The London School of Economics and Political Science

Essays in Semiparametric and High Dimensional Methods

Chen Qiu

A thesis submitted to the Department of Economics of the London School of Economics and Political Science for the degree of Doctor of Philosophy

London, June 2020

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it). The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without my prior written consent. I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party. I declare that my thesis consists of 32860 words.

Statement of co-authored work

I confirm that Chapter 1 was jointly co-authored with Professor Taisuke Otsu and I contributed at least 50% of this work.

Acknowledgement

I am deeply indebted to my supervisor, Taisuke Otsu for invaluable guidance and support. I am also grateful to the rest of the econometrics and applications faculty groups at the LSE, including Peter Robinson, Javier Hidalgo, Marcia Schafgans, Vassilis Hajivassiliou, Tatiana Komarova, Yike Wang, Daniel Sturm, Rachael Meager, Steve Pischke, Wouter Den Haan, Tim Besley, Pasquale Schiraldi, for their insightful discussions and great help. I also thank professional services staff at the LSE, especially Mark Wilbor, Anna Watmuff, Deborah Adams and Lakmini Staskus, for their patient and outstanding administrative support.

To my family and friends.

Abstract

Chapter 1 is concerned with estimation of functionals of a latent weight function that satisfies possibly high dimensional multiplicative moment conditions. A leading example is functionals of the stochastic discount factor in asset pricing. This chapter proposes to estimate the latent weight function by an information theoretic approach combined with the ℓ_1-penalization technique to deal with high dimensional moment conditions under sparsity. This chapter derives asymptotic properties of the proposed estimator and illustrates the proposed method by a theoretical example on treatment effect analysis and empirical example on the stochastic discount factor.

In Chapter 2, I introduce a semiparametric framework called the average regression functional, defined as a continuous linear function of a conditional expectation function. This framework is relevant to many empirical problems, including estimating average treatment effects, regression discontinuity design away from cut-off and measurement error with auxiliary data. I develop a new minimax methodology to estimate average regression functionals. Embedded in a penalized series space, this new strategy exploits a minimax property of a key nonparametric component of the average regression functional and aims to directly control main remainder bias. I then construct a new class of estimators, called minimax learners and show they are straightforward to implement due to their minimum distance representation.

In Chapter 3, I separately study in detail asymptotic properties of minimax learners as the ratio of controls to sample size goes to zero, constant and infinity. Root-n normality is established under weak conditions for all three cases. In simulations where selection bias is mild, minimax learners behave stably, maintain small mean square error and do not over control; if selection bias is substantial, minimax learners are able to correctly reduce mean square error as more relevant controls are added.

As an empirical illustration, Chapter 4 revisits the work of Ferraz and Finan (2011) that studies the effect of electoral accountability on corruption. With plausibly exogenous treatment, one of their main empirical strategies is OLS with many controls. I find estimates from OLS change considerably as more covariates are sequentially added to the regression. Minimax learners, on the other hand, perform stably and lead to economically coherent conclusions, even when the number of controls is much larger. Other popular off-the-shelf shrinkage methods do not work as well as minimax learners.

Contents

1 Information theoretic approach to multiplicative models 11

1.1 Introduction . . . 11
1.1.1 Motivation . . . 11
1.1.2 Methodology . . . 13
1.1.3 Related literature . . . 16
1.2 Low dimensional case . . . 17
1.3 High dimensional case . . . 23
1.3.1 Estimation of ω_0 . . . 24
1.3.2 Debiased estimator for θ_0 . . . 26
1.3.3 Post selection estimator for θ_0 . . . 28
1.3.4 Targeted debiasing estimator for θ_0 . . . 30
1.4 Theoretical application: treatment effect . . . 31
1.5 Empirical application: stochastic discount factor . . . 33
1.5.1 Methodology . . . 34
1.5.2 Data . . . 34
1.5.3 Empirical results . . . 35
1.5.3.1 Low dimensional and intermediate cases . . . 35
1.5.3.2 High dimensional case . . . 36
1.5.3.3 Time series property of penalized SDF estimates 37
1.5.4 Tables and figures . . . 38

2 Minimax learning for average regression functionals: framework 46

2.1 Introduction . . . 46
2.1.1 Related literature . . . 49
2.1.2 Notations and definitions for the rest of the thesis . . . 51
2.2 Average regression functional and related examples . . . 52
2.3 Minimax learning in penalized series space . . . 57
2.3.1 Identification: three ways . . . 57
2.3.2 Calibration of the Riesz Representer . . . 58
2.3.3 Implementation . . . 60
2.3.4 Construction of minimax learners . . . 61

2.3.4.1 When $\frac{k}{n} \to c$ for some $c \in [0, 1)$ 61
2.3.4.2 When $\frac{k}{n} \to \infty$ 62

3 Minimax learning for average regression functionals: theory 64
3.1 Theory: minimax BP 64
3.1.1 Asymptotic normality and semiparametric efficiency when $\frac{k}{n} \to 0$ 65
3.1.2 Consistent estimation of variance 69
3.1.3 $\sqrt{n}$ normality when $\frac{k}{n} \to c < 1$ 70
3.2 Theory: minimax DR 72
3.2.1 Data-driven selection of penalties 78
3.3 Monte Carlo exercises 79
3.3.1 Performance of minimax BP learner under moderately high dimensions 79
3.3.1.1 Baseline result with mild selection bias 80
3.3.1.2 Robustness check: considerable selection bias . . 82
3.3.1.3 Robustness check: sensitivity to penalty coefficient 82
3.3.2 Performance of GMEN learner under high dimensions . . . 83
3.3.3 Tables and figures 84

4 Minimax learning for average regression functionals: application 96
4.1 Main empirical framework 97
4.2 Main empirical results 97
4.3 Controlling for ability and experience 99
4.4 Accounting for many more controls 99
4.5 Tables and figures 100

A Supplementary materials for Chapter 1 110
A.1 Mixing 110
A.2 Proofs for low dimensional case 111
A.2.1 Lemmas 111
A.2.2 Proof of Theorem 1.1 115
A.2.3 Proof of Theorem 1.2 116
A.2.4 Proof of Proposition 1.1 119
A.3 Proofs for high dimensional case 119
A.3.1 Proof of Theorem 1.3 119
A.3.2 Proof of Theorem 1.4 120
A.3.3 Proof of Theorem 1.5 122
A.3.4 Proof of Theorem 1.6 124
A.3.5 Lemmas 126

B Supplementary materials for Chapter 2 **131**
B.1 Derivations of RR in some examples . . . 131
B.2 Proof of Proposition 2.1 . . . 133
B.3 Measure of design uncertainty when $\frac{k}{n} \to c < 1$ . . . 134

C Supplementary materials for Chapter 3 **135**
C.1 Basic lemmas . . . 135
C.1.1 Useful maximal inequalities . . . 135
C.1.2 More results on least square projection . . . 136
C.1.3 Asymptotic linear forms . . . 136
C.1.4 Lemmas for term R_{1BP} . . . 137
C.1.5 Lemma for term R_{1DR}. . . . 138
C.1.6 Lemmas for term R_2 . . . 139
C.2 Proofs for main results when $\frac{k}{n} \to 0$ . . . 140
C.2.1 Additional convergence results when $\frac{k}{n} \to 0$ . . . 140
C.2.2 Additional results for controlling stochastic equicontinuity terms . . . 144
C.2.3 Additional results for Theorem 3.1 . . . 146
C.2.4 Additional results for Theorem 3.2 . . . 149
C.2.5 Proofs for main results when $\frac{k}{n} \to 0$. . . . 152
C.2.5.1 Proof of Theorem 3.1 . . . 152
C.2.5.2 Proof of Corollary 3.1 . . . 152
C.2.5.3 Proof of Theorem 3.2 . . . 153
C.3 Proofs for main results when $\frac{k}{n} \to c < 1$ . . . 155
C.3.1 Additional results on asymptotic boundedness . . . 155
C.3.2 Additional results for Theorem 3.3. . . . 159
C.3.3 Proofs for main results when $\frac{k}{n} \to c < 1$ . . . 160
C.3.3.1 Proof of Theorem 3.3 . . . 160
C.3.3.2 Proof of Corollary 3.2 . . . 161
C.3.4 Sufficient conditions . . . 162
C.4 Proofs for main results when $\frac{k}{n} \to \infty$ . . . 164
C.4.1 Additional convergence results when $\frac{k}{n} \to \infty$ . . . 165
C.4.2 Additional results for Theorem 3.4 . . . 167
C.4.3 Additional results for Theorem 3.5 . . . 170
C.4.4 Proof of main results when $\frac{k}{n} \to \infty$ . . . 173
C.4.4.1 Proof of Theorem 3.4 . . . 173
C.4.4.2 Proof of Theorem 3.5 . . . 173
C.4.5 Sufficient conditions . . . 174
C.5 Proofs for basic lemmas and other related results . . . 177

C.5.1 Proof of Lemma C.1 177
C.5.2 Proof of Lemma C.2 177
C.5.3 Proof of Lemma C.3 178
C.5.4 Proof of Lemma C.4 178
C.5.5 Proofs of Lemmas C.5 and C.6 179
C.5.6 Proof of Lemma C.7 179
C.5.7 Proof of Lemma C.8 179
C.5.8 Proof of Lemma C.9 180
C.5.9 Proof of Lemma C.10 181
C.5.10 Proof of Lemma C.11 181
C.5.11 Proof of Lemma C.12 182
C.5.12 Proof of Lemma C.28 183

Bibliography **187**

List of Tables

1.1 Cross sectional regression in low dimensional case 38
1.2 Cross sectional regression in intermediate case 39
1.3 Cross sectional regression in high dimensional case 43
1.4 Time series properties of estimated SDF from high dimensional case 45

3.1 Construction of B-splines: mild selection bias 84
3.2 Construction of orthogonal polynomials: mild selection bias 84
3.3 Bias and RMSE using B-splines, 10000 Monte Carlo, $\lambda_1 = 0.002$, mild selection bias 85
3.4 Bias and RMSE using orthogonal polynomials, 10000 Monte Carlo, $\lambda_1 = 0.001$, mild selection bias 85
3.5 Coverage probability using B-splines, 10000 Monte Carlo, $\lambda_1 = 0.002$, mild selection bias 88
3.6 Coverage probability using orthogonal polynomials, 10000 Monte Carlo, $\lambda_1 = 0.001$, mild selection bias 88
3.7 Construction of B-splines: considerable selection bias 89
3.8 Bias and RMSE using B-splines, 10000 Monte Carlo, $\lambda_1 = 0.002$, considerable selection bias 89
3.9 Sensitivity of $\hat{\theta}_{BP}$ to λ_1 using B-splines, 10000 Monte Carlo, mild selection bias 91

4.1 Effect of reelection incentives on corruption: baseline results . . . 101
4.2 Effect of reelection incentives on alternative measure of corruption 104
4.3 Effect of reelection incentives on alternative measure of corruption 105
4.4 Effect of reelection incentives on corruption: controlling for political experience 106
4.5 Effect of reelection incentives on corruption: controlling for political ability 107
4.6 Effect of reelection incentives on corruption: minimax DR with many controls 108
4.7 Effect of reelection incentives on corruption: minimax BP with many controls 109

List of Figures

1.1 Summary of cross sectional regression against different penalty levels in high dimension case ($K = 300$ or 425; $T = 360$) 40
1.2 Number of active portfolios selected under 300 portfolios case . . 41
1.3 Number of active portfolios selected under 425 portfolios case . . 42
1.4 Time series plot of estimated SDF in high dimensional case: July 1993 - December 2010.
Grey shaded area represents NBER recessions 44

3.1 Bias and RMSE, B-splines, mild selection bias 86
3.2 Bias and RMSE, orthogonal polynomials, mild selection bias . . . 87
3.3 Bias and RMSE, B-splines, considerable selection bias 90
3.4 Sensitivity of $\hat{\theta}_{BP}$ to λ_1 using B-splines, mild selection bias 92
3.5 Performance of GMEN learner under high dimensions (1) 93
3.6 Performance of GMEN learner under high dimensions (2) 94
3.7 Performance of GMEN learner under high dimensions (3) 95

4.1 Effect of a lame duck mayor on a 5.5 million transfer: OLS 102
4.2 Effect of a lame duck mayor on a 5.5 million transfer: minimax BP learner 103

Chapter 1

Information theoretic approach to multiplicative models

1.1 Introduction

1.1.1 Motivation

In empirical analysis, economic information and other statistical information are commonly characterized by moment conditions on observables. The generalized method of moments provides a unified framework to analyze the moment condition models and numerous extensions have been proposed in the econometrics literature. In this chapter, we consider the following moment condition model taking a multiplicative form:

$$\mathbb{E}[\omega(X)g(X)] = r, \tag{1.1}$$

where X is a vector of observables, $\omega : \mathcal{X} \to (0, \infty)$ is an *unknown* weight function, g is a vector of *known* functions of X, and r is a vector of known constants or moments of observables (say, $r = \mathbb{E}[r(X)]$ for some known $r(\cdot)$). We are interested in the situation where the observables X and/or functions g are high dimensional (possibly larger than the sample size).

In general, there exists a non-trivial set of $\omega \in W$ that satisfies (1.1). In this chapter, we introduce an information theoretic approach to select a particular element $\omega_0 \in W$, and define the object of interest as its linear functional:

$$\theta_0 = \mathbb{E}[\omega_0(X)h(X, Y)], \tag{1.2}$$

where Y is another vector of observables and h is a vector of known functions of (X, Y). This chapter develops a general estimation and inference method for the

parameter θ_0 under possibly high dimensional moment conditions (1.1).

Interestingly, this setup can be motivated by somewhat distant economic problems: inference on stochastic discount factors (SDF) and missing data problems including treatment effect analysis. The latent weight ω plays the role of the SDF for the former example, and the (reciprocal of) missing probability or propensity score for the latter.

Example 1.1 (Stochastic discount factor)**.** In a discrete time economy with no arbitrage, there exists a strictly positive SDF m_t such that

$$\mathbb{E}[m_t R_{j,t+1}] = 1, \tag{1.3}$$

where $R_{j,t+1}$ is the short term return of asset j between time t and $t+1$. This equation says that any asset j in the market would share the same expected return when discounted by the SDF m_t (see Cochrane, 2009, for a review). Let $X_t = (R_{1,t+1}, \ldots, R_{K,t+1})'$ be a K-vector of observable short term returns. Then (1.3) implies $\mathbb{E}[m_t X_t] = 1$. The object of our interest in this example is the *projected* SDF onto the space of X_t defined as $\omega(X_t) = \mathbb{E}[m_t|X_t]$ so that $\mathbb{E}[\omega(X_t)X_t] = 1$. This setup can be considered as a special case of (1.1) with $g(X) = X$ and $r = 1$. Note that $\omega(X_t)$ could have the same pricing implications as m_t (Rosenberg and Engle, 2002; Cochrane, 2009). Unless market is complete, ω is generally set identified from the moment condition $\mathbb{E}[\omega(X_t)X_t] = 1$.

Inference on the SDF is one of the central topics in financial economics. For example, Christensen (2016) investigated extraction of permanent and transitory components of the SDF process, which requires estimation of $\mathbb{E}[m_t s(X_t)s(X_{t+1})']$ for a vector of known basis functions $s(\cdot)$. Christensen (2016) considered two cases: (i) m_t is directly observable, and (ii) m_t is replaced with a (parametric or nonparametric) preliminary estimator. Our information theoretic approach will provide a nonparametric estimator for some particular choice of ω and an alternative estimator for $\mathbb{E}[m_t s(X_t)s(X_{t+1})']$ designed for high dimensional setups. □

Example 1.2 (Missing data)**.** Consider the problem of estimating a population mean from incomplete outcome data (see Roderick et al., 2002, for a survey). For each unit $i = 1, \ldots, N$, we observe an indicator variable D_i ($D_i = 1$ if unit i responds and $D_i = 0$ otherwise), outcome variable $Y_i = D_i Y_i^*$ ($Y_i = 0$ means Y_i^* is missing), and vector of covariates X_i. We are interested in the population mean $\theta = \mathbb{E}[Y_i^*]$. Under conditional independence of Y^* and D given X and certain overlap assumptions, the parameter of interest is identified as $\theta = \mathbb{E}[\omega(X)YD]$, where $\omega(X) = 1/\mathbb{P}\{D = 1|X\}$. In this setup, many estimation and inference methods for θ and their generalizations have been proposed (e.g. Tsiatis, 2007),

including the inverse probability weighted estimator $n^{-1}\sum_{i=1}^{n}\tilde{\omega}(X_i)Y_iD_i$, where $\tilde{\omega}(x)$ is a nonparametric estimator of $1/\mathbb{P}\{D=1|X=x\}$.

Our information theoretic approach can be applied in this setup to develop an alternative estimator of θ. By the law of iterated expectations, the moment conditions (1.1) may be given by

$$\mathbb{E}[\omega(X)g(X)D] = \mathbb{E}[g(X)], \tag{1.4}$$

for any vector of known functions g. Then the estimation problem of θ can be formulated as a special case of ours by replacing the expectations in (1.1) and (1.2) with the conditional expectations given $D=1$ and setting $r=\mathbb{E}[g(X)]$ and $h(X,Y)=Y$. In the recent literature of missing data analysis and causal inference, so-called the balancing covariates approach explores the moment conditions in (1.4) to find adjusting weights used for estimation of θ (e.g., Zubizarreta, 2015; Chan et al., 2016). This chapter proposes an alternative estimation method that may be considered as an extension of these papers toward high dimensional environments. □

1.1.2 Methodology

In this chapter, we propose an information theoretic approach to select an element ω_0 satisfying (1.1) and to estimate the parameters θ_0 in (1.2). Our method allows high dimensional observables and/or moment functions (possibly higher than the sample size). This feature is particularly desirable for our motivating examples. For Example 1.1, the number of assets may be very large. For Example 1.2, the number of covariates tends to be large so that the conditional independence assumption (unconfoundedness or ignorability in causal analysis) is likely to be satisfied.

A key issue for estimation of θ is how to pin down a particular weight function ω_0 satisfying (1.1). In this chapter, we address this issue by an information theoretic approach. More precisely, we regard the latent weight function as the Radon-Nikodym derivative $\omega = d\mathbb{Q}/d\mathbb{P}$, where $\mathbb{P}$ is the data generating measure of X and $\mathbb{Q}$ is a tilted model-based measure. Then the moment condition (1.1) is written as $\mathbb{E}_{\mathbb{Q}}[g(X)] = r$, where $\mathbb{E}_{\mathbb{Q}}[\cdot]$ means expectation under $\mathbb{Q}$. To pin down the tilted measure $\mathbb{Q}$, we apply the information projection based on the ϕ-divergence (e.g., Csiszar, 1975; Liese and Vajda, 1987). In particular, we consider the minimization problem using some strictly convex divergence function ϕ :

$(0, \infty) \to \mathbb{R}$, that is

$$\min_{\mathbb{Q}} \mathbb{E}_{\mathbb{P}} \left[\phi \left(\frac{d\mathbb{Q}}{d\mathbb{P}} \right) \right], \quad \text{s.t. } \mathbb{E}_{\mathbb{Q}}[g(X)] = r. \tag{1.5}$$

Under some regularity conditions,[1] it is known that this minimization problem is associated with the following dual problem (see, Theorem 2.4 and Corollary 2.6 in Borwein and Lewis, 1991, and Section 5.2.3 in Boyd et al., 2004).

$$\min_{\lambda} \mathbb{E}_{\mathbb{P}}[\phi_*(\lambda' g(X)) - \lambda' r], \tag{1.6}$$

where $\phi_*(a) = \sup_{b \in (0,\infty)} \{ab - \phi(b)\}$ is the convex conjugate of ϕ. Furthermore, the solution $\mathbb{Q}_*$ of (1.5) can obtained by the solution of the dual problem as

$$\frac{d\mathbb{Q}_*}{d\mathbb{P}}(\cdot) = \phi_*^{(1)}(\lambda_*' g(\cdot)), \tag{1.7}$$

where λ_* is the solution of (1.6) and $\phi_*^{(1)}$ is the first derivative of ϕ_*.

We now define the weight function ω_0 satisfying (1.1) of our interest. Since the dimension of g, denoted by K, grows as the sample size increases, we define ω_0 as the limit of the information projection in (1.7), i.e.,

$$\omega_0(\cdot) = \lim_{K \to \infty} \frac{d\mathbb{Q}_*}{d\mathbb{P}}(\cdot) = \lim_{K \to \infty} \phi_*^{(1)}(\lambda_*' g(\cdot)). \tag{1.8}$$

Based on this uniquely defined ω_0, our object of interest is defined as

$$\theta_0 = \mathbb{E}_{\mathbb{P}}[\omega_0(X) h(X, Y)]. \tag{1.9}$$

Let $\mathbb{E}_n[\cdot]$ be the sample mean and $\left\| \cdot \right\|_1$ be the ℓ_1-norm for a vector. Also let $\mathbb{I}\{x \in \mathcal{X}_n\}$ be a trimming term for an increasing sequence $\{\mathcal{X}_n\}$ to $\mathcal{X}$ to deal with technical problems for unbounded support of X (cf. Chen and Christensen, 2015). By taking sample counterparts for the trimmed moment functions, our information theoretic estimator of θ_0 is obtained as

$$\hat{\theta} = \mathbb{E}_n[\phi_*^{(1)}(\hat{\lambda}' g(X) \mathbb{I}\{X \in \mathcal{X}_n\}) h(X, Y)], \tag{1.10}$$

[1] Precisely, suppose $\mathbb{E}[|g_j(X)|^2] < \infty$ for each $j = 1, \ldots, K$ and there exists $\tilde{\mathbb{Q}}$ such that its Radon-Nikodym derivative $q(x) = \frac{d\tilde{\mathbb{Q}}}{d\mathbb{P}}(x)$ is strictly positive for almost every x, $\mathbb{E}_{\mathbb{P}}[q(X)^2] < \infty$, and $\mathbb{E}_{\mathbb{P}}[g(X)q(X)] = r$.

where

$$\hat{\lambda} = \begin{cases} \underset{\text{(low dimensional case)}}{\arg\min_{\lambda} \mathbb{E}_n[\phi_*(\lambda' g(X)\mathbb{I}\{X \in \mathcal{X}_n\}) - \lambda' r(X)\mathbb{I}\{X \in \mathcal{X}_n\}]} \\ \\ \underset{\text{(high dimensional case)}}{\arg\min_{\lambda} \mathbb{E}_n[\phi_*(\lambda' g(X)\mathbb{I}\{X \in \mathcal{X}_n\}) - \lambda' r(X)\mathbb{I}\{X \in \mathcal{X}_n\}] + \alpha_n \left\|\lambda\right\|_1} \end{cases}, \tag{1.11}$$

α_n is a penalty level chosen by the researcher, and $r(X)$ may be a vector of known constants (as in Example 1.1). The ℓ_1 penalty term for the high dimensional case is introduced to regularize behaviors of the estimator $\hat{\lambda}$. Although this chapter focuses on the ℓ_1-penalization (Tibshirani, 1996), other penalization methods (such as the smoothly clipped absolute deviation by Fan and Li (2001), and minimax concave penalty by Zhang, 2010) may also be applied.

It should be noted that not only the population objects ω_0 and θ_0 but also the estimator $\hat{\theta}$ depend on the choice of the divergence function ϕ. Popular choices are (i) Kullback-Leibler divergence (or relative entropy) $\phi(x) = x \log x$ with $\phi_*(y) = e^{y-1}$, (ii) reverse Kullback-Leibler divergence (or Berg entropy) $\phi(x) = -\log x$ with $\phi_*(y) = -1 - \log(-y)$, and Pearson's χ^2 divergence $\phi(x) = (x-1)^2$ with $\phi_*(y) = \frac{y^2}{4} + y$. When ω is set identified by the moment conditions (1.1) (as in Example 1.1), different choices of ϕ typically select different elements in the identified set for $\omega \in W$. An optimal choice of ϕ is beyond the scope of this chapter: such analyses typically require additional criteria, such as higher-order properties of the estimator, Bayesian interpretations, and economic motivations.[2]

We emphasize that although the construction of $\hat{\lambda}$ in (1.11) is analogous to the generalized empirical likelihood estimator for overidentified moment condition models (Newey and Smith, 2004), our setup and properties of the estimator are significantly different from theirs due to three reasons. First, our moment conditions (1.1) contain the latent weight function ω, and the information projection is applied to estimate ω_0. Second, the interpretation and property of $\hat{\lambda}$ are different from theirs. In the conventional generalized empirical likelihood estimator, $\hat{\lambda}$ plays the role of the Lagrange multiplier or shadow price for the moment conditions, and converges to zero as the sample size increases if model is correctly specified. On the other hand, in our approach, $\hat{\lambda}$ is an estimator for the

[2]For example, the estimator derived via Kullback-Leibler divergence has a quasi maximum likelihood interpretation and is consistent with the maximum entropy principle in Bayesian methods (Stutzer, 1995). Also, in asset pricing literature, the SDF estimated by the Kullback-Leibler divergence is particularly attractive since it is: (i) intrinsically related to the concept of entropy of pricing kernels, (ii) adapted to the popular log-linear modeling of the SDF, and (iii) consistent with the optimal portfolio choice with an expected utility maximizing investor who has constant absolute risk aversion utility. See Backus et al. (2014) and Hansen (2014)for further details.

dual parameter λ_* and typically does not converge to zero (even if the moment conditions are correctly specified). With this respect, our method is more in line with the sieve estimation methodology. Finally, we allow the moment conditions (1.1) to be high dimensional (possibly larger than the sample size), where the estimator $\hat{\lambda}$ has to be regularized as in (1.11).

1.1.3 Related literature

The construction of our estimator is related to the literature of exponential tilting, empirical likelihood, and its variants (for example, Kitamura and Stutzer 1997; Smith 1997; Imbens et al. 1998; see, Owen, 2001a; Kitamura, 2006, for surveys). In spite of similarity of the construction of the estimator, however, our setup and property of the Lagrange multiplier $\hat{\lambda}$ are quite different from this literature as discussed in Section 1.1.2. Indeed our treatment on the Lagrange multiplier shares more similarities with coefficients for basis functions in series or sieve estimation (see Chen, 2007, for a review).

In order to deal with high dimensional moment conditions, we adapt the general theory of the lasso with convex loss functions by Van de Geer (2008); Bühlmann and Van De Geer (2011) to our setup. In terms of inference, the debiasing method adopted in Section 1.3 of this chapter is similar to Zhang and Zhang (2014a); Van de Geer et al. (2014). We note this complements the literature on high dimensional semiparametric inference with locally/doubly robust moment conditions (e.g., Farrell, 2015; Belloni et al., 2017a). Our method can also be compared to high dimensional versions of empirical likelihood methods, such as Hjort et al. (2009); Tang and Leng (2010); Lahiri and Mukhopadhyay (2012). Again, however, our setup and treatment on $\hat{\lambda}$ are intrinsically different from this literature (typically $\hat{\lambda}$ converges to λ_* in our setup, not zero).

The main applications of our method are inference on missing data models, treatment effects, and stochastic discount factors. Here we only mention closely related papers to clarify our contributions in these fields. See Imbens and Rubin (2015); Cochrane (2009) for overview of these topics.

In the context of missing data and treatment effect analysis, the proposed method, illustrated in Section 1.4, is closely related to the recent literature on balancing weights (Zubizarreta 2015; Chan et al. 2016; Athey et al. 2018). Compared to Zubizarreta (2015); Chan et al. (2016), this chapter is considered as an extension toward a high dimensional setup. Compared to Athey et al. (2018), this chapter proposes an alternative estimation method for treatment effects under high dimensional covariates by utilizing an information theoretic approach.

In the realm of asset pricing, this chapter is closely related to information

theoretic approaches for semi-nonparametric analysis on the SDF (e.g., Kitamura and Stutzer 2002; Ghosh et al. 2015, 2016). In this context, we make three contributions. First, our method can be regarded as an extension of these existing methods to high dimensional environments (especially for a large number of assets). Second, our theoretical analysis for the low dimensional case in Section 1.2 provides a theoretical background for the analysis in Ghosh et al. (2015, 2016). Third, as mentioned in Example 1.1, this chapter can provide an alternative method to extract permanent and transitory components of the SDF process (Christensen, 2016).

Notations for Chapter 1

We work with triangular array data $\{X_{i,n}, Y_{i,n}\}_{i=1}^n$, which are considered as the first n elements of the infinite sequence $\{X_{i,n}, Y_{i,n}\}_{i=1}^\infty$ generated from a probability measure $\mathbb{P}_n$. Our asymptotic analysis is based on the array asymptotics. To simplify the notation, we suppress the subscripts and denote by $\{X_i, Y_i\}_{i=1}^n$ and $\mathbb{P}$. Also, let $\mathbb{E}[\cdot] = \mathbb{E}_{\mathbb{P}}[\cdot]$ be expectation under $\mathbb{P}$, $\mathbb{E}_n[\cdot]$ be the empirical average, $\mathbb{I}\{A\}$ be the indicator function for an event A, $|B| = \sqrt{\lambda_{\max}(B'B)}$ be the L_2 norm for a scalar, vector, or matrix B, and $a \vee b = \max\{a, b\}$. For a matrix C, let $\lambda_{\max}(C)$ and $\lambda_{\min}(C)$ be its maximum and minimum eigenvalues, respectively, and denote $\|C\|_\infty = \max_{ij} |c_{ij}|$ and $\|C\|_1 = \max_{1\le j\le n} \sum_{i=1}^m |c_{ij}|$. The convergence "$\to$" is understood as the one for $n \to \infty$. Finally for two sequences of numbers A_n and B_n, "$A_n \lesssim B_n$" means there exists some constant C that does not depend on n such that $A_n \le B_n C$ all n large enough.

1.2 Low dimensional case

In this section, we present asymptotic properties of our information theoretic estimator $\hat{\theta}$ for the low dimensional case, where the dimension K of function g in (1.1) grows slowly compared to the sample size n. In this case, computation of $\hat{\lambda}$ in (1.11) does not involve the ℓ_1-penalization. We first impose the following conditions.

Condition D.

1. *$\{X_i, Y_i\}_{i=1}^n$ is a strictly stationary and ergodic triangular array, and $\{X_i\}_{i=1}^n$ is α-mixing with mixing coefficients $\{\alpha_{X,m}\}$ satisfying $\sum_{m=1}^n \alpha_{X,m}^{1/2-1/q} \lesssim 1$ for some $q > 2$.*

2. *The support $\mathcal{X} \subseteq \mathbb{R}^p$ of X is a Cartesian product of p convex intervals with nonempty interiors. $\{\mathcal{X}_n\}$ is an increasing sequence of compact, convex,*

and nonempty subsets of $\mathcal{X}$, and satisfies $\mathbb{P}\{X \notin \mathcal{X}_n\} = o(n^{-1})$.

3. *ω_0 exists and is a continuous function bounded from above and away from zero with $\mathbb{E}[\omega_0(X)^2] < \infty$. h is a scalar-valued continuous function with $\mathbb{E}[h(X,Y)^2] < \infty$. There exists $\tilde{\mathbb{Q}}$ such that its Radon-Nikodym derivative satisfies $\varrho(x) = \frac{d\tilde{\mathbb{Q}}}{d\mathbb{P}}(x) > 0$ for almost every x, $\mathbb{E}[\varrho(X)^2] < \infty$, and $\mathbb{E}[g(X)\varrho(X)] = \mathbb{E}[r(X)]$.*

Condition D contains standard assumptions on the data and functions in (1.1) and (1.2). Condition D(1) allows data to be weakly dependent, which covers independent and identically distributed (iid) data as a special case.[3] Condition D(2) is on the support of X and the trimming set $\mathcal{X}_n$. For example, the condition $\mathbb{P}\{X \notin \mathcal{X}_n\} = o(n^{-1})$ is satisfied with $\mathcal{X}_n = \{x \in \mathbb{R}^d : |x| \leq n^{1/a}\}$ for $a \in (0, a_1)$ with $\mathbb{E}[|X|^{a_1}] < \infty$. Condition D(3) is on the functions ω_0 in (1.8) and h in (1.2), and constraint qualifications to guarantee the duality results in (1.6) and (1.7). If the underlying model that implies (1.1) uniquely identifies ω as $K \to \infty$ (as in Example 1.2), ω_0 is considered as this identified ω. If the underlying model that implies (1.1) partially or set identifies ω even when $K \to \infty$ (as in Example 1.1), ω_0 is considered as a particular element in the identified set of ω defined by the limit of the information projection in (1.8). To simplify the presentation, we focus on the case where h (and thus θ_0) is scalar. An extension to the case of vector θ_0 is straightforward. It is also possible to extend our method to the case where θ_0 is implicitly defined as a solution of moment conditions $\mathbb{E}[h(Z, \theta_0, \omega_0(X))] = 0$ for $Z = (Y, X')'$ and a linear map h (in ω_0).

Let $g_n(X) = \mathbb{E}[g(X)g(X)'\mathbb{I}\{X \in \mathcal{X}_n\}]^{-1/2} g(X)\mathbb{I}\{X \in \mathcal{X}_n\}$ be the orthonormalized version of g after trimming. We impose the following assumptions.

Condition S.

1. *All eigenvalues of $\mathbb{E}[g(X)g(X)'\mathbb{I}\{X \in \mathcal{X}_n\}]$ are strictly positive for each n, and $|\mathbb{E}_n[g_n(X)g_n(X)'] - I| = o_p(1)$.*

2. *There exists some $\lambda_b \in \mathbb{R}^K$ such that*

$$\sup_{x \in \mathcal{X}_n} |[\phi_*^{(1)}]^{-1}(\omega_0(x)) - \lambda_b' g_n(x)| \lesssim \eta_{K,n}, \tag{1.12}$$

$$\sqrt{\mathbb{E}[\{\omega_0(X) - \phi_*^{(1)}(\lambda_b' g_n(X))\}^2]} \lesssim \varsigma_{K,n}, \tag{1.13}$$

for some $\eta_{K,n} \to 0$ and $\varsigma_{K,n} \to 0$.

[3] Also it is interesting to extend our approach to introduce some blocking scheme for efficiency gain as in Kitamura and Stutzer (1997).

Condition S lists requirements for the functions g and g_n. Condition S(1) contains eigenvalue conditions on $\mathbb{E}[g(X)g(X)'\mathbb{I}\{X \in \mathcal{X}_n\}]$ to guarantee existence of g_n, and the convergence of the matrix $\mathbb{E}_n[g_n(X)g_n(X)']$. This convergence is satisfied if $\{X_i\}_{i=1}^n$ is iid and $\zeta_{K,n}^2 \log K \to 0$, where $\zeta_{K,n} = \sup_{x\in\mathcal{X}} |g_n(x)|$ (see, Lemma A.3 (i)). This can be satisfied for dependent data as well. For example, by Chen and Christensen (2015, Lemma 2.2), if $\{X_i\}_{i=1}^n$ is stationary and β-mixing with mixing coefficients $\{\beta_m\}$ and is such that $\beta_m n/m \to 0$ for some integer $m \leq n/2$, then $|\mathbb{E}_n[g_n(X)g_n(X)'] - I| = O_p(\sqrt{m\zeta_{K,n}^2 \log K/n})$ provided $m\zeta_{K,n}^2 \log K/n \to 0$.[4] Condition S(2) imposes assumptions on series approximations by g_n for $[\phi_*^{(1)}]^{-1}(\omega_0)$. The orders of the approximation errors $\eta_{K,n}$ and $\varsigma_{K,n}$ depend on the choices of the basis functions g, trimming set $\mathcal{X}_n$, and smoothness of $[\phi_*^{(1)}]^{-1}(\omega_0(\cdot))$. It can be verified by using results from functional analysis literature (e.g., Lorentz, 1966; Schumaker, 2007). For example, if $\mathcal{X} = [0,1]$ and g is a vector of polynomials or splines, then we can set as $\mathcal{X}_n = \mathcal{X}$, and $\eta_{K,n}$ is of order $O(K^{-s/p})$, where s is the number of continuous derivatives of $[\phi_*^{(1)}]^{-1}(\omega_0(\cdot))$ and p is dimension of X, and $\varsigma_{K,n} \leq \eta_{K,n}$. Note different choice of ϕ_* results in different economic modeling of target object ω_0. For example, if $\phi(x) = \frac{1}{2}x^2$, $[\phi_*^{(1)}]^{-1}(\omega_0(\cdot)) = \omega_0(\cdot)$, it implies $\omega_0(\cdot)$ is approximately linear. On the other hand, for Kullback-Leibler divergence $\phi(x) = x\log x$, $[\phi_*^{(1)}]^{-1}(\omega_0(\cdot)) = \log\omega_0(\cdot) + 1$. It implies that target $\omega_0(\cdot)$ is log linear, which is consistent with many financial models (for example, Vasicek, 1977) and might be more attractive for financial applications.

Let $r_n(X) = \mathbb{E}[g(X)g(X)'\mathbb{I}\{X \in \mathcal{X}_n\}]^{-1/2} r(X)\mathbb{I}\{X \in \mathcal{X}_n\}$ and

$$
\begin{aligned}
M_{K,n} &= \max_{1\leq j\leq K} \{\mathbb{E}[|g_{nj}(X)|^q]\}^{1/q} \vee \{\mathbb{E}[|r_{nj}(X)|^q]\}^{1/q} \quad \text{for } q \text{ in Condition D(1)},\\
\tilde{\varsigma}_{K,n} &= \sqrt{\frac{1}{n}\left(\varsigma_{K,n}^2 + \varsigma_{K,n}^{1+2/q}\sum_{m=1}^{n} \alpha_{X,m}^{1/2-1/q}\right)},\\
B_{K,n} &= \varsigma_{K,n} + \sqrt{\tilde{\varsigma}_{K,n}}, \qquad \mu_{K,n} = 1 + M_{K,n}\sum_{m=1}^{n} \alpha_{X,m}^{1/2-1/q}.
\end{aligned}
$$

We impose the following assumptions for the convex conjugate function ϕ_*.

Condition I. *$\phi_* : (0,\infty) \to \mathbb{R}$ is strictly convex and three times continuously differentiable. Also, (i) the second derivative $\phi_*^{(2)}$ is bounded from above and away*

[4] Thus object $\zeta_{K,n}$ is important for our analysis. In general, $\zeta_{K,n} \leq \lambda_{\min}\{\mathbb{E}[g(X)g(X)'\mathbb{I}\{X \in \mathcal{X}_n\}]\}^{-1/2} \sup_{x\in\mathcal{X}_n} |g(x)|$. So impact of diminishing eigenvalues is captured by the growth rate of $\lambda_{\min}\{\mathbb{E}[g(X)g(X)'\mathbb{I}\{X \in \mathcal{X}_n\}]\}^{-1/2}$. If $\mathcal{X}$ is compact and rectangular, usually $\lambda_{\min}\{\mathbb{E}[g(X)g(X)'\mathbb{I}\{X \in \mathcal{X}_n\}]\}^{-1/2} \lesssim 1$ (Chen and Christensen, 2015), and if in addition spline or wavelet series are used, $\zeta_{K,n} = O(\sqrt{K})$; and for power series, $\zeta_{K,n} = O(K)$ (Newey, 1997).

from zero, or (ii) $\zeta_{K,n}(\sqrt{K\mu_{K,n}/n} + B_{K,n}) \lesssim 1$.

Three times continuous differentiability of ϕ_* excludes some popular choices. For example, if $\phi(x) = -\log x$ is the reverse Kullback-Leibler divergence (which corresponds to empirical likelihood), then its convex conjugate is $\phi_*(y) = -1 - \log(-y)$ having a discontinuity point at $y = 0$, and this choice is ruled out by Condition I.[5] On the other hand, Kullback-Leibler divergence $\phi(x) = x \log x$ (corresponding to the exponential tilting) and half Pearson's χ^2 divergence $\phi(x) = \frac{1}{2}(x-1)^2$ (corresponding to the continuous updating GMM) satisfy this condition.

Let $\hat{\omega}(x) = \phi_*^{(1)}(\hat{\lambda}' g(x) \mathbb{I}\{x \in \mathcal{X}_n\})$. Based on the above conditions, the convergence rates of $\hat{\omega}(\cdot)$ and consistency of the estimator $\hat{\theta}$ in (1.10) are obtained as follows.

Theorem 1.1. *Suppose that Conditions D, S, and I hold true,* $K\mu_{K,n}/n \to 0$, *and* $B_{K,n} \to 0$. *Then*

$$\sqrt{\mathbb{E}_n[\{\hat{\omega}(X) - \omega_0(X)\}^2]} = O_p(\sqrt{K\mu_{K,n}/n} + B_{K,n}), \tag{1.14}$$

and $\hat{\theta} \xrightarrow{p} \theta_0$. *If we additionally assume* $\zeta_{K,n}\sqrt{K\mu_{K,n}/n} \to 0$ *and* $\zeta_{K,n} B_{K,n} \to 0$, *then*

$$\sup_{x \in \mathcal{X}_n} |\hat{\omega}(x) - \omega_0(x)| = O_p(\zeta_{K,n}\sqrt{K\mu_{K,n}/n} + \zeta_{K,n} B_{K,n} + \eta_{K,n}). \tag{1.15}$$

The consistency of $\hat{\theta}$ is established by showing that of $\hat{\omega}$ under the empirical L_2-norm in (1.14). As a byproduct of the proof of (1.14), we can obtain (1.15), an upper bound of the uniform convergence rate of $\hat{\omega}$ over the trimming set $\mathcal{X}_n$.[6] Interestingly, although our setup is different from standard nonparametric series estimation and ω_0 is not a conditional expectation function, we achieve a similar convergence rate with conventional series estimators for regression models. Indeed, our proof is in line with series estimation methods, where the estimation error of $\hat{\omega}$ can be decomposed into two parts: approximation bias (corresponding to $B_{K,n}$) and sampling error (corresponding to $\sqrt{K\mu_{K,n}/n}$). The approximation error is dealt with Lemma A.2 while the sampling error is controlled by Lemma A.3. In particular, $\mu_{K,n}$ characterizes a slowdown of the convergence rate for

[5] Intuitively, for the case of EL, we are approximating ω_0 using form $\phi_*^{(1)}(\lambda' g(x)) = -\frac{1}{\lambda' g(x)}$. Therefore uniform approximation might fail since a small change of $\lambda' g(x)$ around point 0 can lead to infinitely large variations in terms of approximation quality.

[6] This sup norm rate is admittedly not optimal. But for asymptotic normality of our estimator, such sup norm rate is not needed. It is also an open question whether we can improve the convergence rate in (1.14) to establish the optimality rate as in Belloni et al. (2015); Chen and Christensen (2015). Since our estimator $\hat{\omega}$ and target ω_0 are more complicated than the least squares estimator for the conditional mean, such analysis will be technically more involving.

the sampling error due to weak dependence of the data. If the data is iid, then $\mu_{K,n} = 1$, and the sampling error is of order $\sqrt{K/n}$. On the other hand, $\tilde{\varsigma}_{K,n}$ is an additional term due to weak dependence in the approximation bias $B_{K,n}$. If the data is iid and $\sqrt{n}\varsigma_{K,n} \to \infty$, then the bias term becomes a familiar expression $B_{K,n} = \varsigma_{K,n}$.

We next consider the limiting distribution of our estimator $\hat{\theta}$. To this end, we add the following conditions.

Condition N.

1. *There exists a function* $r^h : \mathcal{X} \to \mathbb{R}$ *such that* $\mathbb{E}[r^h(X)] = \mathbb{E}[\omega_0(X)\mathbb{E}[h(X,Y)|X]]$ *and*

$$\mathbb{E}[\beta'\{\omega_0(X)g_n(X) - r_n(X)\} - \{\omega_0(X)\mathbb{E}[h(X,Y)|X] - r^h(X)\}]^2 = o(n^{-1}), \tag{1.16}$$

   where $\beta = \mathbb{E}[\phi_*^{(2)}(\lambda_b' g_n(X))g_n(X)g_n'(X)]^{-1}\mathbb{E}[\phi_*^{(2)}(\lambda_b' g_n(X))g_n(X)h(X,Y)]$.

2. $|\mathbb{E}_n[\phi_*^{(2)}(\lambda_b' g_n(X))g_n(X)g_n(X)'] - \mathbb{E}[\phi_*^{(2)}(\lambda_b' g_n(X))g_n(X)g_n(X)']| = O_p(\Gamma_{K,n})$ *for some* $\Gamma_{K,n} \to 0$.

3. $\mathbb{E}[h^2(X,Y)|X = \cdot]$ *is bounded from above,* $\mathbb{E}[|h(X,Y)|^{q_1/(1-q_1/q)}] < \infty$ *for some* $q_1 \in (2, q]$ *and* $\mathbb{E}[|r^h(X)|^q] < \infty$.

4. $\{Y_i, X_i\}_{i=1}^n$ *is* α*-mixing with mixing coefficients* $\{\alpha_{XY,m}\}_{m\in\mathbb{N}}$ *satisfying*

$$\sum_{m=1}^{n} \alpha_{XY,m}^{(a/(2+a))\vee(1/2-1/q_1)} \lesssim 1$$

   for some $a > 0$ *and* $\mathbb{E}[|\Phi|^{2+a}] < \infty$*, where*

$$\Phi = \omega_0(X)h(X,Y) - \theta_0 - \{\omega_0(X)\mathbb{E}[h(X,Y)|X] - r^h(X)\}. \tag{1.17}$$

Condition N(1) is considered as the mean square continuity condition (Assumption 5.3 in Newey, 1994b) in our setup, which guarantees the $\sqrt{n}$-consistency of $\hat{\theta}$ even though $\hat{\omega}$ converges at a slower rate. Basically, (1.16) requires that $\mathbb{E}[h(X,Y)|X = \cdot]$ is well approximated by the basis functions $g_n(\cdot)$. This requirement is typically verified by the results in functional analysis. The function r^h should be specified for each application. If $r(X)$ is known constants (as in Example 1.1), we can simply set as $r^h(X) = \theta_0$. For Example 1.2, we can set as $r^h(X) = E[Y^*|X]$. Proposition 1.1 below gives two examples, where (1.16) is satisfied. Condition N(2) is analogous to Condition S(1). The convergence rate $\Gamma_{K,n}$ can be established as $\sqrt{\zeta_{K,n}^2 \log K/n}$ for the iid case (by Lemma A.3

(i)), and $\sqrt{m\zeta_{K,n}^2 \log K/n}$ for the β-mixing case (by adapting Lemma 2.2 of Chen and Christensen, 2015). Condition N(3) contains mild assumptions on h and r^h. Condition N(4) requires α-mixing for the joint $\{X_i, Y_i\}_{i=1}^n$ to apply a central limit theorem to $\frac{1}{\sqrt{n}}\sum_{i=1}^n \Phi_i$, where Φ_i is the influence function for $\hat{\theta}$.

By imposing this additional condition, the limiting distribution of the estimator $\hat{\theta}$ is obtained as follows.

Theorem 1.2. *Suppose that the conditions of Theorem 1.1 and Condition N hold true. In addition,* $\zeta_{K,n}^4 K\mu_{K,n}/\sqrt{n} \to 0$, $\sqrt{n}\zeta_{K,n}B_{K,n} \to 0$, *and* $\sqrt{K\mu_{K,n}}\zeta_{K,n}\Gamma_{K,n} \to 0$. *Then*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V),$$

where $V = \lim_{n\to\infty} Var(\sqrt{n}\mathbb{E}_n[\Phi])$.

This theorem says that our information theoretic estimator $\hat{\theta}$ is $\sqrt{n}$-consistent and asymptotically normal. If the data is iid, the variance term becomes $\mathbb{E}[\Phi^2]$, which is also the semiparametric efficiency bound.

The asymptotic variance V can be estimated by some heteroskedasticity autocorrelation consistent estimator. For example, based on Newey and West (1987), V can be estimated by

$$\hat{V} = \hat{\gamma}_0 + 2\sum_{l=1}^{M_n}\left(\frac{M_n - l}{M_n}\right)\hat{\gamma}_l,$$

where $\hat{\gamma}_l = (n-l)^{-1}\sum_{i=l+1}^n(\hat{\Phi}_i - n^{-1}\sum_{i=1}^n \hat{\Phi}_i)(\hat{\Phi}_{i-l} - n^{-1}\sum_{i=1}^n \hat{\Phi}_i)$ is the sample autocovariance of

$$\hat{\Phi}_i = \mathbb{I}\{X_i \in \mathcal{X}_n\}[\hat{\omega}(X_i)h(X_i, Y_i) - \hat{\theta} - \{\hat{\omega}(X_i)\hat{h}^X(X_i) - \hat{r}^h(X_i)\}],$$

$\hat{h}^X$ and $\hat{r}^h$ are some nonparametric estimators of $\mathbb{E}[h(X,Y)|X = \cdot]$ and r^h, respectively, and M_n is a tuning parameter. By adapting the proof of Newey and West (1987, Theorem 2) to the present context, the consistency of $\hat{V}$ is obtained as follows.

Proposition 1. Suppose that the conditions of Theorem 1.2 hold true. Additionally, assume that $E[|\Phi_i|^{4q_2+\delta}] < \infty$ for some $q_2 > 1$ and $\delta > 0$, $\sum_{m=1}^n \alpha_{XY,m}^{1-1/(2q_2)} \lesssim 1$, $\sup_{x\in\mathcal{X}_n}|\hat{h}^X(x) - \mathbb{E}[h(X,Y)|X = x]| = O_p(R_n)$ and $\sup_{x\in\mathcal{X}_n}|\hat{r}^h(x) - r^h(x)| = O_p(R_n)$ for $R_n = \zeta_{K,n}\sqrt{K\mu_{K,n}/n} + \zeta_{K,n}B_{K,n} + \eta_{K,n}$, $M_n \to \infty$, and $M_n R_n \to 0$. Then $\hat{V} \xrightarrow{p} V$.

Compared to Theorem 1.1, we impose more stringent conditions on K. For example, in an iid setting, suppose $\mathcal{X} = [0,1]^d$, $\mathcal{X}_n = \mathcal{X}$, and g is a vector of

B-spline basis functions with $K \propto n^a$ for $a > 0$. Then it holds $\zeta_K = O(\sqrt{K})$. If ω_0 lies in some Hölder function class, we can typically assume $\varsigma_{K,n} = K^{-a_1}$, where a_1 depends on the smoothness of ω_0 and dimension of X. Then the conditions required for K are satisfied with $a \in (1/2(a_1 - 1), 1/6)$ and $a_1 > 4$. Therefore, for Theorem 1.2, K should grow at a sufficiently slow rate and ω_0 should be sufficiently smooth. Also note that the condition $\zeta_{K,n}^4 K \mu_{K,n}/\sqrt{n} \to 0$ is in fact the worst case scenario due to generality of the divergence function ϕ, and can be weakened if some choices of ϕ, such as Pearson's χ^2 divergence.

We close this section by providing some specific examples that satisfy Condition N(1).

Proposition 1.1. *Suppose the assumptions in Theorem 1.2 except for (1.16) hold true.*

(i) Suppose $r(X)$ is a vector of known constants, $\mathbb{P}\{X \notin \mathcal{X}_n\} = o((Kn)^{-1})$ and

$$\mathbb{E}[\{\mathbb{E}[h(X,Y)|X] - \lambda' g_n(X)\}^2] = o(n^{-1}), \tag{1.18}$$

for some $\lambda \in \mathbb{R}^K$. Then (1.16) is satisfied with $r^h(X) = \theta_0$.

(ii) In Example 1.2 on missing data, suppose

$$\mathbb{E}[\{\mathbb{E}[Y^*|X] - \lambda' g_n(x)\}^2] = o(1),$$

for some $\lambda \in \mathbb{R}^K$. Then (1.16) is satisfied with $r^h(X) = \mathbb{E}[Y^*|X]$.

Based on Proposition 1.1, if $r(X)$ is a vector of known constants, the influence function Φ simplifies to $\Phi = \omega_0(X)\{h(X,Y) - \mathbb{E}[h(X,Y)|X]\}$.

1.3 High dimensional case

In this section, we consider the high dimensional case, where the dimension K of the moment functions g can be larger and grow faster than the sample size n. In this case, $\hat{\lambda}$ in (1.11) is computed by the ℓ_1-penalization. High dimensionality of g can be caused by either high dimensionality of the original data X or many transformations (or basis functions) based on low dimensional X. In either case, as far as the latent weight function ω_0 in (1.8) admits certain sparse representation, our penalized estimator can consistently estimate ω_0 and the parameter of interest θ_0. In Section 1.3.1, we study asymptotic properties of $\hat{\omega}$ to estimate ω_0. Then we consider three estimation approaches, debiasing (Section 1.3.2), post selection (Section 1.3.3), and targeted debiasing (Section 1.3.4), and present conditions to achieve $\sqrt{n}$-normality for the estimators of θ_0.

1.3.1 Estimation of ω_0

We first present asymptotic properties of $\hat{\omega}$. For the high dimensional case, we impose the following assumptions on the data.

Condition D'. *$\{X_i, Y_i\}_{i=1}^n$ is an iid triangular array. The support $\mathcal{X} \subseteq \mathbb{R}^p$ of X is a Cartesian product of p convex intervals with nonempty interiors. Condition D(3) holds true.*

For the high dimensional case, we focus on the case of iid data. An extension to dependent data involves development of empirical process theory for dependent data, and is beyond the scope of this chapter. We also do not use trimming. Impact from possible unbounded support is dealt implicitly by the growth rate of $\sup_{x\in\mathcal{X}}\|g(x)\|_\infty$ and a mild sup-norm approximation assumption over $\mathcal{X}$ (see the statement in Theorem 1.3).

To state additional conditions for the high dimensional case, we introduce further notation. For an index subset $S \subset \{1, \ldots, K\}$, let $|S|$ be its cardinality (with slight abuse of notation), $\lambda_S = (\lambda_{1,S}, \ldots, \lambda_{K,S})'$ be a K dimensional vector with $\lambda_{j,S} = \lambda_j \mathbb{I}\{j \in S\}$ for the j-th component λ_j of λ, and $\lambda_{S^c} = (\lambda_{1,S^c}, \ldots, \lambda_{K,S^c})'$ with $\lambda_{j,S^c} = \lambda_j \mathbb{I}\{j \notin S\}$. So, λ_S and λ_{S^c} have non-zero elements only in the index set S and its complement S^c, respectively. Furthermore, let $\mathscr{S}$ be a class of index sets.[7] We introduce the so-called compatibility condition.

Condition C. *For each $S \in \mathscr{S}$, there exists some constant $\phi_S > 0$ such that for all λ satisfying $\|\lambda_{S^c}\|_1 \leq 3\|\lambda_S\|_1$, it holds $\|\lambda_S\|_1 \leq \phi_S^{-1}\sqrt{\lambda'\mathbb{E}[g(X)g(X)']\lambda}\sqrt{|S|}$.*

This is a high level condition that bounds $\|\lambda_S\|_1$ by the L_2-norm of its corresponding function $\lambda' g(\cdot)$. Such compatibility condition is commonly employed in the high dimensional statistics literature, such as the restricted eigenvalue condition in Bickel et al. (2009). Let $\lambda_* = \arg\min_\lambda \mathbb{E}[\phi_*(\lambda' g(X)) - \lambda' r(X)]$ and

$$\mathscr{E}(\lambda) = \mathbb{E}[\phi_*(\lambda' g(X)) - \lambda' r(X)] - \mathbb{E}[\phi_*(\lambda_*' g(X)) - \lambda_*' r(X)],$$

be the excess risk. Given $\mathscr{S}$ with associated compatibility constants $\{\phi_S : S \in \mathscr{S}\}$ in Condition C, the oracle $\lambda_{\mathbf{o}}$ achieves the best sparse approximation of $\mathscr{E}(\lambda)$ as

$$\lambda_{\mathbf{o}} = \arg\min_{\lambda: S_\lambda \in \mathscr{S}} 2\mathscr{E}(\lambda) + \frac{8\alpha_n^2}{\phi_{S_\lambda}^2 \varrho}|S_\lambda|, \tag{1.19}$$

[7] Knowledge of $\mathscr{S}$ can reflect researchers' priors on what might be important sets of covariates. In the worst case, without any prior knowledge, $\mathscr{S}$ should contain all possible index sets that are subsets of full set of covariates.

where $S_\lambda = \{j : \lambda_j \neq 0\}$, α_n is a penalty level in (1.11), and ϱ is a constant defined in Condition H below. Let $Q_{\mathbf{o}}$ be the minimized value of (1.19) and $\omega_{\mathbf{o}}(x) = \phi_*^{(1)}(\lambda_{\mathbf{o}}' g(x))$. Note that $\mathscr{E}(\lambda_{\mathbf{o}}) \geq \mathscr{E}(\lambda_*) = 0$ and a part of our sparsity assumption is characterized by the convergence rate of $\mathscr{E}(\lambda_{\mathbf{o}})$ toward zero. Let

$$\nu_n(\lambda) = \mathbb{E}_n[\phi_*(\lambda' g(X)) - \lambda' r(X)] - \mathbb{E}[\phi_*(\lambda' g(X)) - \lambda' r(X)],$$

be an empirical process. We impose the following assumptions.

Condition H. *For every $\varepsilon > 0$ small enough and n large enough, there exist positive constants $\sigma_{\varepsilon,n}$, ϱ, and A such that for $M = \frac{Q_{\mathbf{o}}}{2\sigma_{\varepsilon,n}}$,*

1. $\mathbb{P}\left\{\sup_{\left\|\lambda-\lambda_{\mathbf{o}}\right\|_1 \leq M} |\nu_n(\lambda) - \nu_n(\lambda_{\mathbf{o}})| \leq \sigma_{\varepsilon,n} M\right\} \geq 1 - \varepsilon,$
2. *for any λ satisfying $\left\|\lambda - \lambda_{\mathbf{o}}\right\|_1 \leq M$, it holds*

$$\sup_{x \in \mathcal{X}} |(\lambda - \lambda_{\mathbf{o}})' g(x)| \leq A, \quad \varrho(\lambda - \lambda_{\mathbf{o}})' \mathbb{E}[g(X)g(X)'](\lambda - \lambda_{\mathbf{o}}) \leq \mathscr{E}(\lambda),$$

3. $\sigma_{\varepsilon,n} \leq \alpha_n/8$ *and* $\alpha_n \propto \sqrt{\log K/n}$ *for all* $n \in \mathbb{N}$.

Condition H(1) controls the empirical process $\nu_n(\lambda)$ in a neighborhood of the oracle $\lambda_{\mathbf{o}}$. Intuitively, we require that $\nu_n(\lambda) - \nu_n(\lambda_{\mathbf{o}})$ will be small when λ is close to $\lambda_{\mathbf{o}}$ in terms of the l_1-norm. The order of $\sigma_{\varepsilon,n}$, which is typically $O(\sqrt{\log K/n})$, can be derived by empirical process theory.[8] By Condition H(2), the excess risk $\mathscr{E}(\lambda)$ can be bounded from below by a quadratic function of λ when λ is close to $\lambda_{\mathbf{o}}$ in terms of the l_1-norm. Condition H(3) is on the penalty coefficient α_n. First, α_n should be large enough to offset the effect from $\sigma_{\varepsilon,n}$. Second, since $\sigma_{\varepsilon,n}$ is typically of order $O(\sqrt{\log K/n})$, α_n should be the same order to achieve the fastest convergence. [9]

Under these conditions, the convergence rate of $\hat{\omega}$ and consistency of the parameter estimator $\hat{\theta}$ are established as follows. Let $\tilde{\zeta}_K = \sup_{x \in \mathcal{X}} \left\|g(x)\right\|_\infty$, $s = |S_{\lambda_{\mathbf{o}}}|$, $\kappa_{\mathbf{o},n} = \mathscr{E}(\lambda_{\mathbf{o}}) \sqrt{\frac{n}{\log K}} \vee s\sqrt{\frac{\log K}{n}}$, and $\{\xi_n\}$ and $\{\varsigma_{\mathbf{o},n}\}$ be positive sequences such that $\left\|\mathbb{E}_n[g(X)g(X)'\right\|_\infty = O_p(\xi_n)$ and $\sqrt{\mathbb{E}[\{\omega_{\mathbf{o}}(X) - \omega_0(X)\}^2]} \lesssim \varsigma_{\mathbf{o},n}$, respectively.

[8] Since our objective function is Lipschitz in a neighborhood of $\lambda_{\mathbf{o}}$, probabilistic inequalities, such as Bühlmann and Van De Geer (2011, Lemma 14.20), can be applied.

[9] There are in general two ways to select α_n in a data driven way: First, α_n may be chosen by cross validation although it might lack theoretical justification; Second, α_n can also be chosen as the smallest value such that Condition H stands with a large probability. That is, we can set $\alpha_n = 8\hat{\sigma}_{\varepsilon,n}$, where $\hat{\sigma}_{\varepsilon,n}$ is an estimate of $\sigma_{\varepsilon,n}$, making use of empirical process and moderate deviation theory. See Belloni et al. (2012) for further details on this.

Theorem 1.3. *Suppose Conditions D', C, and H hold true, and Condition I holds true with I(ii) replaced by* $\tilde{\zeta}_K \kappa_{\mathbf{o},n} \lesssim 1$. *Furthermore, assume that* $\varsigma_{\mathbf{o},n} \to 0$, $\kappa_{\mathbf{o},n} \xi_n^{1/2} \to 0$, *and* $\sup_{x \in \mathcal{X}} |\omega_{\mathbf{o}}(x) - \omega_0(x)| \lesssim 1$. *Then*

$$\sqrt{\mathbb{E}_n[\{\hat{\omega}(X) - \omega_0(X)\}^2]} = O_p(\kappa_{\mathbf{o}n} \sqrt{\xi_n} + \varsigma_{\mathbf{o},n}), \tag{1.20}$$

and $\hat{\theta} \xrightarrow{p} \theta_0$. *If we additionally assume* $\tilde{\zeta}_K \kappa_{\mathbf{o},n} \to 0$ *and* $\sup_{x \in \mathcal{X}} |\omega_{\mathbf{o}}(x) - \omega_0(x)| \to 0$, *then*

$$\sup_{x \in \mathcal{X}} |\hat{\omega}(x) - \omega_0(x)| \xrightarrow{p} 0. \tag{1.21}$$

This theorem, a counterpart of Theorem 1.1 for the high dimensional case, establishes the empirical L_2 convergence rate of $\hat{\omega}$, which is required for consistency of $\hat{\theta}$. Note that we only require the boundedness of the uniform approximation error $\sup_{x \in \mathcal{X}} |\omega_{\mathbf{o}}(x) - \omega_0(x)|$ by the oracle. Furthermore, this uniform approximation condition can be dropped if Condition I(ii) is satisfied. The object $\tilde{\zeta}_K$ depends on the choice of basis functions g and $\mathcal{X}$. For example, if g is a vector of polynomials over $\mathcal{X} = [0,1]^p$, it holds $\tilde{\zeta}_K = O(1)$. The object ξ_n measures the growth rate of the sup-norm of $\mathbb{E}_n[g(X)g(X)']$. It can be controlled effectively by Hoeffding's inequality, and is typically of order $O(\|\mathbb{E}[g(X)g(X)']\|_\infty)$ (or could be $O(1)$ for certain basis functions). In this case, if we further assume $\mathscr{E}(\lambda_{\mathbf{o}}) = O(s \log K/n)$ and $\varsigma_{\mathbf{o},n} = O(s\sqrt{\log K/n})$, then the empirical L_2 convergence rate of $\hat{\omega}$ is of order $O_p(s\sqrt{\log K/n})$ and the dimension K may grow faster than n even at an exponential rate. For the high dimensional case, the approximation bias for ω_0 tends to be larger and is controlled by the approximate sparsity assumption that requires sufficiently fast decays of the excess risk $\mathscr{E}(\lambda_{\mathbf{o}})$ and approximation error $\varsigma_{\mathbf{o},n}$. On the other hand, the sampling error of the estimator is controlled by Condition C. A byproduct of this theorem is the uniform consistency in (1.21) under additional assumptions.[10]

1.3.2 Debiased estimator for θ_0

In this subsection, we consider a debiased estimation method for θ_0 in the high dimensional setup. It is well known that plug-in methods to estimate finite dimensional objects, where the first step is implemented by the lasso, typically cannot achieve the $\sqrt{n}$-normality. There is a recent literature in statistics (e.g., Zhang and Zhang 2014b; Van de Geer et al. 2014) that develop procedures to

[10]For the high dimensional case in (1.11), alternatively setting first term as $\log \mathbb{E}_n[\phi_*(\lambda' g(X) - \lambda' r(X)]$ might not necessarily work as our theory for high dimensional part relies on the first term being strictly convex and showing quadratic behavior. With log transformation, such requirement might fail. Even if convexity does not fail, Condition H will become increasingly more difficult to analyze after log transformations. Hence, we do not recommend taking such transformations for high dimensional calculations in practice.

debias the lasso estimators to achieve the $\sqrt{n}$-normality for finite dimensional objects of interest. It is natural to ask whether such debias procedures may be applied to our setup. However, in our setting, it seems the debiasing procedure achieves $\sqrt{n}$-normality to estimate θ_0 only under some stringent conditions.

To illustrate this point, suppose $\phi_*^{(2)}(\cdot) = c_* > 0$ for some known constant c_* (for example, by choosing $\phi(x) = \frac{1}{2}x^2$). Let $\hat{\kappa} = (\text{sign}(\hat{\lambda}_1), \ldots, \text{sign}(\hat{\lambda}_K))'$ and $\hat{\Theta}$ be an approximation of the 'inverse' of $\mathbb{E}_n[g(X)g(X)']$ (which may not exist in the high dimensional case). Here we consider the debiased estimator

$$\hat{\theta}_{DB} = \mathbb{E}_n[\{\phi_*^{(1)}(\hat{\lambda}'g(X)) + \alpha_n g(X)'\hat{\Theta}\hat{\kappa}\}h(X,Y)],$$

where the additional term $\alpha_n g(\cdot)'\hat{\Theta}\hat{\kappa}$ corrects the first-order bias from the plug-in estimation by $\hat{\lambda}$. We note that this additional term will be different if we drop the requirement $\phi_*^{(2)}(\cdot) = c_* > 0$. To establish the $\sqrt{n}$-normality of $\hat{\theta}_{DB}$, we impose the following assumptions. Let $\hat{\beta}_{DB} = \hat{\Theta}'\mathbb{E}_n[g(X)h(X,Y)]$.

Condition DB.

1. *There exist functions* $r^h, \tilde{r}^h, \tilde{h}^X : \mathcal{X} \to \mathbb{R}$ *such that* $\mathbb{E}[r^h(X)] = \mathbb{E}[\omega_0(X)\mathbb{E}[h(X,Y)|X]]$, $\mathbb{E}[\tilde{r}^h(X)] = \mathbb{E}[\omega_0(X)\tilde{h}^X(X)]$, *and*

$$\begin{aligned}\mathbb{E}_n[\hat{\beta}_{DB}'\{\omega_0(X)g(X) - r(X)\} - \{\omega_0(X)\tilde{h}^X(X) - \tilde{r}^h(X)\}]^2 &= o_p(n^{-1}),\\ (\varsigma_n^2 + \varsigma_n n^{-1/2})\mathbb{E}_n[\tilde{h}^X(X) - \hat{\beta}_{DB}'g(X)]^2 &= o_p(n^{-1}).\end{aligned}$$

2. $\sqrt{n}\kappa_{\mathbf{o},n}\left\|\mathbb{E}_n[h(X,Y)g(X)]\right\|_\infty \left\|I - \mathbb{E}_n[g(X)g(X)']\hat{\Theta}\right\|_1 = o_p(1)$.

Condition DB highlights two key requirements for achieving $\sqrt{n}$-normality of the debiased estimator $\hat{\theta}_{DB}$. Condition DB(1) is a natural extension of Condition N(1) under the high dimensional case. It requires that $\hat{\beta}_{DB}'g(\cdot)$ should converge fast enough to some function $\tilde{h}^X(\cdot)$. Intuitively, $\tilde{h}^X(\cdot)$ can be understood as an approximation of $\mathbb{E}[h(X,Y)|X =\cdot\,]$. This is a key condition to correct the bias from the second step to compute $\hat{\theta}_{DB}$. On the other hand, Condition DB(2) controls the ℓ_1-regularization bias. It says the matrix $\hat{\Theta}$ should be selected to guarantee $\left\|I - \mathbb{E}_n[g(X)g(X)']\hat{\Theta}\right\|_1$ to be sufficiently small.

The $\sqrt{n}$-normality of the debiased estimator $\hat{\theta}_{DB}$ is obtained as follows. Let $\{\tau_n\}$ be a positive sequence such that

$$\sqrt{\mathbb{E}[\{\mathbb{E}[h(X,Y)|X] - \tilde{h}^X(X)\}^2]} \vee \sqrt{\mathbb{E}[\{r^h(X) - \tilde{r}^h(X)\}^2]} \lesssim \tau_n.$$

Theorem 1.4. *Suppose Conditions D', C, H, and DB hold true and* $\phi_*^{(2)}(\cdot) = c_* > 0$ *for some known constant* c_*. *If* $\sup_{x\in\mathcal{X}}\mathbb{E}[h(X,Y)^2|X = x] \lesssim 1$, $\varsigma_n \to 0$,

$\tau_n \to 0$, *and* $\sqrt{n}\varsigma_n\tau_n \to 0$, *then*

$$\sqrt{n}(\hat{\theta}_{DB} - \theta_0) \xrightarrow{d} N(0, \mathbb{E}[\Phi^2]).$$

Theorem 1.4 gives conditions under which the debiased estimator $\hat{\theta}_{DB}$ can achieve $\sqrt{n}$-normality. It seems the requirements on $\hat{\Theta}$ listed in Condition DB are difficult to avoid. In fact, our debiasing procedure may be considered as an in-between the parametric debiasing of Zhang and Zhang (2014b); Van de Geer et al. (2014), and the complete debiasing of Farrell (2015); Belloni et al. (2017b). It is beyond the scope of this chapter to study a practical way of finding the matrix $\hat{\Theta}$ (for example, by adapting the lasso with nodewise regression in Van de Geer et al., 2014), and we leave this for future research.

1.3.3 Post selection estimator for θ_0

Given that the debiasing procedure in the last subsection requires relatively strong conditions, we propose the following post selection method to obtain a $\sqrt{n}$-consistent estimator for θ_0.

1. Compute $\hat{\lambda}$ in (1.11) for the high dimensional case. Let $\mathbf{s} = |\hat{S}|$ be the cardinality of the selected set $\hat{S} = \{j : \hat{\lambda}_j \neq 0\}$.

2. Let $g_{\mathbf{s}}$ and $r_{\mathbf{s}}$ be the $\mathbf{s}$-dimensional functions corresponding to the selected set $\hat{S}$. Implement (1.11) for the low dimensional case (i.e., without the ℓ_1-penalty) based on $g_{\mathbf{s}}$ and $r_{\mathbf{s}}$. Denote the solution of this step as

$$\hat{\Lambda} = \arg\min_{\Lambda \in \mathbb{R}^{\mathbf{s}}} \mathbb{E}_n[\phi_*(\Lambda' g_{\mathbf{s}}(X)) - \Lambda' r_{\mathbf{s}}(X)]. \tag{1.22}$$

3. Construct the post selection estimator as

$$\tilde{\theta} = \mathbb{E}_n[\phi_*^{(1)}(\hat{\Lambda}' g_{\mathbf{s}}(X)) h(X, Y)].$$

To study asymptotic properties of the post selection estimator $\tilde{\theta}$, we introduce some notation. Let $\Lambda_* = \arg\min_{\Lambda \in \mathbb{R}^{\mathbf{s}}} \mathbb{E}[\phi_*(\Lambda' g_{\mathbf{s}}(X)) - \Lambda' r_{\mathbf{s}}(X)]$ be the population counterpart of (1.22), and $\omega_*(x) = \phi_*^{(1)}(\Lambda_*' g_{\mathbf{s}}(x))$, which is an approximation of ω_0 using the selected vector $g_{\mathbf{s}}$. Note that ω_* could be different from $\omega_{\mathbf{o}}$ selected by the oracle $\lambda_{\mathbf{o}}$. Also, define

$$\beta_{\mathbf{s}} = \mathbb{E}[\phi_*^{(2)}(\Lambda_*' g_{\mathbf{s}}(X)) g_{\mathbf{s}}(X) g_{\mathbf{s}}(X)']^{-1} \mathbb{E}[\phi_*^{(2)}(\Lambda_*' g_{\mathbf{s}}(X)) g_{\mathbf{s}}(X) \mathbb{E}[h(X, Y)|X]]$$

and $\tilde{h}^X(x) = \beta_{\mathbf{s}}' g_{\mathbf{s}}(x)$. We impose the following conditions.

Condition N'. *There exist functions* $r^h, \tilde{r}^h : \mathcal{X} \to \mathbb{R}$ *such that* $\mathbb{E}[r^h(X)] = \mathbb{E}[\omega_0(X)\mathbb{E}[h(X,Y)|X]]$, $\mathbb{E}[\tilde{r}^h(X)] = \mathbb{E}[\omega_0(X)\tilde{h}^X(X)]$, *and*

$$\mathbb{E}[\beta_{\mathbf{s}}'\{\omega_0(X)g_{\mathbf{s}i}(X) - r_{\mathbf{s}}(X)\} - \{\omega_0(X)\tilde{h}^X(X) - \tilde{r}^h(X)\}]^2 = o(1). \tag{1.23}$$

Condition N' can be viewed as an extension of the mean square continuity (as in Assumption 5.3 of Newey, 1994a) for imperfect model selection, where $\tilde{h}^X(\cdot) = \beta_{\mathbf{s}}'g_{\mathbf{s}}(\cdot)$ is understood as an approximation of $\mathbb{E}[h(X,Y)|X = \cdot]$ based on the selected basis functions $g_{\mathbf{s}}$. In the case of imperfect model selection (i.e., $\hat{S} \neq S_{\lambda_{\mathbf{o}}}$), ω_* and $\tilde{h}^X$ may not approximate ω_0 and h^X well enough, respectively. We impose the following conditions for those approximation errors.

Condition S'. *For each* n, *all eigenvalues of* $\mathbb{E}[g_{\mathbf{s}}(X)g_{\mathbf{s}}(X)']$ *are bounded from above and away from zero, conditional on the selected set* $\hat{S}$. *Also, for some positive sequences* $\{\varsigma_{\mathbf{s},n}\}$ *and* $\{\tau_{\mathbf{s},n}\}$,

$$\sqrt{\mathbb{E}[\{\omega_0(X) - \omega_*(X)\}^2]} \lesssim \varsigma_{\mathbf{s},n}, \tag{1.24}$$

$$\sqrt{\mathbb{E}[\{\mathbb{E}[h(X,Y)|X] - \tilde{h}^X(X)\}^2]} \lesssim \tau_{\mathbf{s},n}. \tag{1.25}$$

Because of the imperfect model selection, $\varsigma_{\mathbf{s},n}$ and $\tau_{\mathbf{s},n}$ may not vanish sufficiently fast as in Theorem 1.2. Instead, we only require $\varsigma_{\mathbf{s},n}$ and $\tau_{\mathbf{s},n}$ to be $O(1)$. Let $\zeta_{\mathbf{s}} = \sup_{x\in\mathcal{X}} |g_{\mathbf{s}}(x)|$.

Condition I'. $\phi_* : (0,\infty) \to \mathbb{R}$ *is strictly convex and three times continuously differentiable,* $\sup_{x\in\mathcal{X}} \phi_*^{(2)}(\Lambda_*'g_{\mathbf{s}}(x)) \lesssim 1$, *and* $\sup_{\Lambda\in\mathbb{R}^{\mathbf{s}}:|\Lambda-\Lambda_*|\lesssim\sqrt{\frac{\zeta_{\mathbf{s}}^2}{n}}} \mathbb{E}_n[\phi_*^{(3)}(\Lambda'g_{\mathbf{s}}(X))^2] = O_p(1)$.

Condition I' is a counterpart of Condition I, and imposes additional requirements on the conjugate function ϕ_*. They can be trivially satisfied for some divergence functions, such as $\phi(x) = \frac{1}{2}x^2$. This can also be satisfied if $\sup_{x\in\mathcal{X}} |[\phi_*^{(1)}]^{-1}(\omega_0(x)) - \Lambda_*'g_{\mathbf{s}}(x)| \lesssim 1$, i.e., the selected model $\Lambda_*'g_{\mathbf{s}}(\cdot)$ for $[\phi_*^{(1)}]^{-1}(\omega_0(\cdot))$ is not too far.

Under these conditions, the $\sqrt{n}$-normality of the post selection estimator $\tilde{\theta}$ is obtained as follows.

Theorem 1.5. *Suppose Conditions D', S', I', and N' hold true. In addition,* $\zeta_{\mathbf{s}}^2 \log \mathbf{s}/n \to 0$, $\zeta_{\mathbf{s}}^6/\sqrt{n} \to 0$, *and* $\mathbb{E}[(\Phi + v_1 + v_2 + v_3)^2] < \infty$, *where* Φ *is defined in (1.17). Then*

$$\sqrt{n}(\tilde{\theta} - \theta_0 + b) \xrightarrow{d} N(0, \mathbb{E}[(\Phi + v_1 + v_2 + v_3)^2]), \tag{1.26}$$

where $b = \mathbb{E}[(\omega_0(X) - \omega_*(X))(h^X(X) - \tilde{h}^X(X))]$,

$$\begin{aligned} v_1 &= (\omega_*(X) - \omega_0(X))(h(X,Y) - h^X(X)), \\ v_2 &= \omega_0(X)(h^X(X) - \tilde{h}^X(X)) + \tilde{r}^h(X) - r^h(X), \\ v_3 &= (\omega_*(X) - \omega_0(X))(h^X(X) - \tilde{h}^X(X)) - \mathbb{E}[(\omega_*(X) - \omega_0(X))(h^X(X) - \tilde{h}^X(X))]. \end{aligned}$$

Furthermore, if $\varsigma_{\mathbf{s},n} \to 0$, $\tau_{\mathbf{s},n} \to 0$, *and* $\sqrt{n}\varsigma_{\mathbf{s},n}\tau_{\mathbf{s},n} \to 0$, *then*

$$\sqrt{n}(\tilde{\theta} - \theta_0) \xrightarrow{d} N(0, \mathbb{E}[\Phi^2]). \tag{1.27}$$

This theorem characterizes the effects of the imperfect model selection from the first step lasso procedure. b is an additional bias term, and v_1, v_2, and v_3 are additional variance terms. In particular, v_1 is the additional variance due to imperfect approximation of ω_0 by ω_*, and v_2 is another variance term due to imperfect approximation of h^X by $\tilde{h}^X$, and v_3 is the term due to slow approximation of both h^X and ω_0. For the case of (1.27), we can conduct inference on θ_0 by estimating the asymptotic variance $\mathbb{E}[\Phi^2]$. On the other hand, if the imperfect model selection is severe in the sense that $\varsigma_{\mathbf{s},n} = \tau_{\mathbf{s},n} = O(1)$, the post selection estimator $\tilde{\theta}$ will have the asymptotic bias b and additional terms in the variance as in (1.26). We leave valid inference in this general case for future research.

1.3.4 Targeted debiasing estimator for θ_0

In this subsection, we discuss a targeted debiasing procedure, which is between the debiasing procedure for the whole vector $\hat{\lambda}$ in Section 1.3.2 and post selection procedure in Section 1.3.3.

Without loss of generality, we assume the first $\mathbf{s}$ elements of $\{1, \ldots, K\}$ are selected by $\hat{\lambda}$. Suppose that $\hat{\Theta}_{\mathbf{s}}$ is a good approximation of the inverse of the $\mathbf{s} \times \mathbf{s}$ matrix $\mathbb{E}[\phi_*^{(2)}(\lambda'_{\mathbf{os}} g_{\mathbf{s}}(X)) g_{\mathbf{s}}(X) g_{\mathbf{s}}(X)']$. For example, a practical choice would be the empirical counterpart $(\mathbb{E}_n[\phi_*^{(2)}(\hat{\lambda}'_{\mathbf{s}} g_{\mathbf{s}}(X)) g_{\mathbf{s}}(X) g_{\mathbf{s}}(X)'])^{-1}$. Define the targeted debiasing version $\hat{\lambda}_{TD}$ of $\hat{\lambda}$ as

$$\hat{\lambda}_{TD} = (\hat{\Lambda}'_{\mathbf{s}}, 0'_{K-\mathbf{s}})', \qquad \hat{\Lambda}_{\mathbf{s}} = \hat{\lambda}_{\mathbf{s}} + \hat{\Theta}_{\mathbf{s}} \alpha_n \hat{\kappa}_{\mathbf{s}},$$

and $0_{K-\mathbf{s}}$ is the $(K - \mathbf{s})$-dimensional vector of zeros. That is, we only correct the bias for the selected elements by $\hat{S}$. Then θ_0 is estimated by

$$\hat{\theta}_{TD} = \mathbb{E}_n[\phi_*^{(1)}(\hat{\lambda}'_{TD} g(X)) h(X,Y)].$$

Let $\tilde{\gamma}_n = \kappa_{\mathbf{o},n} \vee \sqrt{\mathbf{s} \log K/n}$, $\omega_{\mathbf{s}}(x) = \phi_*^{(1)}(\lambda'_{\mathbf{os}} g_{\mathbf{os}}(x))$, and $\tilde{h}^X_{TD}(x) = \tilde{\beta}'_{\mathbf{s}} g_{\mathbf{s}}(x)$, where

$$\tilde{\beta}_{\mathbf{s}} = \mathbb{E}[\phi_*^{(2)}(\lambda'_{\mathbf{os}} g_{\mathbf{s}}(X)) g_{\mathbf{s}}(X) g_{\mathbf{s}}(X)']^{-1} \mathbb{E}[\phi_*^{(2)}(\lambda'_{\mathbf{os}} g_{\mathbf{s}}(X)) g_{\mathbf{s}}(X) \mathbb{E}[h(X,Y)|X]].$$

For the limiting distribution of $\hat{\theta}_{TD}$, we add the following assumptions.

Condition TD.

1. *There exists functions* $r^h, \tilde{r}^h_{TD} : \mathcal{X} \to \mathbb{R}$ *such that* $\mathbb{E}[r^h(X)] = \mathbb{E}[\omega_0(X)\mathbb{E}[h(X,Y)|X]]$, $\mathbb{E}[\tilde{r}^h_{TD}(X)] = \mathbb{E}[\omega_0(X)\tilde{h}^X_{TD}(X)]$, *and*

$$\mathbb{E}[\{\tilde{\beta}'_{\mathbf{s}}(\omega_0(X) g_{\mathbf{s}}(X) - r(X)) - (\omega_0(X)\tilde{h}^X_{TD}(X) - \tilde{r}^h_{TD}(X))\}^2] \to 0.$$

2. $|\hat{\Theta} - Q^{(2)}(\lambda_{\mathbf{os}})^{-1}| = O_p(\varrho_n)$ *and* $\sqrt{n}\tilde{\gamma}_n \zeta_{\mathbf{s}} \varrho_n \to 0$.
3. *Condition I' holds true with* Λ_* *and* $\sqrt{\frac{\zeta^2_{\mathbf{s}}}{n}}$ *replaced by* $\lambda_{\mathbf{os}}$ *and* $\tilde{\gamma}_n$, *respectively.*
4. *Condition S' holds true with* ω_* *and* $\tilde{h}^X$ *replaced by* $\omega_{\mathbf{s}}$ *and* $\tilde{h}^X_{TD}$, *respectively*

.

Condition TD(1) is a counterpart of Condition N'(2). The roles of Conditions TD(3)-(4) for the targeted debiasing procedure are same as Conditions I' and S' for the post selection procedure, respectively. Condition TD(2) is concerned with quality of the targeted debiasing procedure. Under these conditions, the targeted debiasing estimator $\hat{\theta}_{TD}$ admits the same asymptotic representation as the post selection estimator as illustrated by the following theorem.

Theorem 1.6. *Suppose Conditions D', C, H, and TD hold true. Additionally assume* $\sqrt{n}\kappa^2_{\mathbf{o},n}\zeta^4_{\mathbf{s}} \to 0$, $\sqrt{n}\zeta^2_{\mathbf{s}}\tilde{\gamma}^2_n \to 0$, *and* $\mathbb{E}[(\Phi + \tilde{v}_1 + \tilde{v}_2 + \tilde{v}_3)^2] < \infty$. *Then*

$$\sqrt{n}(\tilde{\theta} - \theta_0 + \tilde{b}) \xrightarrow{d} N(0, \mathbb{E}[(\Phi + \tilde{v}_1 + \tilde{v}_2 + \tilde{v}_3)^2]).$$

where $\tilde{b}$, $\tilde{v}_1$, $\tilde{v}_2$, *and* $\tilde{v}_3$ *are same as those in Theorem 1.5 with replacements of* ω_*, $\tilde{h}^X$, *and* $\tilde{r}^h$ *with* $\omega_{\mathbf{s}}$, $\tilde{h}^X_{TD}$, *and* $\tilde{r}^h_{TD}$, *respectively.*

1.4 Theoretical application: treatment effect

In this section, we extend Example 1.2 in Section 1.1 and consider estimation of the average treatment effect. Let D_i be the indicator of a treatment for individual $i = 1, \ldots, n$ ($D_i = 1$ and 0 mean treated and not treated, respectively). For each i, there exist two potential outcomes, $Y_i(1)$ if treated and $Y_i(0)$ if not treated.

The observable outcome is $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$. Also, let X_i be covariates of individual i. Based on a random sample $\{D_i, Y_i, X_i\}_{i=1}^n$, we wish to estimate the average treatment effect $\tau = \mathbb{E}[Y(1) - Y(0)]$. Under the unconfoundedness and overlap assumptions, τ can be identified as (Rosenbaum and Rubin, 1983b)

$$\tau = \mathbb{E}[\omega^t(X)DY] - \mathbb{E}[\omega^u(X)(1 - D)Y] \equiv \theta^t - \theta^u,$$

where $\omega^t(x) = \pi(x)^{-1}$, $\omega^u(x) = \{1 - \pi(x)\}^{-1}$, and $\pi(x) = \Pr\{D = 1|X = x\}$ is the propensity score. We treat ω^t and ω^u as latent weight functions, and construct moment conditions as in (1.1) by utilizing the property of the propensity score:

$$\mathbb{E}[D\omega^t(X)g(X)] = \mathbb{E}[(1 - D)\omega^u(X)g(X)] = \mathbb{E}[g(X)], \tag{1.28}$$

for any g. By applying our methodology based on (1.28), the weight function ω^t can be estimated by

$$\begin{cases} \hat{\omega}^t(x) = \phi_*^{(1)}(\hat{\lambda}_1' g(x)) & \text{(low dimensional case)} \\ \tilde{\omega}^t(x) = \phi_*^{(1)}(\hat{\Lambda}_1' g(x)) & \text{(high dimensional case)} \end{cases},$$

where

$$\hat{\lambda}_1 = \begin{cases} \arg\min_{\lambda} \mathbb{E}_n[D\phi_*(\lambda' g(X)) - \lambda' g(X)] & \text{(low dimensional case)} \\ \arg\min_{\lambda} \mathbb{E}_n[D\phi_*(\lambda' g(X)) - \lambda' g(X)] + \alpha_{1n}\|\lambda\|_1 & \text{(high dimensional case)} \end{cases},$$

$$\hat{\Lambda}_1 = \arg\min_{\Lambda \in \mathbb{R}^{\mathbf{s}_1}} \mathbb{E}_n[D\phi_*(\Lambda' g_{\mathbf{s}_1}(X)) - \Lambda' g_{\mathbf{s}_1}(X)],$$

where $g_{\mathbf{s}1}$ is the $\mathbf{s}_1$-dimensional functions corresponding to $\hat{S}_1 = \{j : \hat{\lambda}_{1j} \neq 0\}$.

Then θ^t is estimated by $\hat{\theta}^t = \mathbb{E}_n[\hat{\omega}^t(X)DY]$ for the low dimensional case, or by the post selection estimator $\tilde{\theta}^t = \mathbb{E}_n[\tilde{\omega}^t(X)DY]$ for the high dimensional case. Similarly we can estimate ω^u and θ^u (by replacing D with $(1 - D)$). Then the average treatment effect τ can be estimated by $\hat{\tau} = \hat{\theta}^t - \hat{\theta}^u$ for the low dimensional case, or $\tilde{\tau} = \tilde{\theta}^t - \tilde{\theta}^u$ for the high dimensional case. By applying the results in the previous sections, we obtain the following corollary.

Corollary 1.1. *Consider the setup of this section. Suppose $D \perp (Y(1), Y(0))|X$ (unconfoundedness condition), and the propensity score π is bounded away from 0 and 1 over the compact support $\mathcal{X}$ (overlap condition). Furthermore, assume $\mathbb{E}[Y^2(0)] < \infty$, $\mathbb{E}[Y^2(1)] < \infty$.*

(i) *[Low dimensional case] Under the assumptions of Theorem 1.2, in particular,*

if

$$\begin{aligned}\sup_{x\in\mathcal{X}}|\mathbb{E}[Y(1)|X=x]-\lambda_1' g(x)| &= o(1),\\ \sup_{x\in\mathcal{X}}|\mathbb{E}[Y(0)|X=x]-\lambda_0' g(x)| &= o(1),\end{aligned}$$

for some $\lambda_1, \lambda_0 \in \mathbb{R}^K$*, it holds*

$$\sqrt{n}(\hat{\tau}-\tau) \xrightarrow{d} N(0,\Sigma),$$

where $\Sigma = \mathbb{E}\left[\{\mathbb{E}[Y(1)|X]-\mathbb{E}[Y(0)|X]-\tau\}^2 + \frac{\mathrm{Var}(Y(1)|X)}{\pi(X)} + \frac{\mathrm{Var}(Y(0)|X)}{1-\pi(X)}\right]$.

(ii) *[High dimensional case] Under the assumptions of Theorem 1.5, it holds*

$$\sqrt{n}(\tilde{\tau}-\tau+b_{ps}) \xrightarrow{d} N(0,\Sigma_{ps}),$$

where the formula of $b_{ps} \geq 0$ *and* $\Sigma_{ps} \geq \Sigma$ *can be found accordingly via Theorem 1.5.*

Proofs are similar to those of Theorems 1.2 and 1.5. This corollary may be considered as an extension of Chan et al. (2016) to the high dimensional case by using the ℓ_1-penalized estimator. Note that the asymptotic variance Σ is the semiparametric efficiency bound for τ established in Hahn (1998).

1.5 Empirical application: stochastic discount factor

To illustrate performance of the proposed method, we consider Example 1.1 in Section 1.1 and estimate the SDF in an equity market. We compare out-of-sample performance of the proposed method with other leading factors in empirical finance literature. In particular, the approach adopted by Ghosh et al. (2015) is a special case of ours for the low (and fixed) dimensional case. Our major findings are: (i) in the low dimensional setup where the number of portfolios in the market is small, predictability of our method is at least as good as the Fama-French three factors model, and the cross sectional errors are lower, and (ii) in a relatively high dimensional setup where the number of portfolios is similar to the number of training periods, upon choosing suitable penalty levels, our method outperforms the Fama-French three factors model while Ghosh et al. (2015)'s method shows erratic behaviors.

1.5.1 Methodology

Following the convention in empirical finance, we estimate the normalized SDF using Kullback-Leibler divergence function $\phi(x) = x \log x$ and *excess* asset returns. To be precise, in July year l, we form a training subsample using portfolio returns data of past 30 years. Based on this training subsample, we estimate the normalized SDF using the proposed method.

In particular, the moment condition in (1.1) can be written as

$$\mathbb{E}[\omega(R_{t_1})R_{t_1}] = 0,$$

where R_{t_1} is a vector of *excess* portfolio returns between time t_1 and $t_1 + 1$(a month between year l and $l - 30$) and $\omega(R_{t_1}) = \frac{\mathbb{E}[m_{t_1}|R_{t_1}]}{\mathbb{E}[m_{t_1}]}$ is the the normalized and projected SDF on excess returns. Since $\phi_*(y) = e^{y-1}$, applying our methods yields $\hat{\omega}(R_{t_1}) = \frac{\exp(\hat{\lambda}' R_{t_1})}{T_1^{-1}\sum_{t_1=1}^{T_1}\exp(\hat{\lambda}' R_{t_1})}$,[11] where

$$\hat{\lambda} = \begin{cases} \arg\min_{\lambda} T_1^{-1}\sum_{t_1=1}^{T_1}\exp(\lambda' R_{t_1}) & \text{(low dimensional portfolios)} \\ \arg\min_{\lambda} T_1^{-1}\sum_{t_1=1}^{T_1}\exp(\lambda' R_{t_1}) + \alpha_n \left\|\lambda\right\|_1 & \text{(high dimensional portfolios)} \end{cases}.$$

Based on this $\hat{\lambda}$, we predict the SDF using a testing subsample one year ahead from year l with total time periods of T_2. Then the estimated out-of-sample SDF from July year l to June year $l + 1$ would be $\hat{\omega}(R_{t_2}) = \frac{\exp(\hat{\lambda}' R_{t_2})}{T_2^{-1}\sum_{t_2=1}^{T_2}\exp(\hat{\lambda}' R_{t_2})}$, where t_2 is a month between year l and $l + 1$. We continue to build the estimated SDF time series in this fashion to cover all periods in our sample.

To test the cross-sectional predictability of our estimated out-of-sample SDF, we use the two-pass regression in empirical finance (Fama and MacBeth 1973; Cochrane 2009). In the first step, we run a time series OLS regression of excess returns R_j on our estimated out-of-sample SDF $\hat{\omega}$ for each portfolio j. We record its slope coefficient $\hat{\beta}_j$ as its factor loading. Then in the second step, we run a cross sectional OLS regression from $\bar{R}$ on $\hat{\beta}$, where $\bar{R}$ is a vector of average excess returns for all portfolios, and $\hat{\beta}$ is a vector of estimated factor loadings in the first step. We compare the adjusted R-squared as well as the estimated constant in the second regression to other empirical asset pricing models.

1.5.2 Data

All data are taken from Kenneth French's data library. To make the results comparable with existing literature (e.g., Fama and French 1993; Lewellen et al. 2010; Ghosh et al. 2015), the out-of-sample evaluation covers from July 1963

[11] For simplicity and convenient comparison with other methods, we do not use trimming.

to December 2010. We use monthly data so each training subsample is of size $T_1 = 360$ and each testing subsample is of size $T_2 = 12$ (except for the last rolling window where $T_2 = 6$). We only consider equity portfolios returns, which are quoted in %.

We compare three methods: Our method without penalty (essentially, Ghosh, Julliard and Taylor, 2016), our method with ℓ_1-penalization, and Fama-French three factors model. These methods are compared under the following scenarios.

(i) Low dimensional case: the SDF is constructed from 25 size and book-to-market portfolios, 10 momentum portfolios, 25 size, and long term reversal portfolios, respectively.

(ii) Intermediate case: the SDF is constructed from 100 size and book-to-market portfolios, 49 industry portfolios, and 25 long term reversal and size+25 short term reversal and size+25 momentum portfolios, respectively.

(iii) High dimensional case: use all portfolios available from Kenneth French's data library. Since some data are only available from 1960s, the out-of-sample period can only cover months from July 1993 to December 2010. The SDF is constructed from the two sets of portfolios: **(a)** 300 portfolios that include 100 portfolios based on size and book-to-market, 100 portfolios based on size and operating profitability, and 100 portfolios formed on size and investment, and **(b)** 425 portfolios that include 300 portfolios above, 49 industry portfolios, 25 portfolios on long term reversal and size, 25 portfolios on short term reversal and size, and 25 momentum portfolios.

1.5.3 Empirical results

1.5.3.1 Low dimensional and intermediate cases

Table 1 presents the cross sectional regression results for the low dimensional case. The numbers of portfolios are less than 30 in all panels and the training subsample size is 360. Although penalization seems unnecessary, we present the result when the penalty level is 0.05, a relatively small penalty, for comparison. The numbers in parentheses are t-values for the coefficients above. In all panels, the estimated price of risk is highly significant with the correct sign, either with or without penalty. The adjusted R-squared for the no penalty estimate is larger than the one for the Fama-French model in Panels A and B. The adjusted R-squared for the penalized estimate is worse than the one for the no penalty estimate in these two panels. Since the dimension is low, we expect every portfolio is informative and there is no need for penalization. Moreover, we can see that the

intercept estimates are all much smaller than the Fama-French estimates. This also indicates that our model is better than Fama-French three factor models. Panel C is interesting, where the no penalty estimate is worse than the penalized estimate. This result indicates usefulness of penalization even for the low dimensional case.[12]

Table 2 summarizes the results for the intermediate case, where the number of portfolios ranges from 50 to 100. The results are similar to the low dimensional case in Table 1. Our method (with or without penalty) outperforms the Fama-French model for most cases in terms the intercept estimates and adjusted R-squared.

1.5.3.2 High dimensional case

This case is of our major interest, where the no penalty estimate (essentially Ghosh et al., 2015) is not applicable or performs erratically, and it is crucial to introduce penalization. In this case, the choice of the penalty level becomes more important. We create a grid from 0.1 to 2 with 0.05 increments, estimate the SDF by our method, and implement the cross sectional regression for each penalty level.

The results are summarized in Figure 1. The SDF estimates without penalization perform very badly with the adjusted R-squared close to 0 and relatively large intercept estimates. As the penalty level increases, the performance of our method gets better. When the penalty level is approximately above 0.5, predictability of our method surpasses Fama-French, and the intercept estimates are much smaller. Then performance of our method gets worse when the penalty level continues to increase above 1.5. This is expected because the number of portfolios selected will be too small for too large penalty levels and the performance would deteriorate. Based on these results, we set the penalty level at 0.9 for 300 portfolios and 0.85 for 425 portfolios, and report the results in Table 3. We can see that the adjusted R-squared by the penalized SDF estimate is much higher than the one of Fama-French and that its intercept estimate is close to 0 and insignificant. Therefore, our method shows excellent performance upon choosing suitable penalty levels.

[12]Under-performance of the no penalty estimate (for both low and high dimensional cases) may be due to non-existence of higher moments. Note that both Ghosh et al. (2015) (no penalization) and our method using Kullback-Leibler divergence divergence (with ℓ_1 penalization) rely upon the exponential moments $\mathbb{E}[\exp(\lambda' R_{t_1})]$ whose finiteness implies infinite order of moments of R_{t_1}. If some higher moments of R_{t_1} does not exist, the no penalty estimator will behave erratically. Although formal analysis is beyond the scope of this chapter, we conjecture that our ℓ_1-penalization may effectively remove such problematic components in asset return movements. On the other hand, if non-existence of higher moments is a significant concern, we can flexibly choose a different divergence function, for example, Pearson's χ^2 divergence.

The number of active portfolios chosen in each year for each penalty level is summarized in Figures 2 (300 portfolios) and 3 (425 portfolios). Without penalization, too many portfolios will be used and cause undesirable performance. As the penalty level increases, the number of selected portfolios drops quickly, and at the levels used for Table 3, the number of portfolios selected is around 5-10.

1.5.3.3 Time series property of penalized SDF estimates

We illustrate time series properties of the penalized SDF estimates for 300 and 425 portfolios at the penalty levels used for Table 3. The plot is displayed in Figure 4 and the gray shaded areas correspond to NBER recessions. Our SDF estimates catch those macro events very well. In Table 4 we run a time series regression of our SDF estimates on other key factors in the market including Fama-French three factors and momentum factors. We can see that correlations of our SDF estimates with those leading factors are very small, and the adjusted R-squared is also small. This indicates that our method catches critical information for asset pricing in the market that cannot be explained by Fama-French and momentum factors.

1.5.4 Tables and figures

Table 1.1: Cross sectional regression in low dimensional case

	Const.	λ_{SDF}	λ_{RM}	λ_{SMB}	λ_{HML}	Adjusted R^2
Panel A: 25 size and book-to-market						
SDF: No penalty	0.649	-0.257				0.844
	(13.977)	(-11.438)				
SDF: $\alpha = 0.05$	0.720	-0.124				0.625
	(10.146)	(-6.400)				
3 Factors	1.668		-0.751	0.204	0.437	0.714
	(4.401)		(-2.067)	(3.853)	(6.773)	
Panel B: 10 momentum						
SDF: No penalty	0.752	-0.168				0.918
	(21.715)	(-10.056)				
SDF: $\alpha = 0.05$	0.716	-0.129				0.908
	(18.714)	(-9.493)				
3 Factors	2.365		-1.198	-0.068	-1.485	0.815
	(1.576)		(-0.754)	(-0.057)	(-1.615)	
Panel C: 25 long term reversal and size						
SDF: No penalty	0.741	-0.215				0.505
	(8.023)	(-5.049)				
SDF: $\alpha = 0.05$	0.382	-0.180				0.785
	(4.372)	(-9.416)				
3 Factors	0.702		0.219	0.111	0.633	0.754
	(2.541)		(0.833)	(1.678)	(5.051)	

Note: Cross sectional regression results in the low dimensional case. The estimated SDF is derived in a rolling window out-of-sample fashion from July 1963 to December 2010, using portfolios in each corresponding panel. Panel A presents results using 25 size and book-to-market portfolios, Panel B presents results using 20 momentum portfolios, and Panel C is concerned with results using 25 long term reversal and size portfolios. The second column is the estimated constant in each model, the last column records the adjusted R^2, and the other columns summarize estimated price of risk. Numbers in the bracket are the corresponding t-values. In each panel the first row is about the estimated SDF when no penalty is imposed, the second row is the estimated SDF when penalty level is at 0.05, and the third row is the seminal Fama-French three factor models.

Table 1.2: Cross sectional regression in intermediate case

	Const.	λ_{SDF}	λ_{RM}	λ_{SMB}	λ_{HML}	Adjusted R^2
Panel A: 100 size and book-to-market						
SDF: No penalty	1.033	-0.926				0.581
	(52.744)	(-11.532)				
SDF: $\alpha = 0.1$	0.725	-0.273				0.652
	(20.435)	(-13.367)				
3 Factors	1.575		-0.639	0.190	0.439	0.627
	(8.618)		(-3.670)	(5.577)	(11.175)	
Panel B: 49 industry						
SDF: No penalty	0.800	-0.129				0.329
	(16.239)	(-4.852)				
SDF: $\alpha = 0.1$	0.686	-0.065				0.294
	(0.686)	(-0.065)				
3 Factors	1.064		-0.008	-0.096	-0.109	-0.002
	(6.229)		(-0.047)	(-0.923)	(-1.151)	
Panel C: 25 long term reversal+25 short term reversal+25 momentum						
SDF: No penalty	1.083	-1.919				0.605
	(48.960)	(-10.698)				
SDF: $\alpha = 0.1$	1.130	-0.484				0.441
	(43.162)	(-7.705)				
3 Factors	1.416		-0.432	0.293	0.012	0.153
	(4.489)		(-1.454)	(3.370)	(0.064)	

Note: Cross-sectional regression results in the intermediate case. The estimated SDF is derived in a rolling window out-of-sample fashion from July 1963 to December 2010, using portfolios in each corresponding panel. Panel A presents results using 100 size and book-to-market portfolios, Panel B presents results using 49 industry portfolios, and Panel C presents results using 75 portfolios listed in the beginning of the panel. The second column is the estimated constant in each model, the last column records the adjusted R^2, and the other columns summarize estimated price of risk. Numbers in the bracket are the corresponding t-values. In each panel the first row is about the estimated SDF when no penalty is imposed, the second row is the estimated SDF when penalty level is at 0.1, and the third row is the seminal Fama-French three factor models.

Figure 1.1: Summary of cross sectional regression against different penalty levels in high dimension case ($K = 300$ or 425; $T = 360$)

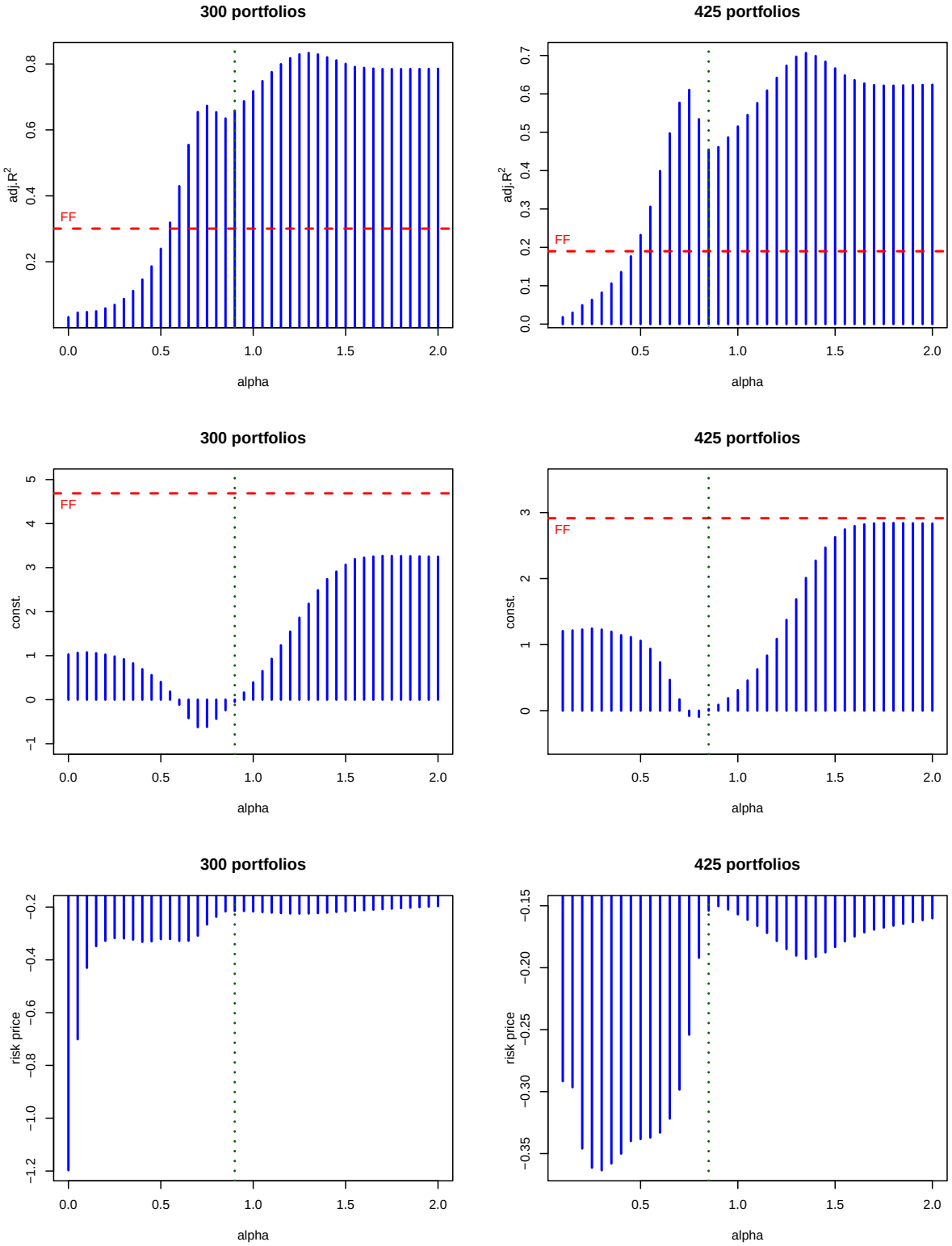


Figure 1.2: Number of active portfolios selected under 300 portfolios case

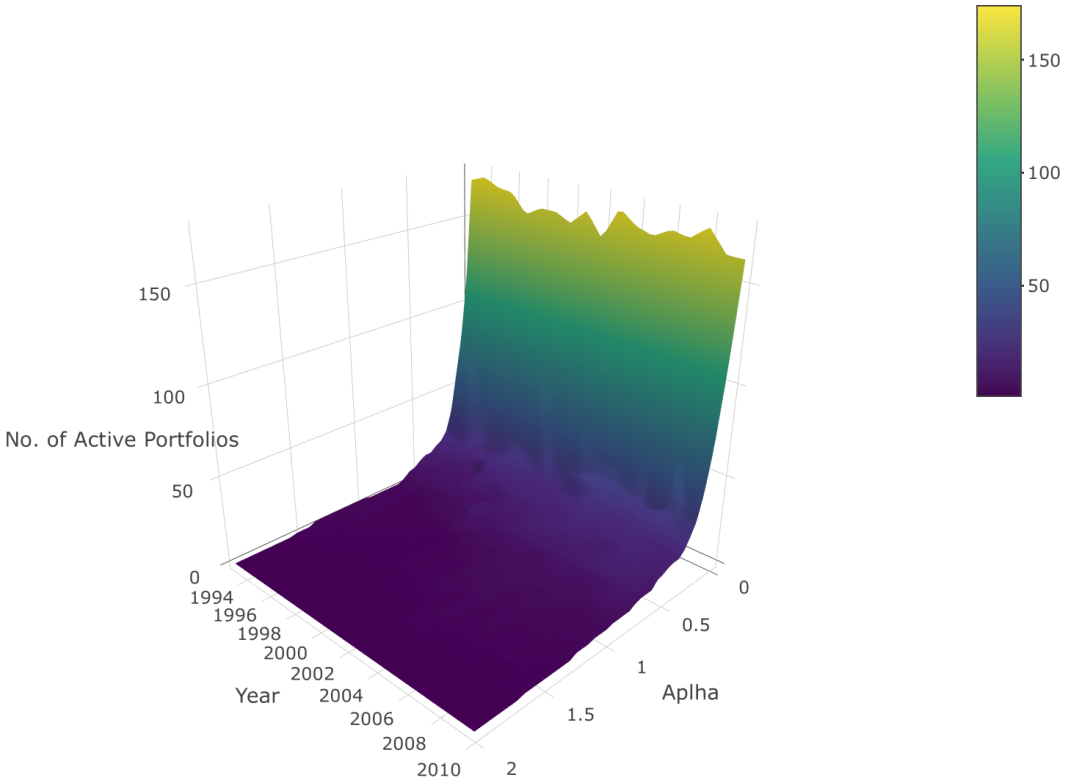


Figure 1.3: Number of active portfolios selected under 425 portfolios case

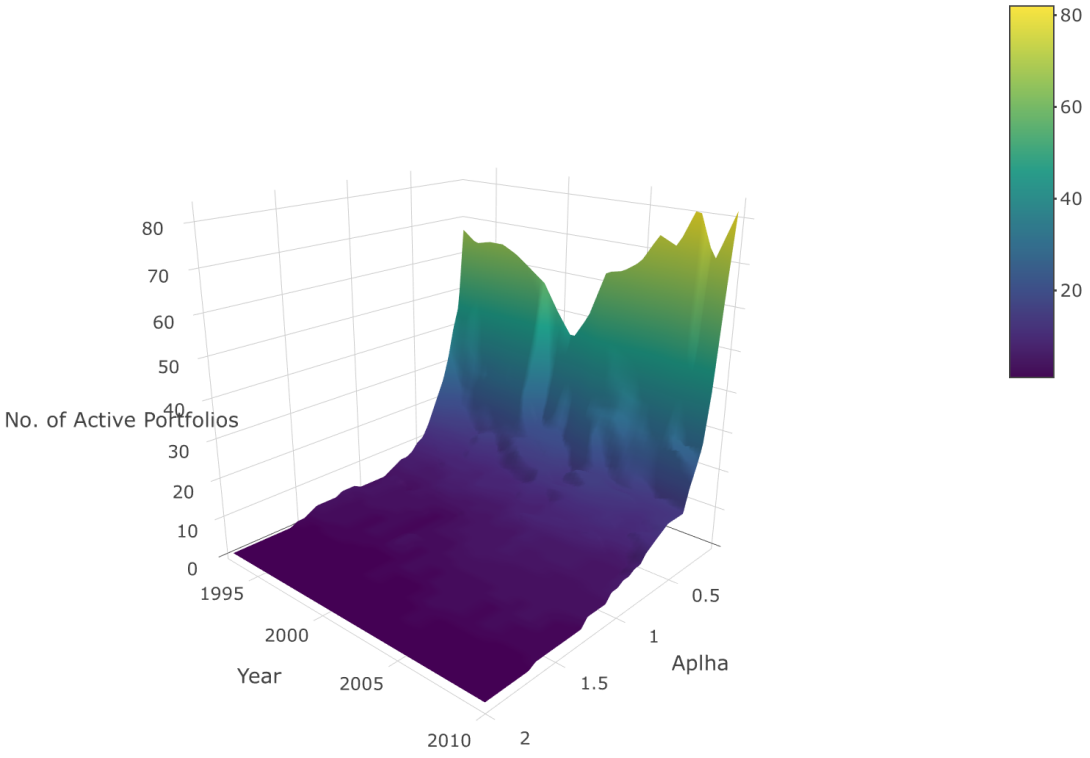


Table 1.3: Cross sectional regression in high dimensional case

	Const.	λ_{SDF}	λ_{RM}	λ_{SMB}	λ_{HML}	Adjusted R^2
Panel A: 300 portfolios 100 size & book-to-market+100 size & operating profitability+100 size & investment						
SDF: $\alpha = 0.1$	1.027 (14.062)	-1.197 (-3.306)				0.032
SDF: $\alpha = 0.9$	-0.050 (-0.851)	-0.214 (-24.017)				0.658
3 Factors	4.687 (10.986)		-3.891 (-9.998)	0.699 (5.295)	-0.517 (-2.900)	0.301
Panel B: 425 portfolios 300 in Panel A+49 industry+25 long term rev.+25 short term rev.+25 momentum						
SDF: $\alpha = 0.1$	1.206 (12.684)	-0.292 (-2.967)				0.018
SDF: $\alpha = 0.85$	0.024 (0.383)	-0.154 (-18.800)				0.455
3 Factors	2.914 (10.507)		-2.121 (-8.339)	0.659 (6.331)	-0.305 (-2.205)	0.190

Note: Cross-sectional regression results in the high dimensional case. The estimated SDF is derived in a rolling window out-of-sample fashion from July 1993 to December 2010, using portfolios in each corresponding panel. Panel A presents results using 300 portfolios, and Panel B presents results using 425 portfolios. The second column is the estimated constant in each model, the last column records the adjusted R^2, and the other columns summarize estimated price of risk. Numbers in the bracket are the corresponding t-values. In each panel the first row is about the estimated SDF when the penalty level is set at 0.1, the second row is the estimated SDF when penalty level is at 0.9 and 0.85, respectively, and the third row is the seminal Fama-French three factor models.

Figure 1.4: Time series plot of estimated SDF in high dimensional case: July 1993 - December 2010.
Grey shaded area represents NBER recessions

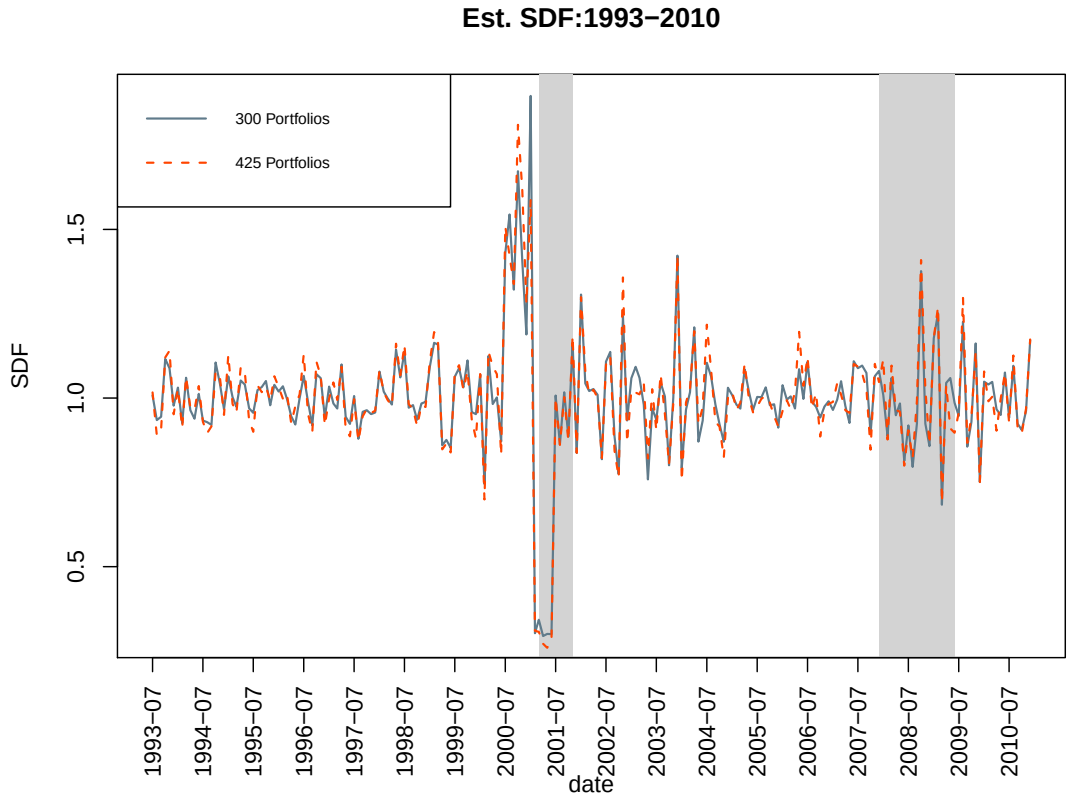


Table 1.4: Time series properties of estimated SDF from high dimensional case

α	β_{RM}	β_{SMB}	β_{HML}	β_{MOM}	Adjusted R^2
Panel A: 300 portfolios, $\alpha = 0.9$					
1.011	-0.004	-0.014	-0.007	-0.007	0.118
(85.846)	(-1.427)	(-4.106)	(-1.851)	(-3.264)	
Panel B: 425 Portfolios, $\alpha = 0.85$					
1.012	-0.006	-0.016	-0.003	-0.008	0.171
(84.730)	(-2.340)	(-4.712)	(-0.879)	(-3.594)	

Note: Time series regression of estimated SDF extracted from the high dimensional case against key factors in the market. The estimated SDF is derived in a rolling window out-of-sample fashion from July 1993 to December 2010, using portfolios and penalty level in each corresponding panel. Panel A presents results using 300 portfolios and when penalty level is 0.9, and Panel B presents results using 425 portfolios and when penalty level is set at 0.85. The first column is the estimated constant (or, "alpha") in each regression, the last column records the adjusted R^2, and the other columns summarize estimated beta for each factor. Numbers in the bracket are the corresponding t-values.

Chapter 2

Minimax learning for average regression functionals: framework

2.1 Introduction

Empirical analyses in many disciplines increasingly rely on a large number of control variables (hereafter controls). A vast majority of research in empirical economics works with observational data (Imbens and Wooldridge, 2009; Angrist and Pischke, 2010; Oster, 2017). So causal inference usually needs some form of exogeneity assumption (Heckman et al., 1999; Imbens, 2004), which is more plausible conditioning on many controls. For example, to get results comparable to experimental studies in job training programs, Heckman et al. (1997, 1998) recommend including a full set of variables related to program participation and labor market outcomes. With the sizable datasets economists are working with nowadays, it is also common to include many fixed effects in regressions to capture widespread heterogeneity in data (Card et al., 2013) or to simulate a quasi-experimental environment (Rossin-Slater, 2017). In some fields, say political economy (Ferraz and Finan, 2011) and macroeconomics (Nakamura and Steinsson, 2018), ideal experimental designs rarely exist. Therefore, whether included controls are rich enough to capture main source of bias becomes crucial. Even in experimental studies, including covariates is able to sometimes reduce variance (Section 4.4, Duflo et al., 2007) or alleviate crossover effect (Linnemayr and Alderman, 2011).

However, the presence of many controls challenges our practice in empirical research. Albeit popular, OLS is subject to a problem of "over control" (or "spurious omitted variable bias"). Without any discipline of the coefficients, adding more regressors can distort estimates. Such distortion can be so large that improvement from addressing omitted variable bias is negligible. Ignoring such effect can lead

to misinterpretation of empirical results. If data do contain considerable bias, this bias issue further intertwines with the many-controls issue in a complicated way. Using OLS is potentially even more dangerous: a correctly specified model with many controls can perform worse than a biased model with fewer controls in terms of mean square error. Off-the-shelf shrinkage methods do not solve "over control" easily. An influential line of research from Belloni et al. (2012, 2014, 2017b, etc.) advances the state by validating inference based on (post) lasso selections. Although allowing a much larger number of controls, their results naturally extend a question of whether some of the technical assumptions are realistic to suit broader scenarios in applied research.

Standard econometric theory also appears less satisfactory when the number of covariates becomes large. To achieve desirable asymptotic properties, bias and variance of many nonparametric methods need to be carefully traded off with the help from tuning parameters (for series method, it is the number of basis functions k in a sample of size n). For example, to establish $\sqrt{n}$ normality, default theory often requires nonparametric part of the model to reach at least $o_p(n^{-1/4})$ convergence rate. Such a condition can be strict for problems with many controls, as it implies k should grow slowly. Even if these technical conditions are asymptotically attainable, they leave little guidance on the actual choice of k for any realistic sample size. As a result, empirical performance of fitted models is often sensitive to k. This can still happen for off-the-shelf shrinkage methods if the number of controls is too large.

In recognition of these empirical and theoretic challenges, the rest of the thesis aims to develop an estimation method that

1. Achieves $\sqrt{n}$ normality under weaker conditions. Indeed, although there are many results on the attainability of $\sqrt{n}$ normality and semiparametric efficiency, it is not yet crystal clear to what extent we can relax those technical conditions.

2. Displays improved finite sample performance. The estimation method should avoid introducing undesirable distortion when dimension increases, i.e., it should not "over control".

3. Is straightforward to implement so applied researchers can apply conveniently.

To achieve these goals, I focus on semiparametric models whose object of interest can be expressed as a population weighted average of a nonparametric regression. This simple setting is relevant to many empirical problems in economics. See Section 2.2 for a detailed discussion. It also allows us to explore interesting facets

of asymptotic theory absent otherwise for general models. The rest of the thesis contributes to the literature mainly in the following two aspects:

1. it proposes a new minimax methodology and develops a new class of estimators called "minimax learners".

2. it studies asymptotic properties of minimax learners and presents a relative complete set of distributional results under various dimensional frameworks and considerably weak conditions.

Embedded in a penalized series space, the estimating strategy exploits a minimax property of a fundamental component (called Riesz Representer, or RR) of the average regression functional and aims to directly control key remainder bias term. This strategy is inspired by two branches of exceptional work in recent literature: series estimation of RR and its high dimensional variants, pioneered by Newey and Robins (2018); Chernozhukov et al. (2018c); and minimax approach to linear semiparametric models, explored in Hirshberg and Wager (2018); Wong and Chan (2018). This chapter deviates from previous literature, calibrating RR in the penalized series space with a criterion different from Hirshberg and Wager (2018). It is carefully designed to accommodate even high dimensional situations. Penalized series space proves a powerful tool, improving finite sample and asymptotic performance. As a first theoretic result, this chapter uncovers an intimate connection between proposed minimax strategy and traditional minimum distance estimation. Minimax exercise embedded in the series space can be analyzed in a weighted minimum distance fashion. It thus links this chapter with earlier fine works of Chen and Pouzo (2012, 2015a), but with inherently different motivations. Due to the minimum distance structure, this minimax exercise is straightforward to implement and computationally competitive compared to other infinite dimensional minimax exercises.

When $\frac{k}{n} \to 0$ (up to log term), I show a kind of plug-in estimator constructed with calibrated RR (called "minimax BP learner") can achieve $\sqrt{n}$ normality and semiparametric efficiency under weak conditions. If the underlying regression function is smooth enough, calibrated RR only needs to be consistent in $\|\cdot\|_{\mathbb{P},2}$ norm without a convergence rate requirement. This is close to the minimal condition so far in the literature, which relies on cross fitting. More interestingly, the minimax BP learner is able to achieve $\sqrt{n}$ normality even when $\frac{k}{n} \to c$ where $0 < c < 1$. Under this asymptotic framework, calibrated RR is not even consistent. $\sqrt{n}$ normality comes from fast approximation of the regression function and a central limit theorem established around sample mean.

This framework also accommodates ultra high dimensional situations when $\frac{k}{n} \to \infty$, suitable if there are many technical controls or interaction terms. I show

that a doubly robust estimator constructed with calibrated RR (called "minimax DR learner") is $\sqrt{n}$ normal and semiparametrically efficient under conditions weaker than Belloni et al. (2017b); Chernozhukov et al. (2018a), and comparable to Chernozhukov et al. (2018c). In particular, calibrated RR only needs to be l_1 consistent with no rate condition per se, if it converges relatively fast under a weak norm, which also does not have to be $o_p(n^{-1/4})$. These results are derived together with two key techniques. First, further l_1 regularization is added to the original penalized series space, making calibrated RR a Generalized Minimax Elastic Net (GMEN) learner. GMEN learner can sometimes converge faster than a pure lasso based method. Second, the first step estimator of the regression function is assumed to be derived from a different random sample, in line with recent literature advocating cross fitting. This also allows inclusion of any machine learning method in the first step. However, GMEN learner is still calibrated from the main sample, weaker than earlier results in the literature.

As an empirical illustration, I revisit Ferraz and Finan (2011)'s work that exploits a natural experiment and studies the effect of electoral accountability on corruption. With plausibly exogenous treatment, one of their main empirical strategies is OLS with many controls. I find estimates from OLS change considerably as more covariates are sequentially added to the regression. Minimax learners, on the other hand, perform stably and produce economically coherent conclusions, even when the number of controls is much larger. Other popular off-the-shelf shrinkage methods do not work as well as minimax learners. This exercise shows the main conclusion of Ferraz and Finan (2011) is robust. However, the over control problem of OLS is salient in this setting. Ignoring this effect leads to an interpretation of spurious omitted variable bias in their data.

2.1.1 Related literature

The rest of the thesis builds on a sequence of fruitful research on semiparametric inference. For general results, see Bickel (1982); Robinson (1988); Newey (1990); Van Der Vaart et al. (1991); Bickel et al. (1993); Andrews (1994); Newey (1994b); Newey and McFadden (1994); Van der Vaart (1998), etc. The workhorse asymptotics is the series method, well studied by at least Newey (1991); Donald and Newey (1994); Newey (1997); Shen (1997); Huang (2003); Ai and Chen (2003); Newey and Powell (2003), etc. For a review, see Chen (2007). And see Belloni et al. (2015); Chen and Pouzo (2015b); Hansen (2015) for recent asymptotic refinement.

RR is a fundamental concept in functional analysis. It stems from the famous Riesz representation theorem that connects Hilbert space and its dual via a simple

inner product structure. While the role of RR has been widely known to establish semiparametric efficiency bound (Newey 1990, 1994b), it is relatively a recent matter to study how RR can be directly approximated without knowing its functional form. The average regression functional set-up in this chapter follows the exemplary work of Newey and Robins (2018), who introduce series estimation of RR and study asymptotic properties of corresponding cross-fitted doubly robust and plug-in estimators. Similar framework is used later by Chernozhukov et al. (2018b,c), extending to high dimensional cases by lasso and Dantzig regularized minimum distance estimators. The ultimate goal of this chapter is quite similar, but the minimax motivation is conceptually different. Compared to Newey and Robins (2018), I work with penalized series space under a different identification strategy, and does not use sample splitting. This in fact connects with the penalized sieve minimum distance (PSMD) estimation studied in Chen and Pouzo (2012, 2015a). Compared to Chernozhukov et al. (2018c), proposed GMEN learner can sometimes attain a better convergence rate. Sample splitting procedure in this thesis is slightly weaker, and a data-driven algorithm for the l_1 penalty coefficient is also proposed.

The rest of the thesis has also been inspired by recent minimax approach to semiparametric models. Specifically, Wong and Chan (2018) study a similar problem for average treatment effect in reproducing kernel Hilbert space, but do not show consistency of their weight (which is closely related to RR). Under the framework of Newey and Robins (2018), Hirshberg and Wager (2018) propose a minimax exercise in general, infinite dimensional penalized space, and focus on asymptotic properties of doubly robust estimators. Their outstanding works motivate us to look at a minimax problem but in penalized series space, which to the best of my knowledge, has not been studied systemically before. While their asymptotics centers around what I think as low dimensional cases, the minimax strategy in this chapter is carefully devised to accommodate both low and high dimensional situations. The simple linear structure of series space enables us to get a richer set of asymptotic results, and offers computational advantage as well. The broader minimax idea is not new in the literature. It has been recently revisited in a range of topics, like kernel methods for linear functionals (Armstrong and Kolesár, 2018b,a) and regression discontinuity designs (Imbens and Wager, 2018).

Minimax BP learner has a close relation with the popular balancing method in statistic literature, for example, Hainmueller (2012); Imai and Ratkovic (2014); Zubizarreta (2015); Chan et al. (2016); Kallus (2016); Athey et al. (2018), etc. In fact, RR is the ideal weight to be used in balancing literature. Asymptotic results when $\frac{k}{n} \to c > 0$ have also been inspired by alternative asymptotic theories devel-

oped in Chao et al. (2012); Cattaneo et al. (2018b,a) and by conditional inference approach discussed in Athey et al. (2018). In ultra high dimensional situations when $\frac{k}{n} \to \infty$, minimax DR learner builds on the popular "doubly robust" or "locally robust" literature, see Belloni et al. (2012); Farrell (2015); Chernozhukov et al. (2016); Rothe and Firpo (2016); Belloni et al. (2017b); Chernozhukov et al. (2018b,a), etc. In terms of high dimensional technical results, I benefit from an outstanding line of research on l_1 regularization methods (Bühlmann and Van De Geer, 2011; Bickel et al., 2009; Koltchinskii et al., 2009; Koltchinskii, 2009, etc.) and extensions to elastic net style regularizations (Bunea, 2008; Hebiri et al., 2011).

2.1.2 Notations and definitions for the rest of the thesis

Triangular array asymptotics

The rest of the thesis works with triangular array data $\{Y_{i,n}, Z_{i,n}, X_{i,n}\}_{i=1}^{n}$, which are first n items of infinite sequences $\{Y_{i,n}, Z_{i,n}, X_{i,n}\}_{i=1}^{\infty}$ generated from probability measure $\mathbb{P} = \mathbb{P}_n$. To simplify notation, use $\{Y_i, Z_i, X_i\}_{i=1}^{n}$ and $\mathbb{P}$ instead. Thus write $\mathbb{E}[\cdot\,] = \mathbb{E}_{\mathbb{P}}[\cdot\,]$ as the expectation operator under $\mathbb{P}$.

Random variables and functions

Capital letters $X, Y \ldots$ usually refer to random variables, while small letters $x, y \ldots$ refer to arguments in their supports or some numbers. For a vector $a = (a_1, a_2 \cdots, a_k)' \in \mathbb{R}^k$, $m(z, a) = [m(z, a_1), m(z, a_2) \cdots, m(z, a_k)]'$ is a $k-$dimensional column vector. $\mathbf{1}\{\cdot\}$ is the indicator function. $\mathbb{E}_n[\cdot\,] = \frac{1}{n}\sum_{i=1}^{n}(\cdot\,)$ is the empirical average.

Norms

For a vector $a = (a_1, a_2 \cdots, a_k)' \in \mathbb{R}^k$, let $\|a\| = \left(\sum_{j=1}^{k} a_j^2\right)^{1/2}$, $\|a\|_1 = \sum_{j=1}^{k} |a_j|$ and $\|a\|_\infty = \max_{1\le j\le k} |a_j|$ denote its l_2, l_1 and sup norms, respectively. For a function $f : \mathcal{X} \to \mathbb{R}$, let $\|f\|_{\mathbb{P},q} = \left[\int |f(x)|^q \, d\mathbb{P}(x)\right]^{1/q}, 1 \le q \le \infty$ denote its $L^q(\mathbb{P})$ norm. In particular, $\|f\|_{\mathbb{P},2}^2 = \mathbb{E}[f(X)^2]$, $\|f\|_{\mathbb{P},\infty} = \sup_{x\in\mathcal{X}} |f(x)|$. For a square matrix $A = \{a_{ij}\}_{i,j=1}^{k}$, let $\lambda_{\max}(A)$, $\lambda_{\min}(A)$ and $tr(A)$ be its largest eigenvalue, smallest eigenvalue and trace, respectively. Thence let $\|A\| = \sqrt{\lambda_{\max}(A'A)}$ be its spectral norm. If A is symmetric, $\|A\| = \max_i |\lambda_i(A)|$. Write $\|A\|_{\max} = \max_{1\le i,j\le k} |a_{ij}|$, $\|A\|_\infty = \max_{1\le i\le k} \sum_{j=1}^{k} |a_{ij}|$.

Numbers and words

For two sequences of numbers a_n and b_n, $a_n \vee b_n = \max\{a_n, b_n\}$, $a_n \wedge b_n = \min\{a_n, b_n\}$; $a_n \lesssim b_n$ means $a_n \leq c b_n$ for some constant c that does not depend on n. Bold $\mathbf{0}$ represents a (k dimensional) vector with each entry valued 0. "Wpa1" means "with probability approaching 1". "LIE" means "Law of Iterated Expectations". "CIA" means "Conditional Independence Assumption".

Empirical process

For a function class $\mathcal{F}$, let $\|Q\|_{\mathcal{F}} = \sup\{|Qf| : f \in \mathcal{F}\}$. An envelope function for $\mathcal{F}$ is defined as some F such that $|f(x)| \leq F(x)$, for each $f \in \mathcal{F}$ and each $x \in \mathcal{X}$. The covering number $N(\mathcal{F}, L^2(Q), \delta)$ is the smallest number of $L^2(Q)$ balls of radius δ to cover $\mathcal{F}$. A Rademacher random variable η_i is such that $\mathbb{P}(\eta_i = -1) = \mathbb{P}(\eta_i = 1) = \frac{1}{2}$.

Convex functions in Hilbert space

Let $\mathcal{H}$ be a Hilbert space with inner product $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$. Write $f(h) : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ as an extended real function. Denote $dom(f) = \{h \in \mathcal{H} : f(h) < +\infty\}$ as the domain of f. Function f is proper if $dom(f) \neq \emptyset$ and is convex if

$$f(\lambda h_1 + (1-\lambda)h_2) \leq \lambda f(h_1) + (1-\lambda) f(h_2) \tag{2.1}$$

for each $h_1, h_2 \in dom(f)$ and $\lambda \in (0,1)$. If inequality in (2.1) is strict whenever $h_1 \neq h_2$ and $\lambda \in (0,1)$, say f is strictly convex. Function f is strongly convex with parameter $\mathbf{c} > 0$ if

$$f(\lambda h_1 + (1-\lambda)h_2) \leq \lambda f(h_1) + (1-\lambda) f(h_2) - \frac{\mathbf{c}}{2}\lambda(1-\lambda)\|h_1 - h_2\|_{\mathbb{P},2}$$

for each $h_1, h_2 \in dom(f)$ and $\lambda \in (0,1)$.

2.2 Average regression functional and related examples

This section introduces the average regression functional framework. Suppose we observe a scalar valued random variable $Y \in \mathbb{R}$, a d_Z-dimensional random vector Z and a d_X-dimensional subvector X of Z

$$Z \in \mathcal{Z} \subseteq \mathbb{R}^{d_Z}, \quad X \in \mathcal{X} \subseteq \mathbb{R}^{d_X}, \quad X \subseteq Z. \tag{2.2}$$

Define $\gamma_0(x) = \mathbb{E}[Y|X = x] \in \Theta_\gamma$ as the conditional expectation function. Assume Θ_γ is a nonempty convex interior of the $L_{\mathbb{P},2}$ space that consists of all $f : \mathcal{X} \to \mathbb{R}$ such that $\mathbb{E}[f(X)^2] < \infty$. Object of interest θ_0 for the rest of the thesis is the continuous linear functional $\mathbb{E}[m(Z, \cdot)] : L_{\mathbb{P},2} \to \mathbb{R}$ evaluated at γ_0

$$\theta_0 = \mathbb{E}[m(Z, \gamma_0(X))], \tag{2.3}$$

where $m(z, \cdot)$ is a known linear function such that for every $\gamma_1, \gamma_2 \in L_{\mathbb{P},2}$ and every constant $r \in \mathbb{R}$

$$m(z, r\gamma_1(x) + \gamma_2(x)) = rm(z, \gamma_1(x)) + m(z, \gamma_2(x)). \tag{2.4}$$

By Riesz representation theorem, $\mathbb{E}[m(Z, \cdot)]$ admits a simple inner product structure in $L_{\mathbb{P},2}$: There exists a unique $\alpha_0 \in \Theta_\alpha \subseteq L_{\mathbb{P},2}$ such that for each $\gamma \in L_{\mathbb{P},2}$

$$\mathbb{E}[m(Z, \gamma(X))] = \mathbb{E}[\gamma(X)\alpha_0(X)]. \tag{2.5}$$

Call α_0 the Riesz Representer (hereafter RR) of $\mathbb{E}[m(Z, \cdot)]$. Assume Θ_α is also a nonempty convex interior of the $L_{\mathbb{P},2}$ space. (2.5) has been commonly used to establish semiparametric variance bound for θ_0 via a mean square continuity condition (for example, see Newey, 1994b). This chapter studies how (2.5) helps identify and approximate α_0 without knowing its functional form, which is an essential purpose for the rest of the thesis. Indeed, by (2.5), we can interpret θ_0 as a RR weighted population average of a regression function (aka "average regression functional")

$$\theta_0 = \mathbb{E}[\underbrace{\alpha_0(X)}_{\text{weight}} \underbrace{\gamma_0(X)}_{\text{regression}}]. \tag{2.6}$$

As shown by examples below, many economic problems display this structure. To save space, detailed derivations of RR in some examples are left in Appendix B.

Example 2.1. Missing data and average treatment effect

Consider a framework of incomplete outcome data studied in Rubin (1974); Rosenbaum and Rubin (1983a). For each individual unit $i = 1 \dots n$, we observe an indicator variable T_i ($T_i = 1$ if unit i responds and $T_i = 0$ if missing), outcome variable $Y_i = T_i Y_i^*$ ($Y_i = 0$ means Y_i^* is missing), and a covariate vector X_i that describes pre response individual characteristics. We are concerned about the population mean $\theta_0 = \mathbb{E}[Y^*]$. Under the assumption that Y^* and T are conditionally independent given X, θ_0 can be identified as

$$\theta_0 = \mathbb{E}[\gamma_0(X, 1)],$$

where $\gamma_0(x,1) = \mathbb{E}[Y|X = x, T = 1]$. Define the inverse propensity score as $\omega(x) = 1/\mathbb{P}\{T = 1|X = x\}$. Further under overlap assumption that

$$0 < \mathbb{P}[T = 1|X = x] < 1 \quad \text{for all } x \in \mathcal{X},$$

we have for each $g \in L_{\mathbb{P},2}$

$$\mathbb{E}[\omega(X)Tg(X)] = \mathbb{E}[g(X)]. \tag{2.7}$$

(2.7) identifies RR as

$$\alpha_0(x,t) = \omega(x)t.$$

This framework can be extended to account for average treatment effect (see for example Qiu and Otsu, 2018), which has become one of the most popular approaches to causal analysis in observational studies (for a review, see Imbens and Rubin, 2015; Imbens and Wooldridge, 2009).

Example 2.2. Regression discontinuity design away from cut-off

Slightly modify Example 2.1 but keep notation (Y, T, X). In addition, researchers understand that T is determined by a scalar running variable R at cut-off point 0 (without loss of generality, hereafter wlog)

$$T = \mathbf{1}\{R \geq 0\},$$

where $\mathbf{1}\{\cdot\}$ is the indicator function. Fix a known boundary point $b > 0$. Object of interest is the population mean of Y^* in a neighborhood around cut-off

$$\theta_0 = \ \mathbb{E}[Y^*| - b \leq R \leq b] \ .$$

This object is helpful for external validity reasons, for example, when we are interested in the population group away from cut-off (say, inframarginal applicants) instead of a group at the immediate neighborhood of cut-off. One way to identify θ_0 is through the Conditional Independence Assumption, similar to Angrist and Rokkanen (2015): Y^* and R are independent conditional on X and $-b \leq R \leq b$. Then it can be shown

$$\theta_0 = \mathbb{E}[\gamma_0(X)| - b \leq R \leq b], \tag{2.8}$$

where $\gamma_0(x) = \mathbb{E}[Y|0 \leq R \leq b, X = x]$. RR is found in a fashion similar to (2.7) under suitable overlap assumption

$$\alpha_0(x,r) = \omega(x)\mathbf{1}\{r \geq 0\}, \tag{2.9}$$

where $\omega(x) = 1/\mathbb{E}\left[\mathbf{1}\{R \geq 0\}|X = x, -b \leq R \leq b\right]$ is the (R-linked) inverse propensity score.

Example 2.3. Weighted average derivative and single index model

Suppose we are interested in the weighted average of some partial derivative of a regression function

$$\theta_0 = \mathbb{E}\left[w(X)\frac{\partial \gamma_0(X)}{\partial X_1}\right], \tag{2.10}$$

where X is a vector of covariates whose first element is X_1, $w(x)$ is a known weight function and $\gamma_0(x) = \mathbb{E}[Y|X = x]$. (2.10) encompasses several interesting models in the literature, like single index model (Stoker, 1986; Härdle and Stoker, 1989; Powell et al., 1989, etc.) and nonseparable model (Imbens and Newey, 2009; Altonji et al., 2012, etc.). See Cattaneo et al. (2013) for a review. To find RR, assume $w(x)$ has value 0 at boundaries. Integration by parts yields

$$\theta_0 = -\mathbb{E}\left[\gamma_0(X)\frac{\partial v(X)/\partial X_1}{f(X)}\right],$$

where $v = wf$ and f is the density of X. RR is then identified as

$$\alpha_0(x) = -\frac{\partial v(x)/\partial X_1}{f(x)}.$$

See Newey and Stoker (1993) and Newey and Robins (2018) for more details.

Example 2.4. Average effect after policy intervention

This set-up was introduced by Stock (1989) and has been further studied by Rothe and Firpo (2016). As usual let $\gamma_0(x) = \mathbb{E}[Y|X = x]$ be the conditional expectation and $\pi(x)$ be a known policy function. Intuitively, the distribution of X is shifted to a new random variable X_π such that $X_\pi(x) = \pi(x)$ after policy intervention. We are interested in predicting average effect on outcome Y after policy intervention, written as

$$\theta_0 = \mathbb{E}[\gamma_0(\pi(X))]. \tag{2.11}$$

Rewriting (2.11) by change of measure, we find RR in this case as $\alpha_0(x) = \frac{f_\pi(x)}{f(x)}$, where f and f_π are densities of X and X_π, respectively.

Example 2.5. Average consumer surplus

This is an example from nonparametric welfare analysis. To introduce the idea, consider a highly simplified version of the problem studied in Hausman and Newey (1995, 2016, 2017). Write the demand function of a commodity as $\gamma_0(p, z) = \mathbb{E}[Q|P = p, Z = z]$ where P is the price, Q is the quantity demanded

and Z is a vector of other characteristics that affect demand. Wlog, denote Z_1, the first variable of Z, as income. Thus define approximate consumer surplus for a price change from p_0 to p_1 as $\int_{p_0}^{p_1} \gamma_0(p, Z)dp$. Object of interest is

$$\theta_0 = \mathbb{E}\left[\omega(Z)\int_{p_0}^{p_1} \gamma_0(p, Z)dp\right],$$

for some known weight function $\omega(z)$. Parameter θ_0 is viewed as the average effect of the price change on certain income (and possibly other observable characteristics) groups. Let $f_{P|Z}(p|z)$ be the conditional density. It can be shown that

$$\theta_0 = \mathbb{E}\left[\frac{\omega(Z)\mathbf{1}\{p_0 \leq P \leq p_1\}}{f_{P|Z}(P|Z)}\gamma_0(P, Z)\right],$$

suggesting that RR admits $\alpha_0(p, z) = \frac{\omega(z)\mathbf{1}\{p_0 \leq p \leq p_1\}}{f_{P|Z}(p|z)}$. Allowing individual heterogeneity, Hausman and Newey (2016) extend the above analysis to derive bounds on average exact consumer surplus, which is also an average regression functional. See Chernozhukov et al. (2018c) for further derivation of RR in these more sophisticated cases.

Example 2.6. Measurement error with auxiliary data

This example is inspired by Chen et al. (2005); Lee and Sepanski (1995). To simplify presentation suppose we are interested in the population mean of a latent variable X^* not directly observable

$$\theta_0 = \mathbb{E}[X^*].$$

However, we have access to a primary data set of random variable X (possibly mismeasured) and an auxiliary data set of random variables $\{X_A^*, X_A\}$. Under strong ignorability assumption that conditional densities $f_{X_A^*|X_A=x} = f_{X^*|X=x}$ for all $x \in \mathcal{X}$, θ_0 can be expressed as

$$\theta_0 = \mathbb{E}[\gamma_0(X)] = \mathbb{E}[\gamma_0^A(X)],$$

where $\gamma_0(x) = \mathbb{E}[X^*|X = x]$ and $\gamma_0^A(x) = \mathbb{E}_A[X_A^*|X_A = x]$, with $\mathbb{E}_A[\cdot\,]$ denoting the expectation operator for auxiliary data set. Let f_X and f_{X_A} be the marginal densities of X and X_A, respectively. We can further write

$$\theta_0 = \mathbb{E}_A\left[\gamma_0^A(X)\frac{f_X(X)}{f_{X_A}(X)}\right],$$

so RR is identified as $\alpha_0(x) = \frac{f_X(x)}{f_{X_A}(x)}$.

Example 2.7. Expected conditional covariance and partly linear model

Given random variables $\{Y, W, X\}$, we are interested in

$$\theta_0 = \mathbb{E}[cov(Y, W|X)] = \mathbb{E}\left[W(Y - \mathbb{E}[Y|X])\right],$$

which has been studied in Robins et al. (2008); Newey and Robins (2018). Note θ_0 is composed of two parts. The first part, $\mathbb{E}[WY]$, is a moment of observables and can be directly estimated by sample averages. The second part, $-\mathbb{E}[W\mathbb{E}[Y|X]]$, is an average regression functional evaluated at $\gamma_0(x) = \mathbb{E}(Y|X = x)$. By Law of Iterated Expectations (hereafter LIE)

$$-\mathbb{E}\left[W\mathbb{E}[Y|X]\right] = -\mathbb{E}\left[\mathbb{E}[W|X]\mathbb{E}[Y|X]\right],$$

identifying RR as $\alpha_0(x) = -\mathbb{E}[W|X = x]$. θ_0 can also be motivated from a partly linear projection, in a fashion similar to the seminal work of Robinson (1988). Indeed, consider a partly linear model

$$Y = \beta_0 W + h(X) + U, \tag{2.12}$$

where $h(x)$ is an unknown function, U is the error term such that $\mathbb{E}[UW] = 0$ and $\mathbb{E}[Uf(X)] = 0$ for any $f \in L_{\mathbb{P},2}$. Parameter θ_0 is then the numerator of the coefficient $\beta_0 = \frac{\mathbb{E}[(Y - \mathbb{E}[Y|X])W]}{\mathbb{E}[(W - \mathbb{E}[W|X])W]}$.

2.3 Minimax learning in penalized series space

2.3.1 Identification: three ways

There are three method-of-moment approaches to identify θ_0 in the literature

$$\begin{aligned}
\theta_0 &= \mathbb{E}[m(Z, \gamma_0(X))]; && \text{(DP)} && (2.13)\\
&= \mathbb{E}[\alpha_0(X)Y]; && \text{(BP)} && (2.14)\\
&= \mathbb{E}[m(Z, \gamma_0(X)) + \alpha_0(X)(Y - \gamma_0(X))]. && \text{(DR)} && (2.15)
\end{aligned}$$

(2.13) directly comes from the original definition of θ_0. It contains one nuisance function γ_0, a conditional expectation of observables. Estimation of θ_0 through (2.13) is straightforward: first estimate γ_0 by any nonparametric method, followed by a simple "plug-in" procedure to construct a sample analogue of (2.13). Thus call any estimator for θ_0 based on (2.13) a "Direct Plug-in" (DP) learner.

To see (2.14), simply apply (2.6) followed by LIE

$$\theta_0 = \mathbb{E}[\alpha_0(X)\gamma_0(X)] = \mathbb{E}[\alpha_0(X)Y].$$

In order to estimate θ_0 via (2.14), we also only need to estimate one nuisance function: α_0 (RR). However, since α_0 often is not a conditional expectation, estimation of α_0 is more involved. To mark the link with statistics literature, call any estimator for θ_0 based on (2.14) a "Balancing Plug-in" (BP) learner. Indeed, α_0 can also be viewed as the weight to balance outcome variable Y in population.

The third identifying moment condition for θ_0, (2.15), has been popular in recent literature. It extends (2.13), taking account of adjustment term $\alpha_0(x)(y - \gamma_0(x))$, which is mean zero by LIE. The form of (2.15) originates from semiparametric theory. For model (2.3), it is well known that any semiparametric efficient estimator $\hat{\vartheta}$, if exists, should have an asymptotic linear form $\sqrt{n}(\hat{\vartheta} - \theta_0) = \sqrt{n}\mathbb{E}_n\phi + o_p(1)$, where ϕ is the influence function defined as

$$\phi(y, z, x) = m(z, \gamma_0(x)) + \alpha_0(x)(y - \gamma_0(x)) - \theta_0. \tag{2.16}$$

By property of influence function, $\mathbb{E}\phi = 0$, which yields (2.15). Note (2.15) contains two nuisance functions: both γ_0 and α_0. Therefore computationally it is the most costly one. However, since (2.15) comes from the influence function, it has a "small bias" property: small changes around either of the two nuisance functions will not adversely affect estimating θ_0 too much. Thus (2.15) is potentially more "robust" to mistakes made in the first step estimation of nuisance functions, aka "Doubly Robust" (DR). Note (2.15) has also been called "locally robust" or "Neyman orthogonal".

Based on the identification strategies discussed above, the next subsection introduces a minimax methodology of calibrating α_0. This calibrated RR can be flexibly used to construct either BP or DR learners for θ_0. When $\frac{k}{n} \to c < 1$, I will show a BP learner usually suffices. These new results complement DP and DR learners studied in Newey and Robins (2018). When $\frac{k}{n} \to \infty$, plug-in learner of either kind usually does not achieve $\sqrt{n}$ normality easily. Hence naturally a DR learner is considered for such high dimensional problems.

2.3.2 Calibration of the Riesz Representer

Given a sample $\{X_i, Y_i, Z_i\}_{i=1}^n$ of size n, a vector of k basis functions $p(x) = (p_1(x), p_2(x) \dots p_k(x))'$, and a $k \times k$ matrix W_n, propose to calibrate α_0 in a penalized series space with a minimax criterion

$$\tilde{\alpha} = \arg\min_{\alpha \in \Theta_n} \left\{ \begin{array}{ccc} \sup_{g \in \mathcal{H}_{W_n}} \{\mathbb{E}_n[\alpha(X)g(X) - m(Z, g(X))]\}^2 & + & \mathcal{P}_n(\alpha(X)) \\ \text{(minimax)} & & \text{(quality control)} \end{array} \right\}, \tag{2.17}$$

where

$$\begin{aligned} \Theta_n &= \left\{\alpha = a'p : a \in \mathbb{R}^k\right\}, && \text{(series space)} \\ \mathcal{H}_{W_n} &= \left\{g = \beta' W_n p : \beta \in \mathbb{R}^k, \|\beta\| \leq 1\right\}, && \text{(calibration space)} \end{aligned}$$

$\mathbb{E}_n[\cdot] = \frac{1}{n}\sum_{i=1}^{n}(\cdot)$ is the sample average, and $\mathcal{P}_n(\alpha(X)) : \Theta_n \to [0, +\infty)$ is a sample based penalty function of α. Θ_n is a standard series space used to approximate RR, while the role of $\mathcal{H}_{W_n}$ is mainly to facilitate minimax calibration (hence called calibration space). Penalty function $\mathcal{P}_n(\alpha(X))$ provides additional regularization to avoid overfitting and to improve finite sample as well as asymptotic performance. It is common to use a Tikhonov penalty, for example, $\mathcal{P}_n(\alpha(X)) = \mathbb{E}_n\alpha^2(X)$. Additional l_1 penalization can also be included in $\mathcal{P}_n(\alpha(X))$ to accommodate high dimensional situations. Minimax exercise (2.17) can be further motivated by two intuitive arguments.

Intuition 1: $\tilde{\alpha}$ exploits a minimax property of RR in population

Observe that $\alpha_0 \in \Theta_\alpha$ is the unique solution of the following minimax problem

$$\min_{\alpha \in \Theta_\alpha} \left[\sup_{g \in L_{\mathbb{P},2}} \{\mathbb{E}[\alpha(X)g(X) - m(Z, g(X))]\}^2 \right].$$

Indeed, definition of RR in (2.5) implies

$$\sup_{g \in L_{\mathbb{P},2}} \{\mathbb{E}[\alpha_0(X)g(X) - m(Z, g(X))]\}^2 = 0,$$

while for any $\alpha \neq \alpha_0$, it must be

$$\sup_{g \in L_{\mathbb{P},2}} \{\mathbb{E}[\alpha(X)g(X) - m(Z, g(X))]\}^2 > 0$$

by uniqueness of α_0. Thus, $\tilde{\alpha}$ can be loosely interpreted as a sample analogue of a population minimax problem, but in computationally feasible spaces Θ_n and $\mathcal{H}_{W_n}$ and subject to a "quality control" term $\mathcal{P}_n(\alpha(X))$.

Intuition 2: minimax learning directly controls key remainder "bias" term

Suppose an econometrician developed some learner $\bar{\theta}$ with generic first step estimators $\bar{\gamma}$ and $\bar{\alpha}$ for γ_0 and α_0, respectively. As long as either BP or DR approach is involved, $\bar{\theta}$ will necessarily admit the following asymptotic linear structure

$$\sqrt{n}(\bar{\theta} - \theta_0) = \sqrt{n}\mathbb{E}_n\phi + R_1 + R_2,$$

where the leading term can be shown to be asymptotically normal with ϕ defined in (2.16), R_1 and R_2 are remainder terms

$$R_1 = \sqrt{n}\mathbb{E}_n\left[\bar{\alpha}(X)g_\gamma(X) - m(Z, g_\gamma(X))\right], \tag{2.18}$$

$$R_2 = \sqrt{n}\mathbb{E}_n\left[\left(\bar{\alpha}(X) - \alpha_0(X)\right)e\right], \tag{2.19}$$

and

$$e = Y - \gamma_0(X). \tag{2.20}$$

The form of g_γ depends on which method is used

$$g_\gamma = \begin{cases} \gamma_0, & \text{(BP approach)} \\ \gamma_0 - \bar{\gamma}. & \text{(DR approach)} \end{cases}$$

(2.19) is concerned with estimation noise from $\bar{\alpha}$. It is mean zero under some conditions and can be usually controlled quite effectively. Even if we do not control (2.19), $\sqrt{n}$ normality turns out to be still possible. On the other hand, (2.18) involves the interplay between unknown functions g_γ and $\bar{\alpha}$. It is not mean zero and much difficult to deal with. Suppose g_γ can be approximated by some basis functions. $\tilde{\alpha}$ then tries to directly control (2.18) by minimizing its largest possible realization in a carefully constructed small ball $\mathcal{H}_{W_n}$. With this respect, $\tilde{\alpha}$ is a minimax estimator with criterion directly targeting (2.18), the key "remainder bias" of any learner for θ_0.

2.3.3 Implementation

The solution of minimax exercise (2.17) can be found straightforwardly, thanks to the simplistic structures of Θ_n and $\mathcal{H}_{W_n}$.

Proposition 2.1. *Given W_n, for each $\alpha \in \Theta_n$,*

$$\sup_{g \in \mathcal{H}_{W_n}} \left\{\mathbb{E}_n[\alpha(X)g(X) - m(Z, g(X))]\right\}^2 = \left\|W_n\mathbb{E}_n[e_\alpha(Z)]\right\|^2,$$

where $e_\alpha(z) = m(z, p(x)) - \alpha(x)p(x)$.

Proof of Proposition 2.1 is simple and left to Appendix B. By Proposition 2.1, $\tilde{\alpha}$ defined in (2.17) has an equivalent weighted minimum distance representation

$$\tilde{\alpha} = \arg\min_{\alpha \in \Theta_n} Q_n(\alpha), \tag{2.21}$$

where

$$Q_n(\alpha) = \underbrace{\mathbb{E}_n[e_\alpha(Z)]' \mathcal{W}_n \mathbb{E}_n[e_\alpha(Z)]}_{\text{minimum distance}} + \underbrace{\mathcal{P}_n(\alpha(X))}_{\text{series penalization}}, \tag{2.22}$$

with $\mathcal{W}_n = W_n' W_n$ symmetric and positive semidefinite by construction and should satisfy some conditions stated below in Assumptions L3, M2 or H3. The first half of the criterion function (2.22) is a $\mathcal{W}_n$ weighted norm of $\mathbb{E}_n[e_\alpha(Z)]$, and the second half is an empirical penalization function of α. Such a structure is reminiscent of the PSMD estimator studied in Chen and Pouzo (2012, 2015a). However, the motivation in this chapter is inherently different. Related to e_α, define

$$e^R(z) = m(z, p(x)) - \alpha_0(x)p(x), \tag{2.23}$$

an object important for later asymptotic analysis. Indeed, by definition of RR, $\mathbb{E}e^R = \mathbf{0}$. The role of e^R is similar to that of e defined in (2.20) for a regression. Thus call e^R the "Riesz error" of the model.

2.3.4 Construction of minimax learners

2.3.4.1 When $\frac{k}{n} \to c$ for some $c \in [0, 1)$

This includes a low dimensional case when $c = 0$, and a moderately high dimensional scenario when $c \neq 0$ but smaller than one. Construct a minimax BP learner for θ_0 as follows

$$\hat{\theta}_{BP} = \mathbb{E}_n[\tilde{\alpha}(X)Y], \tag{2.24}$$

where $\tilde{\alpha}$ is calibrated from (2.17) by setting

$$\mathcal{P}_n(\alpha(X)) = \lambda_1 \mathbb{E}_n \alpha^2(X), \tag{2.25}$$

and $\lambda_1 > 0$ is a coefficient practically determined by the econometrician. The ridge style penalization (2.25) is simple to analyze asymptotically, and has been popular in balancing literature to control how volatile $\tilde{\alpha}$ should be in sample.[1] By the equivalent representation result in Proposition 2.1, $\tilde{\alpha}$ admits an explicit

[1] I suspect the theory can be applied to other penalization functions that display quadratic behaviors, for example, strongly convex functions.

formula

$$\tilde{\alpha}(x) = p(x)'(\hat{G}\mathcal{W}_n\hat{G} + \lambda_1\hat{G})^{-}\hat{G}\mathcal{W}_n\hat{P}, \tag{2.26}$$

where

$$\hat{G} = \mathbb{E}_n[p(X)p(X)'], \quad \hat{P} = \mathbb{E}_n[m(Z, p(X))], \tag{2.27}$$

and $(\cdot)^{-}$ denotes the Moore–Penrose inverse. Indeed, objective function in (2.21) becomes continuously differentiable and strongly convex. First order approach applies. As a result, $\widehat{\theta}_{BP}$ is computationally competitive compared to other methods that carry out minimax exercises in infinite dimensional spaces, which often require numerical optimization.

Special case When $\lambda_1 = 0$ and $\mathcal{W}_n = I$, (2.26) reduces to

$$\hat{\alpha}(x) = p(x)'(\hat{G}\hat{G})^{-}\hat{G}\hat{P}. \tag{2.28}$$

In theory, (2.28) equals the estimator proposed in Newey and Robins (2018)[2]

$$\hat{\alpha}(x) = p(x)'\hat{G}^{-}\hat{P}. \tag{2.29}$$

However, (2.28) and (2.29) can be calculated quite differently by any software due to inherent computational constraint. Such computational discrepancy seems more salient when k becomes large. This fact motivates a new measure of design uncertainty for $\hat{G}$. See Appendix B for more discussions on this issue.

2.3.4.2 When $\frac{k}{n} \to \infty$

This high dimensional situation imposes two major challenges to this minimax methodology. First, magnitude of R_1 will usually be substantially larger for plug-in learners; Second, both $\mathcal{H}_{W_n}$, the calibration space, and Θ_n, the standard series space, grow too quickly and become increasingly complicated. While the first challenge can be addressed by using a DR approach, resolving the second requires more careful control of both Θ_n and $\mathcal{H}_{W_n}$. This is achieved by manipulating $\mathcal{P}_n(\alpha(X))$ and W_n under framework (2.17): The dimension of Θ_n can be effectively reduced by incorporating a lasso term in $\mathcal{P}_n(\alpha(X))$, and $\mathcal{H}_{W_n}$ shall be controlled "small enough" through selecting a suitable (possibly data-driven) weight matrix.[3]

To be specific, a minimax DR learner for θ_0 can be constructed as follows

[2]Since by Corollary 20.5.5 in Harville (1998), $\hat{G}^{-} = (\hat{G}\hat{G})^{-}\hat{G}$.

[3]Another potential way to proceed is to consider the same minimax exercise in an l_1 ball: $\mathcal{H}_{\tilde{\lambda}} = \left\{ g = \beta' p : \beta \in \mathbb{R}^k, \|\beta\|_1 \leq \tilde{\lambda} \right\}$, with $\tilde{\lambda}$ as some small number specified by the researcher. I leave this for future research.

$$\hat{\theta}_{DR} = \mathbb{E}_n[m(Z, \hat{\gamma}(X)) + \tilde{\alpha}(X)(Y - \hat{\gamma}(X))], \tag{2.30}$$

where $\hat{\gamma}$ is any preliminary estimator of γ_0 possibly derived from some machine learning algorithm, and $\tilde{\alpha}$ is still the solution of (2.17), but with an elastic net style penalty

$$\mathcal{P}_n(\alpha(X)) = \lambda_1 \mathbb{E}_n \alpha^2(X) + \lambda_2 \|a\|_1 , \tag{2.31}$$

and λ_1, λ_2 are some penalty loadings.[4] Call $\tilde{\alpha}$ derived with (2.31) a Generalized Minimax Elastic Net (GMEN) learner. By Proposition 2.1, GMEN learner $\tilde{\alpha}$ also has a weighted minimum distance representation: $\tilde{\alpha}(x) = \tilde{a}' p(x)$, where

$$\tilde{a} = \arg\min_{a \in \mathbb{R}^k} \left\{ (\hat{G}a - \hat{P})' \mathcal{W}_n (\hat{G}a - \hat{P}) + \lambda_1 a' \hat{G} a + \lambda_2 \|a\|_1 \right\}, \tag{2.32}$$

where $\hat{G}$ and $\hat{P}$ are defined in (2.27). Solution of (2.32) can be found by fast algorithms, for instance, Orthant-Wise Limited-memory Quasi-Newton (OWL-QN) method.

[4] In fact, the theory can be relatively straightforwardly extended to cover cases with a penalty function $\mathcal{P}_n(\alpha(X)) = \lambda_1 a' \Gamma' \Gamma a + \lambda_2 \|a\|_1$ for some matrix Γ. This Tikhonov plus lasso style penalty could potentially perform even better if Γ is selected to reflect underlying data structure. Also see discussion at the end of Theorem 3.5 for more details.

Chapter 3

Minimax learning for average regression functionals: theory

3.1 Theory: minimax BP

Assumption O.

1. *[Data] Random vectors* $(Y_i, Z_i', X_i')', i = 1 \dots n$, *are independently and identically distributed (hereafter iid) for each* n *and satisfy (2.2);*

2. *[Strong exogeneity]* $\mathbb{E}[e|Z] = 0$, $\mathbb{E}[e^2|Z] < \infty$ *almost surely, where* e *is defined in (2.20);*

3. *[Functional restriction] Function* $m(z, \cdot)$ *is linear in the sense of (2.4), and is bounded such that for any* $\gamma \in L_{\mathbb{P},2}$, $\mathbb{E}[m^2(Z, \gamma(X))] \leq C\mathbb{E}[\gamma^2(X)] < \infty$ *for some constant* C.

Assumption O imposes some basic requirements on the data structure, and applies throughout the rest of the asymptotic analyses. O-(1) flexibly accommodates situations when d_X, the dimension of X, is fixed or growing.[1] O-(2) restricts the behavior of first two conditional moments of e defined in (2.20). Exogeneity condition $\mathbb{E}[e|Z] = 0$ is useful to simplify asymptotics involving remainder term R_2, but is not automatically guaranteed at the generality of (2.3). It is satisfied if

1. $Z = X$, which covers Examples 2.1-2.6, or,

2. Conditional on X, e is independent of $(Z \setminus X)$, the subvector of Z excluding X.

[1] If the dimension of X is growing as $n \to \infty$, d_X should be understood as $d_{X,n}$.

It can also be avoided if some cross-fitting scheme is used so that $\tilde{\alpha}$ is calibrated only using data from a different random sample, which can be potentially applied to Example 2.7.[2] Error term e is additionally assumed to have a finite conditional variance, which is standard but might be weakened by imposing higher unconditional moments for e and X. See for example, Hansen (2015). O-(3) is sufficient for the existence of RR and for finite semiparametric variance bound but is not necessary. This assumption is also key to get consistency of $\tilde{\alpha}$ under weak conditions. Otherwise, functional form of $m(z,\cdot)$ might deteriorate estimation as well as approximation. If this happens, $m(z,\cdot)$ is not continuous, and O-(3) can be modified such that $\mathbb{E}[m^2(Z,\gamma(X))] \leq d_k \mathbb{E}[\gamma^2(X)]$ where d_k is growing as a function of k, similar to Assumption 6 in Newey and Robins (2018). This allows rates slower than the ones presented in the main theorems below.

3.1.1 Asymptotic normality and semiparametric efficiency when $\frac{k}{n} \to 0$

Assumption L1.

1. *[Series space] All eigenvalues of* $\mathbb{E}[p(X)p(X)']$ *are bounded from above and away from zero;*

2. *[Series approximation] There exist some vectors* β_b, a_b *and some numbers* $\mathbf{r}_{\gamma_0}, \mathbf{r}_{\alpha_0}$ *such that*

$$\sup_{x\in\mathcal{X}} |\gamma_0(x) - \beta_b' p(x)| = \mathbf{r}_{\gamma_0}; \quad \sup_{x\in\mathcal{X}} |\alpha_0(x) - a_b' p(x)| = \mathbf{r}_{\alpha_0}.$$

Discussion of Assumption L1

L1 is a prerequisite of series approximation. L1-(1) is a common eigenvalue assumption, typically satisfied in a low dimensional environment.[3] L1-(2) imposes some mild restrictions on approximation quality of series space Θ_n, measured by magnitude of $\mathbf{r}_{\gamma_0}$ and $\mathbf{r}_{\alpha_0}$. If model is correctly specified, $\mathbf{r}_{\gamma_0}$ and $\mathbf{r}_{\alpha_0}$ should go to 0 as $n \to \infty$. When γ_0 and α_0 are within a Hölder class of smoothness order s

$$\mathbf{r}_{\gamma_0} = k^{-\eta_\gamma}, \quad \mathbf{r}_{\alpha_0} = k^{-\eta_\alpha},$$

[2]It is beyond the scope of this chapter to formally establish asymptotics under this case, but results in Newey and Robins (2018) are useful.

[3]For example, it holds if $\mathcal{X} = [0,1]^{d_X}$ with d_X fixed and density of X is bounded away from zero and from above. See also Newey (1988) and Proposition 2.1 in Belloni et al. (2015) for more discussions.

where η_γ and η_α are non negative functions of the following parameters: smoothness s, X's dimension d_X, as well as properties of basis functions p. See DeVore and Lorentz (1993); Newey (1997); Chen (2007) for more details on approximation results.

Following Newey (1997), the following object is important for asymptotic analysis

$$\xi_k = \sup_{x \in \mathcal{X}} \|p(x)\| .$$

Let $f \in \Theta_f$ where Θ_f is some function class. Denote $\mathcal{L}_n f$ as the least square projection of f onto Θ_n. Thus define the so called Lebesgue integral as

$$\ell_k = \sup \left(\frac{\|\mathcal{L}_n f\|_{\mathbb{P},\infty}}{\|f\|_{\mathbb{P},\infty}} : \|f\|_{\mathbb{P},\infty} \neq 0, f \in \Theta_f \right),$$

which has been used in Huang (2003); Belloni et al. (2015); Chen and Pouzo (2015b). For certain basis functions, ℓ_k exploits stability relations between $\|\cdot\|_{\mathbb{P},2}$ and $\|\cdot\|_{\mathbb{P},\infty}$ norms of projections. For more background materials on series estimation and least square projection, see Appendix C.1.

Assumption L2. *As $k \to \infty$ and $n \to \infty$:*

1. *[Dimension restriction] $\frac{\xi_k^2 \log k}{n} = o(1)$;*
2. *Either of the following two conditions holds:*
   (a) *[Approximation quality] $\mathbf{r}_{\gamma_0} = O(\frac{1}{\sqrt{n}})$, $\ell_k \mathbf{r}_{\gamma_0} = O(1)$, $\mathbf{r}_{\alpha_0} = o(1)$;*
   (b) *[Approximation quality] $\sqrt{n}\mathbf{r}_{\alpha_0}\mathbf{r}_{\gamma_0} = o(1)$, $\ell_k \mathbf{r}_{\alpha_0}\mathbf{r}_{\gamma_0} = o(1)$, $\mathbf{r}_{\alpha_0} = o(1)$, [Model complexity] $(\ell_k + 1)\mathbf{r}_{\gamma_0}\left(\sqrt{k \log \xi_k} + \frac{k\xi_k \log \xi_k}{\sqrt{n}}\right) = O(1)$.*

Discussion of Assumption L2

L2 lists key dimension and approximation requirements and is of first order importance. Dimension restriction imposed by L2-(1) is knowingly the weakest possible in the literature. Suppose we choose spline or wavelet series, $\xi_k = \sqrt{k}$ (Newey 1997). L2-(1) reduces to $\frac{k \log k}{n} \to 0$. Under this condition, $\tilde{\alpha}$ will be consistent in $\|\cdot\|_{\mathbb{P},2}$ norm.

L2-(2) is the other factor that crucially determines asymptotic growth of k. For spline and wavelet series, $\ell_k = O(1)$ (Huang, 2003; Chen and Pouzo, 2015b). So if γ_0 has a fast approximation rate $O(\frac{1}{\sqrt{n}})$, L2-(2)-(a) can be invoked with no approximation rate condition imposed on α_0. Thus, it seems worthwhile to give priority to approximating γ_0 when selecting basis functions. Let $\mathbf{r}_{\gamma_0} = k^{-\eta}$ for some $\eta > 0$ and $k = n^r$ for some $r > 0$. Then L2-(2)-(a) and L2-(1) together imply r is allowed in the range $[\frac{1}{2\eta}, 1)$ (ignoring $\log k$ term), as long as $\frac{1}{2} < \eta < 1$.

L2-(2)-(b) can be viewed as an asymptotic refinement of its counterpart (a). It relaxes approximation quality but requires a new model complexity term. Thus a trade-off exists. In fact, approximation quality in L2-(2)-(b) plus L2-(1) forms the minimal condition needed for $\sqrt{n}$ normality and semiparametric efficiency so far in the literature, see Newey and Robins (2018). Also note $\ell_k \mathbf{r}_{\alpha_0} \mathbf{r}_{\gamma_0} = o(1)$ is a very weak condition, trivially satisfied as long as $\sqrt{n}\mathbf{r}_{\alpha_0}\mathbf{r}_{\gamma_0} = o(1)$ and $\frac{\ell_k}{\sqrt{n}} < \infty$. Since no cross fitting is used, model complexity term arises naturally from a stochastic equicontinuity term (or intuitively, "own observation bias") controlled by empirical process theory.[4] If cross fitting is employed, model complexity term can be further relaxed.

It is useful to compare L2 with similar conditions in Newey and Robins (2018). Suppose $\ell_k = O(1), \xi_k = O(\sqrt{k})$ and ignore log terms:

1. If γ_0 is very smooth in the sense that $\eta \geq 1$, model complexity term becomes of order $\mathbf{r}_{\gamma_0} k \sqrt{\frac{k}{n}}$, which is finite trivially so long as $\frac{k}{n} \to 0$. L2 reduces to the minimal rate requirement so far in literature: $\sqrt{n}\mathbf{r}_{\alpha_0}\mathbf{r}_{\gamma_0} = o(1)$ and $\frac{k}{n} = o(1)$.

2. If γ_0 is ordinarily smooth in the sense that $\eta > \frac{1}{2}$, L2-(2)-(a) can be invoked to get (close to but not minimal) condition: $\frac{k}{n} = o(1), \mathbf{r}_{\alpha_0} = o(1)$ and $\mathbf{r}_{\gamma_0} = O(\frac{1}{\sqrt{n}})$. Cross fitted DP and doubly cross fitted DR estimators in Newey and Robins (2018) can satisfy minimal conditions in this case.

3. If γ_0 is not smooth enough such that $\eta \leq \frac{1}{2}$, we need both $\mathbf{r}_{\gamma_0}\sqrt{k}\left(1 + \frac{k}{\sqrt{n}}\right) = O(1)$ and $\sqrt{n}\mathbf{r}_{\alpha_0}\mathbf{r}_{\gamma_0} = o(1)$. Hence k is required to grow slower than $\frac{k}{n} = o(1)$. Additional rate conditions will have to be imposed. Under this case, neither cross fitted DP or doubly cross fitted DR estimator in Newey and Robins (2018) can meet the minimal requirement in general either. But they do have smaller remainder terms due to cross fitting structure.

Assumption L3. *As $k \to \infty$ and $n \to \infty$:*

1. *[Diverging α_0] $(\|\alpha_0\|_{\mathbb{P},\infty} \wedge \ell_k)\mathbf{r}_{\gamma_0} = o(1)$;*

2. *[Weight matrix] $W_n' W_n \hat{G}$ is symmetric and $W_n' W_n - I$ is positive semidefinite;*

3. *[Penalty] $\lambda_1 = o(\frac{1}{n})$.*

[4] I suspect this term might be further weakened. I leave this for future research.

Discussion of Assumption L3

L3 is concerned with a set of regularity conditions. L3-(1) allows diverging α_0, different from a literature that usually assumes boundedness. If $\ell_k = O(1)$, $\|\alpha_0\|_{\mathbb{P},\infty}$ is allowed to grow with no rate requirement, as long as $\mathbf{r}_{\gamma_0} = o(1)$. Otherwise, additional restrictions on $\|\alpha_0\|_{\mathbb{P},\infty}$ might be needed.

L3-(2) is a novel condition on the choice of W_n, which affects convergence rate of $\hat{\theta}_{BP}$. Symmetry of $W_n'W_n\hat{G}$ allows tractable asymptotic behavior of $\tilde{\alpha}$. It requires $\hat{G}$ and $W_n'W_n$ to be commutable. A sufficient condition is that they share the same eigenspace, which can be constructed easily. To achieve $\sqrt{n}$ normality of $\hat{\theta}_{BP}$, all eigenvalues of $W_n'W_n$ needs to be bounded away from 1. Interestingly, this is stricter than what is required for $\|\cdot\|_{\mathbb{P},2}$ consistency of $\tilde{\alpha}$, which only needs $\lambda_{\min}(W_n'W_n)$ to be bounded away from 0. (See Lemma C.16.)

L3-(3) specifies the correct asymptotic order of λ_1. In practice it often suffices to select a small penalty. Other data-driven methods, such as cross-validation or Lepski's method, may be used.

Theorem 3.1. *[Approximate minimax balancing] If O, L1, L2 and L3 hold, then $\hat{\theta}_{BP}$ defined in (2.24) admits*

$$\sqrt{n}\left(\hat{\theta}_{BP} - \theta_0\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[m(X_i, \gamma_0(X_i)) + \alpha_0(X_i)(Y_i - \gamma_0(X_i)) - \theta_0\right] + o_p(1),$$

and

$$\sqrt{n}\left(\hat{\theta}_{BP} - \theta_0\right) \xrightarrow{d} N(0, \Omega),$$

where $\Omega = \mathbb{E}\left[m(X, \gamma_0(X)) + \alpha_0(X)(Y - \gamma_0(X)) - \theta_0\right]^2$.

This is the first main result of the chapter, sharing a spirit similar to the "approximate balancing" literature (For example, Zubizarreta, 2015; Athey et al., 2018). Indeed, $\hat{\theta}_{BP}$ tries to balance a set of covariates only "approximately", trading off between a minimax criterion and a ridge style penalty term. Theorem 3.1 seems to be the first result that establishes semiparametric efficiency for BP methods, without imposing strong convergence rate conditions for nuisance parameters, sample splitting or DR moment conditions. In fact, it seems that as long as γ_0 is estimated by standard series method, DR learner will hardly improve Theorem 3.1. Theorem 3.1 can be extended to the following relevant case.

Corollary 3.1. *[Exact minimax balancing] Set $\lambda_1 = 0$. Then, Theorem 3.1 still stands for any positive definite $W_n'W_n$ if O, L1, L2 and L3-(1) hold.*

Corollary 3.1 pursues a strategy to "exactly balance" all covariates in the minimax sense. Indeed, when $\lambda_1 = 0$ and $\frac{\xi_k^2 \log k}{n} \to 0$, minimax BP learner using

any positive definite $W_n'W_n$ is the same as $\hat{\alpha}$ wpa1. Under this scenario minimax criterion in (2.17) can achieve zero at $\hat{\alpha}$ wpa1.

Overall, upon suitable choice of weight matrix, "approximate minimax balancing" will not do worse than "exact minimax balancing" asymptotically, but might gain finite sample performance due to additional penalization. However, the price to pay for "approximate minimax balancing" is a more limited choice of weight matrix. Simulation shows that setting $W_n = I$ with a small λ_1 can deliver desirable result under moderately high dimensions.

3.1.2 Consistent estimation of variance

Since

$$\begin{aligned}\Omega &= Var\left[m(Z,\gamma_0(X)) + \alpha_0(X)(Y-\gamma_0(X)) - \theta_0\right] \\ &= \left|\mathbb{E}\left[m(Z,\gamma_0(X)) + \alpha_0(X)(Y-\gamma_0(X))\right]^2 - \theta_0^2\right|,\end{aligned}$$

a natural plug-in estimator for Ω admits

$$\hat{\Omega} = \left|\mathbb{E}_n\left[m(Z,\hat{\gamma}^s(X)) + \tilde{\alpha}(X)(Y-\hat{\gamma}^s(X))\right]^2 - \hat{\theta}_{BP}^2\right|, \tag{3.1}$$

where $\hat{\gamma}^s(x) = p(x)'\hat{G}^-\mathbb{E}_n[p(X)Y]$ is the standard series estimator for γ_0. Confidence intervals can be constructed accordingly based on $\hat{\theta}_{BP}$ and standard error $\sqrt{\hat{\Omega}}$.

Theorem 3.2. *Suppose Theorem 3.1 holds. In addition:*

1. $\|\hat{\gamma}^s - \gamma_0\|_{\mathbb{P},\infty} = o_p(1)$;
2. *For some* $\delta > 0$, $\mathbb{E}\left[|e|^{2+\delta}\right] < \infty$ *and* $\xi_k^{\frac{2+\delta}{\delta}}\sqrt{\frac{\log k}{n}} = o(1)$.

Then: $\hat{\Omega} \xrightarrow{p} \Omega$.

Consistency of $\hat{\Omega}$ demands stronger conditions due to presence of second moment. Assumption (1) stipulates that $\hat{\gamma}^s$ should be consistent in sup norm, which can be verified by more primitive conditions. For example, Theorem 4.3 in Belloni et al. (2015) and Lemma 2.4 in Chen and Pouzo (2015b) both establish optimal sup norm convergence for $\hat{\gamma}^s$ under weak conditions, allowing $\frac{k}{n} \to 0$ up to log terms. It is also possible to relax (1) by imposing higher moment conditions for basis functions, see Hansen (2015). (2) sees a trade-off between existence of higher moments for e and growth rate restrictions on k. This condition is mainly used to show convergence of random matrix involving e^2. If $\mathbb{E}[e^4] < \infty$, (2) translates to $\xi_k^2\sqrt{\frac{\log k}{n}} \to 0$, stronger than L2-(1). Both (1) and (2) can be further weakened under cross fitting set-up. Notice sup norm consistency of $\tilde{\alpha}$ is not required.

3.1.3 $\sqrt{n}$ normality when $\frac{k}{n} \to c < 1$

To achieve semiparametric efficiency, $\hat{\theta}_{BP}$ requires at least $\frac{k}{n} \to 0$ (up to log terms). Such a condition currently seems unavoidable as it stems from the essential need to consistently estimate α_0. However, this subsection shows that $\sqrt{n}$ normality can sometimes still be established even when $\tilde{\alpha}$ is not consistent. This relies on two key elements: relatively faster approximation of γ_0 (Assumption M1) and a central limit theorem around $\mathbb{E}_n[\gamma_0(X)]$ (Assumption M2).

Assumption M1. *As $k \to \infty$ and $n \to \infty$:*

1. *[Invertibility and dimension restriction]* $\mathbb{P}\{\lambda_{\min}(\hat{G}) > 0\} \to 1$, $\frac{\xi_k^2 \log k}{n} \to c_1 < \infty$;
2. *[Approximation quality]* $\mathbf{r}_{\gamma_0} = o\left(\frac{1}{\sqrt{n}}\right)$, $\ell_k \mathbf{r}_{\gamma_0} = O(1)$, $\mathbf{r}_{\alpha_0} = O(1)$.

Compared to L2, dimension restriction is relaxed to M1-(1), which guarantees positive definiteness of $\hat{G}$ as well as asymptotic boundedness of $\lambda_{\max}(\hat{G})$. M1-(1) necessarily implies $\frac{k}{n} < 1$, but it does not need to vanish to 0 (aka alternative $\frac{k}{n} \to c$ asymptotics). Under this scenario, $\tilde{\alpha}$ is not consistent. M1-(2) requires γ_0 to be approximated sufficiently fast such that $\mathbf{r}_{\gamma_0} = o\left(\frac{1}{\sqrt{n}}\right)$. However, α_0 does not need to be correctly specified as long as $\mathbf{r}_{\alpha_0} = O(1)$ and $\ell_k \mathbf{r}_{\gamma_0} = O(1)$. M1 is hard to relax at this moment.

Assumption M2. *[Central limit theorem]*

1. $\mathbb{E}_n \hat{\alpha}(X)^2 = O_p(1)$, *where $\hat{\alpha}(X)$ is defined in (2.29);*
2. $\frac{\max_i |\tilde{\alpha}(X_i)|}{\sqrt{n}} = o_p(1)$;
3. $\frac{\lambda_1^2}{\lambda_{\min}^2(W_n' W_n)\lambda_{\min}(\hat{G})} = O_p(1)$;
4. $\|\alpha_0\|_{\mathbb{P},2}^2 - \mathbf{r}_{\alpha_0}^2$ *is bounded away from zero;* $\mathbb{E}[e^2|Z]$ *is bounded away from zero almost surely;* $\mathbb{E}\left[|e|^3 |Z\right] < \infty$ *almost surely.*

M2 is mainly in place to allow a central limit theorem. M2-(1) is an asymptotic boundedness condition so that $\mathbb{E}_n \tilde{\alpha}(X)^2 = O_p(1)$. M2-(2) says $\max_i |\tilde{\alpha}(X_i)|$ should grow strictly slower than rate $\sqrt{n}$. M2-(3) is a new condition that characterizes the relation among degree of penalization, high dimensionality of the design and choice of weight matrix. It makes sure inverse asymptotic variance does not explode. M2-(4) is a regularity condition and fairly weak. Condition related to $\|\alpha_0\|_{\mathbb{P},2}^2 - \mathbf{r}_{\alpha_0}^2$ can be interpreted as a constraint on approximation error: $\mathbf{r}_{\alpha_0}^2$ should not be too large compared to $\|\alpha_0\|_{\mathbb{P},2}^2$. The other conditions on e are quite common in the literature. Except M2-(4), the other three conditions currently seem a bit high level. Below I briefly discuss some of their low level conditions before introducing formal distributional result.

Discussion of M2-(1)

There are some simple situations when M2-(1) holds. See Lemma C.25 for detailed discussions. Basically, if O, L1 and M1 are satisfied, M2-(1) holds if either of the following two conditions is true:

1. There exists some $c_2 > 0$ such that $\mathbb{P}\{\lambda_{\min}(\hat{G}) \geq c_2\} \to 1$;
2. There exists some scalar valued function $\varpi(z)$ such that $\hat{P} = \mathbb{E}_n[\varpi(Z)p(X)]$ and $\mathbb{E}[\varpi(Z)^2] < \infty$.

(1) strengthens M1-(1) so that $\left\|\hat{G}^{-1}\right\|$ will also be $O_p(1)$. Lemma C.26 demonstrates a sufficient condition for existence of such c_2: if M1-(1) is satisfied, we additionally need $\sqrt{2c_1} + \frac{1}{3}c_1 < 1$. Simple calculation yields that c_1 should be small enough so that $c_1 < 0.38$. See Lemma C.26 for details. (2) exploits a simple multiplicative structure of $m(z, \cdot)$ to achieve asymptotic boundedness, leveraging nice properties from empirical projection.

Discussion of M2-(2)

M-(2) is currently difficult to relax except in some special cases. Trivially, it is satisfied if $\frac{\xi_k^2 \log k}{n} = o(1)$ (shown by Lemma C.27), which goes back to $\frac{k}{n} \to 0$ regime. Otherwise, study of $\max_i |\tilde{\alpha}(X_i)|$ is more challenging, almost requiring a bound on $\|\tilde{\alpha}\|_{\mathbb{P},\infty}$. Lemma C.28 uses techniques such as symmetrization and entropy integral bound to explore possible primitive conditions for M-(2). Further restrictions on basis functions and the structure of $m(z, \cdot)$ have to be imposed.

Discussion of M2-(3)

M2-(3) says $\lambda_{\min}(\hat{G})$ should not go to zero faster than $\frac{\lambda_1^2}{\lambda_{\min}^2(W_n'W_n)}$. This condition can be trivially satisfied when $\lambda_1 = 0$, a case with no penalization. Otherwise, λ_1 and W_n should be carefully chosen to counteract high dimensionality issues from possibly diminishing eigenvalues of $\hat{G}$. This seems to confirm the conventional wisdom that while "machine learning" methods perform better empirically, tracking their asymptotic distributions can be more strenuous. When $\lambda_1 \neq 0$, several situations allow M2-(3):

1. If c_1 is small enough, it can be shown by Lemma C.26 that $\hat{G}$ has eigenvalues bounded away from zero. M2-(3) then holds for all W_n satisfying L3-(2);
2. Suppose $\lambda_{\min}(\hat{G})$ shrinks at some rate $\kappa_n \to 0$. Then conditional on L3-(2), M2-(3) becomes $\frac{\lambda_1^2}{\kappa_n} = O(1)$. That is, $\lambda_{\min}(\hat{G})$ is allowed to go to zero slower than rate $\frac{1}{n^2}$;

3. If the behavior of $\lambda_{\min}(\hat{G})$ is uncertain, W_n can be chosen to counteract effect from diminishing $\lambda_{\min}(\hat{G})$. For example, conditional on L3-(2), we can set $W_n'W_n$ such that its smallest eigenvalue is $\left(1+\frac{1}{\lambda_{\min}(\hat{G})}\right)$.

Theorem 3.3. *[$\sqrt{n}$ normality without consistency] Assume O, L1, L3, M1 and M2 hold. Then*

$$\sqrt{n}\left[\hat{\theta}_{BP}-\mathbb{E}_n m(Z,\gamma_0(X))\right]=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\tilde{\alpha}(X_i)(Y_i-\gamma_0(X_i))]+o_p(1),$$

and

$$\sqrt{n}V_n^{-1/2}\left[\hat{\theta}_{BP}-\mathbb{E}_n m(Z,\gamma_0(X))\right]\xrightarrow{d}N(0,1),$$

where $V_n=\frac{1}{n}\sum_{i=1}^{n}\{\tilde{\alpha}^2(X_i)\mathbb{E}\left[e_i^2|Z_i\right]\}$ and $N(0,1)$ is a standard normal random variable. Moreover, $\hat{\theta}_{BP}-\theta_0=O_p(\frac{1}{\sqrt{n}})$.

Theorem 3.3 seems to be the first result in the literature that establishes $\sqrt{n}$ normality for BP learners when $\frac{k}{n}\to c<1$. Object $\mathbb{E}_n m(Z,\gamma_0(X))$ is a sample mean that might be of interest as well, see Athey et al. (2018) on conditional average treatment effect. Theorem 3.3 is also an intermediate step toward $\sqrt{n}$ normality of $\hat{\theta}_{BP}$ around population mean θ_0. Some central limit theorem that allows complicated dependence structure could help. Another way to achieve normality is to further relax conditions in M2 and to make use of the growing variance term. Moreover, when $\frac{k}{n}\to c<1$, consistent estimation of V_n is often not trivial. I leave these interesting aspects for future research.

The following corollary provides a counterpart of the exact minimax balancing result in Corollary 3.1. Similarly, the choice of weight matrix is wider without penalization. It might be more attractive for hypothesis testing.

Corollary 3.2. *Set $\lambda_1=0$. Theorem 3.3 still holds for each positive definite $W_n'W_n$ if O, L1, M1, L3-(1), and M2-(1), (2), (4) hold.*

3.2 Theory: minimax DR

This section studies large sample properties of $\hat{\theta}_{DR}$ defined in (2.30) tailored for ultra high dimensional situations. To proceed, let $\alpha_*(x)=a_*'p(x)$, where

$$a_*=\arg\min_{a\in\mathbb{R}^k}\mathbb{E}[\alpha_0(X)-a'p(X)]^2+\lambda_*\left\|a\right\|_1, \tag{3.2}$$

where λ_* is a theoretic penalty coefficient. α_* can be viewed as the best l_1 regularized approximation of α_0. And some entries of a_* will be 0 (thus a_* is

sparse) when λ_* is large enough. Write $u_* = \alpha_0 - \alpha_*$, the approximation error of α_*.

Assumption H1. *[Series space under high dimensions]*

1. $\mathbb{E}[p_j(X)^2] \leq \tilde{C}^2$ *for all* $j = 1 \dots k$, *where* $0 < \tilde{C} < \infty$;
2. α_* *exists and* $\mathbb{E}u_*^2 = \mu_*^2$ *for some number* $\mu_* > 0$.

Discussion of Assumption H1

H1 is the high dimensional counterpart of L1, and accounts for several key features of high dimensional settings. When we include many covariates in empirical analysis, $\lambda_{\min}\{\mathbb{E}[p(X)p(X)']\}$ can decrease at certain rate and $\lambda_{\max}\{\mathbb{E}[p(X)p(X)']\}$ might be very large. Thus assumption like L1-(1) common in low dimensions is not suitable, and least square approximation usually does not apply. H1-(1) only requires second moments of all basis functions bounded from above, which is weaker but still allows tractable asymptotic analysis. H1-(2) imposes a basic sparse approximation condition for α_0. It can be verified by more primitive conditions, as some functions admit sparse representations with certain basis functions.

Following Belloni et al. (2017b); Qiu and Otsu (2018); Chernozhukov et al. (2018c), let

$$\Lambda_n = \sup_{x \in \mathcal{X}} \left(\max_{1 \leq j \leq k} |p_j(x)| \right).$$

Compared to ξ_k, Λ_n is more useful in high dimensional situations.

Assumption H2. *As* $k \to \infty$ *and* $n \to \infty$*:*

1. *[Dimension restriction]* $\Lambda_n \sqrt{\frac{\log k}{n}} = o(1)$;
2. *[First step]* $\hat{\gamma}$ *is estimated from a different iid sample, which is independent from the main sample used to calibrate* $\tilde{\alpha}$. $\|\hat{\gamma} - \gamma_0\|_{\mathbb{P},2} = O_p(\varphi_n^\gamma)$ *for some* $\varphi_n^\gamma \to 0$;
3. *[Quality of learners]* $\left[(\tilde{a} - a_*)'\hat{G}(\tilde{a} - a_*)\right]^{1/2} = O_p(\varphi_n^\alpha)$ *for some* $\varphi_n^\alpha \to 0$. $\varphi_n^\gamma \mu_* = o(\frac{1}{\sqrt{n}})$, $\varphi_n^\gamma \varphi_n^\alpha = o(\frac{1}{\sqrt{n}})$, $\|\tilde{a} - a_*\|_1 = o_p(1)$;
4. *[Diverging* α_0*]* $\|\alpha_0\|_{\mathbb{P},\infty} \varphi_n^\gamma = o(1)$.

Discussion of Assumption H2

H2-(1) allows k to grow faster than n, up to factor Λ_n and log term (aka high dimensional asymptotics). It slightly improves a similar result in Chernozhukov et al. (2018c) by exploiting a sharper bound using knowledge from second moment. H2-(2) is concerned with first step estimator $\hat{\gamma}$, which should be consistent

and achieve convergence rate $O_p(\varphi_n^\gamma)$ in $\|\cdot\|_{\mathbb{P},2}$ norm. The assumption that $\hat{\gamma}$ is trained from a different iid sample can accommodate generic machine learning estimators for γ_0. It is in line with the recent literature advocating cross fitting (for example, Robins et al., 2009; Newey and Robins, 2018; Chernozhukov et al., 2018b,c) to deal with potentially complicated machine learning algorithms. However, H2-(2) only requires γ_0 to be estimated from a different sample, while $\tilde{\alpha}$ is calibrated from the main sample. With this respect, this is a "single" cross fitting scheme weaker than the "double" cross fitting where both nuisance functions are estimated from different samples. If γ_0 is estimated using lasso, it is possible to omit this cross fitting assumption by imposing more technical conditions on stochastic equicontinuity terms.

H2-(3) is a high level condition showing a trade-off between qualities of $\hat{\gamma}$ and $\tilde{\alpha}$. Such trade-off emerges naturally from the doubly robust structure of $\hat{\theta}_{DR}$, which can effectively control remainder term R_1 in (2.18). RR α_0 needs to be approximately sparse in this set-up since sparse α_* is involved, while γ_0 does not have to be. Compared to the existing literature, H2-(3) is distinct in the followings aspects: It only requires convergence rate of $\tilde{\alpha}$ under a weaker empirical norm $\left[(\tilde{a}-a_*)'\hat{G}(\tilde{a}-a_*)\right]^{1/2}$, so that stronger l_1 convergence rate is not needed per se other than consistency. Since calibrating $\tilde{\alpha}$ does not involve cross fitting, this seems quite weak. This condition also implies that both $\tilde{\alpha}$ and $\hat{\gamma}$ attaining $o_p(n^{-1/4})$ rate is only sufficient but not necessary. Our asymptotics allows a broader scenario when one of them is estimated relatively at faster rate while the other can converge slower than $o_p(n^{-1/4})$, echoing similar recent result in Chernozhukov et al. (2018c). Also note H2-(4) highlights that $\|\alpha_0\|_{\mathbb{P},\infty}$ is permitted to grow up to the convergence rate of $\hat{\gamma}$.

Theorem 3.4. *Let O, H1 and H2 hold. Then*

$$\sqrt{n}\left(\hat{\theta}_{DR}-\theta_0\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[m(Z_i,\gamma_0(X_i))+\alpha_0(X_i)(Y_i-\gamma_0(X_i))-\theta_0\right]+o_p(1),$$

and

$$\sqrt{n}\left(\hat{\theta}_{DR}-\theta_0\right)\xrightarrow{d}N(0,\Omega),$$

where Ω is defined the same way as in Theorem 3.1.

Theorem 3.4 establishes $\sqrt{n}$ normality as well as semiparametric efficiency of $\hat{\theta}_{DR}$ under a set of general assumptions. It underlines a high level convergence hypothesis for $\tilde{\alpha}$ in terms of γ_n^α and $\|\tilde{a}-a_*\|_1$ in H3-(3). These two objects can be further studied with more technical assumptions.

To continue, let A_* be an index set of nonzero elements of a_* and $S_*=|A_*|$ is its cardinality. Thus for each vector $a=(a_1,\cdots,a_k)'\in\mathbb{R}^k$, define $a_{A_*}=$

$(a_{1,A_*}, \dots a_{j,A_*} \dots, a_{k,A_*})' \in \mathbb{R}^k$, where for each $j = 1 \cdots k$, $a_{j,A_*} = a_j \mathbf{1}\{j \in A_*\}$. Similarly, define $a_{A_*^c} = (a_{1,A_*^c}, \dots a_{j,A_*^c} \dots, a_{k,A_*^c})'$, where for each $j = 1 \cdots k$, $a_{j,A_*^c} = a_j \mathbf{1}\{j \notin A_*\}$. In other words, a_{A_*} and $a_{A_*^c}$ have non-zero elements only in A_* and its complement set A_*^c, respectively. Furthermore, write $\mathbf{W}_n = \hat{G} W_n' W_n$, which turns out to be the effective weight matrix; $\hat{\mathcal{G}} = (\mathbf{W}_n \hat{G} + \lambda_1 \hat{G})$, a generalized Gram matrix whose role is similar to that of $\hat{G}$ in a lasso regression.

Assumption H3. *The following four conditions are satisfied simultaneously:*

1. *[Maximum of empirical averages] There exist some $\varepsilon_n^R, \varepsilon_n^m, \varepsilon_n^u$ such that wpa1*

$$\left\| \mathbb{E}_n[e^R] \right\|_\infty = \varepsilon_n^R,$$

$$\left\| \mathbb{E}_n[p(X) u_*] - \mathbb{E}[p(X) u_*] \right\|_\infty = \varepsilon_n^u,$$

$$\left\| \mathbb{E}_n m(Z, p(X)) - \mathbb{E} m(Z, p(X)) \right\|_\infty = \varepsilon_n^m;$$

2. *[Choice of weight matrix] $\hat{\mathcal{G}} - \hat{G}$ is positive semidefinite, $\|\mathbf{W}_n\|_\infty = \varepsilon_n^{\mathbf{W}}$ for some $\varepsilon_n^{\mathbf{W}} > 0$;*

3. *[Compatibility] For every $a \in \mathbb{R}^k$ such that $\left\|a_{A_*^c}\right\|_1 \leq 3 \left\|a_{A_*}\right\|_1$, it follows $\|a_{A_*}\|_1^2 \leq \frac{(a' \hat{\mathcal{G}} a) S_*}{\underline{\kappa}_n}$, where $\underline{\kappa}_n = \underline{\kappa}_n(\mathbf{W}_n, \hat{G}, \lambda_1) > 0$ is a number that depends on $\mathbf{W}_n$, $\hat{G}$ and λ_1;*

4. *[Penalty] $\lambda_2 \geq 2\lambda_0$, where $\lambda_0 = 2\left[(\varepsilon_n^R + \varepsilon_n^u + \tilde{C}\mu_*)(\varepsilon_n^{\mathbf{W}} + \lambda_1) + \lambda_1(\varepsilon_n^m + \tilde{C} C^{1/2})\right] \to 0$ as $k \to \infty$ and $n \to \infty$.*

Discussion of Assumption H3

H3 highlights several aspects we endeavor to manage in challenging high dimensional situations. Each of them can be interpreted intuitively. H3-(1) is in essence an empirical process condition concerning vectors of empirical averages. When the dimension is growing too fast, l_2 counterparts of objects in H3-(1) usually diverge at rate $O\left(\sqrt{\frac{k}{n}}\right)$. However, their l_∞ norms can still be controlled quite well, if each empirical average has tail probability diminishing exponentially. See Appendix C.4 for more detailed treatment of these conditions. In particular, Lemma C.35 establishes $\varepsilon_n^u = O_p\left(\sqrt{\frac{\log k}{n}} \Lambda_n \mu_*\right)$ under O and H1 without subgaussianity of u_*. And a good control of ε_n^R and ε_n^u does require subgaussianity of their corresponding random objects. Lemmas C.34 and C.36 explore several common primitive conditions on the structure of $m(z, p(x))$ and α_0, and show that ε_n^R and ε_n^u are of order $\sqrt{\frac{\log k}{n}}$ up to some factors.

H3-(2) prescribes a correct choice for W_n. To achieve faster convergence, $\hat{\mathcal{G}}$ should not be "too small" in the sense that $\hat{\mathcal{G}} - \hat{G}$ is positive semidefinite, nor

"too big" as a larger $\varepsilon_n^{\mathbf{W}}$ slows down convergence. A simple sufficient condition for $\hat{\mathcal{G}} - \hat{G}$ to be positive semidefinite is that $\mathbf{W}_n$ has eigenvalues no smaller than 1. $\varepsilon_n^{\mathbf{W}}$ summarizes the impact of $\mathbf{W}_n$ on the convergence rate and might or might not grow asymptotically. It shall be suitably chosen such that H3-(3) and (4) are both fulfilled. Since dimension of $\mathbf{W}_n$ is growing with k at rate faster than n, choosing a right $\mathbf{W}_n$ is more delicate. It is beyond the scope of this chapter to give a general form of $\varepsilon_n^{\mathbf{W}}$, but note several simple scenarios:

1. If $\hat{G}$ is invertible wpa1, setting $W_n' W_n = \hat{G}^-$ yields $\mathbf{W}_n = I$ wpa1. Hence $\varepsilon_n^{\mathbf{W}} = O_p(1)$.

2. If $\hat{G}$ is normalized so that $\hat{G} = I$, setting $W_n' W_n = I$ yields $\mathbf{W}_n = I$ and $\varepsilon_n^{\mathbf{W}} = 1$.

3. Since $\hat{G}$ might not be invertible, choosing $\mathbf{W}_n$ through W_n can be difficult. Instead, consider selecting $\mathbf{W}_n$ directly. This opens up potentially many choices of $\mathbf{W}_n$ and diverse behaviors of $\varepsilon_n^{\mathbf{W}}$.

H3-(3) is a modified version of compatibility condition in Van de Geer (2007); Van De Geer et al. (2009). Same as restricted eigenvalue condition (Bickel et al., 2009), compatibility condition mainly alleviates inadequate behavior (non-invertibility) of design matrix in high dimensions. But compatibility condition is usually slightly weaker. For intuition we can interpret $\underline{\kappa}_n$ as the restricted minimum eigenvalue.[5] H3-(3) deviates from the usual compatibility condition in two ways: First, it extends compatibility condition to matrix $\hat{\mathcal{G}}$, which is construed as a generalized Gram matrix, taking into consideration of effective weight matrix and ridge style penalty. Second, $\underline{\kappa}_n$ is allowed to vary according to $\mathbf{W}_n$, $\hat{G}$ and λ_1. This is more realistic and different from a standard lasso where it is often assumed a constant. These modifications naturally adapt to a more complicated $\hat{\mathcal{G}}$. More importantly, it helps to see why sometimes $\tilde{\alpha}$ is able to improve convergence rate when lasso fails. Intuitively, additional penalization from $\lambda_1 \hat{G}$ and a non identity $\mathbf{W}_n$ could lead to a larger $\underline{\kappa}_n$ while a pure lasso might have very small $\underline{\kappa}_n$. See end of this section for further discussions.

H3-(4) gives the correct choice for λ_1 and λ_2. It is much less lucid since we have two tuning parameters to train, and they both affect convergence rate. Coefficient λ_2 should be chosen large enough to dominate λ_0. Coefficient λ_1, on the other hand, usually shall be chosen small enough to counteract effects from term $(\varepsilon_n^m + \tilde{C} C^{1/2})$. But sometimes it pays to set a larger λ_1. See end of this section for further discussions.

[5]For good discussions on compatibility and restricted eigenvalues of matrices, see Sections 6.12 and 6.13 in Bühlmann and Van De Geer (2011).

Theorem 3.5. *[Convergence rate of GMEN learner] Let O, H1 and H3 hold. Then*

$$\varphi_n^{\alpha} = O\left(\lambda_2\sqrt{\frac{S_*}{\underline{\kappa}_n}}\right), \quad \|\tilde{a} - a_*\|_1 = O_p\left(\lambda_2\frac{S_*}{\underline{\kappa}_n}\right).$$

Theorem 3.5 provides the low-level support for key rate condition in H2-(3) of Theorem 3.4. If H3 holds, then $\varphi_n^{\gamma}\varphi_n^{\alpha} = o\left(\frac{1}{\sqrt{n}}\right)$ if $\varphi_n^{\gamma}\lambda_2\sqrt{\frac{S_*}{\underline{\kappa}_n}} = o\left(\frac{1}{\sqrt{n}}\right)$ and $\|\tilde{a} - a_*\|_1 = o_p(1)$ if $\lambda_2\frac{S_*}{\underline{\kappa}_n} \to 0$. Notice φ_n^{α} is converging to zero faster than $\|\tilde{a} - a_*\|_1$.

Further discussion of Theorem 3.5

According to Theorem 3.5, performance of $\tilde{\alpha}$ hinges on three important factors: effective weight matrix $\mathbf{W}_n$, compatibility number $\underline{\kappa}_n$ and penalty coefficients λ_1 and λ_1. Interaction of these three determines the convergence rate of $\tilde{\alpha}$ in a complicated way. Specially, the study of $\underline{\kappa}_n$ is challenging since it can not be verified empirically in general. Though beyond the scope of this chapter to look into this, several cases might be relevant for future research.

1. Benchmark case: $\mathbf{W}_n = I$, $\lambda_1 = 0$ and $\underline{\kappa}_n(I, \hat{G}, 0) = \underline{\kappa}_{0,n}$ is a constant. GMEN learner becomes the minimum distance lasso learner in Chernozhukov et al. (2018c). The correct choice for λ_2 should be $O(\varepsilon_n^R + \varepsilon_n^u + \mu_*)$ so that
$$\varphi_n^{\alpha} = O\left[(\varepsilon_n^R + \varepsilon_n^u + \mu_*)\sqrt{S_*}\right].$$

2. Benchmark case with a small λ_1. If λ_1 is small enough so that $\lambda_1(\varepsilon_n^m \vee \tilde{C}C^{1/2})$ is negligible compared to $(\varepsilon_n^R + \varepsilon_n^u + \tilde{C}\mu_*)(1 + \lambda_1)$, convergence rate is guaranteed the same with benchmark case. A larger λ_1 than what is required leads to slower convergence.

3. Benchmark case with diminishing $\underline{\kappa}_{0,n}$. If $\underline{\kappa}_{0,n}$ is diminishing, the best rate for minimum distance lasso is
$$\varphi_n^{\alpha} = O\left[\frac{(\varepsilon_n^R + \varepsilon_n^u + \mu_*)\sqrt{S_*}}{\underline{\kappa}_{0,n}^{1/2}}\right], \quad \|\tilde{a} - a_*\|_1 = O_p\left[\frac{(\varepsilon_n^R + \varepsilon_n^u + \mu_*)S_*}{\underline{\kappa}_{0,n}}\right], \tag{3.3}$$
slower than case (1). In particular, $\|\tilde{a} - a_*\|_1$ might not even converge if $\underline{\kappa}_{0,n} \to 0$ fast, for example, at rate $O\left[(\varepsilon_n^R + \varepsilon_n^u + \mu_*)S_*\right]$. If this happens, it is worthwhile to choose a non identity $\mathbf{W}_n$ and a larger λ_1. Suppose $\mathbf{W}_n$ is selected such that H3-(2) is met with $\varepsilon_n^{\mathbf{W}} = O(1)$. Since $\lambda_1\hat{G}$ is positive semidefinite, $\underline{\kappa}_n(\mathbf{W}_n, \hat{G}, \lambda_1 > 0) \geq \underline{\kappa}_n(\mathbf{W}_n, \hat{G}, \lambda_1 = 0)$. If λ_1 is large enough

so that $\lambda_1(\varepsilon_n^m \vee \tilde{C}C^{1/2})$ dominates $(\varepsilon_n^R + \varepsilon_n^u + \tilde{C}\mu_*)(1+\lambda_1)$, it can be seen

$$\varphi_n^\alpha = O\left[\frac{\lambda_1\sqrt{S_*}}{\underline{\kappa}_n^{1/2}(\mathbf{W}_n, \hat{G}, \lambda_1 > 0)}\right], \quad \|\tilde{a} - a_*\|_1 = O_p\left[\frac{\lambda_1 S_*}{\underline{\kappa}_n(\mathbf{W}_n, \hat{G}, \lambda_1 > 0)}\right]. \tag{3.4}$$

(3.4) improves (3.3) if $\underline{\kappa}_n(\mathbf{W}_n, \hat{G}, \lambda_1 > 0)$ is significantly larger than $\underline{\kappa}_{0,n}$. Similar results have also been discussed in Hebiri et al. (2011).

3.2.1 Data-driven selection of penalties

Theorem 3.5 shows that depending on the level of λ_1 and structure of $\hat{\mathcal{G}}$, $\tilde{\alpha}$ displays two types of convergence rates: the first best scenario is to select a small λ_1 with a large enough λ_2 (fast convergence regime); the second best scenario is to choose a larger λ_1 with λ_2 adjusted accordingly (slow convergence regime). It is beyond the scope of this chapter to study theoretically justified data-driven choice for both λ_1 and λ_2. However, data-driven selection of λ_2 is possible under the first best scenario. It can also shed some light on the choice of λ_1. Propose to set λ_2 as

$$\hat{\lambda}_2 = \hat{c}\left[\frac{\left\|\hat{\Psi}\right\|_\infty}{\sqrt{n}}\Phi^{-1}\left(1 - \frac{\hat{t}}{2k}\right)\varepsilon_n^{\mathbf{W}}\right], \tag{3.5}$$

where $\hat{\Psi}$ is determined by an iterative algorithm at the end of this section, $\Phi(\cdot)$ is the distribution function of a standard normal variable, and $\hat{c}$ and $\hat{t}$ are some practical loadings. Note $\varepsilon_n^{\mathbf{W}}$ can be directly calculated since the researcher chooses $\mathbf{W}_n$. To achieve the first best rate, apply a small number for λ_1 after finding $\hat{\lambda}_2$. To achieve the second best rate, try resetting both λ_1 and λ_2 at similar magnitude but larger than $\hat{\lambda}_2$.

(3.5) is theoretically motivated and inspired by Belloni et al. (2011, 2012). The intuition is as follows: we aim to find the smallest level of λ_2 such that H3-(4) stands with a large probability. In the first best scenario, the leading term of λ_0 in H3-(4) should be $\varepsilon_n^R\varepsilon_n^{\mathbf{W}}$. Denote $\Psi = diag[\psi_1, \psi_2, \ldots, \psi_k]$ as the $k \times k$ diagonal matrix, where $\psi_j = \left\{\mathbb{E}_n\left[e_j^R\right]^2\right\}^{1/2}$ for each $j = 1 \ldots k$. Then

$$\varepsilon_n^R = \left\|\mathbb{E}_n e^R\right\|_\infty \leq \|\Psi\|_\infty \left\|\tilde{S}\right\|_\infty, \tag{3.6}$$

where

$$\left\|\tilde{S}\right\|_\infty = \left\|\Psi^{-1}\mathbb{E}_n e^R\right\|_\infty = \max_{1\leq j\leq k}\left|\frac{\mathbb{E}_n e_j^R}{\psi_j}\right|.$$

By moderate deviation theory (for example, Belloni et al., 2012; Jing et al.,

2003), for some confidence level $\hat{t}$, we can expect

$$\mathbb{P}\left[\sqrt{n}\left\|\tilde{S}\right\|_{\infty} > \Phi^{-1}\left(1-\frac{\hat{t}}{2k}\right)\right] \leq \hat{t} - o(1).$$

(3.5) thus reflects the idea to bound term $\left\|\tilde{S}\right\|_{\infty}$ in (3.6) with a large probability. In practice, we set $\hat{c} = 1.1, \hat{t} = 0.1/\log(k \vee n)$, in line with recommendations in Belloni et al. (2011, 2012, 2014, 2017b). Finally, Ψ is estimated by $\hat{\Psi}$ using the iterative algorithm below.

Algorithm [Iterative estimation of Ψ]

Step 0: Choose $\hat{c} = 1.1, \hat{t} = 0.1/\log(k \vee n)$ and $\lambda_1 = 0$. Let $L = 15$ be the number of iterations.

Step 1: Let $\hat{\Psi}^1 = diag[\hat{\psi}_1^1, \hat{\psi}_2^1, \ldots \hat{\psi}_k^1]$, where $\hat{\psi}_j^1 = \left\{\mathbb{E}_n\left[p_j(X) - \mathbb{E}_n p_j(X)\right]^2\right\}^{1/2}$ for each $j = 1 \ldots k$. Find $\hat{\lambda}_2$ according to (3.5) and parameters in step 0. Compute $\tilde{\alpha}^1$ with these penalty loadings according to (2.32).

Step 2: For $l = 2 \ldots L$, update $\hat{\lambda}_2$ according to $\hat{\Psi}^l = diag[\hat{\psi}_1^l, \hat{\psi}_2^l, \ldots \hat{\psi}_k^l]$, where

$$\hat{\psi}_j^l = \left\{\mathbb{E}_n\left[m(Z, p_j(X)) - \tilde{\alpha}^{(l-1)}(X) p_j(X)\right]^2\right\}^{1/2} \quad \text{for each } j = 1 \ldots k,$$

where $\tilde{\alpha}^{(l-1)}$ is calibrated in iteration $l-1$ according to (2.32). Repeat the process for L times.

Step 3: Use $\hat{\Psi} = diag[\hat{\psi}_1^{L+1}, \hat{\psi}_2^{L+1}, \ldots \hat{\psi}_k^{L+1}]$ as the final estimate for Ψ, where

$$\hat{\psi}_j^{L+1} = \left\{\mathbb{E}_n\left[m(Z, p_j(X)) - \tilde{\alpha}^{L}(X) p_j(X)\right]^2\right\}^{1/2} \quad \text{for each } j = 1 \ldots k.$$

3.3 Monte Carlo exercises

3.3.1 Performance of minimax BP learner under moderately high dimensions

This subsection assesses finite sample performance of $\hat{\theta}_{BP}$ when $k < n$. The set-up follows closely with the simulation study in Kang et al. (2007), which has become a standard framework for evaluating empirical performance of different estimators.[6] See Example 2.1 for background knowledge. Let $U = \{U_1, U_2, U_3, U_4\}'$ be a vector

[6]Similar experimental studies have also been carried out in Tan (2010); Rotnitzky et al. (2012); Imai and Ratkovic (2014); Zubizarreta (2015); Chan et al. (2016) etc., mainly to evaluate the strength of balancing methods compared to other methods.

of four random variables from multivariate standard normal distribution $N(\mathbf{0}, I_4)$ where I_4 is a 4×4 identity matrix. The outcome variable Y^* is generated as

$$Y^* = 210 + 27.4U_1 + 13.7U_2 + 13.7U_3 + 13.7U_4 + e,$$

where e follows a standard normal distribution $N(0, 1)$ and is independent of U. The true propensity score $\pi(u) = \mathbb{P}\{T = 1|U = u\}$ is set with a logistic fashion

$$\pi(u) = \Lambda(-u_1 + 0.5u_2 - 0.25u_3 - 0.1u_4),$$

where $\Lambda(\cdot) = \frac{\exp(\cdot)}{1+\exp(\cdot)}$. Observed outcome is then $Y = TY^*$. The econometrician does not observe U directly but only its transformed version, denoted as $X = \{X_1, X_2, X_3, X_4\}'$, with

$$X_1 = \exp\left(\frac{U_1}{2}\right), \quad X_2 = \frac{U_2}{1 + \exp(U_1)} + 10,$$
$$X_3 = \left(\frac{U_1 U_3}{25} + 0.6\right)^3, \quad X_4 = (U_2 + U_4 + 20)^2.$$

An iid sample of size $n = 200$ is drawn from observables $\{Y, T, X_1, X_2, X_3, X_4\}$. This mechanism generates a mean response rate of 0.5. Target parameter is $\mathbb{E}[Y^*] = 210$. By Example 2.1, $\alpha_0(x, t) = t/\mathbb{P}\{T = 1|X = x\}$ and the linear function $m(z, g(x))$ is just $g(x)$. For simplicity, set $W_n = I$ throughout the exercise. Hence, given some basis functions p, we can construct

$$\Theta_n = \left\{\alpha = t(a'p) : a \in \mathbb{R}^k\right\},$$

$$\mathcal{H}_I = \left\{g = \beta'p : \beta \in \mathbb{R}^k, \|\beta\| \leq 1\right\}.$$

$\hat{\theta}_{BP}$ can be computed via equation (2.24), with calibrated RR admitting an explicit solution: $\tilde{\alpha}(x, t) = t\tilde{a}'p(x)$, where

$$\tilde{a} = \left[\hat{G}_T\hat{G}_T + \lambda_1\hat{G}_T\right]^- \hat{G}_T\hat{P}, \tag{3.7}$$

and $\hat{G}_T = \mathbb{E}_n[Tp(X)p'(X)]$ and $\hat{P} = \mathbb{E}_n[p(X)]$.

3.3.1.1 Baseline result with mild selection bias

In the first exercise I look at a situation with mild selection bias, where all relevant regressors are included from the beginning. Spaces Θ_n and $\mathcal{H}_I$ are constructed using B-splines or orthogonal polynomials based on X. The dimension k for B-

splines ranges from 5 to 121, covering 11 possible cases. This makes $\frac{k}{n}$ grow from 0.025 to 0.605. As the dimensional restriction on polynomials is stricter, this exercise only considers 9 possible scenarios for polynomials: k starts from 5 to 70, so that $\frac{k}{n}$ increases from 0.025 to 0.35. Detailed procedures for constructing these basis functions can be found in Tables 3.1 and 3.2.

A number of papers in the literature (for example, Hainmueller, 2012; Imai and Ratkovic, 2014; Chan et al., 2016, etc.) have shown in simulations that balancing style estimators usually outperform other methods (for example, matching and inverse propensity score estimators) in terms of bias and root mean square error (RMSE). So, the focus of this exercise is to compare the following three balancing plug-in estimators:

1. Minimax BP learner: RR is computed with coefficient (3.7), where $\lambda_1 = 0.002$ for B-splines and $\lambda_1 = 0.001$ for orthogonal polynomials.

2. "Newey-Robins" (NR) learner: RR is computed with coefficient $\tilde{a}_{NR} = \hat{G}_T^- \hat{P}$, the estimator proposed in Newey and Robins (2018).

3. "Simple Ridge" (SR) learner: RR is computed with coefficient $\tilde{a}_{SR} = (\hat{G}_T \hat{G}_T + \lambda_1 I)^- \hat{G}_T \hat{P}$, with λ_1 same as minimax BP learner. This is the estimator we would get with a naive ridge penalty $\mathcal{P}_n(\alpha(X)) = \|a\|^2$.

Out of interest, performance of the following two baseline estimators are reported as well:

1. A naive estimator when θ_0 is estimated by simply averaging on observed outcomes.

2. A simple averaging estimator when α_0 is known.

Bias and RMSE are computed for each of the five estimators above with 10000 experiments. I also report mean TMU for estimating $\mathbb{E}[Y^*]$ based on design matrix $\hat{G}_T$. Results are collected in Tables 3.3 and 3.4 as well as Figures 3.1 and 3.2.

I also study empirical coverage probabilities of these five estimators, when the variance is estimated by equation (3.1), with $\gamma_0(x, 1) = \mathbb{E}[Y|X = x, T = 1]$ estimated by standard series method with the same basis functions. Nominal coverage probabilities are set at 1% and 5%, respectively. Results after 10000 simulations are reported in Tables 3.5 and 3.6.

Simulation results corroborate both Theorems 3.1 and 3.2. Under this mild selection bias scenario, minimax BP learner with a small penalty maintains a stable performance with a small RMSE over the span of $\frac{k}{n}$ ratios. In fact, it has

the smallest RMSE for most $\frac{k}{n}$ ratios, except in cases when $\frac{k}{n}$ is very small (usually < 0.1). On the other hand, RMSE of NR learner shoots up quite considerably as $\frac{k}{n}$ increases. SR learner also behaves stably, but at the price of large bias and RMSE. As $\frac{k}{n}$ increases, calculated mean TMU also grows, implying there is more computational uncertainty due to higher dimensions. Note the unbiased averaging estimator when α_0 is known also behaves poorly and a naive averaging-on-observed estimator is even more erratic. Similar conclusions can be drawn in terms of inference. Coverage probabilities of minimax BP learner is more stable compared to the other two alternatives, NR and SR learners. When $\frac{k}{n}$ is very small, minimax BP learner does not outperform NR learner. As $\frac{k}{n}$ starts to grow, coverage probabilities of minimax BP learner catch up and behave more steadily. SR learner behaves significantly worse than the other two in all situations.

3.3.1.2 Robustness check: considerable selection bias

I next design a situation with considerable bias. At the beginning only X_4 is used to construct B-splines. So, severe selection bias exists in the specification. But researchers gradually add more and more relevant regressors (X_3, X_2, X_1 and their technical terms) to alleviate bias. This creates a total of 10 cases with dimension k growing from 5 to 121. When the number of regressors exceeds 70, selection bias is very mild. Detailed procedures for constructing these basis functions can be found in Table 3.7. Bias and RMSE are computed for each of the same five estimators used earlier after 10000 experiments. Results are collected in Table 3.8 and Figure 3.3. It is clear that under this considerable bias scenario, minimax BP learner is able to correctly pick up improvement from alleviated OVB. Both bias and RMSE decrease to a desirable level as the model becomes correctly specified. NR or SR learner cannot achieve this easily. In particular, if NR learner is used, a correctly specified model with many regressors can behave much worse than a misspecified model with severe bias but fewer regressors.

3.3.1.3 Robustness check: sensitivity to penalty coefficient

Finally, I check sensitivity of $\hat{\theta}_{BP}$ to the choice of λ_1. Set-up is the same with earlier case of mild selection bias using B-splines. But λ_1 ranges from 0 to 0.005. Results after 10000 simulations are collected in Table 3.9 and Figure 3.4. All of them perform stably and much better than NR learners.

3.3.2 Performance of GMEN learner under high dimensions

The aim of this subsection is to study finite sample performance of GMEN learner $\tilde{\alpha}$ defined in (2.32) under ultra high dimensional regimes. I choose λ_2 in a data-driven way according to (3.5), and look at how estimation error of $\tilde{\alpha}$ changes under various specifications as well as levels of λ_1. Simulation results are in line with Theorem 3.5: when the model is sparse enough, a small λ_1 usually suffices for a good performance; when the model becomes less sparse, sometimes a larger λ_1 leads to a better result.

For simplicity, the set-up still follows the missing data framework in Example (2.1). Since α_0 is defined irrespective of outcome variable Y, I only focus on observed data $\{T, X\}$ excluding Y. Specifically, X is a random vector of $k = 300$ covariates drawn from $N(0, \Sigma)$, where $\Sigma_{(j_1,j_2)} = (0.75)^{|j_1-j_2|}$ for $1 \leq j_1, j_2 \leq k$ is the variance covariance matrix. The propensity score is designed with a logistic fashion

$$\pi_0(x) = \mathbb{P}(T = 1|X = x) = \Lambda\left[1 + A_\pi' x\right],$$

where $\Lambda(\cdot) = \frac{\exp(\cdot)}{1+\exp(\cdot)}$ and

$$A_\pi = \rho_a(1, -2^{-t_a}, 3^{-t_a}, \cdots j^{-t_a}, \cdots - p^{-t_a}).$$

The two parameters (ρ_a, t_a) are chosen by the researcher: t_a controls sparsity and ρ_a controls signal strength of the model. I create a grid of t_a and ρ_a at different values

$$t_a = (0.1, 0.3, 0.5, 0.7, 0.9, 1.1, 1.3, 1.5, 1.7, 1.9);$$
$$\rho_a = (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1).$$

Sample size $n = 200$. Throughout this experiment I choose $\mathbf{W}_n = I$ for simplicity. λ_2 is calculated by the data-driven algorithm discussed in (3.5), and λ_1 takes value in a grid from 0 to 0.1, with an 0.005 increase each step. GMEN learner $\tilde{\alpha}$ is computed according to (2.32) using R package "lbfgs". I report the average of empirical root mean square error $\left\{\mathbb{E}_n\left[\tilde{\alpha}(X) - \alpha_0(X)\right]^2\right\}^{1/2}$ over 1000 simulations. Results are collected in Figures 3.5, 3.6 and 3.7.

The empirical performance of $\tilde{\alpha}$ depends a lot on the data generating process. In general, when the model is more sparse (larger t_a), a small λ_1 usually delivers a good finite sample performance (Figure 3.5). Under these circumstances, a larger λ_2 yields a larger estimation error. On the other hand, in a large-ρ_a specification and when the model is less sparse (say $t_a = 0.1$), estimation error turns out to be

quite big and is decreasing as λ_1 increases (Figures 3.6 and 3.7). This corroborates Theorem 3.5 quite well. In practice, the real data generating process is never known. A practical recommendation for applied work is to start with a small nonzero λ_1 and see how results might change when λ_1 increases. Cross validation could be used for selecting λ_1. More research is needed on theoretically justified data-driven choices for λ_1.

3.3.3 Tables and figures

Table 3.1: Construction of B-splines: mild selection bias

Model	*df* of each X_j	order of each X_j	Aggregating pattern for X_j, $j = 1 \dots 4$	k	$\frac{k}{n}$
(1)	1	1	additive	5	0.025
(2)	2	2	additive	9	0.045
(3)	3	2	additive	13	0.065
(4)	4	3	additive	17	0.085
(5)	5	3	additive	21	0.105
(6)	5	3	(5)+1st and 2nd same degree interactions	$21 + 12 = 33$	0.165
(7)	5	3	(5) + all same degree interactions	$21 + 30 = 51$	0.255
(8)	6	3	additive+same degree interactions	$25 + 36 = 61$	0.305
(9)	6	3	(8)+*df* $= 1$ interactions with *df* $= 2$ and 3	$61 + 24 = 85$	0.425
(10)	6	3	(9)+*df* $= 1$ interactions with *df* $= 4$ and 5	$85 + 24 = 109$	0.545
(11)	6	3	(8)+ *df* $= 1$ interactions with all other *df* terms	$61 + 60 = 121$	0.605

Note: This table shows how we construct B-splines basis functions of different dimensions with R function "bs". Some terminologies: A spline of order r with t interior knots has dimension $r + t$. The degree of freedom (*df*) of B-spline defined on $\mathbb{R}$ is such that it has *df* $- r$ interior knots. Notice R ignores the constant so the real dimension always needs to add 1.

Table 3.2: Construction of orthogonal polynomials: mild selection bias

Model	degree	Aggregating pattern for X_j, $j = 1 \dots 4$	k	$\frac{k}{n}$
(1)	1	all 1st degree polynomials	5	0.025
(2)	2	(1)+ same regressors' 2nd degree polynomials	9	0.045
(3)	2	(1)+all regressor' 2nd degrees polynomials	15	0.075
(4)	3	(3)+half of 3rd degree polynomials	15+10=25	0.125
(5)	3	(3)+all 3rd degree polynomials	15+20=35	0.175
(6)	4	(5)+ first 10 4th degree polynomials	35+10=45	0.225
(7)	4	(6)+ second 10 4th degree polynomials	35+20=55	0.275
(8)	4	(7)+ third 10 4th degree polynomials	35+30=65	0.325
(9)	4	(5)+all 4th degree polynomials	35+35=70	0.35

Note: Orthogonal polynomials are constructed straightforwardly with R package "poly".

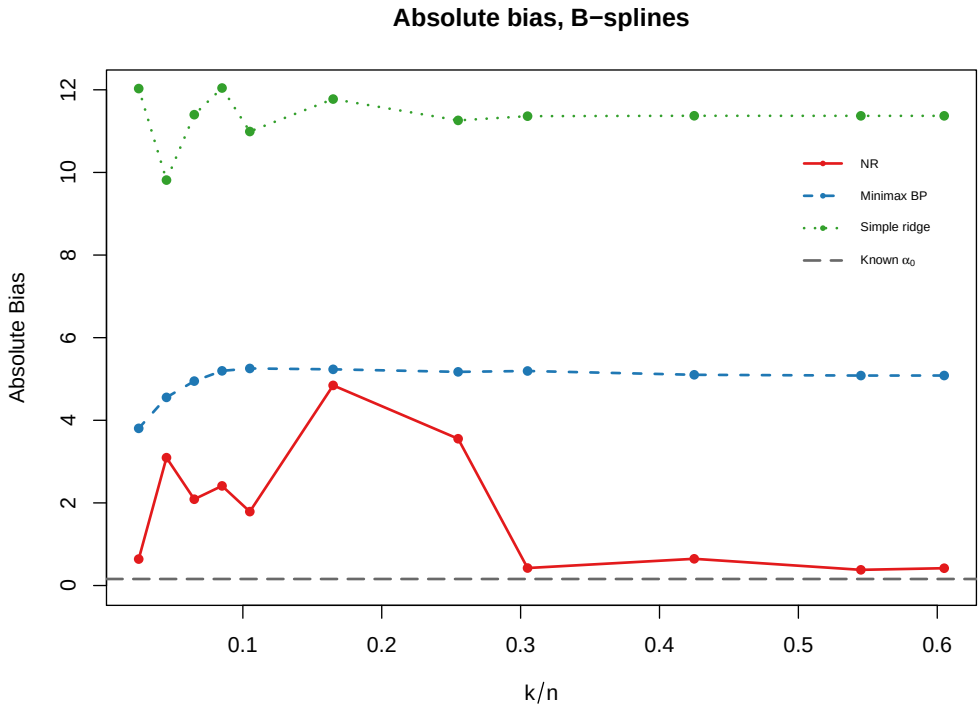
Table 3.3: Bias and RMSE using B-splines, 10000 Monte Carlo, $\lambda_1 = 0.002$, mild selection bias

Model	k	$\frac{k}{n}$	NR		Minimax BP		SR		Mean
			Bias	RMSE	Bias	RMSE	Bias	RMSE	TMU
(1)	5	0.025	-0.6404	3.3775	-3.8049	4.8548	-12.0300	12.4907	0.0000
(2)	9	0.045	-3.0948	4.7457	-4.5542	5.3889	-9.8147	10.3187	0.0020
(3)	13	0.065	-2.0888	4.2762	-4.9485	5.8034	-11.3977	11.8625	0.0050
(4)	17	0.085	-2.4118	11.1004	-5.1964	6.0223	-12.0436	12.5116	0.0249
(5)	21	0.105	-1.7891	17.7536	-5.2546	6.0873	-10.9903	11.4777	0.0449
(6)	33	0.165	-4.8446	14.8438	-5.2335	6.0627	-11.7762	12.2495	0.0370
(7)	51	0.255	-3.5531	16.5214	-5.1723	5.9932	-11.2596	11.7459	0.0492
(8)	61	0.305	-0.4233	14.8353	-5.1933	6.0162	-11.3621	11.8633	0.0450
(9)	85	0.425	-0.6467	15.3074	-5.1013	5.9352	-11.3714	11.8766	0.0519
(10)	109	0.545	-0.3799	14.4902	-5.0826	5.9193	-11.3699	11.8769	0.0516
(11)	121	0.605	-0.4192	14.7215	-5.0849	5.9207	-11.3696	11.8768	0.0523
(12)	Naive average		-10.0458	10.6347					
(13)	$\alpha_0(X)$ known		0.1584	23.6996					

Table 3.4: Bias and RMSE using orthogonal polynomials, 10000 Monte Carlo, $\lambda_1 = 0.001$, mild selection bias

Model	k	$\frac{k}{n}$	NR		Minimax BP		SR		Mean
			Bias	RMSE	Bias	RMSE	Bias	RMSE	TMU
(1)	5	0.025	-0.6404	3.3775	-6.0196	6.6971	-11.0890	11.6207	0.0000
(2)	9	0.045	-3.0948	4.7457	-5.1090	5.9144	-11.3495	11.8623	0.0143
(3)	15	0.075	-1.5043	3.8659	-6.1856	6.8593	-11.1085	11.6401	0.0128
(4)	25	0.125	5.9640	15.5018	-6.3648	7.0406	-11.1049	11.6353	0.0519
(5)	35	0.175	6.4579	16.0746	-6.3638	7.0419	-11.1100	11.6401	0.0557
(6)	45	0.225	-6.6549	27.2024	-6.4208	7.0899	-11.1180	11.6479	0.1096
(7)	55	0.275	-8.1283	26.9325	-6.4274	7.0940	-11.1197	11.6493	0.3311
(8)	65	0.325	-8.0678	26.5464	-6.4276	7.0942	-11.1200	11.6496	0.1960
(9)	70	0.35	-7.7737	26.3677	-6.4321	7.1002	-11.1243	11.6537	0.1026
(12)	Naive average		-10.0458	10.6347					
(13)	$\alpha_0(X)$ known		0.1584	23.6996					

Figure 3.1: Bias and RMSE, B-splines, mild selection bias



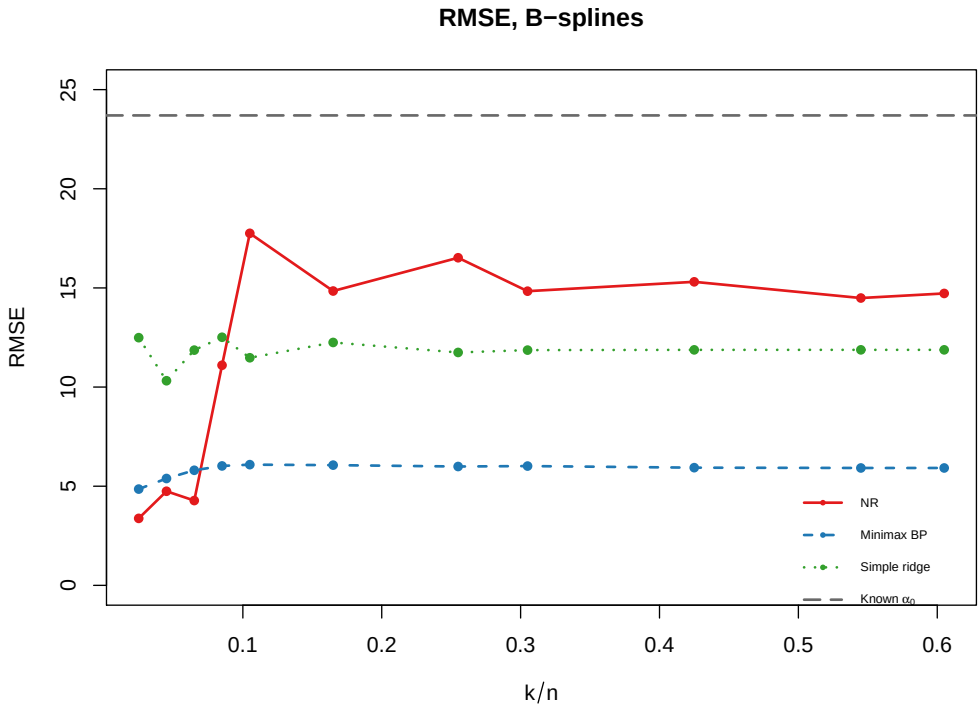
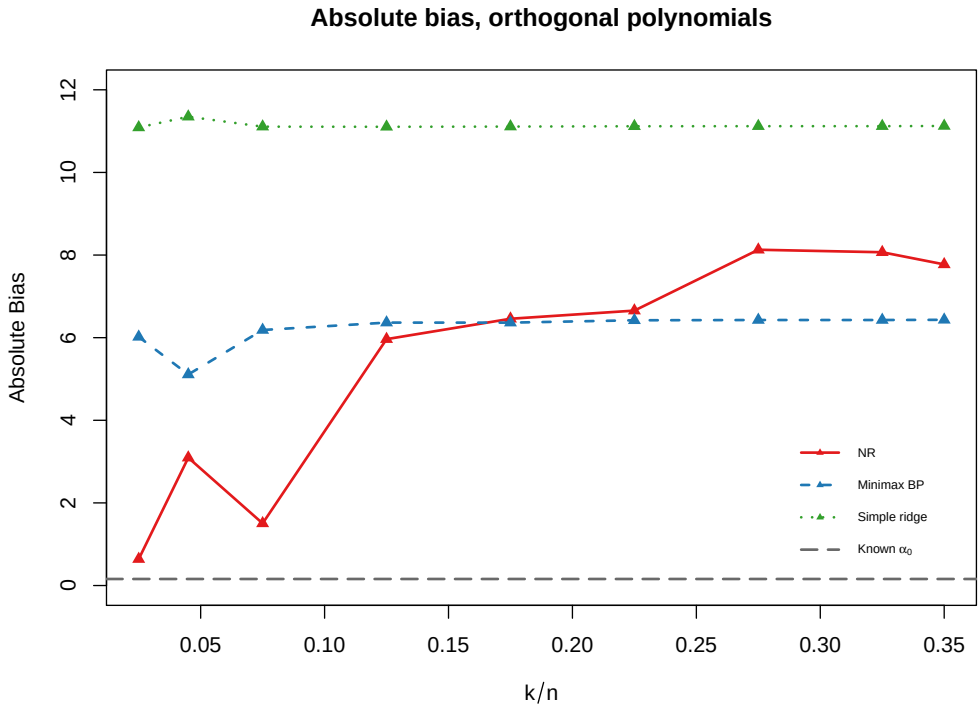


Figure 3.2: Bias and RMSE, orthogonal polynomials, mild selection bias



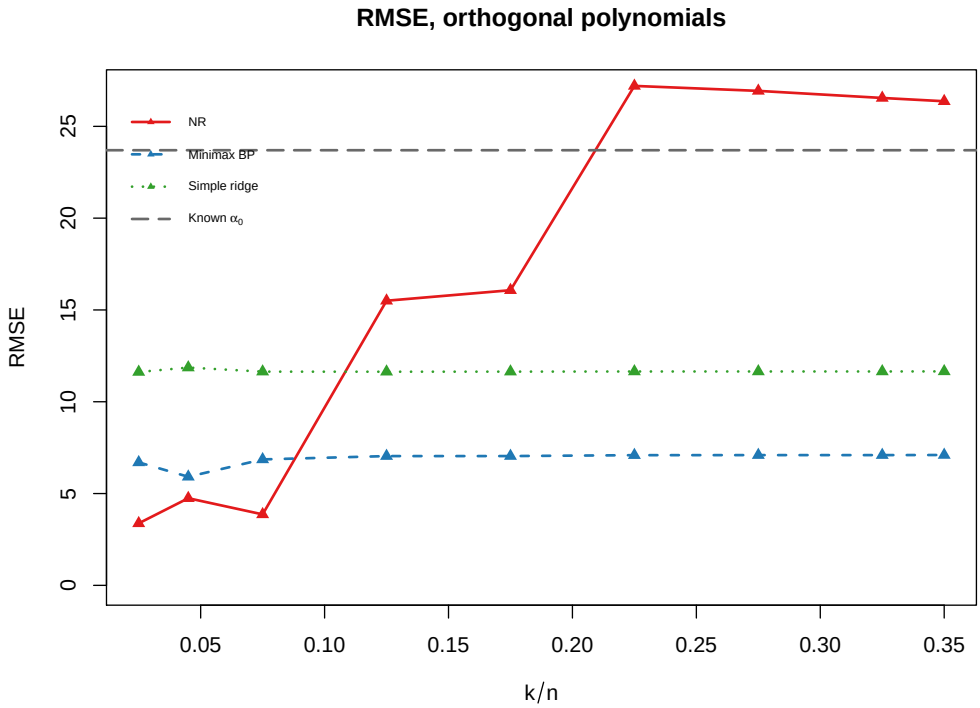


Table 3.5: Coverage probability using B-splines, 10000 Monte Carlo, $\lambda_1 = 0.002$, mild selection bias

Model	k	$\frac{k}{n}$	NR		Minimax BP		SR	
			5%	1%	5%	1%	5%	1%
(1)	5	0.025	0.9348	0.9833	0.8946	0.9753	0.3695	0.7945
(2)	9	0.045	0.8053	0.9214	0.7262	0.8766	0.3904	0.7309
(3)	13	0.065	0.8613	0.9462	0.7666	0.9125	0.3313	0.7094
(4)	17	0.085	0.8959	0.9708	0.8090	0.9257	0.4267	0.7491
(5)	21	0.105	0.9236	0.9795	0.8408	0.9382	0.5631	0.8249
(6)	33	0.165	0.9268	0.9822	0.8271	0.9431	0.4721	0.7927
(7)	51	0.255	0.9008	0.9755	0.8455	0.9365	0.5887	0.8450
(8)	61	0.305	0.9317	0.9851	0.8854	0.9561	0.6234	0.8755
(9)	85	0.425	0.9113	0.9791	0.9266	0.9674	0.7394	0.9211
(10)	109	0.545	0.8831	0.9651	0.9450	0.9755	0.7926	0.9402
(11)	121	0.605	0.8805	0.9630	0.9460	0.9740	0.8013	0.9404
(12)	Naive		0.1741	0.3766				
(13)	$\alpha_0(X)$ known		0.9342	0.9772				

Table 3.6: Coverage probability using orthogonal polynomials, 10000 Monte Carlo, $\lambda_1 = 0.001$, mild selection bias

Model	k	$\frac{k}{n}$	NR		Minimax BP		Simple ridge	
			5%	1%	5%	1%	5%	1%
(1)	5	0.025	0.9348	0.9833	0.8204	0.9608	0.4500	0.8310
(2)	9	0.045	0.8053	0.9214	0.6998	0.8679	0.2881	0.6472
(3)	15	0.075	0.8940	0.9664	0.7557	0.9269	0.4021	0.7750
(4)	25	0.125	0.9351	0.9901	0.9211	0.9813	0.7308	0.9307
(5)	35	0.175	0.9296	0.9906	0.9295	0.9808	0.7601	0.9341
(6)	45	0.225	0.9482	0.9936	0.9576	0.9867	0.8536	0.9608
(7)	55	0.275	0.9572	0.9946	0.9560	0.9869	0.8570	0.9605
(8)	65	0.325	0.9575	0.9964	0.9614	0.9886	0.8626	0.9648
(9)	70	0.35	0.9559	0.9955	0.9586	0.9871	0.8622	0.9616
(10)	Naive		0.1741	0.3766				
(11)	$\alpha_0(X)$ known		0.9342	0.9772				

Table 3.7: Construction of B-splines: considerable selection bias

Model	Regressors used	df of each X_j	order of each X_j	Aggregating pattern for X_j	k	$\frac{k}{n}$
(1)	X_4 only	4	3	additive	5	0.025
(2)	X_4 only	5	3	additive	6	0.03
(3)	X_3, X_4 only	5	3	additive	11	0.055
(4)	X_3, X_4 only	5	3	(3)+ same degree interactions	16	0.08
(5)	X_2, X_3, X_4 only	5	3	additive+same degree interactions	31	0.155
(6)	X_2, X_3, X_4 only	5	3	(5)+ $df = 1$ interactions with all other df terms	55	0.275
(7)	X_2, X_3, X_4 only	6	3	additive+ same degree interactions+$df = 1$ interactions with all other df terms	67	0.335
(8)	All	6	3	(7) $+X_1$	73	0.365
(9)	All	6	3	additive+same degree interactions+$df = 1$ interactions with $df = 5$ and 6	85	0.425
(10)	All	6	3	additive+same degree interaction+$df = 1$ interactions with all other df terms	121	0.605

Note: This table shows how we construct B-splines basis functions of different dimensions with R function "bs". Some terminologies: A spline of order r with t interior knots has dimension $r + t$. The degree of freedom (df) of B-spline defined on $\mathbb{R}$ is such that it has $df - r$ interior knots. Notice R ignores the constant so the real dimension always needs to add 1.

Table 3.8: Bias and RMSE using B-splines, 10000 Monte Carlo, $\lambda_1 = 0.002$, considerable selection bias

Model	k	$\frac{k}{n}$	NR		Minimax BP		Simple ridge	
			Bias	RMSE	Bias	RMSE	Bias	RMSE
(1)	5	0.025	-12.3471	12.7659	-12.8001	13.1886	-12.2811	12.7143
(2)	6	0.03	-12.3616	12.8130	-12.8344	13.2205	-12.6920	13.1125
(3)	11	0.055	-11.5525	14.1700	-12.6433	13.0143	-12.1266	12.5324
(4)	16	0.08	-19.1261	23.0365	-20.2015	20.6683	-20.0538	20.5281
(5)	31	0.155	-17.9294	23.7115	-19.4334	19.9311	-20.0314	20.4959
(6)	55	0.275	-9.1457	20.3253	-11.5796	11.9982	-12.2974	12.6762
(7)	67	0.335	-10.6529	23.3291	-11.3636	11.8042	-12.6588	13.0421
(8)	73	0.365	-1.2343	20.8580	-5.2934	6.1331	-11.1601	11.6531
(9)	85	0.425	-0.6845	14.9800	-5.1839	6.0073	-11.3643	11.8660
(10)	121	0.605	-0.3799	14.4902	-5.0826	5.9193	-11.3699	11.8769
(12)	Naive average		-10.0458	10.6347				
(13)	$\alpha_0(X)$ known		0.1584	23.6996				

Figure 3.3: Bias and RMSE, B-splines, considerable selection bias

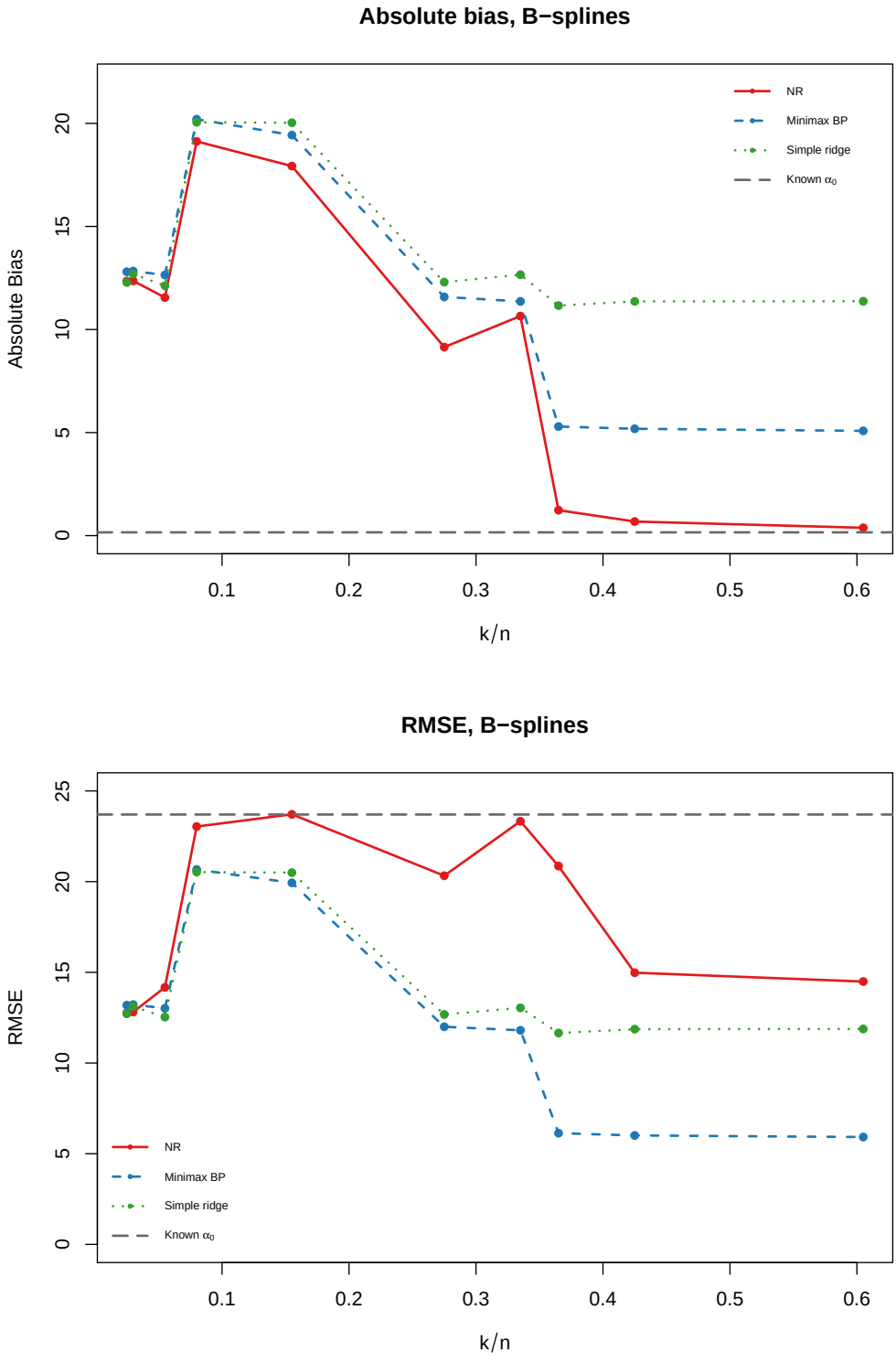
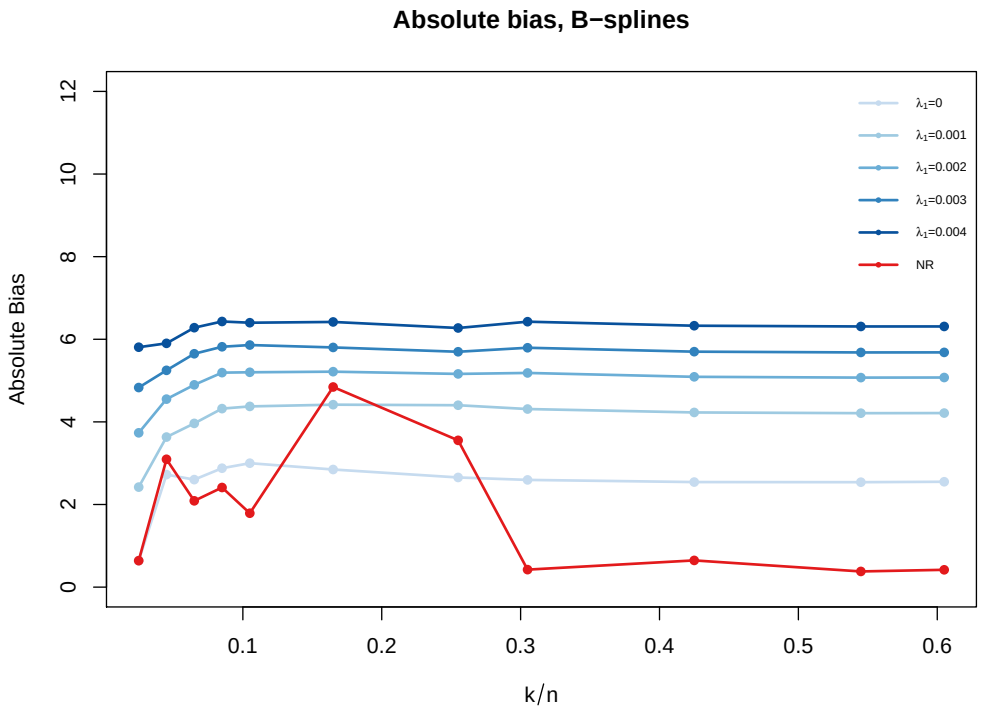


Table 3.9: Sensitivity of $\hat{\theta}_{BP}$ to λ_1 using B-splines, 10000 Monte Carlo, mild selection bias

Model	k	$\frac{k}{n}$	$\lambda_1 = 0$		$\lambda_1 = 0.001$		$\lambda_1 = 0.002$		$\lambda_1 = 0.003$		$\lambda_1 = 0.004$	
			Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
(1)	5	0.025	-0.6404	3.3775	-2.4214	3.8804	-3.7352	4.8249	-4.8314	5.7219	-5.8081	6.5660
(2)	9	0.045	-2.7246	4.2714	-3.6326	4.6738	-4.5510	5.3910	-5.2480	5.9982	-5.9014	6.5619
(3)	13	0.065	-2.6026	4.1235	-3.9635	4.9994	-4.8965	5.7670	-5.6489	6.4320	-6.2802	6.9773
(4)	17	0.085	-2.8785	4.2637	-4.3213	5.3010	-5.1920	6.0275	-5.8184	6.5771	-6.4304	7.1101
(5)	21	0.105	-3.0002	4.3796	-4.3746	5.3580	-5.2010	6.0466	-5.8601	6.6309	-6.4006	7.0975
(6)	33	0.165	-2.8472	4.2580	-4.4166	5.3672	-5.2155	6.0543	-5.8022	6.5738	-6.4192	7.1271
(7)	51	0.255	-2.6549	4.0324	-4.4033	5.3293	-5.1614	5.9882	-5.6977	6.4622	-6.2732	6.9823
(8)	61	0.305	-2.5957	3.9553	-4.3113	5.2456	-5.1847	6.0134	-5.7946	6.5625	-6.4253	7.1304
(9)	85	0.425	-2.5434	3.9136	-4.2287	5.1759	-5.0911	5.9317	-5.6998	6.4786	-6.3278	7.0430
(10)	109	0.545	-2.5401	3.9075	-4.2107	5.1607	-5.0720	5.9155	-5.6812	6.4628	-6.3096	7.0271
(11)	121	0.605	-2.5504	3.9102	-4.2136	5.1620	-5.0740	5.9165	-5.6824	6.4634	-6.3103	7.0274

Figure 3.4: Sensitivity of $\hat{\theta}_{BP}$ to λ_1 using B-splines, mild selection bias



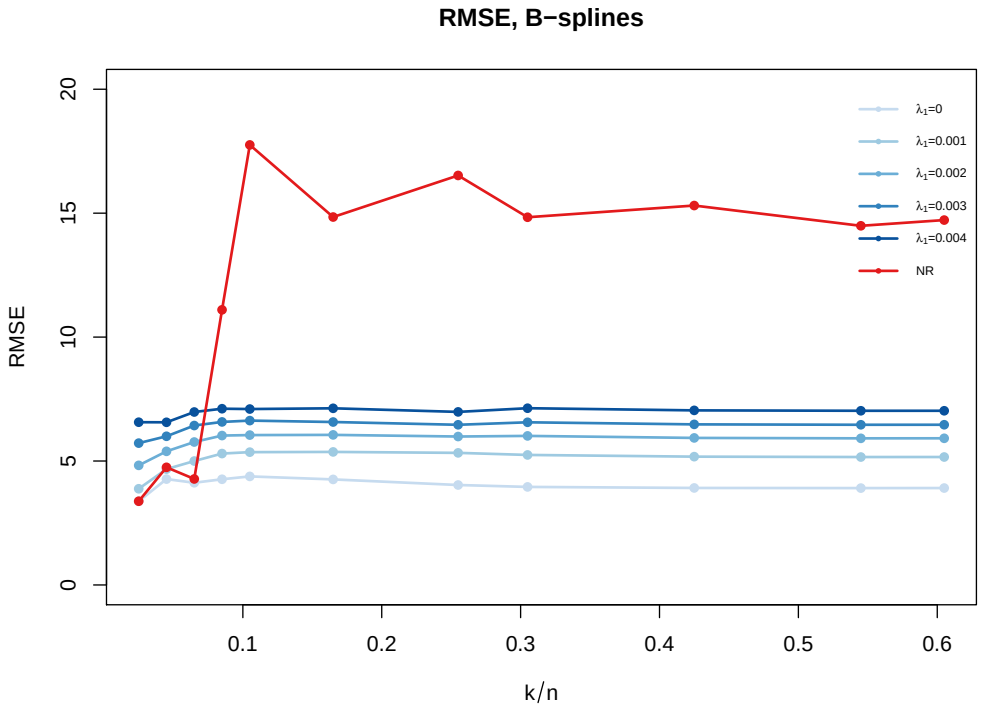
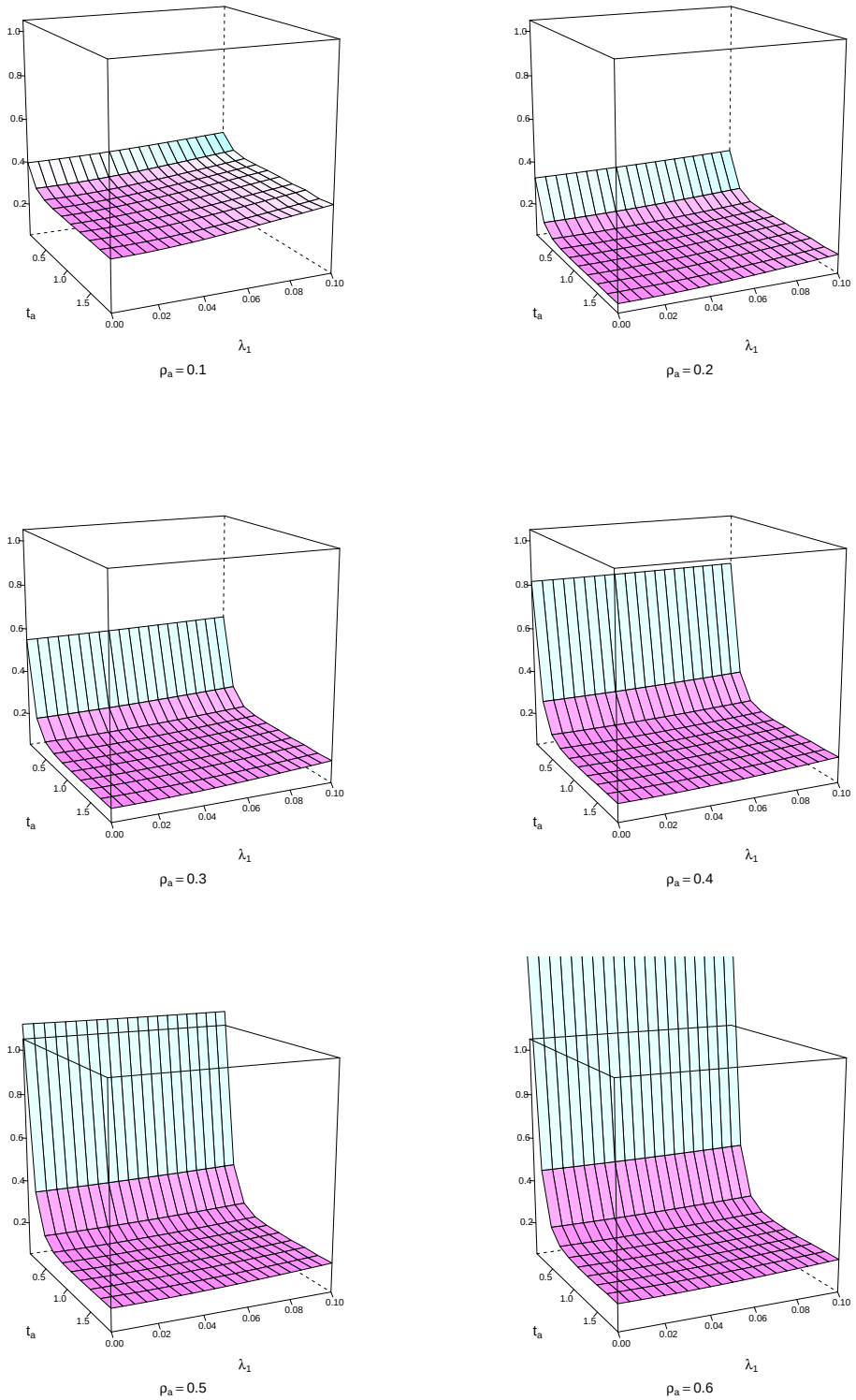
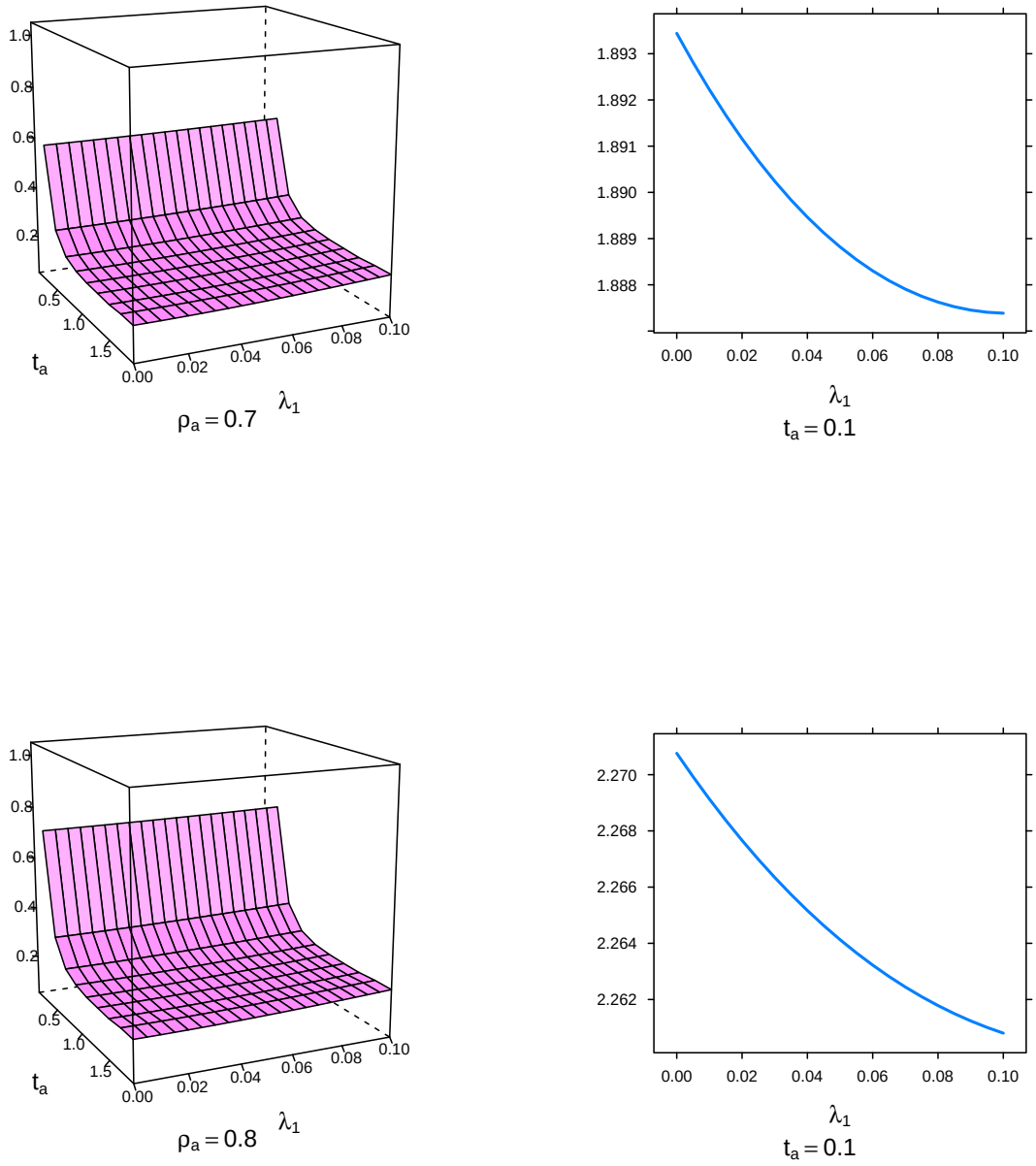


Figure 3.5: Performance of GMEN learner under high dimensions (1)



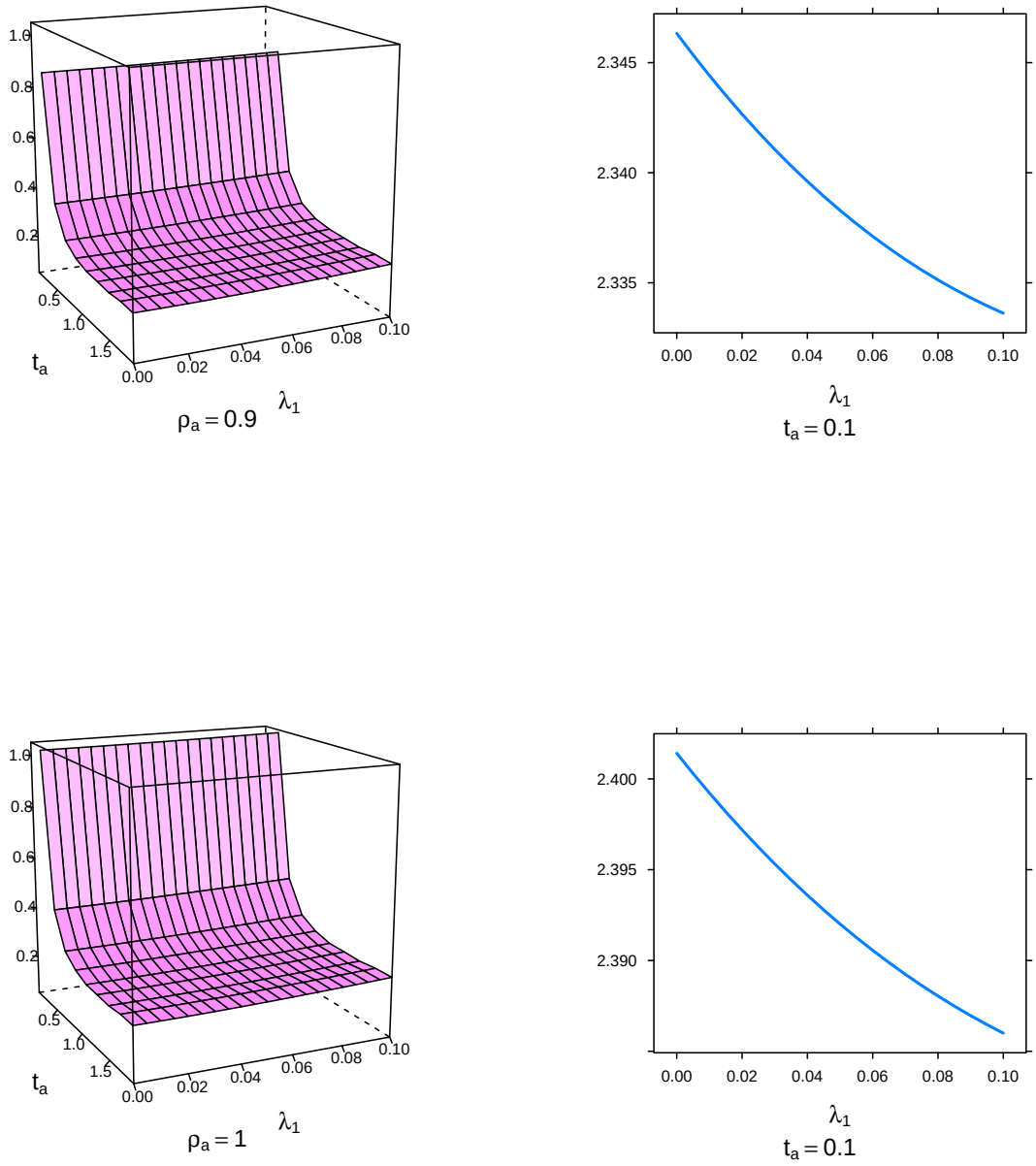
Note: Vertical axis represents average of empirical RMSE. λ_2 is selected by data-driven algorithm. t_a represents sparsity of the model, and ρ_a is signal strength. Results are reported after 1000 simulations.

Figure 3.6: Performance of GMEN learner under high dimensions (2)



Note: Vertical axis in the left panel represents average of empirical RMSE. When $t_a = 0.1$, empirical RMSE is usually too large to be effectively included in the left, and is hence reported in the corresponding right plot. λ_2 is selected by proposed data-driven algorithm. t_a represents sparsity of the model and ρ_a is the signal strength. Results are reported after 1000 simulations.

Figure 3.7: Performance of GMEN learner under high dimensions (3)



Note: Vertical axis in the left panel represents average of empirical RMSE. When $t_a = 0.1$, empirical RMSE is usually too large to be effectively included in the left, and is hence reported in the corresponding right plot. λ_2 is selected by proposed data-driven algorithm. t_a represents sparsity of the model and ρ_a is the signal strength. Results are reported after 1000 simulations.

Chapter 4

Minimax learning for average regression functionals: application

In this chapter I apply minimax learners to the work of Ferraz and Finan (2011). They study the effect of electoral accountability on corruption using a unique municipality-level dataset from Brazil. What is special about their dataset is that: (1) it is observed at subnational level; (2) it is perceived to be a natural experiment so treatment is plausibly close to random assignment. Hence one of their main empirical strategies is OLS with controls of many mayoral and municipal characteristics. They find in municipalities where mayors are serving first term, share of resources involving corruption is significantly lower than in municipalities with second-term mayors. Their results lend convincing evidence to political agency models that underscore how re-election incentives affect political behavior (see Besley, 2006 for a review), and are consistent with other empirical works in this field (for example, Besley and Case, 1995; List and Sturm, 2006).

Within this context, the objective of this exercise is to investigate the performance of minimax learners as well as other popular methods in the literature. I find the main result of Ferraz and Finan (2011) very robust. However, OLS suffers from a problem of "over control", distorting estimates considerably due to dimensionality issues of many controls. Ignoring this effect can lead to misinterpretation of the dataset. Minimax learners, on the other hand, do not over control, perform stably and lead to economically coherent conclusion. It also seems to confirm indeed selection bias in the data of Ferraz and Finan (2011) is mild. In line with the theory developed in this thesis, $\hat{\theta}_{BP}$ works well for moderately high dimensional cases, while $\hat{\theta}_{DR}$ can handle ultra high dimensional situations nicely. Other off-the-shelf shrinkage methods (ridge, lasso, etc.) do not work as well as minimax learners. In particular, when there are many controls, performance of doubly robust estimators with (post) lasso selected propensity scores appears less satisfactory.

4.1 Main empirical framework

Ferraz and Finan (2011) use the following linear regression as one of the main empirical strategies

$$Y_i = \theta_0 T_i + X_i'\beta + Z_i'\gamma + \varepsilon_i, \tag{4.1}$$

where Y_i stands for the share of resources related to corrupt activities in municipality i, as collected from audit reports, T_i is the treatment

$$T_i = \begin{cases} 1 & \text{if mayor i's term limit is not binding,} \\ 0 & \text{if mayor i's term limit is binding,} \end{cases}$$

and ε_i is the error term. In (4.1), object of interest is θ_0, which can be interpreted as the average treatment effect of reelection incentives on corruption if individual treatment effect is a constant (see Angrist, 1998). To solve potential endogeneity, Ferraz and Finan (2011) distinguish two important sets of controls

$$X : \text{municipal characteristics}, \quad Z : \text{mayor characteristics}.$$

The key identification assumption is that all potential confounders have been included in X and Z such that $\mathbb{E}[e_i|T, X, Z] = 0$ for each observation i.

This exercise deviates from the slightly restrictive controlled regression approach (4.1) to a more flexible semiparametric framework. Let us denote $Y(1)$ as the level of corruption when mayor's term limit is not binding (or equivalently, when mayor has reelection incentives), and denote $Y(0)$ as the level of corruption when mayor's term limit is binding (no reelection incentives). The object of interest is then the expected difference between the two corruption levels $\theta_0 = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$. Example 2.1 applies. Under conditional independence and overlap assumptions, θ_0 is identified as

$$\theta_0 = \mathbb{E}\left[\mathbb{E}[Y|X, Z, T = 1]\right] - \mathbb{E}[\mathbb{E}[Y|X, Z, T = 0]]\,. \tag{4.2}$$

Similar to the idea of (2.7), RR can be found for $\mathbb{E}[Y(1)]$ and $\mathbb{E}[Y(0)]$, respectively.

4.2 Main empirical results

As one of their primary empirical analyses, Ferraz and Finan (2011) explore how estimate of θ_0 in (4.1) changes when different sets of covariates are included. They start with a plain vanilla mean comparison, and gradually add five sets of relevant controls: mayor characteristics, municipal characteristics, political

and judicial institutions, lottery and state dummy variables. A full specification includes a total of 67 controls versus a sample size of 476. This is moderately high dimensional with $\frac{k}{n} = 0.14$. I compare estimates using OLS and minimax BP learner with $W_n = I$ and $\lambda_1 = 0.001$. I also report results from six other popular methods in the literature: ridge regression based on (4.1) with fixed penalty 0.001; ridge regression based on (4.1) with 10 fold cross validation; linear partialling out method for (4.1) with post lasso selection; double selection for (4.1) with post lasso selection; doubly robust methods based on (4.2) with lasso and post lasso selected propensity scores, respectively. Results are collected in Table 4.1.

From Table 4.1 we see OLS estimates are quite unstable, sensitive to which controls are included. A simple mean comparison yields a point estimate of -0.0188, meaning lame duck mayors on average steal 1.88 percentage points more resources. As we add more regressors, magnitude of the estimate increases sometimes quite substantially. Once all 67 regressors are included, the point estimate becomes -0.0275. Though all significant at 5% level, such a disparity is quite evident considering the unit of outcomes is percentage point. To put these numbers into perspective, on average each municipality receives an annual federal transfer of R$5.5 million. A simple mean comparison would predict that every lame duck mayor steals approximately R$100,000 more than first term mayor. With all 67 regressors, the prediction becomes close to R$150,000, an increase of almost 50%. See Figure 4.1 for a detailed illustration. On the other hand, minimax BP learner produces quantitatively stable estimate at around -0.018 throughout six specifications, and all of them are statistically significant at at least 10% level. Also see Figure 4.2. Estimates from minimax BP learner are coherent with a dataset that has mild selection bias. While what we see from OLS with many controls are probably a result of "over control": adding more regressors only distorts estimates so much that improvement from addressing omitted variable bias is tiny. This problem is not easily solved by off-the-shelf shrinkage methods. Performance of naive ridge regression with a penalty $\lambda_1 = 0.001$ improves a little bit but overall quite similar to OLS. Ridge regression with 10 fold cross validation almost completely breaks down, yielding nothing stable or significant at all. The four other lasso based selection methods perform well overall but not as stably as minimax BP learner. In this case, the most competitive one seems to be the doubly robust method based on (4.2) with lasso selected propensity score. Note linear regression based methods (double selection and partialling out) do not work as well as those using doubly robust approach based on (4.2).

Following Ferraz and Finan (2011), Tables 4.2 and 4.3 report effects of reelection incentives using two alternative measures of corruption as observed outcome:

number of irregularities associated with corruption and share of service items involving corruption. Results are also similar to those in Table 4.1.

4.3 Controlling for ability and experience

So far added covariates do not account for unobserved characteristics of individual politicians, such as political ability and experience. These personal characteristics are hard to measure, but omitting them might lead to sizable bias. For instance, if we believe more experienced and/or more able politicians are more corrupt, we over estimate the effect of reelection incentives. Ferraz and Finan (2011) carefully devise several strategies to tackle this potential source of bias. In this subsection I illustrate some of them and see how results would look like if minimax BP learner is applied.

To control for political experience, specifications (1) to (6) of Table 4.4 reestimate all original specifications (1)-(6) in Table 4.1, but with one additional proxy variable for each specification. This additional proxy indicates whether first term mayor was in power in one of the previous three terms. To account for possible nonlinearity, specification (7) further adds interaction terms of political experience proxy with 11 other continuous variables on top of specification (6). Results show clearly that OLS produces even more unstable results when this political experience proxy is added. On the other hand, minimax BP learner with same level of penalty $\lambda_1 = 0.001$ still behaves robustly, with statistically significant estimates at around -0.017. Other methods do not behave as well as minimax BP learner. In particular, doubly robust with post lasso selected propensity score performs a bit erratically, especially in specification (7).

Following Ferraz and Finan (2011), mayor's political ability can be controlled by comparing second term mayors with a subset of first term mayors who are reelected in subsequent elections. This reduces sample size from 476 to 313. Results are reported in Table 4.5. All methods, except cross validated ridge regression, deliver very significant results with increased magnitude compared to full sample size. Minimax BP learner is still quite stable compared to OLS and most other off-the-shelf methods.

4.4 Accounting for many more controls

Finally I explore how minimax learners perform in the presence of many more technical controls. B-splines are first created based on 11 continuous regressors used in Table 4.1. Adding interaction and second order terms, I get four specifications with many technical controls. The number of controls in these specifications

ranges from 67 to 254.[1] Under this high dimensional situation, $\hat{\theta}_{DR}$ is more suitable than $\hat{\theta}_{BP}$. To construct $\hat{\theta}_{DR}$, first step $\hat{\gamma}$ is estimated using "rlasso" in "hdm" R package and $\tilde{\alpha}$ is calibrated with $\mathbf{W}_n = I$ and λ_2 selected by proposed data-driven algorithm. Results are reported in Table 4.6 against different levels of λ_1, ranging from 0 to 0.1. Standard errors are calculated using simple plug-in method. Table 4.6 also reports results using the other four lasso based methods mentioned earlier. Finally, for comparison purpose, Table 4.7 collects results using $\hat{\theta}_{BP}$ with a series of small penalties.

We find $\hat{\theta}_{DR}$ behaves very well. The estimate is stable and significant across the four high dimensional specifications. The choice of λ_1 has a minimal impact on point estimate, but a larger λ_1 leads to smaller standard error and thus more significant estimate. Doubly robust estimators using (post) lasso selected propensity scores do not behave well especially when k becomes too large. This might signal erratic behavior of inverse of estimated propensity score under high dimensions, even if it is regularized. Linear partialling out and linear double selection methods on the other hand, behave stably under this case and produce significant estimates throughout four specifications. Their simple linear structure might be the reason behind these results. There are some other interesting findings. $\hat{\theta}_{BP}$ still behaves very stably even in the presence of many controls. However, when there are too many regressors ($k = 188, 254$), computed standard errors become too large to make estimates significant. This is well expected. See discussions for Theorems 3.2 and 3.3. Dimensionality has a more adverse effect on variance estimation. Nevertheless, these exercises under many technical controls further support the view that Ferraz and Finan (2011)'s data are close to random assignment. Their main result stays robust.

4.5 Tables and figures

[1] See footnote of Table 4.6 for a detailed procedure on how to construct these controls.

Table 4.1: Effect of reelection incentives on corruption: baseline results

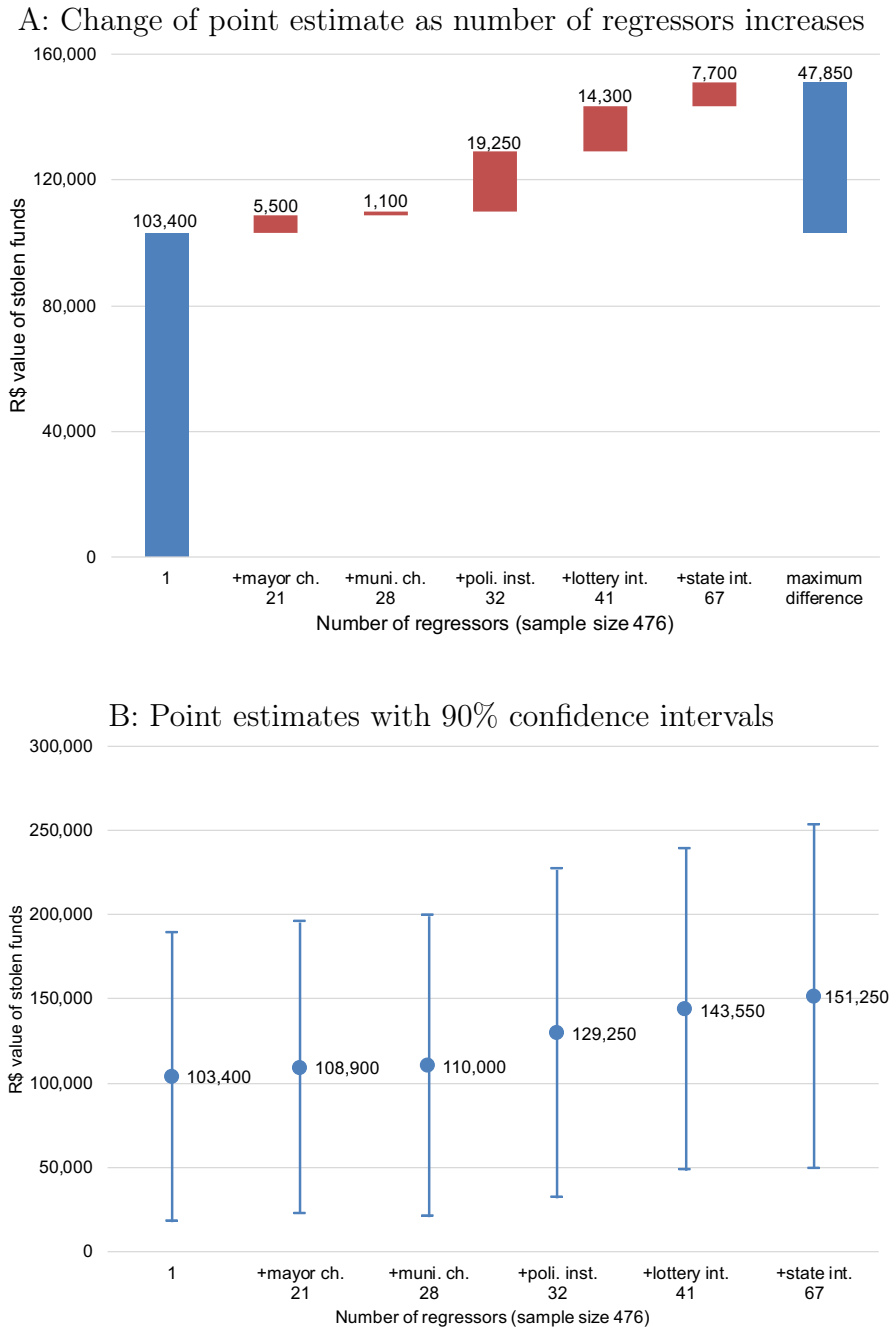
Specification		(1)	(2)	(3)	(4)	(5)	(6)
k		1	21	28	32	41	67
n		476	476	476	476	476	476
TMU: $\mathbb{E}Y(1)$		0	0.0330	0.0198	0.0509	0.0718	0.0989
TMU: $\mathbb{E}Y(0)$		0	0.0248	0.0273	0.0313	0.0404	0.0557
OLS	Effect	-0.0188**	-0.0198**	-0.0200**	-0.0235**	-0.0261**	-0.0275**
	S.E.	(0.0095)	(0.0096)	(0.0099)	(0.0108)	(0.0106)	(0.0113)
Minimax BP	Effect	-0.0187**	-0.0186**	-0.0162*	-0.0182*	-0.0182*	-0.0182**
($W_n = I, \lambda = 0.001$)	S.E.	(0.0094)	(0.0087)	(0.0088)	(0.0097)	(0.0095)	(0.0092)
OLS+ridge	Effect		-0.0195**	-0.0197**	-0.0233**	-0.0256**	-0.0263**
$\lambda = 0.001$	S.E.		(0.0096)	(0.0099)	(0.0108)	(0.0106)	(0.0113)
OLS+ridge	Effect		-0.0070	-0.0078	-0.0010	-0.0076	-0.0053
10 fold CV	S.E.		(0.0097)	(0.0100)	(0.0110)	(0.0109)	(0.0119)
Doubly robust	Effect		-0.0180*	-0.0177*	-0.0252**	-0.0252**	-0.0214*
post lasso selected p.s.†	S.E.		(0.0094)	(0.0096)	(0.0111)	(0.0111)	(0.0110)
Doubly robust	Effect		-0.0188**	-0.0181*	-0.0225**	-0.0225**	-0.0219**
lasso selected p.s.	S.E.		(0.0095)	(0.0095)	(0.0100)	(0.0100)	(0.0100)
Linear partialling out	Effect		-0.0177*	-0.0198**	-0.0248***	-0.0259***	-0.0216**
post lasso selection	S.E.		(0.0093)	(0.0093)	(0.0096)	(0.0095)	(0.0096)
Linear double selection	Effect		-0.0180*	-0.0200**	-0.0248**	-0.0260**	-0.0224**
post lasso selection	S.E.		(0.0096)	(0.0095)	(0.0104)	(0.0103)	(0.0105)
Mayor characteristics		No	Yes	Yes	Yes	Yes	Yes
Municipal characteristics		No	No	Yes	Yes	Yes	Yes
Political and judicial characteristics		No	No	No	Yes	Yes	Yes
Lottery dummy		No	No	No	No	Yes	Yes
State dummy		No	No	No	No	No	Yes

Note: k refers to the number of regressors and n is the sample size. Numbers in parentheses are computed standard errors. (1)-(6) use the same controls as those in Table 4 of Ferraz and Finan (2011). Mayor characteristics include: age, gender, education, party affiliation; Municipal characteristics include: log population, percentage of the population that has at least a secondary education, percentage of the population that lives in the urban sector, new municipality, log GDP per capita in 2002, Gini coefficient, log amount of resources sent to the municipality; Political and judicial characteristics include: effective number of political parties in the legislature, the number of legislators divided by the number of voters, the share of the legislature that is of the same party as the mayor, and whether the municipality is judiciary district. Two ridge methods use R package "glmnet"; Four lasso based methods use R package "hdm".

*** Significant at 1%. ** Significant at 5 %. * Significant at 10%.

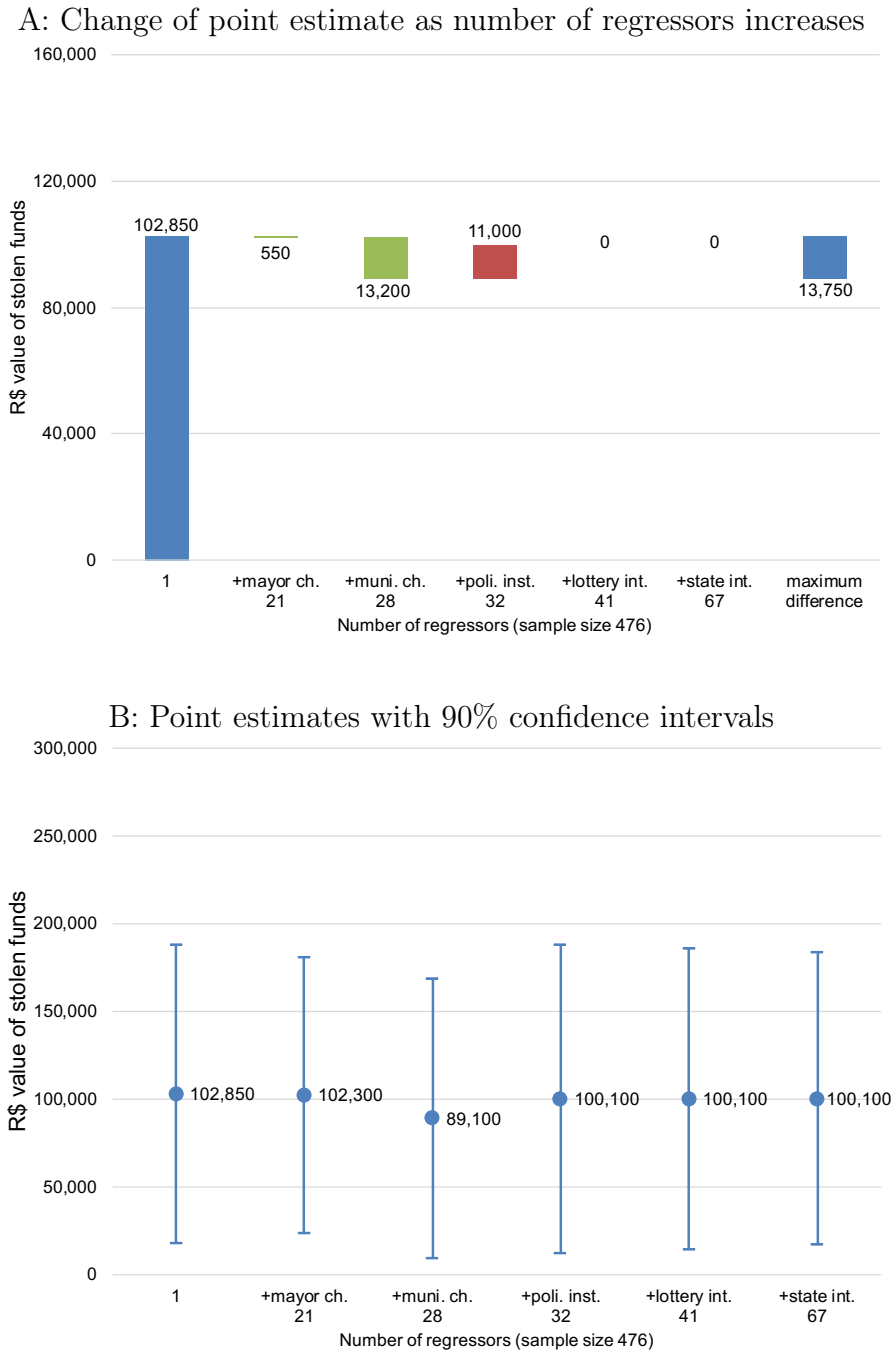
† propensity score.

Figure 4.1: Effect of a lame duck mayor on a 5.5 million transfer: OLS



Note: This figure illustrates how much more value out of a R$5.5 million transfer would be stolen by a second term mayor according to estimations of OLS in Table 4.1. Panel A shows how point estimate changes as we add more regressors from specification (1) to (6) in Table 4.1. Red areas represent an increase and green areas represent a decrease. Panel B shows point estimate with 90% confidence interval for each specification (1)-(6) in Table 4.1.

Figure 4.2: Effect of a lame duck mayor on a 5.5 million transfer: minimax BP learner



Note: This figure illustrates how much more value out of a R$5.5 million transfer would be stolen by a second term mayor according to estimations of minimax BP learner in Table 4.1. Panel A shows how point estimate changes as we add more regressors from specification (1) to (6) in Table 4.1. Red areas represent an increase and green areas represent a decrease. Panel B shows point estimate with 90% confidence interval for each specification (1)-(6) in Table 4.1.

Table 4.2: Effect of reelection incentives on alternative measure of corruption

Specification		(1)	(2)	(3)	(4)	(5)	(6)
Corruption measure		numbers of irregularities involving corruption					
k		1	21	28	32	41	67
n		476	476	476	476	476	476
OLS	Effect	-0.3875**	-0.4297***	-0.3641**	-0.3947**	-0.4470***	-0.4710***
	S.E.	(0.1583)	(0.1549)	(0.1525)	(0.1530)	(0.1506)	(0.1478)
Minimax BP	Effect	-0.3857**	-0.4320***	-0.3477**	-0.3528**	-0.3568***	-0.3568***
($\lambda = 0.001$)	S.E.	(0.1576)	(0.1495)	(0.1421)	(0.1409)	(0.1343)	(0.1225)
OLS+ridge	Effect		-0.4222***	-0.3451***	-0.3804**	-0.4226***	-0.4584***
$\lambda = 0.001$	S.E.		(0.1549)	(0.1524)	(0.1527)	(0.1506)	(0.1490)
OLS+ridge	Effect		-0.2072	-0.2405	-0.2628*	-0.3155**	-0.2484
10 fold CV	S.E.		(0.1562)	(0.1534)	(0.1533)	(0.1523)	(0.1541)
Doubly robust	Effect		-0.3612**	-0.2864*	-0.4123***	-0.4123***	-0.4035***
post lasso selected p.s.†	S.E.		(0.1589)	(0.1507)	(0.1540)	(0.1540)	(0.1514)
Doubly robust	Effect		-0.3873**	-0.3445**	-0.4402***	-0.4408***	-0.4323***
lasso selected p.s.	S.E.		(0.1581)	(0.1568)	(0.1554)	(0.1553)	(0.1551)
Linear partialling out	Effect		-0.4359***	-0.3048**	-0.3797**	-0.3581**	-0.3872***
post lasso selection	S.E.		(0.1572)	(0.1442)	(0.1494)	(0.1493)	(0.1393)
Linear double selection	Effect		-0.4367***	-0.3152**	-0.3872***	-0.3618**	-0.3998***
post lasso selection	S.E.		(0.1579)	(0.1452)	(0.1456)	(0.1460)	(0.1406)
Mayor characteristics		No	Yes	Yes	Yes	Yes	Yes
Municipal characteristics		No	No	Yes	Yes	Yes	Yes
Political and judicial characteristics		No	No	No	Yes	Yes	Yes
Lottery dummy		No	No	No	No	Yes	Yes
State dummy		No	No	No	No	No	Yes

Note: k refers to the number of regressors and n is the sample size. Numbers in parentheses are computed standard errors. (1)-(6) use the same controls as those in Table 4 of Ferraz and Finan (2011). Mayor characteristics include: age, gender, education, party affiliation; Municipal characteristics include: log population, percentage of the population that has at least a secondary education, percentage of the population that lives in the urban sector, new municipality, log GDP per capita in 2002, Gini coefficient, log amount of resources sent to the municipality; Political and judicial characteristics include: effective number of political parties in the legislature, the number of legislators divided by the number of voters, the share of the legislature that is of the same party as the mayor, and whether the municipality is judiciary district. Two ridge methods use R package "glmnet"; Four lasso based methods use R package "hdm".

*** Significant at 1%.

** Significant at 5 %.

* Significant at 10%.

† propensity score.

Table 4.3: Effect of reelection incentives on alternative measure of corruption

Specification		(1)	(2)	(3)	(4)	(5)	(6)
Corruption measure		share of audited items involving corruption					
k		1	21	28	32	41	67
n		476	476	476	476	476	476
OLS	Effect	-0.0076	-0.0100***	-0.0077	-0.0081*	-0.0100**	-0.0105**
	S.E.	(0.0048)	(0.0045)	(0.0047)	(0.0047)	(0.0044)	(0.0044)
Minimax BP	Effect	-0.0076	-0.0091**	-0.0062	-0.0057	-0.0055	-0.0055
($\lambda = 0.001$)	S.E.	(0.0048)	(0.0044)	(0.0044)	(0.0044)	(0.0039)	(0.0036)
OLS+ridge	Effect		-0.0098**	-0.0074	-0.0080*	-0.0096**	-0.0103**
$\lambda = 0.001$	S.E.		(0.0045)	(0.0047)	(0.0047)	(0.0044)	(0.0043)
OLS+ridge	Effect		-0.0030	-0.0035	-0.0043	-0.0066	-0.0058
10 fold CV	S.E.		(0.0046)	(0.0047)	(0.0047)	(0.0045)	(0.0045)
Doubly robust	Effect		-0.0067	-0.0043	-0.0069	-0.0074*	-0.0049
post lasso selected p.s.†	S.E.		(0.0048)	(0.0049)	(0.0048)	(0.0045)	(0.0048)
Doubly robust	Effect		-0.0076	-0.0063	-0.0080*	-0.0080*	-0.0078*
lasso selected p.s.	S.E.		(0.0048)	(0.0048)	(0.0047)	(0.0047)	(0.0047)
Linear partialling out	Effect		-0.0099**	-0.0055	-0.0073	-0.0082*	-0.0073*
post lasso selection	S.E.		(0.0048)	(0.0046)	(0.0048)	(0.0045)	(0.0041)
Linear double selection	Effect		-0.0099**	-0.0058	-0.0074	-0.0081*	-0.0077*
post lasso selection	S.E.		(0.0047)	(0.0046)	(0.0046)	(0.0043)	(0.0041)
Mayor characteristics		No	Yes	Yes	Yes	Yes	Yes
Municipal characteristics		No	No	Yes	Yes	Yes	Yes
Political and judicial characteristics		No	No	No	Yes	Yes	Yes
Lottery dummy		No	No	No	No	Yes	Yes
State dummy		No	No	No	No	No	Yes

Note: k refers to the number of regressors and n is the sample size. Numbers in parentheses are computed standard errors. (1)-(6) use the same controls as those in Table 4 of Ferraz and Finan (2011). Mayor characteristics include: age, gender, education, party affiliation; Municipal characteristics include: log population, percentage of the population that has at least a secondary education, percentage of the population that lives in the urban sector, new municipality, log GDP per capita in 2002, Gini coefficient, log amount of resources sent to the municipality; Political and judicial characteristics include: effective number of political parties in the legislature, the number of legislators divided by the number of voters, the share of the legislature that is of the same party as the mayor, and whether the municipality is judiciary district. Two ridge methods use R package "glmnet"; Four lasso based methods use R package "hdm".

*** Significant at 1%.

** Significant at 5 %.

* Significant at 10%.

† propensity score.

Table 4.4: Effect of reelection incentives on corruption: controlling for political experience

Specification		(1)	(2)	(3)	(4)	(5)	(6)	(7)
k		1+1	21+1	28+1	32+1	41+1	67+1	67+1+11
n		476	476	476	476	476	476	476
TMU: $\mathbb{E}Y(1)$		0	0.0622	0.0064	0.0222	0.0413	0.0739	0.0764
TMU: $\mathbb{E}Y(0)$		0	0.0417	0.0435	0.0629	0.0749	0.1037	0.0351
OLS	Effect	-0.0164*	-0.0178*	-0.0179*	-0.0217*	-0.0243**	-0.0262**	-0.0246**
	S.E.	(0.0099)	(0.0101)	(0.0103)	(0.0113)	(0.0110)	(0.0116)	(0.0122)
Minimax BP	Effect	-0.0179*	-0.0185**	-0.0159*	-0.0179*	-0.0179*	-0.0178*	-0.0177*
($\lambda = 0.001$)	S.E.	(0.0098)	(0.0087)	(0.0088)	(0.0096)	(0.0095)	(0.0092)	(0.0093)
OLS+ridge	Effect		-0.0195**	-0.0176*	-0.0233**	-0.0238**	-0.0250**	-0.0243**
$\lambda = 0.001$	S.E.		(0.0096)	(0.0103)	(0.0108)	(0.0110)	(0.0117)	(0.0122)
OLS+ridge	Effect		-0.0070	-0.0070	-0.0100	-0.0073	-0.0051	-0.0041
10 fold CV	S.E.		(0.0097)	(0.0105)	(0.0110)	(0.0113)	(0.0123)	(0.0130)
Doubly robust	Effect		-0.0180*	-0.0188*	-0.0252**	-0.0230**	-0.0173	-0.0486
post lasso selected p.s.†	S.E.		(0.0094)	(0.0101)	(0.0111)	(0.0111)	(0.0111)	(0.0346)
Doubly robust	Effect		-0.0188**	-0.0176**	-0.0225**	-0.0214**	-0.0210**	-0.0249*
lasso selected p.s.	S.E.		(0.0095)	(0.0094)	(0.0100)	(0.0099)	(0.0098)	(0.0132)
Linear partialling out	Effect		-0.0177*	-0.0169*	-0.0248***	-0.0231**	-0.0202**	-0.0196**
post lasso selection	S.E.		(0.0093)	(0.0095)	(0.0096)	(0.0099)	(0.0099)	(0.0099)
Linear double selection	Effect		-0.0180*	-0.0170*	-0.0248**	-0.0232**	-0.0210*	-0.0204*
post lasso selection	S.E.		(0.0096)	(0.0100)	(0.0104)	(0.0110)	(0.0111)	(0.0112)
Mayor characteristics		No	Yes	Yes	Yes	Yes	Yes	Yes
Municipal characteristics		No	No	Yes	Yes	Yes	Yes	Yes
Political and judicial characteristics		No	No	No	Yes	Yes	Yes	Yes
Lottery dummy		No	No	No	No	Yes	Yes	Yes
State dummy		No	No	No	No	No	Yes	Yes

Note: k refers to the number of regressors and n is the sample size. Numbers in parentheses are computed standard errors. (1)-(6) use the same controls as those in Table 4 of Ferraz and Finan (2011). Mayor characteristics include: age, gender, education, party affiliation; Municipal characteristics include: log population, percentage of the population that has at least a secondary education, percentage of the population that lives in the urban sector, new municipality, log GDP per capita in 2002, Gini coefficient, log amount of resources sent to the municipality; Political and judicial characteristics include: effective number of political parties in the legislature, the number of legislators divided by the number of voters, the share of the legislature that is of the same party as the mayor, and whether the municipality is judiciary district. The one additional regressor used in each specification is a proxy of political experience, indicating whether a first term mayor was a mayor in one of previous three terms. In specification (7), the 11 additional regressors are interactions of non-dummy regressors in (6) with the political experience indicator, respectively.
*** Significant at 1%. ** Significant at 5 %. * Significant at 10%.
† propensity score.

Table 4.5: Effect of reelection incentives on corruption: controlling for political ability

Specification		(1)	(2)	(3)	(4)	(5)	(6)
k		1	21	28	32	41	67
n		313	313	313	313	313	313
TMU: $\mathbb{E}Y(1)$		0	0.0403	0.1636	0.1377	0.0570	0.0055
TMU: $\mathbb{E}Y(0)$		0	0.0542	0.0114	0.0267	0.0358	0.0259
OLS	Effect	-0.0345***	-0.0356***	-0.0358***	-0.0411***	-0.0418***	-0.0398***
	S.E.	(0.0097)	(0.0103)	(0.0109)	(0.0118)	(0.0122)	(0.0130)
Minimax BP	Effect	-0.0345***	-0.0303***	-0.0310***	-0.0330***	-0.0330***	-0.0329***
($\lambda = 0.001$)	S.E.	(0.0097)	(0.0087)	(0.0091)	(0.0092)	(0.0092)	(0.0091)
OLS+ridge	Effect		-0.0349***	-0.0351***	-0.0402***	-0.0409***	-0.0385***
$\lambda = 0.001$	S.E.		(0.0104)	(0.0110)	(0.0119)	(0.0123)	(0.0131)
OLS+ridge	Effect		-0.0160	-0.0140	-0.0167	-0.0125	-0.0102
10 fold CV	S.E.		(0.0104)	(0.0113)	(0.0123)	(0.0127)	(0.0140)
Doubly robust	Effect		-0.0344***	-0.0351***	-0.0405***	-0.0405***	-0.0377***
post lasso selected p.s.†	S.E.		(0.0100)	(0.0103)	(0.0109)	(0.0109)	(0.0112)
Doubly robust	Effect		-0.0338***	-0.0337***	-0.0357***	-0.0357***	-0.0353***
lasso selected p.s.	S.E.		(0.0097)	(0.0097)	(0.0097)	(0.0097)	(0.0096)
Linear partialling out	Effect		-0.0326***	-0.0305***	-0.0359***	-0.0359***	-0.0371***
post lasso selection	S.E.		(0.0111)	(0.0111)	(0.0114)	(0.0114)	(0.0112)
Linear double selection	Effect		-0.0338***	-0.0314***	-0.0370***	-0.0370***	-0.0385***
post lasso selection	S.E.		(0.0097)	(0.0098)	(0.0107)	(0.0107)	(0.0108)
Mayor characteristics		No	Yes	Yes	Yes	Yes	Yes
Municipal characteristics		No	No	Yes	Yes	Yes	Yes
Political and judicial characteristics		No	No	No	Yes	Yes	Yes
Lottery dummy		No	No	No	No	Yes	Yes
State dummy		No	No	No	No	No	Yes

Note: This table only uses a subsample of second term mayors and first term mayors who were later reelected, as a control for political ability. k refers to the number of regressors and n is the sample size. Numbers in parentheses are computed standard errors. (1)-(6) use the same controls as those in Table 4 of Ferraz and Finan (2011). Mayor characteristics include: age, gender, education, party affiliation; Municipal characteristics include: log population, percentage of the population that has at least a secondary education, percentage of the population that lives in the urban sector, new municipality, log GDP per capita in 2002, Gini coefficient, log amount of resources sent to the municipality; Political and judicial characteristics include: effective number of political parties in the legislature, the number of legislators divided by the number of voters, the share of the legislature that is of the same party as the mayor, and whether the municipality is judiciary district.

*** Significant at 1%.

** Significant at 5 %.

* Significant at 10%.

† propensity score.

Table 4.6: Effect of reelection incentives on corruption: minimax DR with many controls

Specification		(1)	(2)	(3)	(4)
k		67	122	188	254
n		476	476	476	476
Minimax DR w. lasso	Effect	-0.0192**	-0.0201**	-0.0204**	-0.0178*
$\lambda_1 = 0$	S.E.	(0.0091)	(0.0091)	(0.0091)	(0.0091)
Minimax DR w. lasso	Effect	-0.0192**	-0.0201**	-0.0204**	-0.0178**
$\lambda_1 = 0.03$	S.E.	(0.0089)	(0.0088)	(0.0088)	(0.0088)
Minimax DR w. lasso	Effect	-0.0192**	-0.0201**	-0.0204**	-0.0178**
$\lambda_1 = 0.06$	S.E.	(0.0086)	(0.0086)	(0.0086)	(0.0086)
Minimax DR w. lasso	Effect	-0.0192**	-0.0201**	-0.0204**	-0.0178**
$\lambda_1 = 0.1$	S.E.	(0.0083)	(0.0083)	(0.0083)	(0.0083)
Doubly robust	Effect	-0.0214*	-0.0225*	0.0409	0.0011
post lasso selected p.s.†	S.E.	(0.0110)	(0.0125)	(0.1052)	(0.0240)
Doubly robust	Effect	-0.0219**	-0.0223**	-0.0203	-0.0157
lasso selected p.s.	S.E.	(0.0100)	(0.0101)	(0.0140)	(0.0110)
linear partialling out	Effect	-0.0216**	-0.0211**	-0.0198**	-0.0211**
post lasso selection	S.E.	(0.0096)	(0.0095)	(0.0096)	(0.0095)
linear double selection	Effect	-0.0224*	-0.0221**	-0.0205*	-0.0221**
post lasso selection	S.E.	(0.0105)	(0.0104)	(0.0106)	(0.0104)

Note: k refers to the number of regressors and n is the sample size. Numbers in parentheses are computed standard errors. Controls in each specification are constructed as follows:
56 dummy variables in specification (6) from Table 4.1 are directly used in all specifications. In addition, we collect all 11 continuous regressors used in specification (6) from Table 4.1. Based on these 11 continuous regressors, we generate B-splines in the following way: (1), degree of freedom 1 and order 1; (2), degree of freedom 1 and order 1, with all interactions; (3) degree of freedom 2 and order 2, with all same degree interactions; (4), degree of freedom 3 and order 2, with all same degree interactions.
*** Significant at 1%.
** Significant at 5 %.
* Significant at 10%.
† propensity score.

Table 4.7: Effect of reelection incentives on corruption: minimax BP with many controls

Specification		(1)	(2)	(3)	(4)
k		67	122	188	254
n		476	476	476	476
OLS	Effect	-0.0275**			
	S.E.	(0.0113)			
$\lambda_1 = 0$	Effect	-0.0277***	-0.0241**	-0.0248	-0.0249
	S.E.	(0.0103)	(0.0096)	(0.0213)	(0.0339)
$\lambda_1 = 0.002$	Effect	-0.0245**	-0.0245**	-0.0244	-0.0240
	S.E.	(0.0097)	(0.0094)	(0.0212)	(0.0339)
$\lambda_1 = 0.004$	Effect	-0.0233**	-0.0238**	-0.0234	-0.0232
	S.E.	(0.0095)	(0.0093)	(0.0212)	(0.0339)
$\lambda_1 = 0.006$	Effect	-0.0225**	-0.0232**	-0.0228	-0.0227
	S.E.	(0.0094)	(0.0093)	(0.0212)	(0.0339)
$\lambda_1 = 0.008$	Effect	-0.0220**	-0.0228**	-0.0223	-0.0222
	S.E.	(0.0093)	(0.0093)	(0.0212)	(0.0339)
$\lambda_1 = 0.01$	Effect	-0.0215**	-0.0224**	-0.0219	-0.0218
	S.E.	(0.0092)	(0.0092)	(0.0212)	(0.0339)

Note: k refers to the number of regressors and n is the sample size. Numbers in parentheses are computed standard errors. Controls in each specification are constructed as follows:
56 dummy variables in specification (6) from Table 4.1 are directly used in all specifications. In addition, we collect all 11 continuous regressors used in specification (6) from Table 4.1. Based on these 11 continuous regressors, we generate B-splines in the following way: (1), degree of freedom 1 and order 1; (2), degree of freedom 1 and order 1, with all interactions; (3) degree of freedom 2 and order 2, with all same degree interactions; (4), degree of freedom 3 and order 2, with all same degree interactions.
*** Significant at 1%.
** Significant at 5 %.
* Significant at 10%.

Appendix A

Supplementary materials for Chapter 1

A.1 Mixing

The probability space is $(\Omega, \mathcal{F}, \mathbb{P})$. For any two $\sigma-$fields $\mathcal{A} \in \mathcal{F}$ and $\mathcal{B} \in \mathcal{F}$, define the following measures of dependence:

$$\alpha(\mathcal{A}, \mathcal{B}) = \sup \left| \mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B) \right|, A \in \mathcal{A}, B \in \mathcal{B};$$
$$\beta(\mathcal{A}, \mathcal{B}) = \sup \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} \left| \mathbb{P}(A_i \cap B_j) - \mathbb{P}(A_i)\mathbb{P}(B_j) \right|,$$

where the supremum is taken over all pairs of (finite) partitions $\{A_1, \ldots, A_I\}$ and $\{B_1, \ldots, B_J\}$ of Ω such that $A_i \in \mathcal{A}$ for each i and $B_j \in \mathcal{B}$ for each j. For a sequence $\{X_i\}_{i=-\infty}^{\infty}$, the $m-$th $\alpha-$mixing coefficient is defined as

$$\alpha_m = \sup_i \alpha(\sigma(\ldots, X_{i-1}, X_i), \sigma(X_{i+m}, X_{i+m+1}, \ldots)),$$

and it is said to be $\alpha-$mixing if $\alpha_m \to 0$ as $m \to \infty$. Similarly, its $m-$th $\beta-$mixing coefficient is defined as

$$\beta_m = \sup_i \beta(\sigma(\ldots, X_{i-1}, X_i), \sigma(X_{i+m}, X_{i+m+1}, \ldots)),$$

and $\{X_i\}_{i=-\infty}^{\infty}$ is called $\beta-$mixing if $\beta_m \to 0$ as $m \to \infty$. See Bradley et al. (2005) for more details.

A.2 Proofs for low dimensional case

A.2.1 Lemmas

Lemma A.1. *Let $f(x) = (f_1(x), \ldots, f_K(x))'$ be a K-dimensional vector of functions, and $M_q = \max_{1\le j\le K}\{\mathbb{E}|f_j(X)|^q\}^{1/q}$. Suppose $\{X_i\}_{i=1}^n$ is α-mixing with mixing coefficient $\{\alpha_m\}_{m\in\mathbb{N}}$ satisfying $KM_2(M_2 + M_q\sum_{m=1}^n \alpha_m^{1/2-1/q})/n \to 0$ for some $q \in (2,\infty]$. Then*

$$|\mathbb{E}_n[f(X)] - \mathbb{E}[f(X)]| = O_p\left(\sqrt{\frac{KM_2}{n}\left(M_2 + M_q\sum_{m=1}^n \alpha_m^{1/2-1/q}\right)}\right).$$

Lemma A.2. *Suppose Conditions D, S, and I hold true. Then*

(i) *for all $x \in \mathcal{X}$ and n large enough, $\lambda_b' g_n(x) \in \mathcal{C}$, where $\mathcal{C}$ is a compact set in $\mathbb{R}$,*

(ii) $\sup_{x\in\mathcal{X}_n} |\omega_0(x) - \phi_*^{(1)}(\lambda_b' g_n(x))| = O(\eta_{K,n})$.

Lemma A.3. *Suppose Conditions for Theorem 1.1 hold true. Then*

(i) *If we additionally assume that $\{X_i\}_{i=1}^n$ is iid and $\zeta_{K,n}^2 \log K/n \to 0$, then $|\mathbb{E}_n[g_n(X)g_n(X)'] - I| = O_p(\sqrt{\zeta_{K,n}^2 \log K/n})$, and thus $\lambda_{\min}(\mathbb{E}_n[g_n(X)g_n(X)'])$ is bounded away from zero and from above with probability approaching to one.*

(ii) $|\mathbb{E}_n[r_n(X) - \omega_0(X)g_n(X)]| = O_p(\sqrt{K\mu_{K,n}/n})$,

(iii) $|\mathbb{E}_n[\{\omega_0(X) - \phi_*^{(1)}(\lambda_b' g_n(X))\}g_n(X)]| = O_p(B_{K,n})$.

(iv) $|\tilde{\lambda} - \lambda_b| = O_p(\sqrt{K\mu_{K,n}/n} + B_{K,n})$, *where* $\tilde{\lambda} = \arg\min_{\lambda} \mathbb{E}_n[\phi_*(\lambda' g_n(X)) - \lambda' r_n(X)]$.

Proof of Lemma A.1

Let $W(X) = f(X) - \mathbb{E}[f(X)]$. Note that

$$\mathbb{E}[|\mathbb{E}_n[W(X_i)]|^2] = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^K \mathbb{E}[W_j(X_i)^2] + \frac{1}{n^2}\sum_{i\neq l}^n\sum_{j=1}^K \mathbb{E}[W_j(X_i)W_j(X_l)].$$

The first term is bounded as $\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^K \mathbb{E}[W_j^2(X_i)] \le KM_2^2/n$. For the second term, (Hall et al., 1980, Corollary A.2) implies

$$|\mathbb{E}[W_j(X_i)W_j(X_l)]| \lesssim \{\mathbb{E}[|W_j(X_i)|^q]\}^{1/q}\sqrt{\mathbb{E}[W_j(X_l)^2]}\alpha_{i-l}^{1/2-1/q} \le M_q M_2 \alpha_{i-l}^{1/2-1/q},$$

and thus $\frac{1}{n^2}\sum_{i\neq l}^{n}\sum_{j=1}^{K}\mathbb{E}[W_j(X_i)W_j(X_l)] \lesssim KM_qM_2\sum_{m=1}^{n}\alpha_m^{1/2-1/q}$. Therefore, the conclusion follows by Markov's inequality.

Proof of Lemma A.2 (i)

By boundedness of ω_0 (Condition D) and continuity of $[\phi_*^{(1)}]^{-1}(\cdot)$ (Condition I), both $\underline{\gamma} = \inf_{x\in\mathcal{X}}[\phi_*^{(1)}]^{-1}(\omega_0(x))$ and $\overline{\gamma} = \sup_{x\in\mathcal{X}}[\phi_*^{(1)}]^{-1}(\omega_0(x))$ are finite. Thus, by (1.12) in Condition S, there exists $C_1 > 0$ such that

$$\lambda_b' g_n(x) \in [\underline{\gamma} - C_1\eta_{K,n}, \overline{\gamma} + C_1\eta_{K,n}], \tag{A.1}$$

for all $x \in \mathcal{X}_n$ (and thus for all $x \in \mathcal{X}$) and n large enough. The conclusion follows by the requirement $\eta_{K,n} \to 0$.

Proof of Lemma A.2 (ii)

Note (A.1) also guarantees

$$\begin{aligned}\omega_0(x) - \phi_*^{(1)}(\lambda_b' g_n(x)) \in [&\phi_*^{(1)}(\lambda_b' g_n(x) - C_1\eta_{K,n}) - \phi_*^{(1)}(\lambda_b' g_n(x)),\\ &\phi_*^{(1)}(\lambda_b' g_n(x) + C_1\eta_{K,n}) - \phi_*^{(1)}(\lambda_b' g_n(x))],\end{aligned}$$

for all $x \in \mathcal{X}_n$ and n large enough. By applying the mean value theorem to the upper and lower bounds under Condition I, there exist $c_1, c_2 > 0$ such that

$$\begin{aligned}\phi_*^{(1)}(\lambda_b' g_n(x) + C_1\eta_{K,n}) - \phi_*^{(1)}(\lambda_b' g_n(x)) &\leq c_1 C_1\eta_{K,n},\\ \phi_*^{(1)}(\lambda_b' g_n(x) - C_1\eta_{K,n}) - \phi_*^{(1)}(\lambda_b' g_n(x)) &\geq -c_2 C_1\eta_{K,n},\end{aligned}$$

for all $x \in \mathcal{X}_n$ and n large enough. Combining these results, the conclusion follows.

Proof of Lemma A.3 (i)

This follows directly from (Belloni et al., 2015, Lemma 6.2) or (Chen and Christensen, 2015, Lemma 2.1).

Proof of Lemma A.3 (ii)

Let $f(x) = r_n(x) - \omega_0(x)g_n(x)$. By (1.1) and Cauchy-Schwarz inequality, we have

$$\begin{aligned}|\mathbb{E}[f(X)]| &\lesssim |\mathbb{E}[\{\omega_0(X)g(X) - r(X)\}\mathbb{I}\{X \notin \mathcal{X}_n\}]|\\ &\leq \sqrt{\mathbb{E}[|\omega_0(X)g(X) - r(X)|^2]}\sqrt{\mathbb{P}\{X \notin \mathcal{X}_n\}} = o(\sqrt{K/n}),\end{aligned} \tag{A.2}$$

where the equality follows from Condition S. Condition S guarantees $\max_{1\leq j\leq K}\{\mathbb{E}[|f_j(X)|^q]\}^{1/q} \lesssim M_{K,n}$. Thus, Lemma A.1 implies

$$|\mathbb{E}_n[f(X)] - \mathbb{E}[f(X)]| = O_p(\sqrt{K\mu_{K,n}/n}). \tag{A.3}$$

The conclusion follows by (A.2) and (A.3).

Proof of Lemma A.3 (iii)

Let $\xi(X) = \{\omega_0(X) - \phi_*^{(1)}(\lambda_b' g_n(X))\}$ and $\hat{\rho} = (\mathbb{E}_n[g_n(X)g_n(X)'])^{-1}\mathbb{E}_n[g_n(X)\xi(X)]$. By the assumption $|\mathbb{E}_n[g_n(X)g_n(X)'] - I| = o_p(1)$, it holds $(\mathbb{E}_n[g_n(X)g_n(X)'])^{-1} = O_p(1)$, and then

$$|\mathbb{E}_n[g_n(X)\xi(X)]| \leq |\mathbb{E}_n[g_n(X)g_n(X)']||\hat{\rho}| \lesssim |\hat{\rho}| \lesssim \sqrt{\mathbb{E}_n[(\hat{\rho}' g_n(X))^2]}, \tag{A.4}$$

with probability approaching one, where the last inequality follows from Condition S. Since $\hat{\rho}$ is the empirical projection coefficient from $\xi(X)$ on $g_n(X)$, we have

$$\mathbb{E}_n[(\hat{\rho}' g_n(X))^2] \leq \{\mathbb{E}_n[\xi(X)^2] - \mathbb{E}[\xi(X)^2]\} + \mathbb{E}[\xi(X)^2] = O_p(B_{K,n}^2), \tag{A.5}$$

where the equality follows from (1.13) in Condition S and Lemma A.1 (note that $\mathbb{E}[|\xi(X)|^q] \lesssim \varsigma_{K,n}^{2/q}$ under Conditions D and S). The conclusion follows from (A.4) and (A.5).

Proof of Lemma A.3 (iv)

Recall that $\hat{\omega}(X) = \phi_*^{(1)}(\hat{\lambda}' g(X)\mathbb{I}\{X \in \mathcal{X}_n\}) = \phi_*^{(1)}(\tilde{\lambda}' g_n(X))$, where $\tilde{\lambda} = \arg\max_{\lambda} \hat{Q}(\lambda)$ and

$$\hat{Q}(\lambda) = \lambda' \mathbb{E}_n[r_n(X)] - \mathbb{E}_n[\phi_*(\lambda' g_n(X))].$$

By Condition I, $\hat{Q}(\lambda)$ is strictly concave in λ. Let $\hat{Q}^{(1)}(\lambda)$ and $\hat{Q}^{(2)}(\lambda)$ be the first and second derivatives of $\hat{Q}(\lambda)$, respectively. The proof is split into several steps.

Step 1: Show $\hat{Q}^{(1)}(\lambda_b) = O_p(\delta_n)$, where $\delta_n = \sqrt{K\mu_{K,n}/n} + B_{K,n}$. Since $\hat{Q}^{(1)}(\lambda_b) = \mathbb{E}_n[r_n(X) - \phi_*^{(1)}(\lambda_b' g_n(X))g_n(X)]$, the triangle inequality yields

$$|\hat{Q}^{(1)}(\lambda_b)| \leq |\mathbb{E}_n[r_n(X) - \omega_0(X)g_n(X)]| + |\mathbb{E}_n[\{\omega_0(X) - \phi_*^{(1)}(\lambda_b' g_n(X))\}g_n(X)]|$$

Thus, Lemma A.3 (ii) and (iii) imply $\hat{Q}^{(1)}(\lambda_b) = O_p(\delta_n)$.

Step 2: Show that for any $C > 0$, there exists some $c > 0$ such that $\eta_C = \inf_{|\lambda - \lambda_b| \leq C\delta_n, x \in \mathcal{X}} \phi_*^{(2)}(\lambda' g_n(x)) > c$. This is trivially true under Condition I (i), so we

focus on the case of Condition I (ii) (i.e., $\delta_n \zeta_{K,n} = O(1)$). Pick any $C > 0$. Note that

$$|\lambda' g_w(x)| \leq |\lambda_b' g_n(x)| + |\lambda - \lambda_b||g_n(x)| \leq |\lambda_b' g_n(x)| + C\delta_n \zeta_{K,n},$$

for all λ satisfying $|\lambda - \lambda_b| \leq C\delta_n$. Thus, by Lemma A.2 (i), $\lambda' g_w(x)$ lies in some compact set $\tilde{\mathcal{C}}$ for all λ satisfying $|\lambda - \lambda_b| \leq C\delta_n$ and $x \in \mathcal{X}$. Since $\phi_*^{(2)}$ is continuous (by Condition I), Wierstrass theorem guarantees $\eta_C > c = \min_{a \in \tilde{\mathcal{C}}} \phi_*^{(2)}(a) > 0$.

Step 3: Show there exists some $C^* > 0$ such that $\hat{Q}(\lambda) < \hat{Q}(\lambda_b)$ with probability approaching one for all λ satisfying $|\lambda - \lambda_b| = C^* \delta_n$. Pick any $\epsilon > 0$. By Step 1, we can take $C^* > 0$ such that

$$\mathbb{P}\{|\hat{Q}^{(1)}(\lambda_b)| < cC^*\delta_n/4\} \geq 1 - \epsilon, \tag{A.6}$$

for all n large enough, where $c > 0$ is chosen in Step 2. An expansion of $\hat{Q}(\lambda)$ around $\lambda = \lambda_b$ yields

$$\hat{Q}(\lambda) - \hat{Q}(\lambda_b) = \hat{Q}^{(1)}(\lambda_b)'(\lambda - \lambda_b) + \frac{1}{2}(\lambda - \lambda_b)' \hat{Q}^{(2)}(\dot{\lambda})(\lambda - \lambda_b),$$

for some $\dot{\lambda}$ on the line joining λ and λ_b. By Step 2,

$$\hat{Q}^{(2)}(\dot{\lambda}) = -\mathbb{E}_n[\phi_*^{(2)}(\dot{\lambda}' g_n(X)) g_n(X) g_n(X)'] \leq_{\text{psd}} -c\mathbb{E}_n[g_n(X) g_n(X)'],$$

and Condition S(2) implies

$$\frac{1}{2}(\lambda - \lambda_b)' \hat{Q}^{(2)}(\dot{\lambda})(\lambda - \lambda_b) \leq -\frac{c}{4}|\lambda - \lambda_b|^2,$$

with probability approaching one. Combining these results, for all λ satisfying $|\lambda - \lambda_b| = C^* \delta_n$,

$$\hat{Q}(\lambda) - \hat{Q}(\lambda_b) \leq |\hat{Q}^{(1)}(\lambda_b)||\lambda - \lambda_b| - \frac{c}{4}|\lambda - \lambda_b|^2 \leq \left(|\hat{Q}^{(1)}(\lambda_b)| - \frac{cC^*\delta_n}{4} \right) |\lambda - \lambda_b|.$$

Thus, (A.6) implies that $\hat{Q}(\lambda) < \hat{Q}(\lambda_b)$ with probability approaching one.

Step 4: By continuity of $\hat{Q}(\lambda)$, it has a maximum on the compact set $\{\lambda : |\lambda - \lambda_b| \leq C^* \delta_n\}$. By Step 3, the maximum $\tilde{\lambda}_{C^*}$ on set $\{\lambda : |\lambda - \lambda_b| \leq C^* \delta_n\}$ must satisfy $|\tilde{\lambda}_{C^*} - \lambda_b| < C^* \delta_n$. By concavity of $\hat{Q}(\lambda)$, $\tilde{\lambda}_{C^*}$ also maximizes $\hat{Q}(\lambda)$ over $\mathbb{R}^k$. The conclusion follows by the same argument used at the end of the proof of (Newey and McFadden, 1994, Theorem 2.7).

A.2.2 Proof of Theorem 1.1

Proof of (1.14)

Let $\omega_b(x) = \phi^{(1)}{}_*(\lambda_b' g_n(x))$. By an expansion around $\tilde{\lambda} = \lambda_b$,

$$\hat{\omega}(x) - \omega_b(x) \;=\; \phi_*^{(2)}(\bar{\lambda}_x' g_n(x))(\tilde{\lambda} - \lambda_b)' g_n(x), \tag{A.7}$$

where $\bar{\lambda}_x$ is a point on the line joining $\tilde{\lambda}$ and λ_b. Pick any $C > 0$. From Step 2 in the proof of Lemma A.3 (iv), $\lambda' g_n(x)$ lies in some compact set $\tilde{\mathcal{C}}$ for all $x \in \mathcal{X}$ and λ satisfying $|\lambda - \lambda_b| \leq C\delta_n$. Weierstrass theorem and Condition I imply

$$\sup_{|\lambda - \lambda_b| \leq C\delta_n, x \in \mathcal{X}} \phi_*^{(2)}(\lambda' g_n(x)) < C_1 < \infty, \tag{A.8}$$

for some $C_1 > 0$.

For the compact set $\tilde{\mathcal{C}}$ defined above, let $\mathcal{E}_n$ be the event that $\bar{\lambda}_x' g_n(x) \in \tilde{\mathcal{C}}$ for all $x \in \mathcal{X}$. Lemma A.3 (iv) guarantees $\mathbb{P}\{\mathcal{E}_n\} \to 1$. Observe that

$$\begin{aligned} \mathbb{E}_n[\{\hat{\omega}(X) - \omega_b(X)\}^2] &= (\tilde{\lambda} - \lambda_b)' \mathbb{E}_n[\{\phi_*^{(2)}(\bar{\lambda}_X' g_n(X))\}^2 g_n(X) g_n(X)'](\tilde{\lambda} - \lambda_b) \\ &\leq C_1 |\tilde{\lambda} - \lambda_b|^2 |\mathbb{E}_n[g_n(X) g_n(X)']| = O_p(|\tilde{\lambda} - \lambda_b|^2), \end{aligned} \tag{A.9}$$

where the inequality follows from (A.8) and $\mathbb{P}\{\mathcal{E}_n\} \to 1$, and the second equality follows from Condition S and Lemma A.3 (iv). Now, the argument in the proof of Lemma A.3 (iii) for (A.5) yields

$$\mathbb{E}_n[\{\omega_b(X) - \omega_0(X)\}^2] = O_p(B_{K,n}^2). \tag{A.10}$$

The conclusion follows by (A.9), (A.10), and the triangle inequality.

Proof of $\hat{\theta} \xrightarrow{p} \theta_0$

Observe that

$$\begin{aligned} & |\hat{\theta} - \theta_0| \\ \leq\; & |\mathbb{E}_n[\hat{\omega}(X) h(X,Y)] - \mathbb{E}_n[\omega_0(X) h(X,Y)]| + |\mathbb{E}_n[\omega_0(X) h(X,Y)] - \mathbb{E}[\omega_0(X) h(X,Y)]| \\ \leq\; & \sqrt{\mathbb{E}_n[\{\hat{\omega}(X) - \omega_0(X)\}^2]} \sqrt{\mathbb{E}_n[h(X,Y)^2]} + |\mathbb{E}_n[\omega_0(X) h(X,Y)] - \mathbb{E}[\omega_0(X) h(X,Y)]| \\ =\; & O_p(\sqrt{K\mu_{K,n}/n} + B_{K,n}) + o_p(1) \end{aligned}$$

where the first inequality follows from the triangle inequality, the second inequality follows from Cauchy-Schwarz inequality, the final equality follows from the law of large numbers (under Condition D) for stationary and ergodic processes

and (1.14) in Theorem 1.1.

Proof of (1.15)

By the triangle inequality,

$$\sup_{x\in\mathcal{X}_n}|\hat{\omega}(x)-\omega_0(x)| \le \sup_{x\in\mathcal{X}_n}|\hat{\omega}(x)-\omega_b(x)| + \sup_{x\in\mathcal{X}_n}|\omega_b(x)-\omega_0(x)|.$$

From the proof of (1.14), it is easy to see that $\sup_{x\in\mathcal{X}_n}|\hat{\omega}(x)-\omega_b(x)| = O_p(\zeta_{K,n}(\sqrt{K\mu_{K,n}/n}+B_{K,n}))$. Thus, the conclusion follows by Lemma A.2 (ii).

A.2.3 Proof of Theorem 1.2

Let

$$\begin{aligned} &h_i = h(X_i,Y_i), \qquad h_i^X = \mathbb{E}[h_i|X_i], \qquad \omega_{0i} = \omega_0(X_i),\\ &g_{ni} = g_n(X_i), \qquad \omega_{bi} = \phi_*^{(1)}(\lambda_b' g_{ni}), \qquad \hat{\omega}_i = \phi_*^{(1)}(\tilde{\lambda}' g_{ni}),\\ &r_{ni} = r_n(X_i), \qquad r_i^h = r^h(X_i). \end{aligned} \tag{A.11}$$

By an expansion of $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n \phi_*^{(1)}(\tilde{\lambda}' g_{ni})h_i$ around $\tilde{\lambda}=\lambda_b$, we decompose

$$\sqrt{n}(\hat{\theta}-\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n(\omega_{0i}h_i-\theta_0) + T_1+T_2+T_3+T_4,$$

where

$$\begin{aligned} T_1 &= \mathbb{E}[\phi_*^{(2)}(\lambda_b' g_{ni})h_i g_{ni}']\sqrt{n}(\tilde{\lambda}-\lambda_b),\\ T_2 &= \frac{1}{\sqrt{n}}\sum_{i=1}^n\{\phi_*^{(2)}(\lambda_b' g_{ni})h_i g_{ni}' - \mathbb{E}[\phi_*^{(2)}(\lambda_b' g_{ni})h_i g_{ni}']\}(\tilde{\lambda}-\lambda_b),\\ T_3 &= \frac{1}{2}(\tilde{\lambda}-\lambda_b)'\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n\phi_*^{(3)}(\dot{\lambda}' g_{ni})h_i g_{ni}g_{ni}'\right)(\tilde{\lambda}-\lambda_b), \qquad T_4 = \frac{1}{\sqrt{n}}\sum_{i=1}^n(\omega_{bi}h_i-\omega_{0i}h_i), \end{aligned}$$

and $\dot{\lambda}$ lies on the line joining $\tilde{\lambda}$ and λ_b.

First, we consider T_2. Since Lemma A.2 (i) and Assumption N imply

$$\max_{1\le j\le K}\{\mathbb{E}[|\phi_*^{(2)}(\lambda_b' g_n)h g_{nj}|^2]\} \lesssim 1$$

and $\max_{1\leq j\leq K}\{\mathbb{E}[|\phi_*^{(2)}(\lambda_b' g_w)h g_{nj}|^{q_1}]\}^{1/q_1} \lesssim M_{K,n}$, Lemma A.1 yields

$$\left|\frac{1}{n}\sum_{i=1}^{n}\{\phi_*^{(2)}(\lambda_b' g_{ni})h_i g_{ni}' - \mathbb{E}[\phi_*^{(2)}(\lambda_b' g_{ni})h_i g_{ni}']\}\right| = O_p\left(\sqrt{\frac{K\mu_{K,n}}{n}}\right).$$

Thus, Cauchy Schwarz inequality and Lemma A.3 (iv) imply $T_2 = O_p(\sqrt{K\mu_{K,n}}(\sqrt{K\mu_{K,n}/n}+ B_{K,n}))$.

Next, we consider T_3. The definitions of $\zeta_{K,n}$ and matrix L_2 norm, Lemmas A.2 (i) and A.3 (iv), and Condition I imply $|\frac{1}{n}\sum_{i=1}^{n}\phi_*^{(3)}(\dot{\lambda}' g_{ni})h_i g_{ni} g_{ni}'| = O_p(\zeta_{K,n}^2)$. Thus, Cauchy Schwarz inequality and Lemma A.3 (iv) imply

$$T_3 = O_p(\sqrt{n}\zeta_{K,n}^2(K\mu_{K,n}/n + B_{K,n}^2)).$$

Third, we consider T_4. From the proof of Lemma A.3 (iii), and the law of large numbers, we have $T_4 = O_p(\sqrt{n}B_{K,n})$.

We now consider T_1. By expanding the first order condition of $\tilde{\lambda}$,

$$0 = \frac{1}{n}\sum_{i=1}^{n}\{\phi_*^{(1)}(\tilde{\lambda}' g_{ni})g_{ni} - r_{ni}\} = \frac{1}{n}\sum_{i=1}^{n}(\omega_{bi}g_{ni} - r_{ni}) + \frac{1}{n}\sum_{i=1}^{n}\phi_*^{(2)}(\bar{\lambda}' g_{ni})g_{ni}g_{ni}'(\tilde{\lambda} - \lambda_b), \tag{A.12}$$

where $\bar{\lambda}$ lies on the line joining $\tilde{\lambda}$ and λ_b. Let $\psi = \mathbb{E}[\phi_*^{(2)}(\lambda_b' g_{ni})h_i g_{ni}']$, $\Sigma = \mathbb{E}[\phi_*^{(2)}(\lambda_b' g_{ni})g_{ni}g_{ni}']$, and $\bar{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\phi_*^{(2)}(\bar{\lambda}' g_{ni})g_{ni}g_{ni}'$. By solving this for $\tilde{\lambda} - \lambda_b$ and inserting to T_1, we have

$$T_1 = -\psi\bar{\Sigma}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\omega_{bi}g_{ni} - r_{ni}) = T_{11} + T_{12} + T_{13},$$

where

$$\begin{aligned} T_{11} &= -\psi(\bar{\Sigma}^{-1} - \Sigma^{-1})\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\omega_{bi}g_{ni} - r_{ni}), \\ T_{12} &= -\psi\Sigma^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\omega_{bi} - \omega_{0i})g_{ni}, \qquad T_{13} = -\psi\Sigma^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\omega_{0i}g_{ni} - r_{ni}). \end{aligned}$$

For T_{12}, note that

$$|T_{12}| \leq |\psi|\frac{1}{\lambda_{\min}(\Sigma)}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\omega_{bi} - \omega_{0i})g_{ni}\right|.$$

It is easy to see $|\psi| = O(\zeta_{K,n})$ due to the definition of $\zeta_{K,n}$. Lemma A.3 (iii) yields $\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\omega_{bi} - \omega_{0i})g_{ni}\right| = O_p(\sqrt{n}B_{K,n})$. Since $\lambda_{\min}(\Sigma)$ is bounded away from zero by Condition D and Lemma A.2 (i), we have $T_{12} = O_p(\sqrt{n}\zeta_{K,n}B_{K,n})$.

For T_{11}, note that (A.12) implies

$$T_{11} = \sqrt{n}\psi(\bar{\Sigma}^{-1} - \Sigma^{-1})\bar{\Sigma}(\tilde{\lambda} - \lambda_b) = \sqrt{n}\psi\Sigma^{-1}(\Sigma - \bar{\Sigma})(\tilde{\lambda} - \lambda_b),$$

which can be bounded as $|T_{11}| \leq \sqrt{n}|\psi|\frac{1}{\lambda_{\min}(\Sigma)}|\Sigma - \bar{\Sigma}| \cdot |\tilde{\lambda} - \lambda_b|$. By the triangle inequality and Condition (1) in Theorem 1.2,

$$|\Sigma - \bar{\Sigma}| \leq |\mathbb{E}_n[(\phi_*^{(2)}(\bar{\lambda}'g_n) - \phi_*^{(2)}(\lambda_b'g_n))g_n g_n']| + O_p(\Gamma_{K,n}).$$

By an expansion of $\phi_*^{(2)}(\bar{\lambda}'g_{ni})$, Lemma A.2 (i) and Lemma A.3 (iv), we have $|\mathbb{E}_n[(\phi_*^{(2)}(\bar{\lambda}'g_n) - \phi_*^{(2)}(\lambda_b'g_n))g_n g_n']| = O_p(\zeta_K^3(\sqrt{K\mu_{K,n}/n} + B_{K,n}))$. Therefore, we obtain

$$|\Sigma - \bar{\Sigma}| = O_p(\zeta_{K,n}^3(\sqrt{K\mu_{K,n}/n} + B_{K,n}) + \Gamma_{K,n}).$$

Also by $|\psi| = O(\zeta_{K,n})$ and Lemma A.3 (iv), we have

$$|T_{11}| = O_p\left(\sqrt{n}\zeta_{K,n}^4(K\mu_{K,n}/n + B_{K,n}^2) + \sqrt{n}\zeta_{K,n}\Gamma_{K,n}(\sqrt{K\mu_{K,n}/n} + B_{K,n})\right).$$

Now consider T_{13}. Note that

$$\begin{aligned} T_{13} &= -\frac{1}{\sqrt{n}}\sum_{i=1}^n(\omega_{0i}h_i^X - r_i^h) - \frac{1}{\sqrt{n}}\sum_{i=1}^n\{\beta'(\omega_{0i}g_{ni} - r_{ni}) - (\omega_{0i}h_i^X - r_i^h)\} \\ &= -\frac{1}{\sqrt{n}}\sum_{i=1}^n(\omega_{0i}h_i^X - r_i^h) + o_p(1), \end{aligned}$$

where the second equality follows from Lemma A.1 and the condition (1.16).

Combining these results, we obtain

$$\sqrt{n}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n\{\omega_{0i}h_i - \theta_0 - (\omega_{0i}h_i^X - r_i^h)\} + O_p(r_n),$$

where $r_n = (\sqrt{n}(\zeta_{K,n}^4 K\mu_{K,n}/n + \zeta_{K,n}B_{K,n} + \sqrt{K\mu_{K,n}/n}\zeta_{K,n}\Gamma_{K,n}))$. Since $r_n \to 0$ by the assumption, the central limit theorem for α mixing processes (for example, Theorem 0 in Bradley et al., 1985) yields the conclusion.

A.2.4 Proof of Proposition 1.1

Proof of (i)

In this case, $r(X)$ is a constant vector $r = \mathbb{E}[\omega_{0i} g_i]$. We set $r^h(X)$ as a constant vector $r^h = \mathbb{E}[\omega_{0i} h_i^X]$. Observe that

$$\mathbb{E}[\beta'(\omega_{0i} g_{ni} - r_{ni}) - (\omega_{0i} h_i^X - \mathbb{E}[\omega_{0i} h_i^X])]^2 \leq N_1 + N_2 + N_3,$$

where

$$\begin{aligned} N_1 &= \mathbb{E}[\beta'(\omega_{0i} g_{ni} - \mathbb{E}[\omega_{0i} g_{ni}]) - (\omega_{0i} h_i^X - \mathbb{E}[\omega_{0i} h_i^X])]^2, \\ N_2 &= \mathbb{E}[\beta'(\mathbb{E}[\omega_{0i} g_{ni}] - \mathbb{E}[r_{ni}])]^2, \qquad N_3 = \mathbb{E}[\beta'(\mathbb{E}[r_{ni}] - r_{ni})]^2. \end{aligned}$$

For N_1,

$$N_1 \leq \mathbb{E}[\omega_{0i}^2 (h_i^X - \beta' g_{ni})^2] \leq \left(\sup_{x \in \mathcal{X}} \frac{\omega_0^2(x)}{\phi_*^{(2)}(\lambda_b' g_n(x))} \right) \mathbb{E}[(\tilde{h}_i - \beta_p' \tilde{g}_{ni})^2],$$

where $\tilde{h}_i = \sqrt{\phi_*^{(2)}(\lambda_b' g_{ni})} h_i^X$, $\tilde{g}_i = \sqrt{\phi_*^{(2)}(\lambda_b' g_{ni})} g_{ni}$, and $\beta_p = \mathbb{E}[\tilde{g}_{ni} \tilde{g}_{ni}']^{-1} \mathbb{E}[\tilde{g}_{ni} \tilde{h}_i]$. Since β_p is the projection coefficient that solves $\min_b \mathbb{E}[(\tilde{h}_i - b' \tilde{g}_{ni})^2]$, the assumption in (1.18) guarantees $N_1 = o(n^{-1})$. For N_2, (A.2) implies $|\beta| = O(1)$ (because β is a projection coefficient). By (1.18), we have

$$N_2 \lesssim \mathbb{E}[|\omega_0(X) g(X) - r(X)|^2] \mathbb{P}\{X \notin \mathcal{X}_n\} = o(n^{-1}).$$

For N_3, the definition of r_{ni}, $|\beta| = O(1)$, and (1.18) imply

$$N_3 = \mathbb{E}[\beta'(r_{ni} - \mathbb{E}[r_{ni}])]^2 \lesssim |\beta|^2 K \mathbb{P}\{X \in \mathcal{X}_n\} \mathbb{P}\{X \notin \mathcal{X}_n\} = o(n^{-1}).$$

Combining these results, the conclusion follows.

Proof of (ii)

This follows by a standard projection argument and thus the proof is omitted.

A.3 Proofs for high dimensional case

A.3.1 Proof of Theorem 1.3

By the mean value theorem, there exists $t_x \in [0, 1]$ such that

$$\hat{\omega}(x) - \omega_{\mathbf{o}}(x) = \phi_*^{(2)}(\lambda_{\mathbf{o}}' g(x) + t_x(\hat{\lambda} - \lambda_{\mathbf{o}})' g(x))(\hat{\lambda} - \lambda_{\mathbf{o}})' g(x), \quad \text{(A.13)}$$

for each $x \in \mathcal{X}$.

First, consider the case of Condition I (ii), i.e., $\tilde{\zeta}_K \kappa_{\mathbf{o},n} \lesssim 1$. Hölder's inequality and Lemma A.4 (ii) imply

$$\sup_{x \in \mathcal{X}} |t_x(\hat{\lambda} - \lambda_{\mathbf{o}})' g(x)| \leq \left\|\hat{\lambda} - \lambda_{\mathbf{o}}\right\|_1 \tilde{\zeta}_K = O_p(\tilde{\zeta}_K \kappa_{\mathbf{o},n}) = O_p(1). \tag{A.14}$$

The assumption $\sup_{x \in \mathcal{X}} |\omega_{\mathbf{o}}(x) - \omega_0(x)| \lesssim 1$ and (A.14) imply $\mathbb{P}\{\mathcal{E}_n\} \to 1$, where $\mathcal{E}_n$ is the event that $\phi_*^{(2)}(\lambda_{\mathbf{o}}' g(x) + t_x(\hat{\lambda} - \lambda_{\mathbf{o}})' g(x))$ lies in a bounded set for all $x \in \mathcal{X}$. On the event $\mathcal{E}_n$, (A.13) and (A.14) imply

$$\begin{aligned} \mathbb{E}_n[\{\hat{\omega}(X) - \omega_{\mathbf{o}}(X)\}^2] &\lesssim (\hat{\lambda} - \lambda_{\mathbf{o}})' \mathbb{E}_n[g(X)g(X)'](\hat{\lambda} - \lambda_{\mathbf{o}}) \\ &\leq \left\|\hat{\lambda} - \lambda_{\mathbf{o}}\right\|_1^2 \left\|\mathbb{E}_n[g(X)g(X)']\right\|_\infty = O_p(\kappa_{\mathbf{o}n}^2 \xi_n), \end{aligned}$$

where the second inequality follows from Hölder's inequality and the equality follows from Lemma A.4 (ii) and the definition of ξ_n.

Now consider the case of Condition I(i), i.e., $\phi_*^{(2)}$ is bounded from above and away from zero. In this case, it is easy to see that we still have $\mathbb{E}_n[\{\hat{\omega}(X) - \omega_{\mathbf{o}}(X)\}^2] = O_p(\kappa_{\mathbf{o}n}^2 \xi_n)$ from (A.13).

Therefore for both cases, on the event $\mathcal{E}_n$, the triangle inequality, the result $\mathbb{E}_n[\{\hat{\omega}(X) - \omega_{\mathbf{o}}(X)\}^2] = O_p(\kappa_{\mathbf{o}n}^2 \xi_n)$, and the assumption $\sqrt{\mathbb{E}[\{\omega_{\mathbf{o}}(X) - \omega_0(X)\}^2]} \lesssim \varsigma_{\mathbf{o},n}$ yield the conclusion in (1.20).

Proofs of $\hat{\theta} \xrightarrow{p} \theta_0$ and (1.21) are similar to those of Theorem 1.1, and thus omitted.

A.3.2 Proof of Theorem 1.4

We employ the notation in (A.11). By the Karush-Kuhn-Tucker (KKT) condition of $\hat{\lambda}$ in (1.11) for the high dimensional case, an expansion around $\hat{\lambda} = \lambda_{\mathbf{o}}$ yields

$$0 = Q_n^{(1)}(\hat{\lambda}) + \alpha_n \hat{\kappa} = Q_n^{(1)}(\lambda_{\mathbf{o}}) + c_* \mathbb{E}_n[g(X)g(X)'](\hat{\lambda} - \lambda_{\mathbf{o}}) + \alpha_n \hat{\kappa},$$

where $Q_n(\lambda) = \mathbb{E}_n[\phi_*(\lambda' g(X)) - \lambda' r(X)]$ and $Q_n^{(1)}(\lambda) = \mathbb{E}_n[\phi_*^{(1)}(\lambda' g(X)) g(X) - r(X)]$ is its first derivative. Since $\omega_{\mathbf{o}}(\cdot) = \phi_*^{(1)}(\lambda_{\mathbf{o}}' g(\cdot))$, an expansion of $\frac{1}{n}\sum_{i=1}^n \phi_*^{(1)}(\hat{\lambda}' g_i) h_i$ around $\hat{\lambda} = \lambda_{\mathbf{o}}$ yields

$$\hat{\theta}_{DB} = \frac{1}{n}\sum_{i=1}^n \omega_{\mathbf{o}i} h_i + \frac{1}{n}\sum_{i=1}^n c_* h_i g_i' \{(\hat{\lambda} - \lambda_{\mathbf{o}}) + \alpha_n \hat{\Theta} \hat{\kappa}\}.$$

By plugging in the form of $\alpha_n\hat{\kappa}$ from the KKT condition to the above equation, we obtain

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n} c_* h_i g_i'\{(\hat{\lambda}-\lambda_{\mathbf{o}}) + \alpha_n\hat{\Theta}\hat{\kappa}\} \\
= & \frac{1}{\sqrt{n}}\sum_{i=1}^{n} c_* h_i g_i'\{(\hat{\lambda}-\lambda_{\mathbf{o}}) - \hat{\Theta}[Q_n^{(1)}(\lambda_{\mathbf{o}}) + \mathbb{E}_n[g(X)g(X)'](\hat{\lambda}-\lambda_{\mathbf{o}})]\} \\
= & -\frac{1}{\sqrt{n}}\sum_{i=1}^{n} c_* h_i g_i'\hat{\Theta}\mathbb{E}_n[\omega_{\mathbf{o}}(X)g(X) - r(X)] + T_\triangle,
\end{aligned}$$

where $T_\triangle = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} c_* h_i g_i'(I - \mathbb{E}_n[g(X)g(X)']\hat{\Theta})(\hat{\lambda}-\lambda_{\mathbf{o}})$. Combining these results and the definition of $\hat{\beta}_{DB}$, we obtain the following decomposition

$$\sqrt{n}(\hat{\theta}_{DB}-\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{r_i^h - \theta_0 + \omega_{0i}(h_i - h_i^X)\} + T_1 + T_2 + T_3 + T_4 + T_5 + T_\triangle,$$

where

$$\begin{aligned}
T_1 &= -c_*\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\hat{\beta}_{DB}'(\omega_{0i}g_i - r_i) - (\omega_{0i}\tilde{h}_i^X - \tilde{r}_i^h)], \\
T_2 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\omega_{\mathbf{o}i}-\omega_{0i})(\tilde{h}^X - \hat{\beta}_{DB}'g_i), \qquad T_3 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\omega_{\mathbf{o}i}-\omega_{0i})(h_i^X - \tilde{h}_i^X), \\
T_4 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\omega_{\mathbf{o}i}-\omega_{0i})(h_i - h_i^X), \qquad T_5 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\omega_{0i}(h_i^X - \tilde{h}_i^X) + (\tilde{r}_i^h - r_i^h)].
\end{aligned}$$

Condition DB guarantees $T_1 \xrightarrow{p} 0$. By Cauchy-Schwarz inequality,

$$|T_2| \le \sqrt{n}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\omega_{\mathbf{o}i}-\omega_{0i})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\tilde{h}^X - \hat{\beta}_{DB}'g_i)^2} \xrightarrow{p} 0,$$

where the equality follows from Chebychev's inequality for the term $\frac{1}{n}\sum_{i=1}^{n}(\omega_{\mathbf{o}i} - \omega_{0i})^2$ and Condition DB.

For T_3, Cauchy-Schwarz inequality and the assumptions in the theorem imply $\mathbb{E}[T_3] \lesssim \sqrt{n}\varsigma_n\tau_n \to 0$. Also, by Chebychev's inequality implies $T_3 - \mathbb{E}[T_3] \xrightarrow{p} 0$. Combining these results, we obtain $T_3 \xrightarrow{p} 0$. Note that both T_4 and T_5 have zero mean. Thus, Chebychev's inequality implies $T_4 = O_p(\varsigma_n) = o_p(1)$ and $T_5 = O_p(\tau_n) = o_p(1)$. Finally, by Hölder's inequality, we have

$$T_\triangle \lesssim \sqrt{n}\left\|\frac{1}{n}\sum_{i=1}^{n} h_i g_i\right\|_\infty \left\|I - \mathbb{E}_n[g(X)g(X)']\hat{\Theta}\right\|_1 \left\|\hat{\lambda}-\lambda_{\mathbf{o}}\right\|_1 = o_p(1),$$

under the assumptions of this theorem .

Combining these results, we obtain

$$\sqrt{n}(\hat{\theta}_{DB} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{r_i^h - \theta_0 + \omega_{0i}(h_i - h_i^X)\} + o_p(1),$$

and the conclusion follows by a central limit theorem.

A.3.3 Proof of Theorem 1.5

First, we show $\left|\hat{\Lambda} - \Lambda_*\right| = O_p(\gamma_n)$, where $\gamma_n = \sqrt{\frac{\zeta_{\mathbf{s}}^2}{n}}$. Recall $\hat{\Lambda} = \arg\max_{\Lambda \in \mathbb{R}^{\mathbf{s}}} \hat{Q}_{\mathbf{s}}(\Lambda)$, where

$$\hat{Q}_{\mathbf{s}}(\Lambda) = \mathbb{E}_n[\Lambda' r_{\mathbf{s}}(X) - \phi_*(\Lambda' g_{\mathbf{s}}(X))].$$

By Condition I', $\hat{Q}_{\mathbf{s}}(\Lambda)$ is strictly concave in Λ. Note $\hat{Q}_{\mathbf{s}}^{(1)}(\Lambda_*) = O_p\left(\sqrt{\frac{\zeta_{\mathbf{s}}^2}{n}}\right)$. Indeed, $\hat{Q}_{\mathbf{s}}^{(1)}(\Lambda_*) = \mathbb{E}_n[r_{\mathbf{s}}(X) - \phi_*^{(1)}(\Lambda_*' g_{\mathbf{s}}(X)) g_{\mathbf{s}}(X)]$ and since $\Lambda_* = \arg\min_{\Lambda \in \mathbb{R}^{\mathbf{s}}} \mathbb{E}[\Lambda' r_{\mathbf{s}}(X) - \phi_*(\Lambda' g_{\mathbf{s}}(X))]$, $\mathbb{E}[r_{\mathbf{s}}(X) - \phi_*^{(1)}(\Lambda_*' g_{\mathbf{s}}(X)) g_{\mathbf{s}}(X)] = 0$. Recall $\omega_*(x) = \phi_*^{(1)}(\Lambda_*' g_{\mathbf{s}}(x))$. It follows by Assumption S and Chebyshev's inequality that $\hat{Q}_{\mathbf{s}}^{(1)}(\Lambda_*) = O_p\left(\sqrt{\frac{\zeta_{\mathbf{s}}^2}{n}}\right)$. The rest of the proof is similar to steps 2-4 in Lemma A.3 (iv) and thus is omitted. Next, by an expansion of the post info lasso estimator $\tilde{\theta} = \frac{1}{n}\sum_{i=1}^{n} \phi_*^{(1)}(\hat{\Lambda}' g_{\mathbf{s}i}) h_i$ around $\hat{\Lambda} = \Lambda_*$, we obtain

$$\sqrt{n}(\tilde{\theta} - \theta_0 + b) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\Phi_i + v_{1i} + v_{2i} + v_{3i}) + T_1 + T_2 + T_3,$$

where

$$\begin{aligned}
T_1 &= \mathbb{E}[\phi_*^{(2)}(\Lambda_*' g_{\mathbf{s}i}) h_i g_{\mathbf{s}i}]' \sqrt{n}(\hat{\Lambda} - \Lambda_*) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\omega_{*i} \tilde{h}_i^X - \tilde{r}_i^h), \\
T_2 &= \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_*^{(2)}(\Lambda_*' g_{\mathbf{s}i}) h_i g_{\mathbf{s}i} - \mathbb{E}[\phi_*^{(2)}(\Lambda_*' g_{\mathbf{s}i}) h_i g_{\mathbf{s}i}] \right)' (\hat{\Lambda} - \Lambda_*), \\
T_3 &= \frac{1}{2} (\hat{\Lambda} - \Lambda_*)' \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_*^{(3)}(\tilde{\Lambda}' g_{\mathbf{s}i}) h_i g_{\mathbf{s}i} g_{\mathbf{s}i}' \right) (\hat{\Lambda} - \Lambda_*),
\end{aligned}$$

and $\tilde{\Lambda}$ is on the line joining $\hat{\Lambda}$ and Λ_*. By Condition I' and Chebyshev's and Cauchy Schwarz inequalities,

$$|T_2| \leq \sqrt{n} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_*^{(2)}(\Lambda_*' g_{\mathbf{s}i}) h_i g_{\mathbf{s}i} - \mathbb{E}[\phi_*^{(2)}(\Lambda_*' g_{\mathbf{s}i}) h_i g_{\mathbf{s}i}] \right| |\hat{\Lambda} - \Lambda_*| = O_p(\zeta_{\mathbf{s}} \gamma_n).$$

For T_3, similarly we have

$$|T_3| \leq \sqrt{n}|\hat{\Lambda} - \Lambda_*|^2 \left| \frac{1}{n} \sum_{i=1}^{n} \phi_*^{(3)}(\tilde{\Lambda}' g_{\mathbf{s}i}) h_i g_{\mathbf{s}i} g_{\mathbf{s}i}' \right|^2 = O_p(\sqrt{n}\zeta_{\mathbf{s}}^2 \gamma_n^2).$$

We now consider T_1. By expanding the first order condition of $\hat{\Lambda}$,

$$0 = \frac{1}{n} \sum_{i=1}^{n} \{\phi_*^{(1)}(\hat{\Lambda}' g_{\mathbf{s}i}) g_{\mathbf{s}i} - r_{\mathbf{s}i}\} = \frac{1}{n} \sum_{i=1}^{n} (\omega_{*i} g_{\mathbf{s}i} - r_{\mathbf{s}i}) + \frac{1}{n} \sum_{i=1}^{n} \phi_*^{(2)}(\bar{\Lambda}' g_{\mathbf{s}i}) g_{\mathbf{s}i} g_{\mathbf{s}i}' (\hat{\Lambda} - \Lambda_*),$$

where $\bar{\Lambda}$ lies on the line joining $\hat{\Lambda}$ and Λ_*. Similarly denote $\Sigma_{\mathbf{s}} = \mathbb{E}[\phi_*^{(2)}(\Lambda_*' g_{\mathbf{s}i}) g_{\mathbf{s}i} g_{\mathbf{s}i}']$ and $\bar{\Sigma}_{\mathbf{s}} = \frac{1}{n} \sum_{i=1}^{n} \phi_*^{(2)}(\bar{\Lambda}' g_{\mathbf{s}i}) g_{\mathbf{s}i} g_{\mathbf{s}i}'$. By solving this for $\hat{\Lambda} - \Lambda_*$ and inserting to T_1, we have

$$T_1 = -\mathbb{E}[\phi_*^{(2)}(\Lambda_*' g_{\mathbf{s}i}) h_i g_{\mathbf{s}i}]' \bar{\Sigma}_{\mathbf{s}}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\omega_{*i} g_{\mathbf{s}i} - r_{\mathbf{s}i}) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\omega_{*i} \tilde{h}_i^X - \tilde{r}_i^h) = T_{11} + T_{12} + T_{13},$$

where

$$\begin{aligned}
T_{11} &= -\mathbb{E}[\phi_*^{(2)}(\Lambda_*' g_{\mathbf{s}i}) h_i g_{\mathbf{s}i}]' (\bar{\Sigma}_{\mathbf{s}}^{-1} - \Sigma_{\mathbf{s}}^{-1}) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\omega_{*i} g_{\mathbf{s}i} - r_{\mathbf{s}i}), \\
T_{12} &= -\mathbb{E}[\phi_*^{(2)}(\Lambda_*' g_{\mathbf{s}i}) h_i g_{\mathbf{s}i}]' \Sigma_{\mathbf{s}}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\omega_{0i} g_{\mathbf{s}i} - r_{\mathbf{s}i}) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\omega_{0i} \tilde{h}_i^X - \tilde{r}_i^h), \\
T_{13} &= -\mathbb{E}[\phi_*^{(2)}(\Lambda_*' g_{\mathbf{s}i}) h_i g_{\mathbf{s}i}]' \Sigma_{\mathbf{s}}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\omega_{*i} - \omega_{0i}) g_{\mathbf{s}i} + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\omega_{*i} - \omega_{0i}) \tilde{h}_i^X.
\end{aligned}$$

For T_{11}, we apply a similar argument used to bound T_{11} in Theorem 1.2 but for iid data, which yields $|T_{11}| = O_p(\sqrt{n}\zeta_{\mathbf{s}}^4 \gamma_n^2)$. Note that $\mathbb{E}[T_{12}] = 0$. By Condition N'(2) and Chebyshev's inequality, $T_{12} = o_p(1)$. Also, the definition of $\tilde{h}_i^X$ implies $T_{13} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\omega_{*i} - \omega_{0i})(\tilde{h}_i^X - \beta_{\mathbf{s}}' g_{\mathbf{s}i}) = 0$. Combining these results, we have

$$\sqrt{n}(\tilde{\theta} - \theta_0 + b) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\Phi_i + v_{1i} + v_{2i} + v_{3i}) + r_n,$$

where $r_n = O_p(\zeta_{\mathbf{s}}^6 / \sqrt{n}) = o_p(1)$ under the assumptions in this theorem. The conclusion follows by applying a central limit theorem for i.i.d data.

A.3.4 Proof of Theorem 1.6

Recall $\omega_{\mathbf{s}}(x) = \phi_*^{(1)}(\lambda'_{\mathbf{os}} g_{\mathbf{os}}(x))$. By an expansion of the debiased estimator

$$\hat{\theta}_{TD} = \frac{1}{n}\sum_{i=1}^{n} \phi_*^{(1)}(\hat{\lambda}'_{TD} g_i) h_i = \frac{1}{n}\sum_{i=1}^{n} \phi_*^{(1)}(\hat{\Lambda}'_{\mathbf{s}} g_{\mathbf{s}i}) h_i$$

around $\hat{\Lambda}_{\mathbf{s}} = \lambda_{\mathbf{os}}$, we obtain

$$\sqrt{n}(\hat{\theta}_{TD} - \theta_0 + \tilde{b}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\Phi_i + \tilde{v}_{1i} + \tilde{v}_{2i} + \tilde{v}_{3i}) + T_1 + T_2 + T_3,$$

where

$$\begin{aligned}
T_1 &= \sqrt{n}\mathbb{E}[\phi_*^{(2)}(\lambda'_{\mathbf{os}} g_{\mathbf{s}i}) h_i g_{\mathbf{s}i}]'(\hat{\Lambda}_{\mathbf{s}} - \lambda_{\mathbf{os}}) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\omega_{\mathbf{s}i}\tilde{h}^X_{TDi} - \tilde{r}^h_{TDi}), \\
T_2 &= \left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\phi_*^{(2)}(\lambda'_{\mathbf{os}} g_{\mathbf{s}i}) h_i g_{\mathbf{s}i} - \mathbb{E}[\phi_*^{(2)}(\lambda'_{\mathbf{os}} g_{\mathbf{s}i}) h_i g_{\mathbf{s}i}]\}\right]'(\hat{\Lambda}_{\mathbf{s}} - \lambda_{\mathbf{os}}), \\
T_3 &= \frac{1}{2}(\hat{\Lambda}_{\mathbf{s}} - \lambda_{\mathbf{os}})'\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_*^{(3)}(\tilde{\Lambda}'_{\mathbf{s}} g_{\mathbf{s}i}) h_i g_{\mathbf{s}i} g'_{\mathbf{s}i}\right)(\hat{\Lambda}_{\mathbf{s}} - \lambda_{\mathbf{os}}),
\end{aligned}$$

and $\tilde{\Lambda}_{\mathbf{s}}$ is on the line joining $\hat{\Lambda}_{\mathbf{s}}$ and $\lambda_{\mathbf{os}}$. Since Condition TD(3) implies $\mathbb{E}[\phi_*^{(2)}(\lambda'_{\mathbf{os}} g_{\mathbf{s}})h]^2 = O(1)$, Chebyshev's inequality yields

$$\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\phi_*^{(2)}(\lambda'_{\mathbf{os}} g_{\mathbf{s}i}) h_i g_{\mathbf{s}i} - \mathbb{E}[\phi_*^{(2)}(\lambda'_{\mathbf{os}} g_{\mathbf{s}i}) h_i g_{\mathbf{s}i}]\}\right| = O_p(\sqrt{\zeta^2_{\mathbf{s}}/n}).$$

Thus, by Cauchy-Schwarz inequality and Lemma A.5(ii), it follows

$$|T_2| \leq \sqrt{n}\left|\frac{1}{n}\sum_{i=1}^{n}\{\phi_*^{(2)}(\lambda'_{\mathbf{os}} g_{\mathbf{s}i}) h_i g_{\mathbf{s}i} - \mathbb{E}[\phi_*^{(2)}(\lambda'_{\mathbf{os}} g_{\mathbf{s}i}) h_i g_{\mathbf{s}i}]\}\right| |\hat{\Lambda}_{\mathbf{s}} - \lambda_{\mathbf{os}}| = O_p(\zeta_{\mathbf{s}}\tilde{\gamma}_n).$$

For T_3, note that

$$|T_3| \leq \sqrt{n}|g_{\mathbf{s}}|^2 |\hat{\Lambda}_{\mathbf{s}} - \lambda_{\mathbf{os}}|^2 \sqrt{\frac{1}{n}\sum_{i=1}^{n}\phi_*^{(3)}(\tilde{\Lambda}'_{\mathbf{s}} g_{\mathbf{s}i})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n} h_i^2} = O_p(\sqrt{n}\zeta^2_{\mathbf{s}}\tilde{\gamma}^2_n),$$

where the first inequality follows from Cauchy-Schwarz inequality, and the equality follows from the law of large numbers, Condition TD(3), and Lemma A.5 (ii).

Now we consider T_1. By Lemma A.5 (i), we have

$$\hat{\Lambda}_{\mathbf{s}} - \lambda_{\mathbf{os}} = -\hat{\Theta}_{\mathbf{s}} \frac{1}{n} \sum_{i=1}^{n} (\omega_{\mathbf{s}i} g_{\mathbf{s}i} - r_{\mathbf{s}i}) + \tilde{\triangle},$$

where $\tilde{\triangle} = (I_{\mathbf{s}} - \hat{\Theta}_{\mathbf{s}} Q_n^{(2)}(\bar{\lambda}_{\mathbf{s}}))(\hat{\lambda}_{\mathbf{s}} - \lambda_{\mathbf{os}})$ and $Q_n^{(2)}(\bar{\lambda}_{\mathbf{s}}) = \mathbb{E}_n[\phi_*^{(2)}(\bar{\lambda}_{\mathbf{s}}' g_{\mathbf{s}}) g_{\mathbf{s}} g_{\mathbf{s}}']$. Also let $Q^{(2)}(\lambda_{\mathbf{os}}) = \mathbb{E}[\phi_*^{(2)}(\lambda_{\mathbf{os}}' g_{\mathbf{s}}) g_{\mathbf{s}} g_{\mathbf{s}}']$. Note that T_1 is decomposed as $T_1 = T_{11} + \cdots + T_{14}$, where

$$\begin{aligned}
T_{11} &= -\sqrt{n}\mathbb{E}[\phi_*^{(2)}(\lambda_{\mathbf{os}}' g_{\mathbf{s}i}) h_i g_{\mathbf{s}i}]' Q^{(2)}(\lambda_{\mathbf{os}})^{-1} \frac{1}{n} \sum_{i=1}^{n} (\omega_{0i} g_{\mathbf{s}i} - r_{\mathbf{s}i}) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\omega_{0i} \tilde{h}_{TDi}^X - \tilde{r}_{TDi}^h), \\
T_{12} &= -\sqrt{n}\mathbb{E}[\phi_*^{(2)}(\lambda_{\mathbf{os}}' g_{\mathbf{s}i}) h_i g_{\mathbf{s}i}]' Q^{(2)}(\lambda_{\mathbf{os}})^{-1} \frac{1}{n} \sum_{i=1}^{n} (\omega_{\mathbf{s}i} - \omega_{0i}) g_{\mathbf{s}i} + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\omega_{\mathbf{s}i} - \omega_{0i}) \tilde{h}_i^X, \\
T_{13} &= -\sqrt{n}\mathbb{E}[\phi_*^{(2)}(\lambda_{\mathbf{os}}' g_{\mathbf{s}i}) h_i g_{\mathbf{s}i}]' (\hat{\Theta} - Q^{(2)}(\lambda_{\mathbf{os}})^{-1}) \frac{1}{n} \sum_{i=1}^{n} (\omega_{\mathbf{s}i} g_{\mathbf{s}i} - r_{\mathbf{s}i}), \\
T_{14} &= \sqrt{n}\mathbb{E}[\phi_*^{(2)}(\lambda_{\mathbf{os}}' g_{\mathbf{s}i}) h_i g_{\mathbf{s}i}] \tilde{\triangle}.
\end{aligned}$$

For T_{11}, Condition TD and Chebychev's inequality imply

$$T_{11} = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{\tilde{\beta}_{\mathbf{s}}' (\omega_{0i} g_{\mathbf{s}i} - r_{\mathbf{s}i}) - (\omega_{0i} \tilde{h}_i^X - \tilde{r}_i^h) \xrightarrow{p} 0.$$

By the definition, $T_{12} = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\omega_{\mathbf{s}i} - \omega_{0i})(\tilde{\beta}_{\mathbf{s}}' g_{\mathbf{s}i} - \tilde{h}_i^X) = 0$. To bound T_{13}, note $\mathbb{E}[\phi_*^{(2)}(\lambda_{\mathbf{os}}' g_{\mathbf{s}i}) h_i g_{\mathbf{s}i}] = O_p(\zeta_{\mathbf{s}})$. By Cauchy-Schwarz inequality, Lemma A.5 (iv), and Condition TD(2), we have

$$|T_{13}| = \left| \sqrt{n}\mathbb{E}[\phi_*^{(2)}(\lambda_{\mathbf{os}}' g_{\mathbf{s}i}) h_i^X g_{\mathbf{s}i}]' (\hat{\Theta} - Q^{(2)}(\lambda_{\mathbf{os}})^{-1}) \frac{1}{n} \sum_{i=1}^{n} (\omega_{\mathbf{s}i} g_{\mathbf{s}i} - r_{\mathbf{s}i}) \right| = O_p(\sqrt{n} \zeta_{\mathbf{s}} \varrho_n \tilde{\gamma}_n).$$

Similarly, by Cauchy-Schwarz inequality, Lemma A.5 (ii) and (v), and the relation between ℓ_1- and ℓ_2-norms, it holds

$$\begin{aligned}
|T_{14}| &= \left| \sqrt{n}\mathbb{E}[\phi_*^{(2)}(\lambda_{\mathbf{os}}' g_{\mathbf{s}i}) h_i^X g_{\mathbf{s}i}]' (I_{\mathbf{s}} - \hat{\Theta} Q_n^{(2)}(\bar{\lambda}_{\mathbf{s}}))(\hat{\lambda}_{\mathbf{s}} - \lambda_{\mathbf{os}}) \right| \\
&\leq \sqrt{n}\mathbb{E}[\phi_*^{(2)}(\lambda_{\mathbf{os}}' g_{\mathbf{s}i}) h_i^X g_{\mathbf{s}i}]' |I_{\mathbf{s}} - \hat{\Theta} Q_n^{(2)}(\bar{\lambda}_{\mathbf{s}})| \left\| \hat{\lambda}_{\mathbf{s}} - \lambda_{\mathbf{os}} \right\|_1 \\
&= O_p(\sqrt{n} \kappa_{\mathbf{o},n}^2 \zeta_{\mathbf{s}}^4 + \sqrt{n} \zeta_{\mathbf{s}} \kappa_{\mathbf{o},n} \varrho_n).
\end{aligned}$$

Combining these results, we obtain

$$\sqrt{n}(\hat{\theta}_{TD} - \theta_0 + \tilde{b}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\Phi_i + \tilde{v}_{1i} + \tilde{v}_{2i} + \tilde{v}_{3i}) + r_n,$$

where $r_n = O_p(\sqrt{n}\kappa_{\mathbf{o},n}^2\zeta_{\mathbf{s}}^4 + \sqrt{n}\tilde{\gamma}_n\zeta_{\mathbf{s}}\varrho_n + \sqrt{n}\zeta_{\mathbf{s}}^2\tilde{\gamma}_n^2) = o_p(1)$ under the assumptions of this theorem. The conclusion follows by applying a central limit theorem.

A.3.5 Lemmas

Lemma A.4. *Under the conditions of Theorem 1.3, it holds*

(i) $\Pr\left\{\frac{1}{2}\mathscr{E}(\hat{\lambda}) + \alpha_n\left\|\hat{\lambda} - \lambda_{\mathbf{o}}\right\|_1 \leq 4\mathscr{E}(\lambda_{\mathbf{o}}) + \frac{16\alpha_n^2 s}{\phi^2_{S_{\lambda_{\mathbf{o}}}}\varrho}\right\} \geq 1 - \varepsilon,$

(ii) $\mathscr{E}(\hat{\lambda}) = O_p(\kappa_{\mathbf{o}n}\sqrt{\log K/n})$ *and* $\left\|\hat{\lambda} - \lambda_{\mathbf{o}}\right\|_1 = O_p(\kappa_{\mathbf{o}n})$.

Lemma A.5. *Let* $Q(\lambda_{\mathbf{s}}) = \mathbb{E}[\phi_*(\lambda_{\mathbf{s}}'g_{\mathbf{s}}) - \lambda_{\mathbf{s}}'r_{\mathbf{s}}]$ *and* $Q_n(\lambda_{\mathbf{s}}) = \mathbb{E}_n[\phi_*(\lambda_{\mathbf{s}}'g_{\mathbf{s}}) - \lambda_{\mathbf{s}}'r_{\mathbf{s}}]$. *Under the conditions of Theorem 1.6, it holds*

(i) $\hat{\Lambda}_{\mathbf{s}} - \lambda_{\mathbf{os}} = -\hat{\Theta}\frac{1}{n}\sum_{i=1}^n(\omega_{\mathbf{s}i}g_{\mathbf{s}i} - r_{\mathbf{s}i}) + \tilde{\triangle}$, *where* $\tilde{\triangle} = (I_{\mathbf{s}} - \hat{\Theta}_{\mathbf{s}}Q_n^{(2)}(\bar{\lambda}_{\mathbf{s}}))(\hat{\lambda}_{\mathbf{s}} - \lambda_{\mathbf{os}})$, *and* $\bar{\lambda}_{\mathbf{s}}$ *is on the line between* $\hat{\lambda}_{\mathbf{s}}$ *and* $\lambda_{\mathbf{os}}$,

(ii) $|\hat{\Lambda}_{\mathbf{s}} - \lambda_{\mathbf{os}}| = O_p(\tilde{\gamma}_n)$, *where* $\tilde{\gamma}_n = \kappa_{\mathbf{o},n} \vee \sqrt{\mathbf{s}\log K/n}$,

(iii) $|Q_n^{(2)}(\bar{\lambda}_{\mathbf{s}}) - Q^{(2)}(\lambda_{\mathbf{os}})| = O_p(\kappa_{\mathbf{o},n}\zeta_{\mathbf{s}}^3)$,

(iv) $|\frac{1}{n}\sum_{i=1}^n(\omega_{\mathbf{s}i}g_{\mathbf{s}i} - r_{\mathbf{s}i})| = O_p(\tilde{\gamma}_n)$,

(v) $|I_{\mathbf{s}} - \hat{\Theta}_{\mathbf{s}}Q_n^{(2)}(\bar{\lambda}_{\mathbf{s}})| = O_p(\kappa_{\mathbf{o},n}\zeta_{\mathbf{s}}^3 + \varrho_n)$.

Proof of Lemma A.4 (i)

Pick any $\varepsilon > 0$ small enough and $n \in \mathbb{N}$ large enough to satisfy Condition H. Then set $M = \frac{Q_{\mathbf{o}}}{2\sigma_{\varepsilon,n}}$ and take $\bar{\lambda} = t\hat{\lambda} + (1-t)\lambda_{\mathbf{o}}$ with $t = \frac{M}{M + \left\|\hat{\lambda} - \lambda_{\mathbf{o}}\right\|_1}$. Due to the definition of $\hat{\lambda}$ in (1.11) and convexity of its objective function, we have

$$\mathbb{E}_n[\phi_*(\bar{\lambda}'g(X)) - \bar{\lambda}'r(X)] + \alpha_n\left\|\bar{\lambda}\right\|_1 \leq \mathbb{E}_n[\phi_*(\lambda_{\mathbf{o}}'g(X)) - \lambda_{\mathbf{o}}'r(X)] + \alpha_n\left\|\lambda_{\mathbf{o}}\right\|_1,$$

and thus

$$\begin{aligned}\mathscr{E}(\bar{\lambda}) + \alpha_n\left\|\bar{\lambda}\right\|_1 &\leq -\{\nu_n(\bar{\lambda}) - \nu_n(\lambda_{\mathbf{o}})\} + \mathscr{E}(\lambda_{\mathbf{o}}) + \alpha_n\left\|\lambda_{\mathbf{o}}\right\|_1 \\ &\leq \mathscr{E}(\lambda_{\mathbf{o}}) + \alpha_n\left\|\lambda_{\mathbf{o}}\right\|_1 + \frac{Q_{\mathbf{o}}}{2}, \end{aligned} \tag{A.15}$$

with probability at least $1-\varepsilon$, where the second inequality follows from Condition H (i) combined with $\left\|\bar{\lambda} - \lambda_{\mathbf{o}}\right\|_1 = \frac{M\left\|\hat{\lambda} - \lambda_{\mathbf{o}}\right\|_1}{M + \left\|\hat{\lambda} - \lambda_{\mathbf{o}}\right\|_1} \leq M$. Hereafter all inequalities involving $\bar{\lambda}$ hold true with probability at least $1 - \varepsilon$.

Note that $\lambda = \lambda_{S_{\lambda_\mathbf{o}}} + \lambda_{S^c_{\lambda_\mathbf{o}}}$, and particularly $\lambda_{\mathbf{o},S_{\lambda_\mathbf{o}}} = \lambda_\mathbf{o}$ and $\lambda_{\mathbf{o},S^c_{\lambda_\mathbf{o}}} = 0$. Thus, (A.15) and the triangle inequality imply

$$\begin{aligned} \mathscr{E}(\bar{\lambda}) + \alpha_n \big\|\bar{\lambda}_{S^c_{\lambda_\mathbf{o}}}\big\|_1 &\leq \mathscr{E}(\lambda_\mathbf{o}) + \alpha_n \big\|\bar{\lambda}_{S_{\lambda_\mathbf{o}}} - \lambda_\mathbf{o}\big\|_1 + \frac{Q_\mathbf{o}}{2} \\ &\leq Q_\mathbf{o} + \alpha_n \big\|\bar{\lambda}_{S_{\lambda_\mathbf{o}}} - \lambda_\mathbf{o}\big\|_1, \end{aligned} \tag{A.16}$$

where the second inequality follows from $\mathscr{E}(\lambda_\mathbf{o}) \leq \frac{Q_\mathbf{o}}{2}$ (due to the definition of $Q_\mathbf{o}$). Thus, the triangle inequality yields

$$\mathscr{E}(\bar{\lambda}) + \alpha_n \big\|\bar{\lambda} - \lambda_\mathbf{o}\big\|_1 \leq Q_\mathbf{o} + 2\alpha_n \big\|\bar{\lambda}_{S_{\lambda_\mathbf{o}}} - \lambda_\mathbf{o}\big\|_1. \tag{A.17}$$

In order to bound the right hand side of (A.17), we consider two cases: (I) $2\alpha_n \big\|\bar{\lambda}_{S_{\lambda_\mathbf{o}}} - \lambda_\mathbf{o}\big\|_1 < Q_\mathbf{o}$, and (II) $2\alpha_n \big\|\bar{\lambda}_{S_{\lambda_\mathbf{o}}} - \lambda_\mathbf{o}\big\|_1 \geq Q_\mathbf{o}$.

Case (I) $2\alpha_n \big\|\bar{\lambda}_{S_{\lambda_\mathbf{o}}} - \lambda_\mathbf{o}\big\|_1 < Q_\mathbf{o}$**.**

In this case, (A.17) and Condition H (iii) imply

$$\mathscr{E}(\bar{\lambda}) + \alpha_n \big\|\bar{\lambda} - \lambda_\mathbf{o}\big\|_1 < 2Q_\mathbf{o} \leq \frac{\alpha_n M}{2}, \tag{A.18}$$

and thus $\big\|\bar{\lambda} - \lambda_\mathbf{o}\big\|_1 \leq \frac{M}{2}$.

Case (II) $2\alpha_n \big\|\bar{\lambda}_{S_{\lambda_\mathbf{o}}} - \lambda_\mathbf{o}\big\|_1 \geq Q_\mathbf{o}$**.**

In this case, (A.16) and $\lambda_{\mathbf{o},S^c_{\lambda_\mathbf{o}}} = 0$ guarantees

$$\big\|\bar{\lambda}_{S^c_{\lambda_\mathbf{o}}} - \lambda_{\mathbf{o},S^c_{\lambda_\mathbf{o}}}\big\|_1 = \big\|\bar{\lambda}_{S^c_{\lambda_\mathbf{o}}}\big\|_1 \leq 3\big\|\bar{\lambda}_{S_{\lambda_\mathbf{o}}} - \lambda_{\mathbf{o},S_{\lambda_\mathbf{o}}}\big\|_1 \leq \frac{3\sqrt{s}}{\phi_{S_{\lambda_\mathbf{o}}}} \sqrt{(\bar{\lambda} - \lambda_\mathbf{o})' \mathbb{E}[g(X)g(X)'](\bar{\lambda} - \lambda_\mathbf{o})}, \tag{A.19}$$

where the last inequality follows from Condition C. Observe that

$$\mathscr{E}(\bar{\lambda}) + \alpha_n \big\|\bar{\lambda} - \lambda_\mathbf{o}\big\|_1 \leq 4\alpha_n \big\|\bar{\lambda}_{S_{\lambda_\mathbf{o}}} - \lambda_\mathbf{o}\big\|_1 \leq \frac{4\alpha_n\sqrt{s}}{\phi_{S_{\lambda_\mathbf{o}}}} \sqrt{(\bar{\lambda} - \lambda_\mathbf{o})' \mathbb{E}[g(X)g(X)'](\bar{\lambda} - \lambda_\mathbf{o})},$$

where the first inequality follows from (A.17) and the condition of Case (II), the second inequality follows from (A.19) (note $\lambda_\mathbf{o} = \lambda_{\mathbf{o},S_{\lambda_\mathbf{o}}}$). Now by using $xy \leq x^2 + \frac{y^2}{4}$ for any $x, y \in \mathbb{R}$, we obtain

$$\begin{aligned} &\frac{4\alpha_n\sqrt{s}}{\phi_{S_{\lambda_\mathbf{o}}}} \sqrt{(\bar{\lambda} - \lambda_\mathbf{o})' \mathbb{E}[g(X)g(X)'](\bar{\lambda} - \lambda_\mathbf{o})} \\ \leq &\frac{1}{2}\left(\varrho (\bar{\lambda} - \lambda_\mathbf{o})' \mathbb{E}[g(X)g(X)'](\bar{\lambda} - \lambda_\mathbf{o}) + \frac{16\alpha_n s}{\phi^2_{S_{\lambda_\mathbf{o}}} \varrho} \right) \leq \frac{1}{2}\left(\mathscr{E}(\bar{\lambda}) + \frac{16\alpha_n s}{\phi^2_{S_{\lambda_\mathbf{o}}} \varrho} \right), \end{aligned}$$

where the second inequity follows from Condition H (ii). Combining these results

with the definition of $Q_{\mathbf{o}}$,

$$\mathscr{E}(\bar{\lambda}) + \alpha_n \left\| \bar{\lambda} - \lambda_{\mathbf{o}} \right\|_1 \leq \frac{1}{2}\mathscr{E}(\bar{\lambda}) + \frac{8\alpha_n^2 s}{\phi^2_{S_{\lambda_{\mathbf{o}}}} \varrho} \leq \frac{1}{2}\mathscr{E}(\bar{\lambda}) + Q_{\mathbf{o}}, \tag{A.20}$$

which implies (by Condition H (iii)) $\left\| \bar{\lambda} - \lambda_{\mathbf{o}} \right\|_1 \leq \frac{2\sigma_\varepsilon M}{\alpha_n} \leq \frac{M}{4}$.

Therefore, for both cases, it holds $\left\| \bar{\lambda} - \lambda_{\mathbf{o}} \right\|_1 \leq \frac{M}{2}$ and also $\left\| \hat{\lambda} - \lambda_{\mathbf{o}} \right\|_1 \leq M$, i.e., $\hat{\lambda}$ is close enough to $\lambda_{\mathbf{o}}$ to invoke Condition H (i).

Repeat the proof above by replacing $\bar{\lambda}$ with $\hat{\lambda}$. Then we obtain the counterparts of (A.18) and (A.20) with replacements of $\bar{\lambda}$ with $\hat{\lambda}$, i.e.,

$$\frac{1}{2}\mathscr{E}(\hat{\lambda}) + \alpha_n \left\| \hat{\lambda} - \lambda_{\mathbf{o}} \right\|_1 \leq 2Q_{\mathbf{o}},$$

with probability at least $1 - \varepsilon$. Therefore, the conclusion follows.

Proof of Lemma A.4 (ii)

By setting $\alpha_n \propto \sqrt{\frac{\log K}{n}}$, Part (i) of this lemma implies

$$\frac{1}{2}\mathscr{E}(\hat{\lambda}) + \sqrt{\frac{\log K}{n}} \left\| \hat{\lambda} - \lambda_{\mathbf{o}} \right\|_1 = O_p\left(\mathscr{E}(\lambda_{\mathbf{o}}) \vee \frac{s \log K}{n} \right),$$

and the conclusion follows.

Proof of Lemma A.5 (i)

By the KKT conditions for $\hat{\lambda}_{\mathbf{s}}$, an expansion around $\lambda_{\mathbf{os}}$ yields

$$0_{\mathbf{s}} = \frac{1}{n}\sum_{i=1}^{n}(\omega_{\mathbf{s}i} g_{\mathbf{s}i} - r_{\mathbf{s}i}) + \alpha_n \hat{\kappa}_{\mathbf{s}} = \frac{1}{n}\sum_{i=1}^{n}(\omega_{\mathbf{s}i} g_{\mathbf{s}i} - r_{\mathbf{s}i}) + Q_n^{(2)}(\bar{\lambda}_{\mathbf{s}})(\hat{\lambda}_{\mathbf{s}} - \lambda_{\mathbf{os}}) + \alpha_n \hat{\kappa}_{\mathbf{s}}, \tag{A.21}$$

where $\bar{\lambda}_{\mathbf{s}}$ is on the line between $\hat{\lambda}_{\mathbf{s}}$ and $\lambda_{\mathbf{os}}$. Thus, we have

$$\hat{\Lambda}_{\mathbf{s}} - \lambda_{\mathbf{os}} = \hat{\lambda}_{\mathbf{s}} - \lambda_{\mathbf{os}} + \hat{\Theta}_{\mathbf{s}} \alpha_n \hat{\kappa}_{\mathbf{s}} = \hat{\lambda}_{\mathbf{s}} - \lambda_{\mathbf{os}} - \hat{\Theta}_{\mathbf{s}} \left[\frac{1}{n}\sum_{i=1}^{n}(\omega_{\mathbf{s}i} g_{\mathbf{s}i} - r_{\mathbf{s}i}) + Q_n^{(2)}(\bar{\lambda}_{\mathbf{s}})(\hat{\lambda}_{\mathbf{s}} - \lambda_{\mathbf{os}}) \right],$$

where $I_{\mathbf{s}}$ is an $\mathbf{s} \times \mathbf{s}$ identity matrix, the first equality follows from the definition of $\hat{\Lambda}_{\mathbf{s}}$, and the second equality follows from (A.21). The conclusion follows by the definition of $\tilde{\triangle}$.

Proof of Lemma A.5 (ii)

By the definition of $\hat{\Lambda}_{\mathbf{s}}$,

$$\begin{aligned} |\hat{\Lambda}_{\mathbf{s}} - \lambda_{\mathbf{os}}| &\leq |\hat{\lambda}_{\mathbf{s}} - \lambda_{\mathbf{os}}| + |\hat{\Theta}_{\mathbf{s}}\alpha_n\hat{\kappa}_{\mathbf{s}}| \leq \left\|\hat{\lambda}_{\mathbf{s}} - \lambda_{\mathbf{os}}\right\|_1 + |\hat{\Theta}_{\mathbf{s}}\alpha_n\hat{\kappa}_{\mathbf{s}}| \\ &\lesssim \kappa_{\mathbf{o},n} + \sqrt{\frac{\mathbf{s}\log K}{n}} = O_p\left(\kappa_{\mathbf{o},n} \vee \sqrt{\frac{\mathbf{s}\log K}{n}}\right), \end{aligned}$$

where the first inequality follows from the triangle inequality, the second inequality follows from the relationship between the ℓ_1- and ℓ_2-norms, and the third inequality follows from Lemma A.4 (ii) and the assumption $|\hat{\Theta}_{\mathbf{s}}| = O_p(1)$.

Proof of Lemma A.5 (iii)

Note that

$$Q^{(2)}(\lambda_{\mathbf{os}}) = \mathbb{E}[\phi_*^{(2)}(\lambda_{\mathbf{os}}'g_{\mathbf{s}})g_{\mathbf{s}}g_{\mathbf{s}}'], \quad Q_n^{(2)}(\bar{\lambda}_{\mathbf{s}}) = \mathbb{E}_n[\phi_*^{(2)}(\bar{\lambda}_{\mathbf{s}}'g_{\mathbf{s}})g_{\mathbf{s}}g_{\mathbf{s}}'],$$

and further denote $Q_n^{(2)}(\lambda_{\mathbf{os}}) = \mathbb{E}_n[\phi_*^{(2)}(\lambda_{\mathbf{os}}'g_{\mathbf{s}})g_{\mathbf{s}}g_{\mathbf{s}}']$. By Lemma A.5 (ii) and Condition TD(3), we have

$$\begin{aligned} &|Q_n^{(2)}(\bar{\lambda}_{\mathbf{s}}) - Q_n^{(2)}(\lambda_{\mathbf{os}})| = |\mathbb{E}_n[\{\phi_*^{(2)}(\lambda_{\mathbf{os}}'g_{\mathbf{s}}) - \phi_*^{(2)}(\bar{\lambda}_{\mathbf{s}}'g_{\mathbf{s}})\}g_{\mathbf{s}}g_{\mathbf{s}}']| \\ \leq\ & \zeta_{\mathbf{s}}^2 \left\{\sup_{\Lambda:\|\Lambda-\lambda_{\mathbf{os}}\|_1\lesssim\tilde{\gamma}_n} \frac{1}{n}\sum_{i=1}^n \phi_*^{(3)}(\lambda_{\mathbf{s}}'g_{\mathbf{s}i})^2\right\}^{1/2} \left\{\frac{1}{n}\sum_{i=1}^n\{(\bar{\lambda}_{\mathbf{s}} - \lambda_{\mathbf{os}})'g_{\mathbf{s}}\}^2\right\}^{1/2} = O_p(\kappa_{\mathbf{o},n}\zeta_{\mathbf{s}}^3). \end{aligned}$$

Thus, the triangle inequality and Lemma A.3 (i) imply

$$\begin{aligned} |Q_n^{(2)}(\bar{\lambda}_{\mathbf{s}}) - Q^{(2)}(\lambda_{\mathbf{os}})| &\leq |Q_n^{(2)}(\bar{\lambda}_{\mathbf{s}}) - Q_n^{(2)}(\lambda_{\mathbf{os}})| + |Q_n^{(2)}(\lambda_{\mathbf{os}}) - Q^{(2)}(\lambda_{\mathbf{os}})| \\ &= O_p(\kappa_{\mathbf{o},n}\zeta_{\mathbf{s}}^3) + O_p\left(\sqrt{\frac{\zeta_{\mathbf{s}}^2\log\mathbf{s}}{n}}\right) = O_p(\kappa_{\mathbf{o},n}\zeta_{\mathbf{s}}^3). \end{aligned}$$

Proof of Lemma A.5 (iv)

By (A.21), we have

$$\begin{aligned} |\frac{1}{n}\sum_{i=1}^n(\omega_{\mathbf{s}i}g_{\mathbf{s}i} - r_{\mathbf{s}i})| &\leq |Q_n^{(2)}(\bar{\lambda}_{\mathbf{s}})(\hat{\lambda}_{\mathbf{s}} - \lambda_{\mathbf{os}})| + |\alpha_n\hat{\kappa}_{\mathbf{s}}| \\ &\leq \left|Q_n^{(2)}(\bar{\lambda}_{\mathbf{s}})\right| \left\|\hat{\lambda}_{\mathbf{s}} - \lambda_{\mathbf{os}}\right\|_1 + |\alpha_n\hat{\kappa}_{\mathbf{s}}| \\ &\lesssim \left\|\hat{\lambda}_{\mathbf{s}} - \lambda_{\mathbf{os}}\right\|_1 + |\alpha_n\hat{\kappa}_{\mathbf{s}}| = O_p\left(\kappa_{\mathbf{o},n} \vee \sqrt{\frac{\mathbf{s}\log K}{n}}\right), \end{aligned}$$

where the second inequality follows from the definition of the matrix norm $|\cdot|$ and the relationship between the ℓ_1- and ℓ_2-norms, and the third inequality uses Lemma A.4 (iii) and Condition TD.

Proof of Lemma A.5 (v)

By triangle inequality, we have

$$|I_{\mathbf{s}} - \hat{\Theta}_{\mathbf{s}} Q_n^{(2)}(\bar{\lambda}_{\mathbf{s}})| \leq |\{Q^{(2)}(\lambda_{\mathbf{os}})^{-1} - \hat{\Theta}_{\mathbf{s}}\} Q^{(2)}(\lambda_{\mathbf{os}})| + |\hat{\Theta}_{\mathbf{s}}\{Q^{(2)}(\lambda_{\mathbf{os}}) - Q_n^{(2)}(\bar{\lambda}_{\mathbf{s}})\}|.$$

Condition TD guarantees $Q^{(2)}(\lambda_{\mathbf{os}}) = O(1)$ and $\hat{\Theta}_{\mathbf{s}} = O_p(1)$. Thus, the conclusion follows by Lemma A.5 (iii).

Appendix B

Supplementary materials for Chapter 2

B.1 Derivations of RR in some examples

Example 2.2

First, show (2.8). By LIE and CIA:

$$\begin{aligned}
\mathbb{E}[Y^*|-b\leq R\leq b] &= \mathbb{E}[\mathbb{E}[Y^*|-b\leq R\leq b,X]|-b\leq R\leq b] \\
&= \mathbb{E}[\mathbb{E}[Y^*|0\leq R\leq b,X]|-b\leq R\leq b] \\
&= \mathbb{E}[\mathbb{E}[Y^*T|0\leq R\leq b,X]|-b\leq R\leq b] \\
&= \mathbb{E}[\mathbb{E}[Y|0\leq R\leq b,X]|-b\leq R\leq b] \\
&= \mathbb{E}[\gamma_0(X)|-b\leq R\leq b].
\end{aligned}$$

To see the continuous linear structure of the underlying functional, note θ_0 can be further rewritten as

$$\theta_0 = \mathbb{E}[\varsigma(R)\gamma_0(X)],$$

where $\varsigma(r) = \frac{\mathbf{1}\{-b\leq r\leq b\}}{\mathbb{E}[\mathbf{1}\{-b\leq R\leq b\}]}$. Indeed, by definition of conditional expectation

$$\begin{aligned}
\mathbb{E}[\gamma_0(X)|-b\leq R\leq b] =& \frac{\mathbb{E}[\mathbf{1}\{-b\leq R\leq b\}\gamma_0(X)]}{\mathbb{P}[-b\leq R\leq b]} = \frac{\mathbb{E}[\mathbf{1}\{-b\leq R\leq b\}\gamma_0(X)]}{\mathbb{E}[\mathbf{1}\{-b\leq R\leq b\}]} \\
=& \mathbb{E}\left[\frac{\mathbf{1}\{-b\leq R\leq b\}}{\mathbb{E}[\mathbf{1}\{-b\leq R\leq b\}]}\gamma_0(X)\right] = \mathbb{E}[\varsigma(R)\gamma_0(X)].
\end{aligned}$$

Finally derive RR in the form of (2.9). Assume

$$0 < \mathbb{E}\left[\mathbf{1}\{R\geq 0\}|X=x,-b\leq R\leq b\right] < 1 \quad \text{for all } x\in\mathcal{X} \text{ (overlap)}.$$

Then by LIE and CIA, for each $g \in L_{\mathbb{P},2}$

$$\begin{aligned}\mathbb{E}\left[\omega(X)\mathbf{1}\{R \geq 0\}g(X)| - b \leq R \leq b\right] &= \mathbb{E}\left[\frac{\mathbf{1}\{R \geq 0\}g(X)}{\mathbb{E}\left[\mathbf{1}\{R \geq 0\}|X, -b \leq R \leq b\right]}| - b \leq R \leq b\right] \\ &= \mathbb{E}\left[g(X)| - b \leq R \leq b\right].\end{aligned}$$

Example 2.3

Let $v(x) = w(x)f(x)$. Assume $w(x)$ has value 0 at boundaries and integrate by parts

$$\begin{aligned}\theta_0 &= \int w(x)f(x)\frac{\partial\gamma_0(x)}{\partial X_1}dx = \int v(x)\frac{\partial\gamma_0(x)}{\partial X_1}dx \\ &= -\int \gamma_0(x)\frac{\partial v(x)}{\partial X_1}dx = -\mathbb{E}\left[\gamma_0(X)\frac{\partial v(X)/\partial X_1}{f(X)}\right].\end{aligned}$$

Example 2.4

Let $F_\pi(x)$ and $f_\pi(x)$ be the cumulative distribution and probability density functions of X_π, respectively. Similarly, let the cumulative distribution and probability density functions of X be $F(x)$ and $f(x)$, respectively. Apply change of measure

$$\begin{aligned}\mathbb{E}[\gamma_0(\pi(X))] &= \int \gamma_0(x)dF_\pi(x) = \int \gamma_0(x)f_\pi(x)dx \\ &= \int \gamma_0(x)\frac{f_\pi(x)}{f(x)}f(x)dx = \int \gamma_0(x)\frac{f_\pi(x)}{f(x)}dF(x).\end{aligned}$$

Example 2.5

Let $f(z), f_{P,Z}(p,z), f_{P|Z}(p|z)$ be the marginal, joint and conditional densities, respectively. Then by definition of marginal density and apply double integration

$$\begin{aligned}\theta_0 &= \mathbb{E}\left[\omega(Z)\int_{p_0}^{p_1}\gamma_0(p,Z)dp\right] = \mathbb{E}\left[\omega(Z)\int \mathbf{1}\{p_0 \leq p \leq p_1\}\gamma_0(p,Z)dp\right] \\ &= \int\int \omega(z)\mathbf{1}\{p_0 \leq p \leq p_1\}\gamma_0(p,z)f(z)dpdz \\ &= \int\int \omega(z)\mathbf{1}\{p_0 \leq p \leq p_1\}\gamma_0(p,z)\frac{f_{P,Z}(p,z)}{f_{P|Z}(p|z)}dpdz \\ &= \mathbb{E}\left[\frac{\omega(Z)\mathbf{1}\{p_0 \leq P \leq p_1\}}{f_{P|Z}(P|Z)}\gamma_0(P,Z)\right].\end{aligned}$$

Example 2.6

Apply change of measure

$$\theta_0 = \int \gamma_0^A(x) f_X(x) dx = \int \gamma_0^A(x) \frac{f_X(x)}{f_{X_A}(x)} f_{X_A}(x) dx = \mathbb{E}_A \left[\gamma_0^A(X) \frac{f_X(X)}{f_{X_A}(X)} \right].$$

B.2 Proof of Proposition 2.1

For each $\alpha \in \Theta_n$, let

$$\begin{aligned} \text{(I)} = & \sup_{g \in \mathcal{H}_{W_n}} \{\mathbb{E}_n[\alpha(X)g(X) - m(Z, g(X))]\}^2, \\ \text{(II)} = & \quad \|W_n \mathbb{E}_n[e_\alpha(Z)]\|^2 . \end{aligned}$$

Step 1: show (I) $\leq$ (II).

For each $g \in \mathcal{H}_{W_n}$

$$\begin{aligned} \{\mathbb{E}_n[\alpha(X)g(X) - m(Z, g(X))]\}^2 &= \{\beta' \mathbb{E}_n[\alpha(X) W_n p(X) - m(Z, W_n p(X))]\}^2 \\ &\leq \|\beta\|^2 \|\mathbb{E}_n[\alpha(X) W_n p(X) - m(Z, W_n p(X))]\|^2 \\ &\leq \|W_n \mathbb{E}_n[e_\alpha(Z)]\|^2 , \end{aligned} \tag{B.1}$$

where the first equality is by linearity of $m(z, \cdot)$, the second line follows from Cauchy-Schwarz inequality, and the third relation uses $\|\beta\| \leq 1$ and linearity of $m(z, \cdot)$ again. Since (B.1) stands for each $g \in \mathcal{H}_{W_n}$, this direction is proved when sup is applied on both sides of display (B.1).

Step 2: show (II) $\leq$ (I).

Further denote $\mathcal{E}_{\alpha,W} = W_n \mathbb{E}_n[e_\alpha(Z)]$. Then

$$\text{(I)} = \sup_{\|\beta\| \leq 1} \beta' \mathcal{E}_{\alpha,W} \mathcal{E}'_{\alpha,W} \beta \geq \sup_{\|\beta\| = 1} \beta' \mathcal{E}_{\alpha,W} \mathcal{E}'_{\alpha,W} \beta = \lambda_{\max}(\mathcal{E}_{\alpha,W} \mathcal{E}'_{\alpha,W}) \geq \|\mathcal{E}_{\alpha,W}\|^2 ,$$

where the last relation follows since $\|\mathcal{E}_{\alpha,W}\|^2$ is one of the eigenvalues of $\mathcal{E}_{\alpha,W} \mathcal{E}'_{\alpha,W}$. To see this, write λ as one of the eigenvalues of $\mathcal{E}_{\alpha,W} \mathcal{E}'_{\alpha,W}$, with $\mathbf{v}$ denoted as its corresponding eigenvector. By definition, $\mathcal{E}_{\alpha,W} \mathcal{E}'_{\alpha,W} \mathbf{v} = \lambda \mathbf{v}$. Premultiplying both sides by $\mathcal{E}_{\alpha,W}$ yields

$$\mathcal{E}'_{\alpha,W} \mathcal{E}_{\alpha,W} \mathcal{E}'_{\alpha,W} \mathbf{v} = \mathcal{E}'_{\alpha,W} \lambda \mathbf{v} = \lambda \mathcal{E}'_{\alpha,W} \mathbf{v},$$

or $\left(\|\mathcal{E}_{\alpha,W}\|^2 - \lambda\right)\mathcal{E}'_{\alpha,W}\mathbf{v} = 0$. Therefore $\|\mathcal{E}_{\alpha,W}\|^2$ must be one of its eigenvalues. Proof of this direction is completed by recalling definition of $\mathcal{E}_{\alpha,W}$.

Step 3: conclusion follows by steps 1 and 2.

B.3 Measure of design uncertainty when $\frac{k}{n} \to c < 1$

Motivated by the fact that (2.28) and (2.29) can be calculated very differently, propose the following measure of design uncertainty. Let $\hat{\theta}_1 = \mathbb{E}_n[Yp(X)'](\hat{G}\hat{G})^-\hat{G}\hat{P}$, $\hat{\theta}_2 = \mathbb{E}_n[Yp(X)']\hat{G}^-\hat{P}$, define

$$\pi_\theta = \left|\frac{\hat{\theta}_1 - \hat{\theta}_2}{\min(\hat{\theta}_1, \hat{\theta}_2)}\right|. \tag{B.2}$$

Measure π_θ evaluates the percentage difference between computed $\hat{G}^-$ and $(\hat{G}\hat{G})^-\hat{G}$ in a given sample targeted at learning θ_0. Thus call π_θ Targeted Measure of Uncertainty (TMU) for design $\hat{G}$. Simulation shows that magnitude of π_θ is usually associated with dimension k and complexity of basis functions. Hence, a larger π_θ often signals a more uncertain specification due to more taxing computations. If this happens, it seems more appropriate to apply penalization.

Appendix C

Supplementary materials for Chapter 3

C.1 Basic lemmas

This section presents some useful lemmas to facilitate proofs of main theorems. Proofs for these lemmas can be found in Appendix C.5.

C.1.1 Useful maximal inequalities

The following maximal inequality is useful to control stochastic equicontinuity terms in low dimensional cases. Proved by Giné and Koltchinskii (2006), it has also been adapted by Belloni et al. (2015).

Lemma C.1. *[Theorem 3.1, Giné and Koltchinskii, 2006] Let $\varsigma_1 \cdots \varsigma_n$ be iid random variables taking values in a measurable space $(S, \mathscr{S})$ with common distribution $\mathbf{P}$ defined on the underlying $n-$fold product probability space. Let $\mathcal{F}$ be a suitable measurable class of functions mapping S to $\mathbb{R}$ with a measurable envelope F. Let σ^2 be a constant such that $\sup_{f\in\mathcal{F}} var(f) \leq \sigma^2 \leq \|F\|^2_{\mathbf{P},2}$. Suppose that there exist constants $A > \mathrm{e}^2$ and $V \geq 2$ such that*

$$\sup_Q N(\mathcal{F}, L^2(Q), \delta \|F\|_{Q,2}) \leq \left(\frac{A}{\delta}\right)^V$$

for all $0 < \delta \leq 1$. Then

$$\mathbb{E}\left[\left\|\sum_{i=1}^n \{f(\varsigma_i) - \mathbb{E}[f(\varsigma_i)]\}\right\|_{\mathcal{F}}\right] \leq \mathbf{C}\left[\sqrt{n\sigma^2 V \log\left(\frac{A\|F\|_{\mathbf{P},2}}{\sigma}\right)} + V\|F\|_{\mathbf{P},\infty}\log\left(\frac{A\|F\|_{\mathbf{P},2}}{\sigma}\right)\right],$$

where $\mathbf{C}$ is a universal constant.

The next lemma is particularly helpful in high dimensional asymptotics and has been proved in Dümbgen et al. (2010) and Lemma 14.24 in Bühlmann and Van De Geer (2011). In this thesis it is mainly used to bound term $\|\mathbb{E}_n[p(X)u_*] - \mathbb{E}[p(X)u_*]\|_\infty$ without assuming subgaussianity of u_*. The key idea is to use a symmetrization argument. Similar techniques have been explored in, for example, Lemma S4 of Belloni et al. (2012).

Lemma C.2. *[Nemirovski moment inequality] Let $\{X_1 \dots X_n\}$ be iid random variables, and let $\{f_1(x) \dots f_k(x)\}$ be a class of k functions. Then*

$$\mathbb{E}\max_{1\le j\le k}\left|\sum_{i=1}^{n}[f_j(X_i) - \mathbb{E}f_j(X_i)]\right| = O\left(\sqrt{\log k}\mathbb{E}\left[\max_{1\le j\le k}\sum_{i=1}^{n} f_j(X_i)^2\right]^{1/2}\right).$$

C.1.2 More results on least square projection

Suppose L1 holds. Write $\mathcal{L}_n\alpha_0 = a_l'p$ as the least square projection of α_0 onto Θ_n, where a_l is the projection coefficient. Denote $u_{\alpha_0} = \alpha_0 - \mathcal{L}_n\alpha_0$ as the corresponding projection error. Similarly, let $\mathcal{L}_n\gamma_0 = \beta_l'p$ be the least square projection of γ_0 onto Θ_n, with β_l as the coefficient and $u_{\gamma_0} = \gamma_0 - \mathcal{L}_n\gamma_0$ as the error. On the other hand, let $\alpha_b = a_b'p$ be the infeasible best approximation of α_0 in Θ_n, where a_b is the approximation coefficient and $r_{\alpha_0} = \alpha_0 - \alpha_b$ is the corresponding approximation error. Also define $\gamma_b = \beta_b'p$ as the infeasible best approximation for γ_0 in Θ_n, with β_b as the coefficient and $r_{\gamma_0} = \gamma_0 - \gamma_b$ as the error. By L1, $\|r_{\alpha_0}\|_{\mathbb{P},\infty} = \mathbf{r}_{\alpha_0}, \|r_{\gamma_0}\|_{\mathbb{P},\infty} = \mathbf{r}_{\gamma_0}$.

Lemma C.3. *If L1 holds, then:*

1. $\mathbb{E}[u_{\alpha_0}p(X)] = \mathbf{0}$, $\mathbb{E}[u_{\gamma_0}p(X)] = \mathbf{0}$;
2. $\|u_{\alpha_0}\|_{\mathbb{P},2} \le \mathbf{r}_{\alpha_0}$; $\|u_{\gamma_0}\|_{\mathbb{P},2} \le \mathbf{r}_{\gamma_0}$;
3. $\|u_{\alpha_0}\|_{\mathbb{P},\infty} \le (\ell_k + 1)\mathbf{r}_{\alpha_0}$; $\|u_{\gamma_0}\|_{\mathbb{P},\infty} \le (\ell_k + 1)\mathbf{r}_{\gamma_0}$.

Lemma C.4. *Let O and L1 hold. If $\mathbf{r}_{\alpha_0} = O(1)$, then $a_l = O(1)$; if $\mathbf{r}_{\gamma_0} = O(1)$, then $\beta_l = O(1)$.*

C.1.3 Asymptotic linear forms

Lemma C.5. *The asymptotic linear form of $\hat{\theta}_{BP}$ defined in (2.24) admits*

$$\begin{aligned}&\sqrt{n}\mathbb{E}_n\left[\tilde{\alpha}(X)Y - \theta_0\right]\\ =&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[m(Z_i, \gamma_0(X_i)) + \alpha_0(X_i)(Y_i - \gamma_0(X_i)) - \theta_0] + R_{1BP} + R_2,\end{aligned}$$

where

$$R_{1BP} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\tilde{\alpha}(X_i)\gamma_0(X_i) - m(Z_i, \gamma_0(X_i)) \right],$$
$$R_2 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[(\tilde{\alpha}(X_i) - \alpha_0(X_i))\, e_i \right].$$

Lemma C.6. *The asymptotic linear form of $\hat{\theta}_{DR}$ defined in (2.30) admits*

$$\begin{aligned} &\sqrt{n}\mathbb{E}_n \left[m(Z, \hat{\gamma}(X)) + \tilde{\alpha}(X)(Y - \hat{\gamma}(X)) - \theta_0 \right] \\ =&\frac{1}{\sqrt{n}} \sum_{i=1}^{n} [m(Z_i, \gamma_0(X_i)) + \alpha_0(X_i)(Y_i - \gamma_0(X_i)) - \theta_0] + R_{1DR} + R_2, \end{aligned}$$

where

$$R_{1DR} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[m(Z_i, \hat{\gamma}(X_i) - \gamma_0(X_i)) - \tilde{\alpha}(X_i)(\hat{\gamma}(X_i) - \gamma_0(X_i)) \right],$$
$$R_2 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[(\tilde{\alpha}(X_i) - \alpha_0(X_i))\, e_i \right].$$

C.1.4 Lemmas for term R_{1BP}

When $\frac{k}{n} \to c < 1$, R_{1BP} can usually be managed quite effectively. The main idea is to decompose γ_0 into a main part directly controlled by minimax learning, and a "residual part" shown to be sufficiently small. Linearity of $m(z, \cdot)$ then implies that magnitude of R_{1BP} can not be larger than the sum of the two. For some $f \in L_{\mathbb{P},2}$, consider the following decomposition

$$f = \pi_n f + r_n, \tag{C.1}$$

where $\pi_n f \in \Theta_n$ can be regarded as an "approximation mapping" to Θ_n and r_n is the corresponding "approximation error". Often it would be convenient to choose $\pi_n f$ as $\mathcal{L}_n f$, but this might not always be the case.

Lemma C.7. *If there exists a decomposition (C.1) for $f \in L_{\mathbb{P},2}$, then*

$$\left\{ \mathbb{E}_n \left[\tilde{\alpha}(X) f(X) - m(Z, f(X)) \right] \right\}^2 \lesssim T_1 + T_2,$$

where

$$T_1 = \left\{ \mathbb{E}_n [\tilde{\alpha}(X)\pi_n f(X) - m(Z, \pi_n f(X))] \right\}^2;$$
$$T_2 = \left\{ \mathbb{E}_n [\tilde{\alpha}(X) r_n - m(Z, r_n)] \right\}^2.$$

The next lemma gives the stochastic order of term T_2, the part not in the control of minimax learning.

Lemma C.8. *If there exists a decomposition (C.1) for $f \in L_{\mathbb{P},2}$, then*

$$\{\mathbb{E}_n[\tilde{\alpha}(X)r_n - m(Z, r_n)]\}^2 = O_p\left\{\left[\frac{1}{n}\sum_{i=1}^{n}\left(\tilde{\alpha}(X_i) - \alpha_0(X_i)\right)r_{ni}\right]^2\right\} + O_p\left\{\frac{\|r_n\|_{\mathbb{P},\infty}^2 \wedge \|\alpha_0\|_{\mathbb{P},\infty}^2 \|r_n\|_{\mathbb{P},2}^2}{n}\right\}.$$

Lemma C.8 bounds $\{\mathbb{E}_n[\tilde{\alpha}(X)r_n - m(Z, r_n)]\}^2$ by two terms. A simple but not necessarily the best way to control the first term, $\left[\frac{1}{n}\sum_{i=1}^{n}\left(\tilde{\alpha}(X_i) - \alpha_0(X_i)\right)r_{ni}\right]^2$, is by Cauchy-Schwarz inequality, which suffices if γ_0 is smooth enough. Otherwise, further decomposition would help. The order of the second term is pinned down by the magnitude of r_n (measured by its two different norms) as well as divergence of α_0 (measured by its sup norm).

The next lemma presents some inequalities useful for controlling term T_1 when $\frac{k}{n} \to c < 1$.

Lemma C.9. *Let $\tilde{\alpha}$ be calibrated according to display (2.17).*

1. Then, for every $f \in \mathcal{H}_{W_n}$ and $\alpha \in \Theta_n$

$$\begin{aligned}&\{\mathbb{E}_n[\tilde{\alpha}(X)f(X) - m(Z, f(X))]\}^2 + \mathcal{P}_n(\tilde{\alpha}(X))\\ &\leq \sup_{g\in\mathcal{H}_{W_n}}\{\mathbb{E}_n[\alpha(X)g(X) - m(Z, g(X))]\}^2 + \mathcal{P}_n(\alpha(X)),\end{aligned} \tag{C.2}$$

and

$$\mathcal{P}_n(\tilde{\alpha}(X)) \leq \sup_{g\in\mathcal{H}_{W_n}}\{\mathbb{E}_n[\alpha(X)g(X) - m(Z, g(X))]\}^2 + \mathcal{P}_n(\alpha(X)). \tag{C.3}$$

2. *If in addition $W_n'W_n - I$ is positive semidefinite*, then for every $f \in \Theta_n$ and $\alpha \in \Theta_n$

$$\begin{aligned}&\{\mathbb{E}_n\left[\tilde{\alpha}(X)f(X) - m(Z, f(X))\right]\}^2\\ &\lesssim \left\{\sup_{g\in\mathcal{H}_{W_n}}\{\mathbb{E}_n[\alpha(X)g(X) - m(Z, g(X))]\}^2 + \mathcal{P}_n(\alpha(X))\right\}\|f\|_{\mathbb{P},2}^2.\end{aligned} \tag{C.4}$$

C.1.5 Lemma for term R_{1DR}.

The next lemma shows a well trained estimator from a different random sample can simplify asymptotics considerably when $\frac{k}{n} \to \infty$. This mechanism avoids

studying stochastic equicontinuity terms on a case-by-case basis, and allows $\hat{\gamma}$ to be any machine learning estimator when constructing $\hat{\theta}_{DR}$.

Lemma C.10. *[Conditional convergence implies unconditional convergence] Let* $\{X_n\}, \{Y_n\}$ *be two sequences of random vectors, and let* $\{A_n\}$ *be a sequence of positive numbers.*

1. *If conditional on* $\{Y_n\}$, $\|X_n\| = o_p(A_n)$, *then* $\|X_n\| = o_p(A_n)$ *unconditionally;*

2. *If conditional on* $\{Y_n\}$, $\|X_n\| = O_p(A_n)$, *then* $\|X_n\| = O_p(A_n)$ *unconditionally.*

C.1.6 Lemmas for term R_2

The following lemmas are particular useful for studying remainder terms involving e. Write $\mathscr{Z}_n = (Z_1, \dots Z_n) \in \mathcal{Z}^n$, where $\mathcal{Z}^n = \prod_{i=1}^{n} \mathcal{Z}_i$ is a product support. Let $\mathcal{A}_i(\cdot) : \mathcal{Z}^n \to \mathbb{R}$ be a function of $\mathscr{Z}_n$ for each $i = 1 \dots n$. Note $\{\mathcal{A}_i(\cdot)\}_{i=1}^n$ are not necessarily iid distributed.

Lemma C.11. *[Property of e] If O holds, then*

$$\frac{1}{n}\sum_{i=1}^{n}\left[\mathcal{A}_i(\mathscr{Z}_n)e_i\right] = O_p\left(\sqrt{\frac{\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathcal{A}_i^2(\mathscr{Z}_n)\right]}{n}}\right).$$

The next lemma can be invoked to establish a central limit theorem even when $\tilde{\alpha}$ can not consistently estimate α_0.

Lemma C.12. *[Central limit theorem] Suppose O and the following conditions hold:*

1. $\min_i \{\mathbb{E}[e_i^2|Z_i]\}$ is bounded away from zero and $\mathbb{E}\left[|e_i|^3 |Z_i\right] < \infty$ almost surely;

2. $\frac{\max_i|\mathcal{A}_i(\mathscr{Z}_n)|}{\sqrt{n}} = o_p(1)$;

3. $\frac{1}{n}\sum_{i=1}^{n}\mathcal{A}_i^2(\mathscr{Z}_n) = O_p(1)$;

4. $\left[\frac{1}{n}\sum_{i=1}^{n}\mathcal{A}_i^2(\mathscr{Z}_n)\right]^{-1} = O_p(1)$.

Then

$$n^{-1/2}\mathcal{V}_n^{-1/2}\sum_{i=1}^{n}[\mathcal{A}_i(\mathscr{Z}_n)(Y_i-\gamma_0(X_i))]\xrightarrow{d}N(0,1),$$

where $\mathcal{V}_n=\frac{1}{n}\sum_{i=1}^{n}[\mathcal{A}_i^2(\mathscr{Z}_n)\mathbb{E}[e_i^2|Z_i]]$ *and* $N(0,1)$ *is a standard normal random variable.*

C.2 Proofs for main results when $\frac{k}{n}\to 0$

Notations. Write $G=\mathbb{E}[p(X)p(X)']$, $\hat{G}=\mathbb{E}_n[p(X)p(X)']$, $\hat{P}=\mathbb{E}_n[m(Z,p(X))]$, $\mathcal{W}_n=W_n'W_n$, $e^R=m(z,p(x))-\alpha_0(x)p(x)$. For this and next sections let $\tilde{a}=(\hat{G}\mathcal{W}_n\hat{G}+\lambda_1\hat{G})^{-}\hat{G}\mathcal{W}_n\hat{P}$, $\hat{a}=\hat{G}^{-}\hat{P}$. Hence minimax BP learner admits $\tilde{\alpha}(x)=p(x)'\tilde{a}$, while $\hat{\alpha}(x)=p(x)'\hat{a}$. Denote $\Psi_n=\left(\mathcal{W}_n\hat{G}+\lambda_1 I\right)^{-1}\mathcal{W}_n\hat{G}$.

C.2.1 Additional convergence results when $\frac{k}{n}\to 0$

The first lemma below is concerned with basic matrix law of large numbers when $\frac{k}{n}\to 0$.

Lemma C.13. *Let O and L1 hold and* $\frac{\xi_k^2\log k}{n}\to 0$. *Then:*

1. $\mathbb{E}\left[\left\|\hat{G}-G\right\|\right]\lesssim\frac{\xi_k^2\log k}{n}+\sqrt{\frac{\xi_k^2\log k}{n}}$;
2. $\left\|\hat{G}-G\right\|=O_p\left(\sqrt{\frac{\xi_k^2\log k}{n}}\right)$;
3. $\hat{G}$ has eigenvalues bounded away from zero and from above wpa1.

Proof. See Lemma 6.2 in Belloni et al. (2015) for statements (1) and (2). To see (3), note by L1-(1), there exists some $c>0$ such that all eigenvalues of G are larger than or equal to c. If wpa1 $\hat{G}$ has some eigenvalue smaller than $\frac{c}{2}$, then there exists some vector $x\in\mathbb{R}^k$, $\|x\|=1$ such that $x'\hat{G}x<\frac{c}{2}$. It follows

$$\left\|\hat{G}-G\right\|\geq\left|x'(\hat{G}-G)x\right|=\left|x'\hat{G}x-x'Gx\right|\geq\left|\frac{c}{2}-c\right|=\frac{c}{2},$$

contradicting statement (2). Next by statement (2), definition of G and triangle inequality

$$\lambda_{\max}\left(\hat{G}\right)=\left\|\hat{G}\right\|\leq\left(\left\|\hat{G}-G\right\|+\|G\|\right)=O_p(1).$$

Hence all eigenvalues of $\hat{G}$ are bounded from above wpa1. □

The next lemma presents several convergence results useful for showing consistency of $\tilde{\alpha}$.

Lemma C.14. *Let O and L1 hold and* $\frac{\xi_k^2 \log k}{n} \to 0$. *Then:*

1. $\hat{G}^{-1}\mathbb{E}_n e^R = O_p\left(\sqrt{\frac{\xi_k^2}{n}} \wedge \sqrt{\frac{\|\alpha_0\|_{\mathbb{P},\infty} k}{n}}\right)$.

2. $\hat{G}^{-1}\mathbb{E}_n[u_{\alpha_0} p(X)] = O_p(\mathbf{r}_{\alpha_0})$.

Proof. Statement (1): Note by definition of α_0, $\mathbb{E}e^R = \mathbf{0}$.

$$\begin{aligned}\mathbb{E}\left\|\mathbb{E}_n[e^R]\right\|^2 &= \mathbb{E}\left(\mathbb{E}_n[e^R]'\mathbb{E}_n[e^R]\right) = \frac{1}{n}\mathbb{E}[(e^R)'e^R] \\ &= \frac{1}{n}\mathbb{E}\left[\sum_{j=1}^k (e_j^R)^2\right] = \frac{1}{n}\sum_{j=1}^k \mathbb{E}\left[m(Z, p_j(X)) - \alpha_0(X)p_j(X)\right]^2 \\ &\lesssim \frac{1}{n}\sum_{j=1}^k \mathbb{E}m^2(Z, p_j(X)) + \frac{1}{n}\sum_{j=1}^k \mathbb{E}\alpha_0^2(X)p_j^2(X),\end{aligned} \tag{C.5}$$

where the first equality is by definition of vector norm $\|\cdot\|$, the second equality uses iid assumption and $\mathbb{E}e^R = \mathbf{0}$, the third and fourth equality are simply rewriting equations, and final relation follows from triangle inequality. Apply O and standard algebra

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^k \mathbb{E}m^2(Z, p_j(X)) &\leq \frac{C}{n}\sum_{j=1}^k \mathbb{E}p_j^2(X) = \frac{C}{n}tr(G) \leq \frac{Ck\lambda_{\max}(G)}{n} = O\left(\frac{k}{n}\right), \\ \frac{1}{n}\sum_{j=1}^k \mathbb{E}\alpha_0^2(X)p_j^2(X) &= \frac{1}{n}\mathbb{E}\left[\alpha_0^2(X)\sum_{j=1}^k p_j^2(X)\right] = \frac{1}{n}\mathbb{E}\left[\alpha_0^2(X)p(X)'p(X)\right],\end{aligned}$$

with either

$$\mathbb{E}\left[\alpha_0^2(X)p(X)'p(X)\right] \leq \xi_k^2 \mathbb{E}\alpha_0^2(X) \lesssim \xi_k^2,$$

or

$$\begin{aligned}\mathbb{E}\left[\alpha_0^2(X)p(X)'p(X)\right] &\leq \|\alpha_0\|_{\mathbb{P},\infty}\mathbb{E}\left[p(X)'p(X)\right] = \|\alpha_0\|_{\mathbb{P},\infty}\mathbb{E}\left[tr(p(X)'p(X))\right] \\ &= \|\alpha_0\|_{\mathbb{P},\infty} tr(G) = O\left(\|\alpha_0\|_{\mathbb{P},\infty} k\right).\end{aligned}$$

It follows by Markov inequality

$$\begin{aligned}\left\|\mathbb{E}_n[e^R]\right\| &= O_p\left[\sqrt{\frac{k}{n}} \vee \left(\sqrt{\frac{\xi_k^2}{n}} \wedge \sqrt{\frac{\|\alpha_0\|_{\mathbb{P},\infty} k}{n}}\right)\right] = O_p\left[\left(\sqrt{\frac{k}{n}} \vee \sqrt{\frac{\xi_k^2}{n}}\right) \wedge \sqrt{\frac{\|\alpha_0\|_{\mathbb{P},\infty} k}{n}}\right] \\ &= O_p\left(\sqrt{\frac{\xi_k^2}{n}} \wedge \sqrt{\frac{\|\alpha_0\|_{\mathbb{P},\infty} k}{n}}\right).\end{aligned}$$

Further by Lemma C.13, $\left\|\hat{G}^{-1}\right\| = O_p(1)$. Hence conclusion follows

$$\left\|\hat{G}^{-1}\mathbb{E}_n b_0\right\| \leq \left\|\hat{G}^{-1}\right\| \|\mathbb{E}_n b_0\| = O_p\left(\sqrt{\frac{\xi_k^2}{n}} \wedge \sqrt{\frac{\|\alpha_0\|_{\mathbb{P},\infty}\, k}{n}}\right).$$

Statement (2): Denote $\tilde{\mathcal{L}}_n u_{\alpha_0} = p'\hat{G}^{-1}\mathbb{E}_n[u_{\alpha_0} p(X)]$ as the empirical projection of u_{α_0} onto Θ_n. It follows $\left\|\hat{G}^{-1/2}\mathbb{E}_n[u_{\alpha_0} p(X)]\right\| = O_p(\mathbf{r}_{\alpha_0})$ since

$$\left\|\hat{G}^{-1/2}\mathbb{E}_n[u_{\alpha_0} p(X)]\right\|^2 = \mathbb{E}_n\left[(\tilde{\mathcal{L}}_n u_{\alpha_0})u_{\alpha_0}\right] = \mathbb{E}_n\left[(\tilde{\mathcal{L}}_n u_{\alpha_0})^2\right] \leq \mathbb{E}_n\left[u_{\alpha_0}^2\right] \xrightarrow{p} \mathbb{E}\left[u_{\alpha_0}^2\right] \leq \mathbf{r}_{\alpha_0}^2,$$

where the first equality is by rewriting, the second and third relations are by definition of empirical projection, the fourth relation follows from Khinchin law of large numbers, and final relation uses Lemma C.3-(2). Conclusion follows since

$$\left\|\hat{G}^{-1}\mathbb{E}_n[u_{\alpha_0} p(X)]\right\| \leq \left\|\hat{G}^{-1/2}\right\| \left\|\hat{G}^{-1/2}\mathbb{E}_n[u_{\alpha_0} p(X)]\right\| = O_p(\mathbf{r}_{\alpha_0}),$$

where $\left\|\hat{G}^{-1/2}\right\| = O_p(1)$ as well when $\frac{\xi_k^2 \log k}{n} \to 0$ by Lemma C.13. □

Lemma C.15. *Let O and L1 hold and* $\frac{\xi_k^2 \log k}{n} \to 0$. *Then:*

1. $\|\hat{a} - a_l\| = O_p\left[\left(\sqrt{\frac{\xi_k^2}{n}} \wedge \sqrt{\frac{\|\alpha_0(X)\|_{\mathbb{P},\infty} k}{n}}\right) + \mathbf{r}_{\alpha_0}\right]$;
2. $\|\hat{a}\| = O_p(1)$.

Proof. Recall $\alpha_0 = a_l' p + u_{\alpha_0}$. Then definition of $\hat{a}$ and a_l yields that wpa1

$$\begin{aligned}\hat{a} &= \hat{G}^{-}\mathbb{E}_n[m(Z, p(X)) - \alpha_0(X)p(X)] + \hat{G}^{-}\mathbb{E}_n[\alpha_0(X)p(X)] \\ &= \hat{G}^{-}\mathbb{E}_n e^R + \hat{G}^{-}\mathbb{E}_n[u_{\alpha_0} p(X)] + a_l.\end{aligned}$$

Note $\hat{G}^{-} = \hat{G}^{-1}$ wpa1 by Lemma C.13. Thus wpa1, $\hat{a} - a_l = \hat{G}^{-1}\mathbb{E}_n e^R + \hat{G}^{-1}\mathbb{E}_n[u_{\alpha_0} p(X)]$. Statement (1) follows by triangle inequality and Lemma C.14

$$\|\hat{a} - a_l\| \leq \left\|\hat{G}^{-1}\mathbb{E}_n e^R\right\| + \left\|\hat{G}^{-1}\mathbb{E}_n[u_{\alpha_0} p(X)]\right\| = O_p\left[\left(\sqrt{\frac{\xi_k^2}{n}} \wedge \sqrt{\frac{\|\alpha_0\|_{\mathbb{P},\infty}\, k}{n}}\right) + \mathbf{r}_{\alpha_0}\right].$$

Statement (2) follows from triangle inequality, Lemma C.4 and statement (1). □

Lemma C.16. *Let Lemma C.15 holds. If in addition,* $\mathcal{W}_n\hat{G}$ *is symmetric and has eigenvalues bounded away from zero wpa1, then* $\|\tilde{a} - a_l\| = o_p(1)$.

Proof. By Lemma C.13 and assumptions for this lemma, it is easy to see $\hat{G}\left(\mathcal{W}_n\hat{G}+\lambda_1 I\right)$ also has eigenvalues bounded away from zero wpa1. Hence, wpa1 $(\hat{G}\mathcal{W}_n\hat{G}+\lambda_1\hat{G})^- = (\hat{G}\mathcal{W}_n\hat{G}+\lambda_1\hat{G})^{-1}$. Recall $\Psi_n = \left(\mathcal{W}_n\hat{G}+\lambda_1 I\right)^{-1}\mathcal{W}_n\hat{G}$. Standard algebra yields that wpa1

$$\begin{aligned}\tilde{a} &= \left[\hat{G}\left(\mathcal{W}_n\hat{G}+\lambda_1 I\right)\right]^{-1}\hat{G}\mathcal{W}_n\hat{P} = \left(\mathcal{W}_n\hat{G}+\lambda_1 I\right)^{-1}\mathcal{W}_n\hat{P} \\ &= \left(\mathcal{W}_n\hat{G}+\lambda_1 I\right)^{-1}\mathcal{W}_n\hat{G}\hat{G}^{-1}\hat{P} = \Psi_n\hat{G}^{-1}\hat{P} = \Psi_n\hat{a}. \end{aligned} \tag{C.6}$$

Thus by triangle inequality

$$\|\tilde{a}-a_l\| \le \|\Psi_n\hat{a}-\hat{a}\| + \|\hat{a}-a_l\| \le \|\Psi_n - I\|\,\|\hat{a}\| + \|\hat{a}-a_l\|.$$

Since $\|\hat{a}-a_l\| = o_p(1)$ and $\|\hat{a}\| = O_p(1)$ by Lemma C.15, it suffices to show $\|\Psi_n - I\| = o_p(1)$.

By assumption, for some $c>0$, we can write $\left\{\mu_j^{\mathbf{W}}\right\}_{j=1}^k$, where wlog $\mu_1^{\mathbf{W}} \ge \mu_2^{\mathbf{W}} \ldots \ge \mu_k^{\mathbf{W}} > c$, are k real eigenvalues of $\mathcal{W}_n\hat{G}$. Since $\mathcal{W}_n\hat{G}$ is assumed symmetric, $\mathcal{W}_n\hat{G}$ is also orthogonally diagonalizable. That is, there exists some U such that $U'U = I$, and some diagonal matrix Λ with $\left\{\mu_j^{\mathbf{W}}\right\}_{j=1}^k$ on the diagonal such that $\mathcal{W}_n\hat{G} = U\Lambda U'$. Apply standard algebra

$$\begin{aligned}\mathcal{W}_n\hat{G}+\lambda_1 I &= U\Lambda U' + \lambda_1 UU' = U(\Lambda+\lambda_1 I)U', \\ \Psi_n &= \left(\mathcal{W}_n\hat{G}+\lambda_1 I\right)^{-1}\mathcal{W}_n\hat{G} = U(\Lambda+\lambda_1 I)^{-1}U'U\Lambda U' = U(\Lambda+\lambda_1 I)^{-1}\Lambda U', \\ \Psi_n - I &= U(\Lambda+\lambda_1 I)^{-1}\Lambda U' - UU' = U\left[(\Lambda+\lambda_1 I)^{-1}\Lambda - I\right]U'.\end{aligned}$$

$(\Lambda+\lambda_1 I)^{-1}\Lambda - I$ is apparently a diagonal matrix so that for each $j = 1\ldots k$

$$\lambda_j\left[(\Psi_n - I)'(\Psi_n - I)\right] = \left(\frac{\mu_j^{\mathbf{W}}}{\mu_j^{\mathbf{W}}+\lambda_1} - 1\right)^2 = \left(\frac{\lambda_1}{\mu_j^{\mathbf{W}}+\lambda_1}\right)^2.$$

It follows

$$\|\Psi_n - I\| = \left\{\lambda_{\max}\left[(\Psi_n - I)'(\Psi_n - I)\right]\right\}^{1/2} = \frac{\lambda_1}{\mu_k^{\mathbf{W}}+\lambda_1} = \frac{1}{1+\mu_k^{\mathbf{W}}/\lambda_1} \xrightarrow{p} 0,$$

since wpa1 $\frac{\mu_k^{\mathbf{W}}}{\lambda_1} \ge \frac{c}{\lambda_1} \to \infty$. This completes proof for Lemma C.16. □

C.2.2 Additional results for controlling stochastic equicontinuity terms

The following lemmas concern some stochastic equicontinuity terms in the remainder terms of Lemma C.5.

Lemma C.17. *If conditions for Theorem 3.1 hold,* $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[(\tilde{a}-a_l)'p(X_i)e_i] = o_p(1)$.

Proof. Let $\frac{\tilde{a}-a_l}{\|\tilde{a}-a_l\|} = \tilde{\mathbf{a}}$ so that $\|\tilde{\mathbf{a}}\| = 1$. Decompose

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[(\tilde{a}-a_l)'p(X_i)e_i] = \|\tilde{a}-a_l\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\tilde{\mathbf{a}}'p(X_i)e_i].$$

From Lemma C.16, $\|\tilde{a}-a_l\| = o_p(1)$, so it remains to show $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\tilde{\mathbf{a}}'p(X_i)e_i] = O_p(1)$. Note

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\tilde{\mathbf{a}}'p(X_i))^2\right] = \mathbb{E}\left[\tilde{\mathbf{a}}'\hat{G}\tilde{\mathbf{a}}\right] \le \mathbb{E}\left[\|\tilde{\mathbf{a}}\|^2\left\|\hat{G}\right\|\right] \le \mathbb{E}\left\|\hat{G}\right\| \lesssim 1, \tag{C.7}$$

where the last inequality is by Lemma C.13 and triangle inequality

$$\mathbb{E}\left\|\hat{G}\right\| \le \mathbb{E}\left\|\hat{G}-G\right\| + \mathbb{E}\|G\| \lesssim \frac{\xi_k^2\log k}{n} + \sqrt{\frac{\xi_k^2\log k}{n}} + \lambda_{\max}(G) \lesssim 1.$$

Since $\tilde{\mathbf{a}}$ is also a function of $Z_1\cdots Z_n$, Lemma C.11 can be invoked with $\mathcal{A}_i(\mathscr{Z}_n) = \tilde{\mathbf{a}}'p(X_i)$ so that

$$\frac{1}{n}\sum_{i=1}^{n}[\tilde{\mathbf{a}}'p(X_i)e_i] = O_p\left(\sqrt{\frac{\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\tilde{\mathbf{a}}'p(X_i))^2\right]}{n}}\right) = O_p\left(\sqrt{\frac{1}{n}}\right)$$

where the second relation follows from display (C.7). □

Lemma C.18. *If conditions for Theorem 3.1 hold*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\left(\tilde{\alpha}(X_i)-\mathcal{L}_n\alpha_0(X_i)\right)u_{\gamma_0 i}\right] = O_p\left[\|\tilde{a}-a_l\|(\ell_k+1)\mathbf{r}_{\gamma_0}\left(\sqrt{k\log\xi_k}+\frac{k\xi_k\log\xi_k}{\sqrt{n}}\right)\right].$$

Proof. By definition of $\mathcal{L}_n\alpha_0$, $\frac{1}{n}\sum_{i=1}^{n}\left[\left(\tilde{\alpha}(X_i)-\mathcal{L}_n\alpha_0(X_i)\right)u_{\gamma_0 i}\right] = \frac{1}{n}\sum_{i=1}^{n}[(\tilde{a}-a_l)'p(X_i)u_{\gamma_0 i}]$. Let $\frac{\tilde{a}-a_l}{\|\tilde{a}-a_l\|} = \tilde{\mathbf{a}}$ so that $\|\tilde{\mathbf{a}}\| = 1$. Consider function class $\mathcal{F} = \left\{\mathbf{a}'p(x)u_{\gamma_0} : \|\mathbf{a}\| = 1, \mathbf{a}\in\mathbb{R}^k\right\}$.

For any $f \in \mathcal{F}$

$$\mathbb{E}[f(X)] = \mathbb{E}\left[\mathbf{a}'p(X)u_{\gamma_0}(X)\right] = \mathbf{a}'\mathbb{E}[p(X)u_{\gamma_0}] = 0,$$

where the last equality follows from Lemma C.3-(1). Further, by Lemma C.3-(3)

$$var(f(X)) = \mathbb{E}[f(X)^2] = \mathbb{E}\left[\left(\mathbf{a}'p(X)\right)^2 u_{\gamma_0}^2\right] \leq \|u_{\gamma_0}\|_{\mathbb{P},\infty}^2 \mathbb{E}[\mathbf{a}'p(X)p(X)'\mathbf{a}] \lesssim (\ell_k+1)^2\mathbf{r}_{\gamma_0}^2,$$

and

$$|\mathbf{a}'p(x)u_{\gamma_0}| \leq \|\mathbf{a}\|\,\|p(x)\|\,|u_{\gamma_0}| \leq \xi_k \|u_{\gamma_0}\|_{\mathbb{P},\infty} \leq \xi_k(\ell_k+1)\mathbf{r}_{\gamma_0}.$$

Thus pick envelope function $\mathbf{F} = \xi_k(\ell_k+1)\mathbf{r}_{\gamma_0}$. For every pair of $\mathbf{a}_1, \mathbf{a}_2 \in \mathbb{R}^k$ with unit l_2 norm

$$\begin{aligned}|\mathbf{a}_1'p(x)u_{\gamma_0} - \mathbf{a}_2'p(x)u_{\gamma_0}| &= |(\mathbf{a}_1-\mathbf{a}_2)'p(x)u_{\gamma_0}| \leq \|\mathbf{a}_1-\mathbf{a}_2\|\,\|p(x)\|\,|u_{\gamma_0}| \\ &\leq \|\mathbf{a}_1-\mathbf{a}_2\|\,\xi_k(\ell_k+1)\mathbf{r}_{\gamma_0}.\end{aligned}$$

It follows then for some universal constant C_A

$$\sup_Q N(\mathcal{F}, L^2(Q), \delta\|\mathbf{F}\|_{Q,2}) \leq \left(\frac{C_A}{\delta}\right)^k.$$

Pick $\sigma = (\ell_k+1)\mathbf{r}_{\gamma_0}$ and apply Lemma C.1

$$\begin{aligned}\mathbb{E}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n [\tilde{\mathbf{a}}'p(X_i)u_{\gamma_0 i}]\right\|_{\mathcal{F}} &\lesssim (\ell_k+1)\mathbf{r}_{\gamma_0}\sqrt{k\log\xi_k} + \frac{k\xi_k(\ell_k+1)\mathbf{r}_{\gamma_0}\log\xi_k}{\sqrt{n}} \\ &= (\ell_k+1)\mathbf{r}_{\gamma_0}\left(\sqrt{k\log\xi_k} + \frac{k\xi_k\log\xi_k}{\sqrt{n}}\right).\end{aligned}$$

Hence Markov inequality yields

$$\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n [\tilde{\mathbf{a}}'p(X_i)u_{\gamma_0 i}]\right\|_{\mathcal{F}} = O_p\left[(\ell_k+1)\mathbf{r}_{\gamma_0}\left(\sqrt{k\log\xi_k} + \frac{k\xi_k\log\xi_k}{\sqrt{n}}\right)\right].$$

Conclusion follows by noting

$$\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \left[\left(\tilde{\alpha}(X_i) - \mathcal{L}_n\alpha_0(X_i)\right)u_{\gamma_0 i}\right]\right| \leq \|\tilde{a}-a_l\|\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n [\tilde{\mathbf{a}}'p(X_i)u_{\gamma_0 i}]\right\|_{\mathcal{F}}.$$

□

C.2.3 Additional results for Theorem 3.1

Theorem C.1. *If conditions for Theorem 3.1 hold, R_{1BP} defined in Lemma C.5 is $o_p(1)$.*

Proof. Step 0. Consider $\mathcal{L}_n\gamma_0$, the least square projection of γ_0 onto Θ_n. (C.1) is obviously satisfied for γ_0 with

$$\pi_n\gamma_0 = \mathcal{L}_n\gamma_0, \quad r_n = u_{\gamma_0}.$$

Apply Lemma C.7 to R_{1BP}. It follows $R_{1BP}^2 \leq n\hat{T}_1 + n\hat{T}_2$, where

$$\hat{T}_1 = \left\{\mathbb{E}_n\left[\tilde{\alpha}(X)\mathcal{L}_n\gamma_0(X) - m(Z, \mathcal{L}_n\gamma_0(X))\right]\right\}^2; \quad \hat{T}_2 = \left\{\mathbb{E}_n\left[\tilde{\alpha}(X)u_{\gamma_0} - m(Z, u_{\gamma_0})\right]\right\}^2.$$

Step 1: bound $\hat{T}_1$. Note by Lemma C.15, $\|\hat{a}\| = O_p(1)$, so wpa1 $\hat{a}'p(X) \in \Theta_n$. Since by assumption $W_n'W_n - I$ is positive semidefinite, Lemma C.9-(2) can be invoked with $\alpha = \hat{a}'p$ and $f = \mathcal{L}_n\gamma_0$

$$\hat{T}_1 \lesssim \hat{T}_{11} + \hat{T}_{12},$$

where

$$\begin{aligned}
\hat{T}_{11} &= \sup_{g \in \mathcal{H}_{W_n}} \left\{\mathbb{E}_n[\hat{a}'p(X)g(X) - m(Z, g(X))]\right\}^2 \|\mathcal{L}_n\gamma_0\|_{\mathbb{P},2}^2, \\
\hat{T}_{12} &= \lambda_1\mathbb{E}_n[\hat{a}'p(X)]^2 \|\mathcal{L}_n\gamma_0\|_{\mathbb{P},2}^2.
\end{aligned}$$

Notice wpa1 $\hat{T}_{11} = 0$, as for any $g \in \mathcal{H}_{W_n}$

$$\mathbb{E}_n\left[\hat{a}'p(X)g(X) - m(Z, g(X))\right] = \beta'W_n\mathbb{E}_n\left[\left(\hat{a}'p(X)\right)p(X) - m(Z, p(X))\right] = 0,$$

where the first equality follows by linearity of $m(z, \cdot)$ and the second equality is by definition of $\hat{a}$. And $\|\mathcal{L}_n\gamma_0\|_{\mathbb{P},2}^2 = a_l'Ga_l = O(1)$ by L1 and Lemma C.4.

Next, bound $\hat{T}_{12}$. First note

$$\mathbb{E}_n\left[\hat{a}'p(X)\right]^2 = \hat{a}'\hat{G}\hat{a} \leq \|\hat{a}\|^2 \left\|\hat{G}\right\| \leq O_p(1),$$

where the last relation is by Lemmas C.15 and C.13. By assumption $\lambda_1 = o\left(\frac{1}{n}\right)$. It follows

$$\hat{T}_{12} = \lambda_1\mathbb{E}_n\left[\hat{a}'p(X)\right]^2 \|\mathcal{L}_n\gamma_0\|_{\mathbb{P},2}^2 = \lambda_1 O_p(1)O(1) = o_p\left(\frac{1}{n}\right).$$

Step 2: bound $\hat{T}_2$. By Lemma C.8, $\hat{T}_2 = \hat{T}_{21} + \hat{T}_{22}$, where

$$\hat{T}_{21} = O_p\left\{\left[\frac{1}{n}\sum_{i=1}^{n}\left(\tilde{\alpha}(X_i) - \alpha_0(X_i)\right)u_{\gamma_0 i}\right]^2\right\}, \quad \hat{T}_{22} = O_p\left\{\frac{\|u_{\gamma_0}\|_{\mathbb{P},\infty}^2 \wedge \|\alpha_0\|_{\mathbb{P},\infty}^2 \|u_{\gamma_0}\|_{\mathbb{P},2}^2}{n}\right\}.$$

Show both $\hat{T}_{21}$ and $\hat{T}_{22}$ are $o_p(\frac{1}{n})$. First note directly

$$\hat{T}_{22} = O_p\left\{\frac{\left[(\ell_k+1)^2 \wedge \|\alpha_0\|_{\mathbb{P},\infty}^2\right]\mathbf{r}_{\gamma_0}^2}{n}\right\} = o_p\left(\frac{1}{n}\right),$$

where the first relation is by Lemma C.3-(2) and (3), and the second relation is by L3-(1). To bound $\hat{T}_{21}$, we consider two cases.

Case 1: L2-(2)-(a) is satisfied. By Cauchy-Schwarz inequality

$$\hat{T}_{21} = O_p\left[\frac{1}{n}\sum_{i=1}^{n}\left(\tilde{\alpha}(X_i) - \alpha_0(X_i)\right)^2 \frac{1}{n}\sum_{i=1}^{n} u_{\gamma_0 i}^2\right]. \tag{C.8}$$

By iid assumption and Lemma C.3-(2) and (3)

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} u_{\gamma_0 i}^2 - \mathbb{E}u_{\gamma_0}^2\right)^2 = \frac{\mathbb{E}\left[u_{\gamma_0}^2 - \mathbb{E}u_{\gamma_0}^2\right]^2}{n} \leq \frac{\mathbb{E}u_{\gamma_0}^4}{n} \leq \frac{\|u_{\gamma_0}\|_{\mathbb{P},\infty}^2 \|u_{\gamma_0}\|_{\mathbb{P},2}^2}{n} \leq \frac{(\ell_k+1)^2\mathbf{r}_{\gamma_0}^4}{n}.$$

It follows from Markov inequality

$$\frac{1}{n}\sum_{i=1}^{n} u_{\gamma_0 i}^2 - \mathbb{E}u_{\gamma_0}^2 = O_p\left(\frac{\ell_k \mathbf{r}_{\gamma_0}^2}{\sqrt{n}}\right) = O_p\left(\frac{\mathbf{r}_{\gamma_0}}{\sqrt{n}}\ell_k \mathbf{r}_{\gamma_0}\right) = O_p\left(\frac{1}{n}\right),$$

where the last relation is from L2-(2)-(a). It follows

$$\frac{1}{n}\sum_{i=1}^{n} u_{\gamma_0 i}^2 = \frac{1}{n}\sum_{i=1}^{n} u_{\gamma_0 i}^2 - \mathbb{E}u_{\gamma_0}^2 + \mathbb{E}u_{\gamma_0}^2 = O_p\left(\frac{1}{n}\right), \tag{C.9}$$

since $\mathbb{E}u_{\gamma_0}^2 \leq \mathbf{r}_{\gamma_0}^2 = O_p(\frac{1}{n})$ as well by Lemma C.3-(2). Further decompose $\alpha_0 = \mathcal{L}_n\alpha_0 + u_{\alpha_0}$. Triangle inequality yields

$$\frac{1}{n}\sum_{i=1}^{n}\left(\tilde{\alpha}(X_i) - \alpha_0(X_i)\right)^2 \leq \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{\alpha}(X_i) - a_l' p(X_i)\right)^2 + \frac{1}{n}\sum_{i=1}^{n} u_{\alpha_0 i}^2 = o_p(1), \tag{C.10}$$

since from Lemmas C.16 and C.13

$$\frac{1}{n}\sum_{i=1}^{n}\left(\tilde{\alpha}(X_i)-a_l'p(X_i)\right)^2=(\tilde{a}-a_l)'\hat{G}(\tilde{a}-a_l)\leq\|\tilde{a}-a_l\|^2\left\|\hat{G}\right\|\leq o_p(1)O_p(1)=o_p(1),$$

and by Khinchin law of large numbers and Lemma C.3: $\frac{1}{n}\sum_{i=1}^{n}u_{\alpha_0 i}^2\xrightarrow{p}\mathbb{E}u_{\alpha_0}^2=o_p(1)$. Hence $\hat{T}_{21}=o_p(1)O_p(\frac{1}{n})=o_p(\frac{1}{n})$ by displays (C.8), (C.9) and (C.10).

Case 2: L2-(2)-(b) is satisfied. Apply standard decomposition and triangle inequality

$$\hat{T}_{21}\lesssim\left\{\mathbb{E}_n\left[\left(\tilde{\alpha}(X)-\mathcal{L}_n\alpha_0(X)\right)u_{\gamma_0}\right]\right\}^2+\left\{\mathbb{E}_n[u_{\alpha_0}u_{\gamma_0}]\right\}^2. \tag{C.11}$$

By iid assumption and Lemma C.3-(2) and (3)

$$\mathbb{E}\left[\mathbb{E}_n[u_{\alpha_0}u_{\gamma_0}]-\mathbb{E}[u_{\alpha_0}u_{\gamma_0}]\right]^2\leq\frac{\mathbb{E}[u_{\alpha_0}^2u_{\gamma_0}^2]}{n}\leq\frac{\|u_{\gamma_0}\|_{\mathbb{P},\infty}^2\,\mathbb{E}[u_{\alpha_0}^2]}{n}\lesssim\frac{(\ell_k+1)^2\mathbf{r}_{\gamma_0}^2\mathbf{r}_{\alpha_0}^2}{n}.$$

Markov inequality yields

$$\mathbb{E}_n[u_{\alpha_0}u_{\gamma_0}]-\mathbb{E}[u_{\alpha_0}u_{\gamma_0}]=O_p\left(\frac{\ell_k\mathbf{r}_{\gamma_0}\mathbf{r}_{\alpha_0}}{\sqrt{n}}\right)=o_p\left(\frac{1}{\sqrt{n}}\right). \tag{C.12}$$

By Cauchy-Schwarz inequality and Lemma C.3-(2)

$$|\mathbb{E}u_{\alpha_0}u_{\gamma_0}|\leq\|u_{\alpha_0}\|_{\mathbb{P},2}\|u_{\gamma_0}\|_{\mathbb{P},2}\leq\mathbf{r}_{\gamma_0}\mathbf{r}_{\alpha_0}=o_p\left(\frac{1}{\sqrt{n}}\right). \tag{C.13}$$

It follows then by triangle inequality, displays (C.12) and (C.13)

$$\mathbb{E}_n[u_{\alpha_0}u_{\gamma_0}]=\mathbb{E}_n[u_{\alpha_0}u_{\gamma_0}]-\mathbb{E}[u_{\alpha_0}u_{\gamma_0}]+\mathbb{E}[u_{\alpha_0}u_{\gamma_0}]=o_p\left(\frac{1}{\sqrt{n}}\right). \tag{C.14}$$

Finally Lemma C.18 yields

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\left(\tilde{\alpha}(X_i)-\mathcal{L}_n\alpha_0(X_i)\right)u_{\gamma_0 i}\right]\\=&O_p\left[\|\tilde{a}-a_l\|\,(\ell_k+1)\mathbf{r}_{\gamma_0}\left(\sqrt{k\log\xi_k}+\frac{k\xi_k\log\xi_k}{\sqrt{n}}\right)\right],\\=&o_p(1)O_p(1)=o_p(1),\end{aligned} \tag{C.15}$$

by L2-(2)-(b) and Lemma C.16. Hence, displays (C.11), (C.14) and (C.15) together yield $\hat{T}_{21}=o_p(\frac{1}{n})$ as well. Summarizing steps 0-2 concludes that $R_{1BP}^2=n(\hat{T}_1+\hat{T}_2)=o_p(1)$. □

Theorem C.2. *If conditions for Theorem 3.1 hold, R_2 defined in Lemma C.5 is*

$o_p(1)$.

Proof. By standard decomposition, $R_2 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{[\tilde{\alpha}(X_i) - \alpha_0(X_i)]\, e_i\} = \hat{T}_3 + \hat{T}_4$, where

$$\hat{T}_3 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[(\tilde{\alpha}(X_i) - \mathcal{L}_n \alpha_0(X_i))\, e_i\right], \qquad \hat{T}_4 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[(\mathcal{L}_n \alpha_0(X_i) - \alpha_0(X_i))\, e_i\right].$$

It follows by Lemma C.11 (treating $\mathcal{L}_n \alpha_0(X_i) - \alpha_0(X_i) = \mathcal{A}_i(\mathscr{Z}_n)$) and iid assumption

$$\frac{1}{\sqrt{n}} \hat{T}_4 = O_p\left(\sqrt{\frac{\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\mathcal{L}_n \alpha_0(X_i) - \alpha_0(X_i)\right]^2}{n}}\right) = O_p\left(\sqrt{\frac{\mathbb{E} u_{\alpha_0}^2}{n}}\right) = o_p\left(\sqrt{\frac{1}{n}}\right), \tag{C.16}$$

where the last relation is from Lemma C.3-(2) and L2-(2). Next, by definition of $\tilde{\alpha}$ and $\mathcal{L}_n \alpha_0$ and Lemma C.17

$$\hat{T}_3 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[(\tilde{a} - a_l)' p(X_i) e_i\right] = o_p(1). \tag{C.17}$$

Hence $R_2 = o_p(1)$ by displays (C.16) and (C.17). □

C.2.4 Additional results for Theorem 3.2

Lemmas below are understood to hold if conditions for Theorem 3.2 hold.

Lemma C.19. $\mathbb{E}_n[m^2(Z, \hat{\gamma}^s(X) - \gamma_0(X))] = o_p(1)$.

Proof. By standard decomposition and triangle inequality

$$\mathbb{E}_n[m^2(Z, \hat{\gamma}^s(X) - \gamma_0(X))] \lesssim \mathbb{E}_n\left[m^2\left(Z, \hat{\gamma}^s(X) - \mathcal{L}_n \gamma_0(X)\right)\right] + \mathbb{E}_n\left[m^2(Z, u_{\gamma_0})\right].$$

Step 1: bound $\mathbb{E}_n[m^2(Z, u_{\gamma_0})]$. By Khinchin law of large numbers, O-(3), Lemma C.3 and L3-(1)

$$\mathbb{E}_n[m^2(Z, u_{\gamma_0})] \xrightarrow{p} \mathbb{E}[m^2(Z, u_{\gamma_0})] \leq C \mathbb{E} u_{\gamma_0}^2 \leq C \mathbf{r}_{\gamma_0}^2 = o_p(1).$$

Step 2: bound $\mathbb{E}_n[m^2(Z, \hat{\gamma}^s(X) - \mathcal{L}_n \gamma_0(X))]$. Let $\hat{\beta} = \hat{G}^{-} \mathbb{E}_n[p(X) Y]$, and $\tilde{\beta} = \frac{\hat{\beta} - \beta_l}{\left\|\hat{\beta} - \beta_l\right\|}$. It follows by linearity of $m(z, \cdot)$

$$\mathbb{E}_n[m^2(Z, \hat{\gamma}^s(X) - \mathcal{L}_n \gamma_0(X))] = \left\|\hat{\beta} - \beta_l\right\|^2 \mathbb{E}_n m^2(Z, \tilde{\beta}' p(X)) \leq \left\|\hat{\beta} - \beta_l\right\|^2 \sup_{\|\tilde{\beta}\|=1} \mathbb{E}_n m^2(Z, \tilde{\beta}' p(X)).$$

By Theorem 4.1 in Belloni et al. (2015), $\left\|\hat{\beta}-\beta_l\right\| = o_p(1)$. It suffices to show $\sup_{\|\tilde{\beta}\|=1} \mathbb{E}_n m^2(Z,\tilde{\beta}'p(X)) = O_p(1)$.

$$\sup_{\|\tilde{\beta}\|=1} \mathbb{E}_n m^2(Z,\tilde{\beta}'p(X)) \leq \Xi_{11}+\Xi_{12},$$

where

$$\Xi_{11} = \sup_{\|\tilde{\beta}\|=1}\left[\mathbb{E}_n m^2(Z,\tilde{\beta}'p(X)) - \mathbb{E}m^2(Z,\tilde{\beta}'p(X))\right], \quad \Xi_{12} = \sup_{\|\tilde{\beta}\|=1}\left[\mathbb{E}m^2(Z,\tilde{\beta}'p(X))\right].$$

By O-(3), $\Xi_{12} = O(1)$ since

$$\Xi_{12} \leq \sup_{\|\tilde{\beta}\|=1}\left\{C\mathbb{E}\left[\tilde{\beta}'p(X)\right]^2\right\} = \sup_{\|\tilde{\beta}\|=1}\left[C\tilde{\beta}'G\tilde{\beta}\right] \lesssim \sup_{\|\tilde{\beta}\|=1}\left[\left\|\tilde{\beta}\right\|^2\lambda_{\max}(G)\right] < \infty.$$

Next, we show $\Xi_{11} = o_p(1)$ by invoking Corollary 2.2 in Newey (1991).

(1) Compactness. Satisfied since $\tilde{\beta} \in \mathcal{C}_{\tilde{\beta}} = \left\{\beta \in \mathbb{R}^k : \|\beta\| = 1\right\}$.

(2) Pointwise convergence. By O-(3) and L1

$$\mathbb{E}\left[m^2(Z,\tilde{\beta}'p(X))\right] \lesssim \mathbb{E}\left[\tilde{\beta}'p(X)\right]^2 = \tilde{\beta}'G\tilde{\beta} \leq \left\|\tilde{\beta}\right\|^2\|G\| < \infty,$$

it follows by Khinchin law of large numbers that pointwisely $\mathbb{E}_n m^2(Z,\tilde{\beta}'p(X)) \xrightarrow{p} \mathbb{E}m^2(Z,\tilde{\beta}'p(X))$.

(3) Assumption 3A. Let $\mathbf{m}(\beta) = m(z,\beta'p(x))$. Then for any $\tilde{\beta}_1,\tilde{\beta}_2 \in \mathcal{C}_{\tilde{\beta}}$, apply standard decomposition, triangle inequality and linearity of $m(z,\cdot)$:

$$\begin{aligned}&\left|\mathbb{E}_n\mathbf{m}^2\left(\tilde{\beta}_1\right) - \mathbb{E}_n\mathbf{m}^2\left(\tilde{\beta}_2\right)\right| \leq 2\left|\mathbb{E}_n\mathbf{m}\left(\tilde{\beta}_2\right)\mathbb{E}_n\mathbf{m}\left(\tilde{\beta}_1-\tilde{\beta}_2\right)\right| + \left|\mathbb{E}_n\mathbf{m}^2\left(\tilde{\beta}_1-\tilde{\beta}_2\right)\right| \\ \leq& 2\left[\mathbb{E}_n\mathbf{m}^2\left(\tilde{\beta}_2\right)\right]^{1/2}\left[\mathbb{E}_n\mathbf{m}^2\left(\tilde{\beta}_1-\tilde{\beta}_2\right)\right]^{1/2} + \left[\mathbb{E}_n\mathbf{m}^2\left(\tilde{\beta}_1-\tilde{\beta}_2\right)\right]^{1/2}.\end{aligned}$$

By step (2), $\mathbb{E}_n\mathbf{m}^2(\tilde{\beta}_2) \xrightarrow{p} \mathbb{E}\mathbf{m}^2(\tilde{\beta}_2) = O(1)$. By Khinchin law of large numbers and O-(3) again:

$$\begin{aligned}\mathbb{E}_n\mathbf{m}^2\left(\tilde{\beta}_1-\tilde{\beta}_2\right) \xrightarrow{p} \mathbb{E}\mathbf{m}^2\left(\tilde{\beta}_1-\tilde{\beta}_2\right) &= \mathbb{E}\left[m(Z,(\tilde{\beta}_1-\tilde{\beta}_2)'p(X))\right]^2 \\ &\leq C\mathbb{E}\left[(\tilde{\beta}_1-\tilde{\beta}_2)'p(X)\right]^2 \leq C\lambda_{\max}(G)\left\|\tilde{\beta}_1-\tilde{\beta}_2\right\|^2.\end{aligned}$$

Thus $\left|\mathbb{E}_n\mathbf{m}^2(\tilde{\beta}_1) - \mathbb{E}_n\mathbf{m}^2(\tilde{\beta}_2)\right| \leq O_p(1)\left\|\tilde{\beta}_1-\tilde{\beta}_2\right\|$ and Assumption 3A is satisfied. $\Xi_{11} = o_p(1)$ by Corollary 2.2 in Newey (1991). Proof is completed by steps 1 and 2. □

Lemma C.20. $\mathbb{E}_n\left[\tilde{\alpha}(X)(\hat{\gamma}^s(X)-\gamma_0(X))\right]^2 = o_p(1)$.

Proof. By Lemmas C.16 and C.4 and triangle inequality:

$$\|\tilde{a}\| \leq \|\tilde{a}-a_l\| + \|a_l\| = o_p(1) + O(1) = O_p(1).$$

Also by Lemma C.13, $\lambda_{\max}(\hat{G}) = O_p(1)$. Hence

$$\mathbb{E}_n\tilde{\alpha}^2(X) = \tilde{a}'\hat{G}\tilde{a} \leq \|\tilde{a}\|^2 \lambda_{\max}(\hat{G}) = O_p(1).$$

It follows then by Assumption (1) in Theorem 3.2

$$\mathbb{E}_n\left[\tilde{\alpha}(X)(\hat{\gamma}^s(X)-\gamma_0(X))\right]^2 \leq \|\hat{\gamma}^s-\gamma_0\|^2_{\mathbb{P},\infty}\,\mathbb{E}_n\tilde{\alpha}^2(X) = o_p(1)O_p(1) = o_p(1).$$

□

Lemma C.21. $\mathbb{E}_n\left[(\tilde{\alpha}(X)-\alpha_0(X))^2e^2\right] = o_p(1)$.

Proof. Note $\tilde{\alpha}-\alpha_0 = \tilde{\alpha}-\mathcal{L}_n\alpha_0-u_{\gamma_0} = (\tilde{a}-a_l)'p-u_{\gamma_0}$. By standard decomposition

$$\mathbb{E}_n\left[(\tilde{\alpha}(X)-\alpha_0(X))^2e^2\right] \lesssim \Xi_{31}+\Xi_{32},$$

where

$$\Xi_{31} = \mathbb{E}_n\left[\left((\tilde{a}-a_l)'p(X)\right)^2e^2\right], \quad \Xi_{32} = \mathbb{E}_n\left[u^2_{\gamma_0}e^2\right].$$

By Khinchin law of large numbers

$$\Xi_{32} \xrightarrow{p} \mathbb{E}\left[u^2_{\gamma_0}e^2\right] = \mathbb{E}\left[u^2_{\gamma_0}\mathbb{E}[e^2|X]\right] \lesssim \mathbb{E}[u^2_{\gamma_0}] = o_p(1),$$

where the inequality follows from O-(2) so that $\mathbb{E}[e^2|X] < \infty$ since $X \subseteq Z$, and the last last relation is by L2-(2). Next, we show $\Xi_{31} = o_p(1)$ as well. Note

$$\Xi_{31} = (\tilde{a}-a_l)'\mathbb{E}_n\left[p(X)p(X)'e^2\right](\tilde{a}-a_l) \leq \|\tilde{a}-a_l\|^2\left\|\mathbb{E}_n\left[p(X)p(X)'e^2\right]\right\|.$$

Since $\|\tilde{a}-a_l\| = o_p(1)$ by Lemma C.16, it suffices to show $\|\mathbb{E}_n\left[p(X)p(X)'e^2\right]\| = O_p(1)$. To this end, by triangle inequality

$$\left\|\mathbb{E}_n\left[p(X)p(X)'e^2\right]\right\| \leq \left\|\mathbb{E}_n\left[p(X)p(X)'e^2\right]-\mathbb{E}\left[p(X)p(X)'e^2\right]\right\|+\left\|\mathbb{E}\left[p(X)p(X)'e^2\right]\right\|.$$

Since $\mathbb{E}[e^2|X] < \infty$, we have

$$\left\|\mathbb{E}\left[p(X)p(X)'\mathbb{E}[e^2|X]\right]\right\| = \sup_{\|a\|=1}\mathbb{E}\left[\left(a'p(X)\right)^2\mathbb{E}[e^2|X]\right] \lesssim \sup_{\|a\|=1}\mathbb{E}\left[(a'p(X))^2\right] = \|G\| = O(1).$$

And by Lemma 3.1 in Chen and Pouzo (2015b)

$$\left\|\mathbb{E}_n\left[p(X)p(X)'e^2\right] - \mathbb{E}\left[p(X)p(X)'e^2\right]\right\| = o_p(1)$$

by Assumption-(2) in Theorem 3.2. This completes the proof for Lemma C.21. □

C.2.5 Proofs for main results when $\frac{k}{n} \to 0$.

C.2.5.1 Proof of Theorem 3.1

By Lemma C.5, Theorems C.1 and C.2

$$\sqrt{n}\mathbb{E}_n\left[\tilde{\alpha}(X)Y - \theta_0\right] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[m(X_i, \gamma_0(X_i)) + \alpha_0(X_i)(Y_i - \gamma_0(X_i)) - \theta_0] + o_p(1),$$

since both R_{1BP} and R_2 are $o_p(1)$. Conclusion follows by Lindeberg–Lévy central limit theorem.

C.2.5.2 Proof of Corollary 3.1

Similar to proof of Theorem 3.1, we show both R_{1BP} and R_2 are $o_p(1)$. First note by Lemma C.13, $\hat{G}^- = \hat{G}^{-1}$ wpa1. So wpa1

$$\tilde{\alpha}(x) = p(x)'(\hat{G}\mathcal{W}_n\hat{G})^-\hat{G}\mathcal{W}_n\hat{P} = p(x)'(\hat{G}\mathcal{W}_n\hat{G})^{-1}\hat{G}\mathcal{W}_n\hat{P} = p(x)'\hat{G}^{-1}\hat{P}. \quad \text{(C.18)}$$

Thus it suffices to treat $\tilde{\alpha}(x) = p(x)'\hat{G}^{-1}\hat{P}$. Next decompose

$$\gamma_0 = \mathcal{L}_n\gamma_0 + u_{\gamma_0}; \quad \tilde{\alpha} - \alpha_0 = \tilde{\alpha} - \mathcal{L}_n\alpha_0 - u_{\alpha_0}.$$

It follows

$$R_{1BP} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\tilde{\alpha}(X_i)\gamma_0(X_i) - m(Z_i, \gamma_0(X_i))\right] = \hat{\mathcal{T}}_1 + \hat{\mathcal{T}}_2,$$
$$R_2 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\left(\tilde{\alpha}(X_i) - \alpha_0(X_i)\right)e_i\right] = \hat{\mathcal{T}}_3 + \hat{\mathcal{T}}_4,$$

where

$$\hat{\mathcal{T}}_1 = \sqrt{n}\mathbb{E}_n\left[\tilde{\alpha}(X)\mathcal{L}_n\gamma_0(X) - m(Z, \mathcal{L}_n\gamma_0(X))\right], \quad \hat{\mathcal{T}}_2 = \sqrt{n}\mathbb{E}_n\left[\tilde{\alpha}(X)u_{\gamma_0} - m(Z, u_{\gamma_0})\right],$$
$$\hat{\mathcal{T}}_3 = \sqrt{n}\mathbb{E}_n\left[\left(\tilde{\alpha}(X) - \mathcal{L}_n\alpha_0(X)\right)e\right], \quad \hat{\mathcal{T}}_4 = \sqrt{n}\mathbb{E}_n\left[u_{\alpha_0}e\right].$$

Since $\mathcal{L}_n\gamma_0 = \beta_l' p$, $\hat{\mathcal{T}}_1$ admits

$$\begin{aligned}\hat{\mathcal{T}}_1 &= \mathbb{E}_n\left[\tilde{\alpha}(X)\mathcal{L}_n\gamma_0(X) - m(Z, \mathcal{L}_n\gamma_0(X))\right] = \mathbb{E}_n\left[\tilde{\alpha}(X)\left(\beta_l' p(X)\right) - m(Z, \beta_l' p(X))\right] \\ &= \beta_l' \mathbb{E}_n[\tilde{\alpha}(X)p(X) - m(Z, p(X))] = 0,\end{aligned}$$

since $\mathbb{E}_n[m(Z, p(X)) - \tilde{\alpha}(X)p(X)] = \mathbf{0}$ by definition of $\tilde{\alpha}$ when $\lambda_1 = 0$. To show $\hat{\mathcal{T}}_2 = o_p(1)$, note $\hat{\mathcal{T}}_2 = \hat{\mathcal{T}}_{21} + \hat{\mathcal{T}}_{22}$, where

$$\hat{\mathcal{T}}_{21} = \sqrt{n}\mathbb{E}_n\left[\left(\tilde{\alpha}(X) - \alpha_0(X)\right)u_{\gamma_0}\right], \quad \hat{\mathcal{T}}_{22} = \sqrt{n}\mathbb{E}_n\left[\alpha_0(X)u_{\gamma_0} - m(Z, u_{\gamma_0})\right].$$

Same techniques used in Theorem C.1 to bound $\hat{T}_{21}$ and $\hat{T}_{22}$ can be applied for $\hat{\mathcal{T}}_{21}$ and $\hat{\mathcal{T}}_{22}$, respectively. Proofs to show $\hat{\mathcal{T}}_3$ and $\hat{\mathcal{T}}_4$ are $o_p(1)$ are essentially the same with those for terms $\hat{T}_3$ and $\hat{T}_4$ in Theorem C.2. Thus details are omitted. Since both R_{1BP} and R_2 are $o_p(1)$, conclusion follows by Lindeberg–Lévy central limit theorem.

C.2.5.3 Proof of Theorem 3.2

By continuous mapping theorem, $\hat{\Omega} \xrightarrow{p} \Omega$ if

$$\mathbb{E}_n\left[m(Z, \hat{\gamma}^s(X)) + \tilde{\alpha}(X)(Y - \hat{\gamma}^s(X))\right]^2 \xrightarrow{p} \mathbb{E}\left[m(Z, \gamma_0(X)) + \alpha_0(X)(Y - \gamma_0(X))\right]^2$$

and $\hat{\theta}_{BP} \xrightarrow{p} \theta_0$. By Theorem 3.1, $\hat{\theta}_{BP} = \theta_0 + O_p(n^{-1/2})$. It remains to show

$$\mathbb{E}_n\left[m(Z, \hat{\gamma}^s(X)) + \tilde{\alpha}(X)(Y - \hat{\gamma}^s(X))\right]^2 \xrightarrow{p} \mathbb{E}\left[m(Z, \gamma_0(X)) + \alpha_0(X)(Y - \gamma_0(X))\right]^2.$$

To this end, apply standard decomposition

$$\begin{aligned}&\mathbb{E}_n\left[m(Z, \hat{\gamma}^s(X)) + \tilde{\alpha}(X)(Y - \hat{\gamma}^s(X))\right]^2 - \mathbb{E}\left[m(Z, \gamma_0(X)) + \alpha_0(X)(Y - \gamma_0(X))\right]^2 \\ \leq &J_1 + J_2,\end{aligned}$$

where

$$\begin{aligned}J_1 &= \mathbb{E}_n\left[m(Z, \hat{\gamma}^s(X)) + \tilde{\alpha}(X)(Y - \hat{\gamma}^s(X))\right]^2 - \mathbb{E}_n\left[m(Z, \gamma_0(X)) + \alpha_0(X)(Y - \gamma_0(X))\right]^2, \\ J_2 &= \mathbb{E}_n\left[m(Z, \gamma_0(X)) + \alpha_0(X)(Y - \gamma_0(X))\right]^2 - \mathbb{E}\left[m(Z, \gamma_0(X)) + \alpha_0(X)(Y - \gamma_0(X))\right]^2.\end{aligned}$$

The following steps bound J_1 and J_2 respectively and show both of them are $o_p(1)$.

Step 1: bound J_2.

By triangle inequality, O-(2) and (3) and LIE

$$\begin{aligned}\mathbb{E}\left[m(Z,\gamma_0(X))+\alpha_0(X)(Y-\gamma_0(X))\right]^2 &\leq 2\mathbb{E}\left[m^2(Z,\gamma_0(X))\right]+2\mathbb{E}\left[\alpha_0^2(X)e^2\right]\\ &\leq 2C\mathbb{E}[\gamma_0^2(X)]+2\mathbb{E}[\alpha_0^2(X)\mathbb{E}(e^2|X)]\\ &\lesssim \mathbb{E}[\gamma_0^2(X)]+\mathbb{E}[\alpha_0^2(X)]<\infty.\end{aligned}$$

It follows then by Khinchin law of large numbers that $J_2 = o_p(1)$.

Step 2: bound J_1.

Let $\hat{\phi}(z) = m(z,\hat{\gamma}^s(x))+\tilde{\alpha}(x)(y-\hat{\gamma}^s(x))$, $\phi_0(z) = m(z,\gamma_0(x))+\alpha_0(x)(y-\gamma_0(x))$. Decompose $J_1 = J_{11}+J_{12}$, where

$$J_{11} = 2\mathbb{E}_n[\phi_0(\hat{\phi}-\phi_0)], \quad J_{12} = \mathbb{E}_n[\hat{\phi}-\phi_0]^2.$$

By Cauchy-Schwarz inequality, $|J_{11}| \leq 2[\mathbb{E}_n\phi_0^2]^{1/2}\{\mathbb{E}_n[\hat{\phi}-\phi_0]^2\}^{1/2}$. By step 1, $\mathbb{E}_n\phi_0^2 = O_p(1)$. It suffices to show $\mathbb{E}_n[\hat{\phi}-\phi_0]^2 = o_p(1)$ to conclude both $J_{11} = o_p(1)$ and $J_{12} = o_p(1)$.

Step 3: show $\mathbb{E}_n[\hat{\phi}-\phi_0]^2 = o_p(1)$.

By standard decomposition

$$\begin{aligned}\mathbb{E}_n[\hat{\phi}-\phi_0]^2 &= \mathbb{E}_n[m(Z,\hat{\gamma}^s(X))-m(Z,\gamma_0(X))+\tilde{\alpha}(X)(Y-\hat{\gamma}^s(X))-\alpha_0(X)(Y-\gamma_0(X))]^2\\ &= \mathbb{E}_n[m(Z,\hat{\gamma}^s(X)-\gamma_0(X))-\tilde{\alpha}(X)(\hat{\gamma}^s(X)-\gamma_0(X))\\ &\quad+(\tilde{\alpha}(X)-\alpha_0(X))(Y-\gamma_0(X))]^2\\ &\lesssim \mathbb{E}_n[m(Z,\hat{\gamma}^s(X)-\gamma_0(X))-\tilde{\alpha}(X)(\hat{\gamma}^s(X)-\gamma_0(X))]^2\\ &\quad+\mathbb{E}_n[(\tilde{\alpha}(X)-\alpha_0(X))(Y-\gamma_0(X))]^2\\ &\lesssim \Xi_1+\Xi_2+\Xi_3,\end{aligned}$$

where

$$\begin{aligned}&\Xi_1 = \mathbb{E}_n[m^2(Z,\hat{\gamma}^s(X)-\gamma_0(X))], \quad \Xi_2 = \mathbb{E}_n[\tilde{\alpha}(X)(\hat{\gamma}^s(X)-\gamma_0(X))]^2,\\ &\Xi_3 = \mathbb{E}_n[(\tilde{\alpha}(X)-\alpha_0(X))^2e^2].\end{aligned}$$

Conclusion follows by Lemmas C.19, C.20 and C.21.

C.3 Proofs for main results when $\frac{k}{n} \to c < 1$

Notations. Main notations are the same as those in Appendix C.2.

C.3.1 Additional results on asymptotic boundedness

Lemma C.22. *If O, L1 and M1-(1) hold,* $\left\|\hat{G}\right\| = O_p(1)$.

Proof. Note $\hat{G} = \sum_{i=1}^{n} \frac{p(X_i)p(X_i)'}{n}$ and for each X_i, $\lambda_{\min}\left(\frac{p(X_i)p(X_i)'}{n}\right) \geq 0$, $\lambda_{\max}\left(\frac{p(X_i)p(X_i)'}{n}\right) = \left\|\frac{p(X_i)p(X_i)'}{n}\right\| \leq \frac{\xi_k^2}{n}$. Furthermore, $\lambda_{\max}(G) < \infty$ by L1. Thence Theorem 5.1 in Tropp (2015) on matrix Chernoff bounds can be invoked: for every $\varrho > 0$

$$\mathbb{E}\lambda_{\max}(\hat{G}) \leq \frac{e^{\varrho}-1}{\varrho}\lambda_{\max}(G) + \frac{\xi_k^2 \log k}{n\varrho} \to \frac{e^{\varrho}-1}{\varrho}\lambda_{\max}(G) + \frac{c_1}{\varrho} < \infty.$$

It follows by Markov inequality that $\left\|\hat{G}\right\| = \lambda_{\max}\left(\hat{G}\right) = O_p\left[\mathbb{E}\lambda_{\max}\left(\hat{G}\right)\right] = O_p(1)$. □

Lemma C.23. *If conditions for Theorem 3.3 hold, then*

1. $\mathbb{E}_n\tilde{\alpha}^2(X) = O_p(1)$;
2. $\left\|\hat{P}\right\|^{-1} = O_p(1)$;
3. $\{\mathbb{E}_n[\tilde{\alpha}^2(X)]\}^{-1} = O_p(1)$;
4. $V_n^{-1} = O_p(1)$.

Proof. Statement (1): Apply Lemma C.9-(1) with $\alpha = \hat{\alpha} = \hat{a}'p$

$$\lambda_1 \frac{1}{n}\sum_{i=1}^{n}\tilde{\alpha}^2(X_i) \leq \sup_{g\in\mathcal{H}_{W_n}}\{\mathbb{E}_n[\hat{\alpha}(X)g(X) - m(Z, g(X))]\}^2 + \lambda_1\frac{1}{n}\sum_{i=1}^{n}[\hat{\alpha}(X)]^2.$$

By M1-(1), $\hat{G}^- = G^{-1}$ wpa1. Hence $\sup_{g\in\mathcal{H}_{W_n}}\{\mathbb{E}_n[\hat{a}'p(X)g(X) - m(Z, g(X))]\}^2 = 0$ by definition of $\hat{a}$. Further by M2-(1), it follows $\frac{1}{n}\sum_{i=1}^{n}\tilde{\alpha}^2(X_i) \leq \frac{1}{n}\sum_{i=1}^{n}\hat{\alpha}^2(X_i) = O_p(1)$.

Statement (2): Let $K = \{1 \dots k\}$ be the index set of basis function $p(x) = \{p_1(x), \dots, p_k(x)\}'$, and $\tilde{K} \subseteq K$ be some index set such that its cardinality $\left|\tilde{K}\right| = \tilde{k}$, $\tilde{k} \to \infty$ and $\frac{\tilde{k}}{k} \to 0$. Denote $p^{\tilde{K}}(x)$ as a vector of $\tilde{k}$ basis functions selected by $\tilde{K}$: $p_j(x) \in p^{\tilde{K}}(x)$ if and only if $j \in \tilde{K}$. Thus write $\hat{P}^{\tilde{K}} = \frac{1}{n}\sum_{i=1}^{n} m(Z_i, p^{\tilde{K}}(X_i))$.

By construction of $p^{\tilde{K}}(x)$, $\left\|\hat{P}\right\| \geq \left\|\hat{P}_{\tilde{K}}\right\|$, so it suffices to show $\left\|\hat{P}_{\tilde{K}}\right\|^{-1} = O_p(1)$. By decomposition $\hat{P}^{\tilde{K}} = \hat{M}_1^{\tilde{K}} + \hat{M}_2^{\tilde{K}}$, where

$$\hat{M}_1^{\tilde{K}} = \frac{1}{n}\sum_{i=1}^{n} m(Z_i, p^{\tilde{K}}(X_i)) - \mathbb{E}m(Z, p^{\tilde{K}}(X)); \quad \hat{M}_2^{\tilde{K}} = \mathbb{E}m(Z, p^{\tilde{K}}(X)).$$

It suffices to show that $\left\|\hat{M}_2^{\tilde{K}}\right\|$ is bounded away from zero and $\hat{M}_1^{\tilde{K}} = o_p(1)$.

Step 1: show $\hat{M}_1^{\tilde{K}} = o_p(1)$. By iid assumption and O-(3)

$$\mathbb{E}\left\|\hat{M}_1^{\tilde{K}}\right\|^2 \leq \frac{\mathbb{E}\left[m(Z, p^{\tilde{K}}(X))' m(Z, p^{\tilde{K}}(X))\right]}{n} = \frac{\sum_{j\in\tilde{K}} \mathbb{E}m^2(Z, p_j(X))}{n} \lesssim \frac{\tilde{k}}{n} = \frac{\tilde{k}}{k}\frac{k}{n} \to 0,$$

where the last relation follows from $\frac{\tilde{k}}{k} \to 0$ by construction and $\frac{k}{n} \leq \frac{\xi_k^2 \log k}{n} \to c_1$ by M1-(1). It follows by Markov inequality that $\hat{M}_1^{\tilde{K}} = o_p(1)$.

Step 2: show $\left\|\hat{M}_2^{\tilde{K}}\right\|$ *is bounded away from zero.* Consider the least square projection of α_0 on $p^{\tilde{K}}$: $\alpha_0 = \mathcal{L}_n^{\tilde{K}}\alpha_0 + u_{\alpha_o}^{\tilde{K}}$, where $\mathcal{L}_n^{\tilde{K}}\alpha_0 = p^{\tilde{K}\prime} a_l^{\tilde{K}}$, $a_l^{\tilde{K}}$ is the projection coefficient and $u_{\alpha_0}^{\tilde{K}}$ is the projection error. Note by construction and L1, $\mathbb{E}\left[p^{\tilde{K}}(X)p^{\tilde{K}}(X)'\right]$ has eigenvalues bounded from above and away from zero as well. Since

$$\left\|\mathcal{L}_n^{\tilde{K}}\alpha_0\right\|_{\mathbb{P},2}^2 = (a_l^{\tilde{K}})'\mathbb{E}\left[p^{\tilde{K}}(X)p^{\tilde{K}}(X)'\right]a_l^{\tilde{K}} \leq \left\|a_l^{\tilde{K}}\right\|^2 \lambda_{\max}\left\{\mathbb{E}\left[p^{\tilde{K}}(X)p^{\tilde{K}}(X)'\right]\right\},$$

$$\left\|a_l^{\tilde{K}}\right\|^2 \geq \frac{\left\|\mathcal{L}_n^{\tilde{K}}\alpha_0\right\|_{\mathbb{P},2}^2}{\lambda_{\max}\left\{\mathbb{E}\left[p^{\tilde{K}}(X)p^{\tilde{K}}(X)'\right]\right\}} = \frac{\|\alpha_0\|_{\mathbb{P},2}^2 - \left\|u_{\alpha_0}^{\tilde{K}}\right\|_{\mathbb{P},2}^2}{\lambda_{\max}\left\{\mathbb{E}\left[p^{\tilde{K}}(X)p^{\tilde{K}}(X)'\right]\right\}} \geq \frac{\|\alpha_0\|_{\mathbb{P},2}^2 - \left(\mathbf{r}_{\alpha_0}^{\tilde{K}}\right)^2}{\lambda_{\max}\left\{\mathbb{E}\left[p^{\tilde{K}}(X)p^{\tilde{K}}(X)'\right]\right\}},$$

where the second relation is by Pythagoras' theorem and the last relation is by Lemma C.3. It follows $\left\|a_l^{\tilde{K}}\right\|^2$ is bounded away from zero by $\lambda_{\max}\left\{\mathbb{E}\left[p^{\tilde{K}}(X)p^{\tilde{K}}(X)'\right]\right\}$ bounded from above, M2-(4) and construction of $p^{\tilde{K}}$. Moreover, by definition of projection

$$\begin{aligned}
\left\|a_l^{\tilde{K}}\right\| &= \left\|\mathbb{E}\left[p^{\tilde{K}}(X)p^{\tilde{K}}(X)'\right]^{-1}\mathbb{E}\left[\alpha_0(X)p^{\tilde{K}}(X)\right]\right\| \\
&\leq \left\|\mathbb{E}\left[p^{\tilde{K}}(X)p^{\tilde{K}}(X)'\right]^{-1}\right\| \left\|\mathbb{E}\left[\alpha_0(X)p^{\tilde{K}}(X)\right]\right\| \\
&= \lambda_{\min}^{-1}\left\{\mathbb{E}\left[p^{\tilde{K}}(X)p^{\tilde{K}}(X)'\right]\right\} \left\|\mathbb{E}\left[\alpha_0(X)p^{\tilde{K}}(X)\right]\right\|.
\end{aligned}$$

It follows

$$\left\|\hat{M}_2^{\tilde{K}}\right\| = \left\|\mathbb{E}\left[\alpha_0(X)p^{\tilde{K}}(X)\right]\right\| \geq \lambda_{\min}\left\{\mathbb{E}\left[p^{\tilde{K}}(X)p^{\tilde{K}}(X)'\right]\right\}\left\|a_l^{\tilde{K}}\right\|.$$

So $\left\|\hat{M}_2^{\tilde{K}}\right\|$ is also bounded away from zero since $\lambda_{\min}\left\{\mathbb{E}\left[p_{\tilde{K}}(X)p_{\tilde{K}}(X)'\right]\right\}$ is bounded away from zero as well. Conclusion follows by steps 1 and 2.

Statement (3): Recall $\tilde{\alpha} = p'\tilde{a}$, where $\tilde{a} = (\hat{G}\mathcal{W}_n\hat{G} + \lambda_1\hat{G})^{-}\hat{G}\mathcal{W}_n\hat{P}$. Note by M1-(1), $\hat{G}^{-} = \hat{G}^{-1}$ wpa1. Since both $\hat{G}$ and $\mathcal{W}_n$ are invertible wpa1, display (C.6) still holds. That is, wpa1 $\tilde{a} = \Psi_n\hat{G}^{-1}\hat{P}$ where $\Psi_n = \left(\mathcal{W}_n\hat{G} + \lambda_1 I\right)^{-1}\mathcal{W}_n\hat{G}$. Therefore

$$\left[\frac{1}{n}\sum_{i=1}^{n}\tilde{\alpha}^2(X_i)\right]^{-1} = \left(\tilde{a}'\hat{G}\tilde{a}\right)^{-1} = \left(\hat{P}'\hat{G}^{-1}\Psi_n'\hat{G}\Psi_n\hat{G}^{-1}\hat{P}\right)^{-1} \leq \left\|\hat{P}\right\|^{-2}\lambda_{\min}^{-1}\left(\hat{G}^{-1}\Psi_n'\hat{G}\Psi_n\hat{G}^{-1}\right),$$

where the last inequality is by property of the positive definite matrix $\hat{G}^{-1}\Psi_n'\hat{G}\Psi_n\hat{G}^{-1}$. By statement (2), $\left\|\hat{P}\right\|^{-1} = O_p(1)$. It suffices to show $\lambda_{\min}^{-1}\left(\hat{G}^{-1}\Psi_n'\hat{G}\Psi_n\hat{G}^{-1}\right) = O_p(1)$. To this end, write

$$\hat{G}^{-1}\Psi_n'\hat{G}\Psi_n\hat{G}^{-1} = \hat{G}^{-1}G\mathcal{W}_n\left(G\mathcal{W}_n + \lambda_1 I\right)^{-1}\hat{G}\left(\mathcal{W}_n\hat{G} + \lambda_1 I\right)^{-1}\mathcal{W}_n\hat{G}\hat{G}^{-1} = \mathscr{G}_n, \tag{C.19}$$

where $\mathscr{G}_n = \mathcal{W}_n\left(G\mathcal{W}_n + \lambda_1 I\right)^{-1}\hat{G}\left(\mathcal{W}_n\hat{G} + \lambda_1 I\right)^{-1}\mathcal{W}_n$. Thus

$$\begin{aligned}
\lambda_{\min}^{-1}\left(\hat{G}^{-1}\Psi_n'\hat{G}\Psi_n\hat{G}^{-1}\right) &= \lambda_{\max}\left(\mathscr{G}_n^{-1}\right) = \lambda_{\max}\left[\mathcal{W}_n^{-1}\left(\mathcal{W}_n\hat{G} + \lambda_1 I\right)\hat{G}^{-1}\left(G\mathcal{W}_n + \lambda_1 I\right)\mathcal{W}_n^{-1}\right] \\
&= \left\|\hat{G} + 2\lambda_1\mathcal{W}_n^{-1} + \lambda_1^2\mathcal{W}_n^{-1}\hat{G}^{-1}\mathcal{W}_n^{-1}\right\| \\
&\leq \left\|\hat{G}\right\| + \left\|2\lambda_1\mathcal{W}_n^{-1}\right\| + \left\|\lambda_1^2\left(\mathcal{W}_n\hat{G}\mathcal{W}_n\right)^{-1}\right\|,
\end{aligned}$$

where the first equality follows from property of $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$, the second equality applies display (C.19), the third equality is by definition of matrix norm $\|\cdot\|$ and rewriting, and final inequality follows from triangle inequality. By Lemma C.22, $\left\|\hat{G}\right\| = O_p(1)$; By L3-(2), $\mathcal{W}_n$ has eigenvalues bounded away from 1, so $\|\mathcal{W}_n^{-1}\| = \frac{1}{\lambda_{\min}(\mathcal{W}_n)} = O_p(1)$ and it follows $\|2\lambda_1\mathcal{W}_n^{-1}\| = 2\lambda_1\|\mathcal{W}_n^{-1}\| = O_p(\lambda_1) = o_p(1)$. Finally, $\left\|\lambda_1^2\left(\mathcal{W}_n\hat{G}\mathcal{W}_n\right)^{-1}\right\| \leq \frac{\lambda_1^2}{\lambda_{\min}^2(\mathcal{W}_n)\lambda_{\min}\left(\hat{G}\right)} = O_p(1)$ by M2-(3). Conclusion follows.

Statement (4): It follows by definition of V_n, $\mathbb{E}\left[e_i^2|Z_i\right]$ bounded from below almost surely and statement (3)

$$V_n^{-1} = \left\{\frac{1}{n}\sum_{i=1}^{n}\left[\tilde{\alpha}^2(X_i)\mathbb{E}[e_i^2|Z_i]\right]\right\}^{-1} \lesssim \left\{\frac{1}{n}\sum_{i=1}^{n}\left[\tilde{\alpha}^2(X_i)\right]\right\}^{-1} = O_p(1).$$

□

Lemma C.24. *If O, L1 and M1 hold, and in addition $\lambda_{\min}(\hat{G})$ is bounded away from zero wpa1, then:*

1. $\left\|\hat{G}^{-1}\right\| = O_p(1)$.
2. $\|\hat{a} - a_l\| = O_p(1)$.
3. $\|\hat{a}\| = O_p(1)$.

Proof. Statement (1): By M1, $\hat{G}$ is invertible wpa1. Also since $\lambda_{\min}(\hat{G})$ is bounded away from zero wpa1, $\left\|\hat{G}^{-1}\right\| = \frac{1}{\lambda_{\min}(\hat{G})} = O_p(1)$.

Statement (2): By L1, $a_l = G^{-1}\mathbb{E}[p(X)\alpha_0(X)]$ is well defined. By M1-(1), $\hat{G}$ is invertible wpa1, so $\hat{a} = \hat{G}^{-1}\hat{P}$ wpa1. Hence similar to Lemma C.15, the following decomposition still stands

$$\hat{a} - a_l = \hat{G}^{-1}\mathbb{E}_n e^R + \hat{G}^{-1}\mathbb{E}_n[u_{\alpha_0} p(X)],$$

where recall $e^R(z) = m(z, p(x)) - \alpha_0(x)p(x)$. The same idea for proof of Lemma C.14 can be used to get

$$\mathbb{E}\left\|\mathbb{E}_n e^R\right\|^2 \lesssim \frac{\xi_k^2}{n} \to c_1(\text{up to log term}).$$

By Markov inequality

$$\left\|\mathbb{E}_n e^R\right\| = O_p\left(\sqrt{\frac{\xi_k^2}{n}}\right) = O_p(1). \tag{C.20}$$

By statement (1) and display (C.20)

$$\left\|\hat{G}^{-1}\mathbb{E}_n e^R\right\| \leq \left\|\hat{G}^{-1}\right\| \left\|\mathbb{E}_n e^R\right\| = O_p(1)O_p(1). \tag{C.21}$$

By statement (1) and a similar argument with Lemma C.14-(2)

$$\begin{aligned}\left\|\hat{G}^{-1}\mathbb{E}_n[u_{\alpha_0} p(X)]\right\| &= \left\|\hat{G}^{-1/2}\hat{G}^{-1/2}\mathbb{E}_n[u_{\alpha_0} p(X)]\right\| \leq \left\|\hat{G}^{-1/2}\right\| \left\|\hat{G}^{-1/2}\mathbb{E}_n[u_{\alpha_0} p(X)]\right\| \\ &\leq O_p(1)O_p(\mathbf{r}_{\alpha_0}) = O_p(1). \end{aligned} \tag{C.22}$$

Statement (2) then follows by triangle inequality, displays (C.21) and (C.22).

Statement (3): This directly follows from statement (2), Lemma C.4 and triangle inequality

$$\|\hat{a}\| \leq \|\hat{a} - a_l\| + \|a_l\| = O_p(1) + O(1) = O_p(1). \tag{C.23}$$

□

C.3.2 Additional results for Theorem 3.3.

Theorem C.3. *If conditions for Theorem 3.3 hold*

$$R_{1BP} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [\tilde{\alpha}(X_i)\gamma_0(X_i) - m(Z_i, \gamma_0(X_i))] = o_p(1).$$

Proof. Decompose $\gamma_0 = \mathcal{L}_n\gamma_0 + u_{\gamma_0}$, with $\mathcal{L}_n\gamma_0 = \beta_l' p$. Since $\beta_l' p \in \Theta_n$, Lemma C.7 can be invoked to yield $R_{1BP} = \tilde{T}_1 + \tilde{T}_2$, where

$$\tilde{T}_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [m(Z_i, \beta_l' p(X_i)) - \tilde{\alpha}(X_i)(\beta_l' p(X_i))], \quad \tilde{T}_2 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [m(Z_i, u_{\gamma_{0i}}) - \tilde{\alpha}(X_i)u_{\gamma_{0i}}].$$

Step 1: bound $\tilde{T}_1$. Apply Lemma C.9-(2) with $\alpha = \hat{\alpha} = \hat{a}'p$ and $f = \beta_l' p$

$$\frac{1}{n}\tilde{T}_1^2 \lesssim \left\{ \sup_{g \in \mathcal{H}_{W_n}} [\mathbb{E}_n (\hat{\alpha}(X)g(X) - m(Z, g(X)))]^2 + \lambda_1 \mathbb{E}_n \hat{\alpha}^2(X) \right\} \|\beta_l' p\|_{\mathbb{P},2}^2, \tag{C.24}$$

Note $\hat{G}^- = \hat{G}^{-1}$ wpa1 by M1-(1). It follows $\sup_{g \in \mathcal{H}_{W_n}} \{\mathbb{E}_n [\hat{a}'p(X)g(X) - m(Z, g(X))]\}^2 = 0$ wpa1 by definition of $\hat{a}$. Moreover, by L1 and Lemma C.4

$$\|\beta_l' p\|_{\mathbb{P},2}^2 = \beta_l' G \beta_l \lesssim \|\beta_l\|^2 \lambda_{\max}(G) = O(1)O(1) = O(1).$$

By M2-(1), $\frac{1}{n}\sum_{i=1}^{n} \hat{\alpha}(X_i)^2 = O_p(1)$. Finally since $\lambda_1 = o(\frac{1}{n})$, conclude

$$\tilde{T}_1^2 \lesssim no(\frac{1}{n})O_p(1)O(1) = o_p(1).$$

Step 2: bound $\tilde{T}_2$. Apply Lemmas C.8 and C.3-(3)

$$\frac{1}{n}\tilde{T}_2^2 = O_p\left\{[\mathbb{E}_n (\tilde{\alpha}(X) - \alpha_0(X)) u_{\gamma_0}]^2\right\} + O_p\left\{ \frac{\left[(\ell_k + 1)^2 \wedge \|\alpha_0\|_{\mathbb{P},\infty}^2\right] \mathbf{r}_{\gamma_0}^2}{n} \right\},$$

where

$$O_p\left\{ \frac{\left[(\ell_k + 1)^2 \wedge \|\alpha_0\|_{\mathbb{P},\infty}^2\right] \mathbf{r}_{\gamma_0}^2}{n} \right\} = o_p\left(\frac{1}{n}\right)$$

directly under L3-(1). By iid assumption, Lemma C.3, Markov inequality and

M1-(2)

$$\frac{1}{n}\sum_{i=1}^{n} u_{\gamma_0 i}^2 - \mathbb{E}u_{\gamma_0}^2 = O_p\left(\frac{\ell_k \mathbf{r}_{\gamma_0}^2}{\sqrt{n}}\right) = O_p\left(\frac{\ell_k \mathbf{r}_{\gamma_0}}{\sqrt{n}}\mathbf{r}_{\gamma_0}\right) = o_p\left(\frac{1}{n}\right),$$
$$\mathbb{E}u_{\gamma_0}^2 = \|u_{\gamma_0}\|_{\mathbb{P},2}^2 \leq \mathbf{r}_{\gamma_0}^2 = o_p\left(\frac{1}{n}\right).$$

Hence $\frac{1}{n}\sum_{i=1}^{n} u_{\gamma_0 i}^2 = \frac{1}{n}\sum_{i=1}^{n} u_{\gamma_0 i}^2 - \mathbb{E}u_{\gamma_0}^2 + \mathbb{E}u_{\gamma_0}^2 = o_p\left(\frac{1}{n}\right)$. By triangle inequality

$$\frac{1}{n}\sum_{i=1}^{n}\left(\tilde{\alpha}(X_i) - \alpha_0(X_i)\right)^2 \lesssim \frac{1}{n}\sum_{i=1}^{n}\tilde{\alpha}^2(X_i) + \frac{1}{n}\sum_{i=1}^{n}\alpha_0^2(X_i) = O_p(1),$$

since $\frac{1}{n}\sum_{i=1}^{n}\alpha_0^2(X_i) \xrightarrow{p} \mathbb{E}\alpha_0^2(X) < \infty$ by Khinchin law of large numbers and $\frac{1}{n}\sum_{i=1}^{n}\tilde{\alpha}^2(X_i) = O_p(1)$ by Lemma C.23. It follows from Cauchy-Schwarz inequality

$$\left\{\mathbb{E}_n\left[\left(\tilde{\alpha}(X) - \alpha_0(X)\right)u_{\gamma_0}\right]\right\}^2 \leq \left(\frac{1}{n}\sum_{i=1}^{n} u_{\gamma_0 i}^2\right)\frac{1}{n}\sum_{i=1}^{n}\left(\tilde{\alpha}(X_i) - \alpha_0(X_i)\right)^2 = o_p\left(\frac{1}{n}\right).$$

Hence $\tilde{T}_2 = o_p(1)$. Proof is completed by combining above two steps. □

Theorem C.4. *Under conditions of Theorem 3.3*

$$n^{-1/2}V_n^{-1/2}\sum_{i=1}^{n}[\tilde{\alpha}(X_i)(Y_i - \gamma_0(X_i))] \xrightarrow{d} N(0,1),$$

where $V_n = \frac{1}{n}\sum_{i=1}^{n}\{\tilde{\alpha}^2(X_i)\mathbb{E}[e_i^2|Z_i]\}$ *and* $N(0,1)$ *is a standard normal random variable.*

Proof. Let $\mathcal{A}_i(\mathscr{Z}_n) = \tilde{\alpha}(X_i)$ and verify conditions of Lemma C.12: *Conditions (1) and (2)* follow from M2-(4). *Conditions (3) and (4)* are satisfied by Lemma C.23. Proof is completed by directly invoking Lemma C.12. □

C.3.3 Proofs for main results when $\frac{k}{n} \to c < 1$

C.3.3.1 Proof of Theorem 3.3

By standard decomposition

$$\sqrt{n}V_n^{-1/2}\left[\hat{\theta}_{BP} - \mathbb{E}_n m(Z, \gamma_0(X))\right] = n^{-1/2}V_n^{-1/2}\sum_{i=1}^{n}[\tilde{\alpha}(X_i)(Y_i - \gamma_0(X_i))] + V_n^{-1/2}R_{1BP},$$

where $R_{1BP} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [\tilde{\alpha}(X_i)\gamma_0(X_i) - m(Z_i, \gamma_0(X_i))]$. By Theorem C.3 and Lemma C.23 $V_n^{-1/2} R_{1BP} = O_p(1)o_p(1) = o_p(1)$. Conclusion then follows from Theorem C.4. To see $\hat{\theta}_{BP} - \theta_0 = O_p\left(\frac{1}{\sqrt{n}}\right)$, note by decomposition: $\hat{\theta}_{BP} - \theta_0 = \hat{\theta}_{BP} - \mathbb{E}_n[m(Z, \gamma_0(X))] + \mathbb{E}_n[m(Z, \gamma_0(X))] - \theta_0$. Then $\hat{\theta}_{BP} - \mathbb{E}_n[m(Z, \gamma_0(X))] = O_p\left(\frac{1}{\sqrt{n}}\right)$ directly from previous statement and $\mathbb{E}_n[m(Z, \gamma_0(X))] - \theta_0 = O_p\left(\frac{1}{\sqrt{n}}\right)$ by Markov inequality.

C.3.3.2 Proof of Corollary 3.2

The main decomposition in the proof of Theorem 3.3 still holds

$$\sqrt{n} V_n^{-1/2} \left[\hat{\theta}_{BP} - \mathbb{E}_n m(Z, \gamma_0(X))\right] = n^{-1/2} V_n^{-1/2} \sum_{i=1}^{n} [\tilde{\alpha}(X_i)(Y_i - \gamma_0(X_i))] + V_n^{-1/2} R_{1BP},$$

with $R_{1BP} = \tilde{\mathcal{T}}_1 + \tilde{\mathcal{T}}_2$, where

$$\tilde{\mathcal{T}}_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[m(Z_i, \beta_l' p(X_i)) - \tilde{\alpha}(X_i)\left(\beta_l' p(X_i)\right)\right], \quad \tilde{\mathcal{T}}_2 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[m(Z_i, u_{\gamma_{0i}}) - \tilde{\alpha}(X_i) u_{\gamma_{0i}}\right].$$

By M1-(1), $\hat{G}^- = \hat{G}^{-1}$ wpa1. Hence $\tilde{\alpha} = p'\hat{G}^{-1}\hat{P}$ wpa1. See also (C.18) in the proof of Corollary 3.1. It follows by linearity of $m(z, \cdot)$, $\beta_l = O(1)$ and definition of $\tilde{\alpha}$

$$\begin{aligned} \tilde{T}_1 &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[m(Z_i, \beta_l' p(X_i)) - \tilde{\alpha}(X_i)\left(\beta_l' p(X_i)\right)\right] \\ &= \beta_l' \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[m(Z_i, p(X_i)) - \tilde{\alpha}(X_i) p(X_i)\right] \right\} = 0. \end{aligned}$$

$\tilde{T}_2$ can be controlled the same way as in Theorem C.3 to conclude $\tilde{T}_2 = o_p(1)$. Next, since now $\tilde{\alpha} = p'\hat{G}^{-1}\hat{P}$ wpa1, it follows

$$\left[\frac{1}{n} \sum_{i=1}^{n} \tilde{\alpha}^2(X_i)\right]^{-1} = \left(\hat{P}'\hat{G}^{-1}\hat{P}\right)^{-1} \leq \left\|\hat{P}\right\|^{-2} \lambda_{\min}^{-1}(\hat{G}^{-1}) = \left\|\hat{P}\right\|^{-2} \lambda_{\max}(\hat{G}).$$

Thus, $\left[\frac{1}{n} \sum_{i=1}^{n} \tilde{\alpha}^2(X_i)\right]^{-1} = O_p(1)$ by Lemmas C.22 and C.23. It follows then by M2-(4) that

$$V_n^{-1} = \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[\tilde{\alpha}^2(X_i) \mathbb{E}[e_i^2 | Z_i]\right] \right\}^{-1} \lesssim \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[\tilde{\alpha}^2(X_i)\right] \right\}^{-1} = O_p(1).$$

Hence $V_n^{-1/2} R_{1BP} = o_p(1)$. The rest of the proof is similar to Theorem 3.3. Details are omitted.

C.3.4 Sufficient conditions

Lemma C.25. *Suppose O, L1 and M1 hold. Then either of the two conditions below is sufficient for M2-(1).*

1. $\mathbb{P}\{\lambda_{\min}(\hat{G}) \geq c_2\} \to 1$ for some $c_2 > 0$;
2. $\hat{P} = \mathbb{E}_n[\varpi(Z)p(X)]$ for some scalar valued function $\varpi(z)$ and $\mathbb{E}[\varpi(Z)^2] < \infty$.

Proof. If condition (1) holds, by Lemmas C.24 and C.23 and $\hat{\alpha} = \hat{a}'p$

$$\left|\frac{1}{n}\sum_{i=1}^n \hat{\alpha}^2(X_i)\right| = \left|\hat{a}'\hat{G}\hat{a}\right| \leq \|\hat{a}\|^2 \left\|\hat{G}\right\| = O_p(1)O_p(1) = O_p(1).$$

If condition (2) holds, note $\hat{a} = \hat{G}^- \hat{P} = \hat{G}^{-1}\hat{P}$ wpa1 by M1-(1). Hence wpa1

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \hat{\alpha}^2(X_i) &= \frac{1}{n}\sum_{i=1}^n [p'(X_i)\hat{a}]^2 = \frac{1}{n}\sum_{i=1}^n \left[p'(X_i)\hat{G}^{-1}\hat{P}\right]^2 \\ &= \frac{1}{n}\sum_{i=1}^n \left[p'(X_i)\hat{G}^{-1}\mathbb{E}_n[\varpi(Z)p(X)]\right]^2 = \frac{1}{n}\sum_{i=1}^n \left[\tilde{\mathcal{L}}_n\varpi(X_i)\right]^2,\end{aligned}$$

where $\tilde{\mathcal{L}}_n\varpi$ is the empirical projection of ϖ onto Θ_n. But

$$\frac{1}{n}\sum_{i=1}^n \left[\tilde{\mathcal{L}}_n\varpi(X_i)\right]^2 \leq \frac{1}{n}\sum_{i=1}^n \varpi(Z_i)^2 \xrightarrow{p} \mathbb{E}\varpi^2(Z) < \infty,$$

where the first inequality is by definition of empirical projection, the second relation follows by Khinchin law of large numbers, and the last inequality is by assumption. As a result, $\frac{1}{n}\sum_{i=1}^n \hat{\alpha}^2(X_i) = O_p(1)$ as well. □

The following lemma gives a primitive condition so that condition (1) in Lemma C.25 holds.

Lemma C.26. *If O, L1 and M1-(1) hold, and in addition $G = I$ and $\sqrt{2c_1} + \frac{1}{3}c_1 < 1$, then $\lambda_{\min}(\hat{G}) \geq c_2 = 1 - \sqrt{2c_1} - \frac{1}{3}c_1 > 0$ wpa1.*

Proof. Let $S_i = \frac{p(X_i)p(X_i)'}{n}$. Hence $\sum_{i=1}^n \mathbb{E}S_i = G$ by iid assumption, and $\mathbb{E}\left[S_i - \mathbb{E}S_i\right]$ is a zero matrix by construction. It follows by triangle inequality that $\lambda_{\max}(S_i -$

$\mathbb{E}S_i) \leq \frac{\xi_k^2}{n} + \frac{\lambda_{\max}(G)}{n}$. Invoke matrix Bernstein inequality for symmetric matrix (Theorem 6.6.1 of Tropp, 2015)

$$\mathbb{E}\left\|\hat{G} - G\right\| = \mathbb{E}\lambda_{\max}\left[\sum_{i=1}^{n}(S_i - \mathbb{E}S_i)\right] \leq \sqrt{2v_* \log k} + \frac{1}{3}\left(\frac{\xi_k^2}{n} + \frac{\lambda_{\max}(G)}{n}\right)\log k, \tag{C.25}$$

where

$$\begin{aligned} v_* &= \left\|\sum_{i=1}^{n}\mathbb{E}\left(S_i - \mathbb{E}S_i\right)\left(S_i - \mathbb{E}S_i\right)\right\| \leq \sum_{i=1}^{n}\left\|\mathbb{E}\left(S_i - \mathbb{E}S_i\right)\left(S_i - \mathbb{E}S_i\right)\right\| \qquad \text{(C.26)} \\ &= \sum_{i=1}^{n}\left\|\mathbb{E}S_i^2 - \mathbb{E}S_i\mathbb{E}S_i\right\| \leq \sum_{i=1}^{n}\left\|\mathbb{E}S_i^2\right\| = \frac{1}{n^2}n\left\|\mathbb{E}\left[(p(X)p(X)')^2\right]\right\| \leq \frac{\xi_k^2}{n}\lambda_{\max}(G). \end{aligned}$$

The first relation of (C.26) is by definition v_* in Theorem 6.6.1 of Tropp (2015), and the first inequality is by triangle inequality, the second equality is a direct calculation, the second inequality is due to the positive semidefiniteness of $\mathbb{E}S_i\mathbb{E}S_i$, the third equality is by directly rewriting $\mathbb{E}S_i^2$ and identical distribution assumption; the final inequality uses the property that for any $a \in \mathbb{R}^k$ with $\|a\| = 1$

$$a'\left[p(X)p(X)'p(X)p(X)'\right]a \leq \xi_k^2 a'\left[p(X)p(X)'\right]a.$$

Combining (C.25) and (C.26)

$$\begin{aligned} \mathbb{E}\left\|\hat{G} - G\right\| &\leq \sqrt{2\frac{\xi_k^2}{n}\lambda_{\max}(G)\log k} + \frac{1}{3}\left(\frac{\xi_k^2}{n} + \frac{\lambda_{\max}(G)}{n}\right)\log k \\ &\to \sqrt{2c_1\lambda_{\max}(G)} + \frac{1}{3}c_1 = \sqrt{2c_1} + \frac{1}{3}c_1 < 1, \qquad \text{(C.27)} \end{aligned}$$

where the last inequality follows from assumption $G = I$ so $\lambda_{\max}(G) = 1$. Now suppose wpa1, $\lambda_{\min}(\hat{G}) < c_2$. Then there exists $a \in \mathbb{R}^k$ with $\|a\| = 1$ such that $a'\hat{G}a < c_2$. Hence

$$\left\|\hat{G} - G\right\| \geq \left|a'(\hat{G} - G)a\right| = \left|a'\hat{G}a - a'Ga\right| = \left|a'\hat{G}a - 1\right| > 1 - c_2 > \sqrt{2c_1} + \frac{1}{3}c_1.$$

That is, wpa1, $\left\|\hat{G} - G\right\|$ will be larger than $\sqrt{2c_1} + \frac{1}{3}c_1$, which violates (C.27) since we always have $\left\|\hat{G} - G\right\| \geq 0$. Therefore wpa1 all eigenvalues of $\hat{G}$ are no smaller than c_2. □

The next two lemmas discusses several scenarios under which $\frac{\max_i|\tilde{\alpha}(X_i)|}{\sqrt{n}} = o_p(1)$ can be satisfied. The first one is trivial.

Lemma C.27. *If conditions for Theorem 3.3 hold and* $\frac{\xi_k^2 \log k}{n} \to 0$*, then* $\frac{\max_i|\tilde{\alpha}(X_i)|}{\sqrt{n}} = o_p(1)$.

Proof. If $\frac{\xi_k^2 \log k}{n} \to 0$, $\lambda_{\min}(\hat{G})$ is bounded away from zero wpa1 by Lemma C.13. Then by Lemma C.24, $\|\hat{a}\| = O_p(1)$. Proof of Lemma C.16 can be recycled so

$$\|\tilde{a}\| \leq \|\tilde{a} - \hat{a}\| + \|\hat{a}\| \leq \|\Psi_n - I\| \|\hat{a}\| + \|\hat{a}\| = o_p(1) + O_p(1) = O_p(1).$$

Since $\tilde{\alpha} = \tilde{a}' p$, apply Cauchy-Schwarz inequality and definition of ξ_k

$$\begin{aligned}\max_i |\tilde{\alpha}(X_i)| &\leq \sup_{x \in \mathcal{X}} |\tilde{\alpha}(x)| = \sup_{x \in \mathcal{X}} |\tilde{a}' p(x)| \leq \sup_{x \in \mathcal{X}} |\|\tilde{a}\| \, \|p(x)\||\\ &= \|\tilde{a}\| \sup_{x \in \mathcal{X}} \|p(x)\| = \|\tilde{a}\| \, \xi_k = O_p(\xi_k).\end{aligned}$$

Therefore $\frac{\max_i |\tilde{\alpha}(X_i)|}{\sqrt{n}} = O_p\left(\frac{\xi_k}{\sqrt{n}}\right) = o_p(1)$. □

To introduce the next lemma, let $\tilde{p}(x) = \frac{p'(x)\hat{G}^{-1}}{\left\|p'(x)\hat{G}^{-1}\right\|}$.

Lemma C.28. *Suppose conditions for Theorem 3.3 hold. Then* $\frac{\max_i |\tilde{\alpha}(X_i)|}{\sqrt{n}} = o_p(1)$ *if:*

1. There exists some $c_2 > 0$ such that $\mathbb{P}\{\lambda_{\min}(\hat{G}) \geq c_2\} \to 1$;
2. $\lambda_1 = 0$, $\frac{\ell_k^2 \mathbf{r}_{\alpha 0}}{\sqrt{n}} = o(1)$;
3. There exists some ξ_k^L such that wpa1,

$$\sup_{x_1, x_2 \in \mathcal{X}, x_1 \neq x_2} \|\tilde{p}(x_1) - \tilde{p}(x_2)\| \leq \xi_k^L \|x_1 - x_2\|.$$

4. One of the following two conditions holds:
   (a) There exists some $\Delta_i = \Delta(Z_i)$ and $m > 2$ such that $\mathbb{E}_n[e^R(e^R)'] = \mathbb{E}_n[\Delta^2 p(X) p(X)']$ and wpa1 $\mathbb{E}\left[\max_i |\Delta_i| \, |X_1 \cdots X_n\right] \lesssim n^{1/m}$, and $\frac{n^{1/m}\sqrt{\log \xi_k^L}}{n^{1/2}} \to 0$.
   (b) For some $t < 1$, $\left\|e^R(e^R)'\right\| \leq \xi_k^2 n^t$ for each $i = 1 \ldots n$, $\lambda_{\max}\left\{\mathbb{E}[e^R(e^R)']\right\} \leq \frac{\xi_k^2 n^t}{n} \log k$, and $\frac{n^{t/2}}{n^{1/2}}\sqrt{\log \xi_k^L} \to 0$.

Proof. The proof is long and deferred to Appendix C.5. □

C.4 Proofs for main results when $\frac{k}{n} \to \infty$

Notations. Main notations are the same as those in Appendix C.2. But now $\tilde{\alpha} = p'\tilde{a}$, where $\tilde{a}$ is the solution of display (2.32). Further write

$$\mathbf{W}_n = \hat{G} W_n' W_n, \quad \hat{\mathcal{G}} = \hat{G} W_n' W_n \hat{G} + \lambda_1 \hat{G} = \mathbf{W}_n \hat{G} + \lambda_1 \hat{G}, \quad \hat{\mathcal{M}} = \mathbf{W}_n \hat{P}.$$

C.4.1 Additional convergence results when $\frac{k}{n} \to \infty$

Lemma C.29. *If O and H1 hold and* $\Lambda_n \sqrt{\frac{\log k}{n}} \to 0$*, then*

1. $\mathbb{E}\left[\left\|\hat{G} - G\right\|_{\max}\right] = O\left(\Lambda_n \sqrt{\frac{\log k}{n}}\right)$;
2. $\left\|\hat{G} - G\right\|_{\max} = O_p\left(\Lambda_n \sqrt{\frac{\log k}{n}}\right)$;
3. $\left\|\hat{G}\right\|_{\max} = O_p(1)$.

Proof. Statement (1): By definition, $\left\|\hat{G} - G\right\|_{\max} = \max_{1\le j,l\le k} \left|\frac{1}{n}\sum_{i=1}^{n} \Delta_{j,l}^{G}(i)\right|$, where $\Delta_{j,l}^{G}(i) = p_j(X_i)p_l(X_i) - \mathbb{E}p_j(X_i)p_l(X_i)$. Note $\mathbb{E}\Delta_{j,l}^{G}(i) = 0$, $\left|\Delta_{j,l}^{G}(i)\right| \le |p_j(X_i)p_l(X_i)| + |\mathbb{E}p_j(X_i)p_l(X_i)| \le 2\Lambda_n^2$. Let $\sigma_{\max}^2 = \max_{1\le j,l\le k} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left|\Delta_{j,l}^{G}(i)\right|^2$. It follows

$$\sigma_{\max}^2 = \max_{1\le j,l\le k} \mathbb{E}\left|\Delta_{j,l}^{G}\right|^2 \le \max_{1\le j,l\le k} \mathbb{E}[p_j(X)p_l(X)]^2 \le \Lambda_n^2 \max_{1\le j\le k} \mathbb{E}[p_j^2(X)] = O(\Lambda_n^2),$$

where the first relation is by identical distribution assumption, the second inequality follows by property of variance, the third relation is by definition of Λ_n, and the last relation is due to H1. Further denote $\tilde{\Delta}_{j,l}^{G}(i) = \frac{\Delta_{j,l}^{G}(i)}{\sigma_{\max}}$. Note for all constant $L > 0$

$$\begin{aligned}
\mathbb{E}\max_{1\le j,l\le k}\left|\frac{1}{n}\sum_{i=1}^{n}\tilde{\Delta}_{j,l}^{G}(i)\right| &= \frac{L}{n}\mathbb{E}\max_{1\le j,l\le k}\left|\frac{1}{L}\sum_{i=1}^{n}\tilde{\Delta}_{j,l}^{G}(i)\right| \\
&\le \frac{L}{n}\log\left\{\mathbb{E}\exp\left[\max_{1\le j,l\le k}\left|\frac{1}{L}\sum_{i=1}^{n}\tilde{\Delta}_{j,l}^{G}(i)\right|\right]\right\} \\
&\le \frac{L}{n}\log\left\{\mathbb{E}\sum_{j,l=1}^{k}\exp\left[\left|\frac{1}{L}\sum_{i=1}^{n}\tilde{\Delta}_{j,l}^{G}(i)\right|\right]\right\} \\
&= \frac{L}{n}\log\left\{\sum_{j,l=1}^{k}\mathbb{E}\exp\left[\left|\frac{1}{L}\sum_{i=1}^{n}\tilde{\Delta}_{j,l}^{G}(i)\right|\right]\right\}, \qquad \text{(C.28)}
\end{aligned}$$

where the first relation is simply by rewriting equation, the second inequality is by Jensen's inequality, the third relation is by property of $\max(\cdot)$ and the final relation is due to linearity. To bound $\mathbb{E}\exp\left[\left|\frac{1}{L}\sum_{i=1}^{n}\tilde{\Delta}_{j,l}^{G}(i)\right|\right]$ for each $j, l = 1 \dots k$, let $\frac{2\Lambda_n^2}{\sigma_{\max}} = \tilde{L}$. Claim

$$\text{For any } L > \tilde{L}, \mathbb{E}\exp\left[\left|\frac{1}{L}\sum_{i=1}^{n}\tilde{\Delta}_{j,l}^{G}(i)\right|\right] \le 2\exp\left[\frac{n}{2(L^2 - L\tilde{L})}\right]. \qquad \text{(C.29)}$$

By claim (C.29) and display (C.28), for each $L > \tilde{L}$

$$\mathbb{E} \max_{1\le j,l\le k} \left| \frac{1}{n} \sum_{i=1}^{n} \tilde{\Delta}^G_{j,l}(i) \right| \le \frac{L}{n} \log \left\{ 2k^2 \exp \left[\frac{n}{2(L^2 - L\tilde{L})} \right] \right\} = \frac{L \log(2k^2)}{n} + \frac{1}{2(L - \tilde{L})}.$$

Now picking $L = \tilde{L} + \sqrt{\frac{n}{\log 2k^2}}$ yields

$$\mathbb{E} \max_{1\le j,l\le k} \left| \frac{1}{n} \sum_{i=1}^{n} \tilde{\Delta}^G_{j,l}(i) \right| \le \frac{\tilde{L} \log 2k^2}{n} + \frac{3}{2} \sqrt{\frac{\log 2k^2}{n}}.$$

Finally recall $\tilde{L} = \frac{2\Lambda_n^2}{\sigma_{\max}}$, conclusion follows

$$\begin{aligned}
\mathbb{E} \left[\max_{1\le j,l\le k} \left| \frac{1}{n} \sum_{i=1}^{n} \Delta^G_{j,l}(i) \right| \right] &= \sigma_{\max} \mathbb{E} \left[\max_{1\le j,l\le k} \left| \frac{1}{n} \sum_{i=1}^{n} \tilde{\Delta}^G_{j,l}(i) \right| \right] = \sigma_{\max} O \left(\frac{\Lambda_n^2}{\sigma_{\max}} \frac{\log k}{n} + \sqrt{\frac{\log k}{n}} \right) \\
&= O \left(\frac{\Lambda_n^2 \log k}{n} + \sigma_{\max} \sqrt{\frac{\log k}{n}} \right) = O \left(\frac{\Lambda_n^2 \log k}{n} + \Lambda_n \sqrt{\frac{\log k}{n}} \right) \\
&= O \left[\Lambda_n \sqrt{\frac{\log k}{n}} \left(\Lambda_n \sqrt{\frac{\log k}{n}} + 1 \right) \right] = O \left(\Lambda_n \sqrt{\frac{\log k}{n}} \right),
\end{aligned}$$

where the last relation follows from assumption that $\Lambda_n \sqrt{\frac{\log k}{n}} \to 0$.

Now show claim (C.29) holds. By construction, $\mathbb{E}\tilde{\Delta}^G_{j,l}(i) = 0$, for each $i = 1 \dots n$. By iid assumption and definition of $\sigma^2_{\max}$

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left| \tilde{\Delta}^G_{j,l}(i) \right|^2 = \mathbb{E} \left| \tilde{\Delta}^G_{j,l}(i) \right|^2 = \frac{\mathbb{E} \left| \tilde{\Delta}^G_{j,l}(i) \right|^2}{\sigma^2_{\max}} \le \frac{\max_{1\le j,l\le k} \mathbb{E} \left| \Delta^G_{j,l} \right|^2}{\sigma^2_{\max}} \le 1;$$

While for $t = 3, 4 \dots$

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left| \tilde{\Delta}^G_{j,l}(i) \right|^t = \mathbb{E} \left| \tilde{\Delta}^G_{j,l}(i) \right|^t \le \mathbb{E} \left| \tilde{\Delta}^G_{j,l}(i) \right|^2 \left(\tilde{L} \right)^{t-2} \le \left(\tilde{L} \right)^{t-2}.$$

So by Lemma 14.1 in Bühlmann and Van De Geer (2011) and standard algebra, for any $L > \tilde{L}$

$$\begin{aligned}
\mathbb{E}\exp\left[\frac{1}{L}\sum_{i=1}^{n}\tilde{\Delta}_{j,l}^{G}(i)\right] &= \mathbb{E}\prod_{i=1}^{n}\left\{\exp\left[\frac{\tilde{\Delta}_{j,l}^{G}(i)}{L}\right]\right\} = \prod_{i=1}^{n}\mathbb{E}\exp\left[\frac{\tilde{\Delta}_{j,l}^{G}(i)}{L}\right] \\
&= \exp\left\{\sum_{i=1}^{n}\log\mathbb{E}\exp\left[\frac{\tilde{\Delta}_{j,l}^{G}(i)}{L}\right]\right\} \\
&\leq \exp\left\{\sum_{i=1}^{n}\left[\mathbb{E}\exp\left|\frac{\tilde{\Delta}_{j,l}^{G}(i)}{L}\right| - 1 - \mathbb{E}\left|\frac{\tilde{\Delta}_{j,l}^{G}(i)}{L}\right|\right]\right\} \\
&\leq \exp\left[\sum_{t=2}^{\infty}\sum_{i=1}^{n}\left(\frac{\mathbb{E}\left|\tilde{\Delta}_{j,l}^{G}(i)\right|^{t}}{t!L^{t}}\right)\right] = \exp\left[n\sum_{t=2}^{\infty}\left(\frac{\mathbb{E}\left|\tilde{\Delta}_{j,l}^{G}(i)\right|^{t}}{t!L^{t}}\right)\right] \\
&\leq \exp\left[\frac{n}{2L^{2}}\sum_{t'=0}^{\infty}\left(\frac{\tilde{L}}{L}\right)^{t'}\right] = \exp\left[\frac{n}{2L^{2}}\frac{1}{1-\frac{\tilde{L}}{L}}\right] = \exp\left[\frac{n}{2(L^{2}-L\tilde{L})}\right].
\end{aligned} \tag{C.30}$$

Hence for each $L > \tilde{L}$

$$\begin{aligned}
\mathbb{E}\exp\left|\frac{1}{L}\sum_{i=1}^{n}\tilde{\Delta}_{j,l}^{G}(i)\right| &\leq \mathbb{E}\exp\left[\frac{1}{L}\sum_{i=1}^{n}\tilde{\Delta}_{j,l}^{G}(i)\right] + \mathbb{E}\exp\left[\frac{1}{L}\sum_{i=1}^{n}\left(-\tilde{\Delta}_{j,l}^{G}(i)\right)\right] \\
&\leq 2\exp\left[\frac{n}{2(L^{2}-L\tilde{L})}\right],
\end{aligned}$$

where the first relation is because $e^{|X|} \leq e^{X} + e^{-X}$ for any X, and the second relation follows from display C.30 (which applies to $\mathbb{E}\exp\left[-\frac{1}{L}\sum_{i=1}^{n}\tilde{\Delta}_{j,l}^{G}(i)\right]$ trivially as well).

Statement (2): This follows from statement (1) and Markov inequality.

Statement (3): This follows from triangle inequality, statement (2) and $\|G\|_{\max} < \infty$ by H1. □

C.4.2 Additional results for Theorem 3.4

Lemma C.30. *If conditions for Theorem 3.4 hold,* then R_{1DR} defined in Lemma C.6 is $o_p(1)$.

Proof. The proof follows from three steps.

Step 1: preparations. By standard decomposition

$$R_{1DR} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[m(Z_i, \hat{\gamma}(X_i) - \gamma_0(X_i)) - \tilde{\alpha}(X_i)(\hat{\gamma}(X_i) - \gamma_0(X_i))\right] = \bar{T}_1 + \bar{T}_2,$$

where

$$\bar{T}_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[m(Z_i, \hat{\gamma}(X_i) - \gamma_0(X_i)) - \alpha_0(X_i)(\hat{\gamma}(X_i) - \gamma_0(X_i)) \right],$$
$$\bar{T}_2 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[(\alpha_0(X_i) - \tilde{\alpha}(X_i))(\hat{\gamma}(X_i) - \gamma_0(X_i)) \right].$$

In the following we show both $\bar{T}_1$ and $\bar{T}_2$ are $o_p(1)$.

Step 2: bound $\bar{T}_1$. By H2-(2), $\hat{\gamma}$ is estimated from a different iid sample (wlog, call it $\mathcal{S}$) independent from the main sample. Then conditional on $\mathcal{S}$, function $\tilde{f}(z) = m(z, \hat{\gamma}(x) - \gamma_0(x)) - \alpha_0(x)(\hat{\gamma}(x) - \gamma_0(x))$ is iid distributed. Further by definition of α_0, $\mathbb{E}[\tilde{f}|\mathcal{S}] = 0$. So by O and H2-(4)

$$\begin{aligned}
\mathbb{E}[\tilde{f}^2|\mathcal{S}] &\lesssim \mathbb{E}[\alpha_0(X)^2(\hat{\gamma}(X) - \gamma_0(X))^2|\mathcal{S}] + \mathbb{E}[m(Z, \hat{\gamma}(X) - \gamma_0(X))^2|\mathcal{S}] \\
&\lesssim \|\alpha_0\|_{\mathbb{P},\infty} \mathbb{E}[(\hat{\gamma}(X) - \gamma_0(X))^2|\mathcal{S}] + C\mathbb{E}[(\hat{\gamma}(X) - \gamma_0(X))^2|\mathcal{S}] \\
&\lesssim \|\alpha_0\|_{\mathbb{P},\infty} \|\hat{\gamma} - \gamma_0\|_{\mathbb{P},2}^2 = o(1).
\end{aligned}$$

It follows $\mathbb{E}\left[\left(\mathbb{E}_n \tilde{f} \right)^2 |\mathcal{S} \right] = \frac{\mathbb{E}[\tilde{f}^2|\mathcal{S}]}{n} = o(\frac{1}{n})$. By a conditional version of Markov inequality, $\mathbb{E}_n \left[\tilde{f}|\mathcal{S} \right] = o_p(\frac{1}{\sqrt{n}})$. Then Lemma C.10 yields $\mathbb{E}_n \tilde{f} = o_p(\frac{1}{\sqrt{n}})$, or $\bar{T}_1 = \sqrt{n}\mathbb{E}_n \tilde{f} = o_p(1)$.

Step 3: bound $\bar{T}_2$. By Cauchy-Schwarz inequality

$$\begin{aligned}
\bar{T}_2^2 &= \frac{1}{n} \left(\sum_{i=1}^{n} \left[(\alpha_0(X_i) - \tilde{\alpha}(X_i))(\hat{\gamma}(X_i) - \gamma_0(X_i)) \right] \right)^2 \\
&\leq n \frac{1}{n} \sum_{i=1}^{n} \left[\alpha_0(X_i) - \tilde{\alpha}(X_i) \right]^2 \frac{1}{n} \sum_{i=1}^{n} \left[\hat{\gamma}(X_i) - \gamma_0(X_i) \right]^2 = n \bar{T}_{21} \bar{T}_{22}, \qquad \text{(C.31)}
\end{aligned}$$

where

$$\bar{T}_{21} = \frac{1}{n} \sum_{i=1}^{n} \left[\alpha_0(X_i) - \tilde{\alpha}(X_i) \right]^2, \quad \bar{T}_{22} = \frac{1}{n} \sum_{i=1}^{n} \left[\hat{\gamma}(X_i) - \gamma_0(X_i) \right]^2.$$

Step 3-1: bound $\bar{T}_{22}$. Similar to that of $\bar{T}_1$, by H-(2), function $\hat{\gamma} - \gamma_0$ is iid distributed. So conditional on $\mathcal{S}$, $\mathbb{E}\left[\left| \bar{T}_{22} \right| |\mathcal{S} \right] = \bar{T}_{\Delta 1} + \bar{T}_{\Delta 2}$, where

$$\bar{T}_{\Delta 1} = \mathbb{E}\left[\left| \mathbb{E}_n[\hat{\gamma}(X_i) - \gamma_0(X_i)]^2 - \mathbb{E}[(\hat{\gamma}(X_i) - \gamma_0(X_i))^2|\mathcal{S}] \right| |\mathcal{S} \right],$$
$$\bar{T}_{\Delta 2} = \mathbb{E}\left[(\hat{\gamma}(X_i) - \gamma_0(X_i))^2 |\mathcal{S} \right].$$

By H2-(2)

$$\bar{T}_{\Delta 2} = \|\hat{\gamma} - \gamma_0\|_{\mathbb{P},2}^2 = O_p\left[(\varphi_n^\gamma)^2\right] \to 0.$$

Also by iid assumption, $\bar{T}_{\Delta 1} \lesssim \mathbb{E}\left[(\hat{\gamma}(X) - \gamma_0(X))^2 | \mathcal{S}\right] = O_p\left[(\varphi_n^\gamma)^2\right]$ as well. Hence conditional on $\mathcal{S}$

$$\mathbb{E}\left[\left|\bar{T}_{22}\right| | \mathcal{S}\right] = O_p\left[(\varphi_n^\gamma)^2\right] + O_p\left[(\varphi_n^\gamma)^2\right] = O_p\left[(\varphi_n^\gamma)^2\right].$$

Conditional Markov inequality yields $(\bar{T}_{22}|\mathcal{S}) = O_p[(\varphi_n^\gamma)^2]$. It follows by Lemma C.10 again

$$\bar{T}_{22} = O_p[(\varphi_n^\gamma)^2]. \tag{C.32}$$

Step 3-2: bound $\bar{T}_{21}$. Since by definition $u_* = \alpha_0 - \alpha_*$, standard decomposition yields

$$\bar{T}_{21} = \frac{1}{n}\sum_{i=1}^n \left[\alpha_0(X_i) - \tilde{\alpha}(X_i)\right]^2 \leq \frac{1}{n}\sum_{i=1}^n \left[\tilde{\alpha}(X_i) - \alpha_*(X_i)\right]^2 + \frac{1}{n}\sum_{i=1}^n u_*^2.$$

By iid assumption, $\frac{1}{n}\sum_{i=1}^n u_{*i}^2 \xrightarrow{p} \mathbb{E}[u_*^2] = \|u_*\|_{\mathbb{P},2}^2 = \mu_*^2$. So $\frac{1}{n}\sum_{i=1}^n u_{*i}^2 = O_p(\mu_*^2)$. Moreover, by definition of φ_n^α

$$\frac{1}{n}\sum_{i=1}^n \left(\tilde{\alpha}(X_i) - \alpha_*(X_i)\right)^2 = (\tilde{a} - a_*)'\hat{G}(\tilde{a} - a_*) \leq O_p\left[(\varphi_n^\alpha)^2\right]. \tag{C.33}$$

Summarizing displays (C.31), (C.32) and (C.33), we conclude

$$\bar{T}_2 = \sqrt{n}\left[O_p(\varphi_n^\gamma \varphi_n^\alpha) + O_p(\varphi_n^\gamma \mu_*)\right] = o_p(1).$$

Conclusion then follows from steps 1-3. □

Lemma C.31. *If conditions for Theorem 3.4 hold, R_2* defined in Lemma C.6 is $o_p(1)$.

Proof. By standard decomposition and definition of α_*

$$R_2 = \frac{1}{\sqrt{n}}\sum_{i=1}^n \left[(\tilde{\alpha}(X_i) - \alpha_0(X_i))e_i\right] = \bar{T}_3 + \bar{T}_4,$$

where

$$\bar{T}_3 = \frac{1}{\sqrt{n}}\sum_{i=1}^n \left[(\tilde{\alpha}(X_i) - \alpha_*(X_i))e_i\right], \quad \bar{T}_4 = -\frac{1}{\sqrt{n}}\sum_{i=1}^n u_{*i}e_i.$$

By Lemma C.11 and iid assumption

$$\frac{1}{\sqrt{n}}\bar{T}_4 = O_p\left(\sqrt{\frac{\frac{1}{n}\sum_{i=1}^n \mathbb{E}[u_{*i}e_i]^2}{n}}\right) = O_p\left(\sqrt{\frac{\mathbb{E}u_*^2}{n}}\right) = O_p\left(\sqrt{\frac{\mu_*^2}{n}}\right) = o_p\left(\sqrt{\frac{1}{n}}\right), \tag{C.34}$$

since $\mu_* \to 0$ due to H1. To bound $\bar{T}_3$, decompose

$$\bar{T}_3 = \frac{1}{\sqrt{n}}\sum_{i=1}^n [(\tilde{a}-a_*)'p(X_i)e_i] = \|\tilde{a}-a_*\|_1 \frac{1}{\sqrt{n}}\sum_{i=1}^n [\tilde{\nu}'p(X_i)e_i],$$

where $\tilde{\nu} = \frac{\tilde{a}-a_*}{\|\tilde{a}-a_*\|_1}$. By H2-(3), $\|\tilde{a}-a_*\|_1 = o_p(1)$. It remains to bound $\frac{1}{\sqrt{n}}\sum_{i=1}^n [\tilde{\nu}'p(X_i)e_i]$. Notice $\tilde{\nu}$ is a function of $(Z_1 \cdots Z_n)$, so Lemma C.11 can be invoked with $\mathcal{A}_i(\mathscr{Z}_n) = \tilde{\nu}'p(X_i)$.

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n A_i^2(\mathscr{Z}_n)\right] = \mathbb{E}\left[\tilde{\nu}'\hat{G}\tilde{\nu}\right] \lesssim \mathbb{E}\left[\|\tilde{\nu}\|_1^2 \left\|\hat{G}\right\|_{\max}\right] = \mathbb{E}\left[\left\|\hat{G}\right\|_{\max}\right] \lesssim 1,$$

where the first inequality uses Cauchy-Schwarz inequality, the second equality is by $\|\tilde{\nu}\|_1 = 1$ and the last relation follows $\mathbb{E}\left[\|G]\|_{\max}\right] < \infty$. To see this, note by triangle inequality, Lemma C.29-(3) and H1

$$\mathbb{E}\left[\left\|\hat{G}\right\|_{\max}\right] \le \mathbb{E}\left[\left\|\hat{G}-G\right\|_{\max}\right] + \mathbb{E}\left[\|G]\|_{\max}\right] < \infty.$$

Therefore, Lemma C.11 yields $\frac{1}{n}\sum_{i=1}^n [\tilde{\nu}'p(X_i)e_i] = O_p\left(\frac{1}{\sqrt{n}}\right)$ and

$$\bar{T}_3 = \sqrt{n}o_p(1)O_p\left(\frac{1}{\sqrt{n}}\right) = o_p(1). \tag{C.35}$$

Final conclusion follows from (C.34) and (C.35). □

C.4.3 Additional results for Theorem 3.5

Lemma C.32. *If conditions for Theorem 3.5 hold,* then $2(\tilde{a}-a_*)'(\hat{\mathcal{M}}-\hat{\mathcal{G}}a_*) \le \|\tilde{a}-a_*\|_1 \lambda_0$ wpa1, where $\lambda_0 = 2\left[(\varepsilon_n^R + \varepsilon_n^u + \tilde{C}\mu_*)(\varepsilon_n^{\mathbf{W}} + \lambda_1) + \lambda_1(\varepsilon_n^m + \tilde{C}C^{1/2})\right]$.

Proof. By Holder's inequality

$$2(\tilde{a}-a_*)'(\hat{\mathcal{M}}-\hat{\mathcal{G}}a_*) \leq \left|2(\tilde{a}-a_*)'(\hat{\mathcal{M}}-\hat{\mathcal{G}}a_*)\right| \leq 2\left\|\tilde{a}-a_*\right\|_1 \left\|\hat{\mathcal{M}}-\hat{\mathcal{G}}a_*\right\|_\infty.$$

Hence it suffices to show $2\left\|\hat{\mathcal{M}}-\hat{\mathcal{G}}a_*\right\|_\infty \leq \lambda_0$. Firstly, note

$$\left\|\mathbb{E}[p(X)u_*]\right\|_\infty = \max_{1\leq j\leq k}\left|\mathbb{E}[p_j(X)u_*]\right| \leq \max_{1\leq j\leq k}\left[\mathbb{E}p_j^2(X)\right]^{1/2}\left[\mathbb{E}u_*^2\right]^{1/2} \leq \tilde{C}\mu_*. \tag{C.36}$$

Secondly, standard decomposition yields

$$\begin{aligned}\hat{P}-\hat{G}a_* &= \mathbb{E}_n[m(Z,p(X))] - \mathbb{E}_n[p(X)\alpha_*(X)] \\ &= \mathbb{E}_n[m(Z,p(X)) - p(X)\alpha_0(X)] + \mathbb{E}_n[p(X)u_* - \mathbb{E}(p(X)u_*)] + \mathbb{E}[p(X)u_*].\end{aligned}$$

Hence by H3, display (C.36) and triangle inequality

$$\left\|\hat{P}-\hat{G}a_*\right\|_\infty \leq \varepsilon_n^R + \varepsilon_n^u + \tilde{C}\mu_*. \tag{C.37}$$

Therefore by definition of induced $\|\cdot\|_\infty$ matrix norm and H3-(2), wpa1

$$\left\|(\mathbf{W}_n+\lambda_1 I)(\hat{P}-\hat{G}a_*)\right\|_\infty \leq \left\|\mathbf{W}_n+\lambda_1 I\right\|_\infty \left\|\hat{P}-\hat{G}a_*\right\|_\infty \leq (\varepsilon_n^R+\varepsilon_n^u+\tilde{C}\mu_*)(\varepsilon_n^{\mathbf{W}}+\lambda_1). \tag{C.38}$$

Thirdly, by O and H1

$$\left\|\mathbb{E}m(Z,p(X))\right\|_\infty = \max_{1\leq j\leq k}\left|\mathbb{E}[m(Z,p_j(X))]\right| \leq \max_{1\leq j\leq k}\left[\mathbb{E}p_j^2(X)\right]^{1/2} C^{1/2} \leq \tilde{C}C^{1/2} < \infty.$$

Triangle inequality and H3 yield

$$\left\|\hat{P}\right\|_\infty \leq \left\|\mathbb{E}_n[m(Z,p(X)) - \mathbb{E}m(Z,p(X))]\right\|_\infty + \left\|\mathbb{E}m(Z,p(X))\right\|_\infty \leq \varepsilon_n^m + \tilde{C}C^{1/2}. \tag{C.39}$$

Final conclusion follows by triangle inequality, displays (C.37), (C.38) and (C.39)

$$\begin{aligned}\left\|\hat{\mathcal{M}}-\hat{\mathcal{G}}a_*\right\|_\infty &= \left\|\mathbf{W}_n\hat{P} - (\mathbf{W}_n\hat{G}+\lambda_1\hat{G})a_*\right\|_\infty = \left\|(\mathbf{W}_n+\lambda_1 I)(\hat{P}-\hat{G}a_*) - \lambda_1\hat{P}\right\|_\infty \\ &\leq \left\|(\mathbf{W}_n+\lambda_1 I)(\hat{P}-\hat{G}a_*)\right\|_\infty + \lambda_1\left\|\hat{P}\right\|_\infty \\ &\leq (\varepsilon_n^R+\varepsilon_n^u+\tilde{C}\mu_*)(\varepsilon_n^{\mathbf{W}}+\lambda_1) + \lambda_1(\varepsilon_n^m+\tilde{C}C^{1/2}).\end{aligned}$$

□

Lemma C.33. *If conditions for Theorem 3.5 hold, then* $2(\tilde{a}-a_*)'\hat{\mathcal{G}}(\tilde{a}-a_*)+$

$\lambda_2 \left\|\tilde{a}_{A_*^c}\right\|_1 \leq 3\lambda_2 \left\|\tilde{a}_{A_*} - a_{*A_*}\right\|_1$ *wpa1.*

Proof. Start from definition of $\tilde{a}$. Since

$$\tilde{a} = \arg\min_{a\in\mathbb{R}^k} (\hat{G}a - \hat{P})' W_n' W_n (\hat{G}a - \hat{P}) + \lambda_1 a' \hat{G} a + \lambda_2 \left\|a\right\|_1,$$

it follows

$$\begin{aligned} &(\hat{G}\tilde{a} - \hat{P})' W_n' W_n (\hat{G}\tilde{a} - \hat{P}) + \lambda_1 \tilde{a}' \hat{G} \tilde{a} + \lambda_2 \left\|\tilde{a}\right\|_1 \\ \leq &(\hat{G}a_* - \hat{P})' W_n' W_n (\hat{G}a_* - \hat{P}) + \lambda_1 a_*' \hat{G} a_* + \lambda_2 \left\|a_*\right\|_1, \end{aligned}$$

or by symmetry of $W_n' W_n$ and $\hat{G}$

$$\begin{aligned} &\tilde{a}' \hat{G} W_n' W_n \hat{G} \tilde{a} - 2\tilde{a}' \hat{G} W_n' W_n \hat{P} + \lambda_1 \tilde{a}' \hat{G} \tilde{a} + \lambda_2 \left\|\tilde{a}\right\|_1 \\ \leq &a_*' \hat{G} W_n' W_n \hat{G} a_* - 2a_*' \hat{G} W_n' W_n \hat{P} + \lambda_1 a_*' \hat{G} a_* + \lambda_2 \left\|a_*\right\|_1. \end{aligned} \tag{C.40}$$

Re-write display (C.40) using notation $\hat{\mathcal{G}}$ and $\hat{\mathcal{M}}$

$$\tilde{a}' \hat{\mathcal{G}} \tilde{a} - 2\tilde{a}' \hat{\mathcal{M}} + \lambda_2 \left\|\tilde{a}\right\|_1 \leq a'_* \hat{\mathcal{G}} \tilde{a} - 2a_*' \hat{\mathcal{M}} + \lambda_2 \left\|a_*\right\|_1. \tag{C.41}$$

With $\tilde{a} = \tilde{a} - a_* + a_*$ trivially, (C.41) becomes

$$(\tilde{a} - a_* + a_*)' \hat{\mathcal{G}} (\tilde{a} - a_* + a_*) - 2(\tilde{a} - a_* + a_*)' \hat{\mathcal{M}} + \lambda_2 \left\|\tilde{a}\right\|_1 \leq a_*' \hat{\mathcal{G}} a_* - 2a_*' \hat{\mathcal{M}} + \lambda_2 \left\|a_*\right\|_1,$$

or by symmetry of $\hat{\mathcal{G}}$ as well

$$(\tilde{a} - a_*)' \hat{\mathcal{G}} (\tilde{a} - a_*) + \lambda_2 \left\|\tilde{a}\right\|_1 \leq 2(\tilde{a} - a_*)' (\hat{\mathcal{M}} - \hat{\mathcal{G}} a_*) + \lambda_2 \left\|a_*\right\|_1.$$

Let $\mathcal{E}_n$ be the event such that $2(\tilde{a} - a_*)' (\hat{\mathcal{M}} - \hat{\mathcal{G}} a_*) \leq \left\|\tilde{a} - a_*\right\|_1 \lambda_0$, it follows by Lemma C.32 that $\mathbb{P}(\mathcal{E}_n) \to 1$. So wpa1

$$2(\tilde{a} - a_*)' \hat{\mathcal{G}} (\tilde{a} - a_*) + 2\lambda_2 \left\|\tilde{a}\right\|_1 \leq 2\lambda_0 \left\|\tilde{a} - a_*\right\|_1 + 2\lambda_2 \left\|a_*\right\|_1,$$

or, since $\lambda_2 \geq 2\lambda_0$

$$2(\tilde{a} - a_*)' \hat{\mathcal{G}} (\tilde{a} - a_*) + 2\lambda_2 \left\|\tilde{a}\right\|_1 \leq \left\|\tilde{a} - a_*\right\|_1 \lambda_2 + 2\lambda_2 \left\|a_*\right\|_1. \tag{C.42}$$

To this end, note by triangle inequality

$$\begin{aligned}\|\tilde{a}\|_1 &= \|\tilde{a}_{A_*}\|_1 + \left\|\tilde{a}_{A_*^c}\right\|_1 = \|a_{*A_*} - (\tilde{a}_{A_*} + a_{*A_*})\|_1 + \left\|\tilde{a}_{A_*^c}\right\|_1 \\ &\geq \|a_{*A_*}\|_1 - \|\tilde{a}_{A_*} - a_{*A_*}\|_1 + \left\|\tilde{a}_{A_*^c}\right\|_1 .\end{aligned}$$

So (C.42) yields

$$2(\tilde{a}-a_*)'\hat{\mathcal{G}}(\tilde{a}-a_*)-2\lambda_2 \|\tilde{a}_{A_*} - a_{*A_*}\|_1+2\lambda_2 \|a_{*A_*}\|_1+2\lambda_2 \left\|\tilde{a}_{A_*^c}\right\|_1 \leq \|\tilde{a} - a_*\|_1 \lambda_2+2\lambda_2 \|a_*\|_1 .$$

Note $\left\|a_{*A_*^c}\right\|_1 = 0$, so

$$\|\tilde{a} - a_*\|_1 = \|\tilde{a}_{A_*} - a_{*A_*}\|_1 + \left\|\tilde{a}_{*A_*^c} - a_{*A_*^c}\right\|_1 = \|\tilde{a}_{A_*} - a_{*A_*}\|_1 + \left\|\tilde{a}_{*A_*^c}\right\|_1 .$$

Also $\|a_*\|_1 = \|a_{*A_*}\|_1$, it follows

$$2(\tilde{a} - a_*)'\hat{\mathcal{G}}(\tilde{a} - a_*) + 2\lambda_2 \|a_{*A_*}\|_1 + \lambda_2 \left\|\tilde{a}_{A_*^c}\right\|_1 \leq 3\lambda_2 \|\tilde{a}_{A_*} - a_{*A_*}\|_1 + 2\lambda_2 \|a_{*A_*}\|_1 . \tag{C.43}$$

Simplifying (C.43) yields

$$2(\tilde{a} - a_*)'\hat{\mathcal{G}}(\tilde{a} - a_*) + \lambda_2 \left\|\tilde{a}_{A_*^c}\right\|_1 \leq 3\lambda_2 \|\tilde{a}_{A_*} - a_{*A_*}\|_1 ,$$

which is the desired result. □

C.4.4 Proof of main results when $\frac{k}{n} \to \infty$

C.4.4.1 Proof of Theorem 3.4

By Lemma C.6

$$\begin{aligned}&\sqrt{n}\mathbb{E}_n \left[m(Z, \hat{\gamma}(X)) + \tilde{\alpha}(X)(Y - \hat{\gamma}(X)) - \theta_0\right] \\ =&\frac{1}{\sqrt{n}} \sum_{i=1}^{n} [m(Z_i, \gamma_0(X_i)) + \alpha_0(X_i)(Y_i - \gamma_0(X_i)) - \theta_0] + R_{1DR} + R_2,\end{aligned}$$

where $R_{1DR} = o_p(1)$ by Lemma C.30 and $R_2 = o_p(1)$ by Lemma C.31. Conclusion follows from Lindeberg–Lévy central limit theorem.

C.4.4.2 Proof of Theorem 3.5

Note

$$
\begin{aligned}
&2(\tilde{a}-a_*)'\hat{\mathcal{G}}(\tilde{a}-a_*)+\lambda_2\left\|\tilde{a}-a_*\right\|_1\\
=&2(\tilde{a}-a_*)'\hat{\mathcal{G}}(\tilde{a}-a_*)+\lambda_2\left\|\tilde{a}_{A_*}-a_{*A_*}\right\|_1+\lambda_2\left\|\tilde{a}_{A_*^c}\right\|_1\\
\leq&3\lambda_2\left\|\tilde{a}_{A_*}-a_{*A_*}\right\|_1+\lambda_2\left\|\tilde{a}_{A_*}-a_{*A_*}\right\|_1=4\lambda_2\left\|\tilde{a}_{A_*}-a_{*A_*}\right\|_1,
\end{aligned}
$$

where the second relation follows wpa1 by Lemma C.33. Since $a_{*A_*^c}=\mathbf{0}$, Lemma C.33 also implies that $\left\|\tilde{a}_{A_*^c}-a_{*A_*^c}\right\|_1\leq 3\left\|\tilde{a}_{A_*}-a_{*A_*}\right\|_1$ wpa1. So H-(3) can be invoked for vector $(\tilde{a}-a_*)$ and $\hat{\mathcal{G}}$. It follows wpa1

$$
\begin{aligned}
2(\tilde{a}-a_*)'\hat{\mathcal{G}}(\tilde{a}-a_*)+\lambda_2\left\|\tilde{a}-a_*\right\|_1&\leq 4\left[(\tilde{a}-a_*)'\hat{\mathcal{G}}(\tilde{a}-a_*)\right]^{1/2}\lambda_2\sqrt{\frac{S_*}{\underline{\kappa}_n}}\\
&\leq(\tilde{a}-a_*)'\hat{\mathcal{G}}(\tilde{a}-a_*)+\frac{4\lambda_2^2S_*}{\underline{\kappa}_n},
\end{aligned}
$$

since $4ab\leq a^2+4b^2$ for any number a and b. Rearrange above inequality

$$
(\tilde{a}-a_*)'\hat{\mathcal{G}}(\tilde{a}-a_*)+\lambda_2\left\|\tilde{a}-a_*\right\|_1\leq\frac{4\lambda_2^2S_*}{\underline{\kappa}_n},
$$

and it must be

$$
(\tilde{a}-a_*)'\hat{\mathcal{G}}(\tilde{a}-a_*)=O_p\left(\frac{\lambda_2^2S_*}{\underline{\kappa}_n}\right),\quad \left\|\tilde{a}-a_*\right\|_1=O_p\left(\frac{\lambda_2S_*}{\underline{\kappa}_n}\right).
$$

By H3-(2), $(\tilde{a}-a_*)'\hat{G}(\tilde{a}-a_*)=O_p\left(\frac{\lambda_2^2S_*}{\underline{\kappa}_n}\right)$ as well since $(\tilde{a}-a_*)'\hat{G}(\tilde{a}-a_*)\leq(\tilde{a}-a_*)'\hat{\mathcal{G}}(\tilde{a}-a_*)$ if $(\hat{\mathcal{G}}-\hat{G})$ is positive semidefinite. Conclusion follows.

C.4.5 Sufficient conditions

Lemma C.34. *Suppose O and H1 hold.*

(1) If there exists some number ρ_{1n} such that $|m(z,p_j(z))-p_j(x)\alpha_0(x)|\leq\rho_{1n}\Lambda_n$ for each $j=1\dots k$, then $\varepsilon_n^R=\rho_{1n}\Lambda_n\sqrt{\frac{\log k}{n}}$.

(2) Suppose (1) holds and in addition $\|\alpha_0\|_{\mathbb{P},\infty}<\infty$. Then $\varepsilon_n^R=\frac{\rho_{1n}\Lambda_n\log k}{n}+\sqrt{\frac{\log k}{n}}$;

(3) If there exists some number ρ_{2n} and sub-gaussian $\mathbf{f}(z)$ such that

$$
|m(z,p_j(x))-p_j(x)\alpha_0(x)|\leq\rho_{2n}\mathbf{f}(z)
$$

for each $j=1\dots k$, then $\varepsilon_n^R=\rho_{2n}\sqrt{\frac{\log k}{n}}$.

Proof. Let $e_j^R(z)=m(z,p_j(x))-\alpha_0(x)p_j(x)$. Note $\left\|\mathbb{E}_n[e^R]\right\|_\infty=\max_{1\leq j\leq k}\left|\mathbb{E}_n[e_j^R]\right|$

and by definition of α_0, $\mathbb{E}e_j^R = 0$ for each $j = 1 \dots k$.

Statement (1): By assumption, $\left|e_j^R\right| \leq \rho_{1n}\Lambda_n$. Then from Lemma 14.14 in Bühlmann and Van De Geer (2011) (i.e., Hoeffding moment inequality)

$$\mathbb{E}\max_{1\leq j\leq k}\left|\mathbb{E}_n[e_j^R]\right| \leq [2\log 2k]^{1/2}\max_{1\leq j\leq k}\left[\sum_{i=1}^n\left(\frac{\rho_{1n}\Lambda_n}{n}\right)^2\right]^{1/2} = O\left(\rho_{1n}\Lambda_n\sqrt{\frac{\log k}{n}}\right).$$

Conclusion follows from Markov inequality.

Statement (2): Let $(\sigma_{\max}^R)^2 = \max_{1\leq j\leq k}\mathbb{E}_n\left[\mathbb{E}\left[\left|e_j^R\right|^2\right]\right]$. Note by O, H1 and $\|\alpha_0\|_{\mathbb{P},\infty} < \infty$

$$\begin{aligned}(\sigma_{\max}^R)^2 = \max_{1\leq j\leq k}\mathbb{E}\left|e_j^R\right|^2 &\leq \max_{1\leq j\leq k}\mathbb{E}m^2(Z, p_j(X)) + \max_{1\leq j\leq k}\mathbb{E}[\alpha_0^2(X)p_j^2(X)] \\ &\leq C\max_{1\leq j\leq k}\mathbb{E}p_j^2(X) + \|\alpha_0\|_{\mathbb{P},\infty}\max_{1\leq j\leq k}\mathbb{E}p_j^2(X) < \infty.\end{aligned}$$

Denote $\tilde{e}_j^R = \frac{e_j^R}{\sigma_{\max}^R}$, it follows

$$\mathbb{E}_n\left[\mathbb{E}\left|\tilde{e}_j^R\right|^2\right] = \frac{\mathbb{E}_n\left[\mathbb{E}\left|e_j^R\right|^2\right]}{(\sigma_{\max}^R)^2} \leq \frac{\max_{1\leq j\leq k}\mathbb{E}_n\left[\mathbb{E}\left|e_j^R\right|^2\right]}{(\sigma_{\max}^R)^2} = 1.$$

And by assumption, for $t = 3, 4 \dots$

$$\mathbb{E}_n\left[\mathbb{E}\left|\tilde{e}_j^R\right|^t\right] \leq \left(\frac{\rho_{1n}\Lambda_n}{\sigma_{\max}^R}\right)^{t-2}\mathbb{E}_n\left[\mathbb{E}\left|\frac{e_j^R}{\sigma_{\max}^R}\right|^2\right] \leq \left(\frac{\rho_{1n}\Lambda_n}{\sigma_{\max}^R}\right)^{t-2}.$$

Thus Lemma 14.12 in Bühlmann and Van De Geer (2011) (i.e., a version of Bernstein's inequality) can be invoked

$$\mathbb{E}\left(\max_{1\leq j\leq k}\left|\mathbb{E}_n[e_j^R]\right|\right) = \sigma_{\max}^R\mathbb{E}\left[\max_{1\leq j\leq k}\left|\mathbb{E}_n[\tilde{e}_j^R]\right|\right] = O\left(\frac{\rho_{1n}\Lambda_n\log k}{n} + \sqrt{\frac{\log k}{n}}\right).$$

Then statement (2) follows by Markov inequality.

Statement (3): We need to show that for any $\delta > 0$, there exists some $C_* < \infty$ large enough such that $\mathbf{P}_n^R = \mathbb{P}\left\{\max_{1\leq j\leq k}\left|\mathbb{E}_n\left[e_j^R\right]\right| > C_*\varepsilon_n^R\right\} < \delta$.

Let $\|X\|_{\psi_2} = \inf\{t > 0 : \mathbb{E}\exp(X^2/t^2) \leq 2\}$ be the sub-gaussian norm of random variable X. Note $\|X\|_{\psi_2} < \infty$ if X is sub-gaussian. By a version of Hoeffding's inequality (for example, Theorem 2.6.3 in Vershynin, 2018), we have for any

C_* and some constant c_*

$$\mathbf{P}_n^R \leq \sum_{j=1}^{k} \mathbb{P}\left\{\left|\mathbb{E}_n\left[e_j^R\right]\right| > C_*\varepsilon_n^R\right\} = \sum_{j=1}^{k} \mathbb{P}\left\{\left|\sum_{i=1}^{n} \frac{1}{n} e_{j,i}^R\right| > C_*\varepsilon_n^R\right\}$$
$$\leq \sum_{j=1}^{k} 2\exp\left\{-\frac{nc_*C_*^2\left(\varepsilon_n^R\right)^2}{\left[\max_i \left\|e_{j,i}^R\right\|_{\psi_2}\right]^2}\right\} = \sum_{j=1}^{k} 2\exp\left[-\frac{nc_*C_*^2\left(\varepsilon_n^R\right)^2}{\left\|e_j^R\right\|_{\psi_2}^2}\right],$$

where the last relation follows by identical distribution assumption. Since also for some $\tilde{c} > 0$

$$\left\|e_j^R\right\|_{\psi_2} \leq \tilde{c}\left\|m(Z, p_j(X)) - p_j(X)\alpha_0(X)\right\|_{\psi_2} = \tilde{c}\left\|\left|m(Z, p_j(X)) - p_j(X)\alpha_0(X)\right|\right\|_{\psi_2}$$
$$\leq \tilde{c}\left\|\rho_{2n}\mathbf{f}(Z)\right\|_{\psi_2} \leq \tilde{c}\rho_{2n}\left\|\mathbf{f}(Z)\right\|_{\psi_2},$$

where the first relation follows from Lemma 2.6.8 in Vershynin (2018), the second equality is by definition of sub-gaussian norm, the third relation is by assumption and definition of sub gaussian norm, and the last inequality uses property of any norm. Hence

$$\mathbf{P}_n^R \leq 2k\exp\left[-\frac{cC_*^2}{\left\|\mathbf{f}(Z)\right\|_{\psi_2}^2\tilde{c}^2}\left(\frac{\sqrt{n}\varepsilon_n^R}{\rho_{2n}}\right)^2\right] = 2\exp\left[\log k\left(1 - \frac{cC_*^2}{\left\|\mathbf{f}(Z)\right\|_{\psi_2}^2\tilde{c}^2}\right)\right],$$

and for any $\delta > 0$, picking $C_* > \tilde{c}\left\|\mathbf{f}(Z)\right\|_{\psi_2}\sqrt{\frac{1}{c_*} - \frac{\log\delta - \log 2}{c_*\log k}}$ yields the desired conclusion. □

Lemma C.35. *Suppose O and H1 hold. Then* $\varepsilon_n^u = O_p\left(\sqrt{\frac{\log k}{n}}\Lambda_n\mu_*\right)$.

Proof. To bound $\left\|\mathbb{E}_n\left[p(X)u_* - \mathbb{E}[p(X)u_*]\right]\right\|_\infty$, use Lemma C.2. Let $f_j = \frac{p_j u_*}{n}$ for each $j = 1 \ldots k$. Then

$$\mathbb{E}\max_{1\leq j\leq k}\left|\sum_{i=1}^{n}\left\{f_j(X_i) - \mathbb{E}[f_j(X_i)]\right\}\right| \leq (8\log 2k)^{1/2}\,\mathbb{E}\left[\max_{1\leq j\leq k}\sum_{i=1}^{n} f_j(X_i)^2\right]^{1/2},$$

but

$$\mathbb{E}\left[\max_{1\leq j\leq k}\sum_{i=1}^{n} f_j(X_i)^2\right]^{1/2} = \mathbb{E}\left[\max_{1\leq j\leq k}\sum_{i=1}^{n}\frac{p_j^2(X_i)u_{*i}^2}{n^2}\right]^{1/2} \leq \mathbb{E}\left[\sum_{i=1}^{n}\frac{\Lambda_n^2 u_{*i}^2}{n^2}\right]^{1/2}$$
$$\leq \frac{\Lambda_n}{\sqrt{n}}\left(\mathbb{E}u_{*i}^2\right)^{1/2} = \frac{\Lambda_n}{\sqrt{n}}\mu_*.$$

Hence

$$\mathbb{E}\max_{1\le j\le k}\left|\sum_{i=1}^{n}\{f_j(X_i)-\mathbb{E}[f_j(X_i)]\}\right|\le(8\log 2k)^{1/2}\,\frac{\Lambda_n}{\sqrt{n}}\mu_*.$$

Conclusion follows by Markov inequality. □

Lemma C.36. *Suppose O and H1 hold.*

(1) If there exists some number ρ_{3n} such that $|m(z,p_j(x))|\le\rho_{3n}\Lambda_n$ for each $j=1\dots k$, then $\varepsilon_n^m=\frac{\rho_{3n}\Lambda_n\log k}{n}+\sqrt{\frac{\log k}{n}}$;

(2) If there exists some number ρ_{4n} and sub-gaussian $\mathbf{h}(z)$ such that $|m(z,p_j(x))|\le\rho_{4n}\mathbf{h}(z)$ for each $j=1\dots k$, then $\varepsilon_n^m=\rho_{4n}\sqrt{\frac{\log k}{n}}$.

Proof. The proofs are similar to those of statements (2) and (3) in Lemma C.34. Thus details are omitted. □

C.5 Proofs for basic lemmas and other related results

C.5.1 Proof of Lemma C.1

See Giné and Koltchinskii (2006).

C.5.2 Proof of Lemma C.2

Let $\eta_1\dots\eta_n$ be a series of independent Rademacher random variables independent of X. By symmetrization inequality (Lemma 2.3.1, Vaart and Wellner, 1996)

$$\mathbb{E}\max_{1\le j\le k}\left|\sum_{i=1}^{n}[f_j(X_i)-\mathbb{E}f_j(X_i)]\right|\le 2\mathbb{E}\max_{1\le j\le k}\left|\sum_{i=1}^{n}[f_j(X_i)\eta_i]\right|. \tag{C.44}$$

Notice $\mathbb{E}\left[f_j(X)\eta\right]=0$ for each $j=1\dots k$, and $|f_j(x)\eta_i|\le|f_j(x)|$ by definition of η_i, $i=1\dots n$. It follows by Lemma 14.14 of Bühlmann and Van De Geer (2011) (or Hoeffding's moment inequality)

$$\mathbb{E}\left(\max_{1\le j\le k}\left|\sum_{i=1}^{n}[f_j(X_i)\eta_i]\right|\right)\le\left[2\log 2k\right]^{1/2}\max_{1\le j\le k}\left[\sum_{i=1}^{n}f_j(X_i)^2\right]^{1/2}. \tag{C.45}$$

Combining (C.44) and (C.45) yields the result.

C.5.3 Proof of Lemma C.3

To save space, only prove results related to α_0. Those related to γ_0 can be shown in the same fashion. By definition

$$a_l = \arg\min_{a \in \mathbb{R}^k} \mathbb{E}[\alpha_0(X) - a'p(X)]^2. \tag{C.46}$$

Statement (1) then follows from first order condition

$$2\mathbb{E}\left[(\alpha_0(X) - a_l'p(X))p(X)\right] = \mathbf{0}.$$

Statement (2) directly follows from definition of least square projection in (C.46)

$$\mathbb{E}[u_{\alpha_0}^2] \leq \mathbb{E}[r_{\alpha_0}^2] \leq \|r_{\alpha_0}\|_{\mathbb{P},\infty}^2 = \mathbf{r}_{\alpha_0}^2.$$

To see statement (3), use ℓ_k to investigate relationship between $\|u_{\alpha_0}\|_{\mathbb{P},\infty}$ and $\|r_{\alpha_0}\|_{\mathbb{P},\infty}$. By standard decomposition and definition of r_{α_0}

$$u_{\alpha_0} = \alpha_0 - a_b'p + a_b'p - a_l'p = r_{\alpha_0} + a_b'p - a_l'p,$$

where

$$\begin{aligned} a_b'p - a_l'p &= p'\mathbb{E}[p(X)p(X)']^{-1}\mathbb{E}[p(X)p(X)']a_b - p'\mathbb{E}[p(X)p(X)']^{-1}\mathbb{E}[p(X)\alpha_0(X)] \\ &= p'\mathbb{E}[p(X)p(X)']^{-1}\mathbb{E}\left[p(X)(p(X)'a_b - \alpha_0(X))\right] \\ &= \mathcal{L}_n(p'a_b - \alpha_0) = \mathcal{L}_n r_{\alpha_0}. \end{aligned}$$

Then statement (3) follows from triangle inequality and definition of ℓ_k.

C.5.4 Proof of Lemma C.4

Apply definition of $\mathcal{L}_n\alpha_0$ and $\lambda_{\min}\left\{\mathbb{E}[p(X)p(X)']\right\}$

$$\|\mathcal{L}_n\alpha_0\|_{\mathbb{P},2}^2 = a_l'\mathbb{E}[p(X)p(X)']a_l \geq \|a_l\|^2 \lambda_{\min}\left\{\mathbb{E}[p(X)p(X)']\right\}.$$

By L1, $\mathbb{E}[p(X)p(X)']$ has all eigenvalues bounded away from zero. It follows

$$\|a_l\|^2 \leq \frac{\|\mathcal{L}_n\alpha_0\|_{\mathbb{P},2}^2}{\lambda_{\min}\left\{\mathbb{E}[p(X)p(X)']\right\}} \leq \frac{\|\alpha_0\|_{\mathbb{P},2}^2 + \|u_{\alpha_0}\|_{\mathbb{P},2}^2}{\lambda_{\min}\left\{\mathbb{E}[p(X)p(X)']\right\}} = O(1),$$

where the second inequality is by triangle inequality, and final relation follows from $\|\alpha_0\|_{\mathbb{P},2} = O(1)$ by O and $\|u_{\alpha_0}\|_{\mathbb{P},2} = O(1)$ by Lemma C.3-(2) and $\mathbf{r}_{\gamma_0} = O(1)$. Proof for the case of β_l is the same and thus omitted.

C.5.5 Proofs of Lemmas C.5 and C.6

Results follow from standard decomposition.

C.5.6 Proof of Lemma C.7

This follows from standard decomposition and linearity of $m(z,\cdot)$

$$\begin{aligned}
&\left\{\mathbb{E}_n\left[\tilde{\alpha}(X)f(X) - m(Z, f(X))\right]\right\}^2 \\
=&\left\{\mathbb{E}_n\left[\tilde{\alpha}(X)\pi_n f(X) - m(Z, \pi_n f(X)) + \tilde{\alpha}(X)r_n - m(Z, r_n)\right]\right\}^2 \\
\leq& 2\left\{\mathbb{E}_n\left[\tilde{\alpha}(X)\pi_n f(X) - m(Z, \pi_n f(X))\right]\right\}^2 + 2\left\{\mathbb{E}_n\left[\tilde{\alpha}(X)r_n - m(Z, r_n)\right]\right\}^2 = 2T_1 + 2T_2.
\end{aligned}$$

C.5.7 Proof of Lemma C.8

By standard decomposition and linearity of $m(z,\cdot)$

$$\begin{aligned}
\left\{\mathbb{E}_n\left[\tilde{\alpha}(X)r_n - m(Z, r_n)\right]\right\}^2 &= \left\{\frac{1}{n}\sum_{i=1}^n \left[(\tilde{\alpha}(X_i) - \alpha_0(X_i))r_{ni} + \alpha_0(X_i)r_{ni} - m(Z_i, r_{ni})\right]\right\}^2 \\
&\leq 2T_{21} + 2T_{22},
\end{aligned}$$

where $T_{21} = \left[\frac{1}{n}\sum_{i=1}^n \left(\tilde{\alpha}(X_i) - \alpha_0(X_i)\right) r_{ni}\right]^2$, and $T_{22} = \left\{\frac{1}{n}\sum_{i=1}^n \left[\alpha_0(X_i)r_{ni} - m(Z_i, r_{ni})\right]\right\}^2$.
Note by Markov inequality

$$T_{22} = O_p\left(\frac{\|r_n\|_{\mathbb{P},\infty}^2 \wedge \|\alpha_0\|_{\mathbb{P},\infty}^2 \|r_n\|_{\mathbb{P},2}^2}{n}\right),$$

since by iid assumption, property of α_0 ($\mathbb{E}[\alpha_0(X)r_n - m(Z, r_n)] = 0$) and O-(3)

$$\begin{aligned}
\mathbb{E}T_{22} &= \mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^n \left[\alpha_0(X_i)r_{ni} - m(Z_i, r_{ni})\right]\right\}^2 = \frac{1}{n}\mathbb{E}\left[\alpha_0(X)r_n - m(Z, r_n)\right]^2 \\
&\lesssim \frac{1}{n}\mathbb{E}\left[\alpha_0(X)r_n\right]^2 + \frac{1}{n}\mathbb{E}m^2(Z, r_n) \lesssim \frac{1}{n}\mathbb{E}\left[\alpha_0^2(X)r_n^2\right] + \frac{C}{n}\mathbb{E}r_n^2 \lesssim \frac{1}{n}\mathbb{E}\left[\alpha_0^2(X)r_n^2\right],
\end{aligned}$$

and $\mathbb{E}\left[\alpha_0^2(X)r_n^2\right]$ can be bounded by either

$$\mathbb{E}\left[\alpha_0^2(X)r_n^2\right] \leq \|r_n\|_{\mathbb{P},\infty}^2 \mathbb{E}\alpha_0^2(X) \lesssim \|r_n\|_{\mathbb{P},\infty}^2,$$

or

$$\mathbb{E}\left[\alpha_0^2(X)r_n^2\right] \leq \|\alpha_0\|_{\mathbb{P},\infty}^2 \mathbb{E}[r_n^2] = \|\alpha_0\|_{\mathbb{P},\infty}^2 \|r_n\|_{\mathbb{P},2}^2.$$

C.5.8 Proof of Lemma C.9

Statement (1)

By definition of $\tilde{\alpha}$ in (2.17)

$$\begin{aligned} & \sup_{g \in \mathcal{H}_{W_n}} \{\mathbb{E}_n[\tilde{\alpha}(X)g(X) - m(Z, g(X))]\}^2 + \mathcal{P}_n(\tilde{\alpha}(X)) \\ \leq & \sup_{g \in \mathcal{H}_{W_n}} \{\mathbb{E}_n[\alpha(X)g(X) - m(Z, g(X))]\}^2 + \mathcal{P}_n(\alpha(X)), \end{aligned}$$

for every $\alpha \in \Theta_n$. (C.2) then follows from definition of sup operator. Indeed, for every $f \in \mathcal{H}_{W_n}$

$$\{\mathbb{E}_n[\tilde{\alpha}(X)f(X) - m(Z, f(X))]\}^2 \leq \sup_{g \in \mathcal{H}_{W_n}} \{\mathbb{E}_n[\tilde{\alpha}(X)g(X) - m(Z, g(X))]\}^2.$$

And (C.3) follows from (C.2) and $\{\mathbb{E}_n[\tilde{\alpha}(X)f(X) - m(Z, f(X))]\}^2 \geq 0$.

Statement (2)

For every $f \in \Theta_n$, write $f(x) = \beta' p(x)$ for some $\beta \in \mathbb{R}^k$ (wlog). By Cauchy-Schwarz inequality

$$\{\mathbb{E}_n\left[\tilde{\alpha}(X)f(X) - m(Z, f(X))\right]\}^2 = \left(\beta' \mathbb{E}_n[e_{\tilde{\alpha}}(Z)]\right)^2 \leq \|\beta\|^2 \left\|\mathbb{E}_n[e_{\tilde{\alpha}}(Z)]\right\|^2, \tag{C.47}$$

where $e_{\tilde{\alpha}}(z) = m(z, p(x)) - \tilde{\alpha}(x)p(x)$. Next, note for every $\alpha \in \Theta_n$

$$\begin{aligned} \left\|\mathbb{E}_n[e_{\tilde{\alpha}}(Z)]\right\|^2 &= \left\|W_n \mathbb{E}_n[e_{\tilde{\alpha}}(Z)]\right\|^2 + \mathbb{E}_n[e_{\tilde{\alpha}}(Z)]'(I - W_n'W_n)\mathbb{E}_n[e_{\tilde{\alpha}}(Z)] \\ &\leq \left\|W_n \mathbb{E}_n[e_{\tilde{\alpha}}(Z)]\right\|^2 = \sup_{g \in \mathcal{H}_{W_n}} \{\mathbb{E}_n[\tilde{\alpha}(X)g(X) - m(Z, g(X))]\}^2 \\ &\leq \sup_{g \in \mathcal{H}_{W_n}} \{\mathbb{E}_n[\tilde{\alpha}(X)g(X) - m(Z, g(X))]\}^2 + \mathcal{P}_n(\tilde{\alpha}(X)) \\ &\leq \sup_{g \in \mathcal{H}_{W_n}} \{\mathbb{E}_n[\alpha(X)g(X) - m(Z, g(X))]\}^2 + \mathcal{P}_n(\alpha(X)), \end{aligned} \tag{C.48}$$

where the first equality follows by writing out $\|\cdot\|$, the second relation is from assumption that $W_n'W_n - I$ is positive semidefinite, the third relation uses Proposition 2.1, the fourth relation is because $\mathcal{P}_n(\tilde{\alpha}(X)) \geq 0$ by assumption, and final relation is due to definition of $\tilde{\alpha}$. On the other hand

$$\|\beta\|^2 \leq \frac{\|f\|_{\mathbb{P},2}^2}{\lambda_{\min}\{\mathbb{E}[p(X)p(X)']\}}, \tag{C.49}$$

since by L1

$$\|f\|_{\mathbb{P},2}^2 = \beta' \mathbb{E}[p(X)p(X)']\beta \geq \|\beta\|^2 \lambda_{\min}\left\{\mathbb{E}[p(X)p(X)']\right\}.$$

Combining (C.47), (C.48) and (C.49) yields the conclusion.

C.5.9 Proof of Lemma C.10

Statement (1)

$\|X_n\| = o_p(A_n)$ conditional on Y_n means for any $\delta > 0$ we have $\mathbb{P}\{\|X_n\| > \delta A_n | Y_n\} \to 0$ as $n \to \infty$. Then by dominated convergence theorem (for example, Theorem 25.12 of Billingsley, 2008)

$$\mathbb{P}\{\|X_n\| > \delta A_n\} \leq \mathbb{E}[P\{\|X_n\| > \delta A_n | Y_n\}] \to 0,$$

since $\mathbb{P}\{\|X_n\| > \delta A_n | Y_n\}$ is uniformly integrable.

Statement (2)

This follows from statement (1). See also Lemma 6.1 in Chernozhukov et al. (2018a) for more details.

C.5.10 Proof of Lemma C.11

By O

$$\mathbb{E}[\mathcal{A}_i(\mathscr{Z}_n)e_i|Z_1, \cdots Z_n] = \mathcal{A}_i(\mathscr{Z}_n)\mathbb{E}[e_i|Z_1, \cdots Z_n] = \mathcal{A}_i(\mathscr{Z}_n)\mathbb{E}[e_i|Z_i] = 0.$$

Hence $\mathbb{E}[\mathcal{A}_i(\mathscr{Z}_n)e_i] = 0$ for each $i = 1 \ldots n$ by LIE. Since also $\mathbb{E}[e^2|Z] < \infty$ almost surely,

$$\begin{aligned}
var\left[\frac{1}{n}\sum_{i=1}^n(\mathcal{A}_i(\mathscr{Z}_n)e_i)\right] &= \mathbb{E}\left[var\left(\frac{1}{n}\sum_{i=1}^n(\mathcal{A}_i(\mathscr{Z}_n)e_i)|Z_1, \cdots Z_n\right)\right] \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[\mathcal{A}_i^2(\mathscr{Z}_n)var(e_i|Z_1, \cdots Z_n)\right] \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[\mathcal{A}_i^2(\mathscr{Z}_n)var(e_i|Z_i)\right] \\
&\lesssim \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\mathcal{A}_i^2(\mathscr{Z}_n) = \frac{1}{n}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \mathcal{A}_i^2(\mathscr{Z}_n)\right].
\end{aligned}$$

Conclusion follows by Markov inequality.

C.5.11 Proof of Lemma C.12

Note $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\mathcal{A}_i(\mathscr{Z}_n)(Y_i-\gamma_0(X_i))] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\mathcal{A}_i(\mathscr{Z}_n)e_i]$. The proof takes four steps.

Step 1

Conditional on $Z_1, \cdots Z_n$,

$$\mathbb{E}[\mathcal{A}_i(\mathscr{Z}_n)e_i|Z_1,\cdots Z_n] = \mathcal{A}_i(\mathscr{Z}_n)\mathbb{E}[e_i|Z_i] = 0.$$

Step 2

Conditional on $Z_1, \cdots Z_n$, similar to proof of Lemma C.11

$$\begin{aligned} &var\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\mathcal{A}_i(\mathscr{Z}_n)e_i]|Z_1,\cdots Z_n\right) = \frac{1}{n}\sum_{i=1}^{n}\left[\mathcal{A}_i^2(\mathscr{Z}_n)var(e_i|Z_1,\cdots Z_n)\right] \\ =&\frac{1}{n}\sum_{i=1}^{n}\left\{\mathcal{A}_i^2(\mathscr{Z}_n)\mathbb{E}[e_i^2|Z_1,\cdots Z_n]\right\} = \frac{1}{n}\sum_{i=1}^{n}\left\{\mathcal{A}_i^2(\mathscr{Z}_n)\mathbb{E}[e_i^2|Z_i]\right\}. \end{aligned}$$

Let $\mathcal{V}_n = var\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\mathcal{A}_i^2(\mathscr{Z}_n)e_i]|Z_1\ldots Z_n\right)$, and $\mathcal{U}_i = n^{-1/2}\mathcal{V}_n^{-1/2}\mathcal{A}_i(\mathscr{Z}_n)e_i$. Then

$$\sum_{i=1}^{n} var\left(\mathcal{U}_i|Z_1,\cdots Z_n\right) = 1.$$

Moreover, by assumption, $\mathbb{E}[e^2|Z]$ is bounded away from zero almost surely and $\left[\frac{1}{n}\sum_{i=1}^{n}\mathcal{A}_i^2(\mathscr{Z}_n)\right]^{-1} = O_p(1)$. So conditional on $Z_1, \cdots Z_n$, $\mathcal{V}_n^{-1} = O(1)$.

Step 3

Show that conditional on $Z_1, \cdots Z_n$, $\sum_{i=1}^{n}\mathcal{U}_i \xrightarrow{d} N(0,1)$. Note conditional on $Z_1, \cdots Z_n$, $\{\mathcal{U}_i\}_{i=1}^n$ are mean zero and independently distributed. Invoke a version of Berry-Esseen inequality (for example, Theorem 3.6 in Chen et al., 2010)

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\sum_{i=1}^{n}\mathcal{U}_i \le t|Z_1,\cdots Z_n\right) - \Phi(t)\right| \le \min\left\{\sum_{i=1}^{n}\left[\mathbb{E}\left|\mathcal{U}_i\right|^3|Z_1,\cdots Z_n\right], 1\right\}.$$

Hence it suffices to show $\sum_{i=1}^{n}\left[\mathbb{E}\left|\mathcal{U}_i\right|^3|Z_1,\cdots Z_n\right] = o_p(1)$. By assumption, $\frac{\max|\mathcal{A}_i(\mathscr{Z}_n)|}{\sqrt{n}} = o(1)$ and $\frac{1}{n}\sum_{i=1}^{n}\mathcal{A}_i^2(\mathscr{Z}_n) = O(1)$ conditional on $Z_1, \cdots Z_n$. It follows

$$\begin{aligned}\sum_{i=1}^{n}\left[\mathbb{E}\left|\mathcal{U}_i\right|^3|Z_1,\cdots Z_n\right] &= \sum_{i=1}^{n} n^{-3/2}\mathcal{V}_n^{-3/2}\left|\mathcal{A}_i(\mathscr{Z}_n)\right|^3\mathbb{E}\left[\left|e_i\right|^3|Z_1,\cdots Z_n\right]\\ &= \frac{1}{n^{3/2}}\mathcal{V}_n^{-3/2}\sum_{i=1}^{n}\left|\mathcal{A}_i(\mathscr{Z}_n)\right|^3\mathbb{E}\left[\left|e_i\right|^3|Z_i\right]\\ &\lesssim \frac{1}{n^{3/2}}\sum_{i=1}^{n}\left|\mathcal{A}_i(\mathscr{Z}_n)\right|^3\mathbb{E}\left[\left|e_i\right|^3|Z_i\right] \text{ (since } \mathcal{V}_n^{-1}=O(1))\\ &\lesssim \frac{1}{n^{3/2}}\sum_{i=1}^{n}\left|\mathcal{A}_i(\mathscr{Z}_n)\right|^3 \text{ (since } \mathbb{E}\left[\left|e_i\right|^3|Z_i\right] \text{ bounded from above a.s.)}\\ &\leq \frac{\max\left|\mathcal{A}_i(\mathscr{Z}_n)\right|}{\sqrt{n}}\frac{1}{n}\sum_{i=1}^{n}\mathcal{A}_i^2(\mathscr{Z}_n)\leq o(1)O(1)=o(1).\end{aligned}$$

Step 4

Conclude unconditionally $\sum_{i=1}^{n}\mathcal{U}_i \xrightarrow{d} N(0,1)$, or $n^{-1/2}\mathcal{V}_n^{-1/2}\sum_{i=1}^{n}[\mathcal{A}_i(\mathscr{Z}_n)e_i] \xrightarrow{d} N(0,1)$ by dominated convergence theorem (for example, Theorem 25.12 of Billingsley, 2008).

C.5.12 Proof of Lemma C.28

Note under stated assumptions, both $\hat{G}$ and $\mathcal{W}_n$ are invertible wpa1. Hence, when $\lambda_1 = 0$, $\tilde{\alpha} = p'(\hat{G}\mathcal{W}_n\hat{G})^{-}\hat{G}\mathcal{W}_n\hat{P} = p'\hat{a}$ wpa1, where $\hat{a} = \hat{G}^{-1}\hat{P}$. It suffices to treat $\tilde{\alpha} = \hat{a}'p$. By triangle inequality

$$\frac{\max_i|\tilde{\alpha}(X_i)|}{\sqrt{n}} \lesssim \frac{\max_i|\tilde{\alpha}(X_i)-\mathcal{L}_n\alpha_0(X_i)|}{\sqrt{n}} + \frac{\max_i|\mathcal{L}_n\alpha_0(X_i)|}{\sqrt{n}}.$$

First, note by L1 and Lemma C.4

$$\mathbb{E}\left[\left(\mathcal{L}_n\alpha_0\right)^2(X)\right] = a_l'Ga_l \leq \|a_l\|^2\lambda_{\max}(G) < \infty.$$

It follows then $\frac{\max_i|\alpha_l(X_i)|}{\sqrt{n}} = o_p(1)$ by Lemma 11.2 in Owen (2001b).

Next bound $\frac{\max_i|\tilde{\alpha}(X_i)-\alpha_l(X_i)|}{\sqrt{n}}$. By Lemma C.15

$$\hat{a} - a_l = \hat{G}^{-1}\mathbb{E}_n e^R + \hat{G}^{-1}\mathbb{E}_n[u_{\alpha_0}p(X)].$$

Thus standard decomposition yields

$$\tilde{\alpha}(X_i) - \mathcal{L}_n\alpha_0(X_i) = p'(X_i)\left\{\hat{G}^{-1}\mathbb{E}_n e^R + \hat{G}^{-1}\mathbb{E}_n[u_{\alpha_0}p(X)]\right\} = p'(X_i)\hat{G}^{-1}\mathbb{E}_n e^R + \tilde{\mathcal{L}}_n u_{\alpha_0 i},$$

where $\tilde{\mathcal{L}}_n u_{\alpha_0} = p'\hat{G}^{-1}\mathbb{E}_n[u_{\alpha_0}p(X)]$ is the empirical projection of u_{α_0} onto Θ_n. Thus

$$\max_i |\tilde{\alpha}(X_i) - \alpha_l(X_i)| \leq \max_i \left|p'(X_i)\hat{G}^{-1}\mathbb{E}_n e^R\right| + \max_i \left|\tilde{\mathcal{L}}_n u_{\alpha_0 i}\right|.$$

Note by definition of ℓ_k and Lemma C.3

$$\max_i \left|\tilde{\mathcal{L}}_n u_{\alpha_0 i}\right| \leq \left\|\tilde{\mathcal{L}}_n u_{\alpha_0}\right\|_{\mathbb{P},\infty} \lesssim \ell_k \left\|u_{\alpha_0}\right\|_{\mathbb{P},\infty} \lesssim \ell_k^2 \mathbf{r}_{\alpha_0}.$$

Hence $\frac{\max_i \left|\tilde{\mathcal{L}}_n(u_{\alpha_0 i})\right|}{\sqrt{n}} = o_p(1)$ under stated assumptions. As a final step we show

$$\frac{\max_i \left|p'(X_i)\hat{G}^{-1}\mathbb{E}_n e^R\right|}{\sqrt{n}} = o_p(1).$$

The proof is long and split into 5 steps.

Step 1

By assumption (1) in Lemma C.28 and Lemma C.24, $\left\|\hat{G}^{-1}\right\| = O_p(1)$. Hence

$$\begin{aligned}
&\max_i \frac{\left|p'(X_i)\hat{G}^{-1}\mathbb{E}_n e^R\right|}{\sqrt{n}} = \max_i \frac{\left\|p'(X_i)\hat{G}^{-1}\right\| \left|\tilde{p}(X_i)'\mathbb{E}_n e^R\right|}{\sqrt{n}} \\
\leq &\max_i \left\|p'(X_i)\hat{G}^{-1}\right\| \max_i \frac{\left|\tilde{p}(X_i)'\mathbb{E}_n e^R\right|}{\sqrt{n}} \leq \frac{\left(\max_i \|p'(X_i)\|\right)\left\|\hat{G}^{-1}\right\|}{\sqrt{n}} \max_i \left|\tilde{p}(X_i)'\mathbb{E}_n e^R\right| \\
\leq &\frac{\xi_k}{\sqrt{n}}\left\|\hat{G}^{-1}\right\| \max_i \left|\tilde{p}(X_i)'\mathbb{E}_n e^R\right| \lesssim O_p(1)\max_i \left|\tilde{p}(X_i)'\mathbb{E}_n e^R\right|.
\end{aligned}$$

It suffices to show $\max_i \left|\tilde{p}(X_i)'\mathbb{E}_n e^R\right| = o_p(1)$.

Step 2

Note

$$\max_i \left|\tilde{p}(X_i)'\mathbb{E}_n e^R\right| \leq \sup_{x\in\mathcal{X}} \left[\tilde{p}(x)'\mathbb{E}_n e^R\right] = \sup_{x\in\mathcal{X}} \mathbb{E}_n \left[\tilde{p}(x)' e^R\right] = \sup_{x\in\mathcal{X}} \mathbb{E}_n \left[\tilde{p}(x)' e^R - \mathbb{E}\tilde{p}(x)' e^R\right],$$

where the last equality follows from $\sup_{x\in\mathcal{X}}\mathbb{E}\tilde{p}(x)'e^R = 0$. To see this, for any $\tilde{p}(x)$

$$\begin{aligned}\mathbb{E}\tilde{p}(x)'e^R &= \mathbb{E}\tilde{p}(x)'[m(Z,p(X)) - \alpha_0(X)p(X)] = \mathbb{E}\sum_{j=1}^{k}\tilde{p}_j(x)[m(Z,p_j(X)) - \alpha_0(X)p_j(X)] \\ &= \mathbb{E}\sum_{j=1}^{k}[m(Z,\tilde{p}_j(x)p_j(X)) - \alpha_0(X)\tilde{p}_j(x)p_j(X)], \\ &= \sum_{j=1}^{k}\mathbb{E}[m(Z,\tilde{p}_j(x)p_j(X)) - \alpha_0(X)\tilde{p}_j(x)p_j(X)] = 0,\end{aligned}$$

since for any $j = 1 \dots k$

$$\mathbb{E}[\tilde{p}_j(x)p_j(X)]^2 = \mathbb{E}[\tilde{p}_j^2(x)p_j^2(X)] \leq \mathbb{E}[\|\tilde{p}(x)\|^2 p_j^2(X)] \leq \mathbb{E}[p_j^2(X)] < \infty.$$

It follows by definition of α_0

$$\mathbb{E}[m(Z,\tilde{p}_j(x)p_j(X)) - \alpha_0(X)\tilde{p}_j(x)p_j(X)] = 0, \forall j = 1 \dots k.$$

Step 3

Bound $\sup_{x\in\mathcal{X}}\left|\mathbb{E}_n\left[\tilde{p}(x)'e^R - \mathbb{E}\tilde{p}(x)'e^R\right]\right|$. Conditional on data, let

$$D := \left\{\mathbf{d} = (\mathbf{d}_1, \mathbf{d}_2, \dots \mathbf{d}_n) \in \mathbb{R}^n : d_i = \tilde{p}(x)'e_i^R, x \in \mathcal{X}\right\}.$$

Then define $\|\cdot\|_{n,2}$ on $\mathbb{R}^n$ as $\|\mathbf{d}\|_{n,2}^2 = \sum_{i=1}^{n}\mathbf{d}_i^2$. By symmetrization inequality (Lemma 2.3.1, Vaart and Wellner, 1996), it follows

$$\begin{aligned}&\mathbb{E}\sup_{x\in\mathcal{X}}\left|\mathbb{E}_n\left[\tilde{p}(x)'e^R - \mathbb{E}\tilde{p}(x)'e^R\right]\right| \\ =&\mathbb{E}\left[\mathbb{E}\left(\sup_{x\in\mathcal{X}}\left|\mathbb{E}_n\left[\tilde{p}(x)'e^R\right]\right| | X_1 \dots X_n\right)\right] \\ =&2\mathbb{E}\left[\mathbb{E}\left(\mathbb{E}_\eta\sup_{x\in\mathcal{X}}\left|\mathbb{E}_n\left[\tilde{p}(x)'e^R\eta\right]\right| | X_1 \dots X_n\right)\right],\end{aligned}$$

where $\eta_1 \dots \eta_n$ are independent Rademacher random variables. Therefore it suffices to bound $\mathbb{E}_\eta\sup_{x\in\mathcal{X}}\left|\mathbb{E}_n[\tilde{p}(x)'e^R\eta]\right|$.

Step 4

By Dudley's inequality (Dudley, 1967, also see proof of Lemma 4.2 in Belloni et al., 2015)

$$\mathbb{E}_\eta \sup_{x\in\mathcal{X}} \left|\mathbb{E}_n\left[\tilde{p}(x)'e^R\eta\right]\right| \lesssim \frac{1}{\sqrt{n}}\int_0^{\mathcal{D}} \sqrt{\log N(D, \|\cdot\|_{n,2}, \delta)}d\delta,$$

where $\mathcal{D} = 2\sup_{\mathbf{d}} \|\mathbf{d}\|_{n,2} = 2\left\{\mathbb{E}_n\left[\tilde{p}(x)'e^R\right]^2\right\}^{1/2} = 2\left\{\mathbb{E}_n\left[\tilde{p}(x)'e^R\right]^2\right\}^{1/2} \leq 2\left[\left\|\mathbb{E}_n e^R e^{R\prime}\right\|\right]^{1/2}$.

For any $x_1, x_2 \in \mathcal{X}$, wpa1

$$\begin{aligned}
&\left\{\mathbb{E}_n\left[\tilde{p}(x_1)'e^R - \tilde{p}(x_2)'e^R\right]^2\right\}^{1/2} \\
&= \left\{\mathbb{E}_n\left[\left(\tilde{p}(x_1) - \tilde{p}(x_2)\right)'e^R\right]^2\right\}^{1/2} \leq \left\{[\tilde{p}(x_1) - \tilde{p}(x_2)]'\mathbb{E}_n\left[e^R e^{R\prime}\right][\tilde{p}(x_1) - \tilde{p}(x_2)]\right\}^{1/2} \\
&\leq \|\tilde{p}(x_1) - \tilde{p}(x_2)\| \left\|\mathbb{E}_n[e^R e^{R\prime}]\right\|^{1/2} \leq \xi_k^L \|x_1 - x_2\| \left\|\mathbb{E}_n\left[e^R e^{R\prime}\right]\right\|^{1/2},
\end{aligned}$$

where the last inequality follows from definition of ξ_k^L. It follows for some universal constant C_D

$$N(D, \|\cdot\|_{n,2}, \delta) \leq \left(\frac{C_D \left\|\mathbb{E}_n[e^R e^{R\prime}]\right\|^{1/2} \xi_k^L}{\delta}\right)^{d_X}$$

and

$$\int_0^{\mathcal{D}} \sqrt{d_X \log\left(\frac{C\left\|\mathbb{E}_n[e^R e^{R\prime}]\right\|^{1/2}\xi_k^L}{\delta}\right)}d\delta \leq \left\|\mathbb{E}_n\left[e^R e^{R\prime}\right]\right\|^{1/2} \int_0^2 \sqrt{d_X \log\left(\frac{C_D\xi_k^L}{\delta}\right)}d\delta.$$

Step 5

Bound $\left\|\mathbb{E}_n[e^R e^{R\prime}]\right\|^{1/2}$. Consider two cases.

Case 1 If Assumption (4)-(a) of Lemma C.28 holds

$$\left\|\mathbb{E}_n[e^R e^{R\prime}]\right\|^{1/2} = \left\|\mathbb{E}_n[\Delta^2 p(X)p(X)']\right\|^{1/2} \leq \max_i |\Delta_i| \hat{G}^{1/2}.$$

Note $\hat{G}^{1/2} = O_p(1)$. So wpa1

$$\mathbb{E}\left\{\mathbb{E}_n\left[e^R e^{R\prime}\right]|X\right\} \lesssim n^{1/m}.$$

Hence

$$\mathbb{E}\left[\sup_{x\in\mathcal{X}}\left|\mathbb{E}_n\left[\tilde{p}(x)'e^R - \mathbb{E}\tilde{p}(x)'e^R\right]\right| | X_1 \ldots X_n\right] \leq \frac{n^{1/m}\sqrt{\log \xi_k^L}}{n^{1/2}} \to 0.$$

Conclusion follows by Markov inequality.

Case *2* If Assumption (4)-(b) of Lemma C.28 holds, it follows from Jensen's inequality and matrix Chernoff bounds (Theorem 5.1, Tropp, 2015) that

$$\mathbb{E}\left[\left\|\mathbb{E}_n[e^R e^{R\prime}]\right\|^{1/2}\right] \leq \left(\mathbb{E}\left\|\mathbb{E}_n[e^R e^{R\prime}]\right\|\right)^{1/2} \lesssim \sqrt{\frac{\xi_k^2 n^t}{n}\log k} \to \sqrt{c_1} n^{t/2}.$$

Hence

$$\mathbb{E}\left[\sup_{x\in\mathcal{X}}\left|\mathbb{E}_n\left[\tilde{p}(x)'e^R - \mathbb{E}\tilde{p}(x)'e^R\right]\right|\right] \lesssim \mathbb{E}\left\|\mathbb{E}_n\left[e^R e^{R\prime}\right]\right\|^{1/2}\sqrt{\log \xi_k^L} \lesssim \frac{n^{t/2}}{\sqrt{n}}\sqrt{\log \xi_k^L} \to 0,$$

and conclusion follows from Markov inequality.

Bibliography

Ai, C. and Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843.

Altonji, J. G., Ichimura, H., and Otsu, T. (2012). Estimating derivatives in nonseparable models with limited dependent variables. *Econometrica*, 80(4):1701–1719.

Andrews, D. W. (1994). Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica: Journal of the Econometric Society*, pages 43–72.

Angrist, J. D. (1998). Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica*, pages 249–288.

Angrist, J. D. and Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of economic perspectives*, 24(2):3–30.

Angrist, J. D. and Rokkanen, M. (2015). Wanna get away? regression discontinuity estimation of exam school effects away from the cutoff. *Journal of the American Statistical Association*, 110(512):1331–1344.

Armstrong, T. and Kolesár, M. (2018a). Finite-sample optimal estimation and inference on average treatment effects under unconfoundedness.

Armstrong, T. B. and Kolesár, M. (2018b). Optimal inference in a class of regression models. *Econometrica*, 86(2):655–683.

Athey, S., Imbens, G. W., and Wager, S. (2018). Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):597–623.

Backus, D., Chernov, M., and Zin, S. (2014). Sources of entropy in representative agent models. *Journal of Finance*, 69(1):51–99.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.

Belloni, A., Chernozhukov, V., Chetverikov, D., and Kato, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345–366.

Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2017a). Program Evaluation and Causal Inference with High-Dimensional Data. *Econometrica*, 85(1):233–298.

Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2017b). Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650.

Belloni, A., Chernozhukov, V., and Wang, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806.

Besley, T. (2006). *Principled agents?: The political economy of good government*. Oxford University Press on Demand.

Besley, T. and Case, A. (1995). Does electoral accountability affect economic policy choices? evidence from gubernatorial term limits. *The Quarterly Journal of Economics*, 110(3):769–798.

Bickel, P. J. (1982). On Adaptive Estimation. *The Annals of Statistics*, 10(3):647–671.

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., and Ritov, Y. (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press Baltimore.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics*, 37(4):1705–1732.

Billingsley, P. (2008). *Probability and measure*. John Wiley & Sons.

Borwein, J. M. and Lewis, A. S. (1991). Duality Relationships for Entropy-Like Minimization Problems. *SIAM Journal on Control and Optimization*, 29(2):325–338.

Boyd, S., Boyd, S. P., and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Bradley, R. C. et al. (1985). On the central limit question under absolute regularity. *The Annals of Probability*, 13(4):1314–1325.

Bradley, R. C. et al. (2005). Basic properties of strong mixing conditions. a survey and some open questions. *Probability surveys*, 2:107–144.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Bunea, F. (2008). Honest variable selection in linear and logistic regression models via l1 and l1+l2 penalization. *Electronic Journal of Statistics*, 2:1153–1194.

Card, D., Heining, J., and Kline, P. (2013). Workplace heterogeneity and the rise of west german wage inequality. *The Quarterly journal of economics*, 128(3):967–1015.

Cattaneo, M. D., Crump, R. K., and Jansson, M. (2013). Generalized jackknife estimators of weighted average derivatives. *Journal of the American Statistical Association*, 108(504):1243–1256.

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

Cattaneo, M. D., Jansson, M., and Newey, W. K. (2018b). Inference in linear regression models with many covariates and heteroscedasticity. *Journal of the American Statistical Association*, 113(523):1350–1361.

Chan, K. C. G., Yam, S. C. P., and Zhang, Z. (2016). Globally efficient nonparametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):673–700.

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

Chen, L. H., Goldstein, L., and Shao, Q.-M. (2010). *Normal approximation by Stein's method*. Springer Science & Business Media.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics*, 6:5549–5632.

Chen, X. and Christensen, T. M. (2015). Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics*, 188(2):447–465.

Chen, X., Hong, H., and Tamer, E. (2005). Measurement error models with auxiliary data. *The Review of Economic Studies*, 72(2):343–366.

Chen, X. and Pouzo, D. (2012). Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, 80(1):277–321.

Chen, X. and Pouzo, D. (2015a). Sieve wald and qlr inferences on semi/nonparametric conditional moment models. *Econometrica*, 83(3):1013–1079.

Chen, X. and Pouzo, D. (2015b). Sieve wald and QLR inferences on semi/nonparametric conditional moment models. *Econometrica*, 83(3):1013–

1079.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018a). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., and Newey, W. K. (2016). Locally Robust Semiparametric Estimation. pages 1–42.

Chernozhukov, V., Newey, W., and Robins, J. (2018b). Double/De-Biased Machine Learning Using Regularized Riesz Representers. pages 1–15.

Chernozhukov, V., Newey, W. K., and Singh, R. (2018c). Learning Continuous Regression Functionals via Regularized Riesz Representers.

Christensen, T. M. (2016). Nonparametric stochastic discount factor decomposition. *Working Paper*, pages 1–53.

Cochrane, J. H. (2009). *Asset Pricing:(Revised Edition)*. Princeton university press.

Csiszar, I. (1975). I divergence geometry of probability distributoins and minimization problems. *Annals of Probability*, 3(1):146–158.

DeVore, R. A. and Lorentz, G. G. (1993). *Constructive approximation*, volume 303. Springer Science & Business Media.

Donald, S. and Newey, W. (1994). Series Estimation of Semilinear Models.

Dudley, R. M. (1967). The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330.

Duflo, E., Glennerster, R., and Kremer, M. (2007). Chapter 61 using randomization in development economics research: a toolkit. *Handbook of development economics*, 4(07):3895–3962.

Dümbgen, L., Van De Geer, S. A., Veraar, M. C., and Wellner, J. A. (2010). Nemirovski's inequalities revisited. *The American Mathematical Monthly*, 117(2):138–160.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of*.

Fama, E. F. and MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of political economy*, 81(3):607–636.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23.

Ferraz, C. and Finan, F. (2011). Electoral accountability and corruption: Evidence from the audits of local governments. *American Economic Review*,

101(4):1274–1311.

Ghosh, A., Julliard, C., and Taylor, A. P. (2015). An information-based one-factor asset pricing model.

Ghosh, A., Julliard, C., and Taylor, A. P. (2016). What is the consumption-CAPM missing? An information-theoretic framework for the analysis of asset pricing models. *The Review of Financial Studies*.

Giné, E. and Koltchinskii, V. (2006). Concentration inequalities and asymptotic results for ratio type empirical processes. *Annals of Probability*, 34(3):1143–1216.

Hahn, J. (1998). On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. *Econometrica*, 66(2):315–331.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46.

Hall, P., Lukacs, E., Birnbaum, Z., and Heyde, C. (1980). Martingale limit theory and its application.

Hansen, B. E. (2015). A unified asymptotic distribution theory for parametric and non-parametric least squares. Technical report, Working paper.

Hansen, L. P. (2014). Nobel lecture: Uncertainty outside and inside economic models. *Journal of Political Economy*, 122(5):945–987.

Härdle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American statistical Association*, 84(408):986–995.

Harville, D. A. (1998). Matrix algebra from a statistician's perspective.

Hausman, J. A. and Newey, W. K. (1995). Nonparametric estimation of exact consumers surplus and deadweight loss. *Econometrica: Journal of the Econometric Society*, pages 1445–1476.

Hausman, J. A. and Newey, W. K. (2016). Individual heterogeneity and average welfare. *Econometrica*, 84(3):1225–1248.

Hausman, J. A. and Newey, W. K. (2017). Nonparametric welfare analysis. *Annual Review of Economics*, 9:521–546.

Hebiri, M., Van De Geer, S., et al. (2011). The smooth-lasso and other l1 + l2-penalized methods. *Electronic Journal of Statistics*, 5:1184–1226.

Heckman, J., Ichimura, H., Smith, J., and Todd, P. (1998). Characterizing selection bias using experimental data. Technical report, National bureau of economic research.

Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The*

review of economic studies, 64(4):605–654.

Heckman, J. J., LaLonde, R. J., and Smith, J. A. (1999). The economics and econometrics of active labor market programs. In *Handbook of labor economics*, volume 3, pages 1865–2097. Elsevier.

Hirshberg, D. A. and Wager, S. (2018). Augmented Minimax Linear Estimation. pages 1–49.

Hjort, N. L., McKeague, I. W., and Van Keilegom, I. (2009). Extending the scope of empirical likelihood. *The Annals of Statistics*, pages 1079–1111.

Huang, J. Z. (2003). Local asymptotics for polynomial spline regression. *Annals of Statistics*, 31(5):1600–1635.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263.

Imbens, G. and Wager, S. (2018). Optimized regression discontinuity designs. *Review of Economics and Statistics*, (0).

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29.

Imbens, G. W. and Newey, W. K. (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, 77(5):1481–1512.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Imbens, G. W., Spady, R. H., and Johnson, P. (1998). Information theoretic approaches to inference in moment condition models. *Econometrica*, pages 333–357.

Imbens, G. W. and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86.

Jing, B.-Y., Shao, Q.-M., Wang, Q., et al. (2003). Self-normalized cramér-type large deviations for independent random variables. *The Annals of probability*, 31(4):2167–2215.

Kallus, N. (2016). Generalized optimal matching methods for causal inference. *arXiv preprint arXiv:1612.08321*.

Kang, J. D., Schafer, J. L., et al. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539.

Kitamura, Y. (2006). Empirical likelihood methods in econometrics : theory and practice. *Cowles foundation discussion paper No. 1569*.

Kitamura, Y. and Stutzer, M. (1997). An information-theoretic alternative to

generalized method of moments estimation. *Econometrica*, 65(4):861–874.

Kitamura, Y. and Stutzer, M. (2002). Connections between entropic and linear projections in asset pricing estimation. *Journal of Econometrics*, 107(1-2):159–174.

Koltchinskii, V. (2009). Sparsity in penalized empirical risk minimization. In *Annales de l'IHP Probabilités et statistiques*, volume 45, pages 7–57.

Koltchinskii, V. et al. (2009). The dantzig selector and sparsity oracle inequalities. *Bernoulli*, 15(3):799–828.

Lahiri, S. and Mukhopadhyay, S. (2012). A penalized empirical likelihood method in high dimensions. *The Annals of Statistics*, 40(5):2511—-2540.

Lee, L.-f. and Sepanski, J. H. (1995). Estimation of linear and nonlinear errors-in-variables models using validation data. *Journal of the American Statistical Association*, 90(429):130–140.

Lewellen, J., Nagel, S., and Shanken, J. (2010). A skeptical appraisal of asset pricing tests. *Journal of Financial Economics*, 96(2):175–194.

Liese, F. and Vajda, I. (1987). *Convex statistical distances*, volume 95. Teubner.

Linnemayr, S. and Alderman, H. (2011). Almost random: Evaluating a large-scale randomized nutrition program in the presence of crossover. *Journal of Development Economics*, 96(1):106–114.

List, J. A. and Sturm, D. M. (2006). How elections matter: Theory and evidence from environmental policy. *The Quarterly Journal of Economics*, 121(4):1249–1281.

Lorentz, G. (1966). Approximation of functions (athena series).

Nakamura, E. and Steinsson, J. (2018). Identification in macroeconomics. *Journal of Economic Perspectives*, 32(3):59–86.

Newey, W. (1994a). The Asymptotic Variance of Semiparametric Estimators. *Econometrica*, 62(6):1349–1382.

Newey, W. K. (1988). Adaptive estimation of regression models via moment restrictions. *Journal of Econometrics*, 38(3):301–339.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of applied econometrics*, 5(2):99–135.

Newey, W. K. (1991). Uniform convergence in probability and stochastic equicontinuity. *Econometrica: Journal of the Econometric Society*, pages 1161–1167.

Newey, W. K. (1994b). The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, pages 1349–1382.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79(1):147–168.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis

testing. *Handbook of econometrics*, 4:2111–2245.

Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578.

Newey, W. K. and Robins, J. R. (2018). Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*.

Newey, W. K. and Smith, R. J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255.

Newey, W. K. and Stoker, T. M. (1993). Efficiency of weighted average derivative estimators and index models. *Econometrica: Journal of the Econometric Society*, pages 1199–1223.

Newey, W. K. and West, K. D. (1987). Hypothesis testing with efficient method of moments estimation. *International Economic Review*, pages 777–787.

Oster, E. (2017). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, pages 1–18.

Owen, A. B. (2001a). *Empirical likelihood*. CRC press.

Owen, A. B. (2001b). *Empirical likelihood*. Chapman and Hall/CRC.

Powell, J. L., Stock, J. H., and Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica: Journal of the Econometric Society*, pages 1403–1430.

Qiu, C. and Otsu, T. (2018). Information theoretic approach to high dimensional multiplicative models: Stochastic discount factor and treatment effect.

Robins, J., Li, L., Tchetgen, E., and van der Vaart, A. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. *Probability and Statistics: Essays in Honor of David A. Freedman*, 2:335–421.

Robins, J., Tchetgen, E. T., Li, L., and van der Vaart, A. (2009). Semiparametric minimax rates. *Electronic journal of statistics*, 3:1305.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954.

Roderick, J., Little, A., and Rubin, D. B. (2002). *Statistical analysis with missing data*.

Rosenbaum, P. R. and Rubin, D. B. (1983a). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rosenbaum, P. R. and Rubin, D. B. (1983b). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1):41–55.

Rosenberg, J. V. and Engle, R. F. (2002). Empirical pricing kernels. *Journal of Financial Economics*, 64(3):341–372.

Rossin-Slater, M. (2017). Signing up new fathers: Do paternity establishment initiatives increase marriage, parental investment, and child well-being? *Amer-*

ican Economic Journal: Applied Economics, 9(2):93–130.

Rothe, C. and Firpo, S. (2016). Properties of doubly robust estimators when nuisance functions are estimated nonparametrically. Technical report, Working paper.

Rotnitzky, A., Lei, Q., Sued, M., and Robins, J. M. (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika*, 99(2):439–456.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Schumaker, L. (2007). *Spline functions: basic theory*. Cambridge University Press.

Shen, X. (1997). On methods of sieves and penalization. *Annals of Statistics*, 25(6):2555–2591.

Smith, R. J. (1997). Alternative semi-parametric likelihood approaches to generalised method of moments estimation. *The Economic Journal*, 107(441):503–519.

Stock, J. H. (1989). Nonparametric policy analysis. *Journal of the American Statistical Association*, 84(406):567–575.

Stoker, T. M. (1986). Consistent estimation of scaled coefficients. *Econometrica: Journal of the Econometric Society*, pages 1461–1481.

Stutzer, M. (1995). A bayesian approach to diagnosis of asset pricing models. *Journal of Econometrics*, 68(2):367–397.

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97(3):661–682.

Tang, C. Y. and Leng, C. (2010). Penalized high-dimensional empirical likelihood. *Biometrika*, 97(4):905–920.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58(1):267–288.

Tropp, J. A. (2015). An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230.

Tsiatis, A. (2007). *Semiparametric theory and missing data*. Springer Science & Business Media.

Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes: with applications to statistics*. Springer.

Van de Geer, S. (2007). The deterministic lasso. Seminar für Statistik, Eidgenössische Technische Hochschule (ETH) Zürich.

Van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The*

Annals of Statistics, 42(3):1166–1202.

Van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 36(2):614–645.

Van De Geer, S. A., Bühlmann, P., et al. (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392.

Van Der Vaart, A. et al. (1991). On differentiable functionals. *The Annals of Statistics*, 19(1):178–204.

Van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3. Cambridge university press.

Vasicek, O. (1977). An equilibrium characterization of the term structure. *Journal of financial economics*, 5(2):177–188.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press.

Wong, R. K. and Chan, K. C. G. (2018). Kernel-based covariate functional balancing for observational studies. *Biometrika*, 105(1):199–213.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, pages 894–942.

Zhang, C.-H. and Zhang, S. S. (2014a). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242.

Zhang, C.-H. and Zhang, S. S. (2014b). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242.

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922.