Approximate Minimax Estimation of Average Regression Functionals

Chen Qiu*

First version: November 2019

This version: October 2022

Abstract

Empirical studies in economics often involve observational data with many controls or technical terms. As a way to seek robustness, a common practice is to compute an estimate using different subsets of controls. However, many conventional estimators ignore the additional mean squared error (MSE) incurred due to the presence of many controls or technical terms, which can cause the empirical results to be misinterpreted. In this paper, we propose a simple, new estimator that minimizes the worst-case finite-sample MSE among a class of plug-in estimators, if the regression function is indeed linear in the parameters. Therefore, our estimator is *approximately minimax optimal* in a finite sample. We characterize our estimator as a solution to a simple minimum-distance problem. We also establish favorable asymptotic properties of our estimator under weak conditions that hold independently of its finite-sample minimax property. Our approach deals with a class of regression functionals, including population average treatment effect under unconfoundedness and overlap as a special case. Compared to other recent approaches, our estimator behaves more robustly for the dataset of Ferraz and Finan (2011) and outperforms in a variety of simulation studies in terms of MSE.

*Department of Economics, Cornell University, Ithaca, New York 14853, US. E-mail: cq62@cornell.edu. I would like to thank the anonymous referees and the editor for their insightful comments, which greatly improved the quality of this paper. I am indebted to Taisuke Otsu for his continuous guidance and support. I benefited from discussions with Jaap Abbring, Debopam Bhattacharyam, Ben Deaner, Wouter Den Haan, Javier Hidalgo, Tatiana Komarova, Jia Li, Rachael Meager, Francesca Molinari, Whitney Newey, Alexey Onatskiy, Peter Robinson, Jörg Stoye, Daniel Sturm, Yike Wang and Yichong Zhang. I also would like to thank the participants in the numerous seminars and conferences for their helpful comments. All errors remain my own.

1 Introduction

Empirical studies in economics often involve observational data with many controls. For example, identifying a causal effect usually exploits quasi-experimental variations in the treatment, which is more plausible when conditioned on many controls. Researchers also frequently include technical terms such as squares and interactions to capture nonlinear or heterogeneous effects in the data (see, e.g., Imbens (2004); Imbens and Wooldridge (2009) for a review). Since economic theory is rarely informative regarding which controls or technical terms should be used, it is common to compare estimates resulting from different subsets of controls (Oster, 2019). If the data indeed come from a good natural experiment, can we estimate a parameter of interest that is not too sensitive to the use of many different controls?

In pursuing this question, we revisit Ferraz and Finan (2011)'s work, which studies the effect of electoral accountability on corruption. With the treatment plausibly randomly assigned, one of their main empirical strategies is to use an ordinary least squares (OLS) regression with many controls. Following common practice, they sequentially add different sets of controls to the regression. The estimated treatment effect changes considerably, jumping almost 50% from a plain vanilla mean comparison to a full specification with more than 60 controls. Such coefficient instability is often perceived only as a concern for the existence of omitted variable bias (OVB). To justify the sensitivity result, one often argues for the direction of the OVB. However, in this setting, the sign of the OVB is far from clear. See Section 2 for a more detailed discussion of this issue. The unsatisfactory performance of the default estimator in this credible natural experiment motivates us to develop a more robust approach. An ideal estimator should produce relatively stable estimates, if the dataset is indeed a good natural experiment with little bias. If, in contrast, the dataset features severe bias, adding important controls should correctly reduce the estimation error.

One way to approach this problem is as follows. Imagine a researcher who is contemplating a set of controls, including technical terms, in their analysis. Given a rich set of conditioning terms, they are willing to assume unconfoundedness and work with a linear-in-parameters specification. In this paper, we propose a simple, new estimator that aims to mitigate the mean squared error (MSE) resulting from the presence of many controls or technical terms. We apply our estimator to the same dataset used by Ferraz and Finan (2011). The estimated treatment effect is significantly more stable. Thus, we suspect that the observed coefficient instability is also associated with suboptimal control of the MSE in the presence of many conditioning terms. A less robust estimator ignores such MSE and can cause many empirical results to be misinterpreted. Our estimator acknowledges the need to trade off the MSE and tries to control it optimally. Compared to other recent approaches, our estimator behaves more robustly for the dataset used by Ferraz and Finan (2011), and it outperforms in a variety of simulation studies in terms of MSE.

We start by noting that many parameters of interest in economics are an ***average regression functional*** written as follows:

$$\theta_0 = \mathbb{E}[m(W, \gamma_0)],$$

where $\gamma_0(\cdot) := \mathbb{E}[Y|W = \cdot]$ is a regression function, and $m(w, \gamma_0)$ is a known linear function for each fixed value w (Newey and Robins, 2018). In many cases, we can also express θ_0 as a population weighted mean of the observed outcome. One leading example is the population average treatment effect under unconfoundedness and overlap conditions, for which the weighting function is the product of binary treatment and inverse propensity score. This weighting function plays a crucial role in semiparametric theory and is, in fact, the Riesz Representer (RR, also see Newey, 1990, 1994). We then consider a class of plug-in estimators in which the weighting function is approximated by some linear function. We call these estimators ***balancing plug-in*** (BP) estimators because they are closely analogous to the balancing method from statistics literature (e.g., Hainmueller, 2012; Zubizarreta, 2015; Chan et al., 2016; Kallus, 2016; Athey et al., 2018).

Our idea is to find a BP estimator that minimizes the worst-case finite-sample MSE, if γ_0 is indeed linear in parameters. Since γ_0 can in fact be nonparametric, our estimator is ***approximately minimax optimal***. Such a treatment is empirically motivated, as many researchers often engage with many conditioning terms. If a researcher works with many controls that do not involve technical terms (like Ferraz and Finan, 2011), our approach provides an estimator with a beneficial MSE property, if γ_0 is indeed linear in the selected controls. For researchers who are more concerned about the nonlinearity of γ_0 and who proceed with many technical terms to approximate γ_0, our estimator provides a way to reduce the MSE caused by using many technical terms. Our approach is simple to implement. The 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 developed by Chen and Pouzo (2012, 2015).

Next, we demonstrate that our estimator also possesses desirable asymptotic properties. These asymptotic results hold independently of the finite-sample minimax property and justify our estimator in large samples. Specifically, building on the literature in series estimation (see, among others, Newey, 1997; Shen, 1997; Huang, 2003; Ai and Chen, 2003; Newey and Powell, 2003; Chen, 2007; Chen et al., 2008; Cattaneo and Farrell, 2013; Belloni et al., 2015; Chen and Christensen, 2015; Hansen, 2015; Cattaneo et al., 2020), we establish the root-n normality of our estimator under weak conditions when $\frac{k}{n} \to 0$, where k is the number of conditioning terms and n is the sample size. Moreover, we show that our estimator can still achieve root-n normality under the many-regressor scenario when k is allowed to grow proportionally to n and the RR is allowed be mis-specified. This result applies the approach of Cattaneo et al. (2018b,a), who focus on partially linear models, to the framework of average regression functionals, which have a different structure.

The estimation of average regression functionals is a part of semiparametric literature dating back to, at least, Bickel (1982); Robinson (1988); Newey (1990); Van Der Vaart et al. (1991); Bickel et al. (1993); Andrews (1994); Newey (1994). See Van der Vaart (Chapter 25, 1998) for a review. The construction of our estimator is motivated by two recent influential approaches that endeavor to select estimators in a more refined way. One approach (*Approach 1*) examines the asymptotic order of the remainder error that arises from a linear expansion. If the remainder error vanishes to zero in a large sample, one can often show that the estimator is asymptotically normal and semiparametrically efficient. This motivates Newey and Robins (2018) and their high-dimensional variants (for example, Chernozhukov et al., 2022b,c) to calculate an estimator with remainders converging asymptotically to zero as quickly as possible. They found that debiasing (or orthogonalization) and cross-fitting in tandem yielded a fast remainder rate. Debiasing is a technique that reduces an estimator's asymptotic asymptotic bias (see, e.g., Chernozhukov et al. (2022a) for a general treatment on constructing debiased moments for a large class of GMM models). With cross-fitting, we estimate nuisance parameters using only observations from a random sample that are independent of the main sample. These techniques have also been studied by Farrell (2015); Rothe and Firpo (2016); Belloni et al. (2017); Chernozhukov et al. (2018), among others.[1] Compared to this approach, the remainder rate of our estimator is just as fast as a plug-in cross-fitted estimator. In addition, debiasing our estimator generally does not further improve the remainder rate. To obtain a remainder rate that is faster than ours, one usually must use both debiasing and cross-fitting simultaneously, which is computationally more costly but which may induce more finite-sample error.[2]

The other approach (*Approach 2*) focuses instead on finite-sample performance. By conditioning on a sample of covariates and prespecifying a convex functional class for γ_0, Donoho (1994) and Armstrong and Kolesár (2018) developed general procedures to obtain optimal statistical decisions for a class of linear functionals. One advantage to this approach is that it can deal with objects that are not root-n estimable, including

[1] Another popular method to correct asymptotic bias is to employ the jackknife method (see, e.g., Cattaneo et al. (2019) for a detailed treatment).

[2] Cattaneo and Jansson (2019) also found that simple plug-in estimators can achieve bootstrap consistency under minimal smoothness conditions, a property that is not achieved by several debiased estimators.

a conditional average binary treatment effect given the value of a continuous covariate (Zimmert and Lechner, 2019; Chernozhukov et al., 2022c; Fan et al., 2022) or an average treatment effect when the overlap is violated (Khan and Tamer, 2010). For the problem of estimation, this approach attempts to find an optimal estimator that minimizes the worst-case finite-sample MSE. Our finite-sample criterion is a special case considered by Donoho (1994); Armstrong and Kolesár (2018) applied to average regression functionals, where γ_0 is restricted as a class of linear functions. Armstrong and Kolesár (2021); Kallus (2020) also applied this approach to estimating a sample average treatment effect, considering general functional classes that may include our linear functions as a special case. But they did not provide the specific finite-sample or asymptotic results in this paper. Wong and Chan (2018) considered a similar criterion but focused on the asymptotic performance of debiased estimators when γ_0 was in a reproducing kernel Hilbert space. Hirshberg and Wager (2021) also proposed similar criteria for debiased estimators. As a result, they examined a case in which γ_0 should lie in a space that was restricted by an initial estimate of γ_0. Imbens and Wager (2019) applied similar minimax linear estimation procedures to their regression discontinuity design.

Our paper contributes to theoretical discussions on whether the finite-sample optimal estimators used in Approach 2 can be as asymptotically efficient as those demonstrated by Approach 1, if the object of interest is indeed root-n estimable. Our results confirm that such a conjecture is true for the specific regularity class of linear functions with growing dimensions when k is not too great compared to n. Thus, our results justify Approach 2 asymptotically as well for that particular class of linear functions. Regarding other general nonparametric classes (e.g., a Hölder class with a sufficiently high smoothness index), this conjecture is plausible but seems to remain an open question to the best of my knowledge. We leave this matter for future research.

When $\frac{k}{n} \geq 1$, the remainder of our estimator can grow asymptotically. In this case, Chernozhukov et al. (2022b) showed that debiasing and cross-fitting are also effective for regularized estimators of the RR using lasso or other types of learners, and their estimators can achieve root-n normality under weak conditions. Their approach is also valid for nonlinear functionals. Chernozhukov et al. (2022c) extended the analyses to nonregular linear functionals. Qiu (2020) modified the minimum-distance estimator for the RR with an elastic-net-style penalty and showed that the modified estimator converged at least as fast as its lasso counterpart. With this modified estimator, Qiu (2020) proposed a debiased estimator that also achieved root-n normality when $\frac{k}{n} \geq 1$ for average regression functionals.

The remainder of this paper is organized as follows: Section 2 demonstrates the performance of our estimator via the work of Ferraz and Finan (2011). Section 3 introduces our general framework and related examples. Section 4 presents our estimator and its finite-sample property. Section 5 establishes the asymptotic proprieties of our estimator. Simulation studies are provided in Section 6. Proofs and additional technical and empirical results can be found in the Appendix.

2 Application: Estimating an average treatment effect with many controls

In this section, we explain the usefulness of our estimator in the context of estimating an average treatment effect with observational data. Section 2.1 introduces the work of Ferraz and Finan (2011) as an example of researchers being interested in estimating a treatment effect that is robust to many different controls. Section 2.2 discusses how a conventional estimator might provide unsatisfactory results and lead to being misinterpreted. Section 2.3 introduces our approach and illustrates the performance of the proposed estimator. Compared to other modern off-the-shelf estimators, we find that our estimator performs more robustly and could be a more suitable estimator for moderately high-dimensional datasets.

2.1 Motivating example: Ferraz and Finan (2011)

Ferraz and Finan (2011) studied the effect of electoral accountability on corruption by exploiting a natural experiment in Brazil. They collected a municipality-level dataset from an anticorruption campaign in which the treatment was plausibly randomly assigned. Therefore, one of their main empirical strategies was to apply OLS using many mayoral and municipal characteristics as controls. They found that in municipalities in which mayors were serving their first term, the share of resources involved in corruption was significantly less than in municipalities with second-term mayors. Ferraz and Finan (2011) used the following controlled regression:

$$Y_i = \theta_0 T_i + X_i' \beta + \varepsilon_i, \tag{2.1}$$

where $T_i = 1$ if mayor i's term limit is not binding (with reelection incentives), $T_i = 0$ if the mayor's term limit is binding (without reelection incentives), Y_i is the observed outcome on the share of resources related to corrupt activities in the mayor's municipality, θ_0 is the object of interest[3], ε_i is the error term, and X_i is a vector of included controls.

As one of their primary empirical analyses, Ferraz and Finan (2011) explored how the estimate of θ_0 in (2.1) changed when different sets of controls were sequentially included.

[3] This can be interpreted as the average treatment effect of reelection incentives on corruption, if the individual treatment effect is homogeneous (Angrist, 1998).

Starting with a plain vanilla mean comparison, they sequentially added five sets of relevant controls. The full specification included a total of 67 conditioning terms versus a sample size of 476, a moderately high-dimensional scenario, with their ratio being approximately 0.14. From Table 1, we may see that the OLS estimates are quite unstable and sensitive to the sets of included controls. A simple mean comparison yields a point estimate of -0.0188, meaning that lame duck mayors, on average, steal 1.88% more resources. As we add additional controls, the magnitude of the estimated effect gradually increases. With all controls included, the point estimate becomes -0.0275, representing an increase of almost 50% over the plain vanilla estimate.

2.2 From unstable coefficients to robust estimates

The approach by Ferraz and Finan (2011) is common among researchers who work with observational data and are willing to assume unconfoundeness (or selection-on-observables). A stable estimate with respect to different controls is considered more reassuring (Chiappori et al., 2012). In the primary analysis by Ferraz and Finan (2011), however, the point estimates do change substantially. For instance, Specification 2 features 21 conditioning terms, and its estimate increases by only about 5% compared to the plain vanilla estimate. Specification 6 features 67 conditioning terms, but its estimate increases almost 40% compared to Specification 2. To justify these unstable estimates, researchers often argue that added controls are relevant and will serve to alleviate OVB. For example, one might claim that: (1) ignoring added controls underestimates the effect of reelection incentives on corruption; (2) adding more controls partially corrects OVB, thereby providing estimates that more closely approximate the true effect. However, such arguments seem unsatisfactory. For example, from Specification 2 to 6, the additional controls include municipal, political, and judicial characteristics as well as some state and lottery dummy variables. How these additional controls are related to the potential corruption level and treatment assignment is far from clear, making the judgement on the sign of OVB difficult. For this arguably credible natural experiment, justifying the unstable estimates purely from the perspective of OVB is not entirely convincing.

Drawing on the example of Ferraz and Finan (2011), we believe it is important to use a more robust estimator when working with observational data: (1) if the dataset is indeed a good natural experiment with mild bias, it should produce relatively stable estimates with different subsets of controls; (2) if, conversely, the dataset features severe bias, adding important controls should correctly reduce the MSE. In this paper, we propose a new estimator that may achieve these goals better than other modern approaches.

2.3 Illustration of our estimator

To introduce our approach, we deviate from (2.1) and consider a more flexible heterogeneous treatment effect model with binary treatment $T_i \in \{1, 0\}$. Let $Y_i(1)$ denote the potential outcome (i.e., the level of corruption) of unit i when $T_i = 1$, and $Y_i(0)$ denote the potential outcome when $T_i = 0$. The average treatment effect is then $\theta_0 := \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]$. In the work of Ferraz and Finan (2011), this is the expected effect of reelection incentives on corruption. A researcher interested in estimating θ_0 collects a random sample $\{(Y_i, T_i, X_i')'\}_{i=1}^n$, where $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$ is the observed outcome, and X_i is a vector of the controls. Given $\{X_i\}_{i=1}^n$, the researcher imposes unconfoundedness and overlap assumptions, under which θ_0 is identified as follows:

$$\theta_0 = \mathbb{E}\left[\gamma_0(X_i, 1) - \gamma_0(X_i, 0)\right], \tag{2.2}$$

where $\gamma_0(x, 1) := \mathbb{E}[Y_i | X_i = x, T_i = 1]$, and $\gamma_0(x, 0) := \mathbb{E}[Y_i | X_i = x, T_i = 0]$. Furthermore, we suppose both $\gamma_0(x, 1)$ and $\gamma_0(x, 0)$ are linear-in-controls:

$$\gamma_0(x, 1) = \beta_1' x, \beta_1 \in \mathbb{R}^k, \ \gamma_0(x, 0) = \beta_0' x, \beta_0 \in \mathbb{R}^k. \tag{2.3}$$

We are concerned about a researcher who is willing to assume (2.3), like Ferraz and Finan (2011), but hopes to find an estimator with low MSE in a finite sample if the linear specification (2.3) were true. Our proposed estimator has a simple form as follows:

$$\tilde{\theta}_{BP} := \tilde{\theta}_1 - \tilde{\theta}_0, \quad \tilde{\theta}_1 := \frac{1}{n} \sum_{i=1}^n \tilde{\alpha}_1(X_i) T_i Y_i, \quad \tilde{\theta}_0 = \frac{1}{n} \sum_{i=1}^n \tilde{\alpha}_0(X_i) \left(1 - T_i\right) Y_i,$$

$$\tilde{\alpha}_1(x) := \tilde{a}_1' x, \quad \tilde{a}_1 := \left[\hat{G}_T \hat{G}_T + \lambda_{n,1} \hat{G}_T\right]^- \hat{G}_T \hat{P}, \quad \hat{G}_T := \frac{1}{n} \sum_{i=1}^n [T_i X_i X_i'], \quad \hat{P} := \mathbb{E}_n[X_i],$$

$$\tilde{\alpha}_0(x) := \tilde{a}_0' x, \quad \tilde{a}_0 := \left[\hat{G}_{1-T} \hat{G}_{1-T} + \lambda_{n,0} \hat{G}_{1-T}\right]^- \hat{G}_{1-T} \hat{P}, \quad \hat{G}_{1-T} := \frac{1}{n} \sum_{i=1}^n [(1 - T_i) X_i X_i'],$$

where $\lambda_{n,1}$ and $\lambda_{n,0}$ are the two penalty coefficients selected by the researcher. As shown in Section 4, $\tilde{\theta}_{BP}$ controls the finite-sample worst-case MSE among a class of plug-in estimators (i.e., it can be treated as a finite sample minimax estimator). In Section 5, we show that $\tilde{\theta}_{BP}$ also possesses desirable asymptotic properties with many controls, even if the minimax property does not hold.

We apply $\tilde{\theta}_{BP}$ with a small, fixed penalty ($\lambda_{n,1} = \lambda_{n,0} = 0.001$) to the same dataset in Ferraz and Finan (2011). The results are reported in the second row of Table 1. In contrast with the controlled OLS regression, our estimator performs more robustly against various sets of included controls. The estimates are quantitatively stable at approximately

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)
Our estimator $\tilde{\theta}_{BP}$	Effect	-0.0187**	-0.0186**	-0.0162*	-0.0182*	-0.0182*	-0.0182**
$\lambda_{n,1} = \lambda_{n,0} = 0.001$	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**
$\lambda_n = 0.001$	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
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***
$\lambda_{n,1} = \lambda_{n,0} = 0.001$	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)
Debiased w.	Effect		-0.0180*	-0.0177*	-0.0252**	-0.0252**	-0.0214*
post lasso selection	S.E.		(0.0094)	(0.0096)	(0.0111)	(0.0111)	(0.0110)
Debiased w.	Effect		-0.0188**	-0.0181*	-0.0225**	-0.0225**	-0.0219**
lasso selection	S.E.		(0.0095)	(0.0095)	(0.0100)	(0.0100)	(0.0100)
Linear partialing 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)
Mayoral 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 dummies		No	No	No	No	Yes	Yes
State dummies		No	No	No	No	No	Yes

Note: k is the number of conditioning terms and n is the sample size. Numbers in parentheses represent computed standard errors. (1)-(6) use the same controls as those used in Table 4 of Ferraz and Finan (2011). Ridge methods use the R package "glmnet"; four lasso based methods use the R package "hdm". For controlled (cross-validated) ridge regressions, standard error is calculated using the sandwich formula with residuals derived from the ridge regression; For plug-in (cross-validated) ridges, standard error is calculated with α_0 estimated using the estimator in Newey and Robins (2018).

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

-0.018 throughout the six specifications, and they are all statistically significant at least at 10% level. We also examine the performance of other modern off-the-shelf methods, the results of which are also reported in Table 1. These estimators do not have the minimax property enjoyed by $\tilde{\theta}_{BP}$, and the majority of them do not perform as robustly as our estimator. Thus, the observed coefficient instability of the other estimators is likely to be associated with the suboptimal control of finite-sample MSE, especially in the presence of many controls.

Admittedly, being stable is not a property that we always desire from an estimator. Indeed, if some important controls are missed, then the coefficient stability actually becomes an erratic property that we hope to avoid. To cope with such concerns, in the simulation studies in Section 6.1.4, we examine the performance of our estimator when the data contain severe bias. We find that our estimator correctly reduces the MSE as additional relevant terms are added, while other methods that fail to control MSE optimally may perform erratically. Based on the performance of the real dataset and the simulation studies, we believe that our estimator is more robust than other existing methods.

We conducted further robustness checks on our estimator. In Appendix E.1, we follow Ferraz and Finan (2011) to use different measures of corruption, and to take measures to control possible selection-on-unobservables. Compared to the other estimators, our estimator still performs more robustly. In Appendix E.2, we use many technical terms constructed from the raw controls. The performance of our estimator as well as the other estimators exhibits a pattern similar to the main baseline results shown in Table 1. Finally, in Appendix E.3, we perform a sensitivity analysis of our estimator with respect to various magnitudes of penalty coefficients. The estimated effect does not change significantly as the penalty varies over a large interval.

3 General framework and related examples

We now introduce a setup that includes the average treatment effect as one example and encompasses many economics problems as well.

3.1 Setup

Let $Y \in \mathbb{R}$ be a random outcome variable and $W \in \mathcal{W} \subseteq \mathbb{R}^{d_W}$ be a random vector of dimension d_W. Let $\mathbb{P}$ denote the joint distribution of $(Y, W')'$. We write the conditional expectation of Y given W as follows:

$$\gamma_0(w) := \mathbb{E}[Y|W = w], \tag{3.1}$$

where $\mathbb{E}[\cdot\,]$ is the expectation operator under $\mathbb{P}$. We assume the following:

$$\gamma_0 \in L_{\mathbb{P},2} := \left\{ f : \mathcal{W} \to \mathbb{R} \mid \int_{w \in \mathcal{W}} f^2(w) d\mathbb{P}(w) < \infty \right\}.$$

Let $m(w, \gamma_0) : \mathbb{R}^{d_W} \times L_{\mathbb{P},2} \to \mathbb{R}$ be a known linear function such that for every $\gamma_1, \gamma_2 \in L_{\mathbb{P},2}$ and every constant $r \in \mathbb{R}$,

$$m(w, r\gamma_1 + \gamma_2) = rm(w, \gamma_1) + m(w, \gamma_2) \tag{3.2}$$

for each $w \in \mathcal{W}$. The object of interest in this paper is the continuous linear functional $\mathbb{E}[m(W, \cdot)] : L_{\mathbb{P},2} \to \mathbb{R}$ evaluated at γ_0[4]:

$$\theta_0 := \mathbb{E}[m(W, \gamma_0)]. \tag{3.3}$$

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

$$\mathbb{E}[m(W, \gamma)] = \mathbb{E}[\gamma(W)\alpha_0(W)]. \tag{3.4}$$

We call α_0 the Riesz Representer (RR) of the functional $\mathbb{E}[m(W, \cdot)]$. According to (3.4), we can interpret θ_0 as an RR-weighted population average of the regression function (or *average regression functional*) as follows:

$$\theta_0 = \mathbb{E}[\underbrace{\alpha_0(W)}_{\text{weight}}\underbrace{\gamma_0(W)}_{\text{regression}}]. \tag{3.5}$$

Furthermore, the Law of Iterated Expectations (LIE) implies that θ_0 can also be written as:

$$\theta_0 = \mathbb{E}[\alpha_0(W)Y],$$

a weighted mean of the observed outcome where the weight function for outcome Y is $\alpha_0(W)$.

3.2 Examples

Example 3.1. Missing data and the average treatment effect

Consider the framework of incomplete outcome data in Rubin (1974) and Rosenbaum and Rubin (1983). We are concerned about the population mean $\theta_0 = \mathbb{E}[Y^*]$ of an outcome variable Y^*. However, we observe only $Y = TY^*$, where $T \in \{0, 1\}$. That is, if $T = 0$, it implies that Y^* is missing. Suppose we also observe a covariate vector X. We

[4] The functionals of the form $\mathbb{E}[m(Z, \cdot)]$ where $W \subseteq Z$ can also be considered at the expense of additional technicality.

define the inverse propensity score as $\omega(x) = 1/\Pr\{T = 1|X = x\}$. Suppose the overlap assumption holds such that $0 < \omega(x) < 1$ for all $x \in \mathcal{X}$. Also, suppose that Y^* and T are conditionally independent given X. Then, θ_0 can be identified as follows:

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

where $\gamma_0(x, 1) = \mathbb{E}[Y|X = x, T = 1]$. Furthermore, for each $g \in L_{\mathbb{P},2}$,

$$\mathbb{E}[\omega(X)Tg(X)] = \mathbb{E}[g(X)]. \tag{3.6}$$

This example fits the general setup by letting $W = (X', T)'$ such that $\gamma_0(w) = \gamma_0(x, t)$ and the linear function is $m(w, \gamma) = \gamma(x, 1)$. Furthermore, (3.6) identifies the RR as $\alpha_0(w) = \alpha_0(x, t) = \omega(x)t$. This framework can be extended to account for the average treatment effect. See Section 2.3 or Qiu and Otsu (2022) for further discussion.

Example 3.2. Regression discontinuity design away from the cut-off

We sightly modify Example 3.1 but keep the notation (Y, T, X). In addition, suppose T is determined by a running variable $R \in \mathbb{R}$ at the cut-off point 0: $T = \mathbf{1}\{R \geq 0\}$. We fix a known boundary point, $b > 0$. The object of interest is defined as:

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

This object is useful when external validity is of concern. For example, we may be interested in the population group away from the cut off (i.e., inframarginal applicants) rather than a group in the immediate vicinity of the cut off. One way to identify θ_0 is by assuming that Y^* and R are independent conditional on X and $-b \leq R \leq b$ (Angrist and Rokkanen, 2015). Then, the following can be shown:

$$\theta_0 = \mathbb{E}[\gamma_0(X)| - b \leq R \leq b], \tag{3.7}$$

where $\gamma_0(x) = \mathbb{E}[Y|X = x, 0 \leq R \leq b]$. Here $W = (X', R)'$, $\gamma_0(w) = \gamma_0(x)$ and $m(w, \gamma) = \gamma(x)$. The RR is found in a fashion similar to (3.6) under suitable overlap assumptions as follows:

$$\alpha_0(w) = \alpha_0(x, r) = \omega(x)\mathbf{1}\{r \geq 0\}, \tag{3.8}$$

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

Example 3.3. Weighted average derivative and single index model

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

$$\theta_0 = \mathbb{E}\left[\omega(W)\frac{\partial \gamma_0(W)}{\partial W_1}\right], \tag{3.9}$$

where W is a vector of the covariates whose first element is W_1, $\omega(w)$ is a known weight function and $\gamma_0(w) = \mathbb{E}[Y|W = w]$. In this example, $m(w, \gamma) = \omega(w)\frac{\partial \gamma(w)}{\partial w_1}$. (3.9) encompasses several models found in the literature, such as the single index model (Stoker, 1986; Härdle and Stoker, 1989; Powell et al., 1989, etc.) and the nonseparable model (Imbens and Newey, 2009; Altonji et al., 2012, etc.). See Cattaneo et al. (2013) for a review. To find the RR, assume that $\omega(w)$ has a value of 0 at the boundaries. Integration by parts yields the following:

$$\theta_0 = -\mathbb{E}\left[\gamma_0(W)\frac{\partial v(W)/\partial W_1}{f(W)}\right],$$

where $v(w) = \omega(w)f(w)$, and f is the marginal density of W. The RR is then found as:

$$\alpha_0(w) = -\frac{\partial v(w)/\partial w_1}{f(w)}.$$

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

Example 3.4. Average effect after policy intervention

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

$$\theta_0 = \mathbb{E}[\gamma_0(\pi(W))]. \tag{3.10}$$

Here, $m(w, \gamma) = \gamma(\pi(w))$. Rewriting (3.10) using a change-of-measure, we find the RR as $\alpha_0(w) = \frac{f_\pi(w)}{f(w)}$, where f and f_π are the marginal densities of W and W_π, respectively.

Example 3.5. Approximate average consumer surplus

This is an example from nonparametric welfare analysis. To introduce the idea, we consider a highly simplified version of the problem studied by Hausman and Newey (1995, 2016, 2017). In our example, the functional of interest is only an approximation of the consumer surplus. See Hausman and Newey (2016) on how to derive bounds on the average exact consumer surplus, which is also an average regression functional. See Chernozhukov et al. (2022b) for the derivation of the RR in the more sophisticated case. Let the demand function of a commodity be $\gamma_0(r, z) = \mathbb{E}[Q|R = r, Z = z]$, where R is the price, Q is the quantity demanded and Z is a vector of other characteristics that affect demand. Let Z_1, the first variable of Z, denote income. We define the approximate consumer surplus for a price change from r_0 to r_1 as $\int_{r_0}^{r_1} \gamma_0(\tilde{r}, Z)d\tilde{r}$. The object of interest is:

$$\theta_0 = \mathbb{E}\left[\omega(Z)\int_{r_0}^{r_1} \gamma_0(\tilde{r}, Z)d\tilde{r}\right],$$

for some known weight function $\omega(z)$. The parameter θ_0 is the average effect of the price change on certain income groups (and possibly on other observable characteristics). Here, $W = (R, Z')'$, $\gamma_0(w) = \gamma_0(r, z)$ and $m(w, \gamma) = \omega(z)\int_{r_0}^{r_1} \gamma(\tilde{r}, z)d\tilde{r}$. Let $f_{R|Z}(r|z)$ be the conditional density. Then the following can be shown:

$$\theta_0 = \mathbb{E}\left[\frac{\omega(Z)\mathbf{1}\{r_0 \leq R \leq r_1\}}{f_{R|Z}(R|Z)}\gamma_0(R, Z)\right],$$

suggesting that the RR is $\alpha_0(w) = \alpha_0(r, z) = \frac{\omega(z)\mathbf{1}\{r_0 \leq r \leq r_1\}}{f_{R|Z}(r|z)}$.

Example 3.6. GMM with auxiliary data

This example was inspired by the work of Chen et al. (2005, 2008) and can be applied to several problems, including measurement error models with validation data and the estimation of the average treatment effect on the treated. To simplify presentation, we focus on the estimation of the average treatment effect on the treated, defined as:

$$\theta_0 = \mu_1 - \mu_0,\ \mu_1 = \mathbb{E}[Y(1)|T = 1], \mu_0 = \mathbb{E}[Y(0)|T = 1],$$

where T is binary treatment, and $Y(1), Y(0)$ are two potential outcomes. Let $Y = TY(1) + (1 - T)Y(0)$ be the observed outcome, X be a vector of pretreatment variables and $\pi(x) = \Pr\{T = 1|X = x\}$ be the propensity score. Under unconfoundedness and overlap conditions similar to those in Example 3.1, we can write the following:

$$\mu_1 = \mathbb{E}[\gamma_0(X, 1)|T = 1] = \mathbb{E}\left[\alpha_1(X, T)Y\right],$$

where $\gamma_0(x, 1) = \mathbb{E}[Y|X = x, T = 1]$, $\alpha_1(x, t) = \alpha_1 t, \alpha_1 = \frac{1}{\mathbb{E}[\pi(X)]}$. Thus, μ_1 is an example of our framework, where $W = (X', T)'$, $\gamma_0(w) = \gamma_0(x, 1)$, $m(w, \gamma) = \gamma(x, 1)$, and the RR is $\alpha_1(x, t)$. Similarly, for μ_0, note:

$$\mu_0 = \mathbb{E}[\gamma_0(X, 0)|T = 1] = \mathbb{E}\left[\alpha_0(X, T)Y\right],$$

where $\gamma_0(x, 0) = \mathbb{E}[Y|X = x, T = 0]$, and where the RR is $\alpha_0(x, t) = \alpha_0(x)(1-t), \alpha_0(x) = \frac{\pi(x)}{\mathbb{E}[\pi(X)](1-\pi(x))}$.

4 The estimator

Now, suppose a random sample $\{(Y_i, W_i')'\}_{i=1}^n$ of size n is drawn from $\mathbb{P}$. Since $\theta_0 = \mathbb{E}[\alpha_0(W)Y]$, we can estimate θ_0 using a BP estimator with some weight function α

$$\tilde{\theta}_{BP}(\alpha) := \mathbb{E}_n[\alpha(W)Y],$$

where $\mathbb{E}_n[f] := \mathbb{E}_n[f(W,Y)] := \frac{1}{n}\sum_{i=1}^n [f(W_i, Y_i)]$ denotes the sample average of function f. In this paper, we are concerned about selecting a weight function α that leads to the least estimation error. We motivate our choice of weight function with the minimax performance criterion.

Slightly abusing notations, hereafter let $\mathbb{E}[\cdot] := \mathbb{E}_{\mathbb{P}^n}[\cdot]$ be the expectation operator under $\mathbb{P}^n$, the sampling distribution of $\{(Y_i, W_i')'\}_{i=1}^n$. That is, from now on and including in our asymptotic analysis, we treat data $\{(Y_i, W_i')'\}_{i=1}^n$ as a triangular array, where n identically and independently distributed (i.i.d.) copies are drawn from the population distribution $\mathbb{P}$, and where n is growing. As a result, $\mathbb{E}[f(W_i, Y_i)] = \mathbb{E}[f(W,Y)]$ for each $i = 1 \ldots n$. Note that $\mathbb{E}[f(W_i, Y_i)]$ should be understood as the expectation under the sampling distribution, whereas $\mathbb{E}[f(W,Y)]$ refers to our expectation according to the population distribution.

4.1 Approximate minimax optimality

From the definition of γ_0 in (3.1), we can write the following expression:

$$Y_i = \gamma_0(W_i) + e_i, \quad \mathbb{E}[e_i | W_i = w] = 0, \quad i = 1 \ldots n. \tag{4.1}$$

To illustrate the finite-sample property of our proposed estimator, assume for now that $\mathbb{E}[e_i^2 | W_i = w] = \sigma^2$ with some known constant $\sigma^2 > 0$.[5] Then, conditional on $\{W_i\}_{i=1}^n$, the MSE of a BP estimator $\tilde{\theta}_{BP}(\alpha)$ with a fixed α is as follows:

$$\mathrm{MSE}_n(\tilde{\theta}_{BP}(\alpha)) := \mathbb{E}\left[\left(\tilde{\theta}_{BP}(\alpha) - \theta_0\right)^2 | W_1, \ldots, W_n\right] = \mathrm{bias}_n^2(\tilde{\theta}_{BP}(\alpha)) + \mathrm{var}_n(\tilde{\theta}_{BP}(\alpha)),$$

where

$$\mathrm{bias}_n(\tilde{\theta}_{BP}(\alpha)) = \mathbb{E}_n[\alpha(W)\gamma_0(W)] - \mathbb{E}[m(W_i, \gamma_0)],$$
$$\mathrm{var}_n(\tilde{\theta}_{BP}(\alpha)) = \frac{\sigma^2}{n}\mathbb{E}_n[\alpha^2(W)].$$

[5] For the validity of the asymptotic property of the estimator analyzed in Section 5, we only require $\sup_{w \in \mathcal{W}} \mathbb{E}[e_i^2 | W_i = w]$ to be bounded from above uniformly for all i.

Conditional on $\{W_i\}_{i=1}^n$, $\text{var}_n(\tilde{\theta}_{BP}(\alpha))$ is a known function of α, while $\text{bias}_n(\tilde{\theta}_{BP}(\alpha))$ depends on two unknown objects: γ_0 and $\mathbb{E}[m(W_i, \gamma_0)]$. Ideally, we hope to evaluate the performance of $\tilde{\theta}_{BP}(\alpha)$ according to its worst-case MSE:

$$\overline{\text{MSE}}_n(\tilde{\theta}_{BP}(\alpha)) := \sup_{\gamma_0 \in \mathcal{H}} \left[\text{bias}_n^2(\tilde{\theta}_{BP}(\alpha))\right] + \text{var}_n(\tilde{\theta}_{BP}(\alpha)), \tag{4.2}$$

where $\mathcal{H}$ is a general functional class to which γ_0 belongs. However, evaluating $\overline{\text{MSE}}_n(\tilde{\theta}_{BP}(\alpha))$ is complicated by two factors: first, even if we know $\mathcal{H}$, we still do not know $\mathbb{E}[m(W_i, \gamma_0)]$; second, the functional space $\mathcal{H}$ is, in fact, unknown and must be prespecified by the researcher.

In this paper, we advocate examining an approximate version of (4.2) in which we first replace $\mathbb{E}[m(W_i, \gamma_0)]$ with its sample counterpart, $\mathbb{E}_n[m(W, \gamma_0)]$, and then choose $\mathcal{H}$ as a sequence of finite dimensional linear spaces, $\mathcal{H}_b$, which will be introduced shortly. Replacing $\mathbb{E}[m(W_i, \gamma_0)]$ with $\mathbb{E}_n[m(W, \gamma_0)]$, we align our criterion (4.2) with the finite-sample MSE criterion studied by Donoho (1994) and Armstrong and Kolesár (2018). See Remark 4.1 for additional discussion on our connection to that study. Our choice of $\mathcal{H}_b$ is motivated by the common practice among researchers who often use many controls or technical terms to approximate γ_0. Also see Remark 4.2 for additional discussion.

To proceed, let $p(w) := (p_1(w), p_2(w), \ldots, p_k(w))'$ be a vector of $k := k(n)$ terms selected by the researcher. The linear span of $p(w)$ is:

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

Our treatment of p and Θ_n includes two cases frequently analyzed in the literature:

1. $p(w)$ could be a vector of basis functions that the researcher specifies to approximate γ_0. In this case, Θ_n is a standard series (i.e., linear sieve) space found in the literature on nonparametric estimation.

2. Alternatively, $p(w) = (1, w')'$ is a scenario in which researchers work with many controls and are willing to assume γ_0 linear-in-controls. In this case, technical terms are not necessarily involved.

Either way, we aim to select an estimator with low finite-sample MSE, if the true γ_0 indeed lies in Θ_n. For a vector $\beta = (\beta_1, \ldots \beta_k)' \in \mathbb{R}^k$, let $\|\beta\| = (\sum_{j=1}^k \beta_j^2)^{1/2}$ be its l_2 norm. Let

$$\mathcal{H}_b := \left\{g : \mathcal{W} \mapsto \mathbb{R} \mid g(w) = \beta'p(w), \beta \in \mathbb{R}^k, \|\beta\| \leq b, b < \infty\right\} \subseteq \Theta_n, \tag{4.3}$$

a small ball contained in Θ_n, where b is an l_2 norm bound of the coefficient that the researcher is willing to impose. We consider an approximate version of (4.2) defined as:

$$\begin{aligned}\overline{\text{AMSE}}_n(\tilde{\theta}_{BP}(\alpha)) &:= \sup_{\gamma_0\in\mathcal{H}_b}\left(\mathbb{E}_n[\alpha(W)\gamma_0(W)] - \mathbb{E}_n[m(W,\gamma_0)]\right)^2 + \frac{\sigma^2}{n}\mathbb{E}_n[\alpha^2(W)] \qquad (4.4)\\ &= b^2\sup_{\gamma_0\in\mathcal{H}_1}\left(\mathbb{E}_n[\alpha(W)\gamma_0(W)] - \mathbb{E}_n[m(W,\gamma_0)]\right)^2 + \frac{\sigma^2}{n}\mathbb{E}_n[\alpha^2(W)].\end{aligned}$$

It follows that the BP estimator,

$$\tilde{\theta}_{BP} := \tilde{\theta}_{BP}(\tilde{\alpha}) = \mathbb{E}_n[\tilde{\alpha}(W)Y], \qquad (4.5)$$

where

$$\overline{\text{AMSE}}_n(\tilde{\theta}_{BP}) \in \min_{\alpha\in\Theta_n}\overline{\text{AMSE}}_n(\tilde{\theta}_{BP}(\alpha))$$

is our proposed ***approximate minimax optimal*** estimator. That is, $\tilde{\theta}_{BP}$ minimizes $\overline{\text{AMSE}}_n$ among a class of BP estimators $\tilde{\theta}_{BP}(\alpha)$ with weight function $\alpha \in \Theta_n$.

Remark 4.1. Donoho (1994) and Armstrong and Kolesár (2018) developed a general methodology to derive finite-sample optimal statistical decisions for a class of linear functionals, including sample averages like $\mathbb{E}_n[m(W_i,\gamma_0)]$. Our criterion (4.4) is a special case of that considered by Donoho (1994) and Armstrong and Kolesár (2018), although the target parameter in our paper is the population average $\mathbb{E}[m(W,\gamma_0)]$. Armstrong and Kolesár (2021) and Kallus (2020) applied the framework by Donoho (1994) and Armstrong and Kolesár (2018) to the minimax linear estimation of sample level treatment effects. Their criterion is similar to (4.4), but with $\mathcal{H}_b$ replaced by a general convex space, $\mathcal{H}$, including a common smooth nonparametric functional class, and also including our $\mathcal{H}_b$ as a special case. Part of their main results focus on particular functional classes. Armstrong and Kolesár (2021) derived finite-sample and asymptotic results when $\mathcal{H}$ is a class of Lipschitz continuous functions. Kallus (2020) considered a functional class similar to $\mathcal{H}_b$ but restricted the weight function, α, to only take nonnegative integer-multiple weights with an upper bound. Hirshberg and Wager (2021) also considered a similar criterion but focused on debiased estimators. As a result, Hirshberg and Wager (2021) studied a case in which $\mathcal{H}$ should bound the error term from an initial estimate of γ_0. Unlike previous studies, our criterion $\overline{\text{AMSE}}_n$ focuses on the finite dimensional linear space $\mathcal{H}_b$, when the weight function α takes values in Θ_n. Our criterion is also similar to the setup considered by Li (1982), who derived a minimax linear estimator for a regression function. We establish asymptotic results that, to the best of my knowledge, have not yet been considered elsewhere.

Remark 4.2. If the researcher truly views γ_0 as a nonparametric function, they can select $\mathcal{H}$ as a nonparametric functional class (for example, by applying Armstrong and Kolesár, 2021). In contrast, our approach only considers the worst-case MSE when γ_0 lies in $\mathcal{H}_b$, which is the linear span of the selected basis functions or controls. Our treatment of $\mathcal{H}_b$ follows the common practice among researchers who view γ_0 as linear in parameters. For example, Ferraz and Finan (2011) specified γ_0 as linear in the controls. Our approach then provides an estimator with a low MSE if the linear specification is correct. Even if γ_0 is not linear in the raw controls, researchers often approximate γ_0 via a linear function of sieve terms constructed from the raw controls. In that case, the asymptotic theory often requires a large number of sieve terms to ***undersmooth*** (see, e.g., Newey et al., 1998), so that the approximation error of the estimator is of a smaller order than its variance asymptotically. Our approach thus offers an estimator that mitigates the finite-sample MSE that may be incurred due to undersmoothing.[6] In Section 6.2, we compare the performance of our estimator and that proposed by Armstrong and Kolesár (2021) in a simulation study. We find that our estimator performs well even though it considers only the MSE in $\mathcal{H}_b$ and not the whole space, $\mathcal{H}$. The estimator by Armstrong and Kolesár (2021) also performs competitively if the space $\mathcal{H}$ and the distance measure on W are selected properly. Thus, our approach may be more attractive when researchers do not have a clear perspective of either $\mathcal{H}$ or the distance measure.

4.2 Implementation

One advantage of focusing on $\mathcal{H}_b$ is that the associated minimax exercise is easy to solve. Let $\lambda_n := \frac{\sigma^2}{b^2 n}$. It follows that the optimal weight function is equivalently characterized as follows:

$$\tilde{\alpha} \in \arg\min_{\alpha \in \Theta_n} \left\{ \sup_{\gamma_0 \in \mathcal{H}_1} \left(\mathbb{E}_n[\alpha(W)\gamma_0(W)] - \mathbb{E}_n[m(W,\gamma_0)]\right)^2 \; + \; \lambda_n \mathbb{E}_n[\alpha^2(W)] \right\}. \quad (4.6)$$

Let $m(W_i, p) = (m(W_i, p_1), \dots m(W_i, p_k))'$ be a column vector for each $i = 1 \dots n$. The following proposition further characterizes the minimax problem (4.6) as a minimum-distance problem.

Proposition 4.1. *For each* $\alpha \in \Theta_n$*,*

$$\sup_{\gamma_0 \in \mathcal{H}_1} \left(\mathbb{E}_n[\alpha(W)\gamma_0(W)] - \mathbb{E}_n[m(W,\gamma_0)]\right)^2 = \left\|\mathbb{E}_n[m(W,p) - \alpha(W)p(W)]\right\|^2.$$

By Proposition 4.1, the optimal weight $\tilde{\alpha}$ can be computed as follows:

$$\tilde{\alpha} \in \arg\min_{\alpha \in \Theta_n} \left\{ \underbrace{\left\|\mathbb{E}_n[m(W,p) - \alpha(W)p(W)]\right\|^2}_{\text{minimum distance}} + \underbrace{\lambda_n \mathbb{E}_n[\alpha^2(W)]}_{\text{ridge-style penalty}} \right\}. \quad (4.7)$$

[6] I would like to thank one of the referees for pointing this out.

The first half of the objective function in (4.7) is the squared Euclidean distance, and the second half is a ridge-style penalty in Θ_n. Interestingly, such a structure is similar to the PSMD estimator examined in Chen and Pouzo (2012, 2015) but with an inherently different motivation. The result in (4.7) also agrees with earlier results derived by Li (1982), who provided a similar characterization of the optimal linear weights in the context of estimating a regression function. Therefore, our results in this section can be viewed as a special case of Li (1982) applied to the average regression functionals, justifying the ridge-style minimum-distance estimator as a finite-sample minimax estimator. Note that the objective function in (4.7) is convex and continuously differentiable. From the first order condition for (4.7), we can further derive the following analytic solution:

$$\begin{aligned} \tilde{\alpha} &= p'\tilde{a}, & \tilde{a} &:= (\hat{G}\hat{G} + \lambda_n \hat{G})^{-}\hat{G}\hat{P}, \\ \hat{G} &:= \mathbb{E}_n[p(W)p(W)'], & \hat{P} &:= \mathbb{E}_n[m(W,p)], \end{aligned} \tag{4.8}$$

and $(\cdot)^{-}$ denotes the Moore-Penrose inverse. By construction, when $\hat{G}$ is invertible, $\tilde{a} = (\hat{G}\hat{G} + \lambda_n\hat{G})^{-1}\hat{G}\hat{P} = (\hat{G} + \lambda_n I)^{-1}\hat{P}$ and $\tilde{\alpha}$ is the unique solution of (4.7); when $\hat{G}$ is not invertible, the set $S := \{a : a \in \mathbb{R}^k \text{ and solves } (4.7)\}$ contains many solutions. It follows from Harville (1998, Theorem 20.6.1) that $\|\tilde{a}\| = \min_{a\in S}\|a\|$. That is, $\tilde{a}$ is a unique solution with a minimum l_2 norm.[7]

Remark 4.3. If b^2 is much larger than σ^2, then λ_n will be very close to zero. The estimator practically becomes one with no penalization. In general, if the ratio $\frac{\sigma^2}{b^2} \in (0,\infty)$, then λ_n approaches zero at a rate of $\frac{1}{n}$ as n becomes large. Since $\frac{\sigma^2}{b^2}$ is, in fact, an unknown object, we recommend two practical methods to select λ_n. If the researcher views (4.7) as a PSMD estimator, they can use a data-driven method to select λ_n (e.g., cross validation). If the researcher views (4.7) as a finite-sample minimax estimator, we recommend conducting a sensitivity analysis against a range of possible values of λ_n. For example, we can gauge the magnitude of b by $\hat{b}$, the l_2 norm of the coefficient from regressing Y_i on $p(W_i)$. In addition, σ^2 can be estimated by the estimator $\hat{\sigma}^2$ from the residuals. We can then explore how $\tilde{\theta}_{BP}$ changes when λ_n takes different values in $\left[\frac{C\hat{\sigma}^2}{n\hat{b}^2}, \frac{1}{C}\frac{\hat{\sigma}^2}{n\hat{b}^2}\right]$, an interval for which $C > 0$ is a scale number selected by the researcher.

Remark 4.4. Our analyses focus on BP estimators of form $\tilde{\theta}_{BP}(\alpha) = \mathbb{E}_n[\alpha(W)Y]$, which include many direct plug-in estimators of form $\hat{\theta}_{DP}(\gamma) := \mathbb{E}_n[m(W,\gamma)]$ with some $\gamma \in \Theta_n$. For example, $\hat{\theta}_{DP}(\hat{\gamma}_B)$ with

$$\hat{\gamma}_B(w) := p(w)'\hat{\beta}_B, \hat{\beta}_B := B\mathbb{E}_n[p(W)Y] \tag{4.9}$$

[7] For asymptotic theory, our conditions always ensure that $\hat{G}$ converges to a positive definite matrix with a probability approaching 1. Therefore, the discontinuity of the Moore–Penrose inverse at singular values does not affect our asymptotic results.

for some $k \times k$ matrix B is numerically equivalent to the BP estimator using $\hat{\alpha}_B(w) := p(w)'B\hat{P}$. In particular, when $B = G^-$, $\hat{\gamma}_B$ becomes the least squares estimator:

$$\hat{\gamma}_{LS}(w) := p(w)'\hat{\beta}_{LS}, \hat{\beta}_{LS} := \hat{G}^-\mathbb{E}_n[p(W)Y]. \tag{4.10}$$

As a result, $\hat{\theta}_{DP}(\hat{\gamma}_{LS})$ is numerically equivalent to $\tilde{\theta}_{BP}(\hat{\alpha}_{NR})$, where

$$\hat{\alpha}_{NR}(w) = p(w)'\hat{G}^-\hat{P} \tag{4.11}$$

is the series estimator for α_0 proposed by Newey and Robins (2018).

5 Asymptotic results

This section studies the asymptotic properties of $\tilde{\theta}_{BP}$ for some $\lambda_n \geq 0$ not necessarily equal to the ideal coefficient $\frac{\sigma^2}{b^2 n}$ in the homoscedastic model discussed in Section 4.2. We introduce the following notation. For a vector $\beta = (\beta_1, \dots \beta_k)' \in \mathbb{R}^k$, let $\|\beta\|_\infty := \max_{1\leq j\leq k} |\beta_j|$. For a function $f : \mathcal{W} \mapsto \mathbb{R}$, let $\|f\|_{\mathbb{P},q} := \left[\int |f(w)|^q \, d\mathbb{P}(w)\right]^{1/q}, 1 \leq q \leq \infty$ denote its $L^q(\mathbb{P})$ norm. In particular, $\|f\|_{\mathbb{P},\infty} := \sup_{w\in\mathcal{W}} |f(w)|$. For two sequences of numbers, a_n and b_n, $a_n \vee b_n := \max\{a_n, b_n\}$, $a_n \wedge b_n := \min\{a_n, b_n\}$, and $a_n \lesssim b_n$ means that $a_n \leq cb_n$ for some constant c that does not depend on n. Furthermore, $\mathcal{L}_n f$ denotes the least squares projection of $f \in \mathcal{H}$ onto Θ_n. For example, $\mathcal{L}_n\gamma_0 = \beta_l'p$, where

$$\beta_l := \arg\min_{\beta\in\mathbb{R}^k} \mathbb{E}[\gamma_0(W_i) - \beta'p(W_i)]^2$$

is the projection coefficient. Then, we can rewrite (4.1) as follows:

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

where $u_{\gamma_0} := \gamma_0 - \beta_l'p$ is the least squares approximation error and $u_{\gamma_0 i} := \gamma_0(W_i) - \beta_l'p(W_i)$. We impose the following assumptions throughout this section.

Assumption O.

1. *For each n, $\left\{(Y_i, W_i')'\right\}_{i=1}^n$ are independently and identically distributed. The function $m(w, \cdot)$ is linear in the sense of (3.2) and $\mathbb{E}[m^2(W_i, \gamma_0)] \lesssim 1$. The RR α_0 exists, satisfies (3.4) and $\mathbb{E}[\alpha_0^2(W_i)] \lesssim 1$.*

2. *$\mathbb{E}[e_i|W_i = w] = 0$, $\sup_{w\in\mathcal{W}} \mathbb{E}[e_i^2|W_i = w] \lesssim 1$ for each $i = 1 \dots n$.*

3. *For each $i = 1 \dots n$ and $j = 1 \dots k$, $\mathbb{E}[m^2(W_i, p_j)] \lesssim \mathbb{E}[p_j^2(W_i)]$, $\mathbb{E}[m^2(W_i, u_{\gamma_0})] \lesssim \mathbb{E}u_{\gamma_0 i}^2$.*

According to Assumption O(1), for each n, data $\left\{(Y_i, W_i')'\right\}_{i=1}^n$ are n i.i.d. copies from the population distribution $\mathbb{P}$ of $(Y, W')'$ and admit the sampling distribution denoted as $\mathbb{P}^n$. Note that our setup allows d_W, the dimension of W, to be either fixed or growing. If d_W is fixed, then Assumption O(1) implies the standard array asymptotics. If the dimension of W is growing as $n \to \infty$, then d_W should be understood as $d_{W,n}$, indexed by n. In the latter case, the population distribution $\mathbb{P}$ should also be understood as being indexed by n as well (i.e., $\mathbb{P} := \mathbb{P}_n$). In addition, the sampling distribution is more precisely denoted $\mathbb{P}^n := \mathbb{P}_n^n$. To reduce the notational burden, we do not differentiate between these two cases. Assumption O(2) restricts the behavior of the first two conditional moments of e_i. The mean independence condition $\mathbb{E}[e_i|W_i = w] = 0$ is automatically satisfied by the definition of γ_0. We also assume that e_i has a finite conditional variance, which is standard but might be weakened by imposing higher unconditional moments for e and W (see, e.g., Hansen (2015)). Assumption O(3) imposes a sufficient degree of continuity on the structure of $\mathbb{E}[m^2(W_i, \cdot\,)]$. If Assumption O(3) is not satisfied, then the functional form of $m(w, \cdot\,)$ might adversely affect the asymptotic performance of our estimator via $d_k^p := \sum_{j=1}^k \mathbb{E}[m^2(W_i, p_j)]$ and $d_k^u := \mathbb{E}[m^2(W_i, u_{\gamma_0})]$. As long as $\frac{d_k^p}{n} \to 0$ and $d_k^u \to 0$ as $n \to \infty$, we can still accommodate a remainder rate that is possibly slower than that presented in Theorem 5.1 below. It is straightforward to see that Assumption O(3) is satisfied in Examples 3.1, 3.2, 3.4 and 3.6 without additional regularity conditions. Proposition 5.1 provides regularity conditions under which Assumption O(3) is met for the other examples and the form of d_k^p and d_k^u when the regularity conditions are not satisfied.

Proposition 5.1. *(i) In Example 3.3, Assumption O(3) is satisfied if:* $\sup_{w\in\mathcal{W}} |\omega(w)| \lesssim 1$, *and there exists some constants* C_1 *and* C_2 *such that* $\mathbb{E}\left[\left(\frac{\partial p_j(W_i)}{\partial W_1}\right)^2\right] \leq C_1 \mathbb{E}\left[p_j(W_i)^2\right]$ *and* $\mathbb{E}\left[\left(\frac{\partial u_{\gamma_0}(W_i)}{\partial W_1}\right)^2\right] \leq C_2 \mathbb{E}\left[u_{\gamma_0}^2(W_i)\right]$ *for all* $i = 1 \dots n$, $j = 1 \dots k$. *Otherwise,* $d_k^p \leq k \max_{j\in\{1\dots k\}} \mathbb{E}\left[\left(\omega(W_i)\frac{\partial p_j(W_i)}{\partial W_1}\right)^2\right]$, $d_k^u = \mathbb{E}\left[\left(\omega(W_i)\frac{\partial u_{\gamma_0}(W_i)}{\partial W_1}\right)^2\right]$.

(ii) In Example 3.5, Assumption O(3) is satisfied if: $\sup_{z\in\mathcal{Z}} |\omega(z)| \lesssim 1$, *and there exists some constants* C_3 *and* C_4 *such that* $\mathbb{E}\left[\left(\int_{r_0}^{r_1} p_j(\tilde{r}, Z_i) d\tilde{r}\right)^2\right] \leq C_3 \mathbb{E}\left[p_j(W_i)^2\right]$ *and* $\mathbb{E}\left[\left(\int_{r_0}^{r_1} u_{\gamma_0}(\tilde{r}, Z_i) d\tilde{r}\right)^2\right] \leq C_4 \mathbb{E}\left[u_{\gamma_0}^2(W_i)\right]$ *for all* $i = 1 \dots n$, $j = 1 \dots k$. *Otherwise,* $d_k^p \lesssim (\sup_{w\in\mathcal{W}} \|p(w)\|)^2$, *and* $d_k^u \lesssim \left(\|u_{\gamma_0}\|_{\mathbb{P},\infty}\right)^2$.

We now introduce an essential but standard condition on the basic quality of the approximating term, p.

Assumption S.

1. *All eigenvalues of $G := \mathbb{E}[p(W_i)p(W_i)']$ are bounded from above and away from zero uniformly over all 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*

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

Assumption S(1) requires that p should not be too collinear or grow too quickly, similar to Chen and Pouzo (Assumption C.1(iii), 2012) and Belloni et al. (Condition A.2, 2015).[8] Assumption S(2) imposes mild restrictions on the approximation quality of the linear span Θ_n. Note that both $\mathbf{r}_{\gamma_0}$ and $\mathbf{r}_{\alpha_0}$ depend on k but we omit indexing by k for notational simplicity. If $\mathbf{r}_{\gamma_0} \to 0$ as $k \to \infty$, then we say that γ_0 is correctly specified; If $\mathbf{r}_{\gamma_0} \nrightarrow 0$ as $k \to \infty$, then we say that γ_0 is mis-specified. The correct specification and mis-specification of α_0 are defined analogously. For example, if p is a vector of basis functions, we often assume that γ_0 and α_0 are within a Hölder smoothness class of order s. It follows that $\mathbf{r}_{\gamma_0} = k^{-\eta_\gamma}, \mathbf{r}_{\alpha_0} = k^{-\eta_\alpha}$ for some non-negative constants η_γ and η_α that depend on s, d_W and p. See DeVore and Lorentz (1993); Newey (1997); Chen (2007), among others, for more details on the approximation results with various basis functions. For another example, if researchers work with a linear specification where p is a vector of controls not involving technical terms, they have implicitly assumed $\mathbf{r}_{\gamma_0} = 0$, as in the case of Ferraz and Finan (2011).

Assumption P. $\lambda_n = o\left(\frac{1}{\sqrt{n}}\right)$.

Assumption P imposes asymptotic order on penalty coefficient λ_n. Recall that if we view our estimator as a finite-sample minimax estimator in a homoscedastic model, then $\lambda_n = \frac{\sigma^2}{b^2 n}$. That is, λ_n should diminish to zero at rate $O(\frac{1}{n})$ (see also Remark 4.3). However, for the validity of the asymptotic results in Theorems 5.1 and 5.2, we only require λ_n to be $o\left(\frac{1}{\sqrt{n}}\right)$. Thus, Assumption P allows λ_n to converge to zero much more slowly than that required by the finite sample minimax optimality.

5.1 Asymptotic remainder rate

For any estimator $\hat{\theta}$, we can mechanically write

$$\sqrt{n}(\hat{\theta} - \theta_0) = \sqrt{n}\mathbb{E}_n\phi + R_n, \tag{5.1}$$

$$\text{where } \phi_i := \phi(W_i, e_i) := m(W_i, \gamma_0) + \alpha_0(W_i)e_i - \theta_0, \tag{5.2}$$

[8] For example, if p is a vector of basis functions, Assumption S(1) is satisfied if p is orthonormal with respect to the Lebesgue measure and the density of W_i is bounded away from zero and from above.

and call R_n the remainder term. Since $\mathbb{E}_n\phi$ is an average of i.i.d. mean zero random variables, $\sqrt{n}\mathbb{E}_n\phi \xrightarrow{d} N(0, \mathbb{E}\phi_i^2)$ under suitable conditions. Thus, for any estimator $\hat{\theta}$, if we can show that its remainder term $R_n = o_p(1)$, then $\hat{\theta}$ is said to be asymptotically linear and normal. Many estimators can achieve asymptotic normality under some conditions. To further select between those estimators, Newey and Robins (2018) proposed finding one such that $R_n \xrightarrow{p} 0$ as quickly as possible as $n \to \infty$. In this section, we derive the asymptotic distribution of $\tilde{\theta}_{BP}$ when $\frac{k}{n} \to 0$ (up to log terms), which enables us to compare the asymptotic quality of our estimator with others in the literature.

Let $\xi_k = \sup_{w\in\mathcal{W}} \|p(w)\|$ (Newey, 1997). Also, we define the following:

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

where $\overline{\mathcal{H}} := \{g + a'p : g \in \mathcal{H}, a \in \mathbb{R}^k\}$. As shown by Huang (2003); Belloni et al. (2015); Chen and Christensen (2015), for certain basis functions, ℓ_k exploits the stability properties of the projection and enables us to impose weaker requirements on the growth rate of k. We now demonstrate the following rate conditions.

Assumption L. $\frac{\xi_k^2 \log k}{n} = o(1)$, $\sqrt{n}\mathbf{r}_{\alpha_0}\mathbf{r}_{\gamma_0} = o(1)$, $\mathbf{r}_{\alpha_0} = o(1)$, $(\|\alpha_0\|_{\mathbb{P},\infty} \wedge \ell_k)\mathbf{r}_{\gamma_0} = o(1)$.

To make sense of Assumption L, suppose the support of W is compact and that we choose a spline or wavelet series as a basis function. It follows that $\xi_k \lesssim \sqrt{k}$ and $\ell_k \lesssim 1$. Then, Assumption L becomes as follows:

$$\frac{k \log k}{n} = o(1), \quad \sqrt{n}\mathbf{r}_{\alpha_0}\mathbf{r}_{\gamma_0} = o(1), \quad \mathbf{r}_{\gamma_0} = o(1), \quad \mathbf{r}_{\alpha_0} = o(1), \tag{5.3}$$

which is the weakest known condition in the literature such that $R_n = o_p(1)$ in (5.1) (Robins et al., 2009; Newey and Robins, 2018). Note that (5.3) allows $\frac{k}{n} = o(1)$ up to the $\log k$ term.

Theorem 5.1. *Suppose Assumptions O, S, P and L hold true and* $\frac{\xi_k^2}{\sqrt{n}}\mathbf{r}_{\gamma_0} = o(1)$. *Then*

(i) $\sqrt{n}(\tilde{\theta}_{BP} - \theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \phi_i + O_p\left(\frac{\xi_k^2}{\sqrt{n}}\mathbf{r}_{\gamma_0}\right)$.

(ii) $\sqrt{n}\left\{\mathbb{E}[\phi_i^2]\right\}^{-\frac{1}{2}}(\tilde{\theta}_{BP} - \theta_0) \xrightarrow{d} N(0,1)$ as $n \to \infty$ if in addition:

$$\mathbb{E}\phi_i^2 \text{ is uniformly bounded away from zero for all } i \text{ and } n. \tag{5.4}$$

Theorem 5.1(i) states that $\tilde{\theta}_{BP}$ is $\sqrt{n}$ consistent and has a remainder rate $O_p\left(\frac{\xi_k^2}{\sqrt{n}}\mathbf{r}_{\gamma_0}\right)$. Since Assumption O guarantees that $\mathbb{E}\phi_i^2 \lesssim 1$, $\sqrt{n}$ estimation of θ_0 is possible. In the context of estimating the average treatment effect, we effectively rule out any situation with limited overlap. As a counterexample, if there exists some $x \in \mathcal{X}$ such that $\Pr\{T_i =$

$1|X_i = x\}$ can become arbitrarily small, then $\mathbb{E}\phi_i^2$ can be infinite (Khan and Tamer, 2010). In such a scenario, Assumption O(1) is violated, $\frac{1}{\sqrt{n}}\sum_{i=1}^n \phi_i$ can diverge and $\sqrt{n}$-consistent estimators of θ_0 might not even exist. Our paper does not involve that case. Theorem 5.1(ii) says that $\tilde{\theta}_{BP}$ is $\sqrt{n}$ normal if we additionally impose (5.4). This is a technical condition to accommodate the situation in which d_W is growing as $n \to \infty$ and $\mathbb{E}[\phi_i^2]$ is indexed by n as a result. Without (5.4), the normalization term $\{\mathbb{E}[\phi_i^2]\}^{-\frac{1}{2}}$ can go to infinity as n becomes large. Thus, (5.4) ensures that the remainder terms are still $o_p(1)$ after normalization. If d_W is fixed, (5.4) can be dropped, and our estimator becomes, in fact, semiparametrically efficient.

Compared to the minimal rate condition, Assumption L, we additionally need $\frac{\xi_k^2}{\sqrt{n}}\mathbf{r}_{\gamma 0} = o(1)$. This additional requirement stems from controlling the ***own-observation bias*** or ***overfitting bias*** of form $\sqrt{n}\mathbb{E}_n[(\tilde{\alpha} - \alpha_0)u_{\gamma_0}]$. In Lemma D.6, we show that this own-observation bias term consists of two parts not converging under Assumption L: a nonlinear approximation bias $\mathbb{E}\big[(m(W,p)-\alpha_0 p)' G p u_{\gamma_0}\big]/\sqrt{n}$ of order $O(\frac{\xi_k^2}{\sqrt{n}}\mathbf{r}_{\gamma_0})$ and a bilinear estimation error $\sqrt{n}\left(\mathbb{E}_n[m(W,p) - \alpha_0 p]\right)'\left(\hat{G}^{-1} - G\right)\mathbb{E}_n[p u_{\gamma_0}]$ of order $O_p\left(\frac{\xi_k^3\sqrt{\log k}}{n}\mathbf{r}_{\gamma_0}\right)$ resulting from estimating G by $\hat{G}$.

Remark 5.1. It is useful to compare our asymptotic result with other similar ones in the literature. Recall the direct plug-in estimator $\hat{\theta}_{DP}(\hat{\gamma}_{LS})$ in Remark 4.4. Let I_l, $l = 1 \dots L$ be a partition of the sample index set $\{1, \dots n\}$ divided into L distinct subsets of about equal size. An $L-$fold cross-fitted direct plug-in estimator is defined as follows:

$$\hat{\theta}_{CFDP}(\hat{\gamma}_{LS}) := \frac{1}{n}\sum_{l=1}^{L}\sum_{i\in l} m(W_i, \hat{\gamma}_{LS}^l),$$

where $\hat{\gamma}_{LS}^l$ is the least squares estimator using only observations ***not*** in I_l. Recall the influence function ϕ_i defined in (5.2). By making use of the moment condition $\mathbb{E}[\phi_i] = 0$, Newey and Robins (2018) considered a doubly cross-fitted debiased estimator as follows:

$$\hat{\theta}_{DCFD}(\hat{\alpha}_{NR}, \hat{\gamma}_{LS}) := \frac{1}{n}\sum_{l=1}^{L}\sum_{i\in l}[m(W_i, \hat{\gamma}_{LS}^l) + \hat{\alpha}_{NR}^l(W_i)(Y_i - \hat{\gamma}_{LS}^l(W_i))],$$

where $\hat{\alpha}_{NR}^l$ is the series estimator defined in (4.11) but only uses observations ***not*** in I_l. Newey and Robins (2018) showed that with certain basis functions

$$\sqrt{n}(\hat{\theta}_{CFDP} - \theta_0) = \sqrt{n}\mathbb{E}_n\phi + O_p\left(\mathbf{r}_{\gamma 0}\frac{k}{\sqrt{n}}\right),$$
$$\sqrt{n}(\hat{\theta}_{DCFD} - \theta_0) = \sqrt{n}\mathbb{E}_n\phi + O_p\left(\mathbf{r}_{\gamma 0}\frac{k^2}{n^{3/2}}\right).$$

Therefore, upon choosing a suitable basis function, the remainder rate of our estimator is as fast as $\hat{\theta}_{CFDP}$ but it is slower than $\hat{\theta}_{DCFD}$.[9]

Remark 5.2. One might be tempted to debias $\tilde{\theta}_{BP}$ in the hope of improving the asymptotic remainder rate. Let $\hat{\gamma} \in \Theta_n$ be some estimator of γ_0. A debiased version of $\tilde{\theta}_{BP}$ (without cross-fitting) is:

$$\hat{\theta}_D(\tilde{\alpha}, \hat{\gamma}) := \mathbb{E}_n[m(W, \hat{\gamma}) + \tilde{\alpha}(W)(Y - \hat{\gamma}(W))].$$

However, in our framework, $\hat{\theta}_D(\tilde{\alpha}, \hat{\gamma})$ actually does not improve the asymptotic performance of $\tilde{\theta}_{BP}$. Specifically, we may write the following:

$$\sqrt{n}(\hat{\theta}_D - \theta_0) = \sqrt{n}\mathbb{E}_n\phi + R_n(\hat{\theta}_D),$$

where $R_n(\hat{\theta}_D)$ is the remainder of $\hat{\theta}_D$. In general, ${}^{R_n(\hat{\theta}_D)}/_{\frac{\xi_k^2}{\sqrt{n}}\mathbf{r}_{\gamma_0}} \overset{p}{\not\to} 0$, implying that the remainder of $\hat{\theta}_D$ will not converge to zero more quickly than the remainder of $\tilde{\theta}_{BP}$. To see this, note that the remainder of $\tilde{\theta}_{BP}$ has a structure as follows:

$$\begin{aligned}
R_n(\tilde{\theta}_{BP}) &= R_{n,1}(\tilde{\theta}_{BP}) + R_{n,2}(\tilde{\theta}_{BP}), \text{ where} \\
R_{n,1}(\tilde{\theta}_{BP}) &= \sqrt{n}\mathbb{E}_n\left[\tilde{\alpha}\gamma_0 - m(W, \gamma_0)\right], \\
R_{n,2}(\tilde{\theta}_{BP}) &= \sqrt{n}\mathbb{E}_n\left[\left(\tilde{\alpha} - \alpha_0\right)e\right],
\end{aligned}$$

while the remainder of $\hat{\theta}_D$ can be decomposed as:

$$\begin{aligned}
R_n(\hat{\theta}_D) &= R_{n,1}(\hat{\theta}_D) + R_{n,2}(\hat{\theta}_D), \text{ where} \\
R_{n,1}(\hat{\theta}_D) &= \sqrt{n}\mathbb{E}_n[\tilde{\alpha}(\gamma_0 - \hat{\gamma}) - m(W, (\gamma_0 - \hat{\gamma}))] \\
R_{n,2}(\hat{\theta}_D) &= R_{n,2}(\tilde{\theta}_{BP}).
\end{aligned}$$

That is, $\tilde{\theta}_{BP}$ and $\hat{\theta}_D$ share the same variance term ($R_{n,2}(\hat{\theta}_D) = R_{n,2}(\tilde{\theta}_{BP})$) and only differ in the main bias terms ($R_{n,1}(\tilde{\theta}_{BP})$ v.s. $R_{n,1}(\hat{\theta}_D)$). If Assumptions O, S, P and L hold true, we can decompose the following:

$$\begin{aligned}
R_{n,1}(\tilde{\theta}_{BP}) &= \underbrace{\sqrt{n}\mathbb{E}_n[\tilde{\alpha}\mathcal{L}_n\gamma_0 - m(W, \mathcal{L}_n\gamma_0)]}_{A_n} + \underbrace{\sqrt{n}\mathbb{E}_n[(\tilde{\alpha} - \alpha_0)u_{\gamma_0}]}_{B_n} + o_p(1), \\
R_{n,1}(\hat{\theta}_D) &= \underbrace{\sqrt{n}\mathbb{E}_n\left[\tilde{\alpha}(\mathcal{L}_n\gamma_0 - \hat{\gamma}) - m\left(W, (\mathcal{L}_n\gamma_0 - \hat{\gamma})\right)\right]}_{C_n} + \underbrace{\sqrt{n}\mathbb{E}_n[(\tilde{\alpha} - \alpha_0)u_{\gamma_0}]}_{B_n} + o_p(1).
\end{aligned}$$

It turns out that $A_n \xrightarrow{p} 0$ and B_n converges to zero more slowly than A_n. Thus, the

[9] We note that cross-fitted plug-in and doubly cross-fitted debiased estimators in Newey and Robins (2018) can meet the minimal requirement (5.3) only in special cases in which the Holder orders of α_0 and γ_0 are low enough and p is a Haar basis function. Doubly cross-fitted debiased estimators can also meet minimal requirements if the functional is the expected conditional covariance.

overall remainder rate of $\tilde{\theta}_{BP}$ is determined by term B_n. In contrast, $\hat{\theta}_D$ has a different remainder bias term C_n, but still faces the same remainder term B_n. By selecting a suitable estimator, $\hat{\gamma} \in \Theta_n$, to ***debias***, one can show that $C_n \xrightarrow{p} 0$. But in general, the term B_n remains, even for the debiased estimator. Thus, term B_n also determines the overall remainder rate of $\hat{\theta}_D$. Since the remainder rates of both $\tilde{\theta}_{BP}$ and $\hat{\theta}_D$ depend on term B_n, $\hat{\theta}_D$ will not perform better than $\tilde{\theta}_{BP}$ asymptotically, in general.

We conclude this section by presenting a simple, consistent estimator for the asymptotic variance by making use of $\tilde{\alpha}$ and $\hat{\gamma}_{LS}$. Other consistent estimators of γ_0 can also be considered, and similar conditions can be established accordingly.

Theorem 5.2. *Let*

$$\hat{\Omega} = \left| \mathbb{E}_n \left[m(W, \hat{\gamma}_{LS}) + \tilde{\alpha}(W)(Y - \hat{\gamma}_{LS}(W)) \right]^2 - \tilde{\theta}_{BP}^2 \right|, \tag{5.5}$$

where $\hat{\gamma}_{LS}$ is defined in (4.10). Suppose all conditions of Theorem 5.1 and the following conditions hold:

(i) For each $\gamma = \beta' p \in \Theta_n$ where $\|\beta\| < \infty$, $\mathbb{E}[m^2(W_i, \gamma)] \lesssim \mathbb{E}[\gamma^2(W_i)]$.

(ii) $\|\hat{\gamma}_{LS} - \gamma_0\|_{\mathbb{P},\infty} = o_p(1)$.

(iii) 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, $\sqrt{n} \left[\hat{\Omega}\right]^{-\frac{1}{2}} (\tilde{\theta}_{BP} - \theta_0) \xrightarrow{d} N(0, 1)$.

Condition (i) in Theorem 5.2 plays a similar role to that of Assumption O(3) and can be verified in a similar fashion. The uniform consistency condition $\|\hat{\gamma}_{LS} - \gamma_0\|_{\mathbb{P},\infty} = o_p(1)$ can be verified by additional preliminary conditions. For example, when d_W is fixed, Belloni et al. (2015, Theorem 4.3) and Chen and Christensen (2015, Lemma 2.4) established an optimal sup-norm convergence for $\hat{\gamma}_{LS}$, allowing $\frac{k}{n} \to 0$ up to log terms. It is also possible to relax the uniform consistency requirement by imposing higher moment conditions for basis functions (see Hansen 2015). Moreover, we see a trade-off between the existence of higher moments for e_i and the 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, a consistent estimation of the variance demands that k must grow even more slowly. Notice that we do not require $\|\tilde{\alpha} - \alpha_0\|_{\mathbb{P},\infty} = o_p(1)$.

5.2 $\sqrt{n}$ normality with many-regressor asymptotics

Another nice asymptotic property of $\tilde{\theta}_{BP}$ is that it can still achieve $\sqrt{n}$ normality with many regressors in the sense that $\frac{k}{n} < 1$ but does not diminish to 0. To work with many-regressor asymptotics, we first present a new set of rate conditions.

Assumption M.

1. $\frac{\xi_k^2}{n} \leq 1$, $\mathbf{r}_{\gamma_0} = o(\frac{1}{\sqrt{n}})$, $\mathbf{r}_{\alpha_0} = O(1)$, $(\|\alpha_0\|_{\mathbb{P},\infty} \wedge \ell_k)\mathbf{r}_{\gamma_0} = o(1)$;
2. *All eigenvalues of $\hat{G}$ are bounded away from zero with a probability approaching one (wpa1);*
3. $\lambda_n = o\left(\frac{1}{\sqrt{n}\log k}\right)$.

Assumption M is the counterpart of Assumptions L and P when we allow k to grow proportionally to the sample size. Note that Assumption M(1) allows $\tilde{\alpha}$ to be inconsistent and even mis-specified. A key condition that makes such a relaxation possible is $\mathbf{r}_{\gamma_0} = o(\frac{1}{\sqrt{n}})$. If p is treated as a vector of basis functions, it requires that γ_0 be supersmooth. If, in instead, the researcher works with many controls not necessarily involving technical terms, then $\mathbf{r}_{\gamma_0} = o(\frac{1}{\sqrt{n}})$ is directly satisfied by imposing a linear-in-control specification for γ_0. Assumption M(2) is similar to the first condition in Cattaneo et al. (Assumption 2, 2018b). In Lemma C.5, we provide a sufficient condition for Assumption M(2), which requires that $\frac{\xi_k^2}{n}$ converges to a constant strictly less than 0.38 (up to log terms). Compared to Assumption P, Assumption M(3) requires λ_n to converge to zero slightly more quickly than $o(\frac{1}{\sqrt{n}})$. This is still a mild condition on the penalty term, but it allows us to control key remainder terms even though $\frac{\xi_k^2}{n} \nrightarrow 0$.

Theorem 5.3. *Suppose Assumptions O, S and M hold true. Let $\sigma_{\tilde{\alpha}}^2 := \mathbb{E}_n\left[\tilde{\alpha}^2(W)\mathbb{E}[e^2|W]\right]$ and $\sigma_m^2 := \mathbb{E}[m^2(W_i, \gamma_0)] - \theta_0^2$. In addition, suppose*

(i) uniformly over all i, k, and n, $\sup_{w\in\mathcal{W}} \mathbb{E}\left[|e_i|^3 |W_i = w\right] \lesssim 1$, $\inf_{w\in\mathcal{W}} \mathbb{E}\left[e_i^2|W_i = w\right]$ is bounded away from zero, $\|\alpha_0\|_{\mathbb{P},2}^2 - \mathbf{r}_{\alpha_0}^2$ is bounded away from zero;

(ii) $\frac{\sigma_m^2}{\sigma_m^2+\sigma_{\tilde{\alpha}}^2}$ converges in probability to some constant $v \in (0, 1)$.

Then, as $n \to \infty$,

$$\sqrt{n}(\sigma_m^2 + \sigma_{\tilde{\alpha},n}^2)^{-1/2}\left(\tilde{\theta}_{BP} - \theta_0\right) \xrightarrow{d} N(0, 1).$$

Remark 5.3. Theorem 5.3 establishes $\sqrt{n}$ normality of $\tilde{\theta}_{BP}$ without strong convergence conditions for nuisance parameters, cross-fitting, or debiasing. We reiterate that Theorem 5.3 allows both $\frac{k}{n} < 1$ and α_0 to be mis-specified. Suppose that researchers select p as a vector of series terms. When the support of W is compact, we can choose spline or wavelet series as basis functions such that $\xi_k \lesssim \sqrt{k}$ and $\ell_k \lesssim 1$. It follows that Assumption M(1) becomes $\frac{k}{n} \lesssim 1, \mathbf{r}_{\gamma_0} = o(\frac{1}{\sqrt{n}})$ and $\mathbf{r}_{\alpha_0} = O(1)$. If the researchers work with many controls, they often specify $\gamma_0(w) = \beta' p(w)$, where $p(w) = (1, w')'$, implying $\mathbf{r}_{\gamma_0} = 0$. If W is compactly supported, all we need for Assumption M(1) to hold is $\frac{k}{n} \lesssim 1$ and $\mathbf{r}_{\alpha_0} = O(1)$. Conditions (i) and (ii) in Theorem 5.3 are standard and mild regularity conditions.

Remark 5.4. To establish Theorem 5.3, we first note that even if $\tilde{\alpha}$ is not consistent, we can still write:

$$\sqrt{n}(\sigma_m^2+\sigma_{\tilde{\alpha}}^2)^{-1/2}\left(\tilde{\theta}_{BP}-\theta_0\right)=\sqrt{n}(\sigma_m^2+\sigma_{\tilde{\alpha}}^2)^{-1/2}\mathbb{E}_n[\tilde{\alpha}(W)e+m(W,\gamma_0)-\theta_0], \\ +(\sigma_m^2+\sigma_{\tilde{\alpha}}^2)^{-1/2}R_{n,1}(\tilde{\theta}_{BP}).$$

Then, we derive several useful intermediate results: $\{\mathbb{E}_n[\tilde{\alpha}^2(W)]\}^{-1}=O_p(1)$ (Lemma C.2), $R_{n,1}(\tilde{\theta}_{BP})=o_p(1)$ (Lemma C.3) and ${}^{\max_i|\tilde{\alpha}(W_i)|}/_{\sqrt{n}}=o_p(1)$ (Lemma C.4). With these results, we can show that the remainder $(\sigma_m^2+\sigma_{\tilde{\alpha}}^2)^{-1/2}R_{n,1}(\tilde{\theta}_{BP})$ is still $o_p(1)$, and that the main term is asymptotically normal by studying the convergence of its characteristic function and invoking the Berry–Esseen inequality.

6 Simulation studies

6.1 Estimation of population mean with missing data

In this section, we assess the finite-sample performance of $\tilde{\theta}_{BP}$ with a small, fixed penalty when the object of interest is the population mean in Example 3.1. Our data-generating process follows Kang et al. (2007). Let

$$U:=\{U_1,U_2,U_3,U_4\}' \tag{6.1}$$

be a vector of four random variables from multivariate standard normal distribution $N(\mathbf{0},I_4)$. The outcome variable Y^* is:

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

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

$$\pi(u)=\Pr\{T=1|U=u\}=\Lambda(-u_1+0.5u_2–0.25u_3–0.1u_4), \tag{6.2}$$

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

$$X_1:=\exp\left(\frac{U_1}{2}\right),X_2:=\frac{U_2}{1+\exp(U_1)}+10,X_3:=\left(\frac{U_1U_3}{25}+0.6\right)^3,X_4:=(U_2+U_4+20)^2. \tag{6.3}$$

A random sample of size $n = 200$ is drawn from observables $\{Y, T, X'\}'$. Our estimator is then $\tilde{\theta}_{BP} = \mathbb{E}_n[\tilde{\alpha}(X)TY]$, where $\tilde{\alpha}(x) = \tilde{a}'p(x)$,

$$\tilde{a} = \left[\hat{G}_T\hat{G}_T + \lambda_n\hat{G}_T\right]^- \hat{G}_T\hat{P}, \quad \hat{G}_T := \mathbb{E}_n[Tp(X)p'(X)], \quad \hat{P} = \mathbb{E}_n[p(X)]. \tag{6.4}$$

6.1.1 Baseline results with mild selection bias

We first choose B-splines as basis functions and let k gradually increase from 5 to 121, including 11 cases with $\frac{k}{n}$ growing from 0.025 to 0.605. We also ensure that the basis functions involving all the X variables are included from the beginning, even when k is small. This simulates a situation in applied research in which the data contains mild selection bias and the researchers try out various technical terms. I compare the performance of the following three BP estimators: (1) our estimator, $\tilde{\theta}_{BP}$, with a small fixed penalty $\lambda_n = 0.002$; (2) the Newey Robins (NR) estimator in which the RR is computed by (4.11); and (3) the "Simple Ridge" (SR) estimator in which the RR is computed with coefficient $\tilde{a}_{SR} = (\hat{G}_T\hat{G}_T + \lambda_n I)^- \hat{G}_T\hat{P}$, with $\lambda_n = 0.002$. The performance of a simple sample average estimator when α_0 is known is also reported. The biases and the root mean square errors (RMSE) of these estimators after 10,000 experiments are collected in Table 2 and Figure 6.1. Empirical coverage probabilities of these estimators when the variance is estimated using (5.5) are reported in Table 3.

6.1.2 Comparison with debiased estimators with no cross-fitting

Now, we compare the finite-sample performance of $\tilde{\theta}_{BP}$ with that of various debiased estimators. I focus on three popular debiased estimators involving the estimation of propensity scores: (1) one in which the regression function and propensity score are estimated using (generalized) linear methods without selection; (2) one in which the regression function and propensity score are estimated using post lasso; and (3) one in which regression function and propensity score are estimated using lasso. Basis functions and simulation specifications are the same as in Section 6.1.1. We only examine cases with smaller $\frac{k}{n}$ ratios for which we know that these debiased estimators would perform relatively well.[10] The results are reported in Figure 6.2. As we can see clearly, in terms of the RMSE, the debiased estimators either perform strictly worse than or similar to $\tilde{\theta}_{BP}$.

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

6.1.3 Comparison with doubly cross-fitted debiased estimators

We continue to investigate the performance and computational cost of $\tilde{\theta}_{BP}$ versus several doubly cross-fitted debiased estimators. As a benchmark, Panel A of Table 4 reports the RMSE and computational cost (in terms of system time for a common laptop) of $\tilde{\theta}_{BP}$ after 1,000 experiments. We consider the three debiased estimators used in Section 6.1.2 but with L-fold cross-fitting, where $L = 2$ and 4. Because cross-fitting reduces the effective number of observations available for each estimate of the propensity score, doubly cross-fitted debiased estimators involving propensity scores are more likely to be affected by extreme propensity score weights. Following Chernozhukov et al. (2018), we trim the propensity scores at 0.01 and 0.99. The simulation results are reported in Panel B of Table 4. We also examine whether it possible to improve the performance of $\tilde{\theta}_{BP}$ by directly debiasing $\tilde{\theta}_{BP}$ and with cross-fitting. A doubly cross-fitted and debiased version of $\tilde{\theta}_{BP}$ is $\hat{\theta}_{DCFD}(\tilde{\alpha}, \hat{\gamma})$, where $\tilde{\alpha}$ is defined in (4.8), and $\hat{\gamma}$ is some estimator of γ_0. Specifically, we consider $\hat{\gamma}$ to be the OLS, lasso, and post lasso estimators, respectively. The performance of $\hat{\theta}_{DCFD}(\tilde{\alpha}, \hat{\gamma})$ is reported in Panel C of Table 4.

From Table 4, it is clear that most doubly cross-fitted estimators, in fact, perform worse than $\tilde{\theta}_{BP}$ and take much longer to compute. While some doubly cross-fitted estimators can improve the performance of $\tilde{\theta}_{BP}$ when $\frac{k}{n}$ is relatively small, the improvement in RMSE is quite limited and comes at a greater computational cost. For example, when $\frac{k}{n} = 0.045$, debiasing $\tilde{\theta}_{BP}$ with 2-fold cross-fitting and lasso can reduce the RMSE of $\tilde{\theta}_{BP}$ from 5.5 to 4.6, but it takes almost 1.2 minutes to compute for 1,000 simulations as opposed to less than 3 seconds for the case of $\tilde{\theta}_{BP}$. Therefore, even though cross-fitted debiased estimators enjoy superior asymptotic properties, they may incur considerable finite-sample MSE and could be less robust to the number of basis functions chosen by the researcher.

6.1.4 Robustness check

Considerable bias We now simulate a situation in applied research in which the data might contain considerable bias. To do so, in the beginning, we only use X_4 to construct B-spline basis functions. Thus, relatively severe bias exists in the specification. Then, we gradually add more and more basis functions involving all of the other relevant regressors (X_3, X_2, X_1) to alleviate bias. This creates a total of 10 cases with k growing from 5 to 121. The bias and RMSE of $\tilde{\theta}_{BP}$ after 10,000 experiments are collected Figure 6.3.

Choice of basis functions Instead of using B-splines, we also construct basis functions with orthogonal polynomials. As the dimensional restrictions on polynomials are stricter, we only consider nine possible scenarios where k grows from 5 to 70 and $\frac{k}{n}$ increases from

0.025 to 0.35. The results of $\tilde{\theta}_{BP}$ are reported in Figure 6.4 and Table 5.

Choice of λ_n Finally, I check the sensitivity of $\tilde{\theta}_{BP}$ to the choice of λ_n. The setup is the same with baseline results using B-splines, but λ_n ranges from 0 to 0.005. Results after 10,000 simulations are collected in Figure 6.5.

6.2 Comparison to Armstrong and Kolesár (2021)

In this section, we compare the performance of our proposed estimator with that of Armstrong and Kolesár (2021) in the context of estimating the average treatment effect on the treated, a special case of Example 3.6. Let U be defined in (6.1) and $\{e_1, e_0\}$ be mutually independent standard normal random variables that are also independent of U. The two potential outcome variables are generated as follows:

$$\begin{aligned} Y(1) &= 210 + 20U_1 - 10U_2 + 5U_3 + 2U_4 + e_1, \\ Y(0) &= 100 + 10U_1 - 5U_2 + 2.5U_3 + U_4 - e_0. \end{aligned}$$

The true propensity score follows the specification in (6.2). Again, we do not observe U directly but only its transformed version X as defined in (6.3). A random sample of size $n = 200$ is drawn from $(Y, T, X')'$. We are interested in estimating $\theta_0 = \mu_1 - \mu_0$, where $\mu_1 = \mathbb{E}[Y_i(1)|T_i = 1]$ and $\mu_0 = \mathbb{E}[Y_i(0)|T_i = 1]$. In this example, we may calculate the true value of θ_0 as 104.7809.

To apply our method, let $p(x)$ be a vector of basis functions. Then, our estimator for the average treatment effect on the treated is:

$$\tilde{\theta}_{BP} = \tilde{\mu}_{1,BP} - \tilde{\mu}_{0,BP},$$

where $\tilde{\mu}_{1,BP} = \mathbb{E}_n[\tilde{\alpha}_1(X)TY]$ is our proposed estimator for μ_1, $\tilde{\mu}_{0,BP} = \mathbb{E}_n[\tilde{\alpha}_0(X)(1-T)Y]$ is our proposed estimator for μ_0, and

$$\begin{aligned} \tilde{\alpha}_1(x) &= p'(x)\tilde{a}_1,\ \tilde{a}_1 = (\hat{G}_T\hat{G}_T + \lambda_{1,n}\hat{G}_T)^{-}\hat{P}_T, \\ \tilde{\alpha}_0(x) &= p'(x)\tilde{a}_0,\ \tilde{a}_0 = (\hat{G}_{1-T}\hat{G}_{1-T} + \lambda_{0,n}\hat{G}_{1-T})^{-}\hat{P}_T, \end{aligned}$$

where $\hat{P}_T := \frac{1}{n_1}\sum_{i=1}^n p(X_i)T_i, n_1 := \sum_{i=1}^n T_i,\ \hat{G}_T := \frac{1}{n}\sum_{i=1}^n p(X_i)p(X_i)'T_i,\ \hat{G}_{1-T} := \frac{1}{n}\sum_{i=1}^n p(X_i)p(X_i)'(1-T_i)$, and $\lambda_{1,n}, \lambda_{0,n}$ are penalty coefficients. For simplicity, I set $\lambda_{1,n} = \lambda_{0,n} = 0.002$ in all simulations. The results with B-splines are reported in the left half of Table 6.

To apply Armstrong and Kolesár (2021), we need to pick a norm $\|\cdot\|_{\mathcal{X}}$ on X and the Lipschitz smoothness constant C. Armstrong and Kolesár (2021) considered the weighted

l_p norm of the form

$$\|X\|_{A,p} = \left(\sum_{j=1}^{\mathsf{dim}(X)} |A_{jj} X_j|^p \right)^{\frac{1}{p}},$$

where A is a diagonal matrix. As argued in Armstrong and Kolesár (2021), the selection of A_{jj} should incorporate the relative importance of different pretreatment variables in explaining the outcome variable. Since here such prior knowledge is unclear, we first try $p = 1$ and fix $A_{jj} = 1$ for all j for simplicity. We also consider an alternative norm that avoids such prior knowledge, where $p = 2$ and $A_{jj} = \frac{1}{\mathsf{std}(X_j)}$, as suggested by Abadie and Imbens (2011). For each of the two norms specified above, I set $C \in \{1, 2, 4, 8\}$. The simulation results are reported in the right half of Table 6, which also includes the performance of the nearest-neighbor matching estimator, which is optimal when C is large enough.

As we can see from Table 6, our proposed method still works very well and is very robust to the number of basis functions used. The performance of the method in Armstrong and Kolesár (2021) seems to be more sensitive to the choice of the norm and the smoothness constant C: if we select $A_{jj} = 1$ with $p = 1$, their performance is considerably dominated by our estimator. We see a much more competitive performance from the estimator of Armstrong and Kolesár (2021) once we switch to the weighted l_2 norm suggested by Abadie and Imbens (2011) and choose a larger C. Therefore, compared to Armstrong and Kolesár (2021), our estimator may be more convenient for practitioners to implement. Our estimator only has one tuning parameter (λ_n), and our theoretical analysis implies that for most scenarios, it suffices to select a small fixed number. On the other hand, for many applied problems, researchers may lack a clear prior on what is a suitable norm and smoothness bound to use for the estimator in Armstrong and Kolesár (2021).

Table 2: Bias and RMSE using B-splines, 10,000 simulations, $\lambda_n = 0.002$, mild bias

Model	k	$\frac{k}{n}$	NR		$\tilde{\theta}_{BP}$		SR	
			Bias	RMSE	Bias	RMSE	Bias	RMSE
(1)	5	0.025	-0.6404	3.3775	-3.8049	4.8548	-12.0300	12.4907
(2)	9	0.045	-3.0948	4.7457	-4.5542	5.3889	-9.8147	10.3187
(3)	13	0.065	-2.0888	4.2762	-4.9485	5.8034	-11.3977	11.8625
(4)	17	0.085	-2.4118	11.1004	-5.1964	6.0223	-12.0436	12.5116
(5)	21	0.105	-1.7891	17.7536	-5.2546	6.0873	-10.9903	11.4777
(6)	33	0.165	-4.8446	14.8438	-5.2335	6.0627	-11.7762	12.2495
(7)	51	0.255	-3.5531	16.5214	-5.1723	5.9932	-11.2596	11.7459
(8)	61	0.305	-0.4233	14.8353	-5.1933	6.0162	-11.3621	11.8633
(9)	85	0.425	-0.6467	15.3074	-5.1013	5.9352	-11.3714	11.8766
(10)	109	0.545	-0.3799	14.4902	-5.0826	5.9193	-11.3699	11.8769
(11)	121	0.605	-0.4192	14.7215	-5.0849	5.9207	-11.3696	11.8768
(12)	Naive average		-10.0458	10.6347				
(13)	α_0 known		0.1584	23.6996				

Figure 6.1: Bias and RMSE using B-splines, 10,000 simulations, $\lambda_n = 0.002$, mild bias

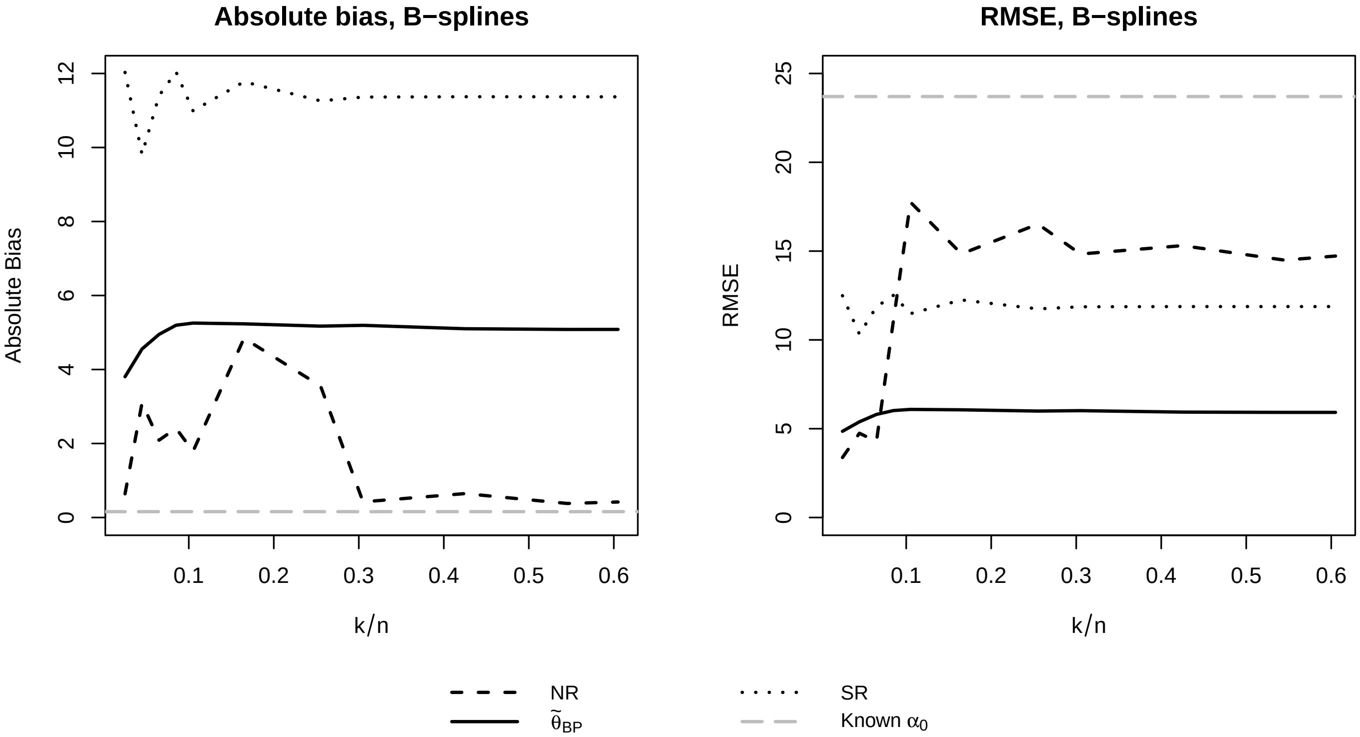

Table 3: Coverage probability using B-splines, 10,000 simulations, $\lambda_n = 0.002$, mild bias

Model	k	$\frac{k}{n}$	NR		$\tilde{\theta}_{BP}$		SR	
			5%	1%	5%	1%	5%	1%
(1)	5	0.025	0.9348	0.9833	0.8946	0.9753	0.3695	0.7945
(2)	9	0.045	0.8053	0.9214	0.7262	0.8766	0.3904	0.7309
(3)	13	0.065	0.8613	0.9462	0.7666	0.9125	0.3313	0.7094
(4)	17	0.085	0.8959	0.9708	0.8090	0.9257	0.4267	0.7491
(5)	21	0.105	0.9236	0.9795	0.8408	0.9382	0.5631	0.8249
(6)	33	0.165	0.9268	0.9822	0.8271	0.9431	0.4721	0.7927
(7)	51	0.255	0.9008	0.9755	0.8455	0.9365	0.5887	0.8450
(8)	61	0.305	0.9317	0.9851	0.8854	0.9561	0.6234	0.8755
(9)	85	0.425	0.9113	0.9791	0.9266	0.9674	0.7394	0.9211
(10)	109	0.545	0.8831	0.9651	0.9450	0.9755	0.7926	0.9402
(11)	121	0.605	0.8805	0.9630	0.9460	0.9740	0.8013	0.9404
(13)	α_0 known		0.9342	0.9772				

Figure 6.2: Bias and RMSE using B-splines and debiased estimators, mild bias, 10,000 simulations

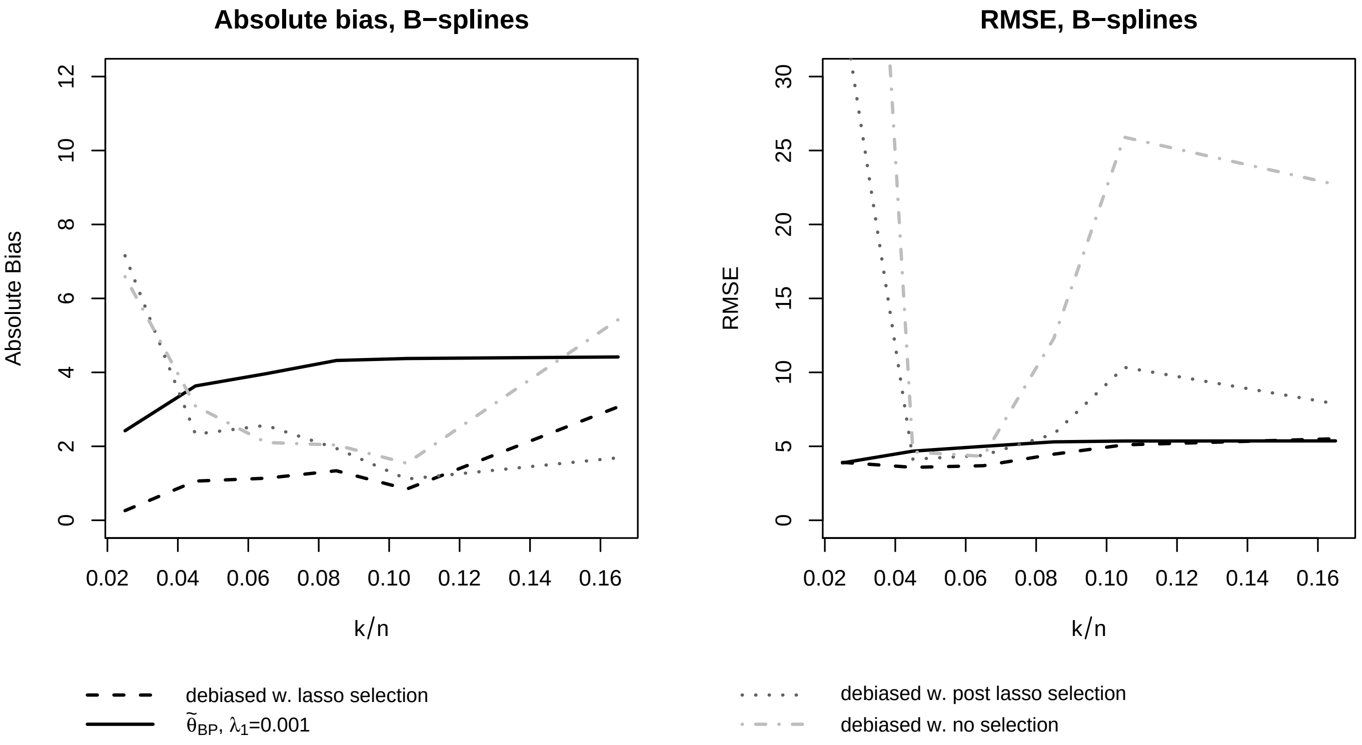

Table 4: RMSE and computational cost of our estimator and other doubly cross-fitted debiased estimators, 1,000 simulations

Panel A: $\tilde{\theta}_{BP}$, $\lambda_n = 0.002$, no cross-fitting

$\frac{k}{n}$	RMSE	system time
0.025	4.9237	2.67 sec
0.045	5.4901	2.65 sec
0.065	5.8963	3.58 sec
0.085	6.0958	4.16 sec
0.105	6.1550	4.19 sec

Panel B: doubly cross-fitted debiased estimators involving propensity score and trimming

$\frac{k}{n}$	RMSE	system time	RMSE	system time	RMSE	system time
	no selection		lasso selection		post lasso selection	
	2-fold		2-fold		2-fold	
0.025	13.2045	4.49 min	9.9478	50.2 sec	45.2243	51.44 sec
0.045	14.2425	1.91 hour	5.2478	1.35 min	32.7772	38.05 sec
0.065	16.2825	38.48 min	5.0133	2.49 min	23.2493	1.02 min
0.085	29.5419	24.39 min	8.7302	32.71 min	35.8430	1.80 min
0.105	110.0712	1.37 hour	6.6776	54.25 min	41.5275	48.94 min
$\frac{k}{n}$	no selection		lasso selection		post lasso selection	
	4-fold		4-fold		4-fold	
0.025	10.6002	50.78 min	10.1280	1.41 min	41.4212	56.29 sec
0.045	11.7668	19.06 min	5.4883	2.48 min	31.1395	1.20 min
0.065	12.3803	18.41 min	5.1988	4.58 min	15.0702	1.97 min
0.085	15.6855	41.67 min	6.1865	8.84 min	32.2416	3.25 min
0.105	31.1820	1.14 hour	5.1956	11.27 min	27.0850	1.58 hour

Panel C: doubly cross-fitted debiased estimators involving $\tilde{\alpha}$

$\frac{k}{n}$	RMSE	system time	RMSE	system time	RMSE	system time
	no selection		lasso selection		post lasso selection	
	2-fold		2-fold		2-fold	
0.025	4.2077	7.95 sec	4.1646	37.63 sec	4.3908	25.69 sec
0.045	6.0349	7.94 sec	4.5939	1.18 min	4.6477	29.71 sec
0.065	9.0701	8.80 sec	5.5366	2.22 min	6.3778	51.80 sec
0.085	39.991	10.57 sec	7.8443	21.73 min	9.9836	1.64 min
0.105	912.32	10.85 sec	6.8910	8.88 min	31.3642	14.35 min
$\frac{k}{n}$	no selection		lasso selection		post lasso selection	
	4-fold		4-fold		4-fold	
0.025	3.5657	12.33 sec	13.7139	1.16 min	14.495	38.13 sec
0.045	4.9074	13.33 sec	13.6422	2.11 min	14.649	54.18 sec
0.065	4.8337	13.87 sec	11.3418	4.05 min	12.5208	1.65 min
0.085	15.2616	14.57 sec	13.3563	8.47 min	14.6055	2.88 min
0.105	37.146	16.62 sec	9.3996	9.80 min	11.5159	9.73 min

Figure 6.3: Bias and RMSE using B-splines, 10,000 simulations, $\lambda_n = 0.002$, considerable bias

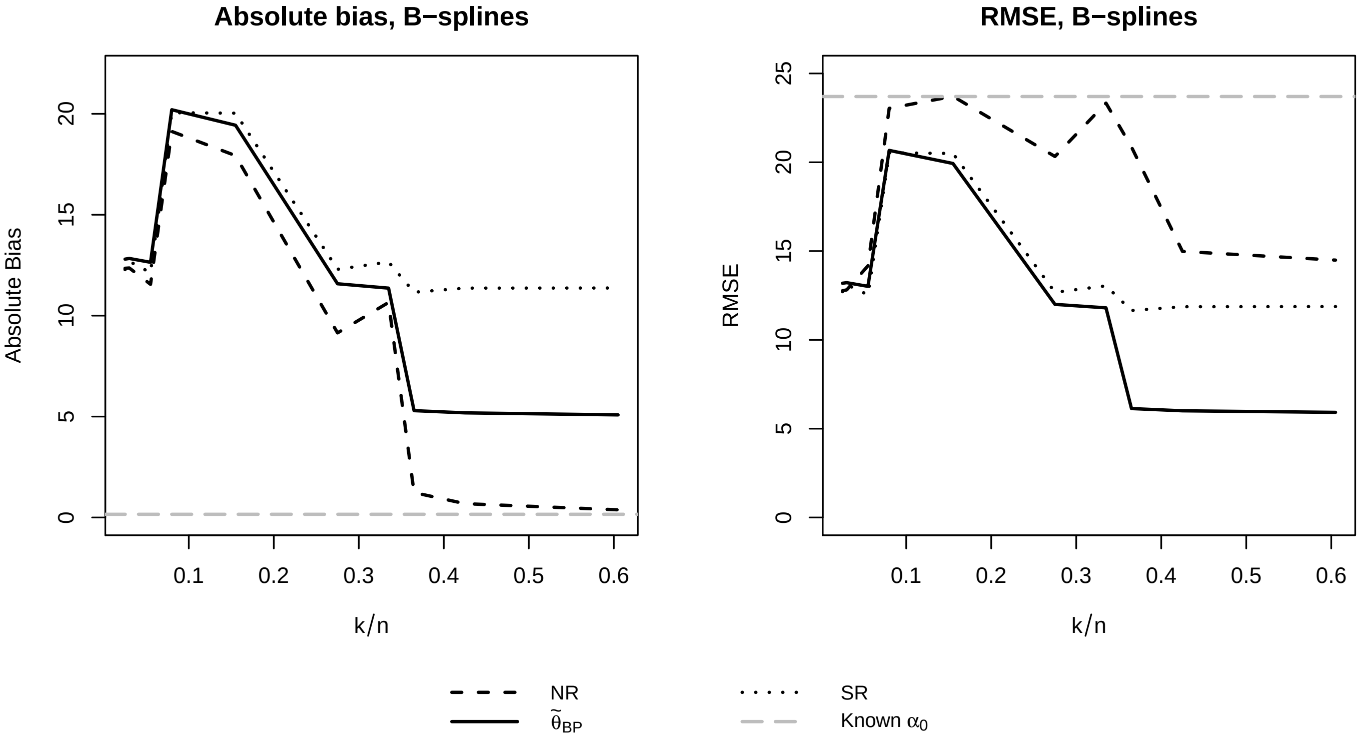

Figure 6.4: Bias and RMSE using orthogonal polynomials, 10,000 simulations, $\lambda_n = 0.001$, mild bias

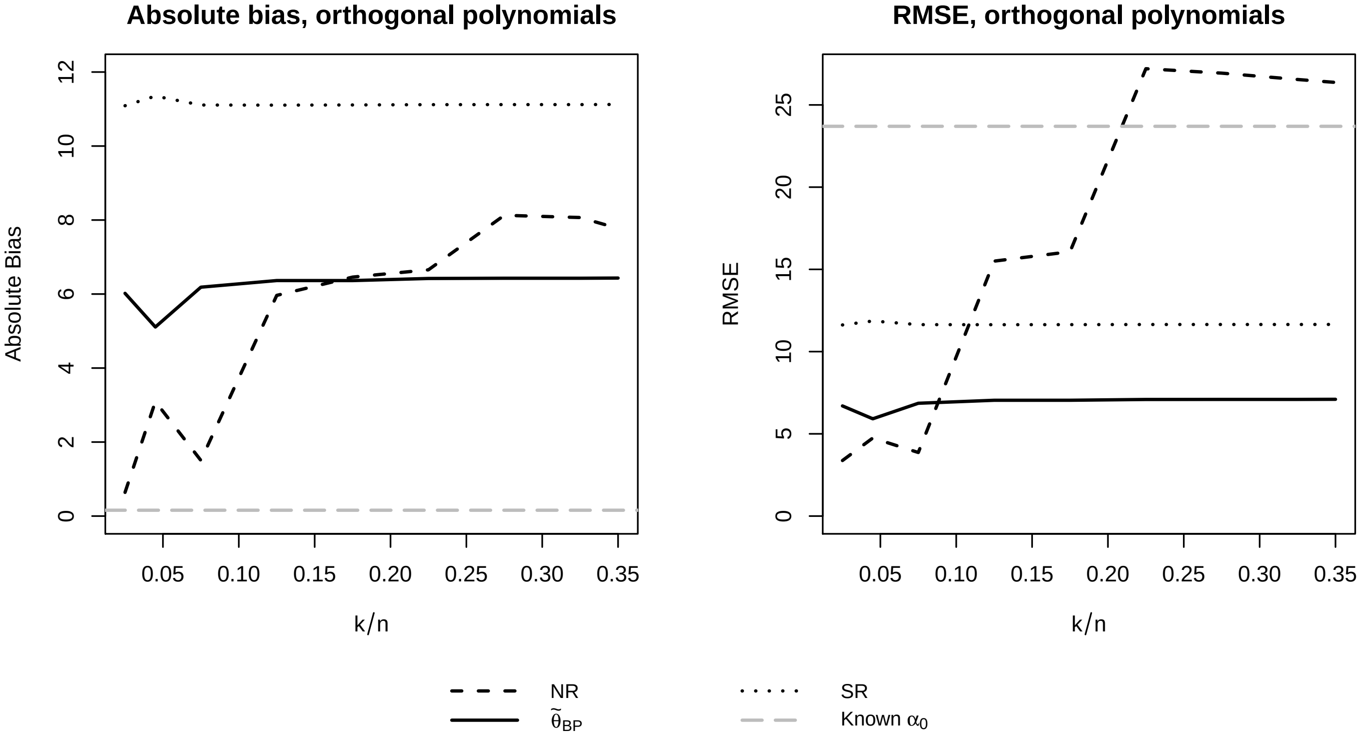

Figure 6.5: Sensitivity of $\tilde{\theta}_{BP}$ to λ_n using B-splines, mild bias

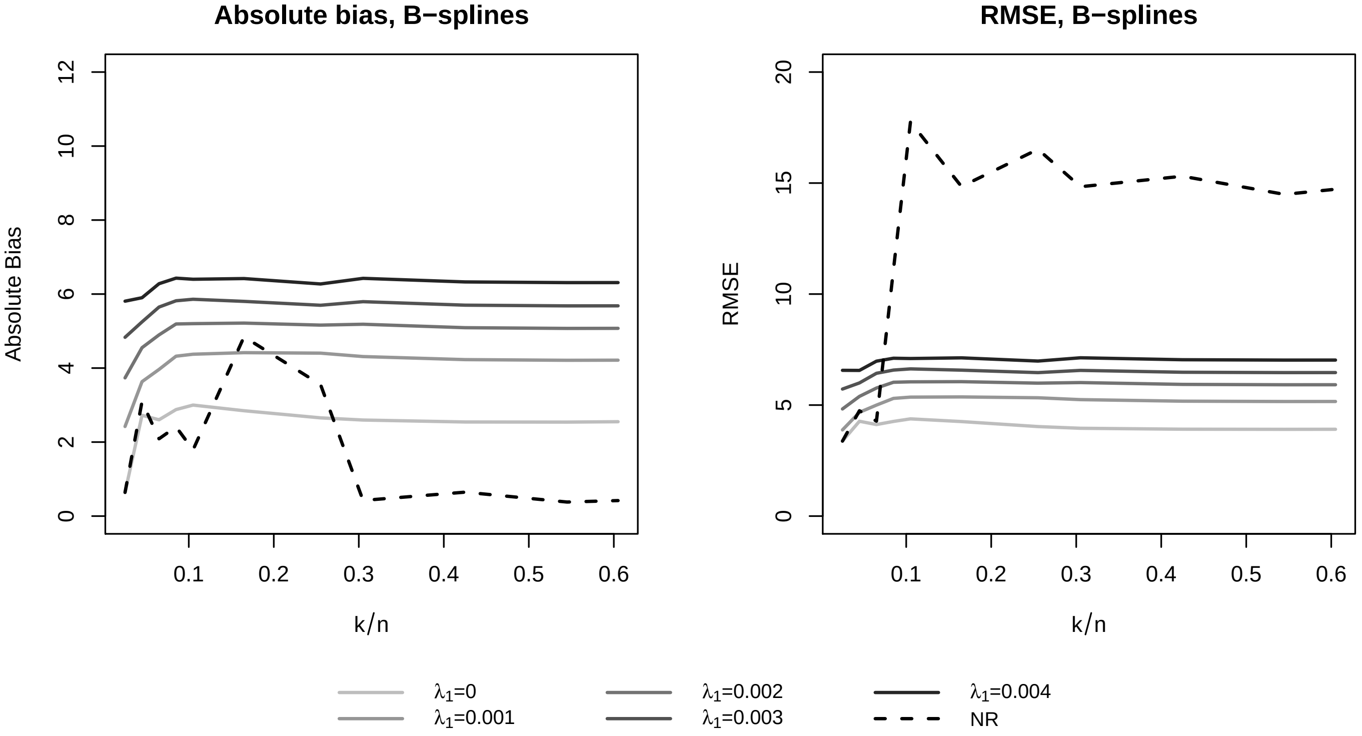

Table 5: Coverage probability using orthogonal polynomials, 10,000 simulations, $\lambda_n = 0.001$, mild bias

Model	k	$\frac{k}{n}$	NR		$\tilde{\theta}_{BP}$		SR	
			5%	1%	5%	1%	5%	1%
(1)	5	0.025	0.9348	0.9833	0.8204	0.9608	0.4500	0.8310
(2)	9	0.045	0.8053	0.9214	0.6998	0.8679	0.2881	0.6472
(3)	15	0.075	0.8940	0.9664	0.7557	0.9269	0.4021	0.7750
(4)	25	0.125	0.9351	0.9901	0.9211	0.9813	0.7308	0.9307
(5)	35	0.175	0.9296	0.9906	0.9295	0.9808	0.7601	0.9341
(6)	45	0.225	0.9482	0.9936	0.9576	0.9867	0.8536	0.9608
(7)	55	0.275	0.9572	0.9946	0.9560	0.9869	0.8570	0.9605
(8)	65	0.325	0.9575	0.9964	0.9614	0.9886	0.8626	0.9648
(9)	70	0.35	0.9559	0.9955	0.9586	0.9871	0.8622	0.9616
(11)	α_0 known		0.9342	0.9772				

Table 6: Bias and RMSE in estimating the average treatment effect on the treated, 1,000 simulations

Our estimator (B splines, $\lambda_n = 0.002$)				Armstrong and Kolesár (2021)			
k	$\frac{k}{n}$	$\|$bias$\|$	RMSE	$\|\cdot\|_{A,p}$	C	$\|$bias$\|$	RMSE
5	0.025	4.6602	5.0058	$A_{jj} = 1$, $p = 1$	1	8.1998	8.5390
9	0.045	2.6975	3.0706		2	8.0529	8.4048
13	0.065	2.4269	2.7920		4	8.0090	8.3654
17	0.085	2.5393	2.9052		8	7.9951	8.3525
21	0.105	2.3435	2.7238		16	7.9913	8.3493
33	0.165	2.3533	2.7348	$A_{jj} = \frac{1}{\mathsf{std}(X_j)}$, $p = 2$	1	5.8555	6.1749
51	0.255	2.2447	2.6229		2	4.3170	4.6725
61	0.305	2.8367	3.2482		4	3.6031	3.9782
85	0.425	2.1222	2.5109		8	3.3579	3.7355
109	0.545	2.1061	2.4945		16	3.2876	3.6670
121	0.605	2.1011	2.4887		nearest neighbor	3.2658	3.6455

Appendix

Notations

For a vector $a = (a_1, a_2 \cdots, a_k)' \in \mathbb{R}^k$, $m(w, a) = (m(w, a_1), m(w, a_2) \cdots, m(w, a_k))'$ 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 $\|a\|_\infty := \max_{1\leq j\leq k} |a_j|$ denote the l_2, l_1 and sup norms of vector a, respectively. For a function $f : \mathcal{W} \mapsto \mathbb{R}$, let $\|f\|_{\mathbb{P},q} := \left[\int |f(w)|^q \, d\mathbb{P}(w)\right]^{1/q}, 1 \leq q \leq \infty$ denote its $L^q(\mathbb{P})$ norm. In particular, $\|f\|_{\mathbb{P},\infty} := \sup_{w\in\mathcal{W}} |f(w)|$. 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, $\|A\| = \max_i |\lambda_i(A)|$. 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 $a_n \leq cb_n$ for some constant c that does not depend on n. Bold $\mathbf{0}$ denotes a k dimensional vector of 0s. Also, 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 projection coefficient and u_{γ_0} is the projection error. Analogously, we may write $\alpha_0 = \mathcal{L}_n\alpha_0 + u_{\alpha_0}$, where $\mathcal{L}_n\alpha_0 = a_l' p$. To simplify notation, let $u_{\gamma_0 i} := \gamma_{0i} - \beta_l' p_i$, $u_{\alpha_0 i} := \alpha_{0i} - a_l' p_i$, where $p_i := p(W_i)$, $\gamma_{0i} := \gamma_0(W_i)$, $\alpha_{0i} := \alpha_0(W_i)$, $i = 1 \ldots n$.

A Proofs of main results

Proof of Proposition 4.1

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

$$\begin{aligned}\text{(I)} &\leq \sup_{\|\beta\|\leq 1} \|\beta\|^2 \left\|\mathbb{E}_n[\alpha(W)p(W) - m(W,p)]\right\|^2 \\ &\leq \left\|\mathbb{E}_n[\alpha(W)p(W) - m(W,p)]\right\|^2 = \text{(II)}.\end{aligned} \tag{A.1}$$

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

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

where the last inequality follows since $\left\|\mathcal{E}_{\alpha,n}\right\|^2$ is one of the eigenvalues of matrix $\mathcal{E}_{\alpha,n}\mathcal{E}_{\alpha,n}'$. To see this, let $\mathcal{E}_{\alpha,n}\mathcal{E}_{\alpha,n}'\mathbf{v} = \lambda\mathbf{v}$ for some $\lambda \geq 0$ and $\mathbf{v} \in \mathbb{R}^k$. Premultiplying both sides by $\mathcal{E}_{\alpha,n}$ yields $\left(\left\|\mathcal{E}_{\alpha,n}\right\|^2 - \lambda\right)\mathcal{E}_{\alpha,n}'\mathbf{v} = 0$. Therefore, $\left\|\mathcal{E}_{\alpha,n}\right\|^2$ must be one of the eigenvalues of $\mathcal{E}_{\alpha,n}\mathcal{E}_{\alpha,n}'$. Combining (A.1) and (A.2) yields the conclusion.

Proof of Theorem 5.1

Proof of statement (i) Note $\tilde{\theta}_{BP}$ may be decomposed as follows:

$$\sqrt{n}\mathbb{E}_n[\tilde{\theta}_{BP} - \theta_0] = \sqrt{n}\mathbb{E}_n\phi + R_{n,1}(\tilde{\theta}_{BP}) + R_{n,2}(\tilde{\theta}_{BP}),$$

where

$$R_{n,1}(\tilde{\theta}_{BP}) = \sqrt{n}\mathbb{E}_n\left[\tilde{\alpha}(W)\gamma_0(W) - m(W,\gamma_0)\right],\ R_{n,2}(\tilde{\theta}_{BP}) = \sqrt{n}\mathbb{E}_n\left[\left(\tilde{\alpha}(W) - \alpha_0(W)\right)e\right].$$

Furthermore, $R_{n,1}(\tilde{\theta}_{BP}) = T_1 + T_2$, where

$$T_1 = \sqrt{n}\mathbb{E}_n[\tilde{\alpha}(W)\mathcal{L}_n\gamma_0(W) - m(W,\mathcal{L}_n\gamma_0)],\ T_2 = \mathbb{E}_n[\tilde{\alpha}(W)u_{\gamma_0} - m(W,u_{\gamma_0})],$$

and $T_2 = T_{21} + T_{22}$, where

$$T_{21} = \sqrt{n}\mathbb{E}_n[\left(\tilde{\alpha}(W) - \alpha_0(W)\right)u_{\gamma_0}],\ T_{22} = \sqrt{n}\mathbb{E}_n\left[\alpha_0(W)u_{\gamma_0} - m(W,u_{\gamma_0})\right].$$

Under Assumptions O, S, P and L, Lemma B.1 shows $R_{n,2}(\tilde{\theta}_{BP}) = o_p(1)$. Lemma B.2 shows $T_1 = o_p(1)$. Lemma B.3 establishes $T_{22} = o_p(1)$. Under Assumptions O, S, P, L and if $\frac{\xi_k^2}{\sqrt{n}}\mathbf{r}_{\gamma_0} = o(1)$, Lemma B.4 shows $T_{21} = O_p\left(\frac{\xi_k^2}{\sqrt{n}}\mathbf{r}_{\gamma_0}\right) = o_p(1)$. Statement (i) follows.

Proof of statement (ii) Note we can write the following:

$$\sqrt{n}\{\mathbb{E}[\phi_i^2]\}^{-\frac{1}{2}}(\tilde{\theta}_{BP} - \theta_0) = \sqrt{n}\{\mathbb{E}[\phi_i^2]\}^{-\frac{1}{2}}\mathbb{E}_n\phi + \{\mathbb{E}[\phi_i^2]\}^{-\frac{1}{2}}R_{n,1}(\tilde{\theta}_{BP}) + \{\mathbb{E}[\phi_i^2]\}^{-\frac{1}{2}}R_{n,2}(\tilde{\theta}_{BP}).$$

Since $\mathbb{E}\phi_i^2$ is uniformly bounded away from zero for all i and n, $\{\mathbb{E}[\phi_i^2]\}^{-\frac{1}{2}} = O(1)$. Thus, it still holds that $\{\mathbb{E}[\phi_i^2]\}^{-\frac{1}{2}}R_{n,1}(\tilde{\theta}_{BP}) + [\mathbb{E}[\phi_i^2]]^{-\frac{1}{2}}\,R_{n,2}(\tilde{\theta}_{BP}) \xrightarrow{p} 0$ by statement (i). To show $\sqrt{n}\{\mathbb{E}[\phi_i^2]\}^{-\frac{1}{2}}\mathbb{E}_n\phi \xrightarrow{d} N(0,1)$, we can apply central limit theorem for i.i.d. data (e.g., Lindeberg–Lévy central limit theorem). Final conclusion follows from the continuous mapping theorem.

Proof of Theorem 5.2

Since $\sqrt{n}\hat{\Omega}^{-\frac{1}{2}}(\tilde{\theta}_{BP} - \theta_0) = \left(\frac{\mathbb{E}[\phi_i^2]}{\hat{\Omega}}\right)^{\frac{1}{2}}\sqrt{n}\left(\mathbb{E}[\phi_i^2]\right)^{-\frac{1}{2}}(\tilde{\theta}_{BP} - \theta_0)$ and $\sqrt{n}\left(\mathbb{E}[\phi_i^2]\right)^{-\frac{1}{2}}(\tilde{\theta}_{BP} - \theta_0) \xrightarrow{d} N(0,1)$ by Theorem 5.1, continuous mapping theorem implies that it suffices to show $\hat{\Omega} \xrightarrow{p} \mathbb{E}[\phi_i^2]$. To this end, note Theorem 5.1 implies $\tilde{\theta}_{BP} \xrightarrow{p} \theta_0$. Thus, we can decompose:

$$\mathbb{E}_n\left[m(W,\hat{\gamma}_{LS}) + \tilde{\alpha}(W)(Y - \hat{\gamma}_{LS}(W))\right]^2 - \mathbb{E}\left[m(W_i,\gamma_0) + \alpha_0(W_i)(Y_i - \gamma_0(W_i))\right]^2 \leq J_1 + J_2,$$

where

$$\begin{aligned}
J_1 &:= \mathbb{E}_n\left[m(W,\hat{\gamma}_{LS}) + \tilde{\alpha}(W)(Y - \hat{\gamma}_{LS}(W))\right]^2 - \mathbb{E}_n\left[m(W,\gamma_0) + \alpha_0(W)(Y - \gamma_0(W))\right]^2, \\
J_2 &:= \mathbb{E}_n\left[m(W,\gamma_0) + \alpha_0(W)(Y - \gamma_0(W))\right]^2 - \mathbb{E}\left[m(W_i,\gamma_0) + \alpha_0(W_i)(Y_i - \gamma_0(W_i))\right]^2.
\end{aligned}$$

By continuous mapping theorem, it suffices to show that both J_1 and J_2 are $o_p(1)$. Note that $J_2 = o_p(1)$ by Assumption O and weak law of large numbers. Then, it suffices to bound J_1. Let $\hat{\varphi}_i := m(W_i,\hat{\gamma}_{LS}) + \tilde{\alpha}(W_i)(Y_i - \hat{\gamma}_{LS}(W_i))$ and $\varphi_i := m(W_i,\gamma_0) + \alpha_0(W_i)(Y_i - \gamma_0(W_i))$. Therefore, we can write $J_1 = J_{11} + J_{12}$, where $J_{11} := 2\mathbb{E}_n[\varphi(\hat{\varphi} - \varphi)]$, $J_{12} := \mathbb{E}_n[\hat{\varphi} - \varphi]^2$. By Cauchy-Schwarz inequality, $|J_{11}| \leq 2[\mathbb{E}_n\varphi^2]^{1/2}\{\mathbb{E}_n[\hat{\varphi} - \varphi]^2\}^{1/2}$. Since we have shown that $\mathbb{E}_n[\varphi^2] \xrightarrow{p} \mathbb{E}[\varphi_i^2]$, $\mathbb{E}_n\varphi^2 = O_p(1)$. If we can further show $\mathbb{E}_n[\hat{\varphi} - \varphi]^2 = o_p(1)$, then $J_{11} = o_p(1)$ and $J_{12} = o_p(1)$. To this end, note $\mathbb{E}_n[\hat{\varphi} - \varphi]^2 \lesssim \Xi_1 + \Xi_2 + \Xi_3$, where

$$\begin{aligned}
&\Xi_1 := \mathbb{E}_n[m^2(W,\hat{\gamma}_{LS} - \gamma_0)], \quad \Xi_2 := \mathbb{E}_n[\tilde{\alpha}(W)(\hat{\gamma}_{LS}(W) - \gamma_0(W))]^2, \\
&\Xi_3 := \mathbb{E}_n[(\tilde{\alpha}(W) - \alpha_0(W))^2 e^2].
\end{aligned}$$

By Lemma B.6, all three terms above are $o_p(1)$, leading to the desired conclusion.

Proof of Theorem 5.3

Note we may write the following:

$$\begin{aligned}\sqrt{n}(\sigma_m^2+\sigma_{\tilde{\alpha}}^2)^{-1/2}\left(\tilde{\theta}_{BP}-\theta_0\right) &= \sqrt{n}(\sigma_m^2+\sigma_{\tilde{\alpha}}^2)^{-1/2}\mathbb{E}_n[\tilde{\alpha}(W)e+m(W,\gamma_0)-\theta_0],\\ &+(\sigma_m^2+\sigma_{\tilde{\alpha}}^2)^{-1/2}R_{n,1}(\tilde{\theta}_{BP}),\end{aligned}$$

where $R_{n,1}(\tilde{\theta}_{BP})=\sqrt{n}\mathbb{E}_n\left[\tilde{\alpha}(W)\gamma_0(W)-m(W,\gamma_0)\right]$. By Lemma C.2, $\left\{\mathbb{E}_n\left[\tilde{\alpha}^2(W)\right]\right\}^{-1}=O_p(1)$. By Lemma C.4, ${}^{\max_i|\tilde{\alpha}(W_i)|}/_{\sqrt{n}}=o_p(1)$. Thus, applying Lemma C.1 yields the following:

$$\sqrt{n}(\sigma_m^2+\sigma_{\tilde{\alpha},n}^2)^{-1/2}\mathbb{E}_n[\tilde{\alpha}(W)e+m(W,\gamma_0)-\theta_0]\xrightarrow{d}N(0,1),\ \text{as } n\to\infty.$$

Furthermore, by Lemma C.3, $R_{n,1}(\tilde{\theta}_{BP})=o_p(1)$. Also note that $(\sigma_m^2+\sigma_{\tilde{\alpha},n}^2)^{-1}\leq\frac{1}{\sigma_{\tilde{\alpha},n}^2}\lesssim\left\{\mathbb{E}_n\left[\tilde{\alpha}^2(W)\right]\right\}^{-1}=O_p(1)$. Thus, the remainder term $(\sigma_m^2+\sigma_{\tilde{\alpha},n}^2)^{-1/2}R_{n,1}(\tilde{\theta}_{BP})=o_p(1)$ as well. The conclusion follows.

B $\frac{k}{n}\to 0$ asymptotics

Additional lemmas supporting Theorem 5.1

Lemma B.1. *If Assumptions O, S, P and L hold true, then*

$$R_{n,2}(\tilde{\theta}_{BP})=\sqrt{n}\mathbb{E}_n\left[\left(\tilde{\alpha}(W)-\alpha_0(W)\right)e\right]=o_p(1).$$

Proof. We may write $R_{n,2}(\tilde{\theta}_{BP})=\sqrt{n}\mathbb{E}_n\left[\left(\tilde{a}-a_l\right)p(W)e\right]+\sqrt{n}\mathbb{E}_n\left[u_{\alpha_0}e\right]$. To bound the first term, apply Lemma D.8 with $\mathcal{A}_j(\mathscr{W}_n)=(\tilde{a}-a_l)'p(W_j), j=1\ldots n$, where $\mathscr{W}_n=\{W_i\}_{i=1}^n$. Note that $\frac{1}{n}\sum_{j=1}^n[(\tilde{a}-a_l)'p(W_j)]^2=(\tilde{a}-a_l)'\hat{G}(\tilde{a}-a_l)\leq\|\tilde{a}-a_l\|^2\left\|\hat{G}\right\|=o_p(1)$ by Lemmas D.2 and B.5. Thus, Lemma D.8 implies $\sqrt{n}\mathbb{E}_n\left[\left(\tilde{a}-a_l\right)p(W)e\right]=o_p(1)$. For the second term, note that $\sup_{w\in\mathcal{W}}\mathbb{E}[e_i^2|W_i=w]\lesssim 1$ by Assumption O(2), and that $\mathbb{E}[u_{\alpha_0 i}^2]\leq\mathbf{r}_{\alpha_0}^2$ by Lemma D.1(ii). Markov inequality and Assumption L yield $\sqrt{n}\mathbb{E}_n\left[u_{\alpha_0}e\right]=O_p(\mathbf{r}_{\alpha_0})=o_p(1)$. □

Lemma B.2. *If Assumptions O, S, P and L hold true, then* $T_1=\sqrt{n}\mathbb{E}_n[\tilde{\alpha}(W)\mathcal{L}_n\gamma_0(W)-m(W,\mathcal{L}_n\gamma_0)]=o_p(1)$.

Proof. Plugging $\mathcal{L}_n\gamma_0=\beta_l'p$ and $\tilde{\alpha}=p'(\hat{G}\hat{G}+\lambda_n\hat{G})^{-}\hat{G}\hat{P}$ in T_1 yields the following:

$$\begin{aligned}|T_1|&=\sqrt{n}\left|\beta_l'\left(\hat{G}(\hat{G}\hat{G}+\lambda_n\hat{G})^{-}\hat{G}-I\right)\hat{P}\right|.\\ &\leq\sqrt{n}\left\|\beta_l\right\|\left\|\hat{G}(\hat{G}\hat{G}+\lambda_n\hat{G})^{-}\hat{G}-I\right\|\left\|\hat{P}\right\|,\end{aligned}$$

where the second line follows from Cauchy-Schwarz inequality. By Lemma D.1(iv), $\|\beta_l\| = O(1)$. By Lemma D.3 and Assumption L, $\left\|\hat{P}\right\| = O_p(1)$ and $\left\|\hat{G}(\hat{G}\hat{G}+\lambda_n\hat{G})^{-}\hat{G}-I\right\| = o_p(\frac{1}{\sqrt{n}})$, as $\lambda_n = o(\frac{1}{\sqrt{n}})$ (Assumption P) and $\frac{1}{\lambda_{\min}(\hat{G})} = O_p(1)$ (Lemma B.5). Thus, $T_1 = \sqrt{n}o_p\left(\frac{1}{\sqrt{n}}\right)O_p(1) = o_p\left(1\right)$. □

Lemma B.3. *If Assumptions O and S hold true and* $\left(\ell_k \wedge \|\alpha_0\|_{\mathbb{P},\infty}\right)\mathbf{r}_{\gamma 0} = o(1)$*, then* $T_{22} = \sqrt{n}\mathbb{E}_n\left[\alpha_0(W)u_{\gamma_0} - m(W,u_{\gamma_0})\right] = O_p\left[\left(\ell_k \wedge \|\alpha_0\|_{\mathbb{P},\infty}\right)\mathbf{r}_{\gamma 0}\right] = o_p(1)$*.*

Proof. Note that $\mathbb{E}[\alpha_{0i}u_{\gamma_0 i} - m(W_i,u_{\gamma_0})] = 0$. Thus, the following holds:

$$\mathbb{E}\left[\alpha_{0i}u_{\gamma_0 i} - m(W_i,u_{\gamma_0})\right]^2 \lesssim \mathbb{E}\left[\alpha_{0i}u_{\gamma_0 i}\right]^2 + \mathbb{E}[m^2(W_i,u_{\gamma_0})] \lesssim \mathbb{E}\left[\alpha_{0i}^2 u_{\gamma_0 i}^2\right],$$

where the first inequality follows from triangle inequality and the second inequality follows from Assumption O. Also, note 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 $\mathbb{E}\left[\alpha_{0i}^2\right] \lesssim 1$ and Lemma D.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 D.1(ii). Then, Markov inequality implies $T_{22} = O_p\left[\left(\ell_k \wedge \|\alpha_0\|_{\mathbb{P},\infty}\right)\mathbf{r}_{\gamma_0}\right] = o_p(1)$. □

Lemma B.4. *If Assumptions O, S, P and L hold true and* $\frac{\xi_k^2}{\sqrt{n}}\mathbf{r}_{\gamma_0} = o(1)$*, then* $T_{21} = \sqrt{n}\mathbb{E}_n[(\tilde{\alpha}(W) - \alpha_0(W))\,u_{\gamma_0}] = O_p\left(\frac{\xi_k^2}{\sqrt{n}}\mathbf{r}_{\gamma_0}\right) = o_p(1)$*.*

Proof. We may write $T_{21} = \sqrt{n}\mathbb{E}_n\left[(\tilde{a}-a_l)'\,p(W)u_{\gamma_0}\right] + \sqrt{n}\mathbb{E}_n[u_{\alpha_0}u_{\gamma_0}]$, where the first term is $o_p(1)$ by Lemma B.5. For the second term, note the following decomposition:

$$\sqrt{n}\mathbb{E}_n[u_{\alpha_0}u_{\gamma_0}] = \sqrt{n}\mathbb{E}_n[u_{\alpha_0}u_{\gamma_0} - \mathbb{E}[u_{\alpha_0 i}u_{\gamma_0 i}]] + \sqrt{n}\mathbb{E}[u_{\alpha_0 i}u_{\gamma_0 i}],$$

where the first term is $o_p(1)$ by Markov inequality and Assumption L, and the second term is $o(1)$ by Assumption L. The conclusion follows. □

Lemma B.5. *Suppose Assumptions O, S, P and L hold true and* $\frac{\xi_k^2}{\sqrt{n}}\mathbf{r}_{\gamma_0} \to 0$*. Then:*

(i) $\lambda_{\min}(\hat{G})$ *is bounded away from zero wpa1.*

(ii) $\|\tilde{a}-a_l\| = o_p(1)$, $\|\tilde{a}\| = O_p(1)$*, and* $\sqrt{n}\mathbb{E}_n\left[(\tilde{a}-a_l)'\,p(W)u_{\gamma_0}\right] = O_p(\frac{\xi_k^2}{\sqrt{n}}\mathbf{r}_{\gamma_0} + \sqrt{n}\mathbf{r}_{\alpha_0}\mathbf{r}_{\gamma_0}) = o_p(1)$*.*

Proof. Applying Tropp (2015, Theorem 5.1.1), we find that $\mathbb{P}\{\lambda_{\min}(\hat{G}) \le 0.5\lambda_{\min}(G)\} \le \exp\{\log k[1 - \frac{0.25\lambda_{\min}(G)}{2\xi_k^2\log k/n}]\} \to 0$ since $\frac{\xi_k^2\log k}{n} \to 0$ and $\lambda_{\min}(G)$ is bounded away from zero. Thus, wpa1 $\lambda_{\min}(\hat{G})$ is bounded away from zero. Statement (ii) follows from Lemmas D.4 and D.5. □

Additional lemmas supporting Theorem 5.2

Lemma B.6. *If the conditions of Theorem 5.2 hold true, then:*

(i) $\mathbb{E}_n[m^2(W,\hat{\gamma}_{LS}-\gamma_0)] = o_p(1)$;

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

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

Proof. Statement (i). Note that $\hat{\beta}_{LS} = \hat{G}^{-}\mathbb{E}_n[p(W)Y]$. By the linearity of $m(w,\cdot)$ and the triangle inequality,

$$\mathbb{E}_n[m^2(W,\hat{\gamma}_{LS}-\gamma_0)] \lesssim \mathbb{E}_n\left[m^2\left(W,(\hat{\beta}_{LS}-\beta_l)'p\right)\right] + \mathbb{E}_n\left[m^2(W,u_{\gamma_0})\right].$$

First, by weak law of large numbers, $\mathbb{E}_n[m^2(W,u_{\gamma_0})] \xrightarrow{p} \mathbb{E}[m^2(W_i,u_{\gamma_0 i})] \lesssim \mathbb{E}u^2_{\gamma_0 i} \lesssim \mathbf{r}^2_{\gamma_0} = o_p(1)$. Second, by the linearity of $m(w,\cdot)$,

$$\mathbb{E}_n\left[m^2\left(W,(\hat{\beta}_{LS}-\beta_l)'p\right)\right] \leq \left\|\hat{\beta}_{LS}-\beta_l\right\|^2 \sup_{\beta\in\mathbb{S}^{k-1}}\mathbb{E}_n m^2(W,\beta'p),$$

where $\mathbb{S}^{k-1} := \{a\in\mathbb{R}^k|\,\|a\|=1\}$. Since $\left\|\hat{\beta}_{LS}-\beta_l\right\| = o_p(1)$ by Belloni et al. (2015, Thereom 4.1), it remains to show that $\sup_{\beta\in\mathbb{S}^{k-1}}\mathbb{E}_n m^2(W,\beta'p) = O_p(1)$. Note the following holds:

$$\sup_{\beta\in\mathbb{S}^{k-1}}\mathbb{E}_n m^2(W,\beta'p) \leq \Xi_{11}+\Xi_{12},$$

where

$$\Xi_{11} := \sup_{\beta\in\mathbb{S}^{k-1}}\left[\mathbb{E}_n m^2(W,\beta'p) - \mathbb{E}m^2(W_i,\beta'p)\right], \quad \Xi_{12} := \sup_{\beta\in\mathbb{S}^{k-1}}\left[\mathbb{E}m^2(W_i,\beta'p)\right].$$

By condition (i) of Theorem 5.2 and S(1), $\Xi_{12} \lesssim \sup_{\beta\in\mathbb{S}^{k-1}}\left\{\mathbb{E}\left[\beta'p(W_i)\right]^2\right\} = \lambda_{\max}(G) \lesssim 1$. To show that $\Xi_{11} = o_p(1)$, we invoke Newey (1991, Corollary 2.2). *Compactness* is satisfied since $\beta\in\mathbb{S}^{k-1}$. *Pointwise convergence* follows from weak law of large numbers and $\Xi_{12}\lesssim 1$. It remains to verify *Assumption 3A*. Let $\mathbf{m}(\beta) := m(w,\beta'p)$. Then for each $\beta_1,\beta_2\in\mathbb{S}^{k-1}$,

$$\left|\mathbb{E}_n\mathbf{m}^2(\beta_1) - \mathbb{E}_n\mathbf{m}^2(\beta_2)\right| \leq 2\left[\mathbb{E}_n\mathbf{m}^2(\beta_2)\right]^{1/2}\left[\mathbb{E}_n\mathbf{m}^2(\beta_1-\beta_2)\right]^{1/2} + \mathbb{E}_n\mathbf{m}^2(\beta_1-\beta_2).$$

By *Pointwise convergence*, it holds that $\mathbb{E}_n\mathbf{m}^2(\beta_2) \xrightarrow{p} \mathbb{E}\mathbf{m}^2(\beta_2) = O(1)$, and

$$\mathbb{E}_n\mathbf{m}^2(\beta_1-\beta_2) \xrightarrow{p} \mathbb{E}\mathbf{m}^2(\beta_1-\beta_2) \lesssim \mathbb{E}\left[(\beta_1-\beta_2)'p(W_i)\right]^2 \lesssim \lambda_{\max}(G)\left\|\beta_1-\beta_2\right\|^2,$$

where the second relation is by condition (i) of Theorem 5.2. Thus, $|\mathbb{E}_n \mathbf{m}^2(\beta_1) - \mathbb{E}_n \mathbf{m}^2(\beta_2)| \leq O_p(1) \|\beta_1 - \beta_2\|$ and *Assumption 3A* is satisfied. Applying Newey (1991, Corollary 2.2) yields $\Xi_{11} = o_p(1)$.

Statement (ii). We note that $\mathbb{E}_n \left[\tilde{\alpha}(W)(\hat{\gamma}_{LS}(W) - \gamma_0(W))\right]^2 \leq \|\hat{\gamma}_{LS} - \gamma_0\|^2_{\mathbb{P},\infty} \mathbb{E}_n[\tilde{\alpha}^2(W)] = o_p(1)O_p(1) = o_p(1)$, as $\mathbb{E}_n[\tilde{\alpha}^2(W)] = \tilde{a}'\hat{G}\tilde{a} \leq \|\tilde{a}\|^2 \left\|\hat{G}\right\| = O_p(1)$ by Lemmas D.2, B.5, and $\|\hat{\gamma}_{LS} - \gamma_0\|_{\mathbb{P},\infty} = o_p(1)$.

Statement (iii). We may write that $\mathbb{E}_n \left[(\tilde{\alpha}(W) - \alpha_0(W))^2 e^2\right] \lesssim \Xi_{31} + \Xi_{32}$, where

$$\Xi_{31} := \mathbb{E}_n \left[\left((\tilde{a} - a_l)'p(W)\right)^2 e^2\right], \quad \Xi_{32} := \mathbb{E}_n \left[u^2_{\gamma_0} e^2\right].$$

By weak law of large numbers and Assumption O, $\Xi_{32} \xrightarrow{p} \mathbb{E}\left[u^2_{\gamma_0 i} {e_i}^2\right] \lesssim \mathbb{E}[u^2_{\gamma_0 i}] = o_p(1)$. Next, we show $\Xi_{31} = o_p(1)$ as well. Note the following:

$$\Xi_{31} = (\tilde{a} - a_l)'\mathbb{E}_n \left[p(W)p(W)'e^2\right](\tilde{a} - a_l) \leq \|\tilde{a} - a_l\|^2 \left\|\mathbb{E}_n \left[p(W)p(W)'e^2\right]\right\|.$$

Since $\|\tilde{a} - a_l\| = o_p(1)$ by Lemma B.5, it suffices to show $\|\mathbb{E}_n \left[p(W)p(W)'e^2\right]\| = O_p(1)$. By the triangle inequality,

$$\left\|\mathbb{E}_n \left[p(W)p(W)'e^2\right]\right\| \leq \left\|\mathbb{E}_n \left[p(W)p(W)'e^2\right] - \mathbb{E}\left[p(W_i)p(W_i)'e_i^2\right]\right\| + \left\|\mathbb{E}\left[p(W_i)p(W_i)'e_i^2\right]\right\|,$$

where the first term is $o_p(1)$ by Chen and Christensen (2015, Lemma 3.1), and the second term is bounded as $\|\mathbb{E}\left[p(W_i)p(W_i)'e_i^2\right]\| \lesssim \sup_{a \in \mathbb{S}^{k-1}} \mathbb{E}\left[(a'p(W_i))^2\right] = \|G\| \lesssim 1$. □

C Many-regressor asymptotics

Lemma C.1 is a new result that may be invoked to establish the asymptotic distribution of $\tilde{\theta}_{BP}$ when $\frac{k}{n} \lesssim 1$.

Lemma C.1. *Suppose Assumption O and the following conditions hold:*

(i) $\left[\mathbb{E}_n \tilde{\alpha}^2(W)\right]^{-1} = O_p(1)$;

(ii) ${}^{\max_i |\tilde{\alpha}(W_i)|}/_{\sqrt{n}} = o_p(1)$;

(iii) $\inf_{w \in \mathcal{W}} \mathbb{E}\left[e_i^2 | W_i = w\right]$ *is bounded away from zero and* $\sup_{w \in \mathcal{W}} \mathbb{E}\left[|e_i|^3 | W_i = w\right] \lesssim 1$ *uniformly over all* i *and* n;

(iv) $\frac{\sigma^2_m}{\sigma^2_m + \sigma^2_{\tilde{\alpha}}}$ *converges in probability to some constant* $v \in (0, 1)$, *where* $\sigma^2_{\tilde{\alpha}} = \mathbb{E}_n \left[\tilde{\alpha}^2(W)\mathbb{E}[e^2|W]\right]$, $\sigma^2_m = \mathbb{E}[m^2(W_i, \gamma_0)] - \theta^2_0$.

Then, the following holds:

$$\sqrt{n}(\sigma^2_m + \sigma^2_{\tilde{\alpha}})^{-1/2} \mathbb{E}_n[\tilde{\alpha}(W)e + m(W, \gamma_0) - \theta_0] \xrightarrow{d} N(0, 1), \text{ as } n \to \infty.$$

Proof. Let $\mathbb{G}_n := \sqrt{n}(\sigma_m^2 + \sigma_{\tilde{\alpha}}^2)^{-1/2}\mathbb{E}_n[\tilde{\alpha}(W)e + m(W,\gamma_0) - \theta_0]$. Also let $\phi_{\mathbb{G}_n}(t)$ be the characteristic function of $\mathbb{G}_n$ and $\phi(t) := e^{-\frac{1}{2}t^2}$ be the characteristic function of a standard normal random variable. To prove the conclusion, it suffices to show that $|\phi_{\mathbb{G}_n}(t) - \phi(t)| \to 0$, as $n \to \infty$ for each t. Notice

$$\begin{aligned}
\phi_{\mathbb{G}_n}(t) &= \mathbb{E}\exp\{\mathbf{i}t\sqrt{n}(\sigma_m^2 + \sigma_{\tilde{\alpha}}^2)^{-1/2}\mathbb{E}_n[\tilde{\alpha}(W)e + m(W,\gamma_0) - \theta_0]\} \\
&= \mathbb{E}\exp\{\mathbf{i}t\sqrt{n}(\sigma_m^2 + \sigma_{\tilde{\alpha}}^2)^{-1/2}\mathbb{E}_n[\tilde{\alpha}(W)e]\}\exp\{\mathbf{i}t\sqrt{n}(\sigma_m^2 + \sigma_{\tilde{\alpha}}^2)^{-1/2}\mathbb{E}_n[m(W,\gamma_0) - \theta_0]\} \\
&= \mathbb{E}\exp\mathbb{D}_{n,1}\exp\mathbb{D}_{n,2},
\end{aligned}$$

where

$$\begin{aligned}
\mathbb{D}_{n,1} &= \{\mathbf{i}t\sqrt{n}(\sigma_m^2 + \sigma_{\tilde{\alpha}}^2)^{-1/2}\mathbb{E}_n[\tilde{\alpha}(W)e]\}, \\
\mathbb{D}_{n,2} &= \{\mathbf{i}t\sqrt{n}(\sigma_m^2 + \sigma_{\tilde{\alpha}}^2)^{-1/2}\mathbb{E}_n[m(W,\gamma_0) - \theta_0]\}.
\end{aligned}$$

Let Z_1 and Z_2 be i.i.d. standard normal also independent of $\mathscr{W}_n := \{W_i\}_{i=1}^n$. It follows that:

$$\begin{aligned}
&|\phi_{\mathbb{G}_n}(t) - \phi(t)| = |\mathbb{E}\exp\mathbb{D}_n^1\mathbb{D}_n^2 - e^{-\frac{1}{2}t^2}| \\
\leq &\left|\mathbb{E}\exp\mathbb{D}_{n,1}\exp\mathbb{D}_{n,2} - \mathbb{E}\exp\left\{\mathbf{i}t\left(\frac{\sigma_{\tilde{\alpha}}^2}{\sigma_m^2 + \sigma_{\tilde{\alpha}}^2}\right)^{1/2} Z_1\right\}\exp\mathbb{D}_{n,2}\right| &\text{(C.1)} \\
+ &\left|\mathbb{E}\exp\left\{\mathbf{i}t\left(\frac{\sigma_{\tilde{\alpha}}^2}{\sigma_m^2 + \sigma_{\tilde{\alpha}}^2}\right)^{1/2} Z_1\right\}\exp\mathbb{D}_{n,2} - \mathbb{E}\exp\left\{\mathbf{i}t\left(\frac{\sigma_{\tilde{\alpha}}^2}{\sigma_m^2 + \sigma_{\tilde{\alpha}}^2}\right)^{1/2} Z_1\right\}\exp\left\{\mathbf{i}t\left(\frac{\sigma_m^2}{\sigma_m^2 + \sigma_{\tilde{\alpha}}^2}\right)^{1/2} Z_2\right\}\right| \\
& &\text{(C.2)} \\
+ &\left|\mathbb{E}\exp\left\{\mathbf{i}t\left(\frac{\sigma_{\tilde{\alpha}}^2}{\sigma_m^2 + \sigma_{\tilde{\alpha}}^2}\right)^{1/2} Z_1\right\}\exp\left\{\mathbf{i}t\left(\frac{\sigma_m^2}{\sigma_m^2 + \sigma_{\tilde{\alpha}}^2}\right)^{1/2} Z_2\right\} - e^{-\frac{1}{2}t^2}\right|. &\text{(C.3)}
\end{aligned}$$

In the following steps we show all three terms above are $o(1)$ and the conclusion follows accordingly.

Step 1: bound term (C.1). Note the following:

$$\begin{aligned}
(C.1) = &\left|\mathbb{E}\exp\mathbb{D}_{n,2}\left\{\exp\{\mathbf{i}t\sqrt{n}(\sigma_m^2 + \sigma_{\tilde{\alpha}}^2)^{-1/2}\mathbb{E}_n\tilde{\alpha}(W)e\} - \exp\{\mathbf{i}t(\sigma_m^2 + \sigma_{\tilde{\alpha}}^2)^{-1/2}\sigma_{\tilde{\alpha}}Z_1\}\right\}\right| \\
\leq &\mathbb{E}\left|\exp\mathbb{D}_{n,2}\right|\left|\mathbb{E}\left[\exp\{\mathbf{i}t\sqrt{n}(\sigma_m^2 + \sigma_{\tilde{\alpha}}^2)^{-1/2}\mathbb{E}_n\tilde{\alpha}(W)e\} - \exp\{\mathbf{i}t(\sigma_m^2 + \sigma_{\tilde{\alpha}}^2)^{-1/2}\sigma_{\tilde{\alpha}}Z_1\} \mid \mathscr{W}_n\right]\right| \\
\leq &\mathbb{E}\left|\mathbb{E}\left[\exp\left\{\mathbf{i}t\left(\frac{\sigma_{\tilde{\alpha}}^2}{\sigma_m^2 + \sigma_{\tilde{\alpha}}^2}\right)^{1/2}\sqrt{n}\sigma_{\tilde{\alpha}}^{-1}\mathbb{E}_n\tilde{\alpha}(W)e\right\} - \exp\left\{\mathbf{i}t\left(\frac{\sigma_{\tilde{\alpha}}^2}{\sigma_m^2 + \sigma_{\tilde{\alpha}}^2}\right)^{1/2} Z_1\right\} \mid \mathscr{W}_n\right]\right|,
\end{aligned}$$

where the first inequality is by LIE and second inequality follows from the properties of a characteristic function. If we can show that for any $s \in \mathbb{R}$,

$$\left|\mathbb{P}\left\{\sqrt{n}\left(\frac{\sigma_{\tilde{\alpha}}^2}{\sigma_m^2 + \sigma_{\tilde{\alpha}}^2}\right)^{1/2}\sigma_{\tilde{\alpha}}^{-1}\mathbb{E}_n\tilde{\alpha}(W)e \leq s|\mathscr{W}_n\right\} - \mathbb{P}\left\{\left(\frac{\sigma_{\tilde{\alpha}}^2}{\sigma_m^2 + \sigma_{\tilde{\alpha}}^2}\right)^{1/2} Z_1 \leq s|\mathscr{W}_n\right\}\right| = o_p(1), \tag{C.4}$$

then the dominated convergence theorem implies that $(C.1) = o(1)$.

Step 2: show (C.4) holds for any $s \in \mathbb{R}$. Note that conditional on $\mathscr{W}_n$,

$$\begin{aligned}(C.4) &= \left|\mathbb{P}\{\sqrt{n}\sigma_{\tilde{\alpha}}^{-1}\mathbb{E}_n\tilde{\alpha}(W)e \leq s\sigma_{\tilde{\alpha}}^{-1}(\sigma_m^2+\sigma_{\tilde{\alpha},n}^2)^{1/2}|\mathscr{W}_n\} - \mathbb{P}\{Z_1 \leq s\sigma_{\tilde{\alpha}}^{-1}(\sigma_m^2+\sigma_{\tilde{\alpha}}^2)^{1/2}|\mathscr{W}_n\}\right| \\ &\leq \sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\sum_{i=1}^n \mathcal{U}_i \leq t|\mathscr{W}_n\right) - \mathbb{P}(Z_1 \leq t)\right|, \qquad \text{(C.5)}\end{aligned}$$

where $\mathcal{U}_i = n^{-1/2}\sigma_{\tilde{\alpha}}^{-1}\tilde{\alpha}(W_i)e_i$. Since $\mathbb{E}[\mathcal{U}_i|\mathscr{W}_n] = 0$ for all i and $\sum\limits_{i=1}^n Var\,(\mathcal{U}_i|\mathscr{W}_n) = 1$, $\{\mathcal{U}_i\}_{i=1}^n$ are mean zero and independent conditional on $\mathscr{W}_n$. It follows that:

$$\begin{aligned}&(C.5) \lesssim \sum_{i=1}^n \mathbb{E}\left[|\mathcal{U}_i|^3\,|\mathscr{W}_n\right] \lesssim \sigma_{\tilde{\alpha}}^{-3}n^{-3/2}\sum_{i=1}^n|\tilde{\alpha}(W_i)|^3 \\ \lesssim &\left[\frac{1}{n}\sum_{i=1}^n\tilde{\alpha}^2(W_i)\right]^{-3/2} n^{-3/2}\sum_{i=1}^n|\tilde{\alpha}(W_i)|^3 = \frac{\max_i|\tilde{\alpha}(W_i)|}{\sqrt{n}}\left[\frac{1}{n}\sum_{i=1}^n\tilde{\alpha}^2(W_i)\right]^{-1/2} = o_p(1),\end{aligned}$$

where the first inequality is by Berry-Esseen inequality, the second inequality is by $\sup_{w\in\mathcal{W}}\mathbb{E}\left[|e_i|^3\,|W_i=w\right] \lesssim 1$, the third inequality is by $\inf_{w\in\mathcal{W}}\mathbb{E}\left[e_i^2|W_i=w\right]$ bounded away from zero, and the final relation is according to conditions (i) and (ii).

Step 3: bound term (C.2).

$$\begin{aligned}(C.2) &= \left|\mathbb{E}\exp\left\{\mathrm{i}t\left(\frac{\sigma_{\tilde{\alpha}}^2}{\sigma_m^2+\sigma_{\tilde{\alpha}}^2}\right)^{1/2}Z_1\right\}\left\{\exp\mathbb{D}_{n,2} - \exp\left\{\mathrm{i}t\left(\frac{\sigma_m^2}{\sigma_m^2+\sigma_{\tilde{\alpha}}^2}\right)^{1/2}Z_2\right\}\right\}\right| \\ &= \left|\mathbb{E}\left\{\mathbb{E}\left[\exp\left\{\mathrm{i}t\left(\frac{\sigma_{\tilde{\alpha}}^2}{\sigma_m^2+\sigma_{\tilde{\alpha}}^2}\right)^{1/2}Z_1\right\}|\mathscr{W}_n\right]\mathbb{E}\left[\exp\mathbb{D}_{n,2} - \exp\left\{\mathrm{i}t\left(\frac{\sigma_m^2}{\sigma_m^2+\sigma_{\tilde{\alpha}}^2}\right)^{1/2}Z_2\right\}|\mathscr{W}_n\right]\right\}\right| \\ &\leq \left|\mathbb{E}\left[\exp\left\{\mathrm{i}t\left(\frac{\sigma_m^2}{\sigma_m^2+\sigma_{\tilde{\alpha}}^2}\right)^{1/2}\sqrt{n}\sigma_m^{-1}\mathbb{E}_n[m(W,\gamma_0)-\theta_0]\right\} - \exp\left\{\mathrm{i}t\left(\frac{\sigma_m^2}{\sigma_m^2+\sigma_{\tilde{\alpha}}^2}\right)^{1/2}Z_2\right\}\right]\right|, \\ &\qquad\qquad \text{(C.6)}\end{aligned}$$

where the second relation is by LIE and conditional independence, the third relation is due to $\mathbb{E}\left[\exp\left\{\mathrm{i}t\left(\frac{\sigma_{\tilde{\alpha}}^2}{\sigma_m^2+\sigma_{\tilde{\alpha}}^2}\right)^{1/2}Z_1\right\}|\mathscr{W}_n\right] = \exp\left\{-\frac{1}{2}\frac{\sigma_{\tilde{\alpha}}^2}{\sigma_m^2+\sigma_{\tilde{\alpha}}^2}t^2\right\} \leq 1$ and LIE.

Step 4: show $(C.6) = o(1)$. This follows from $\sqrt{n}\sigma_m^{-1}\mathbb{E}_n[m(W,\gamma_0)-\theta_0] \xrightarrow{d} Z_2$ according to Lindeberg-Lévy central limit theorem, $\frac{\sigma_m^2}{\sigma_m^2+\sigma_{\tilde{\alpha}}^2} \xrightarrow{p} v \in (0,1)$, and the continuous mapping theorem.

Step 5: show $(C.3) = o(1)$. Since $\frac{\sigma_m^2}{\sigma_m^2+\sigma_{\tilde{\alpha}}^2}$ converges to some constant, and Z_1 and Z_2 are i.i.d. standard normal, it follows by the continuous mapping theorem that:

$$\left(\frac{\sigma_{\tilde{\alpha}}^2}{\sigma_m^2+\sigma_{\tilde{\alpha}}^2}\right)^{1/2}Z_1 + \left(\frac{\sigma_m^2}{\sigma_m^2+\sigma_{\tilde{\alpha}}^2}\right)^{1/2}Z_2 \xrightarrow{d} N(0,1),$$

i.e.,

$$(C.3) = \left|\mathbb{E}\exp\left\{\mathbf{i}t\left[\left(\frac{\sigma^2_{\tilde{\alpha}}}{\sigma^2_m+\sigma^2_{\tilde{\alpha}}}\right)^{1/2} Z_1 + \left(\frac{\sigma^2_m}{\sigma^2_m+\sigma^2_{\tilde{\alpha}}}\right)^{1/2} Z_2\right]\right\} - e^{-\frac{1}{2}t^2}\right| = o(1).$$

□

Lemma C.2. *If the conditions of Theorem 5.3 hold true, then:*

(i) $\|\tilde{a}\|^{-1} = O_p(1)$;

(ii) $\{\mathbb{E}_n[\tilde{\alpha}^2(W)]\}^{-1} = O_p(1)$.

Proof. Statement (i). Let $K = \{1\dots k\}$ be the index set of $p(w) = \{p_1(w),\dots,p_k(w)\}'$, and $\tilde{K} \subseteq K$ be some index set such that its cardinality $\left|\tilde{K}\right| = \tilde{k}$. Denote $p^{\tilde{K}}(w)$ as a vector of $\tilde{k}$ basis functions selected by $\tilde{K}$: $p^{\tilde{K}}_j(w) = p_j(w)$ if and only if $j \in \tilde{K}$. Select a $\tilde{K}$ such that $\tilde{k} \to \infty$ and $\frac{\xi^2_{\tilde{k}}}{n} \to 0$. Denote $\tilde{a}^{\tilde{K}} := \left(\hat{G}^{\tilde{K}}\hat{G}^{\tilde{K}} + \lambda_n\hat{G}^{\tilde{K}}\right)^{-}\hat{G}^{\tilde{K}}\mathbb{E}_n[m(W,p^{\tilde{K}})]$, where $\hat{G}^{\tilde{K}} := \mathbb{E}_n[p^{\tilde{K}}(W)p^{\tilde{K}}(W)']$. Since $\|\tilde{a}\| \geq \left\|\tilde{a}^{\tilde{K}}\right\|$ by construction, it suffices to show $\left\|\tilde{a}^{\tilde{K}}\right\|^{-1} = O_p(1)$. Note we may write $\alpha_0 = \mathcal{L}^{\tilde{K}}_n\alpha_0 + u^{\tilde{K}}_{\alpha_o}$, where $\mathcal{L}^{\tilde{K}}_n\alpha_0 = p^{\tilde{K}\prime}a^{\tilde{K}}_l$ is the least square projection using $p^{\tilde{K}}$, $a^{\tilde{K}}_l$ is the projection coefficient and $u^{\tilde{K}}_{\alpha_0}$ is the projection error. And we may write $\tilde{a}^{\tilde{K}} = \hat{M}^{\tilde{K}}_1 + \hat{M}^{\tilde{K}}_2$, where

$$\hat{M}^{\tilde{K}}_1 := \tilde{a}^{\tilde{K}} - a^{\tilde{K}}_l, \quad \hat{M}^{\tilde{K}}_2 := a^{\tilde{K}}_l.$$

Thus, to show $\left\|\tilde{a}^{\tilde{K}}\right\|^{-1} = O_p(1)$, it suffices to show that $\left\|\hat{M}^{\tilde{K}}_2\right\|$ is bounded away from zero and $\left\|\hat{M}^{\tilde{K}}_1\right\| = o_p(1)$. First, by construction, the conditions of Lemma D.4 are met. Therefore, Lemma D.4 implies $\left\|\hat{M}^{\tilde{K}}_1\right\| = o_p(1)$. Next, we show that $\left\|\hat{M}^{\tilde{K}}_2\right\|$ is bounded away from zero. To this end, note by construction and Assumption S, $\mathbb{E}[p^{\tilde{K}}(W_i)p^{\tilde{K}}(W_i)']$ has eigenvalues bounded from above and away from zero as well. It follows that:

$$\left\|a^{\tilde{K}}_l\right\|^2 \gtrsim \left\|\mathcal{L}^{\tilde{K}}_n\alpha_0\right\|^2_{\mathbb{P},2} = \|\alpha_0\|^2_{\mathbb{P},2} - \left\|u^{\tilde{K}}_{\alpha_0}\right\|^2_{\mathbb{P},2} \geq \|\alpha_0\|^2_{\mathbb{P},2} - \left(\mathbf{r}^{\tilde{K}}_{\alpha_0}\right)^2, \tag{C.7}$$

where the first relation is because $\left\|\mathcal{L}^{\tilde{K}}_n\alpha_0\right\|^2_{\mathbb{P},2} \leq \left\|a^{\tilde{K}}_l\right\|^2\lambda_{\max}\left\{\mathbb{E}\left[p^{\tilde{K}}(W_i)p^{\tilde{K}}(W_i)'\right]\right\}$, the second relation is by Pythagoras' theorem and the last relation is by Lemma D.1(ii). Therefore, $\left\|a^{\tilde{K}}_l\right\|^2$ is bounded away from zero as $\|\alpha_0\|^2_{\mathbb{P},2} - \mathbf{r}^2_{\alpha_0}$ is bounded away from zero uniformly over all k and n by assumption. The conclusion follows.

Statement (ii). Since all eigenvalues of $\hat{G}$ are bounded away from zero wpa1 by assumption, $\lambda_{\min}^{-1}\left(\hat{G}\right) = O_p(1)$. Also, by statement (i), $\|\tilde{a}\|^{-1} = O_p(1)$. Thus,

$$\left[\mathbb{E}_n\tilde{\alpha}^2(W)\right]^{-1} = \left[\tilde{a}'\hat{G}\tilde{a}\right]^{-1} \leq \|\tilde{a}\|^{-2}\lambda_{\min}^{-1}\left[\hat{G}\right] = O_p(1).$$

□

Lemma C.3. *If the conditions of Theorem 5.3 hold true, then $R_{n,1}(\tilde{\theta}_{BP}) = o_p(1)$.*

Proof. We may still decompose $R_{n,1}(\tilde{\theta}_{BP}) = T_1 + T_{21} + T_{22}$, where

$$\begin{aligned} T_1 &= \sqrt{n}\mathbb{E}_n[\tilde{\alpha}(W)\mathcal{L}_n\gamma_0(W) - m(W,\mathcal{L}_n\gamma_0)], \\ T_{21} &= \sqrt{n}\mathbb{E}_n[(\tilde{\alpha}(W) - \alpha_0(W))u_{\gamma_0}], \\ T_{22} &= \sqrt{n}\mathbb{E}_n[\alpha_0(W)u_{\gamma_0} - m(W,u_{\gamma_0})]. \end{aligned}$$

Similar to Lemma B.2, we bound

$$\begin{aligned} |T_1| &= \sqrt{n}\left|\beta_l'\left(\hat{G}(\hat{G}\hat{G} + \lambda_n\hat{G})^{-}\hat{G} - I\right)\hat{P}\right| \\ &\leq \sqrt{n}\,\|\beta_l\|\left\|\hat{G}(\hat{G}\hat{G} + \lambda_n\hat{G})^{-}\hat{G} - I\right\|\left\|\hat{P}\right\| \\ &= \sqrt{n}\,\|\beta_l\|\left\|(\hat{G} + \lambda_n I)^{-1}\hat{G} - I\right\|\left\|\hat{P}\right\|. \end{aligned}$$

By Lemma D.1(iv), $\|\beta_l\| = O(1)$. Applying Lemma D.3, we can show that $\left\|\hat{P}\right\| = O_p(\frac{\xi_k^2\log k}{n} + \sqrt{\frac{\xi_k^2\log k}{n}} + 1)$, and $\left\|(\hat{G} + \lambda_n)^{-1}\hat{G} - I\right\| = O_p(\frac{\lambda_n}{\lambda_{\min}(\hat{G})}) = O_p(\lambda_n) = o_p\left(\frac{1}{\sqrt{n}\log k}\right)$ by Assumption M. Then, the following holds:

$$T_1 = \sqrt{n}O_p\left(\frac{\xi_k^2\log k}{n} + \sqrt{\frac{\xi_k^2\log k}{n}} + 1\right)o_p\left(\frac{1}{\sqrt{n}\log k}\right) = o_p\left(1\right).$$

For T_{21}, note that it still holds $T_{21} = \sqrt{n}\mathbb{E}_n\left[(\tilde{a} - a_l)'p(W)u_{\gamma_0}\right] + \sqrt{n}\mathbb{E}_n[u_{\alpha_0}u_{\gamma_0}]$, where under the stated conditions, $\sqrt{n}\mathbb{E}_n[u_{\alpha_0}u_{\gamma_0}] = o_p(1)$. Furthermore, applying Lemma D.5 yields that $\sqrt{n}\mathbb{E}_n\left[(\tilde{a} - a_l)'p(W)u_{\gamma_0}\right] = o_p(1)$. Thus, $T_{21} = o_p(1)$ as well. The proof for showing $T_{22} = o_p(1)$ is the same as Lemma B.3 and is omitted. □

Lemma C.4. *If the conditions of Theorem 5.3 hold true, then ${}^{\max_i|\tilde{\alpha}(W_i)|}/_{\sqrt{n}} = o_p(1)$.*

Proof. Let $\mathbf{a} := \frac{\tilde{a}}{\|\tilde{a}\|}$. Then, it follows that $\frac{\max_i|\tilde{\alpha}(W_i)|}{\sqrt{n}} = \|\tilde{a}\|\frac{\max_i|\mathbf{a}'p(W_i)|}{\sqrt{n}}$. Since $\|\tilde{a}\| = O_p(1)$ by Lemma D.4, it suffices to show that $\max_i|\mathbf{a}'p(W_i)| = o_p(\sqrt{n})$. To this end, note that $\mathbb{E}[|\mathbf{a}'p(W_i)|^2] \leq \sup_{a\in\mathbb{S}^{k-1}}\mathbb{E}[|a'p(W_i)|^2] = a'Ga \lesssim 1$ by Assumption S. It follows by Markov inequality that $\sum_{i=1}^n \mathbb{P}\{|\mathbf{a}'p(W_i)| > \sqrt{n}\} \leq \sum_{i=1}^n \frac{\mathbb{E}[|\mathbf{a}'p(W_i)|^2]}{n} \lesssim 1$. Thus, by Borel-Cantelli lemma, $|\mathbf{a}'p(W_i)| > \sqrt{n}$ happens only for finitely many n. Therefore, the conclusion follows from the same argument used in the proof of Owen (2001, Lemma 11.2). □

Lemma C.5. *Let $G = I$ and Assumptions O and S hold true. In addition, suppose $\sqrt{\frac{2\xi_k^2 \log 2k}{n}} + \frac{\xi_k^2 \log 2k}{3n} \to c_1 < 1$. Then, there exists a strictly positive constant $c_2 < 1 - c_1$ such that $\lambda_{\min}(\hat{G}) \geq c_2$ wpa1.*

Proof. Let $S_i := {p(W_i)p(W_i)'}/{n}$. Note that $\sum_{i=1}^{n} \mathbb{E}S_i = I$, that $\mathbb{E}\left[S_i - \mathbb{E}S_i\right]$ is a zero matrix by construction, and that $\|S_i - \mathbb{E}S_i\| \leq \|S_i\| \leq \frac{\xi_k^2}{n}$ by the positive semidefiniteness of $\mathbb{E}S_i$. By Tropp (2015, Theorem 6.6.1),

$$\mathbb{E}\left\|\hat{G} - I\right\| \leq \sqrt{2v_* \log 2k} + {\xi_k^2 \log 2k}/{3n} \to c_1, \tag{C.8}$$

where

$$\begin{aligned} v_* &:= \left\| \sum_{i=1}^{n} \mathbb{E}\left(S_i - \mathbb{E}S_i\right)\left(S_i - \mathbb{E}S_i\right) \right\| \leq \sum_{i=1}^{n} \left\|\mathbb{E}\left(S_i - \mathbb{E}S_i\right)\left(S_i - \mathbb{E}S_i\right)\right\| \\ &= \sum_{i=1}^{n} \left\|\mathbb{E}S_i^2 - \mathbb{E}S_i\mathbb{E}S_i\right\| \leq \sum_{i=1}^{n} \left\|\mathbb{E}S_i^2\right\| = \frac{1}{n}\left\|\mathbb{E}\left[(p(W_i)p(W_i)')^2\right]\right\| \leq \frac{\xi_k^2}{n}. \end{aligned} \tag{C.9}$$

The first relation of (C.9) is from the definition of v_*, the second relation is by the triangle inequality, the third relation is from a direct calculation, the fourth relation is due to the positive semidefiniteness of $\mathbb{E}S_i\mathbb{E}S_i$, the fifth relation is by rewriting $\mathbb{E}S_i^2$ and the i.i.d. assumption, and the final relation uses the property that for any $a \in \mathbb{S}^{k-1}$ $a'\left[p(W_i)p(W_i)'p(W_i)p(W_i)'\right]a \leq \xi_k^2 a'\left[p(W_i)p(W_i)'\right]a$. Now, suppose wpa1, $\lambda_{\min}(\hat{G}) < c_2$. Then there exists $a \in \mathbb{S}^{k-1}$ such that $a'\hat{G}a < c_2$. Thus, wpa1, we have the following:

$$\left\|\hat{G} - I\right\| \geq \left|a'(\hat{G} - I)a\right| = \left|a'\hat{G}a - 1\right| > 1 - c_2 > c_1,$$

which is a contradiction since (C.8) implies $\mathbb{P}\left\{\left\|\hat{G} - I\right\| > 1 - c_2\right\} \leq \frac{c_1}{1-c_2} < 1$. Therefore, wpa1, all eigenvalues of $\hat{G}$ are no smaller than c_2. □

D Additional technical results

D.1 Series asymptotics with many terms

The following lemmas present some basic results on series estimation which may be of independent interest.

Lemma D.1. *If Assumptions O and S hold true, then:*

(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 the results related to α_0. Those related to γ_0 can be shown in the same fashion. Note by definition,

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

Statement (i) follows from the first order condition of a_l. Statement (ii) directly follows from (D.1): $\|u_{\alpha 0}\|_{\mathbb{P},2} = \mathbb{E}[u^2_{\alpha 0 i}] \leq \mathbb{E}[(\alpha_{0i} - a_b' p_i)^2] \leq \mathbf{r}^2_{\alpha 0}$. For statement (iii), note that $u_{\alpha 0} = \alpha_0 - a_b' p + a_b' p - a_l' p$, 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, statement (iii) follows from the triangle inequality and the definition of ℓ_k. Finally, to see statement (iv), note that:

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

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

$$\|a_l\|^2 \lesssim \|\mathcal{L}_n \alpha_0\|^2_{\mathbb{P},2} \leq \|\alpha_0\|^2_{\mathbb{P},2} + \|u_{\alpha 0}\|^2_{\mathbb{P},2} = O(1 + \mathbf{r}^2_{\gamma 0}),$$

where the second inequality is by the triangle inequality, and the final relation follows from $\|\alpha_0\|_{\mathbb{P},2} = O(1)$ (Assumption O) and $\|u_{\alpha 0}\|_{\mathbb{P},2} \leq \mathbf{r}_{\gamma 0}$ (Lemma D.1(ii)). □

Lemma D.2. *If Assumptions O and S hold true, then:*

(i) $\mathbb{E}\left[\left\|\hat{G} - G\right\|\right] \lesssim \frac{\xi_k^2 \log k}{n} + \sqrt{\frac{\xi_k^2 \log k}{n}}$, $\left\|\hat{G} - G\right\| = O_p\left(\frac{\xi_k^2 \log k}{n} + \sqrt{\frac{\xi_k^2 \log k}{n}}\right)$, $\left\|\hat{G}\right\| = O_p(\frac{\xi_k^2 \log k}{n} + \sqrt{\frac{\xi_k^2 \log k}{n}} + 1)$.

(ii) If in addition $\frac{\xi_k^2 \log k}{n} = o(1)$, then $\mathbb{E}\left[\left\|\hat{G} - G\right\|\right] = o(1)$, $\left\|\hat{G} - G\right\| = o_p(1)$.

Proof. Statement (i) follows from Belloni et al. (2015, Lemma 6.2) and the triangle inequality. Statement (ii) is a direct application of statement (i). □

Lemma D.3. *Suppose Assumptions O, S hold true, $\frac{\xi_k^2}{n} \leq 1$, and all eigenvalues of $\hat{G}$ are positive wpa1. Then:*

(i) $\left\|\mathbb{E}_n\left[m(W,p)-\alpha_0(W)p(W)\right]\right\| = O_p\left(\sqrt{\frac{\xi_k^2}{n}} \wedge \sqrt{\frac{\|\alpha_0\|_{\mathbb{P},\infty} k}{n}}\right)$;

(ii) $\|\mathbb{E}_n[u_{\alpha_0} p(W)]\| = O_p\left(\mathbf{r}_{\alpha_0}\sqrt{\frac{\xi_k^2}{n}}\right)$, $\|\mathbb{E}_n[u_{\gamma_0} p(W)]\| = O_p\left(\mathbf{r}_{\gamma_0}\sqrt{\frac{\xi_k^2}{n}}\right)$;

(iii) $\hat{P} = \mathbb{E}_n\left[m(W,p)-\alpha_0(W)p(W)\right] + \mathbb{E}_n[u_{\alpha_0} p(W)] + \hat{G}a_l$, and $\left\|\hat{P}\right\| = O_p\left(\frac{\xi_k^2 \log k}{n} + \sqrt{\frac{\xi_k^2 \log k}{n}} + 1\right)$;

(iv) If in addition $\frac{\lambda_n}{\lambda_{\min}(\hat{G})} = o_p(1)$, then $\left\|\left(\hat{G}+\lambda_n I\right)^{-1}\hat{G} - I\right\| = O_p\left(\frac{\lambda_n}{\lambda_{\min}(\hat{G})}\right)$, $\left\|\hat{G}(\hat{G}\hat{G}+\lambda_n\hat{G})^{-}\hat{G} - I\right\|$
$O_p\left(\frac{\lambda_n}{\lambda_{\min}(\hat{G})}\right)$.

Proof. Statement (i). Let $e_i^R = m(W_i,p) - \alpha_0(W_i)p(W_i)$. By the definition of α_0, $\mathbb{E}e_i^R = \mathbf{0}$. By the i.i.d. assumption and the triangle inequality, the following holds:

$$\begin{aligned}\mathbb{E}\left\|\mathbb{E}_n[e^R]\right\|^2 &= \frac{1}{n}\sum_{j=1}^k \mathbb{E}\left[m(W_i,p_j) - \alpha_0(W_i)p_j(W_i)\right]^2 \\ &\lesssim \frac{1}{n}\sum_{j=1}^k \mathbb{E}m^2(W_i,p_j) + \frac{1}{n}\mathbb{E}\left[\alpha_0^2(W_i)p(W_i)'p(W_i)\right],\end{aligned}$$

where the first term is $O\left(\frac{k}{n}\right)$ since $\mathbb{E}m^2(W_i,p_j) \lesssim \mathbb{E}p_j^2(W_i) \lesssim 1$ by Assumptions O(3) and S(1). For the second term, note either $\mathbb{E}\left[\alpha_0^2(W_i)p(W_i)'p(W_i)\right] \le \xi_k^2 \mathbb{E}\alpha_0^2(W_i) \lesssim \xi_k^2$, or $\mathbb{E}\left[\alpha_0^2(W_i)p(W_i)'p(W_i)\right] \le \|\alpha_0\|_{\mathbb{P},\infty}\mathbb{E}\left[p(W_i)'p(W_i)\right] = \|\alpha_0\|_{\mathbb{P},\infty} tr(G) \lesssim \|\alpha_0\|_{\mathbb{P},\infty} k$. It follows by Markov inequality and $\frac{\xi_k^2}{n} \le 1$ that:

$$\left\|\mathbb{E}_n[e^R]\right\| = O_p\left[\sqrt{\frac{k}{n}} \vee \left(\sqrt{\frac{\xi_k^2}{n}} \wedge \sqrt{\frac{\|\alpha_0\|_{\mathbb{P},\infty} k}{n}}\right)\right] = O_p\left(\sqrt{\frac{\xi_k^2}{n}} \wedge \sqrt{\frac{\|\alpha_0\|_{\mathbb{P},\infty} k}{n}}\right).$$

Statement (ii). Note that $\mathbb{E}[u_{\alpha_0} p(W_i)] = 0$. By the i.i.d. assumption, we may show that:

$$\begin{aligned}\mathbb{E}\left\|\mathbb{E}_n[u_{\alpha_0} p(W)]\right\|^2 &= \mathbb{E}\left[\left(\mathbb{E}_n[u_{\alpha_0} p(W)]\right)'\left(\mathbb{E}_n[u_{\alpha_0} p(W)]\right)\right] \\ &= \frac{1}{n}\mathbb{E}[u_{\alpha_0}^2 p'(W)p(W)] \lesssim \left(\mathbf{r}_{\alpha_0}^2 \frac{\xi_k^2}{n}\right).\end{aligned}$$

Then, the conclusion follows from Markov inequality. $\|\mathbb{E}_n[u_{\gamma_0} p(W)]\| = O_p\left(\mathbf{r}_{\gamma_0}\sqrt{\frac{\xi_k^2}{n}}\right)$ can be shown analogously.

Statement (iii). The decomposition result is derived by rewriting the following:

$$\hat{P} = \mathbb{E}_n\left[m(W,p) - \alpha_0(W)p(W)\right] + \mathbb{E}_n\left[\alpha_0(W)p(W)\right],$$

and then by plugging $\alpha_0 = a_l' p + u_{\alpha_0}$ in the above equation. Furthermore, note

$$\begin{aligned}\left\|\hat{P}\right\| &\leq \|\mathbb{E}_n\left[m(W,p) - \alpha_0(W)p(W)\right]\| + \|\mathbb{E}_n[u_{\alpha_0} p(W)]\| + \left\|\hat{G}\right\| \|a_l\| \\ &= O_p(\sqrt{\frac{\xi_k^2}{n}} + \mathbf{r}_{\alpha_0}\sqrt{\frac{\xi_k^2}{n}} + \frac{\xi_k^2 \log k}{n} + \sqrt{\frac{\xi_k^2 \log k}{n}} + 1) \\ &= O_p(\frac{\xi_k^2 \log k}{n} + \sqrt{\frac{\xi_k^2 \log k}{n}} + 1),\end{aligned}$$

where the first relation follows from the triangle inequality, the second relation follows from statements (i), (ii), Lemmas D.2 and D.1(iv), and the last relation follows from the assumption that $\frac{\xi_k^2}{n} \lesssim 1$.

Statement (iv). By assumption, $\hat{G}$ is invertible wpa1. Thus $\Psi_n := \left(\hat{G} + \lambda_n I\right)^{-1} \hat{G}$ exists wpa1. Now, consider the orthogonal diagonalization of $\hat{G}$ such that $\hat{G} = U\Lambda U'$, where $U'U = I$, Λ is a diagonal matrix with $\{\mu_j\}_{j=1}^k$ on the diagonal, and without loss of generality $\mu_1 \geq \mu_2 \ldots \geq \mu_k$ are k real eigenvalues of $\hat{G}$. It follows that:

$$\Psi_n - I = U(\Lambda + \lambda_n I)^{-1}\Lambda U' - UU' = U\left[(\Lambda + \lambda_n I)^{-1}\Lambda - I\right]U',$$

and $(\Lambda+\lambda_n I)^{-1}\Lambda - I$ is a diagonal matrix where for each $j = 1 \ldots k$, $\lambda_j\left[(\Psi_n - I)'(\Psi_n - I)\right] = \left(\frac{\mu_j}{\mu_j+\lambda_n} - 1\right)^2 = \left(\frac{\lambda_n}{\mu_j+\lambda_n}\right)^2$. Thus, the following holds:

$$\|\Psi_n - I\| = \{\lambda_{\max}\left[(\Psi_n - I)'(\Psi_n - I)\right]\}^{1/2} = \lambda_n \frac{1}{\mu_k + \lambda_n} = O_p\left(\frac{\lambda_n}{\lambda_{\min}(\hat{G})}\right).$$

Furthermore, note that if all the eigenvalues of $\hat{G}$ are positive wpa1, then $\hat{G}(\hat{G}\hat{G} + \lambda_n \hat{G})^{-}\hat{G} = \Psi_n$ wpa1. Therefore, $\left\|\hat{G}(\hat{G}\hat{G} + \lambda_n \hat{G})^{-}\hat{G} - I\right\| = O_p\left(\frac{\lambda_n}{\lambda_{\min}(\hat{G})}\right)$ as well. □

Lemma D.4. *Suppose Assumptions O, S, and P hold true, $\frac{\xi_k^2}{n} \leq 1$, and all eigenvalues of $\hat{G}$ are bounded away from zero wpa1. Then, it holds that:*

$$\|\tilde{a} - a_l\| = O_p\left(\left(\sqrt{\frac{\xi_k^2}{n}} \wedge \sqrt{\frac{\|\alpha_0\|_{\mathbb{P},\infty}\, k}{n}}\right) + \mathbf{r}_{\alpha_0}\sqrt{\frac{\xi_k^2}{n}}\right)$$

and $\|\tilde{a}\| = O_p(1)$.

Proof. Under stated assumptions, $\tilde{a} = \left(\hat{G} + \lambda_n I\right)^{-1} \hat{P}$ wpa1. Thus,

$$\begin{aligned}\tilde{a} - a_l = &\left(\hat{G} + \lambda_n I\right)^{-1} \mathbb{E}_n \left[m(W,p) - \alpha_0(W)p(W)\right]\\ &+ \left(\hat{G} + \lambda_n I\right)^{-1} \mathbb{E}_n[u_{\alpha_0} p(W)]\\ &+ \left(\left(\hat{G} + \lambda_n I\right)^{-1} \hat{G} - I\right) a_l.\end{aligned}$$

Note that $\left\|\left(\hat{G} + \lambda_n I\right)^{-1}\right\| = O_p(1)$ since all eigenvalues of $\hat{G}$ are bounded away from zero wpa1. The conclusion follows from applying Lemmas D.3, D.1(iv) and the triangle inequality. □

Lemma D.5. *Suppose the conditions of Lemma D.4 hold true. In addition,* $\frac{\xi_k^2}{\sqrt{n}}\mathbf{r}_{\gamma_0} = o(1)$ *and* $\sqrt{n}\mathbf{r}_{\alpha_0}\mathbf{r}_{\gamma_0} = o(1)$*. Then,* $\sqrt{n}\mathbb{E}_n\left[(\tilde{a} - a_l)' p(W) u_{\gamma_0}\right] = O_p(\frac{\xi_k^2}{\sqrt{n}}\mathbf{r}_{\gamma_0} + \sqrt{n}\mathbf{r}_{\alpha_0}\mathbf{r}_{\gamma_0}) = o_p(1)$.

Proof. Note by Cauchy-Schwarz inequality, Lemmas D.3 and D.4,

$$\begin{aligned}\left|\sqrt{n}\mathbb{E}_n\left[(\tilde{a} - a_l)' p(W) u_{\gamma_0}\right]\right| &\le \sqrt{n}\left\|\tilde{a} - a_l\right\| \left\|\mathbb{E}_n\left[p(W)u_{\gamma_0}\right]\right\|\\ &= O_p(\frac{\xi_k^2}{\sqrt{n}}\mathbf{r}_{\gamma_0} + \sqrt{n}\mathbf{r}_{\alpha_0}\mathbf{r}_{\gamma_0}) = o_p(1).\end{aligned}$$

□

Lemma D.6. *If the conditions of Theorem 5.1 hold true, then,*

$$\sqrt{n}\mathbb{E}_n\left[(\tilde{a} - a_l)' p(W) u_{\gamma_0}\right] = S_{11} + S_{12} + o_p(1),$$

where

$$\begin{aligned}S_{11} &= \sqrt{n}\left(\mathbb{E}_n[e_i^R]\right)'\left(\hat{G}^{-1} - G^{-1}\right)\mathbb{E}_n[p(W)u_{\gamma_0}] = O_p\left(\frac{\xi_k^3\sqrt{\log k}}{n}\mathbf{r}_{\gamma_0}\right),\\ S_{12} &= \frac{1}{\sqrt{n}}\mathbb{E}\left[\left(e_i^R\right)' G p(W_i) u_{\gamma_0 i}\right] = O_p\left(\frac{\xi_k^2}{\sqrt{n}}\mathbf{r}_{\gamma_0}\right).\end{aligned}$$

Proof. Plugging in the form of $\tilde{a}$ yields

$$\sqrt{n}\mathbb{E}_n\left[(\tilde{a} - a_l)' p(W) u_{\gamma_0}\right] = S_1 + S_2 + S_3 + S_4,$$

where

$$
\begin{aligned}
S_1 &= \sqrt{n}\left(\mathbb{E}_n[m(W,p) - \alpha_0(W)p(W)]\right)' \hat{G}^{-1}\mathbb{E}_n[p(W)u_{\gamma 0}], \\
S_2 &= \sqrt{n}\left(\mathbb{E}_n[m(W,p) - \alpha_0(W)p(W)]\right)' ((\hat{G} + \lambda_n I)^{-1}\hat{G} - I)\hat{G}^{-1}\mathbb{E}_n[p(W)u_{\gamma 0}], \\
S_3 &= \sqrt{n}\mathbb{E}_n[u_{\alpha_0} p'(W)](\hat{G} + \lambda_n I)^{-1}\mathbb{E}_n\left[p(W)u_{\gamma 0}\right], \\
S_4 &= \sqrt{n}a_l'\left(\hat{G}(\hat{G} + \lambda_n I)^{-1} - I\right)\mathbb{E}_n\left[p(W)u_{\gamma 0}\right].
\end{aligned}
$$

Under Assumptions O, S, P and L, $S_2 = o_p(1)$, $S_3 = o_p(1)$ and $S_4 = o_p(1)$ by applying Lemmas D.3 and D.1. Since $e_i^R = m(W_i, p) - \alpha_0(W_i)p(W_i)$, we may decompose S_1 as

$$
S_1 = S_{11} + S_{12} + S_{13} + S_{14},
$$

where

$$
S_{11} = \sqrt{n}\left(\mathbb{E}_n[e_i^R]\right)'\left(\hat{G}^{-1} - G^{-1}\right)\mathbb{E}_n[p(W)u_{\gamma 0}], \qquad S_{12} = \frac{1}{\sqrt{n}}\mathbb{E}\left[\left(e_i^R\right)' Gp(W_i)u_{\gamma 0 i}\right],
$$

$$
S_{13} = \frac{1}{\sqrt{n}}\mathbb{E}_n\left[\left(e^R\right)' Gp(W)u_{\gamma_0} - \mathbb{E}[\left(e_i^R\right)' Gp(W_i)u_{\gamma 0 i}]\right] \quad S_{14} = \sqrt{n}\frac{1}{n^2}\sum_{i \neq j}^{n}\left(e_i^R\right)' Gp(X_j)u_{\gamma 0 j},
$$

for which we can show that $S_{13} = o_p(1)$ and $S_{14} = o_p(1)$ by Markov inequality. Note that $\left\|\hat{G}^{-1} - G^{-1}\right\| = O_p(\sqrt{\frac{\xi_k^2 \log k}{n}})$ (Lemma D.2). Therefore, Cauchy-Schwarz inequality and Lemma D.3 further imply $S_{11} = O_p\left(\frac{\xi_k^3\sqrt{\log k}}{n}\mathbf{r}_{\gamma 0}\right)$. It also follows that $S_{12} = O_p\left(\frac{\xi_k^2}{\sqrt{n}}\mathbf{r}_{\gamma 0}\right)$ by Cauchy-Schwarz inequality. □

D.2 Additional convergence results

The next lemma says that conditional convergence implies unconditional convergence. It sometimes helps simplify arguments for the convergence of sample averages containing complicated terms.

Lemma D.7. *Let $\{X_n\}, \{Y_n\}$ be two sequences of random vectors, and let $\{A_n\}$ be a sequence of positive numbers.*

(i) *If conditional on $\{Y_n\}$, $\|X_n\| = o_p(A_n)$, then $\|X_n\| = o_p(A_n)$ unconditionally;*

(ii) *If conditional on $\{Y_n\}$, $\|X_n\| = O_p(A_n)$, then $\|X_n\| = O_p(A_n)$ unconditionally.*

Proof. For statement (i), note that $\|X_n\| = o_p(A_n)$ conditional on Y_n means for any $\delta > 0$, $\mathbb{P}\{\|X_n\| > \delta A_n | Y_n\} \to 0$ as $n \to \infty$. Then, by Billingsley (2008, Theorem 25.12), $\mathbb{P}\{\|X_n\| > \delta A_n\} \leq \mathbb{E}[\mathbb{P}\{\|X_n\| > \delta A_n | Y_n\}] \to 0$ as well since $\mathbb{P}\{\|X_n\| > \delta A_n | Y_n\}$ is uniformly integrable. Statement (ii) follows from Chernozhukov et al. (2018, Lemma 6.1). □

Denote $\mathscr{W}_n = \{W_i\}_{i=1}^n \in \mathcal{W}^n = \prod_{i=1}^n \mathcal{W}_i$. Let $\{\mathcal{A}_j(\cdot) : \mathcal{W}^n \to \mathbb{R}, j = 1 \ldots n\}$ be a class of n functions. The following result can be used to show the convergence of sample averages involving terms e_i, $i = 1 \ldots n$.

Lemma D.8. *If Assumption O holds true, then* $\frac{1}{n}\sum_{j=1}^n [\mathcal{A}_j(\mathscr{W}_n)e_j] = O_p\left(\sqrt{\frac{\sum_{j=1}^n \mathcal{A}_j^2(\mathscr{W}_n)}{n^2}}\right)$.

Proof. Note by Assumption O, $\mathbb{E}[\mathcal{A}_j(\mathscr{W}_n)e_j|\mathscr{W}_n] = 0$ for each $j = 1 \ldots n$, and

$$var\left\{\frac{1}{n}\sum\nolimits_{j=1}^n [\mathcal{A}_j(\mathscr{W}_n)e_j]\,|\mathscr{W}_n\right\} = \frac{1}{n^2}\sum\nolimits_{j=1}^n \left[\mathcal{A}_j^2(\mathscr{W}_n)\mathbb{E}[e_j^2|\mathscr{W}_n]\right] \lesssim \frac{1}{n^2}\sum\nolimits_{j=1}^n [\mathcal{A}_j^2(\mathscr{W}_n)].$$

Conditional Markov inequality then implies $\left\{\frac{1}{n}\sum_{j=1}^n [\mathcal{A}_j(\mathscr{W}_n)e_j]\,|\mathscr{W}_n\right\} = O_p\left(\sqrt{\frac{\sum_{j=1}^n \mathcal{A}_j^2(\mathscr{W}_n)}{n^2}}\right)$. The final conclusion follows by applying Lemma D.7. □

E Further robustness checks for the empirical application

E.1 Alternative measures of corruption and selection-on-unobservables

Following Ferraz and Finan (2011), we provide additional robustness checks for the baseline results. First, Tables 7 and 8 report the effects of reelection incentives by using two alternative measures of corruption as the observed outcome: the number of irregularities associated with corruption and the share of service items involving corruption. Second, we control the ability and experience to see how the results would change if $\tilde{\theta}_{BP}$ is applied. This addresses the concern that unobserved characteristics of individual politicians might drive the main conclusion. To control for the political experience, Specifications (1)-(6) of Table 9 keep all the conditioning terms in Specifications (1)-(6) of Table 1, and add one additional proxy for the experience. This proxy indicates whether a first term mayor was in power in one of the previous three terms. To account for possible nonlinearity, Specification (7) further adds interaction terms of the political experience proxy with the other 11 continuous variables, in addition to all terms in Specification (6). A mayor's political ability may be controlled by comparing the second term mayors with a subset of the first term mayors who are reelected in subsequent elections. This reduces the sample size from 476 to 313. The results are reported in Table 10.

E.2 Many technical terms

The baseline results in Table 1 utilizes raw control variables. We now consider using many technical terms constructed from the controls, including square terms and interactions.

In Table 11, Specification (1) is the same as that in Table 1. Specification (2) keeps all controls in specification (2) of Table 1, and adds age^2, and interactions of mayoral gender, education level and age. Specification (3) keeps all the technical terms in Specification (2), and also considers a second order polynomial of municipal characteristics (raw controls, their square and interaction terms). The technical terms in Specification (4) contain those in Specification (3), and a second order polynomial of political and judicial characteristics. Specification (5) additionally includes the lottery dummies and their interactions with the indicator variable that shows whether the municipality is a judicial district. Specification (6) also considers the state dummies and their interactions with the judiciary district indicator. The estimates exhibit similar patterns with the baseline results. Even with more technical terms, our estimator seems to perform more robustly compared to the other estimators in Table 11.

E.3 Sensitivity analysis with respect to penalty terms

Treating $\tilde{\theta}_{BP}$ as a finite-sample minimax estimator, we now conduct sensitivity analysis of the estimator with respect to different penalty coefficients. Recall that in a homoscedastic model, the ideal penalty coefficient is $\lambda_n = \frac{\sigma^2}{nb^2}$. Thus, we can gauge the magnitude of λ_n by $\hat{\sigma}^2$ and $\hat{b}^2$, where $\hat{b}$ is the l_2 norm of the estimated coefficient of the conditional expectation function, and $\hat{\sigma}^2$ is an average of the associated residual squares. In a full specification with 67 controls, we can calculate that ${\hat{\sigma}_1^2}/{\hat{b}_1^2} = 0.021$ for $\mathbb{E}[Y_i|X_i, T_i = 1]$ and ${\hat{\sigma}_0^2}/{\hat{b}_0^2} = 0.019$ for $\mathbb{E}[Y_i|X_i, T_i = 0]$. Therefore, we apply $\tilde{\theta}_{BP}$ with ${\sigma^2}/{b^2} \in (0, 100)$, assuming the ratio ${\sigma^2}/{b^2}$ is the same for $\mathbb{E}[Y_i|X_i, Z_i, T_i = 1]$ and $\mathbb{E}[Y_i|X_i, Z_i, T_i = 0]$. The results are illustrated in Figure E.1.

Table 7: Effect of reelection incentives on alternative measure of corruption

Specification		(1)	(2)	(3)	(4)	(5)	(6)
Corruption measure		numbers of irregularities involving corruption					
k		1	21	28	32	41	67
n		476	476	476	476	476	476
Controlled OLS	Effect	-0.3875**	-0.4297***	-0.3641**	-0.3947**	-0.4470***	-0.4710***
	S.E.	(0.1583)	(0.1549)	(0.1525)	(0.1530)	(0.1506)	(0.1478)
Our estimator $\tilde{\theta}_{BP}$	Effect	-0.3857**	-0.4320***	-0.3477**	-0.3528**	-0.3568***	-0.3568***
$\lambda_{n,1} = \lambda_{n,0} = 0.001$	S.E.	(0.1576)	(0.1495)	(0.1421)	(0.1409)	(0.1343)	(0.1225)
Controlled ridge	Effect		-0.4222***	-0.3451***	-0.3804**	-0.4226***	-0.4584***
$\lambda_n = 0.001$	S.E.		(0.1549)	(0.1524)	(0.1527)	(0.1506)	(0.1490)
Controlled ridge	Effect		-0.2072	-0.2405	-0.2628*	-0.3155**	-0.2484
10 fold CV	S.E.		(0.1562)	(0.1534)	(0.1533)	(0.1523)	(0.1541)
Debiased w.	Effect		-0.3612**	-0.2864*	-0.4123***	-0.4123***	-0.4035***
post lasso selection	S.E.		(0.1589)	(0.1507)	(0.1540)	(0.1540)	(0.1514)
Debiased w.	Effect		-0.3873**	-0.3445**	-0.4402***	-0.4408***	-0.4323***
lasso selection	S.E.		(0.1581)	(0.1568)	(0.1554)	(0.1553)	(0.1551)
Linear partialing out	Effect		-0.4359***	-0.3048**	-0.3797**	-0.3581**	-0.3872***
post lasso selection	S.E.		(0.1572)	(0.1442)	(0.1494)	(0.1493)	(0.1393)
Linear double selection	Effect		-0.4367***	-0.3152**	-0.3872***	-0.3618**	-0.3998***
post lasso selection	S.E.		(0.1579)	(0.1452)	(0.1456)	(0.1460)	(0.1406)
Mayoral 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 dummies		No	No	No	No	Yes	Yes
State dummies		No	No	No	No	No	Yes

Note: k is the number of conditioning terms and n is the sample size. Numbers in parentheses are computed standard errors. (1)-(6) use the same controls as those used in Table 4 of Ferraz and Finan (2011). Mayoral characteristics include age, gender, education and party affiliation. Municipal characteristics include the log of population, the percentage of the population that has at least a secondary education, the percentage of the population that lives in the urban sector, new municipality, the log of GDP per capita in 2002, the Gini coefficient, and the log amount of resources sent to the municipality. Political and judicial characteristics include the effective number of political parties in the legislature, the number of legislators divided by the number of voters, the share of the legislature that is of the same party as the mayor, and whether the municipality is a judiciary district. Two ridge methods use the R package "glmnet"; four lasso based methods use R package "hdm".
*** Significant at 1%.
** Significant at 5 %.
* Significant at 10%.

Table 8: Effect of reelection incentives on alternative measures of corruption

Specification		(1)	(2)	(3)	(4)	(5)	(6)
Corruption measure		Share of audited items involving corruption					
k		1	21	28	32	41	67
n		476	476	476	476	476	476
Controlled OLS	Effect	-0.0076	-0.0100***	-0.0077	-0.0081*	-0.0100**	-0.0105**
	S.E.	(0.0048)	(0.0045)	(0.0047)	(0.0047)	(0.0044)	(0.0044)
Our estimator $\tilde{\theta}_{BP}$	Effect	-0.0076	-0.0091**	-0.0062	-0.0057	-0.0055	-0.0055
$\lambda_{n,1} = \lambda_{n,0} = 0.001$	S.E.	(0.0048)	(0.0044)	(0.0044)	(0.0044)	(0.0039)	(0.0036)
Controlled ridge	Effect		-0.0098**	-0.0074	-0.0080*	-0.0096**	-0.0103**
$\lambda_n = 0.001$	S.E.		(0.0045)	(0.0047)	(0.0047)	(0.0044)	(0.0043)
Controlled ridge	Effect		-0.0030	-0.0035	-0.0043	-0.0066	-0.0058
10 fold CV	S.E.		(0.0046)	(0.0047)	(0.0047)	(0.0045)	(0.0045)
Debiased w.	Effect		-0.0067	-0.0043	-0.0069	-0.0074*	-0.0049
post lasso selection	S.E.		(0.0048)	(0.0049)	(0.0048)	(0.0045)	(0.0048)
Debiased w.	Effect		-0.0076	-0.0063	-0.0080*	-0.0080*	-0.0078*
lasso selection	S.E.		(0.0048)	(0.0048)	(0.0047)	(0.0047)	(0.0047)
Linear partialing out	Effect		-0.0099**	-0.0055	-0.0073	-0.0082*	-0.0073*
post lasso selection	S.E.		(0.0048)	(0.0046)	(0.0048)	(0.0045)	(0.0041)
Linear double selection	Effect		-0.0099**	-0.0058	-0.0074	-0.0081*	-0.0077*
post lasso selection	S.E.		(0.0047)	(0.0046)	(0.0046)	(0.0043)	(0.0041)
Mayoral 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 dummies		No	No	No	No	Yes	Yes
State dummies		No	No	No	No	No	Yes

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

*** Significant at 1%.

** Significant at 5 %.

* Significant at 10%.

Table 9: Effect of reelection incentives on corruption: Controlling for political experience

Specification		(1)	(2)	(3)	(4)	(5)	(6)	(7)
k		1+1	21+1	28+1	32+1	41+1	67+1	67+1+11
n		476	476	476	476	476	476	476
Controlled OLS	Effect	-0.0164*	-0.0178*	-0.0179*	-0.0217*	-0.0243**	-0.0262**	-0.0246**
	S.E.	(0.0099)	(0.0101)	(0.0103)	(0.0113)	(0.0110)	(0.0116)	(0.0122)
Our estimator $\tilde{\theta}_{BP}$	Effect	-0.0179*	-0.0185**	-0.0159*	-0.0179*	-0.0179*	-0.0178*	-0.0177*
$\lambda_{n,1} = \lambda_{n,0} = 0.001$	S.E.	(0.0098)	(0.0087)	(0.0088)	(0.0096)	(0.0095)	(0.0092)	(0.0093)
Controlled ridge	Effect		-0.0195**	-0.0176*	-0.0233**	-0.0238**	-0.0250**	-0.0243**
$\lambda_n = 0.001$	S.E.		(0.0096)	(0.0103)	(0.0108)	(0.0110)	(0.0117)	(0.0122)
Controlled ridge	Effect		-0.0070	-0.0070	-0.0100	-0.0073	-0.0051	-0.0041
10 fold CV	S.E.		(0.0097)	(0.0105)	(0.0110)	(0.0113)	(0.0123)	(0.0130)
Debiased w.	Effect		-0.0180*	-0.0188*	-0.0252**	-0.0230**	-0.0173	-0.0486
post lasso selection	S.E.		(0.0094)	(0.0101)	(0.0111)	(0.0111)	(0.0111)	(0.0346)
Debiased w.	Effect		-0.0188**	-0.0176**	-0.0225**	-0.0214**	-0.0210**	-0.0249*
lasso selection	S.E.		(0.0095)	(0.0094)	(0.0100)	(0.0099)	(0.0098)	(0.0132)
Linear partialing out	Effect		-0.0177*	-0.0169*	-0.0248***	-0.0231**	-0.0202**	-0.0196**
post lasso selection	S.E.		(0.0093)	(0.0095)	(0.0096)	(0.0099)	(0.0099)	(0.0099)
Linear double selection	Effect		-0.0180*	-0.0170*	-0.0248**	-0.0232**	-0.0210*	-0.0204*
post lasso selection	S.E.		(0.0096)	(0.0100)	(0.0104)	(0.0110)	(0.0111)	(0.0112)
Mayoral characteristics		No	Yes	Yes	Yes	Yes	Yes	Yes
Municipal characteristics		No	No	Yes	Yes	Yes	Yes	Yes
Political and judicial characteristics		No	No	No	Yes	Yes	Yes	Yes
Lottery dummies		No	No	No	No	Yes	Yes	Yes
State dummies		No	No	No	No	No	Yes	Yes

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

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

Table 10: Effect of reelection incentives on corruption: Controlling for political ability

Specification		(1)	(2)	(3)	(4)	(5)	(6)
k		1	21	28	32	41	67
n		313	313	313	313	313	313
Controlled OLS	Effect	-0.0345***	-0.0356***	-0.0358***	-0.0411***	-0.0418***	-0.0398***
	S.E.	(0.0097)	(0.0103)	(0.0109)	(0.0118)	(0.0122)	(0.0130)
Our estimator $\tilde{\theta}_{BP}$	Effect	-0.0345***	-0.0303***	-0.0310***	-0.0330***	-0.0330***	-0.0329***
$\lambda_{n,1} = \lambda_{n,0} = 0.001$	S.E.	(0.0097)	(0.0087)	(0.0091)	(0.0092)	(0.0092)	(0.0091)
Controlled ridge	Effect		-0.0349***	-0.0351***	-0.0402***	-0.0409***	-0.0385***
$\lambda_n = 0.001$	S.E.		(0.0104)	(0.0110)	(0.0119)	(0.0123)	(0.0131)
Controlled ridge	Effect		-0.0160	-0.0140	-0.0167	-0.0125	-0.0102
10 fold CV	S.E.		(0.0104)	(0.0113)	(0.0123)	(0.0127)	(0.0140)
Debiased w.	Effect		-0.0344***	-0.0351***	-0.0405***	-0.0405***	-0.0377***
post lasso selection	S.E.		(0.0100)	(0.0103)	(0.0109)	(0.0109)	(0.0112)
Debiased w.	Effect		-0.0338***	-0.0337***	-0.0357***	-0.0357***	-0.0353***
lasso selection	S.E.		(0.0097)	(0.0097)	(0.0097)	(0.0097)	(0.0096)
Linear partialing out	Effect		-0.0326***	-0.0305***	-0.0359***	-0.0359***	-0.0371***
post lasso selection	S.E.		(0.0111)	(0.0111)	(0.0114)	(0.0114)	(0.0112)
Linear double selection	Effect		-0.0338***	-0.0314***	-0.0370***	-0.0370***	-0.0385***
post lasso selection	S.E.		(0.0097)	(0.0098)	(0.0107)	(0.0107)	(0.0108)
Mayoral 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 dummies		No	No	No	No	Yes	Yes
State dummies		No	No	No	No	No	Yes

Note: This table uses only a subsample of second-term mayors and first-term mayors who were later reelected as a control for political ability. k is the number of conditioning terms and n is the sample size. Numbers in parentheses represent computed standard errors. (1)-(6) use the same controls as those in Table 4 of Ferraz and Finan (2011). Mayoral characteristics include age, gender, education and party affiliation. Municipal characteristics include the log of population, the percentage of the population that has at least a secondary education, the percentage of the population that lives in an urban sector, new municipality, the log of GDP per capita in 2002, the Gini coefficient, the log amount of resources sent to the municipality. Political and judicial characteristics include the effective number of political parties in the legislature, the number of legislators divided by the number of voters, the share of the legislature that is of the same party as the mayor, and whether the municipality is a judiciary district.
*** Significant at 1%.
** Significant at 5 %.
* Significant at 10%.

Table 11: Effect of reelection incentives on corruption: Many technical terms

Specification		(1)	(2)	(3)	(4)	(5)	(6)
k		1	25	59	72	90	142
n		476	476	476	476	476	476
Controlled OLS	Effect	-0.0188**	-0.0186*	-0.0189*	-0.0196*	-0.0214*	-0.0247**
	S.E.	(0.0095)	(0.0096)	(0.0102)	(0.0107)	(0.0110)	(0.0118)
Our estimator $\tilde{\theta}_{BP}$	Effect	-0.0187**	-0.0171**	-0.0179**	-0.0182**	-0.0182**	-0.0182**
$\lambda_{n,1} = \lambda_{n,0} = 0.001$	S.E.	(0.0094)	(0.0081)	(0.0087)	(0.0087)	(0.0086)	(0.0085)
Plug-in ridge	Effect		-0.0181**	-0.0200**	-0.0236**	-0.0252**	-0.0293***
$\lambda_{n,1} = \lambda_{n,0} = 0.001$	S.E.		(0.0092)	(0.0091)	(0.0102)	(0.0099)	(0.0095)
Plug-in ridge	Effect		-0.0188*	-0.0188*	-0.0188*	-0.0188*	-0.0188*
10 fold CV	S.E.		(0.0096)	(0.0098)	(0.0115)	0.0114)	(0.0113)
Debiased w.	Effect		-0.0195*	-0.0173*	-0.0247**	-0.0260**	-0.0222*
post lasso selection	S.E.		(0.0094)	(0.0096)	(0.0117)	0.0120)	(0.0117)
Debiased w.	Effect		-0.0188**	-0.0178*	-0.0223**	-0.0223**	-0.0218**
lasso selection	S.E.		(0.0095)	(0.0094)	(0.0100)	0.0100)	(0.0099)
Linear partialing out	Effect		-0.0194**	-0.0174*	-0.0234**	-0.0247**	-0.0250***
post lasso selection	S.E.		(0.0093)	(0.0094)	(0.0097)	0.0096)	(0.0096)
Linear double selection	Effect		-0.0196**	-0.0175*	-0.0234**	-0.0249**	-0.0262***
post lasso selection	S.E.		(0.0096)	(0.0095)	(0.0103)	0.0103)	(0.0101)
Mayor characteristics and their series terms		No	Yes	Yes	Yes	Yes	Yes
Municipal characteristics and their series terms		No	No	Yes	Yes	Yes	Yes
Political and judicial characteristics and their series terms		No	No	No	Yes	Yes	Yes
Lottery dummies and interactions with judical chacteristics		No	No	No	No	Yes	Yes
State dummies and interactions with judical chacteristics		No	No	No	No	No	Yes

Note: k is the number of technical terms and n is the sample size. Numbers in parentheses represent computed standard errors. Ridge methods use the R package "glmnet"; four lasso based methods use the R package "hdm". For plug-in (cross validated) ridges, standard error is calculated with α_0 estimated using the estimator in Newey and Robins (2018).
*** Significant at 1%. ** Significant at 5 %. * Significant at 10%.

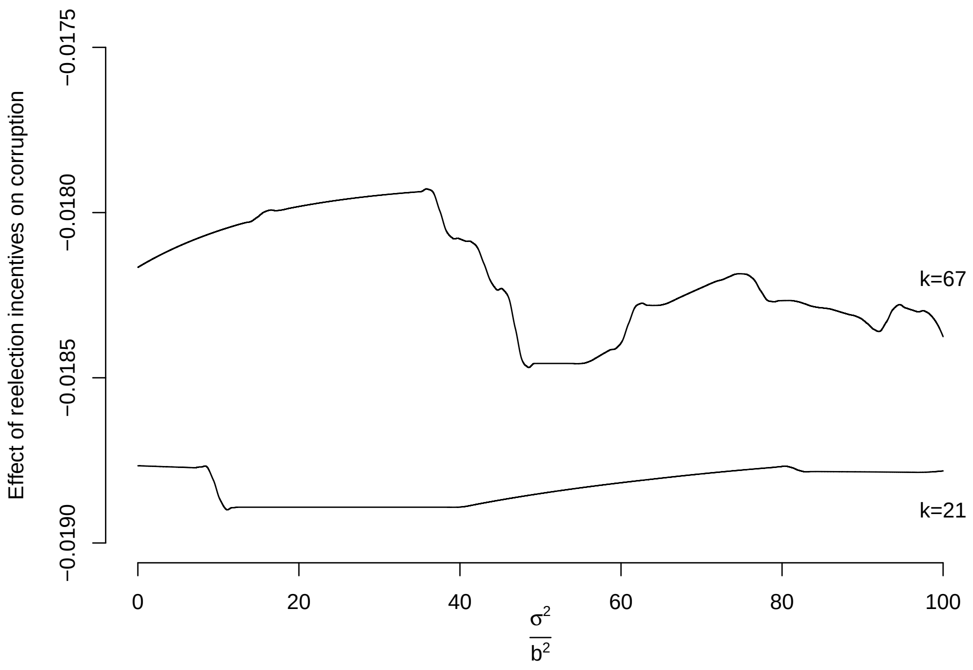

Figure E.1: $\tilde{\theta}_{BP}$ with penalty coefficients optimally selected against a range of $\frac{\sigma^2}{b^2} \in (0, 100)$. The first line in the above figure represents when all controls are added with $k = 67$. The second line represents when only mayoral characteristics are included with $k = 21$. Both curves are smoothed using local polynomials.

References

Abadie, A. and Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1):1–11.

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

Altonji, J. G., Ichimura, H., and Otsu, T. (2012). Estimating derivatives in nonseparable models with limited dependent variables. *Econometrica*, 80(4):1701–1719.

Andrews, D. W. (1994). Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica: Journal of the Econometric Society*, pages 43–72.

Angrist, J. D. (1998). Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica*, pages 249–288.

Angrist, J. D. and 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. B. and Kolesár, M. (2018). Optimal inference in a class of regression models. *Econometrica*, 86(2):655–683.

Armstrong, T. B. and Kolesár, M. (2021). Finite-sample optimal estimation and inference on average treatment effects under unconfoundedness. *Econometrica*, 89(3):1141–1177.

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., 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.

Bickel, P. J. (1982). On Adaptive Estimation. *The Annals of Statistics*, 10(3):647–671.

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., and Ritov, Y. (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press Baltimore.

Billingsley, P. (2008). *Probability and measure*. John Wiley & Sons.

Cattaneo, M. D., Crump, R. K., and Jansson, M. (2013). Generalized jackknife estimators of weighted average derivatives. *Journal of the American Statistical Association*, 108(504):1243–1256.

Cattaneo, M. D. and Farrell, M. H. (2013). Optimal convergence rates, bahadur representation, and asymptotic normality of partitioning estimators. *Journal of Econometrics*, 174(2):127–143.

Cattaneo, M. D., Farrell, M. H., and Feng, Y. (2020). Large sample properties of partitioning-based series estimators. *Annals of Statistics*, 48(3):1718–1741.

Cattaneo, M. D. and Jansson, M. (2019). Average density estimators: Efficiency and bootstrap consistency. *arXiv preprint arXiv:1904.09372*.

Cattaneo, M. D., Jansson, M., and Ma, X. (2019). Two-step estimation and inference with possibly many included covariates. *The Review of Economic Studies*, 86(3):1095–1122.

Cattaneo, M. D., Jansson, M., and Newey, W. K. (2018a). Alternative asymptotics and the partially linear model with many regressors. *Econometric Theory*, 34(2):277–301.

Cattaneo, M. D., Jansson, M., and Newey, W. K. (2018b). Inference in linear regression models with many covariates and heteroscedasticity. *Journal of the American Statistical Association*, 113(523):1350–1361.

Chan, K. C. G., Yam, S. C. P., and Zhang, Z. (2016). Globally efficient 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., Hong, H., and Tarozzi, A. (2008). Semiparametric efficiency in gmm models with auxiliary data. *The Annals of Statistics*, 36(2):808–843.

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. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. (2022a). Locally robust semiparametric estimation. *Econometrica*, 90(4):1501–1535.

Chernozhukov, V., Newey, W. K., and Singh, R. (2022b). Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90(3):967–1027.

Chernozhukov, V., Newey, W. K., and Singh, R. (2022c). Debiased machine learning of global and local parameters using regularized riesz representers. *The Econometrics Journal*.

Chiappori, P.-A., Oreffice, S., and Quintana-Domeque, C. (2012). Fatter attraction: anthropometric and socioeconomic matching on the marriage market. *Journal of Political Economy*, 120(4):659–695.

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

Donoho, D. L. (1994). Statistical estimation and optimal recovery. *The Annals of Statistics*, pages 238–270.

Fan, Q., Hsu, Y.-C., Lieli, R. P., and Zhang, Y. (2022). Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business & Economic Statistics*, 40(1):313–327.

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.

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 non-parametric least squares. Technical report, Working paper.

Härdle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American statistical Association*, 84(408):986–995.

Harville, D. A. (1998). Matrix algebra from a statistician's perspective.

Hausman, J. A. and Newey, W. K. (1995). Nonparametric estimation of exact consumers surplus and deadweight loss. *Econometrica: Journal of the Econometric Society*, pages 1445–1476.

Hausman, J. A. and Newey, W. K. (2016). Individual heterogeneity and average welfare. *Econometrica*, 84(3):1225–1248.

Hausman, J. A. and Newey, W. K. (2017). Nonparametric welfare analysis. *Annual Review of Economics*, 9:521–546.

Hirshberg, D. A. and Wager, S. (2021). Augmented minimax linear estimation. *The Annals of Statistics*, 49(6):3206–3227.

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

Imbens, G. and Wager, S. (2019). Optimized regression discontinuity designs. *Review of Economics and Statistics*, 101(2):264–278.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29.

Imbens, G. W. and Newey, W. K. (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, 77(5):1481–1512.

Imbens, G. W. and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86.

Kallus, N. (2016). Generalized optimal matching methods for causal inference. *arXiv preprint arXiv:1612.08321*.

Kallus, N. (2020). Generalized optimal matching methods for causal inference. *J. Mach. Learn. Res.*, 21:62–1.

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.

Khan, S. and Tamer, E. (2010). Irregular identification, support conditions, and inverse weight estimation. *Econometrica*, 78(6):2021–2042.

Li, K.-C. (1982). Minimaxity of the method of regularization of stochastic processes. *The Annals of Statistics*, 10(3):937–942.

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

Newey, W. K. (1991). Uniform convergence in probability and stochastic equicontinuity. *Econometrica: Journal of the Econometric Society*, pages 1161–1167.

Newey, W. K. (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., Hsieh, F., and Robins, J. (1998). Undersmoothing and bias corrected functional estimation. Technical report.

Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578.

Newey, W. K. and Robins, J. R. (2018). Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*.

Newey, W. K. and Stoker, T. M. (1993). Efficiency of weighted average derivative estimators and index models. *Econometrica: Journal of the Econometric Society*, pages 1199–1223.

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204.

Owen, A. B. (2001). *Empirical likelihood*. Chapman and Hall/CRC.

Powell, J. L., Stock, J. H., and Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica: Journal of the Econometric Society*, pages 1403–1430.

Qiu, C. (2020). Near optimal estimation of average regression functionals. *working paper*.

Qiu, C. and Otsu, T. (2022). Information theoretic approach to high-dimensional multiplicative models: Stochastic discount factor and treatment effect. *Quantitative Economics*, 13(1):63–94.

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

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954.

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.

Stock, J. H. (1989). Nonparametric policy analysis. *Journal of the American Statistical*

Association, 84(406):567–575.

Stoker, T. M. (1986). Consistent estimation of scaled coefficients. *Econometrica: Journal of the Econometric Society*, pages 1461–1481.

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

Van Der Vaart, A. et al. (1991). On differentiable functionals. *The Annals of Statistics*, 19(1):178–204.

Van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3. Cambridge university press.

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

Zimmert, M. and Lechner, M. (2019). Nonparametric estimation of causal heterogeneity under high-dimensional confounding. *arXiv preprint arXiv:1908.08779*.

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.