Near Optimal Estimation of Average Regression Functionals

Chen Qiu*

This version: September 21, 2020

Abstract

This paper considers estimation of a continuous linear regression functional that can be written as a weighted population mean of observed outcome. A leading example is the population average treatment effect under unconfoundedness and overlap conditions, when the weight function is the product of binary treatment and inverse propensity score. We propose a new plug-in estimator of the functional when the weight function is approximated in series space. This estimator is near optimal in the sense that its mean square remainder error is controlled as small as possible in finite sample. We characterize its asymptotic distribution, allowing the number of basis functions k to grow proportionally to sample size n. We compare both finite sample and asymptotic performance of our estimator with doubly robust (DR) estimators. We find DR estimators often do not have materially smaller mean square remainder error in finite sample. DR estimators also do not improve the asymptotic performance when the ratio of k to n is smaller than 1. When the ratio is larger than 1, we propose a modified DR estimator to improve the asymptotic performance of our plug-in estimator. We apply our method to the work of Ferraz and Finan (2011) and conduct simulations to support theoretic findings.

*The Institute for Fiscal Studies, 7 Ridgmount Street London WC1E 7AE. Email: c.qiu.lse@gmail.com. This is a revision of my job market paper entitled *Minimax Learning for Average Regression Functionals*. I am indebted to Taisuke Otsu for patient and continuous guidance and support. This draft benefits from useful discussions with Jaap Abbring, Debopam Bhattacharyam, Ben Deaner, Rachael Meager, Francesca Molinari, Whitney Newey, Alexey Onatskiy, Jörg Stoye and Daniel Sturm. I also thank numerous seminar and conference participants for helpful feedback.

1 Introduction

In many empirical problems, parameter of interest is a continuous linear functional of a regression function in a Hilbert space. In this paper, we call this parameter an *average regression functional*. One leading example is the population average treatment effect under unconfoundedness and overlap conditions. As modern datasets are often high dimensional (for example, in an observational study, the number of pretreatment covariates is possibly non negligible compared to sample size), it is important to find an estimator with estimation error as small as possible.

One standard framework of evaluating estimators is based on the asymptotic order of their remainder error terms (remainders) that arise from the asymptotic linear expansion. If remainders vanish to zero asymptotically, the estimator is $\sqrt{n}$ normal and usually also semiparametrically efficient[1]. This observation motivates Newey and Robins (2018) to find an estimator with remainders converging to zero asymptotically as fast as possible. Their estimator relies on doubly robust (DR, or Neyman orthogonal) moment and cross fitting (also see, among others, Farrell, 2015; Chernozhukov et al., 2016; Rothe and Firpo, 2016; Belloni et al., 2017; Chernozhukov et al., 2018a) to reduce the asymptotic bias. By construction, the asymptotic impact of estimation error is minimal.

Another approach focuses on the finite sample performance. Since many linear functionals are not $\sqrt{n}$ estimable (for example, the value of regression function at a fixed point), this approach directly controls finite sample error by linear minimax exercise (see for example, Armstrong and Kolesár, 2018b; Imbens and Wager, 2018). The resulting estimator and confidence interval are by construction, optimal in the minimax sense among a class of linear estimators. Armstrong and Kolesár (2018a) extend this approach to conditional (or sample based) average treatment effect, and one advantage is they can also deal with scenarios when $\sqrt{n}$ inference is not possible (for example, when the overlap condition is violated).

This paper complements previous two influential approaches. We propose a nearly optimal procedure of estimating average regression functional when its semiparametric efficiency bound is finite. We start by writing the average regression functional alternatively as a population weighted mean of observed outcome. For example, in the case of population average treatment effect, the weight function is the product of the binary treatment and inverse propensity score. This weight function plays a crucial role in semiparametric theory and is in fact the Riesz Representer (RR, also see Newey, 1990, 1994). We consider a class of plug-in estimators where the weight function is approximated in a series (linear sieve) space. These estimators are closely related to the balancing method in

[1]See, among others, Newey (1990); Van Der Vaart et al. (1991); Bickel et al. (1993); Andrews (1994); Newey (1994); Newey and McFadden (1994) for a general treatment of semiparametric inference.

statistics literature (for example, Hainmueller, 2012; Zubizarreta, 2015; Chan et al., 2016; Kallus, 2016; Athey et al., 2018) and we call them *balancing plug-in* (BP) estimators.

Our idea is to find a BP estimator so that its remainders are as small as possible in finite sample. This motivates a new performance benchmark called *maximum mean square remainder* (MMSR, formally defined in Section 3.1) in a given sample. MMSR is analogous to the notion of maximum mean square error but focuses on remainders. Since RR is usually unknown, directly minimizing MMSR is infeasible in general. Instead, we minimize a feasible upper bound of MMSR. The linear structure of series space offers tractability: Our minimax exercise can be solved analytically by considering a minimum distance criterion with a ridge style penalty, a form similar to the penalized sieve minimum distance (PSMD) estimator in Chen and Pouzo (2012, 2015). Our derived estimator is thus near optimal in the sense that it minimizes the MMSR bound among all BP estimators in the series space. Building on a vast literature in series estimation[2], we characterize the asymptotic distribution of our estimator and allow the number of basis functions k to grow proportionally to sample size n. We also compare the finite sample as well as asymptotic performance of our estimator with DR approach.

Our estimator is also inspired by Wong and Chan (2018); Hirshberg and Wager (2018), who investigate how to directly control remainders by minimax exercise. However, our paper is different from theirs in several notable aspects: first, their minimax exercises are embedded in infinite dimensional space while this paper focuses on series space; second, their main analyses are centered around the asymptotic performance of DR estimators, while this paper focuses on BP method; third, we formalize the minimax exercise as a finite sample performance criterion regarding remainders, and compare finite sample as well as asymptotic performance with DR estimators. Our paper is also related to Chernozhukov et al. (2018b,c). One of their results show that DR moment and cross fitting are also effective for regularized estimators of the RR in high dimensional cases, and their estimators can achieve $\sqrt{n}$ normality under weak conditions.

In terms of finite sample results, we show that DR estimators do not achieve smaller MMSR bound unless the estimator for the regression function is precise enough in finite sample. The improvement from DR estimators is also likely to be limited. When the sample Gram matrix of basis functions is orthogonal, we derive a sharp lower bound on the relative efficiency of DR estimators compared to our estimator. These theoretic results are supported by simulations in Appendix E.3 where we compare finite sample performance of our estimator with various DR estimators.

In terms of asymptotic results, our estimator can achieve $\sqrt{n}$ normality without a

[2]See, among others, Newey (1997); Shen (1997); Huang (2003); Ai and Chen (2003); Newey and Powell (2003); Chen (2007) and more recently, Belloni et al. (2015); Chen and Christensen (2015); Hansen (2015).

consistent estimator of the weight function. The remainder of our estimator also converges to zero asymptotically at a rate fast enough, in the sense that it achieves the minimal condition for semiparametric efficiency in Robins et al. (2009) if the regression function is smooth enough.[3] One important implication is that, if $\frac{k}{n} < 1$, the remainder of DR estimator (without cross fitting) does not converge faster than that of our estimator. Therefore, if our estimator does not attain semiparametric efficiency when $\frac{k}{n} < 1$, resorting to DR estimators alone will not help. In this scenario, cross fitting (with or without DR structure) can improve the *asymptotic* remainder rate but does not have the finite sample property established in this paper, to the best of my knowledge. As a technical contribution, we also allow the minimum eigenvalue of the sample Gram matrix to diminish to 0 at a fast rate for certain functionals.

When $\frac{k}{n} \geq 1$, the remainder of our estimator is often growing asymptotically. In this case, it is indeed possible to improve the asymptotic performance by a DR structure. Based on this observation, we propose a modified minimum distance estimator for the weight function with an elastic net style penalty. This modified estimator converges at least as fast as its lasso counterpart. We also develop algorithms to choose penalty coefficients in a data-driven way. With this modified estimator, we can build a DR estimator, which achieves semiparametric efficiency when $\frac{k}{n} \geq 1$ under conditions weaker than those in Belloni et al. (2017); Chernozhukov et al. (2018a,c) and comparable to Chernozhukov et al. (2018c).

Our procedure is computationally convenient, and is also of interest to applied researchers who often encounter datasets large both in terms of sample size and controls. As an empirical illustration, I revisit Ferraz and Finan (2011)'s work that studies the effect of electoral accountability on corruption. With plausibly exogenous treatment, one of their main empirical strategies is OLS with many controls. As a standard practice, they sequentially add different sets of relevant controls to the regression. However, for this dataset, point estimates do change considerably, jumping almost 50% from a plain vanilla mean comparison to a full specification with the ratio $\frac{k}{n} = 0.14$. While such coefficient instability is usually interpreted as a bias correction, we show that there is another probable interpretation of the results.[4] Applying our estimator to the same dataset, estimated treatment effect behaves stably. Thus observed coefficient instability can also be associated with sub optimal control of mean square remainders in the presence of many controls. This is a deficiency of the estimation method itself. Ignoring such effect could lead to misinterpretation of many empirical results. On the other hand, our estimator tries to control remainders in an optimal manner and should be more robust for

[3] Recently, Bradic et al. (2019) also derive the minimax condition for $\sqrt{n}$ consistent and efficient estimation of some average regression functional under sparsity assumption.

[4] Also see Oster (2017) for a different reason why the common interpretation could be wrong.

moderately high dimensional datasets.

The rest of the paper is organized as follows: Section 2 introduces the framework and running example. Section 3 presents the main methodology and compares the finite sample performance with DR estimators. Section 4 develops the asymptotic theory and compares the asymptotic performance with DR estimators when $\frac{k}{n} < 1$. Section 5 shows that DR structure can improve the asymptotic performance of our estimator when $\frac{k}{n} \geq 1$, and proposes a modified elastic net style estimator of the weight function. Empirical application is presented in Section 6. Additional technical results, examples, simulations, tables and figures can be found in Appendices and Supplementary Materials.

2 Framework

2.1 Set-up

Our set-up is adapted from Newey and Robins (2018). Let $Y \in \mathbb{R}$ be a random outcome variable and $X \in \mathcal{X} \subseteq \mathbb{R}^{d_X}$ be a random vector. Suppose a random sample $\{(Y_i, X_i')'\}_{i=1}^n$ of size n is drawn from the distribution of $(Y, X')'$. Let $\mathbb{E}[\cdot\,] := \mathbb{E}_{\mathbb{P}}[\cdot\,]$ be the expectation operator under $\mathbb{P}$, where $\mathbb{P} := \mathbb{P}_n$ is the sampling distribution of $\{(Y_i, X_i')'\}_{i=1}^n$. Write

$$Y_i = \gamma_0(X_i) + e_i, \quad \mathbb{E}[e_i|X_i = x] = 0, \quad i = 1 \dots n, \tag{2.1}$$

where $\gamma_0(x) := \mathbb{E}[Y_i|X_i = x]$ is the conditional expectation (regression) function, $\gamma_0 \in L_{\mathbb{P},2} := \left\{f : \mathcal{X} \mapsto \mathbb{R},\ \int_{x\in\mathcal{X}} f^2(x)d\mathbb{P}(x) < \infty\right\}$ and $\sup_{x\in\mathcal{X}} \mathbb{E}[e_i^2|X_i = x] \lesssim 1$. The object of interest in this paper is the continuous linear functional $\mathbb{E}[m(X_i, \cdot)] : L_{\mathbb{P},2} \mapsto \mathbb{R}$ evaluated at γ_0[5]

$$\theta_0 := \mathbb{E}[m(X_i, \gamma_0(X_i))], \tag{2.2}$$

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

$$m(x, r\gamma_1(x) + \gamma_2(x)) = rm(x, \gamma_1(x)) + m(x, \gamma_2(x)). \tag{2.3}$$

By Riesz representation theorem, there exists a unique $\alpha_0 \in L_{\mathbb{P},2}$ such that for each $\gamma \in L_{\mathbb{P},2}$

$$\mathbb{E}[m(X_i, \gamma(X_i))] = \mathbb{E}[\gamma(X_i)\alpha_0(X_i)]. \tag{2.4}$$

[5]We can also extend our method to functional $\mathbb{E}[m(Z_i, \cdot)]$ where $X_i \subseteq Z_i$ at the expense of more technicalities.

Call α_0 the Riesz Representer (RR) of $\mathbb{E}[m(X_i, \cdot)]$. By (2.4), we can interpret θ_0 as RR weighted population average of a regression function (aka "average regression functional")

$$\theta_0 = \mathbb{E}[\underbrace{\alpha_0(X_i)}_{\text{weight}} \underbrace{\gamma_0(X_i)}_{\text{regression}}]. \tag{2.5}$$

2.2 A running example

Many economic problems fit into the average regression functional framework. One leading example is the missing data and average treatment effect model stated below, which is also the workhorse of simulation and empirical application. Other well-known examples include weighted average derivative and single index model, average effect after policy intervention and average consumer surplus, etc. See Appendix A.2 for two additional examples.

Example 2.1. Missing data and average treatment effect

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

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

where $\gamma_0(x, 1) := \mathbb{E}[Y_i | X_i = x, T_i = 1]$. Define the inverse propensity score as $\omega(x) := 1/\mathbb{P}\{T_i = 1 | X_i = x\}$. Further under overlap assumption that $0 < \mathbb{P}\{T_i = 1 | X_i = x\} < 1$ for all $x \in \mathcal{X}$, we have for each $g \in L_{\mathbb{P},2}$

$$\mathbb{E}[\omega(X_i) T_i g(X_i)] = \mathbb{E}[g(X_i)]. \tag{2.6}$$

(2.6) 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).

3 Methodology with a finite-sample motivation

Estimation of θ_0 can be based on any of the following three moment equations

$$\begin{aligned}
\theta_0 :&= \mathbb{E}[m(X_i, \gamma_0(X_i))]; && \text{(DP)} && (3.1)\\
&= \mathbb{E}[\alpha_0(X_i) Y_i]; && \text{(BP)} && (3.2)\\
&= \mathbb{E}[m(X_i, \gamma_0(X_i)) + \alpha_0(X_i)(Y_i - \gamma_0(X_i))]. && \text{(DR)} && (3.3)
\end{aligned}$$

(3.1) is the original definition of θ_0, which we call a "Direct Plug-in" (DP) approach. (3.2) comes from (2.5) and Law of Iterated Expectations (LIE). We call it a "Balancing Plug-in" (BP) approach, treating RR α_0 as a balancing weight function for outcome. (3.3) is the Doubly Robust (DR) moment condition.

To approximate γ_0 and α_0, let $p(x) := (p_1(x), p_2(x) \dots p_k(x))'$ be a vector of $k := k(n)$ basis functions. As $n \to \infty$ and $k \to \infty$, we consider a series (linear sieve) space

$$\Theta_n := \left\{ g : \mathcal{X} \mapsto \mathbb{R}, g(x) = a'p(x), a \in \mathbb{R}^k \right\}.$$

For some $f \in L_{\mathbb{P},2}$, denote $\mathcal{L}_n f$ as the least square projection of f onto Θ_n. Thus write $\mathcal{L}_n \gamma_0 = \beta_l' p$, where $\beta_l := \arg\min_{\beta \in \mathbb{R}^k} \mathbb{E}[\gamma_0(X_i) - \beta' p(X_i)]^2$ is the projection coefficient. Hence (2.1) can be rewritten as

$$Y_i = \beta_l' p(X_i) + u_{\gamma_0 i} + e_i, \qquad u_{\gamma_0 i} = \gamma_0(X_i) - \beta_l' p(X_i), \quad i = 1 \dots n.$$

To motivate the finite sample property of our estimator, assume for now

$$\|\beta_l\| = b, \qquad \mathbb{E}[e_i^2 | X_i = x] = \sigma^2, \tag{3.4}$$

for some constants σ^2 and b[6]. Parameter b reflects the "size" of $\mathcal{L}_n \gamma_0$. Thus define a small ball $\mathcal{H}_b \subseteq \Theta_n$ that contains $\mathcal{L}_n \gamma_0$

$$\mathcal{H}_b := \left\{ g : \mathcal{X} \mapsto \mathbb{R}, g(x) = \beta' p(x), \beta \in \mathbb{R}^k, \|\beta\| \le b \right\}. \tag{3.5}$$

3.1 Near optimal BP estimator in series space

Let $\mathbb{E}_n[f] := \mathbb{E}_n[f(W)] := \frac{1}{n} \sum_{i=1}^n [f(W_i)]$, $W_i = (Y_i, X_i')'$. Given $\{W_i\}_{i=1}^n$ and Θ_n, we focus on a class of BP estimators $\theta_{BP}(\alpha) := \mathbb{E}_n[\alpha(X)Y]$, where $\alpha \in \Theta_n$ is some balancing weight. To find an "optimal" weight function in Θ_n, notice $\theta_{BP}(\alpha)$ necessarily admits the following asymptotic linear structure

$$\sqrt{n}(\theta_{BP}(\alpha) - \theta_0) = \sqrt{n}\mathbb{E}_n \phi + \sqrt{n} R_1(\alpha, \gamma_0) + \sqrt{n} R_2(\alpha), \tag{3.6}$$

$$\text{where } \sqrt{n}\mathbb{E}_n \phi \xrightarrow{d} N(0, \mathbb{E}\phi_i^2),$$

$$\phi_i := \phi(X_i, e_i) := m(X_i, \gamma_0(X_i)) + \alpha_0(X_i) e_i - \theta_0. \tag{3.7}$$

[6] By Lemma B.1(iv), b is finite if all eigenvalues of $\mathbb{E}[p(X_i)p(X_i)']$ are bounded away from 0 and $\mathbb{E}[u_{\gamma_0 i}^2] \lesssim 1$. Also, if e_i is conditionally heteroscedastic, σ^2 can be alternatively interpreted as $\sup_{x \in \mathcal{X}} \mathbb{E}[e_i^2 | X_i = x]$.

Remainder error terms $R_1(\alpha, \gamma_0)$ and $R_2(\alpha)$ admit

$$R_1(\alpha, \gamma_0) := \mathbb{E}_n \left[\alpha(X)\gamma_0(X) - m(X, \gamma_0(X))\right], \tag{3.8}$$

$$R_2(\alpha) := \mathbb{E}_n \left[(\alpha(X) - \alpha_0(X))\, e\right]. \tag{3.9}$$

Since $\mathbb{E}\phi_i^2$ is also the semiparametric efficiency bound (Newey, 1994), ideally we hope to select some $\alpha \in \Theta_n$ such that $\sqrt{n}R_1$ and $\sqrt{n}R_2$ are as small as possible in a given sample. With such motivation, one natural performance criterion is the maximum mean square remainder (MMSR) for $\gamma_0 \in \mathcal{H}_b$[7] and conditional on $\mathscr{X}_n := \{X_i\}_{i=1}^n$

$$\begin{aligned} MMSR_{\mathcal{H}_b}(\theta_{BP}(\alpha)) &:= \sup_{\gamma_0 \in \mathcal{H}_b} \mathbb{E}\left[(R_1(\alpha, \gamma_0) + R_2(\alpha))^2 | \mathscr{X}_n\right] \\ &= \underbrace{b^2 \sup_{\gamma_0 \in \mathcal{H}_1} R_1^2(\alpha, \gamma_0)}_{\text{maximum square bias}} + \underbrace{\frac{\sigma^2}{n}\mathbb{E}_n[(\alpha(X) - \alpha_0(X))^2]}_{\text{variance}}, \end{aligned} \tag{3.10}$$

where (3.10) follows since $\mathbb{E}[R_2(\alpha)|\mathscr{X}_n] = 0$ and $\sup_{\gamma_0 \in \mathcal{H}_b} R_1^2(\alpha, \gamma_0) = b^2 \sup_{\gamma_0 \in \mathcal{H}_1} R_1^2(\alpha, \gamma_0)$. In general, α_0 is unknown[8], so we replace $MMSR_{\mathcal{H}_b}(\theta_{BP}(\alpha))$ with a feasible MMSR upper bound

$$\overline{MMSR}_{\mathcal{H}_b}(\theta_{BP}(\alpha)) := \underbrace{b^2 \sup_{\gamma_0 \in \mathcal{H}_1} R_1^2(\alpha, \gamma_0)}_{\text{maximum square bias}} + \underbrace{\frac{2\sigma^2}{n}\left\{\mathbb{E}_n[\alpha^2(X)] + \mathbb{E}_n[\alpha_0^2(X)]\right\}}_{\text{variance bound}}.$$

Since $\overline{MMSR}_{\mathcal{H}_b}(\theta_{BP}(\alpha))$ depends on α only through $\sup_{\gamma_0 \in \mathcal{H}_1} R_1^2(\alpha, \gamma_0)$ and $\mathbb{E}_n[\alpha^2(X)]$, it suffices to consider the following Lagrange problem

$$\tilde{\alpha} := \tilde{\alpha}_{\lambda_1} := \arg\min_{\alpha \in \Theta_n} \left\{ \sup_{\gamma_0 \in \mathcal{H}_1} R_1^2(\alpha, \gamma_0) \quad + \quad \lambda_1 \mathbb{E}_n[\alpha^2(X)] \right\}, \tag{3.11}$$

where $\lambda_1 \geq 0$ is a penalty coefficient. (3.11) calibrates α_0 in a penalized series space with a minimax criterion. Our proposed BP estimator is then

$$\tilde{\theta}_{BP} := \theta_{BP}(\tilde{\alpha}) = \mathbb{E}_n[\tilde{\alpha}(X)Y]. \tag{3.12}$$

Notice the solution of (3.11) traces out the bias-variance frontier of $\overline{MMSR}_{\mathcal{H}_b}(\theta_{BP}(\alpha))$. Thus, a practical way to select optimal λ_1 is to pick one that minimizes $\overline{MMSR}_{\mathcal{H}_b}(\tilde{\theta}_{BP})$ in a grid of possible values.[9] Let $\lambda_1^* := \arg\min_{\lambda_1} \overline{MMSR}_{\mathcal{H}_b}(\theta_{BP}(\tilde{\alpha}_{\lambda_1}))$ and $\tilde{\theta}_{BP}^* := \mathbb{E}_n[\tilde{\alpha}_{\lambda_1^*}(X)Y]$. Then $\tilde{\theta}_{BP}^*$ is near optimal in the sense that it minimizes the MMSR bound.

[7] This is a single optimality criterion. It would be interesting to extend our approach to a double optimality criterion with respect to both γ_0 and α_0. We leave this for future research.

[8] If α_0 is known, like the case of randomized control trials, our method can be applied to get a sharp estimator. See Appendix A.5 for further discussions.

[9] Since in practice b and σ are unknown, we recommend conducting sensitivity analysis against a series of possible values for the ratio $\frac{\sigma^2}{nb^2}$.

3.2 Implementation

Proposition 3.1. *For each* $\alpha \in \Theta_n$, $\sup_{\gamma_0 \in \mathcal{H}_1} R_1^2(\alpha, \gamma_0) = \|\mathbb{E}_n[m(X, p(X)) - \alpha(X)p(X)]\|^2$.

By Proposition 3.1, the solution of minimax exercise (3.11) can be found analytically and $\tilde{\alpha}$ has an equivalent minimum distance representation

$$\tilde{\alpha} = \arg\min_{\alpha \in \Theta_n} \left\{ \underbrace{\|\mathbb{E}_n[m(X, p(X)) - \alpha(X)p(X)]\|^2}_{\text{minimum distance}} + \underbrace{\lambda_1 \mathbb{E}_n[\alpha^2(X)]}_{\text{ridge style penalty}} \right\}. \tag{3.13}$$

The objective function in (3.13) is convex and continuously differentiable: the first half of the objective function is a squared Euclidean distance, and the second half is a ridge style penalty in series space.[10] This quadratic optimization exercise has an analytic solution

$$\begin{aligned} \tilde{\alpha} &= p'\tilde{a}, & \tilde{a} &:= (\hat{G}\hat{G} + \lambda_1 \hat{G})^- \hat{G}\hat{P}, \\ \hat{G} &:= \mathbb{E}_n[p(X)p(X)'], & \hat{P} &:= \mathbb{E}_n[m(X, p(X))], \end{aligned} \tag{3.14}$$

and $(\cdot)^-$ denotes the Moore–Penrose inverse. By construction, $\tilde{\alpha}$ is the unique solution of (3.13) when $\hat{G}$ is invertible and a solution with minimum norm otherwise.

Remark 3.1. Focusing on BP estimators does not lose out DP estimators. In fact, for each DP estimator $\theta_{DP}(\gamma) := \mathbb{E}_n[m(X, \gamma(X))]$ where $\gamma \in \Theta_n$, there exists at least one $\alpha_\gamma \in \Theta_n$ such that $\theta_{BP}(\alpha_\gamma) = \theta_{DP}(\gamma)$. For example, a DP estimator with standard series estimator

$$\dot{\gamma}(x) := p(x)'\dot{\beta}, \dot{\beta} := \hat{G}^- \mathbb{E}_n[p(X)Y] \tag{3.15}$$

is equivalent to the BP estimator using $\dot{\alpha}(x) := p(x)'\hat{G}^-\hat{P}$, the estimator for α_0 proposed in Newey and Robins (2018). We reserve further comparison between BP and DR estimators in Sections 3.3 and 4.2.

3.3 Scope for finite sample improvement by DR estimators

In this section, we study whether a DR estimator can improve the finite sample performance of $\tilde{\theta}^*_{BP}$, in the sense of achieving a substantively smaller MMSR bound. Consider a class of DR estimator

$$\theta_{DR}(\alpha, \gamma) := \mathbb{E}_n[m(X, \gamma(X)) + \alpha(X)(Y - \gamma(X))],$$

where $\alpha \in \Theta_n$, $\gamma = \beta' p \in \Theta_n$ for some $\beta \in \mathbb{R}^k$. Note the remainder of $\theta_{DR}(\alpha, \gamma)$ shares a similar structure with $\theta_{BP}(\alpha)$:

[10] Such structure is similar to the PSMD estimator studied in Chen and Pouzo (2012, 2015) but with a different motivation.

$$\sqrt{n}(\theta_{DR}(\alpha,\gamma)-\theta_0)=\sqrt{n}\mathbb{E}_n\phi+\sqrt{n}R_1(\alpha,\gamma_0-\gamma)+\sqrt{n}R_2(\alpha). \tag{3.16}$$

Since $\gamma\in\Theta_n$, the MMSR bound for $\theta_{DR}(\alpha,\gamma)$ conditional on $\mathscr{X}_n$ is

$$\overline{MMSR}_{\mathcal{H}_b}(\theta_{DR}(\alpha,\gamma))=b_\gamma^2\sup_{\gamma_0\in\mathcal{H}_1}R_1^2(\alpha,\gamma_0)+\frac{2\sigma^2}{n}\left\{\mathbb{E}_n[\alpha^2(X)]+\mathbb{E}_n[\alpha_0^2(X)]\right\},$$

where $b_\gamma=\|\beta_l-\beta\|$. Note $\overline{MMSR}_{\mathcal{H}_b}(\theta_{DR}(\alpha,\gamma))$ is different from $\overline{MMSR}_{\mathcal{H}_b}(\theta_{BP}(\alpha))$ only in the scale of the bias part b_γ^2. So for each $\alpha\in\Theta_n$, $\overline{MMSR}_{\mathcal{H}_b}(\theta_{DR}(\alpha,\gamma))<\overline{MMSR}_{\mathcal{H}_b}(\theta_{BP}(\alpha))$ only when $b_\gamma<b$. Moreover, irrespective of the magnitude of b_γ, $\tilde{\alpha}$ also traces out the bias-variance frontier of $\overline{MMSR}_{\mathcal{H}_b}(\theta_{DR}(\alpha,\gamma))$. Thus let $\lambda_{1,\gamma}^*:=\arg\min_{\lambda_1}\overline{MMSR}_{\mathcal{H}_b}(\theta_{DR}(\tilde{\alpha},\gamma))$. Then $\theta_{DR}(\tilde{\alpha}_{\lambda_{1,\gamma}^*},\gamma)$ minimizes $\overline{MMSR}_{\mathcal{H}_b}(\theta_{DR}(\alpha,\gamma))$ and is thus the near optimal DR estimator given γ.

Proposition 3.2. *The relative efficiency of $\theta_{DR}(\tilde{\alpha}_{\lambda_{1,\gamma}^*},\gamma)$ compared to $\tilde{\theta}_{BP}^*$ is*

$$RE:=\sqrt{\frac{\overline{MMSR}_{\mathcal{H}_b}(\theta_{DR}(\tilde{\alpha}_{\lambda_{1,\gamma}^*},\gamma))}{\overline{MMSR}_{\mathcal{H}_b}(\tilde{\theta}_{BP}^*)}}.$$

If $b_\gamma<b$, $RE>\frac{b_\gamma}{b}$; otherwise, $RE\geq 1$. In particular, if $\hat{G}=I$, $RE>\sqrt{\frac{1+2\varrho}{1+2\varrho/\left(\frac{b_\gamma}{b}\right)^2}}$, where $\varrho:=\frac{\sigma^2}{nb^2}$ is the variance-size ratio.

Remark 3.2. Proposition 3.2 implies two things. First, $\theta_{DR}(\tilde{\alpha}_{\lambda_{1,\gamma}^*},\gamma)$ will not improve finite sample performance of $\tilde{\theta}_{BP}^*$ in the minimax sense unless $b_\gamma<b$. This is a demanding finite sample requirement for the quality of estimator γ. Otherwise, if $b_\gamma\geq b$, $\theta_{DR}(\tilde{\alpha}_{\lambda_{1,\gamma}^*},\gamma)$ is in fact no better and usually worse than $\tilde{\theta}_{BP}^*$. Thus, using DR estimator might actually risk losing robustness in the minimax sense. Consider a simple example when $\hat{G}=I$ and $u_{\gamma_0}=0$. If we use the series estimator $\dot{\gamma}(x)=p(x)'\dot{\beta}$ for γ_0, it follows $b_{\dot{\gamma}}:=\left\|\beta_l-\dot{\beta}\right\|=\|\mathbb{E}_n[p(X)e]\|$. By Markov inequality, $\mathbb{P}\{b_{\dot{\gamma}}>b|\mathscr{X}_n\}\leq k\varrho$, small only when the product of k and ϱ is small. Second, even if $b_\gamma<b$ indeed holds true in a given sample, the scope for improvement from using a DR estimator is likely limited. In fact, if $b_\gamma<b$, the relative efficiency of any optimal DR estimator is always larger than $\frac{b_\gamma}{b}$. An illustrative case is when $\hat{G}=I$, for which we are able to derive a sharp lower bound of RE. And the room for finite sample improvement by any DR estimator can be quite small. See Figure 3.1.

4 Asymptotic properties of $\tilde{\theta}_{BP}$

This section studies asymptotic properties of $\tilde{\theta}_{BP}$ for some $\lambda_1\geq 0$. We first give a general characterization of the asymptotic distribution when $\frac{k}{n}\leq 1$, followed by discussions on

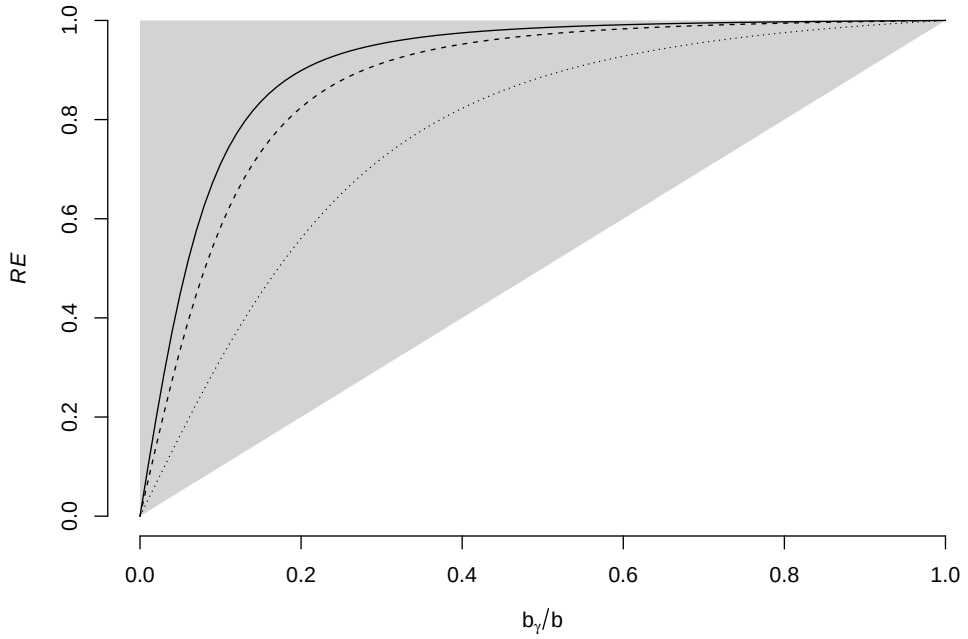


Figure 3.1: Relative efficiency (RE) as a function of $\frac{b_\gamma}{b}$. When $b_\gamma < b$, RE of any near optimal DR estimator always falls in the grey area. Assume $\hat{G} = I$. The solid black line plots the lower bound of RE when $\varrho = 0.005$, corresponding to the case $u_{\gamma_0} = 0$ and $\mathbb{P}\{b_{\hat{\gamma}} > b|\mathscr{X}_n\} \leq 0.005k$; The dashed line plots the lower bound when $\varrho = 0.01$ ($u_{\gamma_0} = 0$ and $\mathbb{P}\{b_{\hat{\gamma}} > b|\mathscr{X}_n\} \leq 0.01k$); The dotted line plots the lower bound when $\varrho = 0.05$ ($u_{\gamma_0} = 0$ and $\mathbb{P}\{b_{\hat{\gamma}} > b|\mathscr{X}_n\} \leq 0.05k$).

whether DR estimator can improve the asymptotic performance of $\tilde{\theta}_{BP}$. Then we look at when semiparametric efficiency is achievable and conditions that further relax some technical conditions in the general theorem.

4.1 General characterization

Assumption O.

1. *Sample* $\{(Y_i, X_i')'\}_{i=1}^n$ *are independently and identically distributed (iid) for each* n. *Function* $m(x, \cdot)$ *is linear in the sense of (2.3). RR* α_0 *exists and satisfies (2.4).*
2. $\mathbb{E}[e_i|X_i = x] = 0$, $\sup_{x\in\mathcal{X}} \mathbb{E}[e_i^2|X_i = x] \lesssim 1$, *for each* $i = 1 \dots n$.
3. *For each* $\gamma \in \Theta_n$ *and* $i = 1 \dots n$, $\mathbb{E}[m^2(X_i, \gamma(X_i))] \lesssim \mathbb{E}[\gamma^2(X_i)]$, $\mathbb{E}[m^2(X_i, u_{\gamma_0 i})] \lesssim \mathbb{E}u_{\gamma_0 i}^2$.

Assumption O(1) is a basic condition on data structure. Note the dimension of X can be either fixed or growing.[11] O(2) restricts the behavior of first two conditional moments

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

of e_i. Exogeneity condition $\mathbb{E}[e_i|X_i = x] = 0$ is automatically satisfied by definition of γ_0. We also assume that e_i has a finite conditional variance. O(3) imposes sufficient degree of continuity on the structure of $\mathbb{E}[m^2(X_i, \cdot)]$ and is satisfied by Examples 2.1, A.1 and A.2.[12]

Assumption L.

1. *All eigenvalues of* $G := \mathbb{E}[p(X_i)p(X_i)']$ *are bounded from above and away from zero uniformly for each* n, k, *and* i;

2. *There exist some vectors* $\beta_b, a_b \in \mathbb{R}^k$ *and finite constants* $\mathbf{r}_{\gamma_0}, \mathbf{r}_{\alpha_0}$ *such that*

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

Assumption L is an essential condition on series approximation. L(1) requires that p should not be too collinear or grow too quickly, similar to Assumption C.1(iii) in Chen and Pouzo (2012) and Condition A.2 in Belloni et al. (2015).[13] L(2) imposes mild restrictions on the approximation quality of Θ_n. Under correct specification, $\mathbf{r}_{\gamma_0} \to 0$ and $\mathbf{r}_{\alpha_0} \to 0$ as $k \to \infty$ and $n \to \infty$. When γ_0 and α_0 are within a Hölder class of smoothness order s, $\mathbf{r}_{\gamma_0} = k^{-\eta_\gamma}, \mathbf{r}_{\alpha_0} = k^{-\eta_\alpha}$ for some non negative constants η_γ and η_α depending on s, d_X and p. See, among others, DeVore and Lorentz (1993); Newey (1997); Chen (2007) for more details on approximation results. Following Newey (1997), let $\xi_k = \sup_{x\in\mathcal{X}} \|p(x)\|$.

Assumption M.

1. $\frac{\xi_k^2 \log k}{n} \leq 1$;

2. *All eigenvalues of* $\hat{G}$ *are bounded away from zero with probability approaching one (wpa1).*

Assumption M(1) allows k to grow proportionally to sample size. As a result, $\tilde{\alpha}$ could be inconsistent. M(2) guarantees that $\hat{G}^{-1} = O_p(1)$ and can be satisfied if the ratio $\frac{k}{n}$ is small enough. For example, by Lemma C.7, if $G = I$ and Assumptions O and L hold true, a sufficient condition for M(2) is that $\frac{\xi_k^2}{n}$ converges to a constant strictly smaller than 0.38 (up to log terms). M(2) can be further relaxed for some special structures of $m(x, \cdot)$. See Section 4.4 for details.

[12] If O(3) is not satisfied, we can modify it such that $\mathbb{E}[m^2(X_i, \gamma(X_i))] \leq d_k \mathbb{E}[\gamma^2(X_i)]$ for $\gamma \in \Theta_n$ and $\mathbb{E}[m^2(X_i, u_{\gamma_0 i})] \lesssim d_k \mathbb{E}u_{\gamma_0 i}^2$, for some d_k growing as a function of k, similar to Assumption 6 in Newey and Robins (2018).

[13] For example, it will be satisfied if p is orthonormal with respect to Lebesgue measure and the density of X_i is bounded away from zero and from above.

For $f \in \Theta_f \subseteq L_{\mathbb{P},2}$, define $\ell_k := \sup\left(\frac{\|\mathcal{L}_n f\|_{\mathbb{P},\infty}}{\|f\|_{\mathbb{P},\infty}} : \|f\|_{\mathbb{P},\infty} \neq 0, f \in \Theta_f\right)$, where $\|f\|_{\mathbb{P},\infty} := \sup_{x\in\mathcal{X}} |f(x)|$.[14] Also let $\|f\|^2_{\mathbb{P},2} := \int_{x\in\mathcal{X}} f^2(x) d\mathbb{P}(x)$. Denote $\delta_n := \left(\sqrt{\frac{\xi_k^2}{n}} \wedge \sqrt{\frac{\|\alpha_0\|_{\mathbb{P},\infty} k}{n}}\right) + \mathbf{r}_{\alpha_0}$.

Theorem 4.1. *Suppose Assumptions O, L and M hold true and* $\mathbf{r}_{\alpha_0} = O(1)$. *In addition, (i)* $\sqrt{n}\mathbf{r}_{\alpha_0}\mathbf{r}_{\gamma_0} = o(1)$; (ii) $\delta_n \mathbf{r}_{\gamma_0} \ell_k \left(\sqrt{k \log \xi_k} + \frac{k\xi_k \log \xi_k}{\sqrt{n}}\right) \wedge \delta_n \mathbf{r}_{\gamma_0}\sqrt{n} = o(1)$; *(iii)* $(\|\alpha_0\|_{\mathbb{P},\infty} \wedge \ell_k)\mathbf{r}_{\gamma_0} = o(1)$; *(iv)* $\lambda_1 = o(\frac{1}{n})$; *(v)* $\max_i|\tilde{\alpha}(X_i)|/\sqrt{n} = o_p(1)$, $\sup_{x\in\mathcal{X}} \mathbb{E}\left[|e_i|^3 |X_i = x\right] \lesssim 1$ *and* $\inf_{x\in\mathcal{X}} \mathbb{E}\left[e_i^2 | X_i = x\right]$ *is bounded away from zero for each* $i = 1 \dots n$, $\|\alpha_0\|^2_{\mathbb{P},2} - \mathbf{r}^2_{\alpha_0}$ is bounded away from zero uniformly for each k and n. *Then:*

$$\sqrt{n}\left(\tilde{\theta}_{BP} - \theta_0\right) \xrightarrow{d} \sigma_{\tilde{\alpha},n} Z_1 + \sigma_m Z_2, \tag{4.1}$$

where $\sigma^2_{\tilde{\alpha},n} := \mathbb{E}_n\left[\tilde{\alpha}^2(X)\mathbb{E}[e^2|X]\right]$, $\sigma^2_m := \mathbb{E}[m^2(X_i, \gamma_0(X_i))] - \theta_0^2$, *and* Z_1 *and* Z_2 *are two iid standard normal random variables independent of* $\{X_i, e_i\}_{i=1}^n$. *If in addition* $\sigma^2_{\tilde{\alpha},n} \xrightarrow{p} \mathcal{V} > 0$, *then*

$$\sqrt{n}\left(\tilde{\theta}_{BP} - \theta_0\right) \xrightarrow{d} N(0, \mathcal{V} + \sigma^2_m). \tag{4.2}$$

Theorem 4.1 establishes $\sqrt{n}$ consistency and the asymptotic distribution of $\tilde{\theta}_{BP}$ without imposing strong convergence conditions for nuisance parameters, cross fitting or DR moment conditions. In general, the asymptotic distribution of $\tilde{\theta}_{BP}$ is a weighted sum of two iid standard normal random variables. If $\sigma^2_{\tilde{\alpha},n}$ converges to some constant $\mathcal{V}$ in large sample, $\tilde{\theta}_{BP}$ is $\sqrt{n}$ normal. Condition (i) is a basic requirement on approximation quality. (ii) is an additional model complexity term from controlling a stochastic equicontinuity term by a maximal inequality (Giné and Koltchinskii 2006, Theorem 3.1; Belloni et al. 2015, Theorem 6.1). (iii) is a regularity condition that permits $\|\alpha_0\|_{\mathbb{P},\infty}$ to be unbounded. If $\frac{\xi_k^2 \log k}{n} \nrightarrow 0$, conditions (i)-(iii) can be satisfied when $\mathbf{r}_{\alpha_0} = O(1)$ and $\mathbf{r}_{\gamma_0} = o(\frac{1}{\sqrt{n}})$. Thus we allow α_0 to be misspecified as long as γ_0 is correctly specified and smooth enough. (iv) says asymptotically λ_1 should decay fast at rate $o(\frac{1}{n})$. Condition $\max_i|\tilde{\alpha}(X_i)|/\sqrt{n} = o_p(1)$ is needed to invoke Berry–Esseen inequality, and is satisfied under mild conditions that Assumption L(1) holds true and $\|\tilde{a}\| = O_p(1)$ by an argument from Borel-Cantelli lemma. Also see Lemma C.8.

4.2 Scope for asymptotic improvement by DR estimators

Let $\mathcal{R}_n$ and $\mathcal{R}_n^\gamma$ be the *asymptotic* orders of the remainders of $\tilde{\theta}_{BP}$ and $\theta_{DR}(\tilde{\alpha}, \gamma)$, respectively. By Theorem 4.1, we can compare the magnitudes of $\mathcal{R}_n$ and $\mathcal{R}_n^\gamma$ and address whether it possible to improve the asymptotic performance of $\tilde{\theta}_{BP}$ by a DR structure.

[14] For certain basis functions, ℓ_k exploits stability relations of projection and allows weaker conditions on the growth rate of k (see also Huang, 2003; Belloni et al., 2015; Chen and Christensen, 2015).

Corollary 4.1. *Suppose Assumptions O, L and M hold true,* $\mathbf{r}_{\gamma_0} = O(1)$, $\mathbf{r}_{\alpha_0} = O(1)$ *and* $\lambda_1 = o(\frac{1}{n})$. *Then,* $\mathcal{R}_n^\gamma \geq \mathcal{R}_n$. *Moreover,* $\mathcal{R}_n \nrightarrow 0$ *as* $n \to \infty$ *implies* $\mathcal{R}_n^\gamma \nrightarrow 0$.

By Corollary 4.1, under the asymptotic framework of $\frac{k}{n} < 1$, we can not improve the asymptotic performance of $\tilde{\theta}_{BP}$ simply by resorting to a DR estimator. If $\tilde{\theta}_{BP}$ can not achieve semiparametric efficiency (i.e., $\mathcal{R}_n \nrightarrow 0$), neither can $\theta_{DR}(\tilde{\alpha}, \gamma)$. The reason is intuitive. Comparing (3.6) with (3.16), the only difference between the remainders of $\theta_{DR}(\tilde{\alpha}, \gamma)$ and $\tilde{\theta}_{BP}$ is in the R_1 part: for $\tilde{\theta}_{BP}$, it writes $R_1(\tilde{\alpha}, \gamma_0)$ and for $\theta_{DR}(\tilde{\alpha}, \gamma)$, it is $R_1(\tilde{\alpha}, \gamma_0 - \gamma)$. Since $\gamma \in \Theta_n$, $\theta_{DR}(\tilde{\alpha}, \gamma)$ only has chance to improve the order of a part in $R_1(\tilde{\alpha}, \gamma_0)$ where $\tilde{\theta}_{BP}$ already controls very well. Another implication of Corollary 4.1 is that if $\frac{k}{n} \geq 1$, DR estimator can improve the asymptotic performance of $\tilde{\theta}_{BP}$. We continue the discussion for the case when $\frac{k}{n} \geq 1$ in Section 5.

4.3 Attainability of semiparametric efficiency

Corollary 4.2. *Suppose* $\frac{\xi_k^2 \log k}{n} = o(1)$ *and* $\mathbf{r}_{\alpha_0} = o(1)$. *Moreover, Assumptions O, L and conditions (i)-(iv) of Theorem 4.1 hold true. Then,* $\sqrt{n}(\tilde{\theta}_{BP} - \theta_0) \xrightarrow{d} N(0, \mathbb{E}[\phi_i^2])$. *Let*

$$\hat{\Omega} = \left| \mathbb{E}_n \left[m(X, \dot{\gamma}(X)) + \tilde{\alpha}(X)(Y - \dot{\gamma}(X)) \right]^2 - \tilde{\theta}_{BP}^2 \right|, \tag{4.3}$$

where $\dot{\gamma}(x)$ *is defined in (3.15). If in addition,* $\mathbb{E}[m^2(X_i, \gamma_0(X_i))] \lesssim \mathbb{E}[\gamma_0^2(X_i)]$, $\|\dot{\gamma} - \gamma_0\|_{\mathbb{P},\infty} = o_p(1)$, *and for some* $q > 0$, $\mathbb{E}\left[|e_i|^{2+q}\right] < \infty$ *and* $\xi_k^{\frac{2+q}{q}} \sqrt{\frac{\log k}{n}} = o(1)$, *then* $\hat{\Omega} \xrightarrow{p} \mathbb{E}[\phi_i^2]$.

If $\frac{\xi_k^2 \log k}{n} = o(1)$ and $\mathbf{r}_{\alpha_0} = o(1)$, $\tilde{\alpha}$ becomes consistent and $\tilde{\theta}_{BP}$ can be semiparametrically efficient. As semiparametric efficiency bound is defined in terms of true RR α_0, the requirement on the consistency of $\tilde{\alpha}$ seems unavoidable. It's useful to compare Corollary 4.2 with other similar results in the literature. Suppose we choose spline or wavelet series as basis functions. Then $\xi_k = \sqrt{k}$ (Newey, 1997) and $\ell_k = O(1)$ (Huang, 2003; Chen and Christensen, 2015). Assume $\mathbf{r}_{\gamma_0} = k^{-\eta}$ for some $\eta > 0$ and ignore log terms:

1. If γ_0 is smooth enough such that $\eta > 1/2$, model complexity condition (ii) in Theorem 4.1 is trivially satisfied as long as $\sqrt{n}\mathbf{r}_{\gamma_0}\mathbf{r}_{\alpha_0} \to 0$. Obtaining semiparametric efficiency requires

$$\frac{k}{n} = o(1), \quad \sqrt{n}\mathbf{r}_{\alpha_0}\mathbf{r}_{\gamma_0} = o(1), \quad \mathbf{r}_{\gamma_0} = o(1), \quad \mathbf{r}_{\alpha_0} = o(1), \tag{4.4}$$

   currently the weakest possible condition in the literature.

2. If γ_0 is not smooth enough in the sense that $\eta \leq \frac{1}{2}$, then in addition to (4.4), we require $\mathbf{r}_{\gamma_0}\frac{k^2}{n} = o(1)$.[15]

[15] Under this case, cross fitted DP and doubly cross fitted DR estimators in Newey and Robins (2018)

Corollary 4.2 also proposes a consistent estimator for the asymptotic variance. Uniform consistency condition $\|\dot{\gamma} - \gamma_0\|_{\mathbb{P},\infty} = o_p(1)$ can be verified by more preliminary conditions.[16] Moreover, we see a trade-off between existence of higher moments for e_i and growth rate restrictions on k. For example, if $\mathbb{E}[e_i^4] < \infty$, we additionally need $\xi_k^2\sqrt{\frac{\log k}{n}} \to 0$. Thus consistent estimation of variance demands slower growth rate of k. Notice sup norm consistency of $\tilde{\alpha}$ is not required.

4.4 Relaxation of Assumption M(2)

For some special linear structure of $m(x,\cdot)$, we can relax Assumption M(2) which may be deemed restrictive in some situations.

Assumption M'.

1. *Function $m(x,\cdot)$ is degenerate: $m(x,p(x)) = p(x)$;*
2. *All eigenvalues of $\hat{G}$ are positive wpa1 and $\lambda_{\min}^{-1}(\hat{G}) = O_p(n^2)$.*

Assumption M' exploits a simple degenerate structure of $m(x,\cdot)$ to leverage nice properties of empirical projection. Such a structure allows $\lambda_{\min}(\hat{G})$ to diminish at an asymptotic rate no faster than $\frac{1}{n^2}$, and is in fact met by Examples 2.1, A.1 and A.2.

Corollary 4.3. *Let Assumptions O, L, M(1), M' and condition (v) in Theorem 4.1 hold true. In addition, $\mathbf{r}_{\alpha_0} = O(1)$, $\mathbf{r}_{\gamma_0} = o\left(\frac{1}{\sqrt{n}}\right)$, $(\|\alpha_0\|_{\mathbb{P},\infty} \wedge \ell_k)\mathbf{r}_{\gamma_0} = o(1)$ and $\lambda_1 = o(\frac{1}{n})$. Then (4.1) and (4.2) still hold true.*

5 Asymptotic improvement when $\frac{k}{n} \geq 1$

Although $\tilde{\theta}_{BP}$ and $\tilde{\theta}^*_{BP}$ are well defined even when $\frac{k}{n} \geq 1$ and their *finite* sample property still holds, the key remainder term $\sqrt{n}R_1(\tilde{\alpha}_{\lambda_1},\gamma_0) \overset{p}{\nrightarrow} 0$ and in fact is usually growing *asymptotically*. Following Corollary 4.1, this less satisfactory asymptotic behavior can be resolved by considering a DR estimator $\hat{\theta}_{DR} := \theta_{DR}(\hat{\alpha},\hat{\gamma})$, where $\hat{\gamma}$ and $\hat{\alpha}$ are some estimators for γ_0 and α_0, respectively.[17] In this section, we first present a general theorem

can meet the minimal requirement (4.4) only in special cases when the Holder orders of α_0 and γ_0 are small enough and p is Haar basis function. Doubly cross fitted DR estimator can also meet minimal requirement if the functional is expected conditional covariance. In general, doubly cross fitted DR estimator has smaller additional term of order $\mathbf{r}_{\gamma_0}\frac{k^2}{n^{3/2}}$ and cross fitted DP estimator has smaller additional term of order $\mathbf{r}_{\gamma_0}\frac{k}{\sqrt{n}}$.

[16] For example, Belloni et al. (2015, Theorem 4.3) and Chen and Christensen (2015, Lemma 2.4) both establish optimal sup norm convergence for $\dot{\gamma}$, allowing $\frac{k}{n} \to 0$ up to log terms. It is also possible to relax this uniform consistency requirement by imposing higher moment conditions for basis functions, see Hansen (2015).

[17] Another potential remedy could be considering minimax exercise in a small ball in terms l_1 norm: $\tilde{\mathcal{H}}_b := \{g = \beta' p : \beta \in \mathbb{R}^k, \|\beta\|_1 \leq b\}$ for some constant b. I leave this for future research.

that establishes semiparametric efficiency of $\hat{\theta}_{DR}$. Then we propose a modified estimator for α_0 based on (3.13), which converges at least as fast as its lasso counterpart and can be used to construct $\hat{\theta}_{DR}$.

5.1 Semiparametric efficiency of $\hat{\theta}_{DR}$

We first impose some assumptions on series space Θ_n suitable when $\frac{k}{n} \geq 1$.

Assumption H. *For each n, k and i:*

1. $\mathbb{E}[p_j(X_i)^2] \lesssim 1$ *for all* $j = 1 \dots k$;
2. *There exists some* $\alpha^* := p'a^*$, $a^* \in \mathbb{R}^k$, *and some finite constant* $\mu_* > 0$ *such that* $\mathbb{E}[\alpha_0(X_i) - \alpha^*(X_i)]^2 \leq \mu_*^2$.

Assumption H is the high dimensional counterpart of Assumption L. H(1) only requires second moments of all basis functions bounded from above uniformly, weaker than Assumption L(1). H(2) imposes a basic approximation condition for α_0. Note α_* is not necessarily sparse. Following Belloni et al. (2017); Qiu and Otsu (2018); Chernozhukov et al. (2018c), let $\Lambda_n := \sup_{x \in \mathcal{X}}(\max_{1 \leq j \leq k} |p_j(x)|)$, more useful in high dimensional situations compared to ξ_k. For $f \in L_{\mathbb{P},2}$, denote $\|f\|_n = \{\mathbb{E}_n[f^2]\}^{1/2}$ as its empirical L^2 norm.

Theorem 5.1. *Let Assumptions O and H hold true. In addition: (i)* $\Lambda_n \sqrt{\frac{\log k}{n}} = o(1)$*; (ii)* $\hat{\gamma}$ *is estimated from a random sample* $\mathcal{S}$ *independent of* $\{(Y_i, X_i')'\}_{i=1}^n$, $\mathbb{E}[m^2(X_i, \hat{\gamma}(X_i) - \gamma_0(X_i))|\mathcal{S}] \lesssim \mathbb{E}[(\hat{\gamma}(X_i) - \gamma_0(X_i))^2|\mathcal{S}]$*; (iii)* $\|\hat{\alpha} - \alpha_*\|_n \mu_* = o_p(\frac{1}{\sqrt{n}})$, $\|\hat{\gamma} - \gamma_0\|_{\mathbb{P},2} \|\hat{\alpha} - \alpha_*\|_n = o_p(\frac{1}{\sqrt{n}})$, $\|\hat{\alpha} - \alpha_*\|_n = o_p(1)$*; (iv) either* $\|\alpha_0\|_{\mathbb{P},\infty} \lesssim 1$ *and* $\|\hat{\gamma} - \gamma_0\|_{\mathbb{P},2} = o_p(1)$*, or* $\|\hat{\gamma} - \gamma_0\|_{\mathbb{P},\infty} = o_p(1)$*. Then* $\sqrt{n}\left(\hat{\theta}_{DR} - \theta_0\right) \xrightarrow{d} N(0, \mathbb{E}[\phi_i^2])$.

Theorem 5.1 establishes semiparametric efficiency of $\hat{\theta}_{DR}$ under a set of general assumptions. Condition (i) in Theorem 5.1 allows k to grow faster than n, up to factor Λ_n and log term. The role of (ii) is twofold. First, it is a simplifying device that can handle any generic estimator $\hat{\gamma}$ (possibly derived from machine learning methods)[18], similar to recent literature advocating cross fitting (for example, Robins et al., 2009; Newey and Robins, 2018; Chernozhukov et al., 2018b,c); Second, it also imposes a mild continuity condition on the functional form of $m(x, \cdot)$. However, we do not require $\hat{\alpha}$ to be estimated from a different random sample, different from the double cross fitting scheme in the literature. (iii) lists key conditions on the quality of $\hat{\gamma}$ and $\hat{\alpha}$. A trade-off between

[18] Otherwise, we need to study the asymptotic order of $\sqrt{n}R_1(\alpha_0, \gamma_0 - \hat{\gamma})$ by empirical process theory on a case-by-case basis, depending on the form of $\hat{\gamma}$. For example, if $\hat{\gamma}$ is estimated by lasso, Belloni et al. (2017, Lemma C.1) can be invoked as long as $\hat{\gamma}$ is sufficiently sparse.

the convergence rates of $\hat{\gamma}$ and $\hat{\alpha}$ exists, so that their product rate shall achieve $o_p(\frac{1}{\sqrt{n}})$. This accommodates a broader scenario when one of them is estimated relatively at a faster rate while the other converges possibly slower than $o_p(n^{-1/4})$. We also only require convergence of $\hat{\alpha}$ under the weak empirical norm $\|\hat{\alpha} - \alpha_*\|_n$. Stronger l_1 convergence rate is not needed per se. Finally (iv) is a regularity condition allowing unbounded α_0 as long as $\hat{\gamma}$ is consistent in sup norm.

5.2 A modified minimum distance estimator for α_0

By Theorem 5.1, semiparametric efficiency of $\hat{\theta}_{DR}$ requires at least a consistent estimator for α_0 in empirical L^2 norm. In this section we modify the original problem (3.13) and derive a new estimator whose asymptotic behavior is simple to characterize when $\frac{k}{n} \geq 1$. Note for $\alpha = a'p \in \Theta_n$, the minimum distance criterion in (3.13) also reads $a' \underbrace{\hat{G}} \hat{G}a - 2a' \underbrace{\hat{G}} \hat{P} + \hat{P}'\hat{P}$, which has an additional weight matrix $\hat{G}$ in the curly bracket adversely affecting estimation when $\frac{k}{n} \geq 1$. Thus, replacing this bracketed $\hat{G}$ with a $k \times k$ identity matrix and omitting $\hat{P}'\hat{P}$, we propose to estimate α_0 by $\widetilde{\alpha} := p'\widetilde{a}$, where

$$\widetilde{a} := \widetilde{a}(\lambda_1, \lambda_2) = \arg\min_{a \in \mathbb{R}^k} \left\{ \underbrace{a'\hat{G}a - 2a'\hat{P}}_{\text{modified minimum distance}} + \underbrace{\lambda_1 a'\Gamma_k a + \lambda_2 \|a\|_1}_{\text{elastic net style penalty}} \right\}, \tag{5.1}$$

Γ_k is a $k \times k$ symmetric and positive semidefinite matrix, and $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are penalty coefficients[19], selected practically by algorithms developed in Appendix D.3. The penalty term in (5.1) is a sum of a weighted l_2 (Tikhonov) penalty and an l_1 penalty, similar to an elastic net regularization.

To describe the asymptotic property of $\widetilde{\alpha}$, let A_* be an index set of nonzero elements of a_*. Denote $S_* := |A_*|$ as the cardinality of A_*. For each $a = (a_1, \ldots, a_k)' \in \mathbb{R}^k$, define $a_{A_*} := (a_{1,A_*}, \ldots a_{j,A_*} \ldots, a_{k,A_*})' \in \mathbb{R}^k$, where for each $j = 1 \ldots k$, $a_{j,A_*} := a_j \mathbf{1}\{j \in A_*\}$. Similarly, define $a_{A_*^c} := (a_{1,A_*^c}, \ldots a_{j,A_*^c} \ldots, a_{k,A_*^c})'$, where $a_{j,A_*^c} := a_j \mathbf{1}\{j \notin A_*\}$ for each $j = 1 \ldots k$. 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 $\hat{\mathcal{G}} := (\hat{G} + \lambda_1 \Gamma_k)$, a generalized

[19] We can alternatively consider a weighted minimum distance problem $\widetilde{a} := \arg\min_{a \in \mathbb{R}^k} \left\{ (\hat{G}a - \hat{P})' \mathcal{W}_n (\hat{G}a - \hat{P}) + \lambda_1 a'\Gamma_k a + \lambda_2 \|a\|_1 \right\}$ with some $k \times k$ positive semidefinite matrix $\mathcal{W}_n$. Its asymptotic property can be analyzed in a similar way but with more technicalities. See the JMP version of this paper.

Gram matrix whose role is similar to that of $\hat{G}$ in a lasso regression. Let

$$\varepsilon_n^m := \|\mathbb{E}_n m(X, p(X)) - \mathbb{E} m(X_i, p(X_i))\|_\infty, \quad \varepsilon_n^\Gamma := \|\Gamma_k a_*\|_\infty,$$
$$\Delta_n := \varepsilon_n^m + \sqrt{\frac{\log k}{n}}\Lambda_n + \mu_* + \lambda_1 \varepsilon_n^\Gamma, \qquad \lambda_0 := 2\left\|\hat{P} - \hat{\mathcal{G}} a_*\right\|_\infty.$$

Theorem 5.2. *Let Assumptions O, H hold true and $\Delta_n = o_p(1)$. Then $\lambda_0 = O_p(\Delta_n)$. If $\lambda_2 = \lambda_0$ and $\Delta_n \|a_*\|_1 = o_p(1)$, it follows $\|\widetilde{\alpha} - \alpha_*\|_n = O_p(\sqrt{\Delta_n \|a_*\|_1})$. If in addition, (i) for every $a \in \mathbb{R}^k$ such that $\|a_{A_*^c}\|_1 \leq 3\|a_{A_*}\|_1$, it holds $\|a_{A_*}\|_1^2 \leq \frac{(a'\hat{\mathcal{G}}a)S_*}{\underline{\kappa}_n}$, where $\underline{\kappa}_n := \underline{\kappa}_n(\Gamma_k, \lambda_1)$ is a sequence of positive numbers; (ii) $\lambda_2 \geq 2\lambda_0$. Then $\|\widetilde{\alpha} - \alpha_*\|_n = O_p\left(\Delta_n \sqrt{\frac{S_*}{\underline{\kappa}_n}}\right)$, $\|\widetilde{a} - a_*\|_1 = O_p\left(\Delta_n \frac{S_*}{\underline{\kappa}_n}\right)$.*

Theorem 5.2 characterizes the convergence rate of $\widetilde{\alpha}$ and provides low level support for Theorem 5.1 when $\hat{\theta}_{DR}$ is constructed with $\widetilde{\alpha}$. In particular, consistency of $\widetilde{\alpha}$ does not require sparsity condition as long as $\Delta_n \|a_*\|_1 \to 0$, which gives a convergence rate of at least $O_p(\sqrt{\Delta_n \|a_*\|_1})$ in empirical L^2 norm. If additional conditions (i)(ii) in Theorem 5.2 are satisfied, then $\widetilde{\alpha}$ admits faster convergence rates of $O_p\left(\Delta_n \sqrt{\frac{S_*}{\underline{\kappa}_n}}\right)$ in empirical L^2 norm and $O_p\left(\Delta_n \frac{S_*}{\underline{\kappa}_n}\right)$ for l_1 norm, respectively. The magnitude of ε_n^m can be well studied by Bernstein or Hoeffding inequalities. See Lemma D.4 for details. The rate of ε_n^Γ shall be studied on a case by case basis. If $\Gamma_k = I$, $\varepsilon_n^\Gamma = \|a_*\|_\infty$; If $\Gamma_k = \hat{G}$, $\varepsilon_n^\Gamma = O_p(1)$ by Lemma D.5.

Because of the additional Tikhonov penalty, $\widetilde{\alpha}$ can converge faster than a pure lasso estimator ($\lambda_1 = 0$). This is due to condition (ii) in Theorem 5.2, which is a modified compatibility condition in Van de Geer (2007); Van De Geer et al. (2009).[20] Compatibility number $\underline{\kappa}_n$ affects the performance of $\widetilde{\alpha}$, as a smaller $\underline{\kappa}_n$ leads to slower convergence. Because of the additional penalty $\lambda_1 a'\Gamma_k a$, the compatibility number of $\widetilde{\alpha}$ is always larger than its pure lasso counterpart. Hence, $\widetilde{\alpha}$ could perform better with a positive λ_1 if $\underline{\kappa}_{0,n} := \underline{\kappa}_n(\Gamma_k, 0)$ is very small. Let $\widetilde{\alpha}_0 := \widetilde{a}_0' p$ be the solution of (5.1) when $\lambda_1 = 0$ and $\Delta_{0,n} := \varepsilon_n^m + \sqrt{\frac{\log k}{n}}\Lambda_n + \mu_*$.

Case 1: $\underline{\kappa}_{0,n}$ is bounded away from zero. Then $\|\widetilde{\alpha}_0 - \alpha_*\|_n = O_p\left(\Delta_{0,n}\sqrt{S_*}\right)$. If λ_1 is small enough so that $\lambda_1 \varepsilon_n^\Gamma$ is negligible compared to leading term $\Delta_{0,n}$, we have $\|\widetilde{\alpha} - \alpha_*\|_n = O_p\left(\Delta_{0,n}\sqrt{S_*}\right)$ as well. Thus $\widetilde{\alpha}$ is as good as $\widetilde{\alpha}_0$ asymptotically.

[20] Intuitively we can interpret $\underline{\kappa}_n$ as the restricted minimum eigenvalue, cf. Bickel et al. (2009). For discussions on compatibility and restricted eigenvalues of matrices, see Bühlmann and Van De Geer (2011, Sections 6.12 and 6.13).

Case 2: $\underline{\kappa}_{0,n}$ is diminishing. Then the convergence rate of $\widetilde{\alpha}_0$ is

$$\|\widetilde{\alpha}_0 - \alpha_*\|_n = O_p\left(\Delta_{0,n}\sqrt{\frac{S_*}{\underline{\kappa}_{0,n}}}\right), \quad \|\widetilde{a}_0 - a_*\|_1 = O_p\left(\frac{\Delta_{0,n}S_*}{\underline{\kappa}_{0,n}}\right), \tag{5.2}$$

slower than case 1. Note $\|\widetilde{a}_0 - a_*\|_1$ might not even converge if $\underline{\kappa}_{0,n} \to 0$ fast, for example, at rate $O(\Delta_{0,n}S_*)$. If this happens, it pays to have a larger $\lambda_1 > 0$. Since $\lambda_1\Gamma_k$ is positive semidefinite, $\underline{\kappa}_n(\Gamma_k, \lambda_1) \geq \underline{\kappa}_{0,n}$. If λ_1 is large enough so that $\lambda_1\varepsilon_n^\Gamma$ dominates $\Delta_{0,n}$, it follows

$$\|\widetilde{\alpha} - \alpha_*\|_n = O_p\left(\lambda_1\varepsilon_n^\Gamma\sqrt{\frac{S_*}{\underline{\kappa}_n(\Gamma_k, \lambda_1)}}\right), \quad \|\widetilde{a} - a_*\|_1 = O_p\left(\frac{\lambda_1\varepsilon_n^\Gamma S_*}{\underline{\kappa}_n(\Gamma_k, \lambda_1)}\right). \tag{5.3}$$

(5.3) is slower than the rate derived in case 1 but improves (5.2) in terms of l_1 norm if $\frac{\underline{\kappa}_n(\Gamma_k,\lambda_1)}{\underline{\kappa}_{0,n}} > \frac{\lambda_1\varepsilon_n^\Gamma}{\Delta_{0,n}}$ and in terms of empirical L^2 norm if $\frac{\underline{\kappa}_n(\Gamma_k,\lambda_1)}{\underline{\kappa}_{0,n}} > \left(\frac{\lambda_1\varepsilon_n^\Gamma}{\Delta_{0,n}}\right)^2$.

6 Application: electoral accountability and corruption

This section applies our method to the work of Ferraz and Finan (2011) that studies the effect of electoral accountability on corruption. They collect a municipality-level dataset from a Brazilian anti-corruption campaign where 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, the share of resources involving corruption is significantly lower than in municipalities with second-term mayors.

Within this context, the objective of this exercise is to investigate the performance of the near optimal BP estimator (and its high dimensional variant) with OLS and other popular methods in the literature. I find the main conclusion of Ferraz and Finan (2011) very robust. However, OLS estimates change considerably when more controls are sequentially added to the regression. Such coefficient instability is commonly interpreted as a bias correction (i.e., alleviated omitted variable bias by adding more confounders). On the other hand, for the same dataset near optimal BP estimator performs stably. This implies coefficient instability is also likely associated with sub optimal control of mean square remainders, especially in the presence of many controls. By controlling remainders near optimally, our method should be more robust. The estimates of other off-the-shelf shrinkage methods, including DR estimators with lasso selection, are less volatile than OLS. But when there are many technical controls, performance of DR estimators with (post) lasso selected propensity scores appears less satisfactory.

6.1 Main empirical framework

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

$$Y_i = \theta_0 T_i + X_i'\beta + Z_i'\gamma + \varepsilon_i, \tag{6.1}$$

where $T_i = 1$ if mayor i's term limit is not binding (with reelection incentives), and $T_i = 0$ if the mayor's term limit is binding (without reelection incentives); Y_i stands for the share of resources related to corrupt activities in the mayor's municipality; θ_0 is the object of interest[21]; ε_i is the error term; X_i and Z_i are controls of municipal and mayor characteristics, respectively. We deviate from (6.1) and adopt a more flexible semiparametric framework. Denote $Y_i(1)$ as the level of corruption when $T_i = 1$, and denote $Y_i(0)$ as the level of corruption when $T_i = 0$. The object of interest is then defined as $\theta_0 := \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]$, the expected effect of reelection incentives on corruption. Example 2.1 applies. Under conditional independence and overlap assumptions, θ_0 is identified as

$$\theta_0 = \mathbb{E}\left[\mathbb{E}[Y_i|X_i, Z_i, T_i = 1]\right] - \mathbb{E}[\mathbb{E}[Y_i|X_i, Z_i, T_i = 0]]\,. \tag{6.2}$$

Similar to (2.6), RR can be found for $\mathbb{E}[Y_i(1)]$ and $\mathbb{E}[Y_i(0)]$, respectively.

6.2 Baseline results: near optimal BP with fixed penalty

As one of their primary empirical analyses, Ferraz and Finan (2011) explore how estimate of θ_0 in (6.1) changes when different sets of controls are sequentially included in the specification. They start with a plain vanilla mean comparison, and sequentially add five sets of relevant controls. A full specification includes a total of 67 regressors versus a sample size of 476. This is moderately high dimensional with $\frac{k}{n} = 0.14$. Table 1 compares estimates from OLS, near optimal BP estimator with fixed penalty $\lambda_1 = 0.001$ and eight other popular methods in the literature.[22]

From Table 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% points more resources. As we add more regressors, magnitude of the estimate gradually increases. Once all 67 regressors are included, the point estimate becomes -0.0275, an increase of almost 50%. On the other hand, near optimal BP estimator produces quantitatively stable estimates at around -0.018 throughout

[21] It can be interpreted as the average treatment effect of reelection incentives on corruption if individual treatment effect is a constant, see Angrist (1998).

[22] Specifically, controlled ridge regression based on (6.1) with fixed penalty 0.001 and penalty selected by cross validation, respectively; DP method based on (6.2) where conditional expectation function is estimated by ridge with fixed penalty 0.001 and cross validation, respectively; linear partialling out based on (6.1) with post lasso selection; double selection for (6.1) with post lasso selection; doubly robust methods based on (6.2) with lasso and post lasso selected propensity scores, respectively.

six specifications, and all of them are statistically significant at at least 10% level. The majority of the other off-the-shelf shrinkage methods do not perform as stably as the near optimal BP estimator, except the DP approach with cross validated ridge. In Supplementary Materials S4, we also conduct robustness checks by using alternative measures of corruption, and by controlling ability and experience.

6.3 Sensitivity analysis with λ_1 optimally selected

To optimally choose λ_1 for the near optimal BP estimator, we need to know the variance-size ratio $\varrho = \frac{\sigma^2}{nb^2}$ for both $\mathbb{E}[Y_i|X_i, Z_i, T_i = 1]$ and $\mathbb{E}[Y_i|X_i, Z_i, T_i = 0]$. Since n is known, we can gauge the magnitude of ϱ by $\hat{\sigma}^2$ and $\hat{b}^2$, where $\hat{b}$ is the norm of the estimated coefficient of the conditional expectation function, and $\hat{\sigma}^2$ is the empirical average of the associated square residuals. In a full specification with 67 regressors, $\hat{\sigma}_1^2/\hat{b}_1^2 = 0.021$ for $\mathbb{E}[Y_i|X_i, Z_i, T_i = 1]$ and $\hat{\sigma}_0^2/\hat{b}_0^2 = 0.019$ for $\mathbb{E}[Y_i|X_i, Z_i, T_i = 0]$. Therefore, as a sensitivity analysis, we implement the optimal penalty selection procedure in Section 3.1 against a range of $\sigma^2/b^2 \in (0, 100)$, assuming this ratio is the same for $\mathbb{E}[Y_i|X_i, Z_i, T_i = 1]$ and $\mathbb{E}[Y_i|X_i, Z_i, T_i = 0]$. Results are illustrated in Figure 6.1.

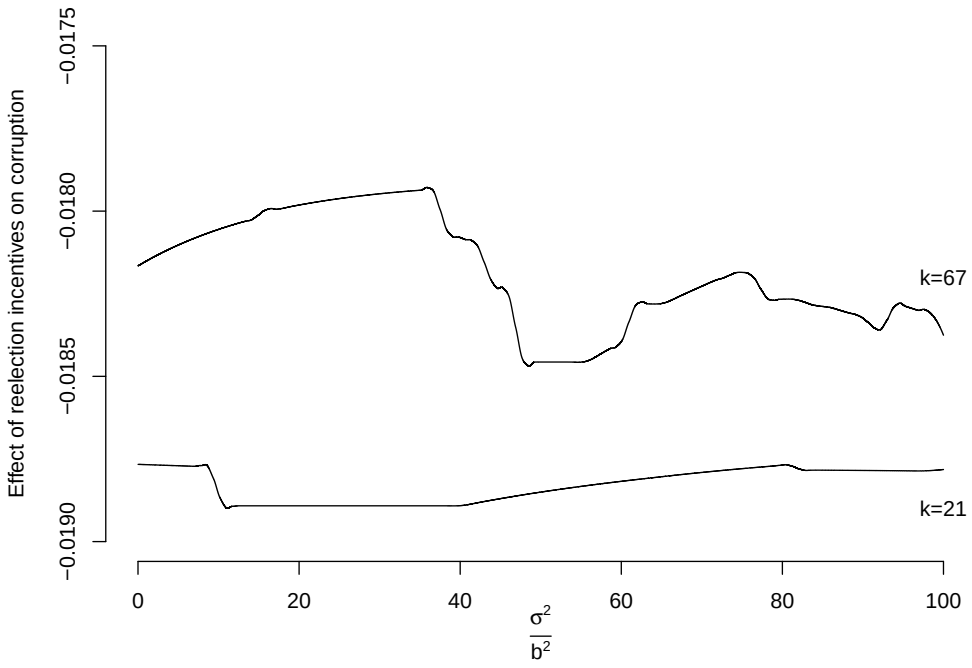


Figure 6.1: Near optimal BP estimator with λ_1 optimally selected against a range of $\frac{\sigma^2}{b^2} \in (0, 100)$. The first line at the above is when all controls are added with $k = 67$. The second line below is when only mayor characteristics are included with $k = 21$. Both curves are smoothed with local polynomials.

6.4 Accounting for many more controls

Finally we explore a scenario with many more technical controls. B-splines are first created based on 11 continuous regressors in Table 1. By adding interaction and second order terms, we get four specifications with number of controls ranging from 67 to 254.[23] We construct $\hat{\theta}_{DR}$ with $\hat{\gamma}$ estimated using "rlasso" in "hdm" R package, and $\widetilde{\alpha}$ calibrated with $\mathbf{\Gamma}_k = \hat{G}$, λ_1 fixed and λ_2 selected by Algorithm 2 in Appendix D.3.[24] Table 2 compares our estimates with four other popular high dimensional methods involving (post) lasso selected propensity scores.[25]

We find $\hat{\theta}_{DR}$ behaves very well. The estimates are stable and significant across the four high dimensional specifications. The choice of λ_1 has a small impact on point estimates. Doubly robust estimators with (post) lasso selected propensity scores do not behave well when k becomes too large. This might signal erratic behavior of the inverse of estimated propensity score with selection under high dimensions. Linear partialling out and linear double selection methods on the other hand, behave relatively stably and produce significant estimates throughout four specifications. 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 conclusion stays robust.

[23] See footnote of Table 2 for details on how these controls are constructed.

[24] Since in this exercise estimated effect barely changes as λ_1 changes, I decided not to report results when λ_1 is optimally chosen via Algorithm 1.

[25] See also Table S5 for the performance of the near optimal BP estimator with a series of small penalties.

Table 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
Controlled 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)
Near optimal BP	Effect	-0.0187**	-0.0186**	-0.0162*	-0.0182*	-0.0182*	-0.0182**
($\lambda = 0.001$)	S.E.	(0.0094)	(0.0087)	(0.0088)	(0.0097)	(0.0095)	(0.0092)
Controlled ridge	Effect		-0.0195**	-0.0197**	-0.0233**	-0.0256**	-0.0263**
w. ridge, $\lambda = 0.001$	S.E.		(0.0096)	(0.0099)	(0.0108)	(0.0106)	(0.0113)
Controlled ridge	Effect		-0.0070	-0.0078	-0.0010	-0.0076	-0.0053
w. ridge, 10 fold CV	S.E.		(0.0097)	(0.0100)	(0.0110)	(0.0109)	(0.0119)
Plug-in ridge	Effect		-0.0186**	-0.0191**	-0.0235**	-0.0250**	-0.0272***
$\lambda = 0.001$	S.E.		(0.0089)	(0.0095)	(0.0102)	(0.0101)	(0.0101)
Plug-in ridge	Effect		-0.0188**	-0.0188*	-0.0188*	-0.0188*	-0.0188*
10 fold CV	S.E.		(0.0092)	(0.0098)	(0.0107)	(0.0108)	(0.0113)
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 is 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). Ridge methods use R package "glmnet"; Four lasso based methods use R package "hdm". For controlled (cross validated) ridge regression, standard error is calculated using sandwich formula with residuals derived from ridge regression; For plug-in (cross validated) ridge, standard error is calculated with α_0 estimated by the estimator in Newey and Robins (2018).

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

† propensity score.

Table 2: Effect of reelection incentives on corruption: many technical controls

Specification		(1)	(2)	(3)	(4)
k		67	122	188	254
n		476	476	476	476
$\hat{\theta}_{DR}$ w. rlasso	Effect	-0.0187**	-0.0195**	-0.0201**	-0.0176*
and $\widetilde{\alpha}(\lambda_1 = 0)$	S.E.	(0.0090)	(0.0089)	(0.0090)	(0.0091)
$\hat{\theta}_{DR}$ w. rlasso	Effect	-0.0187**	-0.0196**	-0.0201**	-0.0176**
and $\widetilde{\alpha}(\lambda_1 = 0.03)$	S.E.	(0.0089)	(0.0087)	(0.0087)	(0.0088)
$\hat{\theta}_{DR}$ w. rlasso	Effect	-0.0187**	-0.0196**	-0.0202**	-0.0178**
$\widetilde{\alpha}(\lambda_1 = 0.06)$	S.E.	(0.0085)	(0.0084)	(0.0085)	(0.0086)
$\hat{\theta}_{DR}$ w. rlasso	Effect	-0.0187**	-0.0196**	-0.0202**	-0.0178**
$\widetilde{\alpha}(\lambda_1 = 0.1)$	S.E.	(0.0082)	(0.0081)	(0.0082)	(0.0083)
Doubly robust	Effect	-0.0214*	-0.0225*	0.0409	0.0012
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**	-0.0192*
post lasso selection	S.E.	(0.0096)	(0.0095)	(0.0096)	(0.0097)
linear double selection	Effect	-0.0224*	-0.0221**	-0.0205*	-0.0197**
post lasso selection	S.E.	(0.0105)	(0.0104)	(0.0106)	(0.0106)

Note: k is 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 1 are directly used in all specifications. In addition, we collect all 11 continuous regressors used in specification (6) from Table 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. $\widetilde{\alpha}$ is calculated with fixed λ_1 and λ_2 selected by Algorithm 2. Consistent with other l_1 penalization methods, we do not include the intercept for the l_1 penalization. Standard errors of $\hat{\theta}_{DR}$ are calculated using simple plug-in method.
*** Significant at 1%.
** Significant at 5 %.
* Significant at 10%.
† propensity score.

A Preliminary Results

A.1 Overall notations

This paper works with triangular array $\{(Y_i, X_i')'\}_{i=1}^n$ generated from probability measure $\mathbb{P} := \mathbb{P}_n$ indexed by sample size n. Thus $\mathbb{E}[\cdot] = \mathbb{E}_{\mathbb{P}}[\cdot]$ is the expectation operator under $\mathbb{P}$. For a vector $a = (a_1, a_2 \cdots, a_k)' \in \mathbb{R}^k$, $m(x, a) = [m(x, a_1), m(x, a_2) \cdots, m(x, a_k)]'$ is a $k-$dimensional column vector. Let $\|a\| := (\sum_{j=1}^k a_j^2)^{1/2}$, $\|a\|_1 := \sum_{j=1}^k |a_j|$ and $\|a\|_\infty := \max_{1\le j\le k} |a_j|$ denote the l_2, l_1 and sup norms of vector a, respectively. For a function $f : \mathcal{X} \mapsto \mathbb{R}$, 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, $\|f\|_{\mathbb{P},\infty} := \sup_{x\in\mathcal{X}} |f(x)|$. For a generic function f, denote $\mathbb{E}_n[f] := \mathbb{E}_n[f(W)] := \frac{1}{n}\sum_{i=1}^n [f(W_i)]$, and $\|f\|_n := \{\mathbb{E}_n[f]^2\}^{1/2}$. 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 $\|A\|_{\max} := \max_{1\le i,j\le k} |a_{ij}|$, $\|A\|_\infty := \max_{1\le i\le k} \sum_{j=1}^k |a_{ij}|$. For a set A, $|A|$ denotes its cardinality. $\mathbf{1}\{\cdot\}$ is the indicator function. 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 $a_n \le c b_n$ for some constant c that does not depend on n. Bold $\mathbf{0}$ denotes a k dimensional vector of 0s.

A.2 Additional examples

Example A.1. Regression discontinuity design away from cut-off

Slightly modify Example 2.1 but keep notation (Y_i, T_i, X_i). In addition, researchers understand that T_i is determined by a running variable $R_i \in \mathbb{R}$ at cut-off point 0: $t := \mathbf{1}\{r \ge 0\}$ for each realization t of T_i and r of R_i. Fix a known boundary point $b > 0$. Object of interest is defined as

$$\theta_b := \mathbb{E}[Y_i^* | -b \le R_i \le b].$$

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_b is by assuming Y_i^* and R_i are independent conditional on X_i and $-b \le R_i \le b$, similar to Angrist and Rokkanen (2015). Then it can be shown

$$\theta_b = \mathbb{E}[\gamma_b(X_i, 0) | -b \le R_i \le b], \tag{A.1}$$

where $\gamma_b(x, 0) = \mathbb{E}[Y_i | X_i = x, 0 \le R_i \le b]$. RR is found in a fashion similar to (2.6) under

suitable overlap assumption

$$\alpha_b(x, r) = \omega_b(x)\mathbf{1}\{r \geq 0\}, \tag{A.2}$$

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

Example A.2. Measurement error with auxiliary data

This example is inspired by Chen et al. (2005). To simplify presentation suppose we are interested in

$$\theta_0 := \mathbb{E}[X_i^*], \quad i = 1 \dots n,$$

where the latent variable X_i^* is not directly observable. However, we have access to a primary dataset of random variable $\{X_i\}_{i=1}^n$ (possibly mismeasured) and an auxiliary dataset of random variables $\{X_{Ai}^*, X_{Ai}\}_{i=1}^n$. 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

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

where $\gamma_0(x) := \mathbb{E}[X_i^*|X_i = x]$ and $\gamma_0^A(x) := \mathbb{E}_A[X_{Ai}^*|X_{Ai} = x]$, with $\mathbb{E}_A[\cdot]$ denoting the expectation operator for auxiliary dataset. Let f_X and f_{X_A} be the marginal densities of X and X_A, respectively. We can further write $\theta_0 = \mathbb{E}_A\left[\gamma_0^A(X_i)\frac{f_X(X_i)}{f_{X_A}(X_i)}\right]$, so that RR is identified as $\alpha_0(x) = \frac{f_X(x)}{f_{X_A}(x)}$.

A.3 Proof of Proposition 3.1

Let (I) $= \sup_{\gamma_0 \in \mathcal{H}_1} R_1^2(\alpha, \gamma_0)$, (II) $= \|\mathbb{E}_n[m(X, p(X)) - \alpha(X)p(X)]\|^2$. By linearity of $m(x, \cdot)$, Cauchy-Schwarz inequality and definition of $\mathcal{H}_1$

$$\begin{aligned}(\text{I}) &= \sup_{\gamma_0 \in \mathcal{H}_1}\{\mathbb{E}_n[\alpha(X)\gamma_0(X) - m(X, \gamma_0(X))]\}^2 \leq \sup_{\|\beta\| \leq 1} \|\beta\|^2 \|\mathbb{E}_n[\alpha(X)p(X) - m(X, p(X))]\|^2 \\ &\leq \|\mathbb{E}_n[\alpha(X)p(X) - m(X, p(X))]\|^2 = (\text{II}).\end{aligned} \tag{A.3}$$

Next, let $\mathcal{E}_{\alpha,n} = \mathbb{E}_n[\alpha(X)p(X) - m(X, p(X))]$. Then

$$(\text{I}) = \sup_{\|\beta\| \leq 1} \beta' \mathcal{E}_{\alpha,n}\mathcal{E}_{\alpha,n}'\beta \geq \sup_{\|\beta\| = 1} \beta' \mathcal{E}_{\alpha,n}\mathcal{E}_{\alpha,n}'\beta = \lambda_{\max}(\mathcal{E}_{\alpha,n}\mathcal{E}_{\alpha,n}') \geq \|\mathcal{E}_{\alpha,n}\|^2 = (\text{II}), \tag{A.4}$$

where the last inequality follows since $\|\mathcal{E}_{\alpha,n}\|^2$ is one of the eigenvalues of $\mathcal{E}_{\alpha,n}\mathcal{E}_{\alpha,n}'$. To see this, suppose $\mathcal{E}_{\alpha,n}\mathcal{E}_{\alpha,n}'\mathbf{v} = \lambda\mathbf{v}$ for some $\lambda \geq$ and $\mathbf{v} \in \mathbb{R}^k$. Premultiplying both sides

by $\mathcal{E}_{\alpha,n}$ yields $\left(\|\mathcal{E}_{\alpha,n}\|^2 - \lambda\right)\mathcal{E}'_{\alpha,n}\mathbf{v} = 0$. Therefore $\|\mathcal{E}_{\alpha,n}\|^2$ must be one of the eigenvalues of $\mathcal{E}_{\alpha,n}\mathcal{E}'_{\alpha,n}$. Combining (A.3) and (A.4) yields the conclusion.

A.4 Proof of Proposition 3.2

Note $b_\gamma < b$ implies $RE^2 > \frac{b_\gamma^2}{b^2}\tilde{\mathcal{R}}$, where $\tilde{\mathcal{R}} = \frac{(1)}{(2)}$, $(1) = \sup_{\gamma_0\in\mathcal{H}_1} R_1^2(\tilde{\alpha}_{\lambda^*_{1,\gamma}}, \gamma_0) + \frac{2\sigma^2}{b_\gamma^2 n}\mathbb{E}_n[\tilde{\alpha}^2_{\lambda^*_{1,\gamma}}(X)]$, and $(2) = \sup_{\gamma_0\in\mathcal{H}_1} R_1^2(\tilde{\alpha}_{\lambda^*_1}, \gamma_0) + \frac{2\sigma^2}{b^2 n}\mathbb{E}_n[\tilde{\alpha}^2_{\lambda^*_1}(X)]$. And $b_\gamma < b$ also implies for each λ_1, $\sup_{\gamma_0\in\mathcal{H}_1} R_1^2(\tilde{\alpha}_{\lambda_1}, \gamma_0) + \frac{2\sigma^2}{b_\gamma^2 n}\mathbb{E}_n[\tilde{\alpha}^2_{\lambda_1}(X)] > \sup_{\gamma_0\in\mathcal{H}_1} R_1^2(\tilde{\alpha}_{\lambda_1}, \gamma_0) + \frac{2\sigma^2}{b^2 n}\mathbb{E}_n[\tilde{\alpha}^2_{\lambda_1}(X)]$. By definition,

$$\lambda^*_{1,\gamma} = \arg\min_{\lambda_1}\{\sup_{\gamma_0\in\mathcal{H}_1} R_1^2(\tilde{\alpha}_{\lambda_1}, \gamma_0) + \frac{2\sigma^2}{b_\gamma^2 n}\mathbb{E}_n[\tilde{\alpha}^2_{\lambda_1}(X)]\},$$
$$\lambda^*_1 = \arg\min_{\lambda_1}\{\sup_{\gamma_0\in\mathcal{H}_1} R_1^2(\tilde{\alpha}_{\lambda_1}, \gamma_0) + \frac{2\sigma^2}{b^2 n}\mathbb{E}_n[\tilde{\alpha}^2_{\lambda_1}(X)]\}.$$

It follows $\tilde{\mathcal{R}} > 1$ and $RE > \frac{b_\gamma}{b}$. If $b_\gamma \geq b$, $RE \geq 1$ follows from $\overline{MMSR}_{\mathcal{H}_b}(\theta_{DR}(\tilde{\alpha}_{\lambda_1}, \gamma)) \geq \overline{MMSR}_{\mathcal{H}_b}(\theta_{BP}(\tilde{\alpha}_{\lambda_1}))$ for each λ_1. If $\hat{G} = I$, the lower bound can be deduced directly since first order conditions yield $\lambda^*_1 = \frac{2\sigma^2}{b^2 n}$ and $\lambda^*_{1,\gamma} = \frac{2\sigma^2}{b_\gamma^2 n}$.

A.5 Optimal estimator when α_0 is known

If α_0 is known, consider the following problem

$$\tilde{\alpha} := \tilde{\alpha}_{\lambda_1} = \arg\min_{\alpha\in\Theta_n}\left\{\ \sup_{\gamma_0\in\mathcal{H}_1} R_1^2(\alpha, \gamma_0)\ \ +\ \ \lambda_1\mathbb{E}_n[(\alpha(X) - \alpha_0(X))^2]\ \right\},$$

which has an analytic solution: $\tilde{\alpha} = \tilde{a}'p$, where $\tilde{a} = (\hat{G}\hat{G} + \lambda_1\hat{G})^-\left\{\hat{G}\hat{P} + \lambda_1\mathbb{E}_n[\alpha_0(X)p(X)]\right\}$. This solution traces out the bias-variance frontier of $MMSR_{\mathcal{H}_b}(\theta_{BP}(\alpha))$ when α_0 is known. The optimal BP estimator can be derived accordingly by selecting an λ_1 that minimizes $MMSR_{\mathcal{H}_b}(\theta_{BP}(\tilde{\alpha}))$. Properties of this estimator can be established analogously and we leave this for future research.

B Basic lemmas

This section provides some basic tools regarding series approximation, minimax calibration, and asymptotic distributions of BP estimators. Recall $\gamma_0 = \mathcal{L}_n\gamma_0 + u_{\gamma_0}$, where $\mathcal{L}_n\gamma_0 = \beta_l'p$ is the least square projection of γ_0 onto Θ_n, β_l is the coefficient and u_{γ_0} is the projection error. Similarly write $\alpha_0 = \mathcal{L}_n\alpha_0 + u_{\alpha_0}$, where $\mathcal{L}_n\alpha_0 = a_l'p$ is the least square projection of α_0 onto Θ_n, a_l is the projection coefficient and u_{α_0} is the projection error. To simplify presentation, let $u_{\gamma_0 i} := \gamma_{0i} - \beta_l'p_i$, $u_{\alpha_0 i} := \alpha_{0i} - a_l'p_i$, where $p_i := p(X_i)$, $\gamma_{0i} := \gamma_0(X_i)$, $\alpha_{0i} := \alpha_0(X_i)$, $i = 1\ldots n$.

Lemma B.1. *If Assumptions O and L hold true, it holds that*

(i) $\mathbb{E}[u_{\alpha_0 i} p_i] = \mathbf{0}$, $\mathbb{E}[u_{\gamma_0 i} p_i] = \mathbf{0}$;

(ii) $\|u_{\alpha_0}\|_{\mathbb{P},2} \leq \mathbf{r}_{\alpha_0}$, $\|u_{\gamma_0}\|_{\mathbb{P},2} \leq \mathbf{r}_{\gamma_0}$;

(iii) $\|u_{\alpha_0}\|_{\mathbb{P},\infty} \leq (\ell_k + 1)\mathbf{r}_{\alpha_0}$, $\|u_{\gamma_0}\|_{\mathbb{P},\infty} \leq (\ell_k + 1)\mathbf{r}_{\gamma_0}$;

(iv) $a_l = O(1 + \mathbf{r}_{\alpha_0})$, $\beta_l = O(1 + \mathbf{r}_{\gamma_0})$.

Proof. We only prove results related to α_0. Those related to γ_0 can be shown in the same fashion. By definition

$$a_l = \arg\min_{a \in \mathbb{R}^k} \mathbb{E}[\alpha_{0i} - a' p_i]^2. \tag{B.1}$$

(i) follows from the first order condition of a_l. (ii) directly follows from (B.1): $\|u_{\alpha_0}\|_{\mathbb{P},2} = \mathbb{E}[u_{\alpha_0 i}^2] \leq \mathbb{E}[(\alpha_{0i} - a_b' p_i)^2] \leq \mathbf{r}_{\alpha_0}^2$. For (iii), note $u_{\alpha_0} = \alpha_0 - a_b' p + a_b' p - a_l' p$, where

$$\begin{aligned} a_b' p - a_l' p &= p' \mathbb{E}[p_i p_i']^{-1} \mathbb{E}[p_i p_i'] a_b - p' \mathbb{E}[p_i p_i']^{-1} \mathbb{E}[p_i \alpha_{0i}] \\ &= p' \mathbb{E}[p_i p_i']^{-1} \mathbb{E}\left[p_i (p_i' a_b - \alpha_{0i})\right] = \mathcal{L}_n (p' a_b - \alpha_0). \end{aligned}$$

Then (iii) follows from triangle inequality and definition of ℓ_k. Finally to see (iv), note

$$\|\mathcal{L}_n \alpha_0\|_{\mathbb{P},2}^2 = a_l' \mathbb{E}[p_i p_i'] a_l \geq \|a_l\|^2 \lambda_{\min}\left\{\mathbb{E}[p_i p_i']\right\}.$$

By Assumption L, $\mathbb{E}[p_i p_i']$ has all eigenvalues bounded away from zero. It follows

$$\|a_l\|^2 \lesssim \|\mathcal{L}_n \alpha_0\|_{\mathbb{P},2}^2 \leq \|\alpha_0\|_{\mathbb{P},2}^2 + \|u_{\alpha_0}\|_{\mathbb{P},2}^2 = O(1 + \mathbf{r}_{\gamma_0}^2),$$

where the second inequality is by triangle inequality, and final relation follows from $\|\alpha_0\|_{\mathbb{P},2} = O(1)$ by Assumption O and $\|u_{\alpha_0}\|_{\mathbb{P},2} \leq \mathbf{r}_{\gamma_0}$ by Lemma B.1(ii). □

The next lemma presents an effective way of controlling $R_1^2(\tilde{\alpha}, \gamma_0)$ when $\hat{G}$ is invertible.

Lemma B.2. *If Assumptions O and L hold true, it holds*

(i) $R_1^2(\tilde{\alpha}, \gamma_0) \lesssim T_1 + T_2$, where

$$T_1 := \left\{\mathbb{E}_n[\tilde{\alpha}(X)\mathcal{L}_n\gamma_0(X) - m(X, \mathcal{L}_n\gamma_0(X)]\right\}^2, \quad T_2 := \left\{\mathbb{E}_n[\tilde{\alpha}(X)u_{\gamma_0} - m(X, u_{\gamma_0})]\right\}^2.$$

(ii) $T_2 = \left\{\mathbb{E}_n\left[\tilde{\alpha}(X) - \alpha_0(X))u_{\gamma_0}\right]\right\}^2 + O_p\left[\left(\ell_k^2 \wedge \|\alpha_0\|_{\mathbb{P},\infty}^2\right)\mathbf{r}_{\gamma_0}^2/n\right]$.

(iii) For every $\alpha \in \Theta_n$,

$$T_1 \lesssim \left\{\|\mathbb{E}_n[m(X, p(X)) - \alpha(X)p(X)]\|^2 + \lambda_1 \mathbb{E}_n \alpha^2(X)\right\} \|\beta_l\|_2^2, \tag{B.2}$$

$$\lambda_1 \mathbb{E}_n \tilde{\alpha}^2(X) \leq \|\mathbb{E}_n[m(X, p(X)) - \alpha(X)p(X)]\|^2 + \lambda_1 \mathbb{E}_n \alpha^2(X). \tag{B.3}$$

Proof. Statement (i) follows from linearity of $m(x,\cdot)$ and triangle inequality. Note $T_2 \lesssim T_{21}+T_{22}$, where $T_{21} := \{\mathbb{E}_n[(\tilde{\alpha}(X)-\alpha_0(X))\,u_{\gamma_0}]\}^2$ and $T_{22} := \{\mathbb{E}_n\left[\alpha_0(X)u_{\gamma_0} - m(X,u_{\gamma_0})\right]\}^2$. By Assumption O and note $\mathbb{E}[\alpha_{0i}u_{\gamma_0 i} - m(X_i,u_{\gamma_0 i})] = 0$:

$$\mathbb{E}T_{22} = \frac{1}{n}\mathbb{E}\left[\alpha_{0i}u_{\gamma_0 i} - m(X_i,u_{\gamma_0 i})\right]^2 \lesssim \frac{1}{n}\mathbb{E}\left[\alpha_{0i}u_{\gamma_0 i}\right]^2 + \frac{1}{n}\mathbb{E}[m^2(X_i,u_{\gamma_0 i})] \lesssim \frac{1}{n}\mathbb{E}\left[\alpha_{0i}^2 u_{\gamma_0 i}^2\right],$$

where either $\mathbb{E}\left[\alpha_{0i}^2u_{\gamma_0 i}^2\right] \lesssim \|u_{\gamma_0}\|_{\mathbb{P},\infty}^2 \lesssim \ell_k^2\mathbf{r}_{\gamma_0}^2$ by $\mathbb{E}\left[\alpha_{0i}^2\right] < \infty$ and Lemma B.1(iii), or $\mathbb{E}\left[\alpha_{0i}^2u_{\gamma_0 i}^2\right] \le \|\alpha_0\|_{\mathbb{P},\infty}^2\|u_{\gamma_0}\|_{\mathbb{P},2}^2 \le \|\alpha_0\|_{\mathbb{P},\infty}^2\mathbf{r}_{\gamma_0}^2$ by Lemma B.1(ii). Then statement (ii) follows by Markov inequality. Finally, (B.2) follows from linearity of $m(x,\cdot)$, Cauchy-Schwarz inequality and definition of $\tilde{\alpha}$. (B.3) follows from definition of $\tilde{\alpha}$ as well. □

Lemma B.3 can be invoked to establish the asymptotic distribution of $\tilde{\theta}_{BP}$ when $\frac{k}{n} \le 1$.

Lemma B.3. *Suppose Assumption O holds true and (i)* $\inf_{x\in\mathcal{X}}\mathbb{E}\left[e_i^2|X_i=x\right]$ *is bounded away from zero,* $\sup_{x\in\mathcal{X}}\mathbb{E}\left[|e_i|^3|X_i=x\right] \lesssim 1$*; (ii)* $\max_i|\tilde{\alpha}(X_i)|/\sqrt{n} = o_p(1)$*; (iii)* $\left[\mathbb{E}_n\tilde{\alpha}^2(X)\right]^{-1} = O_p(1)$*. Then* $\sqrt{n}\mathbb{E}_n[\tilde{\alpha}(X)e+m(X,\gamma_0(X))-\theta_0] \xrightarrow{d} \sigma_{\tilde{\alpha},n}Z_1+\sigma_m Z_2$*, where* $\sigma_{\tilde{\alpha},n}^2 = \mathbb{E}_n\left[\tilde{\alpha}^2(X)\mathbb{E}[e^2|X]\right]$*,* $\sigma_m^2 = \mathbb{E}[m^2(X_i,\gamma_0(X_i))]-\theta_0^2$*, and* Z_1 *and* Z_2 *are two iid standard normal random variables independent of* $\{X_i,e_i\}_{i=1}^n$*.*

Proof. Recall $\mathscr{X}_n := \{X_i\}_{i=1}^n$. Let $\mathcal{U}_i = n^{-1/2}\sigma_{\tilde{\alpha},n}^{-1}\tilde{\alpha}(X_i)e_i$. We split the proof into three steps.

Step 1: show $\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\sum_{i=1}^n\mathcal{U}_i \le t|\mathscr{X}_n\right) - \mathbb{P}(Z_1\le t)\right| = o_p(1)$. Note $\mathbb{E}[\mathcal{U}_i|\mathscr{X}_n] = 0$ for each $i = 1\dots n$ and $\sum_{i=1}^n Var\left(\mathcal{U}_i|\mathscr{X}_n\right) = 1$. Thus conditional on $\mathscr{X}_n$, $\{\mathcal{U}_i\}_{i=1}^n$ are mean zero and independent. It follows:

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\sum_{i=1}^n\mathcal{U}_i \le t|\mathscr{X}_n\right) - \mathbb{P}(Z_1\le t)\right| \lesssim \sum_{i=1}^n\mathbb{E}\left[|\mathcal{U}_i|^3|\mathscr{X}_n\right] \lesssim \sigma_{\tilde{\alpha},n}^{-3}n^{-3/2}\sum_{i=1}^n|\tilde{\alpha}(X_i)|^3$$
$$\lesssim \left[\frac{1}{n}\sum_{i=1}^n\tilde{\alpha}^2(X_i)\right]^{-3/2} n^{-3/2}\sum_{i=1}^n|\tilde{\alpha}(X_i)|^3 = \frac{\max_i|\tilde{\alpha}(X_i)|}{\sqrt{n}}\left[\frac{1}{n}\sum_{i=1}^n\tilde{\alpha}^2(X_i)\right]^{-1/2} = o_p(1),$$

where the first inequality is by Berry-Esseen inequality, the second inequality is by $\sup_{x\in\mathcal{X}}\mathbb{E}\left[|e_i|^3|X_i=x\right] \lesssim 1$, the third inequality is by $\inf_{x\in\mathcal{X}}\mathbb{E}\left[e_i^2|X_i=x\right]$ bounded away from zero, and the final relation uses assumptions (ii) and (iii) in Lemma B.3.

Step 2: show for each $t\in\mathbb{R}$, $\left|\mathbb{P}\left(\sqrt{n}\mathbb{E}_n[\tilde{\alpha}(X)e]\le t|\mathscr{X}_n\right) - \mathbb{P}(\sigma_{\tilde{\alpha},n}Z_1\le t)\right| = o_p(1)$. This follows from $\mathbb{P}\left(\sqrt{n}\mathbb{E}_n[\tilde{\alpha}(X)e]\le t|\mathscr{X}_n\right) = \mathbb{P}\left(\sum_{i=1}^n\mathcal{U}_i \le \sigma_{\tilde{\alpha},n}^{-1}t|\mathscr{X}_n\right)$, $\mathbb{P}(\sigma_{\tilde{\alpha},n}Z_1\le t) = \mathbb{P}(Z_1\le\sigma_{\tilde{\alpha},n}^{-1}t)$ for each $t\in\mathbb{R}$ and conclusion from step 1.

Step 3: show $\mathbb{G}_n \xrightarrow{d} \mathbb{G}^*$, where $\mathbb{G}_n := \sqrt{n}\mathbb{E}_n[\tilde{\alpha}(X)e + m(X,\gamma_0(X)) - \theta_0]$, $\mathbb{G}^* := \sigma_{\tilde{\alpha},n}Z_1 + \sigma_m Z_2$. Let $\phi_{\mathbb{G}_n}(t)$ and $\phi_{\mathbb{G}^*}(t)$ be the characteristic functions of $\mathbb{G}_n$ and $\mathbb{G}^*$,

respectively and $\mathbf{i}^2 = -1$. It suffices to show $\phi_{\mathbb{G}_n}(t) \to \phi_{\mathbb{G}^*}(t)$ for each t. By triangle inequality

$$
\begin{aligned}
&|\phi_{\mathbb{G}_n}(t) - \phi_{\mathbb{G}^*}(t)| = |\mathbb{E}\exp\{\mathbf{i}t\mathbb{G}_n\} - \mathbb{E}\exp\{\mathbf{i}t\mathbb{G}^*\}| \\
\leq &|\mathbb{E}\exp\{\mathbf{i}t\mathbb{G}_n\} - \mathbb{E}\exp\{\mathbf{i}t\{\sigma_{\tilde{\alpha},n}Z_1 + \sqrt{n}\mathbb{E}_n[m(X,\gamma_0(X)) - \theta_0]\}\}| \qquad \text{(B.4)} \\
+ &|\mathbb{E}\exp\{\mathbf{i}t\{\sigma_{\tilde{\alpha},n}Z_1 + \sqrt{n}\mathbb{E}_n[m(X,\gamma_0(X)) - \theta_0]\}\} - \mathbb{E}\exp\{\mathbf{i}t\mathbb{G}^*\}|. \qquad \text{(B.5)}
\end{aligned}
$$

Note

$$
\begin{aligned}
(B.4) = &\left|\mathbb{E}\exp\left\{\mathbf{i}t\left(\sqrt{n}\mathbb{E}_n[m(X,\gamma_0(X)) - \theta_0]\right)\right\}\left\{\exp\left\{\mathbf{i}t\sqrt{n}\mathbb{E}_n\tilde{\alpha}(X)e\right\} - \exp\{\mathbf{i}t\sigma_{\tilde{\alpha},n}Z_1\}\right\}\right| \\
\leq &\mathbb{E}\left|\exp\left\{\mathbf{i}t\left(\sqrt{n}\mathbb{E}_n[m(X,\gamma_0(X)) - \theta_0]\right)\right\}\right|\left|\mathbb{E}\left[\exp\left\{\mathbf{i}t\sqrt{n}\mathbb{E}_n\tilde{\alpha}(X)e\right\} - \exp\left\{\mathbf{i}t\sigma_{\tilde{\alpha},n}Z_1\right\} \mid \mathscr{X}_n\right]\right| \\
\leq &\mathbb{E}\left|\mathbb{E}\left[\exp\left\{\mathbf{i}t\sqrt{n}\mathbb{E}_n\tilde{\alpha}(X)e\right\} - \exp\left\{\mathbf{i}t\sigma_{\tilde{\alpha},n}Z_1\right\} \mid \mathscr{X}_n\right]\right|,
\end{aligned}
$$

where the first inequality is by LIE and second inequality follows from property of characteristic function. Then by dominated convergence theorem and conclusion from step 2, $(B.4) = o(1)$. Next,

$$
\begin{aligned}
(B.5) &= |\mathbb{E}\exp\left\{\mathbf{i}t\sigma_{\tilde{\alpha},n}Z_1\right\}\left\{\exp\left\{\mathbf{i}t\left(\mathbb{E}_n[m(X,\gamma_0(X)) - \theta_0]\right)\right\} - \exp\left\{\mathbf{i}t\sigma_m Z_2\right\}\right\}| \\
&= \left|\mathbb{E}\left\{\mathbb{E}\left[\exp\left\{\mathbf{i}t\sigma_{\tilde{\alpha},n}Z_1\right\}|\mathscr{X}_n\right]\mathbb{E}\left[\left\{\exp\left\{\mathbf{i}t\left(\sqrt{n}\mathbb{E}_n[m(X,\gamma_0(X)) - \theta_0]\right)\right\} - \exp\left\{\mathbf{i}t\sigma_m Z_2\right\}\right\}|\mathscr{X}_n\right]\right\}\right| \\
&\leq \left|\mathbb{E}\left[\exp\left\{\mathbf{i}t\left(\sqrt{n}\mathbb{E}_n[m(X,\gamma_0(X)) - \theta_0]\right)\right\} - \exp\left\{\mathbf{i}t\sigma_m Z_2\right\}\right]\right| = o(1),
\end{aligned}
$$

where the second relation is by LIE and conditional independence, the third relation is because $\mathbb{E}\left[\exp\left\{\mathbf{i}t\sigma_{\tilde{\alpha},n}Z_1\right\}|\mathscr{X}_n\right] = \exp\left\{-\frac{1}{2}\sigma^2_{\tilde{\alpha},n}t^2\right\} \leq 1$ and final relation follows from $\sqrt{n}\mathbb{E}_n[m(X,\gamma_0(X)) - \theta_0] \xrightarrow{d} \sigma_m Z_2$ by Lindeberg-Lévy central limit theorem. This completes proof for the third step and the conclusion follows. □

C Technical results for Section 4

Additional Notations. Let $e_i^R := m(X_i, p(X_i)) - \alpha_0(X_i)p(X_i)$. Recall $\dot{\alpha} = p'\dot{a}$, $\dot{a} := \hat{G}^- \hat{P}$.

C.1 Proofs for main results in Section 4

Proof of Theorem 4.1

Since $\sqrt{n}\mathbb{E}_n[\tilde{\theta}_{BP} - \theta_0] = \sqrt{n}\mathbb{E}_n[\tilde{\alpha}(X)e + m(X,\gamma_0(X)) - \theta_0] + \sqrt{n}R_1(\tilde{\alpha},\gamma_0)$, display (4.1) follows by Theorems C.1. If in addition, $\sigma^2_{\tilde{\alpha},n} \xrightarrow{p} \mathcal{V} > 0$, display (4.2) follows from (4.1) and Slutsky's theorem.

Proof of Corollary 4.1

By Theorem C.1, $\mathcal{R}_n = \epsilon_n + r_n$ where $r_n = \delta_n \ell_k \mathbf{r}_{\gamma_0} \left(\sqrt{k \log \xi_k} + \frac{k \xi_k \log \xi_k}{\sqrt{n}} \right) \wedge \sqrt{n} \delta_n \mathbf{r}_{\gamma_0} + \sqrt{n} \mathbf{r}_{\gamma_0} \mathbf{r}_{\alpha_0} + (\ell_k \wedge \|\alpha_0\|_{\mathbb{P},\infty}) \mathbf{r}_{\gamma_0} + \delta_n$ and $\epsilon_n = o(r_n)$. On the other hand, let $\|\beta_l - \beta\| = O_p(\delta_n^\beta)$ for some $\delta_n^\beta \geq 0$. Then $\mathcal{R}_n^\gamma = \epsilon_n \delta_n^\beta + r_n$. If $\delta_n^\beta \to 0$, r_n is still the leading term for $\mathcal{R}_n^\gamma$. Thus $\mathcal{R}_n = \mathcal{R}_n^\gamma = r_n$. Otherwise, $\delta_n^\beta \nrightarrow 0$ implies $\mathcal{R}_n^\gamma \geq \mathcal{R}_n$. So in either case, $\mathcal{R}_n^\gamma \geq \mathcal{R}_n$. Finally, if $\mathcal{R}_n \nrightarrow 0$, then $r_n \nrightarrow 0$. Hence $\mathcal{R}_n^\gamma \nrightarrow 0$ as well.

Proof of Corollary 4.2

First note by Tropp (2015, Theorem 5.1.1), $\mathbb{P}\{\lambda_{\min}(\hat{G}) \leq 0.5\lambda_{\min}(G)\} \leq \exp\{\log k[1 - \frac{0.25\lambda_{\min}(G)}{2\xi_k^2 \log k/n}]\} \to 0$ since $\frac{\xi_k^2 \log k}{n} \to 0$ and $\lambda_{\min}(G)$ is bounded away from zero. Thus Assumption M is satisfied. So Lemmas C.1-C.4 and Theorem C.1 still hold true. Second, $\sqrt{n}\mathbb{E}_n \left[\tilde{\theta}_{BP} - \theta_0 \right] = \sqrt{n}\mathbb{E}_n \phi + \sqrt{n} R_1(\tilde{\alpha}, \gamma_0) + \sqrt{n} R_2(\tilde{\alpha})$, where $\sqrt{n} R_1(\tilde{\alpha}, \gamma_0) = o_p(1)$ by Theorem C.1, $\sqrt{n} R_2(\tilde{\alpha}) = o_p(1)$ by Lemma C.3(iii), $\frac{\xi_k^2 \log k}{n} = o(1)$ and $\mathbf{r}_{\alpha_0} = o(1)$. Then $\sqrt{n} \left[\tilde{\theta}_{BP} - \theta_0 \right] \xrightarrow{d} N(0, \mathbb{E}[\phi_i^2])$ by Lindeberg–Lévy central limit theorem. Finally $\hat{\Omega} \xrightarrow{p} \Omega$ follows by Lemma C.5(iv).

Proof of Corollary 4.3

The proof is similar to that of Theorem 4.1 by using conclusions from Theorem C.2, so details are omitted.

C.2 Additional results

Lemma C.1 concerns a basic matrix law of large numbers. Lemmas C.2 and C.3 provide tools to establish asymptotic properties of $\tilde{\alpha}$ and its associated remainder terms. Lemma C.4 is useful for verifying primitive conditions of Lemma B.3.

Lemma C.1. *If Assumptions O, L and M(1) hold true, then* $\mathbb{E}\left[\left\| \hat{G} - G \right\| \right] \lesssim \frac{\xi_k^2 \log k}{n} + \sqrt{\frac{\xi_k^2 \log k}{n}}$, $\left\| \hat{G} - G \right\| = O_p\left(\sqrt{\frac{\xi_k^2 \log k}{n}} \right)$, $\left\| \hat{G} \right\| = O_p(1)$.

Lemma C.2. *If Assumptions O, L and M hold true and* $\lambda_1 = o(\frac{1}{n})$, *then*

(i) $\hat{G}^{-1} \mathbb{E}_n e^R = O_p\left(\sqrt{\frac{\xi_k^2}{n}} \wedge \sqrt{\frac{\|\alpha_0\|_{\mathbb{P},\infty} k}{n}} \right)$;

(ii) $\hat{G}^{-1} \mathbb{E}_n[u_{\alpha_0} p(X)] = O_p(\mathbf{r}_{\alpha_0})$;

(iii) $\|\dot{a} - a_l\| = O_p(\delta_n)$, $\|\dot{a}\| = O_p(1 + \delta_n)$;

(iv) $\|\tilde{a} - a_l\| = O_p(\delta_n)$, $\|\tilde{a}\| = O_p(1 + \delta_n)$;

(v) $\mathbb{E}_n\tilde{\alpha}^2(X) = O_p((1+\delta_n)^2)$, $\mathbb{E}_n\dot{\alpha}^2(X) = O_p((1+\delta_n)^2)$.

Lemma C.3. *If Assumptions O, L and M hold true and* $\lambda_1 = o(\frac{1}{n})$*, then*

(i) $\sqrt{n}\mathbb{E}_n\left[(\tilde{a}-a_l)'p(X)u_{\gamma_0}\right] = O_p\left[\delta_n \mathbf{r}_{\gamma_0}\ell_k\left(\sqrt{k\log\xi_k} + \frac{k\xi_k\log\xi_k}{\sqrt{n}}\right) \wedge \delta_n\mathbf{r}_{\gamma_0}\sqrt{n}\right]$.

(ii) $\sqrt{n}\mathbb{E}_n\left[(\tilde{a}-a_l)'p(X)e\right] = O_p(\delta_n)$;

(iii) $\sqrt{n}R_2(\tilde{\alpha}) = O_p(\delta_n)$.

Lemma C.4. *If Assumptions O, L and M(1) hold true, all eigenvalues of* $\hat{G}$ *are positive wpa1,* $\lambda_1 = o(\frac{1}{n})$ *and* $\|\alpha_0\|^2_{\mathbb{P},2} - \mathbf{r}^2_{\alpha_0}$ is bounded away from zero uniformly for each k and n*, then*

(i) $\left\|\hat{P}\right\|^{-1} = O_p(1)$;

(ii) $\left\{\mathbb{E}_n\left[\tilde{\alpha}^2(X)\right]\right\}^{-1} = O_p(1) + o_p\left(\frac{1}{n^2\lambda_{\min}(\hat{G})}\right)$.

Theorem C.1. *If conditions for Theorem 4.1 hold true, then*

(i) $\sqrt{n}R_1(\tilde{\alpha},\gamma_0) = o_p(1)$;

(ii) $\sqrt{n}\mathbb{E}_n[\tilde{\alpha}(X)e + m(X,\gamma_0(X)) - \theta_0] \xrightarrow{d} \sigma_{\tilde{\alpha},n}Z_1 + \sigma_m Z_2$.

Lemma C.5. *If conditions of Corollary 4.2 hold true, then*

(i) $\mathbb{E}_n[m^2(X,\dot{\gamma}(X)-\gamma_0(X))] = o_p(1)$;

(ii) $\mathbb{E}_n\left[\tilde{\alpha}(X)^2(\dot{\gamma}(X)-\gamma_0(X))^2\right] = o_p(1)$;

(iii) $\mathbb{E}_n\left[(\tilde{\alpha}(X)-\alpha_0(X))^2e^2\right] = o_p(1)$;

(iv) $\hat{\Omega} \xrightarrow{p} \mathbb{E}[\phi_i^2]$.

Lemma C.6. *If conditions of Corollary 4.3 hold true, then*

(i) $\mathbb{E}_n\dot{\alpha}^2(X) = O_p(1)$;

(ii) $\mathbb{E}_n\tilde{\alpha}^2(X) = O_p(1)$;

(iii) $\left[\mathbb{E}_n\tilde{\alpha}^2(X)\right]^{-1} = O_p(1)$.

Theorem C.2. *If conditions of Corollary 4.3 hold true, then*

(i) $\sqrt{n}R_1(\tilde{\alpha},\gamma_0) = o_p(1)$;

(ii) $\sqrt{n}\mathbb{E}_n[\tilde{\alpha}(X)e + m(X, \gamma_0(X)) - \theta_0] \xrightarrow{d} \sigma_{\tilde{\alpha},n} Z_1 + \sigma_m Z_2$.

The following lemmas present some sufficient conditions for main assumptions made in Section 4.

Lemma C.7. *Let $G = I$ and Assumptions O and L hold true. In addition, $\sqrt{\frac{2\xi_k^2 \log 2k}{n}} + \frac{\xi_k^2 \log 2k}{3n} \to c_1$, where c_1 is a constant strictly smaller than 1. Then there exists a constant $c_2 < 1 - c_1$ strictly positive such that $\lambda_{\min}(\hat{G}) \geq c_2$ wpa1.*

Lemma C.8. *If Assumption L(1) holds true and $\|\tilde{a}\| = O_p(1)$, then $\max_i |\tilde{\alpha}(X_i)| / \sqrt{n} = o_p(1)$.*

D Technical results for Section 5

Additional Notations. Let $u_{*i} := \alpha_0(X_i) - \alpha^*(X_i)$, $\alpha^*(X_i) = p'(X_i)a^*$.

D.1 Proofs for main results in Section 5

Proof of Theorem 5.1

Since $\sqrt{n}(\hat{\theta}_{DR} - \theta_0) = \sqrt{n}\mathbb{E}_n\phi + \sqrt{n}R_1(\hat{\alpha}, \gamma_0 - \hat{\gamma}) + \sqrt{n}R_2(\hat{\alpha})$, conclusion follows from Theorem D.1 and Lindeberg–Lévy central limit theorem.

Proof of Theorem 5.2

By Lemma D.3(iii), $\lambda_0 = O_p(\Delta_n)$. If $\lambda_2 = \lambda_0$, Lemma D.3(iv) and triangle inequality yields $(\widetilde{a} - a_*)'\hat{\mathcal{G}}(\widetilde{a} - a_*) \leq 2\lambda_0 \|a_*\|_1$. Since $\hat{\mathcal{G}} - \hat{G}$ is positive semidefinite, first conclusion follows as

$$\mathbb{E}_n[\widetilde{\alpha}(X) - \alpha_*(X)]^2 = (\widetilde{a} - a_*)'\hat{G}(\widetilde{a} - a_*) \leq (\widetilde{a} - a_*)'\hat{\mathcal{G}}(\widetilde{a} - a_*) \leq 2\lambda_0 \|a_*\|_1 .$$

If conditions (i)(ii) of Theorem 5.2 also hold true, then

$$\begin{aligned} 2(\widetilde{a} - a_*)'\hat{\mathcal{G}}(\widetilde{a} - a_*) + \lambda_2 \|\widetilde{a} - a_*\|_1 &= 2(\widetilde{a} - a_*)'\hat{\mathcal{G}}(\widetilde{a} - a_*) + \lambda_2 \|\widetilde{a}_{A_*} - a_{*A_*}\|_1 + \lambda_2 \left\|\widetilde{a}_{A_*^c}\right\|_1 \\ &\leq 3\lambda_2 \|\widetilde{a}_{A_*} - a_{*A_*}\|_1 + \lambda_2 \|\widetilde{a}_{A_*} - a_{*A_*}\|_1 = 4\lambda_2 \|\widetilde{a}_{A_*} - a_{*A_*}\|_1 , \end{aligned}$$

where the second relation follows wpa1 by Theorem D.2. Since $a_{*A_*^c} = \mathbf{0}$, Theorem D.2 also implies that $\left\|\widetilde{a}_{A_*^c} - a_{*A_*^c}\right\|_1 \leq 3 \|\widetilde{a}_{A_*} - a_{*A_*}\|_1$ wpa1. So by condition (i) of Theorem

5.2, it holds wpa1 that

$$2(\widetilde{a}-a_*)'\hat{\mathcal{G}}(\widetilde{a}-a_*)+\lambda_2\left\|\widetilde{a}-a_*\right\|_1 \leq 4\left[(\widetilde{a}-a_*)'\hat{\mathcal{G}}(\widetilde{a}-a_*)\right]^{1/2}\lambda_2\sqrt{\frac{S_*}{\underline{\kappa}_n}} \leq (\widetilde{a}-a_*)'\hat{\mathcal{G}}(\widetilde{a}-a_*)+\frac{4\lambda_2^2 S_*}{\underline{\kappa}_n},$$

where the second inequality is due to $4ab \leq a^2+4b^2$ for any number a and b. Rearranging above inequality yields $(\widetilde{a}-a_*)'\hat{\mathcal{G}}(\widetilde{a}-a_*) = O_p\left(\frac{\lambda_2^2 S_*}{\underline{\kappa}_n}\right)$ and $\left\|\widetilde{a}-a_*\right\|_1 = O_p\left(\frac{\lambda_2 S_*}{\underline{\kappa}_n}\right)$. Since $(\hat{\mathcal{G}}-\hat{G})$ is positive semidefinite it follows $\mathbb{E}_n[\widetilde{\alpha}(X)-\alpha_*(X)]^2 \leq (\widetilde{a}-a_*)'\hat{\mathcal{G}}(\widetilde{a}-a_*) = O_p\left(\frac{\lambda_2^2 S_*}{\underline{\kappa}_n}\right)$ as well.

D.2 Additional results

Lemma D.1 is concerned with matrix convergence in max norm.

Lemma D.1. *If Assumptions O and H hold true and* $\Lambda_n\sqrt{\frac{\log k}{n}} \to 0$, *then*

$$\mathbb{E}\left[\left\|\hat{G}-G\right\|_{\max}\right] = O\left(\Lambda_n\sqrt{\frac{\log k}{n}}\right), \left\|\hat{G}-G\right\|_{\max} = O_p\left(\Lambda_n\sqrt{\frac{\log k}{n}}\right), \left\|\hat{G}\right\|_{\max} = O_p(1).$$

Theorem D.1. *If conditions for Theorem 5.1 hold true,* then

(i) $\sqrt{n}R_1(\hat{\alpha},\gamma_0-\hat{\gamma}) = o_p(1)$;

(ii) $\sqrt{n}R_2(\hat{\alpha}) = o_p(1)$.

Lemma D.2. *If Assumptions O and H hold true and* $\Delta_n = o_p(1)$, *then*

(i) $\left\|\mathbb{E}_n[p(X)u_*]-\mathbb{E}[p(X_i)u_{*i}]\right\|_\infty = O_p\left(\sqrt{\frac{\log k}{n}}\Lambda_n\mu_*\right)$;

(ii) $\left\|\mathbb{E}_n[p(X)\alpha_0(X)]-\mathbb{E}[p(X_i)\alpha_0(X_i)]\right\|_\infty = O_p\left(\sqrt{\frac{\log k}{n}}\Lambda_n\right)$;

(iii) $\left\|\mathbb{E}_n[p(X)\alpha_*(X)]-\mathbb{E}[p(X_i)\alpha_*(X_i)]\right\|_\infty = O_p\left(\sqrt{\frac{\log k}{n}}\Lambda_n\right)$.

Lemma D.3. *If Assumptions O and H hold true and* $\Delta_n = o_p(1)$, then wpa1

(i) $\left\|\mathbb{E}_n[m(X,p(X))-p(X)\alpha_0(X)]\right\|_\infty \lesssim \varepsilon_n^m + \sqrt{\frac{\log k}{n}}\Lambda_n$;

(ii) $\left\|\hat{P}-\hat{G}a_*\right\|_\infty \lesssim \varepsilon_n^m + \sqrt{\frac{\log k}{n}}\Lambda_n + \mu_*$;

(iii) $2(\widetilde{a}-a_*)'(\hat{P}-\hat{\mathcal{G}}a_*) \leq \left\|\widetilde{a}-a_*\right\|_1\lambda_0$, where $\lambda_0 = O_p(\Delta_n)$;

(iv) $(\widetilde{a}-a_*)'\hat{\mathcal{G}}(\widetilde{a}-a_*)+\lambda_2\|\widetilde{a}\|_1 \leq \lambda_0\|\widetilde{a}-a_*\|_1+\lambda_2\|a_*\|_1$.

Theorem D.2. *If all conditions of Theorem 5.2 hold true,* then wpa1 $2(\widetilde{a}-a_*)'\hat{\mathcal{G}}(\widetilde{a}-a_*)+\lambda_2\left\|\widetilde{a}_{A_*^c}\right\|_1 \leq 3\lambda_2\|\widetilde{a}_{A_*}-a_{*A_*}\|_1$.

The following lemmas present some sufficient conditions for main assumptions made in Section 5.

Lemma D.4. *Suppose Assumptions O and H hold true.*

(i) If there exists a sequence of numbers ρ_{1n} such that $|m(x,p_j(x))| \leq \rho_{1n}\Lambda_n$ for each $j=1\dots k$, then $\varepsilon_n^m = \frac{\rho_{1n}\Lambda_n \log k}{n} + \sqrt{\frac{\log k}{n}}$;

(ii) If there exists a sequence of number ρ_{2n} and sub-gaussian $\mathbf{h}(X_i)$ such that $|m(X_i,p_j(X_i))| \leq \rho_{2n}\mathbf{h}(X_i)$ for each $j=1\dots k$ and $i=1\dots n$, then $\varepsilon_n^m = \rho_{2n}\sqrt{\frac{\log k}{n}}$.

Lemma D.5. *Suppose Assumption O and H hold true,* $\sqrt{\frac{\log k}{n}}\Lambda_n \to 0$ *and* $\mu_* \to 0$. *Then* $\left\|\hat{G}a_*\right\|_\infty = O_p(1)$.

D.3 Algorithms for selecting penalties

To implement (5.1), we need to specify Γ_k, λ_1 and λ_2. Suppose $\Gamma_k = \hat{G}$, we propose the following procedure: for each λ_1, first select λ_2 conservatively such that $\widetilde{\alpha}$ can achieve the fast convergence rate, and then select λ_1 so that its associated DR or BP estimator has a smaller $\overline{MMSR}_{\mathcal{H}_b}$ given some $b>0$ (thus more robust in the minimax sense).

Algorithm 1. *[Practical selection of λ_1 and λ_2]*

Step 1*: For each λ_1 in a grid of possible values, set $\hat{\lambda}_2 = 2\hat{\lambda}_0$, where $\hat{\lambda}_0 = \hat{\lambda}_0(\lambda_1)$ is an approximation of λ_0 given λ_1. Let $\widetilde{\alpha}(\lambda_1,\hat{\lambda}_2)$ be the solution of (5.1) with λ_1 and $\hat{\lambda}_2$.*

Step 2*: Calculate $\overline{MMSR}_{\mathcal{H}_b}\left[\tilde{\theta}_{BP}(\tilde{\alpha}(\lambda_1,\hat{\lambda}_2))\right]$ with some pre-specified b^2 and σ^2, or calculate $\overline{MMSR}_{\mathcal{H}_b}\left(\theta_{DR}(\tilde{\alpha}(\lambda_1,\hat{\lambda}_2),\gamma)\right)$ for some γ, and pre-specified b_γ^2 and σ^2.*[26]

Step 3*: Pick $\hat{\lambda}_1$ as the minimizer of $\overline{MMSR}_{\mathcal{H}_b}\left[\tilde{\theta}_{BP}(\tilde{\alpha}(\lambda_1,\hat{\lambda}_2))\right]$ or $\overline{MMSR}_{\mathcal{H}_b}\left[\theta_{DR}(\tilde{\alpha}(\lambda_1,\hat{\lambda}_2),\gamma)\right]$.*

Step 1 of Algorithm 1 requires an estimate of λ_0. Note $e_i^R = (e_{1i}^R,\dots,e_{ki}^R)'$, where $e_{ji}^R := m(X_i,p_j(X_i)) - \alpha_0(X_i)p_j(X_i)$ for $j=1\dots k$, $i=1\dots n$. Thus let $\Psi := diag[\psi_1,\psi_2,\dots,\psi_k]$ be a $k\times k$ diagonal matrix where $\psi_j := \left\{\mathbb{E}_n\left[e_{ji}^R\right]^2\right\}^{1/2}$ for $j=1\dots k$. When $\Gamma_k = \hat{G}$, we suggest to set λ_0 as

$$\hat{\lambda}_0 = 2\hat{c}(\lambda_1+1)\|\hat{\Psi}\|_\infty \Phi^{-1}\left(1-\tfrac{t}{2k}\right)/\sqrt{n} + 2\lambda_1\left\|\hat{P}\right\|_\infty, \tag{D.1}$$

[26] In practice, none of b, σ^2 or b_γ is known. We recommend conducting sensitivity analysis against different ratios of σ^2/nb^2 or σ^2/nb_γ^2.

where $\hat{\Psi}$ is an iterative estimate of Ψ stated in Algorithm 2 at the end of this section, $\Phi(\cdot)$ is the distribution function of a standard normal random variable, and $\hat{c} > 1$ is a slack constant and $0 < \hat{t} \leq 1$ is a confidence level.[27] The idea behind (D.1) is as follows: since usually $\|\mathbb{E}_n[p(X)(\alpha_0(X) - \alpha_*(X))]\|_\infty$ converges faster than $\left\|\mathbb{E}_n e^R\right\|_\infty$, we can expect $\lambda_0 \leq \bar{\lambda}_0$ wpa1, where $\bar{\lambda}_0 := 2(1+\lambda_1)\hat{c}\left\|\mathbb{E}_n e^R\right\|_\infty + 2\lambda_1\left\|\hat{P}\right\|_\infty$ for some $\hat{c} > 1$. Therefore, we estimate the conservative upper bound $\bar{\lambda}_0$ which only contains one unknown object $\left\|\mathbb{E}_n e^R\right\|_\infty$. Further note $\mathbb{E}e_i^R = \mathbf{0}$ and $\left\|\mathbb{E}_n e^R\right\|_\infty \leq \|\Psi\|_\infty \left\|\tilde{S}\right\|_\infty$, where $\left\|\tilde{S}\right\|_\infty := \max_{1\leq j\leq k}\left|\frac{\mathbb{E}_n e_{ji}^R}{\psi_j}\right|$. By Belloni et al. (2012, Lemma 5), we can expect $\mathbb{P}\left[\sqrt{n}\left\|\tilde{S}\right\|_\infty > \Phi^{-1}\left(1 - \frac{\hat{t}}{2k}\right)\right] \leq \hat{t} - o(1)$ for confidence level $\hat{t}$. (D.1) reflects the idea to bound term $\left\|\tilde{S}\right\|_\infty$ with a large probability.

Algorithm 2. *[Iterative estimation of Ψ]*

Step 0*: For each λ_1, fix $\hat{c}$ and $\hat{t}$. 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 $\hat{\psi}_j^1 =: \left\{\mathbb{E}_n\left[m(X, p_j(X)) - \mathbb{E}_n m(X, p_j(X))\right]^2\right\}^{1/2}$ for each $j = 1 \ldots k$. Find $\hat{\lambda}_2^1$ according to (D.1) and parameters in step 0. Compute $\widetilde{\alpha}^1(\lambda_1, \hat{\lambda}_2^1)$ according to (5.1) with penalty loadings λ_1 and $\hat{\lambda}_2^1$.*

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

$$\hat{\psi}_j^l := \left\{\mathbb{E}_n\left[m(X, p_j(X)) - \widetilde{\alpha}^{(l-1)}(\lambda_1, \hat{\lambda}_2^{l-1})p_j(X)\right]^2\right\}^{1/2} \quad \text{for each } j = 1 \ldots k,$$

and $\widetilde{\alpha}^{(l-1)}(\lambda_1, \hat{\lambda}_2^{l-1})$ is calibrated in iteration $l-1$ according to (5.1). 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*

$$\hat{\psi}_j^{L+1} := \left\{\mathbb{E}_n\left[m(X, p_j(X)) - \widetilde{\alpha}^{L}(\lambda_1, \hat{\lambda}_2^{L})p_j(X)\right]^2\right\}^{1/2} \quad \text{for each } j = 1 \ldots k.$$

E Simulation

This section assesses finite sample performance of $\tilde{\theta}_{BP}$ with a small fixed penalty when $k < n$. The set-up follows Example 2.1 and is in line with Kang et al. (2007). Let $U := \{U_1, U_2, U_3, U_4\}'$ be a vector of four random variables from multivariate standard normal distribution $N(\mathbf{0}, I_4)$. Outcome variable Y^* is generated as

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

[27] 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, 2017).

where e follows a standard normal distribution and is independent of U. The target parameter is $\mathbb{E}[Y^*] = 210$. The true propensity score is

$$\pi(u) = \mathbb{P}\{T = 1 | U = u\} = \Lambda(-u_1 + 0.5u_2 - 0.25u_3 - 0.1u_4),$$

where $\Lambda(\cdot) := \frac{\exp(\cdot)}{1+\exp(\cdot)}$. This mechanism generates a mean response rate of 0.5. Observed outcome is $Y = TY^*$. We do not observe U directly but only its transformed version $X := \{X_1, X_2, X_3, X_4\}'$, where

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

An iid sample of size $n = 200$ is drawn from observables $\{Y, T, X_1, X_2, X_3, X_4\}$. Calibrated RR is $\tilde{\alpha}(x,t) = t\tilde{a}'p(x)$, where

$$\tilde{a} = \left[\hat{G}_T\hat{G}_T + \lambda_1\hat{G}_T\right]^- \hat{G}_T\hat{P}, \quad \text{where } \hat{G}_T := \mathbb{E}_n[Tp(X)p'(X)], \hat{P} = \mathbb{E}_n[p(X)]. \tag{E.1}$$

E.1 Baseline results with mild selection bias

First we look at a situation with mild selection bias, where all relevant regressors are included even when k is small. We choose B-splines as basis functions, and k ranges from 5 to 121, covering 11 cases with $\frac{k}{n}$ growing from 0.025 to 0.605. We compare the performance of the following three BP estimators: (1) Near Optimal BP estimator with small penalty: RR is computed via (E.1) and small coefficient $\lambda_1 = 0.002$; (2) NR estimator: RR is computed with coefficient $\tilde{a}_{NR} = \hat{G}_T^-\hat{P}$, proposed in Newey and Robins (2018) and numerically equivalent to a plug-in OLS estimator; (3) "Simple Ridge" (SR) estimator: RR is computed with coefficient $\tilde{a}_{SR} = (\hat{G}_T\hat{G}_T + \lambda_1 I)^-\hat{G}_T\hat{P}$, with $\lambda_1 = 0.002$. Performance of a simple sample average estimator when α_0 is known is also reported. Bias and RMSE after 10000 experiments are collected in Figure E.1. Empirical coverage probabilities of these estimators when the variance is estimated by equation (4.3) are reported in Tables S6 and S7.

E.2 Robustness check

Considerable selection bias We consider 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 interactive terms) to alleviate bias. This creates a total of 10 cases with k growing from 5 to 121. Bias and RMSE after 10000 experiments are collected Figure E.2.

Sensitivity to choice of basis functions. Instead of using B-splines, we also construct basis functions with orthogonal polynomials. As the dimensional restriction on polynomials is stricter, we only consider 9 possible scenarios: k grows from 5 to 70, and $\frac{k}{n}$ increases from 0.025 to 0.35. Results are reported in Figure E.3.

Sensitivity to choice of λ_1 Finally, I check sensitivity of $\tilde{\theta}_{BP}$ to the choice of λ_1. Set-up is the same with baseline results using B-splines, but λ_1 ranges from 0 to 0.005. Results after 10000 experiments are collected in Figure E.4.

E.3 Comparison with doubly robust method

In this section we compare the finite sample performance of $\tilde{\theta}_{BP}$ with that of various DR estimators. I focus on three popular doubly robust methods involving estimation of propensity scores: (1) regression function and propensity score are estimated using (generalized) linear methods without selection; (2) regression function and propensity score are estimated using post lasso. (3) regression function and propensity score are estimated using lasso. Basis functions and simulation specifications follow Section E.1 and Table E.1. We only look at cases with smaller $\frac{k}{n}$ ratios, where we know these DR estimators would perform relatively well.[28] Results are reported in Figure E.5. As we can see clearly, the RMSE of DR estimators either perform strictly worse than $\tilde{\theta}_{BP}$, or in the case it does achieve a smaller RMSE, the improvement is not significant. Note these observations hold even though the penalty level for $\tilde{\theta}_{BP}$ is fixed and not optimally chosen in our simulation.

[28] Otherwise, when $\frac{k}{n}$ is too large, fitted propensity scores of numerically 0 or 1 would occur for the DR methods without selection, and convergence is also not guaranteed for the algorithm of lasso even after maximum iterations.

Figure E.1: Bias and RMSE using B-splines, 10000 Monte Carlo, $\lambda_1 = 0.002$, mild selection bias

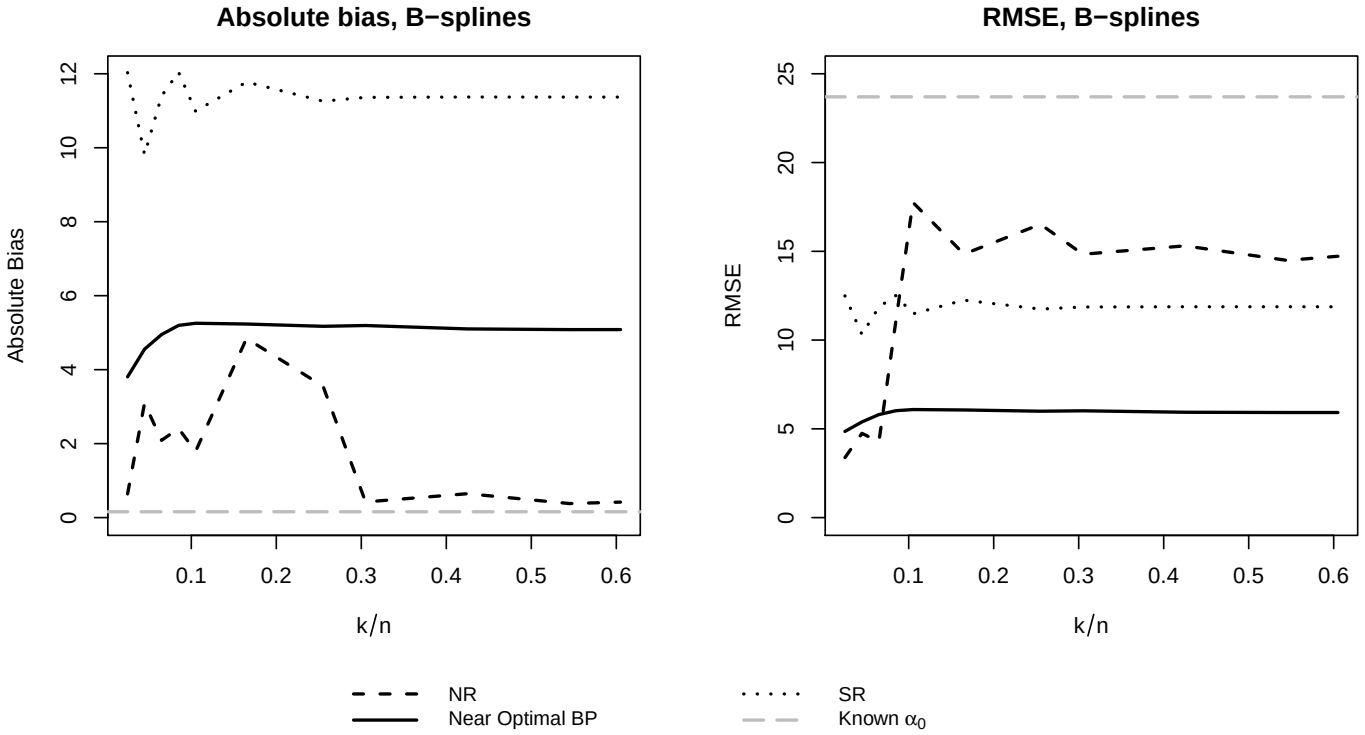


Figure E.2: Bias and RMSE using B-splines, 10000 Monte Carlo, $\lambda_1 = 0.002$, considerable selection bias

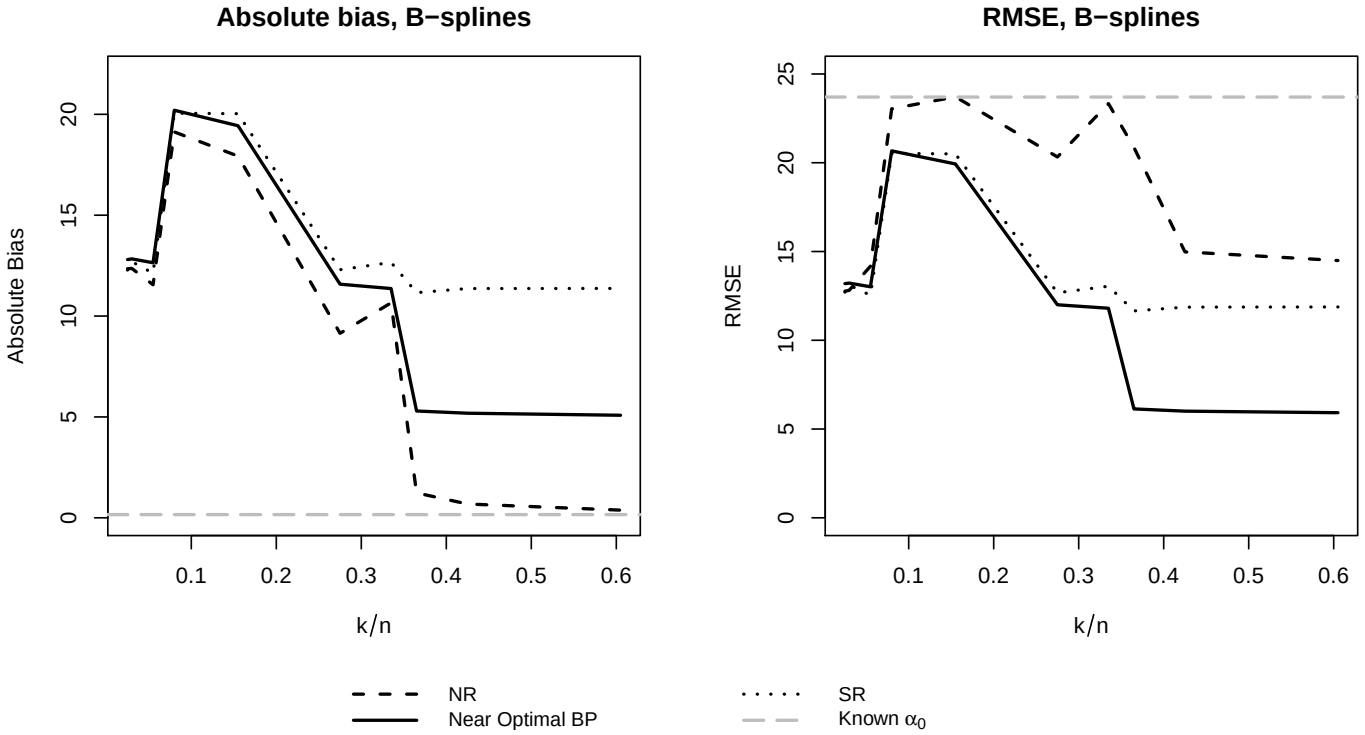


Figure E.3: Bias and RMSE using orthogonal polynomials, 10000 Monte Carlo, $\lambda_1 = 0.001$, mild selection bias

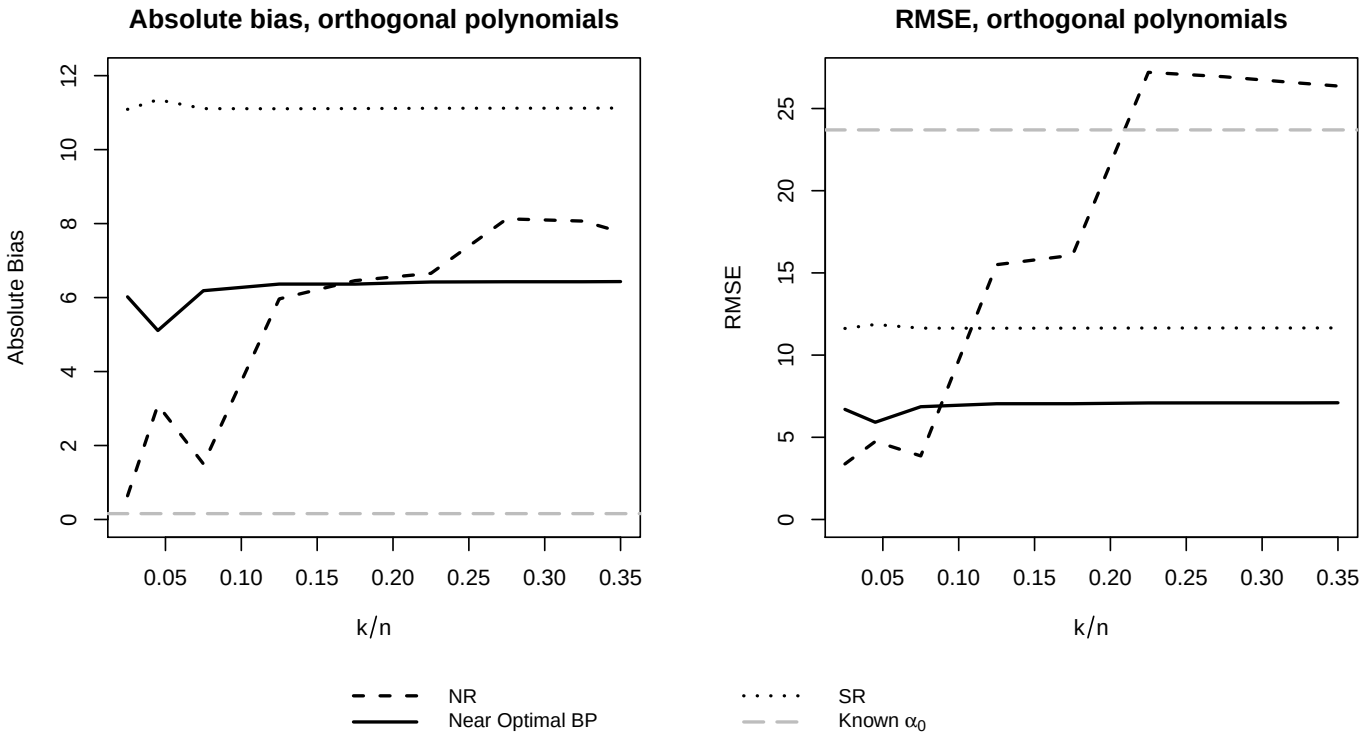


Figure E.4: Sensitivity of $\tilde{\theta}_{BP}$ to λ_1 using B-splines, mild selection bias

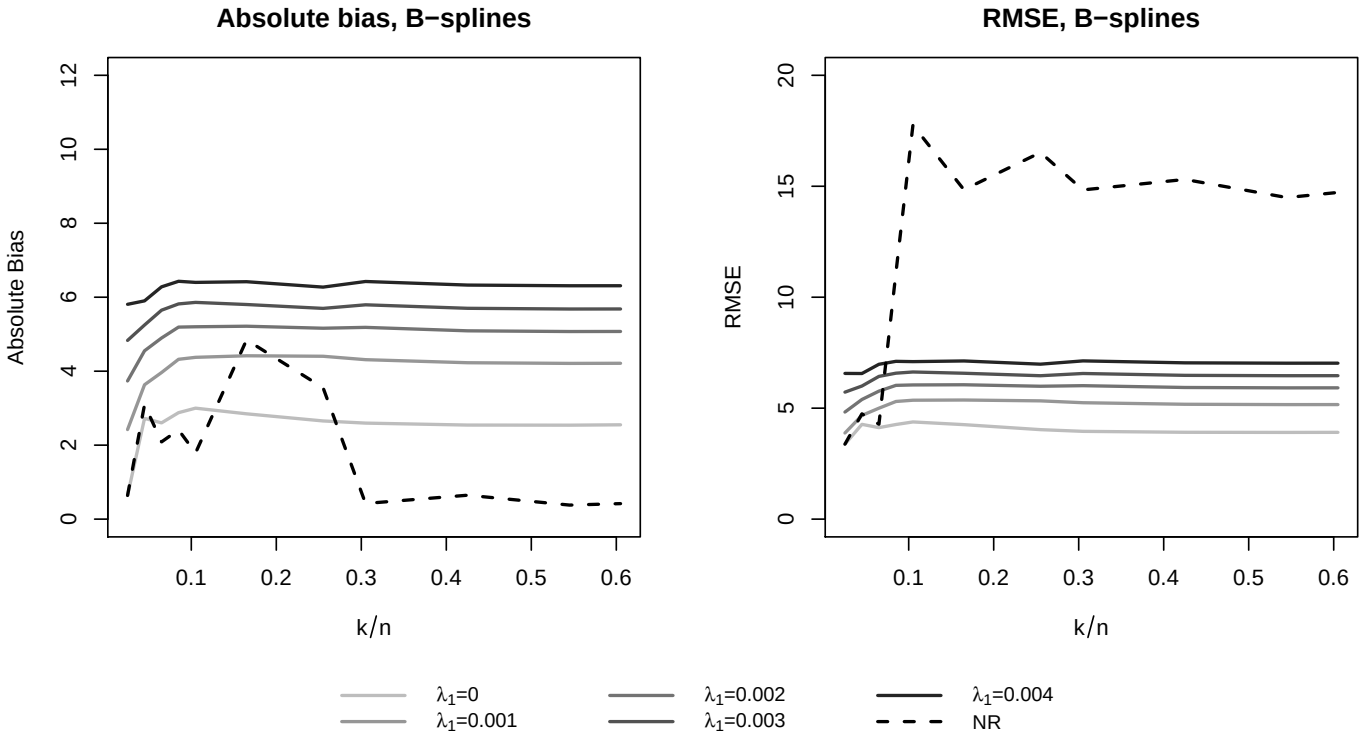
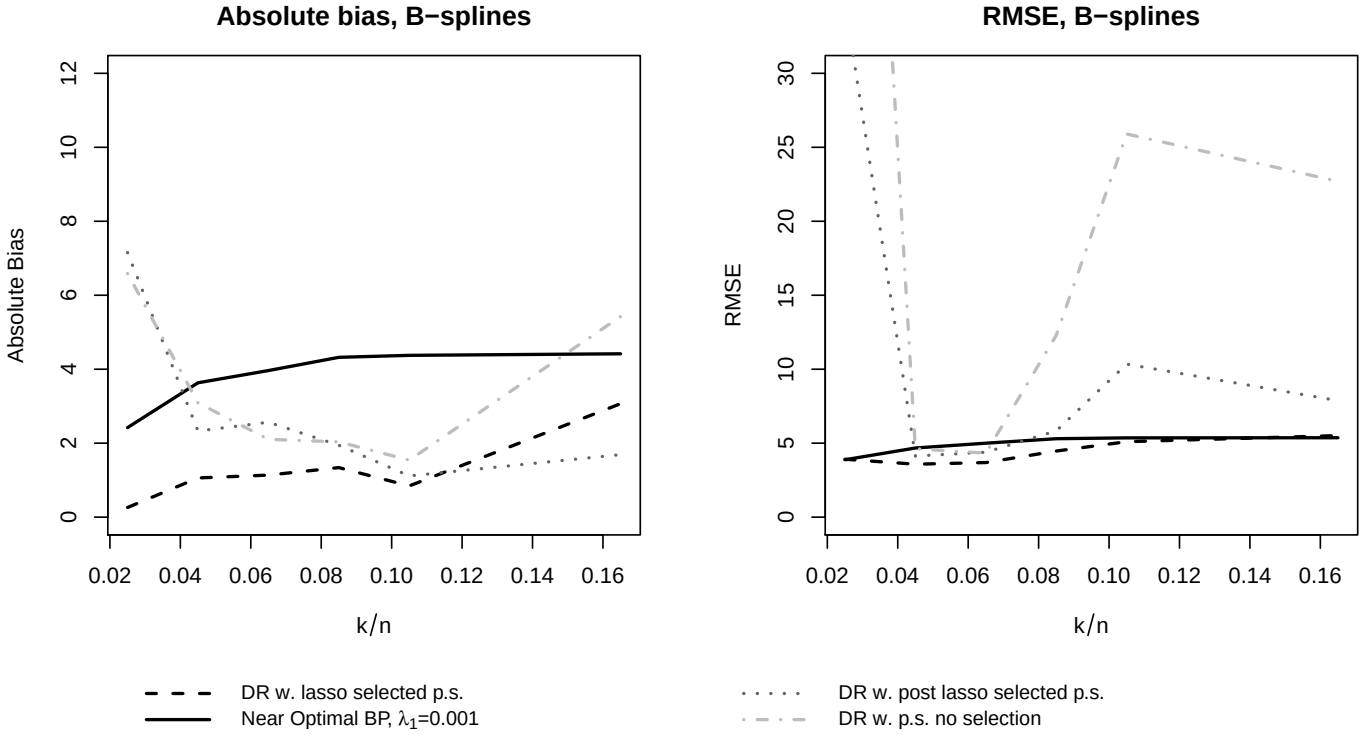


Figure E.5: Bias and RMSE using B-splines and DR methods, mild selection bias, 10000 simulations



References

Ai, C. and Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843.

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 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.

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. (2017). 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.

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.

Bradic, J., Chernozhukov, V., Newey, W. K., and Zhu, Y. (2019). Minimax semiparametric learning with approximate sparsity. *arXiv preprint arXiv:1912.12213*.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science and Business Media.

Chan, K. C. G., Yam, S. C. P., and Zhang, Z. (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):673–700.

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. (2015). 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.

DeVore, R. A. and Lorentz, G. G. (1993). *Constructive approximation*, volume 303. Springer Science and Business Media.

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.

Giné, E. and Koltchinskii, V. (2006). Concentration inequalities and asymptotic results for ratio type empirical processes. *Annals of Probability*, 34(3):1143–1216.

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.

Hansen, B. E. (2015). A unified asymptotic distribution theory for parametric and nonparametric least squares. Technical report, Working paper.

Hirshberg, D. A. and Wager, S. (2018). Augmented Minimax Linear Estimation. pages 1–49.

Huang, J. Z. (2003). Local asymptotics for polynomial spline regression. *Annals of Statistics*, 31(5):1600–1635.

Imbens, G. and Wager, S. (2018). Optimized regression discontinuity designs. *Review of Economics and Statistics*, (0).

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.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of applied econometrics*, 5(2):99–135.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, 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*.

Oster, E. (2017). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, pages 1–18.

Qiu, C. and Otsu, T. (2018). Information theoretic approach to high dimensional multiplicative models: Stochastic discount factor and treatment effect.

Robins, J., Tchetgen, E. T., Li, L., and van der Vaart, A. (2009). Semiparametric minimax rates. *Electronic journal of statistics*, 3:1305.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rothe, C. and Firpo, S. (2016). Properties of doubly robust estimators when nuisance functions are estimated nonparametrically. Technical report, Working paper.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Shen, X. (1997). On methods of sieves and penalization. *Annals of Statistics*, 25(6):2555–2591.

Tropp, J. A. (2015). An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230.

Van de Geer, S. (2007). The deterministic lasso. Seminar für Statistik, Eidgenössische Technische Hochschule (ETH) Zürich.

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.

Wong, R. K. and Chan, K. C. G. (2018). Kernel-based covariate functional balancing for observational studies. *Biometrika*, 105(1):199–213.

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.