^{2K+d}} & \sum_{\substack{k,\tilde{k}\\ [k]=[\tilde{k}]=K}} \frac{1}{k!\tilde{k}!} \left(\prod_{\ell=1}^{d} |\mathcal{X}_\ell|^{k_\ell+\tilde{k}_\ell}\right) \left\{\int_{\mathcal{X}} \left(\partial^k\mu(x)\right)\left(\partial^{\tilde{k}}\mu(x)\right) w(x)dx\right\} \\
& \times \left\{ \int_{\mathcal{U}} z^{k+\tilde{k}} dz - \int_{\mathcal{U}} R(z)' z^{\tilde{k}} dz \left(\int_{\mathcal{U}} R(z)R(z)' dz\right)^{-1} \int_{\mathcal{U}} R(z) z^{k} dz \right\}.
\end{aligned} \tag{B.6}$$

Bahadur Representation and Asymptotic Normality

Proof of Theorem III.7. Using the linearity condition on $\theta(\cdot)$, we express the remainder in Eqn. (3.3) as $\theta(\nu_n) = T_{n1} + T_{n2} + T_{n3} + T_{n4}$, where $T_{n1} = \sum_{j=1}^{J_n^d} \Theta_j' \mathbb{1}_{n,j} \Omega_j^{-1}(\Omega_j -$

$\hat{\Omega}_j)\hat{\Omega}_j^{-1}\tilde{R}_j'(Y-G)/(nq_j)$, $T_{n2} = \sum_{j=1}^{J_n^d} \Theta_j' \mathbb{1}_{n,j} \hat{\Omega}_j^{-1} \tilde{R}_j'(G - \tilde{R}_j \beta_j^0)/(nq_j)$,

$$T_{n3} = \sum_{j=1}^{J_n^d} \mathbb{1}_{n,j}(\Theta_j' \beta_j^0 - \theta(\mu_j)),$$

and $T_{n4} = \sum_{j=1}^{J_n^d} (\mathbb{1}_{n,j} - 1)[\theta(\mu_j) + \Theta_j' \Omega_j^{-1} \tilde{R}_j'(Y-G)/(nq_j)]$.

Further, we can write $T_{n1} = T_{n11} - T_{n12}$, with $T_{n11} = \sum_{i=1}^n \sum_{j=1}^{J_n^d} \Theta_j' \mathbb{1}_{n,j} \Omega_j^{-1}(\Omega_j - \hat{\Omega}_j)\Omega_j^{-1}(\Omega_j - \hat{\Omega}_j)\hat{\Omega}_j^{-1}\tilde{R}_j(X_i)\varepsilon_i/(nq_j)$ and

$$T_{n12} = \sum_{i=1}^n \sum_{j=1}^{J_n^d} \Theta_j' \mathbb{1}_{n,j} \Omega_j^{-1}(\hat{\Omega}_j - \Omega_j)\Omega_j^{-1}\tilde{R}_j(X_i)\varepsilon_i/(nq_j).$$

Applying linearity and then continuity of the functional $\theta(\cdot)$ from Assumption III.6, followed by Lemmas B.1, B.3, B.4, and B.5 we have the following bound on $|T_{n11}|$:

$$\begin{aligned}
|T_{n11}| &= \left|\theta\left(\sum_{j=1}^{J_n^d} (\tilde{R}_j(\cdot))' \mathbb{1}_{n,j} \Omega_j^{-1}(\Omega_j - \hat{\Omega}_j)\Omega_j^{-1}(\Omega_j - \hat{\Omega}_j)\hat{\Omega}_j^{-1} \frac{\tilde{R}_j'(Y-G)}{nq_j}\right)\right| \\
&\leq C \max_{[m]\leq s} \left\| \sum_{j=1}^{J_n^d} (\partial^m \tilde{R}_j(\cdot))' \mathbb{1}_{n,j} \Omega_j^{-1}(\Omega_j - \hat{\Omega}_j)\Omega_j^{-1}(\Omega_j - \hat{\Omega}_j)\hat{\Omega}_j^{-1} \frac{\tilde{R}_j'(Y-G)}{nq_j} \right\|_\infty \\
&\leq C \left(\max_{1\leq j\leq J_n^d} \max_{[m]\leq s} \|\partial^m \tilde{R}_j(\cdot)\|_\infty\right) \left(\max_{1\leq j\leq J_n^d} \left|\Omega_j - \hat{\Omega}_j\right|^2\right) \left(\max_{1\leq j\leq J_n^d} \left|\mathbb{1}_{n,j}\hat{\Omega}_j^{-1}\right|\right) \\
&\qquad \times \left(\max_{1\leq j\leq J_n^d} \left|\Omega_j^{-1}\right|^2\right) \left(\max_{1\leq j\leq J_n^d} \left|\frac{\tilde{R}_j'(Y-G)}{nq_j}\right|\right) \\
&= O_p\left(\frac{J_n^{(2-\xi/2)d+s} \log(J_n^d)^{1+\xi/2}}{n^{3/2}}\right).
\end{aligned}$$

For T_{n12}, begin by defining

$$W_j(i,l) = \mathbb{1}_{n,j} \Omega_j^{-1} \left(\tilde{R}_j(X_i)\tilde{R}_j(X_i)' - \mathbb{E}[\tilde{R}_j(X_i)\tilde{R}_j(X_i)']\right) \Omega_j^{-1} \tilde{R}_j(X_l)\varepsilon_l,$$

so that we write $T_{n12} = \sum_{j=1}^{J_n^d} \sum_{i=1}^n \sum_{l=1}^n \Theta_j' W_j(i,l)/(n^2 q_j^2)$. Observe that $\mathbb{E}[T_{n12}] = 0$ and that unless $i = h$ and $l = m$, $\mathbb{E}\left[W_j(i,l)W_j(h,m)\right] = 0$. By Lemmas B.1 and B.3, Assumption III.1(c), and $q_j \asymp J_n^{-d}$, we have:

$$\max_{1\leq j\leq J_n^d} \mathbb{E}[W_j(i,i)W_j(i,i)']$$

$$\leq C(\max_{1\leq j\leq J_n^d}|\Omega_j^{-1}|^4)(|\tilde{R}_j(\cdot)|_\infty^6)(\sup_{x\in\mathcal{X}}\sigma^2(x))\max_{1\leq j\leq J_n^d}\mathbb{E}[\mathbb{1}_{P_j}(X_i)] = O(J_n^{-d}).$$

Similarly $\max_{1\leq j\leq J_n^d}\mathbb{E}\left[W_j(i,l)W_j(i,l)'\right] = O(J_n^{-2d})$. Further, by Assumption III.6 and Lemma A.1 give that:

$$\max_{1\leq j\leq J_n^d}|\Theta_j| \leq C\max_{1\leq j\leq J_n^d}\max_{[m]\leq s}\|\partial^m\tilde{R}_j(\cdot)\|_\infty = O(J_n^s)$$

. Therefore the variance of T_{n2} is $O_p(J_n^{2d+2s}/n^2)$ because

$$\begin{aligned}\mathbb{E}[T_{n2}^2] &= \sum_{j=1}^{J_n^d}\frac{1}{(nq_j)^4}\sum_{i=1}^n\sum_{l=1}^n\Theta_j'\mathbb{E}\left[W_j(i,l)W_j(i,l)'\right]\Theta_j \\ &\leq \frac{CJ_n^{4d}}{n^4}\left(\max_{1\leq j\leq J_n^d}|\Theta_j|\right)\left(\max_{1\leq j\leq J_n^d}n\mathbb{E}\left[W_j(i,l)W_j(i,l)'\right]+n(n-1)\mathbb{E}\left[W_j(i,l)W_j(i,l)'\right]\right) \\ &\quad\times\left(\max_{[m]\leq s}\max_{1\leq j\leq J_n^d}\sup_{x\in P_j}(\partial^m\tilde{R}_j(\cdot))\right),\end{aligned}$$

using $q_j \asymp J_n^{-d}$, linearity and continuity of $\theta(\cdot)$, and Lemma B.1. Hence $|T_{n2}| = O_p\left(J_n^{d+s}/n\right)$, by Markov's inequality.

Similar steps as employed for T_{n11} give $|T_{n2}| = O_p(J_n^{-((S+\alpha)\wedge K-s)})$ and $|T_{n3}| = O_p(J_n^{-((S+\alpha)\wedge K-s)})$, additionally applying Lemma B.2. Finally, from $\min_{1\leq j\leq J_n^d}\mathbb{1}_{n,j} = 1$ w.p.a. 1 it follows that T_{n4} is smaller order than the other terms. This completes the proof. □

We now demonstrate a version of Theorem III.7 that holds with probability one.

Theorem B.10. *Let Assumption III.6 hold with $s \leq S\wedge(K-1)$, and consider the representation in Eqn.* (3.3). *If the conditions of Theorem B.6 hold, then:*

$$\theta(\nu_n) = O_{as}\left(\frac{J_n^{(3/2-\xi/2)d+s}\log(J_n^d)^{(1+\xi)/2}}{n}+J_n^{-((S+\alpha)\wedge K-s)}\right).$$

Proof of Theorem B.10. Use the same expansion as in the proof of Theorem III.7. Remainders T_{n2}, T_{n3}, and T_{n4} are handled identically, applying the almost sure versions of the same steps, but T_{n1} is bounded directly, using the same steps as for T_{n11} above. □

Proof of Theorem III.8(a). By assumption $\sigma^2(x)$ is bounded away from zero, so under

Assumption III.1(c) we have $\Gamma_j \asymp \Omega_j$. Further by $q_j \asymp J_n^{-d}$ and Lemma B.3 we have:

$$V_n \asymp \mathbb{E}[\Psi_n(X)^2] = \|\Psi_n\|_2^2, \quad \text{and} \quad V_n \asymp \sum_{j=1}^{J_n^d} \Theta_j'\Omega_j^{-1}\Theta_j/q_j \asymp J_n^d \sum_{j=1}^{J_n^d} |\Theta_j|^2. \tag{B.7}$$

The condition that $\theta(\nu_n) = o_p(\sqrt{V_n}/\sqrt{n})$ and the result of Theorem III.7 immediately give the triangular array representation of the Theorem. By construction, $\mathbb{E}[\Psi_n(X_i)\varepsilon_i/\sqrt{nV_n}] = 0$ and $\sum_{i=1}^n \mathbb{E}[\big(\Psi_n(X_i)\varepsilon_i/\sqrt{nV_n}\big)^2] = 1$. It remains to verify the Lindeberg condition. For any $\delta > 0$, by the Hölder and Markov inequalities, Assumption III.1(c), $V_n \asymp \|\Psi_n\|_2^2$ by Eqn. (B.7), and the conditions of the Theorem,

$$\begin{aligned}
&\sum_{i=1}^n \mathbb{E}\left[\left(\frac{\Psi_n(X_i)\varepsilon_i}{\sqrt{nV_n}}\right)^2 \mathbb{1}\left\{\left|\frac{\Psi_n(X_i)\varepsilon_i}{\sqrt{nV_n}}\right| > \delta\right\}\right] \\
&\qquad \leq n \left[\mathbb{E}\left[\left(\frac{\Psi_n(X_i)\varepsilon_i}{\sqrt{nV_n}}\right)^{2+\eta}\right]\right]^{\frac{2}{2+\eta}} \left[\mathbb{P}\left[\left|\frac{\Psi_n(X_i)\varepsilon_i}{\sqrt{nV_n}}\right| > \delta\right]\right]^{\frac{\eta}{2+\eta}} \\
&\qquad \leq \frac{1}{\delta^\eta}\frac{\mathbb{E}\left[|\Psi_n(X_i)|^{2+\eta}\mathbb{E}[|\varepsilon_i|^{2+\eta} \mid X_i]\right]}{n^{\eta/2}V_n^{1+\eta/2}} = O\left(\left(\frac{\|\Psi_n\|_{2+\eta}}{n^{\eta/(4+2\eta)}\|\Psi_n\|_2}\right)^{2+\eta}\right) \to 0.
\end{aligned}$$

Convergence in distribution follows by the Lindeberg-Feller central limit theorem.

For the second conclusion, observe that by $\mathbb{1}_{n,j} = 1$ w.p.a. 1, uniformly in j, we have $\hat{V}_n/V_n - 1 = T_{n1} + T_{n2} + T_{n3} + o_p(1)$, where

$$T_{n1} = V_n^{-1}\hat{V}_n - V_n^{-1}\sum_{j=1}^{J_n^d} \mathbb{1}_{n,j}\Theta_j'\hat{\Omega}_j^{-1}\tilde{\Gamma}_j\hat{\Omega}_j^{-1}\Theta_j/q_j,$$

$$T_{n2} = V_n^{-1}\sum_{j=1}^{J_n^d} \mathbb{1}_{n,j}\Theta_j'(\hat{\Omega}_j^{-1} + \Omega_j^{-1})\tilde{\Gamma}_j\left(\hat{\Omega}_j^{-1} - \Omega_j^{-1}\right)\Theta_j/q_j,$$

$$T_{n3} = V_n^{-1}\sum_{j=1}^{J_n^d} \Theta_j'\Omega_j^{-1}\left(\tilde{\Gamma}_j - \Gamma_j\right)\Omega_j^{-1}\Theta_j/q_j,$$

and $\tilde{\Gamma}_j = \sum_{i=1}^n \tilde{R}_j(X_i)\tilde{R}_j(X_i)'\varepsilon_i^2/(nq_j)$. First, expanding the squared terms, T_{n1} can be split into two terms, and upon applying Lemmas B.1 and B.4, $q_j \asymp J_n^{-d}$, Eqns. (B.2) and (B.7), and the condition of the Theorem, we find that

$$T_{n1} = V_n^{-1}\sum_{j=1}^{J_n^d} \mathbb{1}_{n,j}\Theta_j'\hat{\Omega}_j^{-1}\left(\frac{1}{nq_j}\sum_{i=1}^n \tilde{R}_j(X_i)\tilde{R}_j(X_i)'(\hat{\mu}(X_i) - \mu(X_i))^2\right)\hat{\Omega}_j^{-1}\Theta_j/q_j$$

$$
\begin{aligned}
&- V_n^{-1} \sum_{j=1}^{J_n^d} \mathbb{1}_{n,j} \Theta_j' \hat{\Omega}_j^{-1} \left(\frac{1}{nq_j} \sum_{i=1}^{n} \tilde{R}_j(X_i) \tilde{R}_j(X_i)' 2\varepsilon_i (\hat{\mu}(X_i) - \mu(X_i)) \right) \hat{\Omega}_j^{-1} \Theta_j / q_j \\
&\leq \left(\max_{1 \leq j \leq J_n^d} \mathbb{1}_{n,j} |\hat{\Omega}_j^{-1}|^2 \right) \left(\max_{1 \leq j \leq J_n^d} \|\tilde{R}_j(\cdot)\|_\infty^2 \right) (\|\hat{\mu} - \mu\|_\infty) \\
&\quad \times \left\{ \|\hat{\mu} - \mu\|_\infty \frac{J_n^d}{V_n} \sum_{j=1}^{J_n^d} |\Theta_j|^2 \frac{1}{nq_j} \sum_{i=1}^{n} \mathbb{1}_{P_j}(X_i) + \frac{J_n^d}{V_n} \sum_{j=1}^{J_n^d} |\Theta_j|^2 \frac{1}{nq_j} \sum_{i=1}^{n} \mathbb{1}_{P_j}(X_i) |\varepsilon_i| \right\} \\
&= O_p\left(\|\hat{\mu} - \mu\|_\infty\right) \times \{o_p(1) O(1) O_p(1) + O_p(1)\} = o_p(1),
\end{aligned}
$$

where the final line additionally uses Assumption III.1(c) and the final relation of Eqn. (B.7) to give:

$$
\mathbb{E}\left[\frac{J_n^d}{V_n} \sum_{j=1}^{J_n^d} |\Theta_j|^2 \frac{1}{nq_j} \sum_{i=1}^{n} \mathbb{1}_{P_j}(X_i) |\varepsilon_i| \right] \leq C \frac{J_n^d}{V_n} \sum_{j=1}^{J_n^d} |\Theta_j|^2 \frac{\mathbb{E}\left[\mathbb{1}_{P_j}(X_i) \mathbb{E}\left[|\varepsilon_i| \mid X_i \right] \right]}{q_j} = O(1).
$$

By Lemma B.1 and otherwise identical steps to the above, we get:

$$
\mathbb{E}[V_n^{-1} \sum_{j=1}^{J_n^d} |\Theta_j|^2 |\tilde{\Gamma}_j| / q_j] = O(1).
$$

Therefore, applying Lemmas B.3 and B.4: $|T_{n2}| \leq C(\max_{1 \leq j \leq J_n^d} \mathbb{1}_{n,j} |\hat{\Omega}_j^{-1}|^3 \wedge |\Omega_j^{-1}|^3) \times (\max_{1 \leq j \leq J_n^d} |\hat{\Omega}_j - \Omega_j|) V_n^{-1} \sum_{j=1}^{J_n^d} |\Theta_j|^2 |\tilde{\Gamma}_j| / q_j = o_p(1)$.

Finally, referring to the definitions in Eqn. (3.3), observe that $T_{n3} = \sum_{i=1}^{n} T_{n3}(i)/n$, where $T_{n3}(i) = V_n^{-1}(\Psi_n(X_i)^2 \varepsilon_i^2 - \mathbb{E}[\Psi_n(X_i)^2 \varepsilon_i^2])$, so that $\mathbb{E}\left[T_{n3}(i)\right] = 0$. Consider two cases. First, suppose $\eta < 2$. Then by Burkholder's inequality, the fact that for $\delta \in (0, 1)$, $(a + b)^{(1+\delta)/2} \leq a^{(1+\delta)/2} + b^{(1+\delta)/2}$, the c_r inequality, Jensen's inequality, Assumption III.1(c), and Eqn. (B.7):

$$
\begin{aligned}
&\mathbb{E}\left[\left| \frac{1}{n} \sum_{i=1}^{n} T_{n3}(i) \right|^{1+\eta/2} \right] \\
&\leq \frac{C}{n^{1+\eta/2}} \mathbb{E}\left[\left| \sum_{i=1}^{n} T_{n3}(i)^2 \right|^{(1+\eta/2)/2} \right] \\
&\leq \frac{C}{n^{1+\eta/2}} \mathbb{E}\left[\sum_{i=1}^{n} |T_{n3}(i)|^{1+\eta/2} \right] \\
&\leq \frac{C}{n^{\eta/2}} \frac{\mathbb{E}\left[|\Psi_n(X_i)|^{2+\eta} \mathbb{E}\left[|\varepsilon_i|^{2+\eta} \mid X \right] \right] + \left(\mathbb{E}\left[\Psi_n(X_i)^2 \sigma^2(X) \right] \right)^{1+\eta/2}}{V_n^{1+\eta/2}}
\end{aligned}
$$

$$= O\left(\left(\frac{\|\Psi_n\|_{2+\eta}}{n^{\eta/(4+2\eta)}\|\Psi_n\|_2}\right)^{2+\eta}\right) \to 0.$$

Next, for the case of $\eta \geq 2$ we utilize only the fourth moment to find that:

$$\mathbb{E}\left[\left(\sum_{i=1}^{n} T_{n3}(i)/n\right)^2\right] \leq V_n^{-2}\mathbb{E}\left[\Psi_n(X_i)^4\varepsilon_i^4\right]/n = O\left(\|\Psi_n\|_4^4 n^{-1}\|\Psi_n\|_2^{-4}\right) \to 0,$$

again using Jensen's inequality, Assumption III.1(c), and the first relation of Eqn. (B.7). In either case, $T_{3n} = o_p(1)$ by Markov's inequality. □

Proof of Theorem III.8(b). By Assumption III.1(c), the Cauchy-Schwarz and triangle inequalities, and the conditions of the Theorem: $V_n - V = \mathbb{E}[(\Psi_n(X)^2 - \Psi(X)^2)\sigma^2(X)] \leq C\mathbb{E}[(\Psi_n(X)-\Psi(X))^2]^{1/2}\mathbb{E}[(\Psi_n(X)-\Psi(X)+2\Psi(X))^2]^{1/2} \leq C\|\Psi_n - \Psi\|_2(\|\Psi_n - \Psi\|_2 + 2\|\Psi\|_2) \to 0$, whence the second conclusion.

Convergence in distribution follows under the assumed moment condition on $\Psi(X)$ and a standard central limit theorem, because

$$\begin{aligned}
&\sqrt{n}(\theta(\hat{\mu}) - \theta(\mu))/\sqrt{V_n} - \sum_{i=1}^{n}\Psi(X_i)\varepsilon_i/(\sqrt{nV}) \\
&= \sum_{i=1}^{n}[(\Psi_n(X_i) - \Psi(X_i))\varepsilon_i/(\sqrt{nV}) + \Psi_n(X_i)\varepsilon_i/(\sqrt{nV})(\sqrt{V/V_n} - 1)] \\
&\qquad\qquad + \sqrt{n}\theta(\nu_n)/\sqrt{V_n} \\
&= o_p(1)
\end{aligned}$$

using the above result, the assumed mean-square convergence of $\Psi_n(X)$, and the remainder condition of the Theorem.

For the final conclusion, as in the proof of Theorem III.8(a) write $\hat{V}_n/V_n - 1 = T_{n1} + T_{n2} + T_{n3} + o_p(1)$, for T_{n1}, T_{n2}, and T_{n3} defined there. As above, $T_{n1} = o_p(1)$ and $T_{n2} = o_p(1)$. Next, $T_{n3} = (V_n^{-1} - V^{-1}\sum_{i=1}^{n}\Psi_n(X_i)^2\varepsilon_i^2/n + \sum_{i=1}^{n}[\Psi_n(X_i)^2 - \Psi(X_i)^2]\varepsilon_i^2/(nV) + \sum_{i=1}^{n}(\Psi(X_i)^2\varepsilon_i^2 - V)/(nV)$, where the first two terms are $o_p(1)$ as in the second conclusion and Markov's inequality, and the third by the law of large numbers. □

APPENDIX C

Proofs for Chapter 4

Proofs for Chapter 4

Throughout the appendix, C denotes a generic positive constant that may take different values in different places. All bounds are uniform in $j = 1, \cdots, J_n^d$ unless explicitly noted otherwise. For A, a scalar, vector, or matrix, let $|A|$ denote the Euclidean norm.

Define $\Omega_{j,t} = q_j^{-1}\mathbb{E}[\mathbf{1}_{P_j}(X_i)D_{t,i}R(X_i)R(X_i)']$, $\varepsilon_t = (Y_1(t) - \mu_t(X_1), \cdots, Y_n(t) - \mu_t(X_n))'$, and $E_t = ((e_t(X_1) - D_{t,1})/e_t(X_1), \cdots, (e_t(X_n) - D_{t,n})/e_t(X_n))'$. We now collect several useful results regarding the nonparametric partition regression estimator. Details and proofs may be found in Cattaneo and Farrell (2011a). All results given in the appendix implicitly utilize an appropriate non-singular linear transformation of the polynomial basis, although the same notation is maintained for simplicity. Cattaneo and Farrell (2011a) give details on the appropriate rotation and demonstrate its existence under the conditions imposed in Theorem 1.

Lemma C.1. *Under the conditions of Theorem 1, the following results hold:*

(A-1) $\max_{1\le j\le J_n^d} \sup_{x\in P_j} |R_j(x)| \le C < \infty$.

(A-2) There exists vectors $\gamma_{\mu,j}$ *and* $\gamma_{e,j}$, $j = 1, \cdots, J_n^d$, *such that*

$$\max_{1\le j\le J_n^d} \sup_{x\in P_j} |\mu_t(x) - R_j(x)'\gamma_{\mu,j}| = O(J_n^{-K\wedge S_\mu}),$$

and

$$\max_{1\le j\le J_n^d} \sup_{x\in P_j} \left|\frac{1}{e_t(x)} - R_j(x)'\gamma_{e,j}\right| = O(J_n^{-K\wedge S_e}).$$

(A-3) $\lambda_{\min}(\Omega_{j,t}) \geq C > 0$.

(A-4) $\max_{1\leq j\leq J_n^d} \left|\hat{\Omega}_{j,t} - \Omega_{j,t}\right|^2 = O_p(J_n^d \log(J_n)/n)$.

(A-5) $\max_{1\leq j\leq J_n^d} \left|R_{j,t}'\varepsilon_t/(nq_j)\right|^2 = O_p(J_n^{9d/7} \log(J_n)^{5/7}/n)$.

(A-6) $\max_{1\leq j\leq J_n^d} \left|R_j' E_t/(nq_j)\right|^2 = O_p(J_n^d \log(J_n)/n)$.

(A-7) $\max_{1\leq j\leq J_n^d} \sup_{x\in P_j} \left|\hat{\mu}_{j,t}(x) - \mu_{j,t}(x)\right|^2 = O_p(J_n^{9d/7} \log(J_n)^{5/7}/n + J_n^{-2K\wedge S_\mu})$.

Results (A-3) and (A-4) imply that $\max_{1\leq j\leq J_n^d} |\Omega_{j,t}^{-1}| \leq C$, $\max_{1\leq j\leq J_n^d} |\hat{\Omega}_{j,t}^{-1}| = O_p(1)$, and $\mathbb{P}(\max_{1\leq j\leq J_n^d} |\mathbf{1}_{n,j} - 1| = 0) \to 1$.

Proof of Theorem 1

Let $\gamma_{\mu,j}$ and $\gamma_{e,j}$ be as given in (A-4). Observe that

$$\sqrt{n}(\hat{\mu}_t - \mu_t) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_t(Y_i, X_i, T_i) + \epsilon_{n,1} + \epsilon_{n,2} + \epsilon_{n,3} + \epsilon_{n,4} + \epsilon_{n,5} + \epsilon_{n,6},$$

where

$$\epsilon_{n,1} = \frac{1}{\sqrt{n}}\sum_{j=1}^{J_n^d}\sum_{i=1}^n \left(1 - \frac{D_{t,i}}{e_t(X_i)}\right)\mathbf{1}_{n,j} R_j(X_i)'(R_{j,t}'R_{j,t})^{-1}R_{j,t}'\varepsilon_t,$$

$$\epsilon_{n,2} = \frac{1}{\sqrt{n}}\sum_{j=1}^{J_n^d}\sum_{i=1}^n \mathbf{1}_{n,j}\left(\frac{1}{e_t(X_i)} - \gamma_{e,j}'R_j(X_i)\right) D_{t,i}R_j(X_i)'(R_{j,t}'R_{j,t})^{-1}R_{j,t}'\varepsilon_t,$$

$$\epsilon_{n,3} = \frac{1}{\sqrt{n}}\sum_{j=1}^{J_n^d}\sum_{i=1}^n \mathbf{1}_{n,j}\left(\gamma_{e,j}'R_j(X_i) - \frac{1}{e_t(X_i)}\right)\mathbf{1}_{P_j}(X_i)D_{t,i}(Y_i - \mu_t(X_i)),$$

$$\epsilon_{n,4} = \frac{1}{\sqrt{n}}\sum_{j=1}^{J_n^d}\sum_{i=1}^n\sum_{k=1}^n \mathbf{1}_{n,j}R_j(X_i)'(R_{j,t}'R_{j,t})^{-1}D_{t,k}R_j(X_k)\left(\mu_t(X_k) - R_j(X_k)'\gamma_{\mu,j}\right),$$

$$\epsilon_{n,5} = \frac{1}{\sqrt{n}}\sum_{j=1}^{J_n^d}\sum_{i=1}^n \mathbf{1}_{n,j}\mathbf{1}_{P_j}(X_i)(R(X_i)'\gamma_{\mu,j} - \mu_t(X_i)),$$

$$\epsilon_{n,6} = \frac{1}{\sqrt{n}}\sum_{j=1}^{J_n^d}\sum_{i=1}^n \left(\mathbf{1}_{n,j} - 1\right)\mathbf{1}_{P_j}(X_i)\left\{\frac{D_{t,i}(Y_i - \mu_t(X_i))}{e_t(X_i)} + \mu_t(X_i)\right\}.$$

Consider each reminder $\epsilon_{n,1}$–$\epsilon_{n,6}$. First, $\epsilon_{n,1} = \epsilon_{n,11} + \epsilon_{n,12} + \epsilon_{n,13}$ with

$$\epsilon_{n,11} = \frac{1}{\sqrt{n}} \sum_{j=1}^{J_n^d} \mathbf{1}_{n,j} \left[R_j' E_t\right]' \Omega_{j,t}^{-1} \left[\Omega_{j,t} - \hat{\Omega}_{j,t}\right] \Omega_{j,t}^{-1} \left[\Omega_{j,t} - \hat{\Omega}_{j,t}\right] \hat{\Omega}_{j,t}^{-1} \left[R_{j,t}' \varepsilon_t/(nq_j)\right] = o_p(1),$$

$$\epsilon_{n,12} = -\frac{1}{\sqrt{n}} \sum_{j=1}^{J_n^d} \mathbf{1}_{n,j} \left[R_j' E_t\right]' \Omega_{j,t}^{-1} \left[\hat{\Omega}_{j,t} - \Omega_{j,t}\right] \Omega_{j,t}^{-1} \left[R_{j,t}' \varepsilon_t/(nq_j)\right] = o_p(1),$$

$$\epsilon_{n,13} = \frac{1}{\sqrt{n}} \sum_{j=1}^{J_n^d} \mathbf{1}_{n,j} \left[R_j' E_t\right]' \Omega_{j,t}^{-1} \left[R_{j,t}' \varepsilon_t/(nq_j)\right] = o_p(1),$$

because

$$\begin{aligned} |\epsilon_{n,11}| &\leq \sqrt{n} \max_{1\leq j\leq J_n^d} \left|R_j' E_t/(nq_j)\right| \max_{1\leq j\leq J_n^d} \left|\mathbf{1}_{n,j} \hat{\Omega}_{j,t}^{-1}\right| \max_{1\leq j\leq J_n^d} \left|\hat{\Omega}_{j,t} - \Omega_{j,t}\right|^2 \\ &\quad \times \max_{1\leq j\leq J_n^d} \left|\Omega_{j,t}^{-1}\right|^2 \max_{1\leq j\leq J_n^d} \left|R_{j,t}' \varepsilon_t/(nq_j)\right| \\ &= \sqrt{n} O_p(J_n^{3d/2} \log(J_n)^{3/2}/n^{3/2}) O_p(J_n^{9d/14} \log(J_n)^{5/14}/\sqrt{n}) = o_p(1), \end{aligned}$$

and simple variance bounds give $\mathbb{E}\left[\epsilon_{n,12}^2\right] = O(J_n^{2d}/n^2) = o_p(1)$ and $\mathbb{E}\left[\epsilon_{n,13}^2\right] = O(J_n^d/n) = o_p(1)$, as $\mathbb{E}[R_j(X_i)E_{t,i}|X_i] = 0$, $\mathbb{E}[q_j^{-1} D_{t,i} R_j(X_i) R_j(X_i)' - \Omega_{t,j}] = 0$ and $\mathbb{E}[D_{t,i} R_j(X_i) \varepsilon_{t,i}|X_i, T_i] = 0$.

Next, observe that $\epsilon_{n,2} = \epsilon_{n,21} + \epsilon_{n,22}$ with

$$\begin{aligned} \epsilon_{n,21} = -\frac{1}{\sqrt{n}} \sum_{j=1}^{J_n^d} \mathbf{1}_{n,j} \sum_{i=1}^{n} \left(\frac{1}{e_t(X_i)} - \gamma_{e,j}' R_j(X_i)\right) \\ \times D_{t,i} R_j(X_i)' \hat{\Omega}_{j,t}^{-1} [\hat{\Omega}_{j,t} - \Omega_{j,t}] \Omega_{j,t}^{-1} \left[R_{j,t}' \varepsilon_t/(nq_j)\right] = o_p(1), \end{aligned}$$

$$\epsilon_{n,22} = \frac{1}{\sqrt{n}} \sum_{j=1}^{J_n^d} \mathbf{1}_{n,j} \sum_{i=1}^{n} \left(\frac{1}{e_t(X_i)} - \gamma_{e,j}' R_j(X_i)\right) D_{t,i} R_j(X_i)' \Omega_{j,t}^{-1} \left[R_{j,t}' \varepsilon_t/(nq_j)\right] = o_p(1),$$

because

$$\begin{aligned} |\epsilon_{n,21}| &\leq O(J_n^{-K\wedge S_e}) \max_{1\leq j\leq J_n^d} \left|\mathbf{1}_{n,j} \hat{\Omega}_{j,t}^{-1}\right| \max_{1\leq j\leq J_n^d} \left|\hat{\Omega}_{j,t} - \Omega_{j,t}\right| \\ &\quad \times \max_{1\leq j\leq J_n^d} \left|\Omega_{j,t}^{-1}\right| \max_{1\leq j\leq J_n^d} \left|R_{j,t}' \varepsilon_t/(nq_j)\right| \frac{1}{\sqrt{n}} \sum_{j=1}^{J_n^d} \sum_{i=1}^{n} \mathbf{1}_{P_j}(X_i) \\ &= O(J_n^{-K\wedge S_e}) \sqrt{n} O_p(J_n^{d/2} \log(J_n)^{1/2}/\sqrt{n}) O_p(J_n^{9d/14} \log(J_n)^{5/14}/\sqrt{n}) = o_p(1), \end{aligned}$$

and $\mathbb{E}\left[\epsilon^2_{n,22}\right] = O(J_n^{-2K\wedge S_e}) = o_p(1)$.

Next, $\epsilon_{n,3} = o_p(1)$ because

$$\mathbb{E}\left[\epsilon^2_{n,3}\right] \leq \sum_{j=1}^{J_n^d} \mathbb{E}\left[\mathbf{1}_{P_j}(X_i)\left(\gamma'_{e,j}R(X_i) - \frac{1}{e_t(X_i)}\right)^2 (Y_i(t) - \mu_t(X_i))^2\right]$$
$$= O(J_n^{-2K\wedge S_e}) = o(1).$$

Next, $\epsilon_{n,4} = o_p(1)$ because

$$\begin{aligned} |\epsilon_{n,4}| &\leq \sqrt{n}O(J_n^{-K\wedge S_\mu})\frac{J_n^d}{n^2}\sum_{j=1}^{J_n^d}\sum_{i=1}^{n}\sum_{k=1}^{n}\mathbf{1}_{n,j}R_j(X_i)'\hat{\Omega}^{-1}_{j,t}D_{t,k}R_j(X_k) \\ &\leq \sqrt{n}O_p(J_n^{-K\wedge S_\mu})\frac{J_n^d}{n^2}\sum_{j=1}^{J_n^d}\sum_{i=1}^{n}\sum_{k=1}^{n}\mathbf{1}_{P_j}(X_i)\mathbf{1}_{P_j}(X_k) = \sqrt{n}O_p(J_n^{-K\wedge S_\mu}) = o_p(1). \end{aligned}$$

Next, $\epsilon_{n,5} = o_p(1)$ because

$$|\epsilon_{n,5}| \leq \sqrt{n}O(J_n^{-K\wedge S_\mu})\frac{1}{n}\sum_{j=1}^{J_n^d}\sum_{i=1}^{n}\mathbf{1}_{P_j}(X_i) = \sqrt{n}O(J_n^{-K\wedge S_\mu}).$$

Finally, $\epsilon_{n,6} = o_p(1)$ because $\mathbb{P}(\max_{1\leq j\leq J_n^d}|\mathbf{1}_{n,j} - 1| = 0) \to 1$. ■

Proof of Theorem 2

For $\hat{V}_{W,[t,s]}$, first define $\tilde{\Sigma}_{j,t} = n^{-1}\sum_{i=1}^{n} R_j(X_i)R_j(X_i)'D_{t,i}(Y_i - \mu_t(X_i))^2$ and $\tilde{L}_j = \frac{1}{nq_j}\sum_{i=1}^{n} R_j(X_i)D_{t,i}/e_t(X_i)$, then note that

$$\hat{V}_{W,[t,t]} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{n,j}\frac{D_{t,i}\varepsilon^2_{t,i}}{e_t(X_i)} + \epsilon_{W,n,1} + \epsilon_{W,n,2} + \epsilon_{W,n,3} + \epsilon_{W,n,4} + \epsilon_{W,n,5},$$

where

$$\epsilon_{W,n,1} = \sum_{j=1}^{J_n^d}\mathbf{1}_{n,j}\hat{L}'_j\hat{\Omega}^{-1}_{tj}\left(\hat{\Sigma}_{tj} - \tilde{\Sigma}_{tj}\right)\hat{\Omega}^{-1}_{tj}\hat{L}_j,$$

$$\epsilon_{W,n,2} = \sum_{j=1}^{J_n^d}\mathbf{1}_{n,j}\left\{\hat{L}'_j\hat{\Omega}^{-1}_{tj}\tilde{\Sigma}_{tj}\hat{\Omega}^{-1}_{tj}\left(\hat{L}_j - \tilde{L}_j\right) + \left(\hat{L}_j - \tilde{L}_j\right)'\hat{\Omega}^{-1}_{tj}\tilde{\Sigma}_{tj}\hat{\Omega}^{-1}_{tj}\tilde{L}_j\right\},$$

$$\epsilon_{W,n,3} = \sum_{j=1}^{J_n^d} \mathbf{1}_{n,j} \tilde{L}_j' \hat{\Omega}_{tj}^{-1} \tilde{\Sigma}_{tj} \hat{\Omega}_{tj}^{-1} \left(\frac{1}{nq_j} \sum_{i=1}^{n} R_j(X_i) D_{t,i} \left(\frac{1}{e_t(X_i)} - R_j(X_i)'\gamma_{e,j} \right) \right),$$

$$\epsilon_{W,n,4} = \sum_{j=1}^{J_n^d} \mathbf{1}_{n,j} \left(\frac{1}{nq_j} \sum_{i=1}^{n} R_j(X_i)' D_{t,i} \left(\frac{1}{e_t(X_i)} - R_j(X_i)'\gamma_{e,j} \right) \right)' \hat{\Omega}_{tj}^{-1} \tilde{\Sigma}_{tj} \gamma_{e,j},$$

$$\epsilon_{W,n,5} = \frac{1}{n} \sum_{j=1}^{J_n^d} \mathbf{1}_{n,j} \sum_{i=1}^{n} \mathbf{1}_{P_j}(X_i) D_{t,i} \varepsilon_i^2 \left(R_j(X_i)'\gamma_{e,j} - \frac{1}{e_t(X_i)} \right) \left(R_j(X_i)'\gamma_{e,j} + \frac{1}{e_t(X_i)} \right).$$

Now, $|\epsilon_{W,n,1}| \leq |\epsilon_{W,n,11}| + |\epsilon_{W,n,12}| = o_p(1)$, where applying (A-7),

$$\begin{aligned} |\epsilon_{W,n,11}| &= \left| \sum_{j=1}^{J_n^d} \mathbf{1}_{n,j} \hat{L}_j' \hat{\Omega}_{tj}^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} 2R_j(X_i) R_j(X_i)' D_{t,i} \varepsilon_i \left(\hat{\mu}_{tj}(X_i) - \mu_t(X_i) \right) \right) \hat{\Omega}_{tj}^{-1} \hat{L}_j \right| \\ &\leq C \max_{1 \leq j \leq J_n^d} \left| \hat{L}_j \right|^2 \max_{1 \leq j \leq J_n^d} \left| \mathbf{1}_{n,j} \hat{\Omega}_{tj}^{-1} \right|^2 \sup_{x \in \mathcal{X}} |\hat{\mu}_t(x) - \mu(x)| \sum_{j=1}^{J_n^d} \sum_{i=1}^{n} \mathbf{1}_{P_j}(X_i) D_{t,i} |\varepsilon_{t,i}|/n \\ &= o_p(1), \end{aligned}$$

and similarly $|\epsilon_{W,n,12}| = o_p(1)$.

Next, $\epsilon_{W,n,2} = o_p(1)$ because

$$\begin{aligned} |\epsilon_{W,n,2}| &\leq 2 \left(\max_{1 \leq j \leq J_n^d} \left| \hat{L}_j \right| + \left| \tilde{L}_j \right| \right) \left(\max_{1 \leq j \leq J_n^d} \left| \mathbf{1}_{n,j} \hat{\Omega}_{tj}^{-1} \right|^2 \right) \\ &\quad \times \left(\max_{1 \leq j \leq J_n^d} \left| \frac{1}{nq_j} \sum_{i=1}^{n} R_j(X_i) \left(1 - \frac{D_{t,i}}{e_t(X_i)} \right) \right| \right) \\ &\quad \times \sum_{j=1}^{J_n^d} \sum_{i=1}^{n} \left| R_j(X_i) R_j(X_i)' D_{t,i} \varepsilon_i^2 \right| / n \\ &= O_p(1) O_p(1) O_p \left(\left(\frac{J_n^d \log J_n}{n} \right)^{1/2} \right) O_p(1) = o_p(1), \end{aligned}$$

where

$$\begin{aligned} &\mathbb{E} \left[\sum_{j=1}^{J_n^d} \sum_{i=1}^{n} \left| R_j(X_i) R_j(X_i)' D_{t,i} \varepsilon_i^2 \right| / n \right] \\ &\qquad \leq C \left(\sup_{x \in \mathcal{X}} |R(x)|^2 \right) \frac{1}{n} \sum_{j=1}^{J_n^d} \sum_{i=1}^{n} \mathbb{E} \left[\mathbf{1}_{P_j}(X_i) \right] = O(1). \end{aligned}$$

Next, $\epsilon_{W,n,3} = o_p(1)$ because

$$\begin{aligned}|\epsilon_{W,n,3}| &\leq \left(\max_{1\leq j\leq J_n^d}\left|\tilde{L}_j\right|\right)\left(\max_{1\leq j\leq J_n^d}\left|\mathbf{1}_{n,j}\hat{\Omega}_{tj}^{-1}\right|^2\right)\\ &\quad\times\sum_{j=1}^{J_n^d}\frac{1}{n^2q_j}\sum_{i=1}^{n}\sum_{l=1}^{n}\left|R_j(X_l)R_j(X_l)'\right|D_{t,l}\varepsilon_l^2\left|R_j(X_i)\left(\frac{1}{e_t(X_i)}-R_j(X_i)'\gamma_{e,j}\right)\right|\\ &= O_p\left(J_n^{-K\wedge S_e}\right) = o_p(1),\end{aligned}$$

since

$$\begin{aligned}&\mathbb{E}\left[\sum_{j=1}^{J_n^d}\frac{1}{n^2q_j}\sum_{i=1}^{n}\sum_{l=1}^{n}\left|R_j(X_l)R_j(X_l)'\right|D_{t,l}\varepsilon_l^2\left|R_j(X_i)\left(\frac{1}{e_t(X_i)}-R_j(X_i)'\gamma_{e,j}\right)\right|\right]\\ &= O_p\left(\left(1+\frac{J_n^d}{n}\right)J_n^{-K\wedge S_e}\right) = O\left(J_n^{-K\wedge S_e}\right).\end{aligned}$$

Identical reasoning shows $|\epsilon_{W,n,4}| = o_p(1)$ and $|\epsilon_{W,n,5}| = o_p(1)$. Hence $\hat{V}_{W,[t,s]} = V_{W,[t,s]} + o_p(1)$, as $\mathbb{P}(\min_{1\leq j\leq J_n^d}\mathbf{1}_{n,j} = 1)\to 1$.

Now consider the "between" term of the variance estimator. For $\hat{V}_{B,[t,s]}$, note that

$$\begin{aligned}\hat{V}_{B,[t,s]} &= \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_t(X_i)\hat{\mu}_s(X_i) - \hat{\mu}_s\frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_t(X_i) - \hat{\mu}_t\frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_s(X_i) + \hat{\mu}_s\hat{\mu}_t\\ &= \frac{1}{n}\sum_{i=1}^{n}\mu_t(X_i)\mu_s(X_i) - \hat{\mu}_s\frac{1}{n}\sum_{i=1}^{n}\mu_t(X_i) - \hat{\mu}_t\frac{1}{n}\sum_{i=1}^{n}\mu_s(X_i) + \hat{\mu}_s\hat{\mu}_t\\ &\qquad+\epsilon_{B,n,1}+\epsilon_{B,n,2}+\epsilon_{B,n,3}+\epsilon_{B,n,4}+\epsilon_{B,n,5},\end{aligned}$$

where

$$\epsilon_{B,n,1} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_t(X_i)-\mu_t(X_i)\right)\left(\hat{\mu}_s(X_i)-\mu_s(X_i)\right),$$

$$\epsilon_{B,n,2} = \frac{1}{n}\sum_{i=1}^{n}\mu_t(X_i)\left(\hat{\mu}_s(X_i)-\mu_s(X_i)\right),\qquad \epsilon_{B,n,3} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_t(X_i)-\mu_t(X_i)\right)\mu_s(X_i),$$

$$\epsilon_{B,n,4} = -\hat{\mu}_s\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_t(X_i)-\mu_t(X_i)\right),\qquad \epsilon_{B,n,5} = -\hat{\mu}_t\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_s(X_i)-\mu_s(X_i)\right).$$

Thus, because $\hat{\mu}-\mu = o_p(1)$ and Result (A-7) holds under the assumptions of the theorem, $\epsilon_{B,n,k} = o_p(1)$ for $k = 1,\cdots,5$, and $\hat{V}_{B,[t,s]} = V_{B,[t,s]} + o_p(1)$ as stated. ■

BIBLIOGRAPHY

BIBLIOGRAPHY

Abadie, A. (2005): "Semiparametric difference-in-differences estimators," *Review of Economic Studies*, 72, 1–19.

Abadie, A., D. Drukker, J. L. Herr, and G. W. Imbens (2004): "Implementing Matching Estimators for Average Treatment Effects in Stata," *The Stata Journal*, 4(3), 290–311.

Abadie, A., and G. W. Imbens (2006): "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 74(1), 235–267.

Andrews, D. W. K. (1991): "Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Regression Models," *Econometrica*, 59(2), 307–345.

Andrews, D. W. K., and P. Guggenberger (2009): "Incorrect asymptotic size of subsampling procedures based on post-consistent model selection estimators," *Journal of Econometrics*, 152, 19–27.

Bach, F. R. (2008): "Consistency of the Group Lasso and Multiple Kernel Learning," *Journal of Machine Learning Research*, 9, 1179–1225.

——— (2010): "Self-concordant analysis for logistic regression," *Electronic Journal of Statistics*, 4, 384–414.

Banerjee, A. N. (2007): "A method of estimating the average derivative," *Journal of Econometrics*, 136, 65–88.

Bang, H., and J. M. Robins (2005): "Doubly Robust Estimation in Missing Data and Causal Inference Models," *Biometrics*, 61, 962–972.

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., X. Chen, V. Chernozhukov, and K. Kato (2012): "On the Asymptotic Theory for Least Squares Series: Pointwise and Uniform Results," *Arxiv preprint arXiv:1212.0442*.

Belloni, A., and V. Chernozhukov (2011a): "ℓ_1-Penalized quantile regression in high-dimensional sparse models," *Annals of Statistics*, 39(1), 82–130.

——— (2011b): "Least Squares After Model Selection in High-dimensional Sparse Models," *Arxiv preprint arXiv:1001.0188v4*.

Belloni, A., V. Chernozhukov, and C. Hansen (2013): "Inference on Treatment Effects after Selection Amongst High-Dimensional Controls," *cemmap working paper CWP26/13*.

Belloni, A., V. Chernozhukov, and Y. Wei (2013): "Honest Confidence Regions for Logistic Regression with a Large Number of Controls," *arXiv:1304.3969v1*.

Berk, R., L. Brown, A. Buja, K. Zhang, and L. Zhao (2013): "Valid Post-Selection Inference," *Annals of Statistics*, 4(2), 802–837.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009): "Simultaneous Analysis of LASSO and Dantzig Selector," *Annals of Statistics*, 37(4), 1705–1732.

Braun, W. J., and L.-S. Huang (2005): "Kernel Spline Regression," *The Canadian Journal of Statistics*, 33(2), 259–278.

Buhlmann, P., and S. van de Geer (2011): *Statistics for High-Dimensional Data*, Springer Series in Statistics. Springer-Verlag, Berlin.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2012): "Robust Data-Driven Inference in the Regression-Discontinuity Design," *working paper available at http://www.umich.edu/ cattaneo*.

Cattaneo, M. D. (2010): "Efficient Semiparametric Estimation of Multi-valued Treatment Effects under Ignorability," *Journal of Econometrics*, 155(2), 138–154.

Cattaneo, M. D., D. M. Drukker, and A. D. Holland (forthcoming): "Estimation of multivalued treatment effects under conditional independence," *The Stata Journal*.

Cattaneo, M. D., and M. Farrell (2011a): "Optimal Convergence Rates, Bahadur Representation, and Asymptotic Normality of Partitioning Estimators," working paper.

Cattaneo, M. D., and M. H. Farrell (2011b): "Efficient Estimation of the Dose Response Function under Ignorability using Subclassification on the Covariates," in *Advances in Econometrics: Missing Data Methods*, ed. by D. Drukker, vol. 27A, pp. 93–127. Emerald Group Publishing Limited.

——— (2013): "Optimal Convergence Rates, Bahadur Representation, and Asymptotic Normality of Partitioning Estimators," *Journal of Econometrics*, 174, 127–143.

Cattaneo, M. D., G. W. Imbens, C. Pinto, and G. Ridder (2009): "Subclassification on the Propensity Score: Large Sample Properties," work in progress.

Chen, X. (2007): "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 6B of *Handbook of Econometrics*, chap. 76. Elsevier.

Chen, X., and T. M. Christensen (2013): "Optimal Uniform Convergence Rates for Sieve Nonparametric Instrumental Variables Regression," *Cowles Foundation discussion paper no. 1923.*

Chen, X., H. Hong, and A. Tarozzi (2004): "Semiparametric Efficiency in GMM Models of Nonclassical Measurement Errors, Missing Data and Treatment Effects," Cowles Foundation Discussion Paper No. 1644.

——— (2008): "Semiparametric Efficiency in GMM Models With Auxiliary Data," *Annals of Statistics*, 36(2), 808–843.

Chen, X., and J. Huang (2003): "Sup norm convergence rate and asymptotic normality for a class of linear sieve estimators," *Work in progress, shared in personal communication.*

Cochran, W. G. (1968): "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies," *Biometrics*, 24(2), 295–313.

de Jong, R. M. (2002): "A note on 'Convergence rates and asymptotic normality for series estimators': uniform convergence rates," *Journal of Econometrics*, 11, 1–9.

de la Peña, V. H., T. L. Lai, and Q.-M. Shao (2009): *Self-Normalized Processes: Limit Theory and Statistical Applications*, Probability and Its Applications. Springer.

Dehejia, R. H., and S. Wahba (1999): "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94(448), 1053–1062.

——— (2002): "Propensity Score-Matching Methods for Nonexperimental Causal Studies," *The Review of Economics and Statistics*, 84(1), 151–161.

Efron, B. (2013): "Estimation and Accuracy after Model Selection," *Stanford University working paper.*

Eggermont, P. P. B., and V. N. LaRiccia (2009): *Maximum Penalized Likelihood Estimation, Volume II: Regression.* Springer.

Fama, E. F., and K. R. French (2008): "Dissecting Anomalies," *The Journal of Finance*, 63(4), 1653–1678.

Fan, J., and I. Gijbels (1996): *Local polynomial modelling and its applications.* Chapman and Hall, London.

Firpo, S. (2007): "Efficient Semiparametric Estimation of Quantile Treatment Effects," *Econometrica*, 75(1), 259–276.

Györfi, L., M. Kohler, A. Krzyżak, and H. Walk (2002): *A Distribution-Free Theory of Nonparametric Regression.* Springer-Verlag.

Györfi, L., M. Kohler, A. Krzyżak, and H. Walk (2002): *A Distribution-Free Theory of Nonparametric Regression.* Springer, New York.

Hahn, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66(2), 315–331.

He, X., and Q.-M. Shao (2000): "On Parameters of Increasing Dimensions," *Journal of Multivariate Analysis*, 73, 1201–35.

Heckman, J., H. Ichimura, and P. Todd (1997): "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *Review of Economic Studies*, 64, 605–654.

Heckman, J. J., and E. J. Vytlacil (2007): "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation," in *Handbook of Econometrics, vol. VI*, ed. by J. Heckman, and E. Leamer, pp. 4780–4874. Elsevier Science B.V.

Hirano, K., G. W. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score," *Econometrica*, 71(4), 1161–1189.

Holland, P. W. (1986): "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81(396), 945–960.

Horowitz, J. L., and C. F. Manski (2000): "Nonparametric Analysis of Randomized Experiments With Missing Covariate and Outcome Data," *Journal of the American Statistical Association*, 95(449), 77–84.

Huang, J., and T. Zhang (2010): "The Benefit of Group Sparsity," *Annals of Statistics*, 38(4), 1978–2004.

Huang, J. Z. (2003): "Local asymptotics for polynomial spline regression," *Annals of Statistics*, 31(5), 1600–1635.

Ichimura, H., and P. E. Todd (2007): "Implementing Nonparametric and Semiparametric Estimators," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 6B of *Handbook of Econometrics*, chap. 74. Elsevier.

Imai, K., and D. A. van Dyk (2004): "Causal Inference With General Treatment Regimes: Generalizing the Propensity Score," *Journal of the American Statistical Association*, 99(467), 854–866.

IMBENS, G. W. (2000): "The Role of the Propensity Score in Estimating Dose-Response Functions," *Biometrika*, 87(3), 706–710.

——— (2004): "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *Review of Economics and Statistics*, 86(1), 4–29.

IMBENS, G. W., AND T. LEMIEUX (2008): "Regression discontinuity designs: A guide to practice," *Journal of Econometrics*, 142, 615–635.

IMBENS, G. W., W. K. NEWEY, AND G. RIDDER (2007): "Mean-Squared-Error Calculations for Average Treatment Effects," *working paper.*

IMBENS, G. W., AND J. M. WOOLDRIDGE (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47(1), 5–86.

KANG, J. D. Y., AND J. L. SCHAFER (2007): "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data," *Statistical Science*, 22(4), 523–539.

KOHLER, M., A. KRZYŻAK, AND H. WALK (2006): "Rates of convergence for partitioning and nearest neighbor regression estimates with unbounded data," *Journal of Multivariate Analysis*, 97, 311–323.

——— (2009): "Optimal global rates of convergence for nonparametric regression with unbounded data," *Journal of Statistical Planning and Inference*, 139, 1286–1296.

KOLAR, M., J. LAFFERTY, AND L. WASSERMAN (2011): "Union Support Recovery in Multi-task Learning," *Journal of Machine Learning Research*, 12, 2415–2435.

KONG, E., O. LINTON, AND Y. XIA (2010): "Uniform Bahadur representation for local polynomial estimates of M-regression and its application to the additive model," *Econometric Theory*, 26, 1529–1564.

LALONDE, R. J. (1986): "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, 76(4), 604–620.

LECHNER, M. (2001): "Identification and estimation of causal effects of multiple treatments under the conditional independence assumption," in *Econometric Evaluations of Active Labor Market Policies*, ed. by M. Lechner, and E. Pfeiffer, pp. 43–58. Physica, Heidelberg.

LEEB, H., AND B. M. PÖTSCHER (2005): "Model Selection and Inference: Facts and Fiction," *Econometric Theory*, 21, 21–59.

——— (2008a): "Can One Estimate the Unconditional Distribution of Post-Model-Selection Estimators?," *Econometric Theory*, 24, 338–376.

——— (2008b): “Sparse estimators and the oracle property, or the return of Hodges’ estimator,” *Journal of Econometrics*, 142, 201–211.

Lounici, K., M. Pontil, S. van de Geer, and A. B. Tsybakov (2009): “Taking advantage of sparsity in multi-task learning,” in *Proceedings of the 22nd Annual Conference on Learning Theory (COLT 2009)*, pp. 73–82. Omnipress.

——— (2011): “Oracle Inequalities and Optimal Inference under Group Sparsity,” *Annals of Statistics*, 39(4), 2164–2204.

Masry, E. (1996): “Multivariate Local Polynomial Regression for Time Series: Uniform Strong Consistency and Rates,” *Journal of Time Series Analysis*, 17(6), 571–599.

Negahban, S. N., P. Ravikumar, M. J. Wainwright, and B. Yu (2012): “A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers,” *Statistical Science*, 27(4), 538–557.

Newey, W. K. (1994a): “The Asymptotic Variance of Semiparametric Estimators,” *Econometrica*, 62(6), 1349–1382.

——— (1994b): “Kernel Estimation of Partial Means and a General Variance Estimator,” *Econometric Theory*, 10, 233–253.

——— (1997): “Convergence rates and asymptotic normality for series estimators,” *Journal of Econometrics*, 79, 147–168.

Newey, W. K., and D. L. McFadden (1994): “Large sample estimation and hypothesis testing,” in *Handbook of Econometrics*, ed. by R. F. Engle, and D. McFadden, vol. 4 of *Handbook of Econometrics*, chap. 36, pp. 2111–2245. Elsevier.

Obozinski, G., M. J. Wainwright, and M. I. Jordan (2011): “Support Union Recovery in High-Dimensional Multivariate Regression,” *Annals of Statistics*, 39(1), 1–47.

Pötscher, B. M. (2009): “Confidence Sets Based on Sparse Estimators Are Necessarily Large,” *Sankhyā*, 71-A, 1–18.

Pötscher, B. M., and H. Leeb (2009): “On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding,” *Journal of Multivariate Analysis*, 100, 2065–2085.

Raskutti, G., M. J. Wainwright, and B. Yu (2010): “Restricted Eigenvalue Properties for Correlated Gaussian Designs,” *Journal of Machine Learning Research*, 11, 2241–2259.

Robins, J. M., and A. Rotnitzky (1995): “Semiparametric Efficiency in Multivariate Regression Models with Missing Data,” *Journal of the American Statistical Association*, 90(429), 122–129.

Romano, J. P. (2004): "On non-parametric testing, the uniform behaviour of the t-test, and related problems," *Scandinavian Journal of Statistics*, 31(4), 567–584.

Rosenbaum, P. R., and D. B. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70(1), 41–55.

——— (1984): "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score," *Journal of the American Statistical Association*, 79(1), 516–524.

Rudelson, M., and S. Zhou (2011): "Reconstruction from anisotropic random measurements," *Arxiv preprint arXiv:1106.1151.*

Ruppert, D., and M. P. Wand (1994): "Multivariate Locally Weighted Least Squares Regression," *Annals of Statistics*, 22(3), 1346–1370.

Smith, J. A., and P. E. Todd (2005): "Does matching overcome LaLonde's critique of nonexperimental estimators?," *Journal of Econometrics*, 125, 305–353.

Stoker, T. (1986): "Consistent estimation of scaled coefficients," *Econometrica*, 54(6), 1461–1481.

Stone, C. J. (1982): "Optimal Global Rates of Convergence for Nonparametric Regression," *Annals of Statistics*, 10(4), 1040–1053.

Tan, Z. (2010): "Bounded, efficient and doubly robust estimation with inverse weighting," *Biometrik*, 97, 661–682.

Tanabe, K., and M. Sagae (1992): "An Exact Cholesky Decomposition and the Generalized Inverse of the Variance-Covariance Matrix of the Multinomial Distribution, with Applications," *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(1), 211–219.

Tsiatis, A. A. (2006): *Semiparametric Theory and Missing Data*. Springer, New York.

Tukey, J. W. (1947): "Non-Parametric Estimation II. Statistically Equivalent Blocks and Tolerance Regions–The Continuous Case," *Annals of Mathematical Statistics*, 18(4), 529–539.

van de Geer, S. (2008): "High-Dimensional Generalized Linear Models and the Lasso," *Annals of Statistics*, 36, 614–645.

van de Geer, S., and P. Buhlmann (2009): "On the conditions used to prove oracle results for the Lasso," *Electronic Journal of Statistics*, 3, 1360–1392.

van der Laan, M., and J. M. Robins (2003): *Unified Methods for Censored Longitudinal Data and Causality*. Springer-Verlag.

van der Vaart, A. (1991): “On Differentiable Functionals,” *Annals of Statistics*, 19(1), 178–204.

von Bahr, B., and C.-G. Esseen (1965): “Inequalities for the rth absolute moment of a sum of random variables, $1 \leqq r \leqq 2$,” *Annals of Mathematical Statistics*, 36(1), 299–303.

Wei, F., and J. Huang (2010): “Consistent group selection in high-dimensional linear regression,” *Bernoulli*, 16(4), 1369–1384.

White, H., and X. Lu (2011): “Causal Diagrams for Treatment Effect Estimation with Application to Efficient Covariate Selection,” *Review of Economics and Statistics*, 93(4), 1453–1459.

Wooldridge, J. M. (2007): “Inverse probability weighted estimation for general missing data problems,” *Journal of Econometrics*, 141, 1281–1301.

——— (2010): *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, 2 edn.

Yuan, M., and Y. Lin (2006): “Model selection and estimation in regression with grouped variables,” *Journal of the Royal Statistical Society. Series B*, 68(1), 46–67.

Zhao, P., and B. Yu (2006): “On Model Selection Consistency of Lasso,” *Journal of Machine Learning Research*, 7, 2541–2563.

Zhou, S., and D. A. Wolfe (2000): “On Derivative Estimation in Spline Regression,” *Statistica Sinica*, 10, 93–108.

Zou, H. (2006): “The Adaptive Lasso and Its Oracle Properties,” *Journal of the American Statistical Association*, 101(476), 1418–1429.