ESSAYS ON HIGH-DIMENSIONAL ECONOMETRICS WITHOUT SPARSITY AND BOUNDS ON STANDARD ERRORS

By

Jooyoung Cha

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Economics

May 9, 2025

Nashville, Tennessee

Approved:

Committee Chair, Yuya Sasaki, Ph.D.

Committee member, Atsushi Inoue, Ph.D.

Committee member, Tong Li, Ph.D.

Committee member, Ke-Li Xu, Ph.D.

ACKNOWLEDGMENTS

I am deeply grateful to my advisor, Yuya, for his unwavering support throughout my Ph.D. journey. His guidance, patience, and insightful feedback have been invaluable in shaping this dissertation and my growth as a researcher. I would also like to extend my sincere appreciation to my committee members, Atsushi, Tong, and Ke-Li, for their thoughtful comments, invaluable support, and guidance throughout this process. Their expertise and encouragement have been instrumental in shaping my work. Beyond my committee, I am grateful to Harold for his mentorship and collaboration during our coauthorship. His insights and support have been greatly appreciated throughout this journey.

I have also greatly benefited from discussions and feedback received at various conferences, including SEA, KAEA, MEG, CESG, Bristol ESG, IAAE, SETA, and NY Camp. I also want to acknowledge the MEG Women's Mentoring Session, which provided a valuable space for connecting with other women in the field. Additionally, I appreciate the thoughtful feedback and discussions I received during my visits to various institutions, which provided valuable perspectives on my work. I am particularly grateful to Vadim for his thoughtful advice, to Chuanping for his keen insights on our shared research, and to Andrii for his invaluable guidance and detailed feedback.

I am also thankful for the friendships that have enriched my time at Vanderbilt. My colleagues have provided not only intellectual engagement but also a sense of camaraderie that made this journey truly meaningful. Special thanks to my wonderful cohorts, whose support and friendship have made this experience all the more rewarding. I am also deeply grateful to Yukun, Terry, and Jihye for their invaluable support and encouragement throughout my time at Vanderbilt. I deeply appreciate the faculty members at Vanderbilt, whose guidance, in ways big and small, has been instrumental in completing this process. I am also deeply thankful to Sarah and to Yulong from Syracuse for their invaluable support in helping me refine my arguments and articulate my ideas more effectively. Their thoughtful advice has been instrumental in sharpening my research and broader academic thinking.

My deepest gratitude goes to my family. My mother, Jinsun Chun, and my father, Junhee Cha, have been my unwavering pillars of support, and my brother, Sanghyun Cha, has always been there for me. Their love and encouragement have been the foundation of my resilience. Finally, I am especially grateful to those who have been by my side beyond academia, particularly my dearest friends Yoomin and Hyomin, whose friendship has been invaluable. And to Jooyoung, thank you for always listening, understanding, and standing by me through it all.

This dissertation is a culmination of the support, guidance, and kindness of many. To everyone who has been part of this journey, I extend my deepest gratitude.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Local Projections Inference with High-dimensional Covariates without Sparsity

1.1 Introduction

Impulse responses analyze how a shock, such as a change in interest rates, affects various aspects of the economy over time. Knowing how different parts of the economy respond to shocks, such as policy changes or natural shocks, helps policymakers make informed decisions and businesses prepare for potential impacts. In recent years, local projections (LPs), introduced by Jordà (2005), have gained considerable attention as a flexible alternative to traditional Vector Autoregression (VAR) models. LPs directly estimate the relationship between shocks and future outcomes, offering simplicity and intuitive interpretability.

As the use of LPs has grown, so has the recognition of the need for high-dimensional controls in these models. While methods like LASSO have been proposed to handle high-dimensional covariates, they heavily rely on sparsity assumptions——the idea that most parameters are exactly zero. Recent work by Adamek et al. (2023, 2024) proposed using a debiased or desparsified LASSO in time series and local projections with weaker form of sparsity, but their approach still faces challenges in scenarios with dense data generating processes (DGPs). Recent studies by Giannone et al. (2021) and Kolesár et al. (2023) have raised significant concerns about the validity and reliability of these assumptions in economic models, challenging their applicability in many empirical settings.

This paper presents a novel approach to local projections that incorporates high-dimensional covariates without relying on sparsity constraints. The proposed method adapts the Orthogonal Greedy Algorithm with a High-Dimensional Akaike Information Criterion (OGA+HDAIC), originally developed by Ing (2020), to the context of local projections. In contrast to LASSO-based methods, the proposed approach allows parameters to be nonzero but assumes they decay toward zero as dimensionality grows. This assumption is more flexible and is automatically satisfied by autoregressive processes, making it especially suitable for time series settings. This approach offers a more versatile framework that can handle both sparse and dense scenarios, which is ideal for economic time series analysis.

The Orthogonal Greedy Algorithm proposed in this paper offers several advantages over existing methods, particularly in terms of interpretability. Unlike Principal Component methods (Stock and Watson, 2002), the proposed method prioritizes variables based on their cross-sectional explanatory power. This feature not only enhances the model's interpretability but also provides valuable insights into the most influential covariates in impulse response dynamics. Such information is crucial for economists and policymakers seeking to understand shock propagation mechanisms in the economy. Recent work by Dinh et al. (2024) has applied random subspace methods in this domain, sharing the

motivation to address high-dimensional controls in local projections. While their approach offers empirical insights, the proposed method provides a statistical inference theory in addition to practical applicability.

From a theoretical perspective, this paper contributes to the literature by extending the OGA+HDAIC method of Ing (2020) to the context of local projections. My coauthors and I previously adapted this method to a cross-sectional double machine learning framework (Cha et al., 2023). In the current paper, I further develop this approach for the time series context by incorporating explicit assumptions on the dependence structure, using the near epoch dependence (NED) assumption as employed by Adamek et al. (2023, 2024). A key advantage of using the NED assumption is that it inherits the mixingale properties that I employ in deriving the error bounds.[1] I leverage the triplex inequality from Jiang (2009) in conjunction with the results in Ing (2020) to derive the inference of the proposed method. In deriving the inference, I incorporate a double selection method (Belloni et al., 2013a), which effectively addresses potential overfitting issues in high-dimensional dynamics involving lagged covariates and outcomes. This framework enables the derivation of error bounds and asymptotic Gaussianity, providing both theoretical guarantees and practical implementation strategies.

Building on the solid statistical background, the proposed method's ability to handle high-dimensional controls has important implications for identification, addressing several key challenges. From a causal identification perspective, Angrist et al. (2018) proposed interpreting local projections as causal parameters, with the key identifying assumption being unconfoundedness. The inclusion of large sets of covariates, which the proposed method efficiently manages, enhances the plausibility of this assumption (D'Amour et al., 2021; Rosenbaum and Rosenbaum, 2002). By incorporating a comprehensive set of potential confounders, we can more confidently assume unconfoundedness, strengthening the causal interpretation of the estimates.

In terms of identifying structural impulse responses, the proposed method addresses the invertibility condition by essentially solving an omitted variable bias problem (Stock and Watson, 2016, p. 450). When relying on the LP-VAR equivalence results from Plagborg-Møller and Wolf (2021), we must consider the bias term arising from finite lag approximations. This bias decreases as the number of included lags increases, further justifying the use of high-dimensional controls.

In a Difference-in-Differences (DiD) framework, Dube et al. (2023) extend local projections to estimate DiD parameters, with the parallel trends assumption the key assumption for identification. Building on this identification scheme, my framework introduces high-dimensional controls, which greatly enhance the credibility of the conditional parallel trends assumption. As highlighted by Heckman et al. (1997, 1998), factors like demographic differences or regional economic conditions can lead to non-parallel trends, introducing bias into traditional DiD estimates. By

[1] I have benefited enormously from the comprehensive foundation of dependence concepts and properties in the exquisite textbook Davidson (1994).

incorporating a rich set of covariates, high-dimensional methods address these potential biases, making the conditional parallel trends assumption more plausible.

By addressing these identification challenges through the inclusion of high-dimensional controls, the proposed method enhances their interpretability and potential for causal inference. In the following Section 1.2, I will demonstrate how these high-dimensional methods are applied to various identification schemes, including structural impulse response estimation and DiD approaches, providing a comprehensive framework for modern empirical analysis.

To demonstrate the effectiveness of this approach, I conduct illustrative simulation studies comparing the proposed method with conventional LP and LASSO-based approaches. For the simulation design, I adapt from Montiel Olea and Plagborg-Møller (2021), where they proposed lag augmentation—adding one more lag as controls than the true autoregressive model suggests, to achieve robust inference with persistent data at longer horizons. I adopt lag augmentation for all the methods to see if such robust inference holds in finite samples. Simulation results suggest that the proposed method performs consistently well, especially in dense and more persistent scenarios. This is particularly relevant for macroeconomic applications where data often exhibit high persistence and complex interdependencies.

In empirical applications, I apply the proposed method to the following two studies. First, I revisit the analysis of Acemoglu et al. (2019) on the causal effects of democracy on economic growth. The proposed method demonstrates significant efficiency gains over the conventional local projections approach. It also shows robustness to high-dimensional controls. This application highlights the method's practical use in empirical analysis, where high-dimensional confounders play a critical role.

Next, I reexamine the study by Bhandari et al. (2024), which explores how subjective beliefs, particularly pessimism, influence macroeconomic aggregates like inflation and unemployment. The proposed method shows robustness even as model complexity increases, particularly when incorporating a moderate number of variables with extended lags. This robustness across varying model specifications is crucial in empirical analysis, where the true underlying model structure is often unknown.

The remainder of this paper is organized as follows. Section 1.2 presents the notations and implementation details of the proposed method with two main applications for the identification schemes. Section 1.3 illustrates its finite sample performance through simulation studies. Section 3.4 outlines the theoretical framework, including definitions, assumptions, and main results. Section 1.5 presents two empirical applications: one examining the effect of democratization on GDP growth, and the other investigating the impact of pessimism on macroeconomic aggregates. Finally, Section 1.6 includes concluding remarks.

Related Literature. The literature on local projection provides important insights into estimating dynamic causal effects. Following the initial introduction as a tool for impulse responses by Jordà (2005), Angrist et al. (2018) developed

an identification scheme for dynamic causal effects, further developed into DiD in Dube et al. (2023). Montiel Olea and Plagborg-Møller (2021) proposed a lag-augmentation, assuming a finite lag order for the DGP, while Xu (2023) put forth a uniform inference framework for local projections with increasing lag order. Most recently, Adamek et al. (2024) proposed adjustments for high-dimensional settings.

While the local projection framework has been extensively studied, the literature has also explored high-dimensional methods that can handle dependent data structures: see, for example, Masini et al. (2023) for a review of machine learning techniques in time series forecasting. Chernozhukov et al. (2021) explored the use of double/debiased LASSO with temporal and cross-sectional dependence. Babii et al. (2022) tackled both high-dimensional and mixed-frequency settings using sparse group-LASSO.

Beyond LASSO based approaches, the literature has also explored high-dimensional methods that do not rely on sparsity assumptions. A Prominent example is Principal Component (PC) type methods (Stock and Watson, 2002). Borrowing arguments from Feng et al. (2020), albeit in a different framework, model selection methods have the advantage of selecting the important variables based on their importance in the "cross-section" of variables, as opposed to other PC-type methods that select factors based on their ability to explain the time-series variation of responses.

In line of this literature, Ing (2020) proposed the OGA+HDAIC method, which allows for high-dimensioanl inference without the exact or L^p sparsity. While he derived error bounds in his paper, Cha et al. (2023) have worked on deriving inference on the low-dimensional parameter of interest in a cross-sectional setting. Additionally, Barnichon and Brownlees (2019) utilized splines to smooth the impulse responses, and Dinh et al. (2024) applied a random subspace method to local projections. While these papers propose practical tools for LP estimation, I further develop statistical foundation for the proposed estimator.

Building on these strands from the existing literature, my proposed approach integrates high-dimensional controls into LP framework, leveraging the OGA+HDAIC method (Ing, 2020). This allows me to account for the time series dependence in the data (NED) and derive error bounds as well as the asymptotic normality for the parameter of interest. Additionally, I am complementing the methodology proposed by Adamek et al. (2024) to address robust estimation in dense DGP settings.

1.2 Overview of the Method

I will first introduce a general estimation framework and then provide further details on the estimation procedures in the latter part of this section. Consider the following $h-$step ahead local projection regression model

$$y_{t+h} = \beta_h x_t + \delta'_{h,0} \boldsymbol{r}_t + \sum_{\ell=1}^{L} \delta'_{h,\ell} \boldsymbol{z}_{t-\ell} + u_{t,h}, \tag{1.2.1}$$

where y_{t+h} is the h-step ahead response variable, x_t the innovation, $\boldsymbol{r}_t$ are contemporaneous controls, and $\boldsymbol{z}_t$ are lagged controls. The index $t = 1,\ldots,\bar{T}$ is an index for the observations, $h = 1,\ldots,H_{\max}$ is an index for horizons, and $\ell = 1,\ldots,L$ is the number of lags included in the model. This is a general representation of local projections in the literature, as introduced in Plagborg-Møller and Wolf (2021). Following the spirit of Jordà (2005), I fix h and focus on each horizon of interest.

We are interested in the response of y_{t+h} with respect to a shock in x_t, and our parameter of interest β_h, is defined as

$$\beta_h = E\left[y_{t+h}|x_t = 1, \boldsymbol{r}_t, \{\boldsymbol{z}_{t-\ell}\}_{\ell=1}^{L}\right] - E\left[y_{t+h}|x_t = 0, \boldsymbol{r}_t, \{\boldsymbol{z}_{t-\ell}\}_{\ell=1}^{L}\right].$$

The parameter can be interpreted as impulse responses or treatment effects according to different identification assumptions. Below I introduce two applications which include the identification schemes.

1.2.1 Applications

Impulse response analysis with local projections

As was originally proposed by Jordà (2005), local projections are tools to estimate impulse responses with correctly specified controls. Plagborg-Møller and Wolf (2021) present a detailed review on how the LPs can be used to estimate the structural impulse responses, using their equaivalence results between the VAR and LPs. I will briefly introduce this approach following Section 3.1. of Plagborg-Møller and Wolf (2021).

Denote the data as $w_t = (\mathbf{r}_t', x_t, y_t, \mathbf{z}_t')'$ with the dimension of $\mathbf{r}_t$ as n_r. Consider the following Structural Vector Moving Average (SVMA) as the DGP

$$w_t = \mu + \Theta(L)\varepsilon_t, \quad \Theta(L) = \sum_{\ell=0}^{\infty} \Theta_\ell L^\ell,$$

and assume normality for the structural shocks: $\varepsilon_t \overset{i.i.d.}{\sim} N(0, I_{n_\varepsilon})$. Under some regularity conditions, the (i,j)th element of $\Theta_{i,j,h}$ is the impulse response of variable i-th element of w_t to the shock of j-th element of ε_t at horizon h. Consider the case where we are interested in the response of y_t with respect to a shock in $\varepsilon_{1,t}$. The parameter of interest is then $\theta_h := \Theta_{n_r+2,1,h}$ for $h = 0, 1, \ldots$.

Assuming that the structural shock can be recovered as a function of both current and past data, denoted as $\varepsilon_{1,t} \in \text{span}(\{w_\tau\}_{\tau \le t})$, the structural shock can be identified as a linear combination of the Wold forecast errors using an SVAR identification scheme. This can be expressed as $\tilde{\varepsilon}_{1,t} = b' e_t$, where b is obtained as a function of reduced-form VAR parameters depending on identification schemes. The LP approach involves using reduced-form LP parameters instead of VAR counterparts to generate structural impulse responses. Using this method, the population estimand

would be equivalent. For a detailed explanation with different identification schemes, please refer to Plagborg-Møller and Wolf (2021).

In cases where the model is non-invertible, the straightforward approach would be to use instruments. I cannot efficiently deal with such cases since this paper does not provide a theory for the instrumental variable LP. However, as mentioned in Känzig (2021), non-invertibility is essentially the omitted variable bias problem. In that sense, I can argue that high-dimensional controls still help because the model spans more information.

The above results leverage the VAR-LP equivalence with infinite lags. In practical applications, we rely on a finite lag approximation. The use of finite lags introduces approximation errors, which must be carefully considered in empirical applications. A detailed discussion on the implications of using finite lags is provided in Appendix A.3.

In the empirical analysis in Section 1.5.2, I apply the impulse response analysis scheme outlined above to study the effects of subjective beliefs on some key economic aggregates. I consider several model specifications with an increasing number of lags, revisiting the issues of invertibility and finite lag approximations. This allows me to assess the robustness of the results to potential non-invertibility and to explore the trade-offs associated with different lag lengths in capturing the dynamic effects of belief shocks.

Local projections approach to difference-in-differences with high-dimensional controls

A recent paper by Dube et al. (2023) proposes a local projections approach to DiD event studies. Following their identification schemes, the parameter of interest β_h can be interpreted as a DiD estimator. In the following, I will briefly introduce their identification schemes with application to high-dimensional controls.

Consider a potential outcomes framework (Rubin (1974)) in a panel setting with a binary treatment. Denote $y_{it}(0)$ and $y_{it}(p)$ as the potential outcomes with respect to the control and treatment at time $p \neq \infty$ status. Units are divided into groups, $g \in \{0,1,\ldots,G\}$, where each group g shares the same treatment timing, p_g. Denote the group $g = 0$ as the never treated group with $p_0 = \infty$. Treatment is absorbing, meaning that once a unit receives treatment, it remains treated in all subsequent periods.

The group-specific average treatment effect on the treated (ATT) at horizon h for group g, which starts treatment at time p, is given by:

$$\tau_g(h) = E[y_{i,p+h}(p) - y_{i,p+h}(0)|p_i = p].$$

The DiD approach hinges on two main assumptions: parallel trends and no anticipation. The no anticipation assumption implies that for any time t before the treatment time p, the expected difference in potential outcomes is ($E[y_{it}(p) - y_{it}(0)] = 0$). The parallel trend assumption states that $E[y_{i,t}(0) - y_{i,1}(0)|p_i = p] = E[y_{i,t}(0) - y_{i,1}(0)]$, for all $t \geq 2$

and for all $p \in \{1, \ldots, T, \infty\}$. Also assume a simple structure to the never treated outcome to be

$$E[y_{i,t}(0)] = \alpha_i + \delta_t.$$

Now consider the following estimating equation.

$$y_{i,t+h} - y_{i,t-1} = \beta_h^{\text{LP-DiD}} \Delta D_{it} + \delta_t^h + e_{it}^h,$$

where δ_t^h are time specific controls and e_{it}^h the error terms. The LP-DiD parameter then identifies

$$\begin{aligned}
E[\beta_h^{\text{LP-DiD}}] =& E[\Delta_h y_{it} | t, \Delta D_{it} = 1] - E[\Delta_h y_{it} | t, \Delta D_{it} = 0] \\
=& E\Big[\sum_{g=1}^{G} (\tau_g(h) \times \mathbf{1}\{t = p_g\}) \Big] \\
& - E\Big[\sum_{g=1}^{G} \Big[\sum_{j=1}^{\infty} (\tau_g(h+j) - \tau_g(j-1)) \times \Delta D_{i,t-j} \times \mathbf{1}\{t = p_g + j\} \Big] \Big] \\
& - E\Big[\sum_{g=1}^{G} \Big[\sum_{j=1}^{\infty} \tau_g(h-j) \times \Delta D_{i,t+j} \times \mathbf{1}\{t = p_g - j\} \Big] \Big],
\end{aligned}$$

From this expression, we can clearly see the two bias terms. The contribution of the referenced paper is to restrict the sample such that both bias terms become zero.

The two key terms contributing to the biases $\Delta D_{i,t+j}$ for $j \leq h$ and $\Delta D_{i,t-j}$ for $1 \geq j$. These terms go to zero if $\Delta D_{i,t-j} = 0$ for $-h < j < \infty$, which simplifies to $D_{t+h} = 0$ under the assumption of absorbing treatment. This is where the sample restriction plays a critical role.

By restricting to the sample that are either

$$\begin{cases} \text{newly treated} & \Delta D_{it} = 1, \\ \text{or clean control} & D_{i,t+h} = 0, \end{cases}$$

the LP-DiD parameter identifies

$$\begin{aligned}
E[\beta_h^{\text{LP-DiD}}] =& E[\Delta_h y_{it} | t, \Delta D_{it} = 1] - E[\Delta_h y_{it} | t, \Delta D_{it} = 0, D_{i,t+h=0}] \\
=& E[\sum_{g=1}^{G} (\tau_g(h) \times \mathbf{1}\{t = p_g\})],
\end{aligned}$$

where it provides a convex combination of all group-specific effects $\tau_g(h)$ by removing previously treated observations

and observations treated between $t+1$ and $t+h$ from the control group.

Now, consider the scenario where the researcher believes the parallel trend holds only after controling for a vector of covariates, $\mathbf{w}_{it}$. Including controls becomes necessary for the identification purpose, resulting in the following conditional parallel trend assumption:

$$E[y_{i,t}(0) - y_{i,1}(0)|p_i = p, \mathbf{w}_{it}] = E[y_{i,t}(0) - y_{i,1}(0)|\mathbf{w}_{it}].$$

The conditional parallel trends assumption becomes crucial when allowing for variations in outcome dynamics across different groups, provided that these differences can be fully accounted for by the covariates. This makes the assumption far more realistic than the conventional parallel trends assumption, especially in settings where untreated outcomes are not expected to evolve similarly across treatment groups (Heckman et al., 1997; Abadie, 2005; Callaway and Sant'Anna, 2021).

Adopting my estimation framework introduces a new layer to this argument: by accounting for high-dimensional controls, the conditional parallel trends assumption becomes even more realistic and robust. By leveraging these controls, we ensure that any remaining variation in trends across groups is largely orthogonal to the treatment, further strengthening the identification strategy. This approach makes the conditional parallel trends assumption not only more plausible but also more applicable to modern datasets, where a wide range of observable factors can be accounted for in high-dimensional frameworks.

Then the estimating equation becomes

$$y_{i,t+h} - y_{i,t-1} = \beta_h^{\text{LP-DiD}^{\mathbf{w}}} \Delta D_{i,t} + \mathbf{w}_{i,t}' \gamma^h + \delta_t^h + e_{i,t}^h,$$

where $\mathbf{w}_{i,t}$ can include lagged terms of the dependent variable and other controls. By restricting the sample to be newly treated group and the clean control group as before, the LP-DiD$^{\mathbf{w}}$ parameter identifies

$$\begin{aligned} E[\beta_h^{\text{LP-DiD}^{\mathbf{w}}}|\mathbf{w}_{it}] =& E[\Delta_h y_{it}|t, \Delta D_{it} = 1, \mathbf{w}_{it}] - E[\Delta_h y_{it}|t, \Delta D_{it} = 0, D_{i,t+h=0}, \mathbf{w}_{it}] \\ =& E[\sum_{g=1}^{G} (\tau_g(h) \times \mathbf{1}\{t = p_g\})]. \end{aligned}$$

This framework enables us to focus on the portion of outcome evolution that is not explained by the covariates, isolating the treatment effect more effectively. I will revisit this identification framework and its implications with the empirical application in Section 1.5.1.

1.2.2 Estimation Procedures

For notation simplicity, stack the covariates except for x_t into $\boldsymbol{w}_t = (\boldsymbol{r}_t, \boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-L})$ and write

$$y_{t+h} = \beta_h x_t + \beta'_{-h} \boldsymbol{w}_t + u_{t,h}, \tag{1.2.2}$$

where $\beta_{-h} = (\delta'_{h,0}, \delta'_{h,1}, \ldots, \delta'_{h,L})'$. To address the bias that arises from excluding higher-order lag controls, the dimension of $\boldsymbol{w}_t$ expands as the lag order, L, increases. Turning to the causal identification scheme of Angrist et al. (2018), it is intuitive to add a large set of covariates as controls to account for unobserved counfounding variables, which also contributes to increasing the dimension of $\boldsymbol{w}_t$.

A straightforward approach to (1.2.2) with high dimensionality would be to apply a regularization method such as the LASSO. However, it is now well known that these methods come with an associated cost known as regularization bias. One way to address this challenge is a debiasing strategy using node-wise regression, regressing each covariate on the remaining covariates, as illustrated by Van de Geer et al. (2014) and Zhang and Zhang (2014). Using the estimates from the node-wise regressions, they construct a remedy for the bias term and control for it. An alternative approach is a double selection method that establishes an orthogonality condition in a similar spirit to Belloni et al. (2013a). This method is analogous to the debiasing process of the LASSO, as noted in Semenova et al. (2023) and Chernozhukov et al. (2021). In the context of (1.2.2), the main idea of both approaches is to incorporate another set of regressions of the covariates to control for the regularization bias, either by debiasing or by using the covariates selected in both regression models.

In this paper, I consider the double selection method with a model selection method OGA+HDAIC by Ing (2020), where it consists of two steps. OGA first orders the covariates $\boldsymbol{w}_t$ by their explanatory power, and then we select the number of covariates that minimizes the high-dimensional AIC criteria. The selections come from the following regression models,

$$y_{t+h} = \lambda'_h \boldsymbol{w}_t + e_{t,h}, \tag{1.2.3}$$

$$x_t = \gamma'_h \boldsymbol{w}_t + v_{t,h}. \tag{1.2.4}$$

Although not necessary in (1.2.4), I used the subscript h for consistency across both equations. The concept involves applying OGA+HDAIC on both equations (1.2.3) and (1.2.4). To remove the regularization bias, I use the union of two selected covariates as $\boldsymbol{w}_t$ to finally run the local projection regression in (1.2.2). For a clear presentation of the OGA+HDAIC procedure, I fix some notations below.

Notations. Let t be an index for observations and j for covariates, so that $\boldsymbol{w}_t = (w_{t,j})_{j=1}^{p}$ and $\boldsymbol{w}_j = (w_{t,j})_{t=1}^{T}$, where $p := dim(\boldsymbol{w}_t)$ and $T := \bar{T} - h - L$ is the effective sample size. Let $[p] = 1, \ldots, p$ be the set of all covariate indices. Let J be a set of covariate indices and let the subset of the covariates be $\boldsymbol{w}_t(J) := (w_{t,j})_{j \in J}$. Stack all $t = 1, \ldots, T$ observations of $\boldsymbol{w}_t(J)$ and denote $W(J) := (\boldsymbol{w}_t(J))_{t=1}^{T}$. Similarly define $\boldsymbol{y}_h = (y_{t+h})_{t=1}^{T}$. Denote the projection matrix using $W(J)$ as $P_J := W(J)(W(J)'W(J))^{-1}W(J)'$.

First, consider the OGA part of y_{t+h} on $\boldsymbol{w}_t$. The ordering of the covariates requires the following definition, which indicates the explanatory power of the covariate $\boldsymbol{w}_i$.

$$\mu_i(\emptyset) = \frac{\boldsymbol{w}_i' \boldsymbol{y}_h}{\sqrt{T}\,\|\boldsymbol{w}_i\|} = \frac{\frac{1}{T}\boldsymbol{w}_i' \boldsymbol{y}_h}{\sqrt{\frac{1}{T}\boldsymbol{w}_i'\boldsymbol{w}_i}}, \tag{1.2.5}$$

As it looks, it works as a scaled version of a single regressor regression of $\boldsymbol{y}_h$ on each regressor $\boldsymbol{w}_i$. Since it is scaled by the square root of the L_2 norm of each regressor, the scale of $\boldsymbol{w}_i$ does not affect the magnitude of $\mu_i(J)$. To choose the one with the most explanatory power, we start with the setting $J = \emptyset$ and choose the one with the largest value of $|\mu_i(\emptyset)|$. Let this covariate index as

$$\widehat{j_1} := \arg\max_{i \in [p]} |\mu_i(\emptyset)|,$$

and define the first chosen set as $\widehat{J_1} := \{\widehat{j_1}\}$. To select the second order covariate, we use the residuals from the first step, using $\mu_i(\widehat{J_1})$ to measure explanatory power. We choose the second order covariate from the remaining covariates, $\widehat{j_2} := \arg\max_{i \in [p] \setminus \widehat{j_1}} \left|\mu_i(\widehat{J_1})\right|$, then update the chosen set of covariates, $\widehat{J_2} = \widehat{J_1} \cup \{\widehat{j_2}\}$. Generalizing to the $m-$th order covariate, suppose we have the previously chosen set of covariates, $\widehat{J}_{m-1}$. Compute the following coefficient for all $i \in [p] \setminus \widehat{J}_{m-1}$, where

$$\mu_i(\widehat{J}_{m-1}) = \frac{\boldsymbol{w}_i'(I - P_{\widehat{J}_{m-1}})\boldsymbol{y}_h}{\sqrt{T}\,\|\boldsymbol{w}_i\|}, \tag{1.2.6}$$

and select the one with the largest absolute value of $\mu_i(\widehat{J}_{m-1})$,

$$\widehat{j}_m = \arg \max_{i \in [p] \setminus \widehat{J}_{m-1}} \left|\mu_i(\widehat{J}_{m-1})\right|, \tag{1.2.7}$$

and update the chosen set of covariates, $\widehat{J}_m = \widehat{J}_{m-1} \cup \{\widehat{j}_m\}$. Repeating this procedure orders the covariates in descending order of their explanatory power, conditioning on the previously chosen set of covariates. Now that the covariates are ordered, the remaining task is to select the threshold for the number of covariates. The information criterion we use is

HDAIC[2], deinfed as

$$\text{HDAIC}(J) = \left(1 + \frac{C^*|J|\log p}{T}\right)\widehat{\sigma}_J^2, \tag{1.2.8}$$

where $\widehat{\sigma}_J^2 = \boldsymbol{y}_h'(I - P_J)\boldsymbol{y}_h/T$. The number of covariates to be included in the model is the one which minimizes the information criteria,

$$\widehat{m} = \arg\min_{1\le m\le M_T^*} \text{HDAIC}(\widehat{J}_m). \tag{1.2.9}$$

Denote the chosen set of covariates as $\widehat{J}^{[1]} := \widehat{J}_{\widehat{m}}$. Next, repeat the OGA+HDAIC for (1.2.4) and obtain the set of covariates $\widehat{J}^{[2]}$. Finally, denote the union set as $\widetilde{J} = \widehat{J}^{[1]} \cup \widehat{J}^{[2]}$ and run the local projection regression (1.2.2) with the selected covariates:

$$y_{t+h} = \beta_h x_t + \beta_{-h}'\boldsymbol{w}_t(\widetilde{J}) + u_{t,h}.$$

The least squares estimator for β_h in this final model is our proposed estimator. For the variance estimator, define $\psi_t = v_{t,h}e_{t,h}$ and $\tau^2 = E[\sum_{t=1}^T v_{t,h}^2/T]$. The variance estimator is then defined as

$$\widehat{\sigma}_h^2 = \frac{1}{\widehat{\tau}^2}\widehat{\Omega}, \tag{1.2.10}$$

$$\widehat{\Omega} = \sum_{\ell=-(K-1)}^{K-1}\left(1 - \frac{\ell}{K}\right)\frac{1}{T-\ell}\sum_{t=\ell+1}^{T}\widehat{\psi}_t\widehat{\psi}_{t-\ell}, \tag{1.2.11}$$

where $\widehat{\Omega}$ is the Newey-West estimator with a bandwidth parameter K, which is assumed to be increasing with respect to increasing sample size. We borrow arguments from Andrews (1991) for the choice of K. $\widehat{\psi}_t$ is the sample analogue of ψ_t, where $\widehat{v}_{t,h}$ and $\widehat{e}_{t,h}$ are the residuals from (1.2.3) and (1.2.4).

The proposed algorithm is detailed in Algorithm 1 after a remark on tuning parameters.

Remark. Unlike the OGA procedure, the HDAIC procedure has two unknown parameters, M_T^* and C^*. A detailed description of both definitions can be found in Appendix A.3. M_T^* is a parameter concerning the maximum number of covariates to include in the model and increases with $(T/\log p)$, as defined in (A.3.1). C^*, specified in (A.3.2), adjusts the penalty for dimensionality and is assumed to be greater than certain constants. While C^* could be viewed as a tuning parameter, it's more appropriately considered a hyperparameter, analogous to the hyperparameter of 2 in traditional

[2]Note that it is called HD because it takes into account the penalization on $\log p/T$. This notion, unlike the traditional AIC vs. BIC framework, takes into account the size of the entire feature space p relative to the available data T. This holistic approach to model complexity resonates in the high-dimensional setting, where the effect of the penalty is amplified as $\log p/T$ diverges with increasing dimensionality.

AIC. Aside from the plug-in option, there is a data-driven way to choose the value of C^* by setting candidates, e.g. $\bar{C} = \{1.6, 1.8, 2, 2.2, 2.4\}$, and choosing the one that yields the smallest prediction errors in each regression in (1.2.3) and (1.2.4).

Algorithm 1 Double-OGA+HDAIC

1. Compute $\mu_i(\emptyset)$ in (1.2.5) for all $i \in [p]$ and select the covariate with the largest $|\mu_i(\emptyset)|$. Denote the index of covariate as $\widehat{j}_1$ and define $\widehat{J}_1 = \widehat{j}_1$.
2. Given $\widehat{J}_1$, compute $\mu_i(\widehat{J}_1)$ for all $i \in [p] \setminus \widehat{J}_1$ and select the covariate with the largest $|\mu_i(\widehat{J}_1)|$. Denote the index of covariate as $\widehat{j}_2$ and define $\widehat{J}_2 = \widehat{J}_1 \cup \widehat{j}_2$.
3. For $m > 2$, compute $\mu_i(\widehat{J}_{m-1})$ for all $i \in [p] \setminus \widehat{J}_{m-1}$ and select the covariate with the largest $|\mu_i(\widehat{J}_{m-1})|$. Denote the index of covariate as $\widehat{j}_m$ and define $\widehat{J}_m = \widehat{J}_{m-1} \cup \widehat{j}_m$.
4. Compute HDAIC in (1.2.8) and select m that minimizes HDAIC$(\widehat{J}_m)$. Denote it as $\widehat{m}^y$ and define $\widehat{J}_y := \widehat{J}_{\widehat{m}^y}$.
5. Run steps 1–4 by replacing y_{t+h} with x_t. Obtain $\widehat{J}_x$ and the residuals $\widehat{v}_{t,h}$.
6. Run OLS of y_{t+h} on $[x_t : \boldsymbol{w}_t(\widetilde{J})]$, where $\widetilde{J} = \widehat{J}_x \cup \widehat{J}_y$. Obtain the final estimates $\widehat{\beta}_h$ and the residuals $\widehat{u}_{t,h} = y_{t+h} - \widehat{\beta}_h x_t + \widehat{\beta}'_{-h} \boldsymbol{w}_t(\widetilde{J})$.
7. Calculate the variance estimator $\hat{\sigma}_h^2$ defined in (1.2.10).

With the above algorithm, one can construct an $100(1-\alpha)\%$ confidence interval for $h-$step ahead impulse response estimator of the following form

$$[\widehat{\beta}_h - z_\alpha \hat{\sigma}_h / \sqrt{T},\ \widehat{\beta}_h + z_\alpha \hat{\sigma}_h / \sqrt{T}],$$

where $z_\alpha = \Phi^{-1}(1-\alpha/2)$ is the $(1-\alpha/2)$ quantile of the standard normal distribution.

1.3 Illustrative Simulation Study

This section highlights the advantages of the proposed method over the commonly used LASSO approach, especially when dealing with unknown sparsity assumptions and the degree of persistence. I present a comparative analysis of the proposed estimator, the standard local projection estimator, and the debiased LASSO[3], as described in Adamek et al. (2024). For the DGP, I adapt from Montiel Olea and Plagborg-Møller (2021). Consider the following DGP

$$\begin{aligned} y_{1,t} =& \rho y_{1,t-1} + u_{1,t}, \\ \boldsymbol{y}_t =& B_1 \boldsymbol{y}_{t-1} + \cdots + B_{12} \boldsymbol{y}_{t-12} + \boldsymbol{u}_t, \quad \boldsymbol{u}_t \sim N(0, \Sigma), \quad \Sigma_{i,j} = \tau^{|i-j|}, \end{aligned}$$

[3]For the debiased lasso estimation, I use the R package *desla* provided by Adamek et al. (2024).

where $\boldsymbol{y}_t \in \mathbb{R}^n$. We are interested in estimating the reduced-form impulse response of $y_{2,t}$ with respect to the innovation $u_{1,t}$. I evaluate the finite sample properties with 95% coverage probabilities and the median widths of the confidence intervals over 1000 iterations.

For parameter settings, I set τ to be 0.3, $n = 10$, and I consider the sample size of $T = 300$. For estimation settings, I adopt the lag-augmentation and set the number of lags included in the model as $L = 21$, while the lags included in the DGP is 12. To evaluate performance over longer horizons, I estimate the reduced-form impulse responses for horizons from 1 to 60. The simulations explore different levels of persistence and sparsity by varying ρ and B_ℓ for $\ell = 1, \ldots, 12$. Two levels of persistence are considered: $\rho = 0.5$ and $\rho = 0.95$. For sparsity, I define the values of B_ℓ using a vector $a \in \mathbb{R}^{n-1}$ of different magnitudes. I defined the even elements of each row $B\ell_{(j,)}$ as polynomials of a_j with alternating signs. Different values of a are considered to generate different levels of sparsity, where I set $a^s = (0.4, -0.36, ..., , -0.094, 0.05)'$ for a sparse scenario and $a^d = (0.8, -0.73, ..., , -0.28, 0.2)'$ for a dense scenario.

The resulting coefficients of the corresponding local projection equations are depicted in Figure 1.1. The coefficients are ordered in descending order of their absolute magnitude. The left panel displays the regression coefficients from the local projection equation using the sparse vector a_s, while the right panel shows the coefficients using the dense vector a_d. The top panels correspond to data with lower persistence ($\rho = 0.5$), and the bottom panels correspond to data with higher persistence ($\rho = 0.95$). Each line represents the regression coefficients from the h-step ahead local projection equations, with horizons spanning from 3 to 59. Note that h ranges from 1 to 60, and I have selected 9 points within this range for illustration. In the left panel, the coefficients are sparse, with fewer than 10 coefficients being nonzero. In contrast, the right panel shows a dense set of coefficients; although the absolute magnitudes decay, a larger number of coefficients remain nonzero. The more persistent the data, the more amplified the largest ordered coefficients become.

Figures 1.2 and 1.3 show the simulation results for different levels of persistence. In the left panel of Figure 1.2, the model is sparse and the data are less persistent, making it easier for all estimation methods to perform well. Most methods achieve satisfactory coverage probabilities, except for debiased LASSO, which performs slightly worse. Notably, while the conventional LP achieves 95% coverage rates, it does so at the cost of significantly larger confidence interval widths compared to the high-dimensional methods. Additionally, the conventional LP method shows slight underperformance for the last 5 horizons, despite its large standard errors. This can be attributed to the sample size of $T = 300$ and the fact that the number of covariates included in the model is about 70% of the sample size. As a result, the standard erorr gets larger when fitting all covariates, and this issue worsens with increasing horizons due to the reduced effective sample size.

We can observe a similar pattern for the dense scenario in the right panel, except for a sharp failure in the initial horizons using the debiased LASSO method. The proposed method achieves the coverage rates with small standard

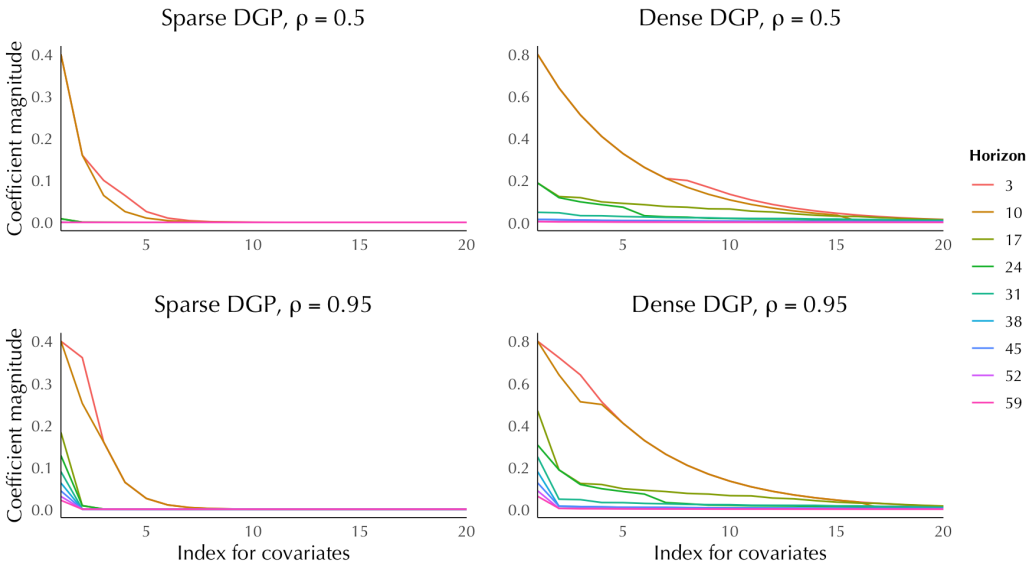


Figure 1.1: Magnitude of coefficients in the estimating equations

Notes. Each line represents h-step ahead local projection coefficients for horizons from 3 to 59. The coefficients are ordered in their absolute magnitudes. The left panel shows coefficients using the sparse vector a_s, while the right panel uses the dense vector a_d. The top panels are for lower persistence ($\rho = 0.5$) and the bottom panels for higher persistence ($\rho = 0.95$).

errors across the different specifications. The huge gain in efficiency comes from the model selection, and it still achieves efficiency in the longer horizons. This is true for both high-dimensional methods.

Figure 1.3 presents the results for the higher persistence scenario. There is a huge failure of the debiased LASSO for either sparse or dense scenarios, especially when it gets to longer horizons. The reason for this failure can be inferred from the median width plot, where LASSO rules out too many regressors. The conventional LP performs similarly to the less persistent case, where it achieves the coverage rates but with larger standard errors. Overall, the proposed method maintains the coverage rates at a much lower cost compared to the conventional LP.

The simulation results illustrate that the proposed method maintains desirable coverage rates with more stable and narrower confidence intervals across different specifications. While the conventional LP also achieves the coverage probabilities, it suffers from larger standard errors, especially as the horizon increases. As the debiased LASSO consistently gives the smallest standard errors throughout all scenarios, there might be an appropriate DGP where it outperforms the proposed method at least in terms of efficiency, but less likely in dense scenarios, as the performance of debiased LASSO tends to decline with denser DGPs. Given the uncertainties regarding the underlying sparsity and persistence of the data, the proposed method provides a robust and reliable alternative that performs well across various conditions.

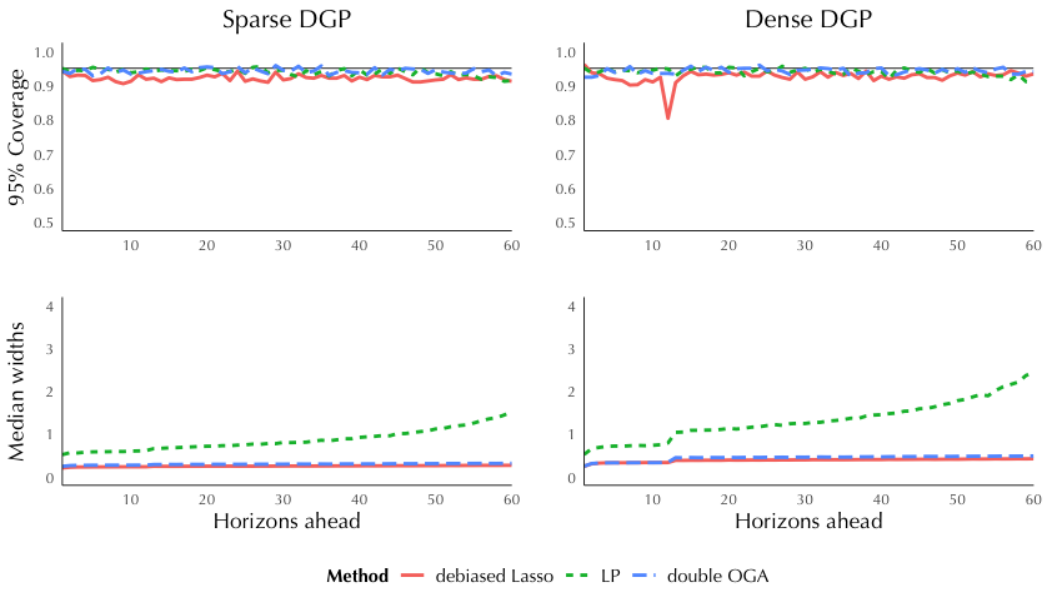


Figure 1.2: 95% Coverage probabilities and median widths, less persistent case

Notes. The figure compares 95% coverage probabilities (top panels) and median widths (bottom panels) over 60 forecast horizons for sparse and dense DGPs, generated with a_s (left panels) and a_d (right panels) in the less persistent case ($\rho = 0.5$). The results are based on 1,000 iterations. The red solid line represents the debiased Lasso method, the green dotted line indicates conventional LP, and the blue dashed line corresponds to the proposed method.

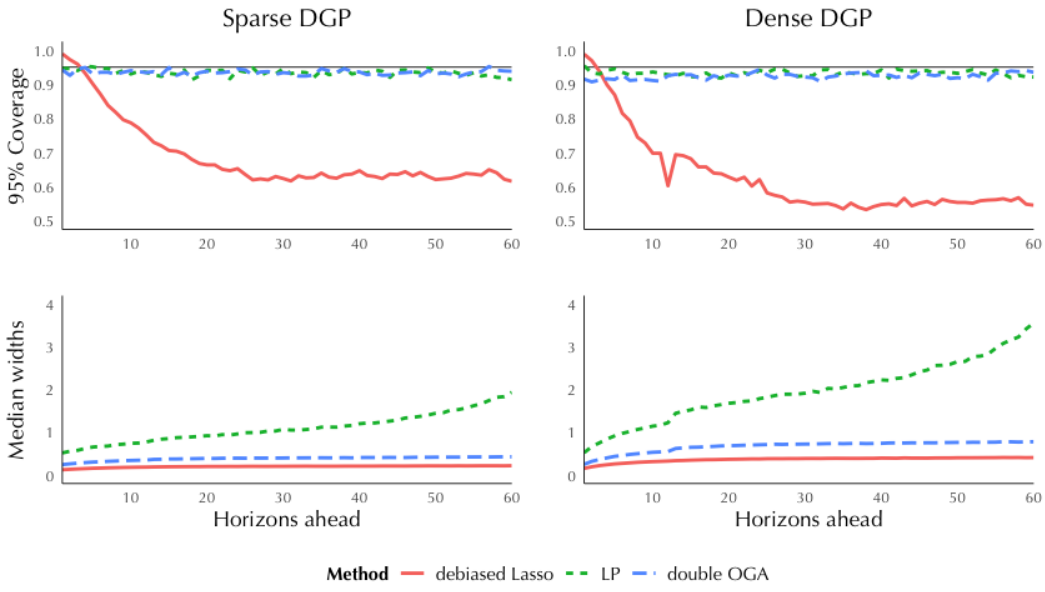


Figure 1.3: 95% Coverage probabilities and median widths, more persistent case

Notes. The figure compares 95% coverage probabilities (top panels) and median widths (bottom panels) over 60 forecast horizons for sparse and dense DGPs, generated with a_s (left panels) and a_d (right panels) in the less persistent case ($\rho = 0.95$). The results are based on 1,000 iterations. The red solid line represents the debiased Lasso method, the green dotted line indicates conventional LP, and the blue dashed line corresponds to the proposed method.

1.4 Theory

1.4.1 Preliminaries

In this subsection, I introduce the definitions used throughout the theory. Most of the definitions and explanations are taken from Davidson (2021). I will use these definitions in the context of the main text in the following subsection.

Consider a probability space $(\Omega, \mathscr{F}, P)$. With time series data, time flows in one direction: The past is known while the future is unknown. It is hence important to condition on previous information set in the context of time series analysis. The accumulation of information is represented by an increasing sequence of sub $\sigma-$fields, $\{\mathscr{F}_t\}_{-\infty}^{\infty}$, where $\mathscr{F}_s \subseteq \mathscr{F}_t$ for $s \leq t$. With the uncertainty of the future, the natural next step is to take expectations. If X_t is $\mathscr{F}_t-$measurable for each t, $\{X_t, \mathscr{F}_t\}_{-\infty}^{\infty}$ is called an adapted sequence, and $E[X_t|\mathscr{F}_{t-1}]$ is defined. Also, if $E[X_{t+s}|\mathscr{F}_t] = X_t$ a.s. for all $s \geq 0$ under adaptation, the sequence is identified by its history alone, and $\{X_t\}_{-\infty}^{\infty}$ is called a causal stochastic sequence. I start by introducing the most widely used dependence concept, the martingale difference (m.d.) sequences.

Definition 1 (Martingale difference sequence, Davidson (2021))**.** The process $\{X_t\}_{-\infty}^{\infty}$ is a martingale difference sequence if

$$E[X_t] < \infty$$
$$E[X_t|\mathscr{F}_{-\infty}^{t-1}] = 0 \quad a.s.$$

As is clear from the definition, m.d. assumes one-step-ahead unpredictability. It follows that it is uncorrelated with any measurable function of its lagged values, and thus behaves like independent sequences in classical limit results. As shown in Davidson (2021), many limit theorems that hold under independence also hold under the m.d. assumption with few additional assumptions about the marginal distributions, making it a preferred dependence assumption for econometricians. To extend our understanding beyond one-step-ahead unpredictability, the next definition introduces the concept of asymptotic unpredictability.

Definition 2 (Mixingale, Davidson (2021))**.** For $q \geq 1$, the sequence $\{X_t, \mathscr{F}_t\}_{-\infty}^{\infty}$ is an L_q-mixingale if for a sequence of non-negative constants $\{c_t\}_{-\infty}^{\infty}$ and $\{\rho_m\}_0^{\infty}$ such that $\rho_m \to 0$ as $m \to \infty$,

$$\left(E\left[\left|E[X_t|\mathscr{F}_{-\infty}^{t-m}]\right|^q\right]\right)^{1/q} \leq c_t \rho_m, \tag{1.4.1}$$
$$\left(E\left[\left|X_t - E[X_t|\mathscr{F}_{-\infty}^{t+m}]\right|^q\right]\right)^{1/q} \leq c_t \rho_{m+1}, \tag{1.4.2}$$

hold for all t and $m \geq 0$.

A mixingale generalizes a m.d. as a special case where $\rho_m = 0$ for all $m > 0$. The equations show a diminishing

effect of past information on the present, while suggesting eventual complete knowledge of the present in the distant future. Note that if the sequence is adapted, then $E[X_t|F_{-\infty}^{t+m}] = X_t$ for all $m \geq 0$ and (1.4.2) is satisfied. Just as m.d.s behave like independent sequences in limit theorems, mixingales behave like mixing sequences, which implies asymptotic independence with the following definitions.

Definition 3 ($\alpha-$mixing coefficients, Hansen (1991))**.** The $\alpha-$mixing coefficients of a sequence $\{X_t\}_{-\infty}^{\infty}$ are given by

$$\alpha_m = \sup_j \sup_{\{F \in \mathscr{F}_{-\infty}^{j}, G \in \mathscr{F}_{j+m}^{\infty}\}} |P(G \cap F) - P(G)P(F)|,$$

and the process X_t is said to be $\alpha-$mixing if $\lim_{m\to\infty} \alpha_m = 0$.

The $\alpha-$mixing coefficient α_m measures the dependence between the sequences separated by a lag of m. If a process X_t is $\alpha-$mixing, the dependence diminishes as the lag increases. Mixing conditions often play a crucial role in further establishing asymptotic properties. Although the mixingale condition offers several advantageous properties, an important limitation arises: even if a function of mixingales is independent, it loses its mixing property with an infinite number of lags (Davidson (2021), Chapter 18). The following definition introduces a mapping from a mixing process to a random sequence. This transformation enables the sequence to inherit some desired mixingale properties.

Definition 4 (Near Epoch Dependence (NED), Davidson (2021))**.** For a stochastic process $\{V_t\}_{-\infty}^{\infty}$, let $\mathscr{F}_{t-m}^{t} = \sigma\{V_{t-m}, \ldots, V_t\}$, such that $\{\mathscr{F}_{t-m}^{t}\}_{m=0}^{\infty}$ is an increasing sequence of $\sigma-$fields. If for $q > 0$ an adapted sequence $\{X_t\}_{-\infty}^{\infty}$ satisfies

$$\left(E\left[\left|X_t - E(X_t|\mathscr{F}_{t-m}^{t})\right|^q\right]\right)^{1/q} \leq d_t \zeta_m, \tag{1.4.3}$$

where $\zeta_m \to 0$ as $m \to \infty$ and $\{d_t\}_{-\infty}^{\infty}$ is a sequence of positive constants, X_t is said to be near-epoch dependent in L_q-norm (L_q-NED) on $\{V_t\}_{-\infty}^{\infty}$.

The NED condition is the main dependence assumption I use on the data, following Adamek et al. (2023) and Adamek et al. (2024). In the rest of the subsection, I introduce some useful lemmas using the aforementioned definitions.

The following lemma gives a concentration bound without a specific assumption on the dependence structure, which will be used both in Theorems 1.4.1 and 1.4.3.

Lemma 1.4.1 (Triplex inequality, Jiang (2009))**.** *Let $\{X_t\}_{-\infty}^{\infty}$ be a causal stochastic sequence and $\{\mathscr{F}_{t-m}^{t}\}_{m=0}^{\infty}$ be an increasing sequence of $\sigma-$fields. Let $\{X_t\}$ be $\mathscr{F}_t-$measurable for each t. Then for any $\varepsilon, M > 0$ and positive integers*

m,

$$
\begin{aligned}
P\left(\left|\frac{1}{T}\sum_{t=1}^{T}X_t - EX_t\right| > \varepsilon\right) \leq\ & 2m\exp\left(-\frac{T\varepsilon^2}{288m^2M^2}\right) \\
& + (6/\varepsilon)\frac{1}{T}\sum_{t=1}^{T}E[E[X_t|\mathscr{F}_{t-m}] - EX_t] \\
& + (15/\varepsilon)\frac{1}{T}\sum_{t=1}^{T}E\left[|X_t|\,\mathbb{1}\{|X_t| > M\}\right]
\end{aligned}
\tag{1.4.4}
$$

This inequality is called the triplex inequality because it has three components. The first element is a Bernstein-type bound, the second deals with dependence, and the last component is on the tail. This bound is a central theory used in the proofs, where I use the mixingale property inherited by the NED assumption to simplify the dependence and the tail bounds. The following lemma provides the simplified bounds by imposing the mixingale assumptions on $\{X_t\}_{-\infty}^{\infty}$.

Lemma 1.4.2 (Triplex inequality for mixingales)**.** *Suppose the assumptions in Lemma 1.4.1 hold. Further assume that the moment generating function for X_t exists and $\{X_t\}_{-\infty}^{\infty}$ is an $L_{\bar{q}}-$bounded L_q-mixingale with constants c_t and $\rho_m \to 0$ for some $1 \leq \bar{q} \leq q$, so that $E[|X_t|^{\bar{q}}] \leq C$.Then for any $\varepsilon, M > 0$ and positive integers m,*

$$
\begin{aligned}
P\left(\left|\frac{1}{T}\sum_{t=1}^{T}X_t - EX_t\right| > \varepsilon\right) \leq\ & 2m\exp\left(-\frac{T\varepsilon^2}{288m^2M^2}\right) \\
& + \frac{6\bar{c}_T}{\varepsilon}\rho_m + \frac{15}{\varepsilon}C\exp(-M(\bar{q}-1)),
\end{aligned}
\tag{1.4.5}
$$

where $\bar{c}_T := \sum_{t=1}^{T} c_t/T$.

Proof. The proof can be found in Appendix A.2.0.1. □

1.4.2 Assumptions and Lemmas

The definitions of the variables and the parameters come from the baseline models (1.2.2) – (1.2.4). I start by introducing the assumptions on the data generating processes.

Assumption 1 (NED)**.** *Denote $\varepsilon_{t,h}$ as a genereric notation for the error terms $u_{t,h}, e_{t,h}$, and $v_{t,h}$. There exist some constants $\bar{q} > q > 2$, and $b \geq \max\{1, (\bar{q}/q-1)/(\bar{q}-2)\}$ such that*

(a) *$(x_t, \boldsymbol{w}_t, u_{t,h}, e_{t,h})$ are zero-mean causal stochastic process with $E[x_t u_{t,h}] = E[x_t e_{t,h}] = 0$ and $E[\boldsymbol{w}_t u_{t,h}] = E[\boldsymbol{w}_t e_{t,h}] = \mathbf{0}$.*

(b) *$\max_{1\leq j\leq p} E[|w_{t,j}|^{2\bar{q}}] \leq C$ and $\max_{1\leq h\leq H_{\max}} E[|\varepsilon_{t,h}|^{2\bar{q}}] \leq C$.*

(c) *Denote the moment generating function of X as* $M_X(t)$. *Assume* $M_X(t) < \infty$ *for* $X = \{w_{t,j}\varepsilon_{t,h}, w_{t,j}w_{t,k}\}$ *for all* $j,k = 1,\dots,p, j \neq k$.

(d) *Let* $\{\mathrm{Y}_{T,t}\}$ *denote a triangular array that is* $\alpha-$*mixing of size* $-b/(1/q-1/\bar{q})$ *with* $\sigma-$*field* $\mathscr{F}_t^{\mathrm{Y}} := \sigma\{\mathrm{Y}_{T,t}, \mathrm{Y}_{T,t-1}, \dots\}$ *such that* $(x_t, \boldsymbol{w}_t, u_{t,h}, e_{t,h})$ *is* $\mathscr{F}_t^{\mathrm{Y}}-$*measurable. The processes* x_t, $w_{t,j}$, $u_{t,h}$, *and* $e_{t,h}$ *are* $L_{2q}-$*near-epoch-dependent (NED) of size* $-b$ *on* $\mathrm{Y}_{T,t}$ *with positive constants* $\{d_t\}$ *and a sequence* $\zeta_m \to 0$ *as* $m \to \infty$, *uniformly over* $j = 1,\dots,p$.

Assumption 1 (a) assumes that the error terms in (1.2.2) and (1.2.3) are not correlated with the contemporaneous regressors. Note that I impose the adaptation assumption to use asymptotic martingale difference property the Bernstein blocks must have (Davidson (1994), page 387). According to Davidson (2021), it is a widely employed assumption in time series econometrics, assuming that future shock information cannot help in predicting X_t given its history (Davidson (2021), page 383). Assumption 1 (b) ensures the processes have bounded $2\bar{q}-$th moments. Assumption 1 (d) is on the dependence structure. While $(x_t, \boldsymbol{w}_t, u_{t,h}, e_{t,h})$ themselves are not mixing processes, they depend almost entirely on 'near epoch' of $\{\mathrm{Y}_t\}$ (Davidson (2021), page 368), which is $\alpha-$mixing. This allows $(x_t, \boldsymbol{w}_t, u_t, e_{t,h})$ to inherit some mixingale properties, as will be stated in the following lemma.

Lemma 1.4.3 (Mixingale Property)**.** *Denote* $\varepsilon_{t,h}$ *as a genereric notation for the error terms* $u_{t,h}, e_{t,h}$, *and* $v_{t,h}$. *Under Assumption 1, the following holds. Note that* $\{c_t\}$ *and* ρ_m *are generic notations for the mixingale constants.*

(a) $\{v_{t,h}\}_{t=-\infty}^{\infty}$ *is causal* $L_{2q}-$*NED of size* $-b$ *on* $\mathrm{Y}_{T,t}$ *with positive bounded constants uniformly over* $h = 1,\dots,H_{\max}$.

(b) $\{w_{t,j}\varepsilon_t\}$ *is a causal* L_q-*mixingale with non-negative constants* $\{c_t\}$ *and sequences* ρ_m *for all* $j = 1,\dots,p$ *and* $h = 1,\dots,H_{\max}$.

(c) $\{w_{t,j}w_{t,k} - E[w_{t,j}w_{t,k}]\}$ *is a causal* L_q-*mixingale with non-negative constants* $\{c_t\}$ *and sequences* ρ_m *for all* $j \neq k$ *where* $j,k = 1,\dots,p$.

Proof. The proof can be found in Appendix A.2.1. □

This lemma shows that some transformations of NEDs and their demeaned processes are mixingales. It thus allows us to use Lemma 1.4.2. Next, I impose some assumptions on the mixingale constants to simplify some bounds used in the proof of the main theorems.

Assumption 2. *Recall the NED constants* $\{d_t\}$ *and a sequence* $\zeta_m \to 0$ *defined in Assumption 1 (d). Let* $\widetilde{\zeta}_m = 2\zeta_m + \zeta_m^2$ *and* $k := [q/2]$. *Define* $\rho_m = 6\alpha_k^{1/q-1/\bar{q}} + 2\widetilde{\zeta}_k$, *where* α_k *is the mixing coefficient for the sequences*

$X = \{\varepsilon_{t,h}, v_{t,h}, w_{t,j}\varepsilon_{t,h}, w_{t,j}w_{t,k}\}$ *for all* $j = 1,\ldots,p$ *and* $h = 1,\ldots,H_{\max}$. *Assume that*

$$\rho_m \leq \exp(-m\kappa)$$

for some $\kappa > c_\kappa \log p$, *where* $c_\kappa > 2$.

This assumption is on the mixingale constants, where I define the constants to align with the constants that inherit the mixingale properties. The sequence $\rho_m \to 0$ is hence the mixingale sequence for the transformed variables in Lemma 1.4.3. Assumption 2 gives the assumption of how fast ρ_m should decay. The parameter κ governs the strength of the dependence, with larger κ values indicating a weaker dependence. Note that this is a technical assumption I need to prove Thoerem 1.4.1, where it simplifies the dependence bound in the triplex inequality (1.4.4). The following is the assumptions on the degree of sparseness of the underlying coefficients, λ_h and γ_h.

Assumption 3. *Let* $|\beta_h| \leq C$. *Let* ξ *be a generic notation that represents* β_{-h}, λ_h, *and* γ_h *for all* $h = 1,\ldots,H_{\max}$. *Then* ξ *follows the following assumption. Suppose* $\sum_{j=1}^{p} \xi_j^2 \leq C$. *Also assume that there exists* $\delta > 1$ *and* $0 < C_\delta < \infty$ *that for any* $J \subseteq \mathfrak{P}$,

$$\sum_{j\in J} \left|\xi_j\right| \leq C_\delta \left(\sum_{j\in J} \xi_j^2\right)^{(\delta-1)/(2\delta-1)}, \tag{1.4.6}$$

where C refers to a generic constant $0 < C < \infty$. $\mathfrak{P}$ *refers to the power set of J, where J is a set of covariates.*

This assumption is the main difference between this method and LASSO, in that it doesn't limit the number of nonzero coefficients, but rather restricts the magnitudes of the coefficients. While we do not require that the parameters to have a natural order, these conditions can be expressed more simply by rearranging the parameters in descending order of their magnitude $|\xi_j|$. Denote the rearrangement as $|\xi_{(1)}| \geq |\xi_{(2)}| \geq \cdots \geq |\xi_{(p)}|$. Then, Assumption 3 3 implies

$$C_1 j^{-\delta} \leq \left|\xi_{(j)}\right| \leq C_2 j^{-\delta}, 0 < C_1 \leq C_2 < \infty, \delta > 1, \tag{1.4.7}$$

where from (1.4.7) we call Assumption 3 3 as a polynomial decay case: if (1.4.7) holds for some $\delta > 1$, then (1.4.6) holds for the same δ, as shown in Lemma A.2 in Ing (2020). Apart from (1.4.7), it can be shown that Assumption 3 3 also implies

$$\sum_{j=1}^{p} \left|\xi_j\right|^{1/\delta} < C, 0 < C < \infty, \delta > 1, \tag{1.4.8}$$

which is a frequently adopted assumption in the high-dimensional literature, as in Wang et al. (2014). Assume that

(1.4.8) holds for some δ. Applying Hölder's inequality,

$$\begin{aligned}\sum_{j\in J}\left|\xi_j\right| &\leq \left(\sum_{j\in J}\left|\xi_j\right|^{\frac{1}{2\delta-1}\frac{2\delta-1}{\delta}}\right)^{\delta/(2\delta-1)}\left(\sum_{j\in J}\xi_j^{\frac{2\delta-2}{2\delta-1}\frac{2\delta-1}{\delta-1}}\right)^{(\delta-1)/(2\delta-1)}\\ &=\left(\sum_{j\in J}\left|\xi_j\right|^{1/\delta}\right)^{\delta/(2\delta-1)}\left(\sum_{j\in J}\xi_j^2\right)^{(\delta-1)/(2\delta-1)}\\ &\leq C^{\delta/(2\delta-1)}\left(\sum_{j\in J}\xi_j^2\right)^{(\delta-1)/(2\delta-1)},\end{aligned}$$

and we can see that by setting $C_\delta = C^{\delta/(2\delta-1)}$, (1.4.6) holds. The parameter δ governs the degree of sparseness, with larger values indicating a faster decay of the coefficients. Thus, we can see that Assumption 3 covers a wider class of sparsity conditions where the condition trivially implies the exact sparsity case.

Remark. There are many variations of sparsity-based assumptions that extend the exact/strong sparse assumption. First, (1.4.8) is called a soft sparsity, as opposed to strong sparsity, in that L_q norm is bounded for $0 < q \leq 1$, whereas strong sparsity means that L_0 norm is bounded by a small constant. Another concept is the approximate sparsity (Chernozhukov et al. (2021), Comment 3.1). The concept is to add a fast enough approximation error based on the error bounds, so that exact sparsity can "approximate" a less sparse DGP such as (1.4.7).

Assumption 4. *Assume the followings.*

(a) For some positive constant C_1 and C_2, it holds that

$$\max_{1\leq|J|\leq C_1(T/\log p)^{1/2}, i\notin J}\left\|\Gamma^{-1}(J)E[w_{t,i}\boldsymbol{w}_t(J)]\right\|_1 < C_2,$$

where $\Gamma(J) = E[\boldsymbol{w}_t(J)\boldsymbol{w}_t(J)']$.

(b) Let $\Sigma = E[\sum_{t=1}^T \boldsymbol{w}_t\boldsymbol{w}_t'/T]$. Assume $\max_{j\in[p]}\Sigma_{j,j} < C$ and $1/C \leq \Lambda_{\min}$, where $\Lambda_{\min}$ is the minimum eigenvalue of Σ and $0 < C < \infty$.

(c) $\log p = o(T^{\frac{\delta-1}{3(2\delta-1)}})$ and $(T/(\log p)^3)^{1/2}p^{-\underline{c}} = o(1)$ for some $\underline{c} = \min\{c_\kappa - 2, c_M(\bar{q}-1)-2\} > 0$, where $c_M > 2/(\bar{q}-1)$ and $c_\kappa > 2$.

Assumption 4 are some additional assumptions on the covariates and growth rates of the dimensionality $dim(\boldsymbol{w}_t) = p$. Assumption 4 (a) limits strong correlations between covariates, ensuring that the OLS coefficients of one covariate on any set of other covariates remain finite. Assumption 4 (b) is on minimum eigenvalue of the Gram matrix, with a remark that this condition is on the population level, not on the sample matrix as in the restricted eigenvalue conditions

in Bickel et al. (2009). The third condition addresses the relationship between the growth of the covariate dimension p and sample size T. It specifies that $\log p$ should not grow too quickly compared to T. This condition is crucial for bounding the triplex inequality in the proof of Theorem 1.4.1. Note that this assumption is stronger than the usual growth rates in the high-dimensional literature with i.i.d. data, for example $\log p = o(T^{1/3})$ in Belloni et al. (2013a). While bounding either the Bernstein bound or the tail bound would require $\log p = o(T^{(\delta-1)/(2\delta-1)})$, bounding all the terms simultaneously requires a stricter condition $(\log p)^3 = o(T^{(\delta-1)/(2\delta-1)})$. More detailed explanations are given in the proofs of Theorems 1.4.1 and 1.4.2. While this may sound restrictive, note that this assumption is on the growth rate of $\log p$, where p can grow at much faster rates.

1.4.3 Inference

In this subsection I establish the inference for the parameter of interest with each horizon h. I use a matrix representation as the baseline instead of (1.2.2) – (1.2.4) for simplicity. Let $\boldsymbol{y}_h$, $\boldsymbol{x}_h$, $\boldsymbol{u}_h$, and $\boldsymbol{v}_h$ be the $T \times 1$ vector and W_h be the $T \times p$ matrix, where $p = dim(\boldsymbol{w}_t)$. Then the models can be written in a compact matrix form:

$$\begin{aligned} \boldsymbol{y}_h =& \beta_h \boldsymbol{x}_h + W_h \beta_{-h} + \boldsymbol{u}_h, \\ \boldsymbol{x}_h =& W_h \gamma_h + \boldsymbol{v}_h. \end{aligned} \tag{1.4.9}$$

Theorem 1.4.1 (Error bounds for Double OGA+HDAIC)**.** *Denote ξ as a representative vector of parameters, β_{-h} and γ_h, for all $h = 1, \ldots, H_{\max}$. Denote $E_T[.] = E[.|(y_{t+h}, x_t, \boldsymbol{w}_t)_{t=1}^{T-h}]$. Under Assumptions 1 – 4, the following holds for all $h = 1, \ldots, H_{\max}$.*

$$E_T\left[\left\|W_h(\widehat{\xi} - \xi)\right\|^2\right] = O_p\left(\left(\frac{(\log p)^3}{T}\right)^{1-1/2\delta}\right). \tag{1.4.10}$$

Proof. The proof can be found in Appendix A.1.1. □

Note that the convergence rates are not the fastest rates proposed in the main theorem in Ing (2020): the error bounds are $(\log p/T)^{1-1/2\delta}$ for the polynomial decay case. The main results require assumptions (A1) and (A2) in Ing (2020), and with Assumption 1 and Lemma 1.4.2 those assumptions are not satisfied. I hence use the bounds from weakened assumptions in equations **(2.33)** and **(2.34)** in Ing (2020), resulting in the slower converge rates. As noted in equation **(2.35)** of Ing (2020), these relaxed assumptions require $\log p = o(T^{1/3})$, which is implied by Assumption 4 (c). While Theorem 1.4.1 itself is useful in deriving error bounds for the high-dimensional nuisance parameters, our parameter of interest is β_h in equation (1.2.2). The following theorem derives asymptotic distribution of $\widehat{\beta}_h$ using the error bounds derived in Theorem 1.4.1.

Theorem 1.4.2. *Under Assumptions 1 – 4,*

$$\sqrt{T}(\widehat{\beta}_h - \beta_h) \xrightarrow{d} N(0, \sigma_h^2),$$

where $\sigma_h^2 = \lim_{T\to\infty} \Omega_h / \tau_h^4$, Ω_h *is the long-run covariance matrix following Newey-West, and* $\tau^2 = E[v_h' v_h / T]$.

Proof. The proof can be found in Appendix A.1.2. □

By Theorem 1.4.2, the estimator of interest achieves the square root T convergence rate regardless of the slower convergence rate of the nuisance parameters presented in Theorem 1.4.1. The following theorem establishes the validity of the proposed variance estimator.

Theorem 1.4.3. *Define* $\psi_{h,t} = v_{h,t} u_t$ *and* $\tau_h^2 = E[v_h v_h / T]$. *Let* $\sigma_h^2 = \lim_{T\to\infty} \tau_h^{-4} \Omega_h$, *where* $\Omega_h = \sum_{\ell=-(T-1)}^{T-1} 1/T \sum_{t=\ell+1}^{T} E[\psi_{h,t} \psi_{h,t-\ell}]$ *with a bandwidth parameter K. Denote the sample analogues as* $\hat{\tau}_h$, $\hat{\sigma}_h^2$ *and* $\hat{\Omega}_h$. *Under Assumptions 1 – 4,*

$$\left|\hat{\sigma}_h^2 - \sigma_h^2\right| \xrightarrow{p} 0.$$

Proof. The proof can be found in Appendix A.1.3. □

With Theorems 1.4.2 and 1.4.3, we can now construct $100(1-\alpha)\%$ confidence intervals for each horizon. Given a confidence level α, an asymptotic $100(1-\alpha)\%$ confidence interval $\widehat{\mathscr{I}}_{\alpha,h}$ is given by

$$\widehat{\mathscr{I}}_{\alpha,h} := [\widehat{\beta}_h - z_\alpha \hat{\sigma}_h / \sqrt{T},\ \widehat{\beta}_h + z_\alpha \hat{\sigma}_h / \sqrt{T}], \tag{1.4.11}$$

where $z_\alpha = \Phi^{-1}(1-\alpha/2)$ is the $(1-\alpha/2)$ quantile of the standard normal distribution and $\hat{\sigma}_h$ is defined in Theorem 1.4.2.

1.5 Empirical Applications

1.5.1 The Effect of Democracy on GDP

I illustrate the performance of the proposed method through an empirical investigation of the response of GDP growth to democratization, revisiting the LP framework in Acemoglu et al. (2019). They develop a binary index of democracy by gathering information from several different sources of data. The data covers 184 countries from 1960 to 2010. In part of their analysis, they adopt the LP specification to see how the effect of democratization evolves over time. I adopt

the baseline LP specification in Acemoglu et al. (2019)

$$y_{c,t+h} - y_{c,t-1} = \beta_h^{\text{LP-DiD}} \Delta D_{ct} + \delta_t^h + \sum_{j=1}^{p} \gamma_j^h y_{c,t-j} + \varepsilon_{ct}^h, \tag{1.5.1}$$

where $y_{c,t}$ denotes the log GDP per capita in country c at time t, D_{ct} is the binary democracy variable, and δ_t^h the time dummy. The baseline model includes the lagged terms of the GDP to address the selection into democracy, as the descriptive data suggests the dip in GDP preceding democratizations. The identifying assumption is that conditional on the lags of GDP, countries that democratize do not follow a different GDP trend relative to other nondemocracies. This specification can be viewed as a version of the LP-DiD identification scheme by Dube et al. (2023), where the sample is restricted to

$$\begin{cases} \text{democratizations} & D_{it} = 1, D_{i,t-1} = 0 \\ \text{clean controls} & D_{it} = D_{i,t-1} = 0. \end{cases}$$

I ran (1.5.1) with the conventional LP and with my method, with $p = 4$ lags of $y_{c,t}$ included as controls. The results are depicted in Figure 1.4. While the point estimates are very similar, the proposed method has efficiency gain overall, especially for the longer horizon estimates. As expected, the standard errors increase with the horizon due to the decreasing effective sample size. However, even after accounting for long-run variance, the proposed method results in smaller standard errors, attributable to the model selection process. The empirical findings align with the original results, indicating significant GDP growth of 20 percent higher than the controls over a 20-year period.

However, as noted in their paper, even after controlling for fixed effects and GDP dynamics, there are possible sources of bias coming from unobservables related to future GDP and the change in democracy. To investigate these factors, the authors show the robustness of the results with different sets of controls. The sets of controls include potential trends related to differences in the level of GDP in the initial period, dummies for the democracy of Soviet and Soviet satellite countries, lags of trade and financial flows, lags of demographic structure, and the full set of region×initial regime×year dummies, and more in the appendix. To see the robustness against high-dimensional controls, I have combined these controls into a union set. This brings the number of covariates included in the model to 276 with a sample size of 774 at horizon 20.

The results are displayed in Figure 1.5. As expected for the conventional LP, the standard errors are larger due to the inclusion of all regressors, and the estimates are generally insignificant. Even with the proposed method, most estimates remain insignificant, although it manages to maintain moderate standard errors. However, the proposed method identifies significant GDP growth at longer horizons, starting from year 25 onward. The percentage increase

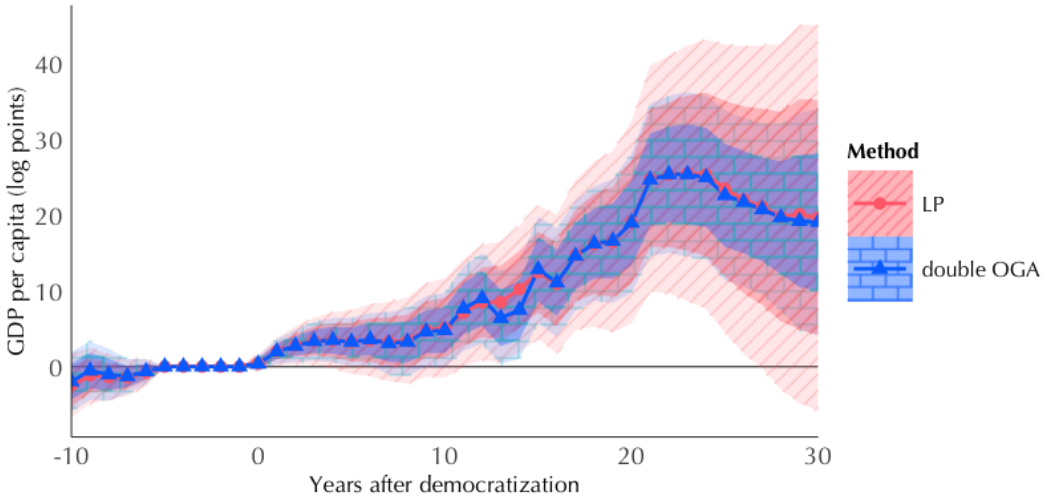


Figure 1.4: The effect of democratization on GDP growth, baseline model

Notes. This model follows the baseline model in (1.5.1), where 4 lags of GDP are included as controls. Solid lines represent the point estimates, derived using the LP (red, diagonal shading) and double OGA (blue, brick-shaped shading) methods. Light-shaded areas indicate 90% confidence intervals; dark shades 68%.

reaches 17.33 percent in year 30, which is consistent with the original findings using baseline controls. This result suggests that the proposed method provides reliable estimates even when considering a large number of potential confounders, which is essential for the causal interpretation of the parameter.

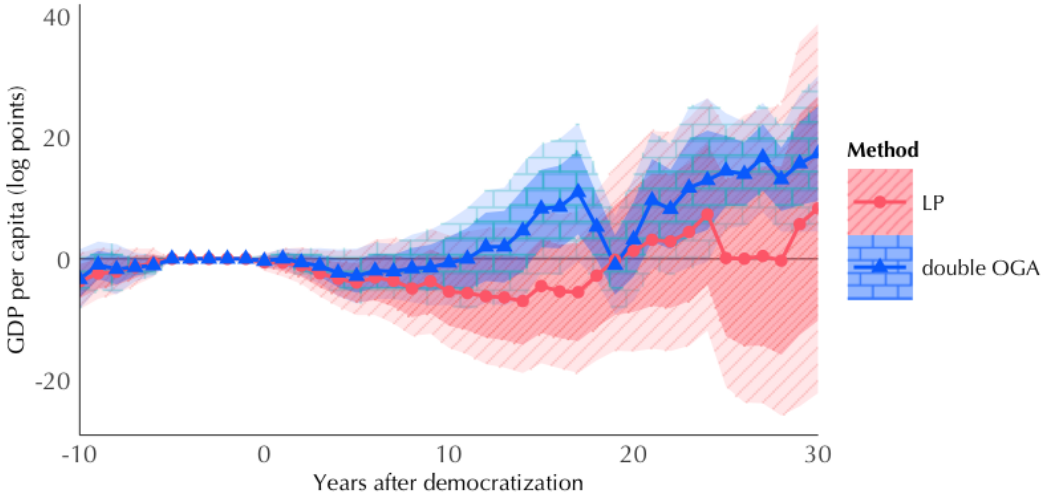


Figure 1.5: The effect of democratization on GDP growth, high-dimensional controls

Notes. The model includes a total of 276 covariates at horizon 20, which are listed in the main text. Solid lines represent the point estimates, derived using the LP (red, diagonal shading) and double OGA (blue, brick-shaped shading) methods. Light-shaded areas indicate 90% confidence intervals; dark shades 68%.

1.5.2 Subjective Beliefs in Business Cycle Models

In this subsection, I revisit the empirical study by Bhandari et al. (2024), which explores the impact of subjective beliefs on macroeconomic aggregates. Their research develops a theory on how the bias of subjective beliefs from rational beliefs affects key economic indicators, formalizing these departures using model-consistent notions of pessimism and optimism.

The theoretical framework posits that fluctuations in beliefs, particularly increases in pessimism, have significant effects on the macroeconomy. This pessimism is hypothesized to be contractionary and to increase belief biases in both inflation and unemployment forecasts. The mechanism operates through various channels: pessimistic households lower current demand due to consumption smoothing, firms adjust their pricing and hiring strategies based on expectations of future economic conditions, and labor market frictions amplify these effects. This shared pessimistic outlook creates a positive correlation between the biases in inflation and unemployment forecasts.

Figure 1.6 replicates the dynamic responses originally presented in Figure 10 of Bhandari et al. (2024). It compares the dynamic responses under two scenarios: one where all agents, including firms, hold subjective beliefs, and another where firms adopt rational expectations while households retain subjective beliefs. The solid line represents the case where all agents follow subjective beliefs, while the dashed line reflects the responses when firms adopt rational expectations. This comparison highlights important differences, where rational firms exhibit more muted fluctuations compared to the scenario where all agents are subjective. Specifically, rational firms keep inflation lower on impact, as they perceive an increase in pessimism as contractionary but do not anticipate higher future marginal costs, leading to smaller price adjustments. This figure provides a baseline understanding of the impulse responses under different belief structures, which will serve as the foundation for further analysis.

To test this theory empirically, the authors utilize data from multiple sources. The key variable is the belief bias, which they term the ”belief wedge”—defined as the difference between subjective beliefs and rational forecasts. Subjective beliefs are measured using the University of Michigan Surveys of Consumers, while rational predictions are constructed using both VAR predictions and forecasts from the Survey of Professional Forecasters (SPF). The study focuses on quarterly data from 1982Q1 to 2019Q4.

Since their model predicts a one-factor structure of the belief wedges, they define the belief shock as the first principal component derived from the standardized inflation belief wedge and unemployment belief wedge. To address a possible misspecification of VAR forecasts and hence the belief wedges, the authors construct a belief shock using the principal component of the belief wedges between the Michigan and SPF forecasts. The question here is how positive deviation to this belief shock (pessimism) affects the economy.

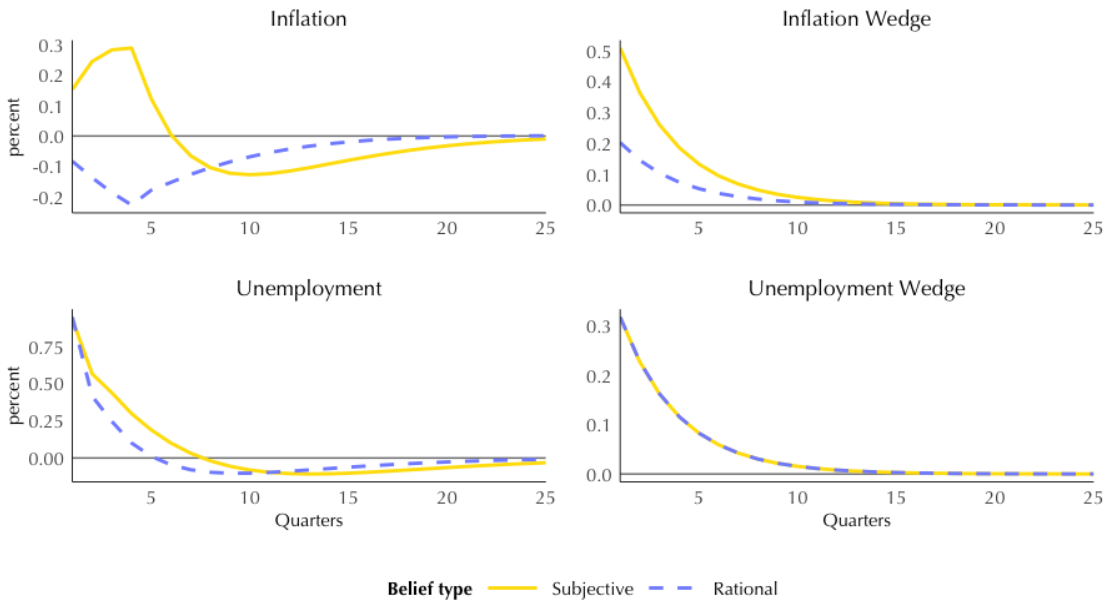


Figure 1.6: Impulse responses under subjective and rational beliefs

Notes. Replicated from Figure 10 in Bhandari et al. (2024). This figure compares impulse responses under two scenarios: subjective beliefs for all agents (solid line) and rational expectations for firms (dashed line).

The general estimating equation can be written as:

$$y_{t+h} = \beta_h \theta_t + \sum_{\ell=1}^{L} \delta_\ell \theta_{t-\ell} + \mathbf{w}_t' \gamma + u_{t,h}, \tag{1.5.2}$$

where θ_t is the belief shock, the first PC of the inflation and unemployment belief wedges, $\mathbf{w}_t$ includes additional controls, and $u_{t,h}$ the error term.

Baseline Analysis and Robustness Across Specifications

Figures 1.7–1.8 present the responses of the inflation, unemployment, and the belief wedges to a positive innovation in the belief shock across different model specifications. Each set of responses is estimated using the conventional LP and the proposed method. To provide context, the model-implied impulse response functions are included, where all agents are assumed to hold subjective beliefs. This provides a theoretical benchmark against which the empirical results can be evaluated.

I begin by presenting the results for the baseline model, where no additional controls are included and the lag order is set to 4. As shown in Figure 1.7 (a), the empirical results generally align with the model's predictions. Both LPs and the model show initial increases in inflation. While the model does not generate the hump-shaped responses by both

methods, the authors say the magnitude is comparable to what the model implies. These outcomes align with increased pessimism, resulting in higher unemployment and inflation wedges. This alignment provides initial support for the robustness of the original study's conclusions.

To account for the influence of key economic indicators on beliefs and to ensure robustness, alternative model specifications are considered, including adding variables used to create the VAR forecasts as controls and increasing their lags. These controls, $\mathbf{w}_t$ in (1.5.2), include nine key economic indicators: inflation between end-of-quarter months, real GDP growth, the unemployment rate, the federal funds rate, the relative price of investment, capital utilization, hours worked per person, the consumption rate, and the investment rate.

Invoking the recursive identification scheme, this approach assumes that these economic factors influence the belief shock variable, while the belief shock does not, in turn, affect these factors. This assumption aligns intuitively, as both subjective beliefs and rational forecasts are derived from the current state of economic indicators. The picture becomes more complex when additional controls are introduced. When I added the 9 variables used for the VAR prediction with 4 lags (Figure 1.7 (b)), the standard errors for the conventional LP increased substantially. This increase in uncertainty makes it challenging to interpret some results coming from LP, particularly for inflation, where no significant effects are observed.

Further increasing the lags to 8 and 10 (Figures 1.8 (a) and (b), the conventional LP produces increasingly erratic results with inconsistent patterns. In contrast, the proposed method maintains consistent results with narrower confidence intervals across different specifications. Notably, the proposed method closely mimics the model-implied impulse responses, with the exception of unemployment. While neither methods do not comply with the model-implied responses for unemployment, the proposed method consistently find an approximately 0.5 percent increase in unemployment in magnitude across all model specifications.

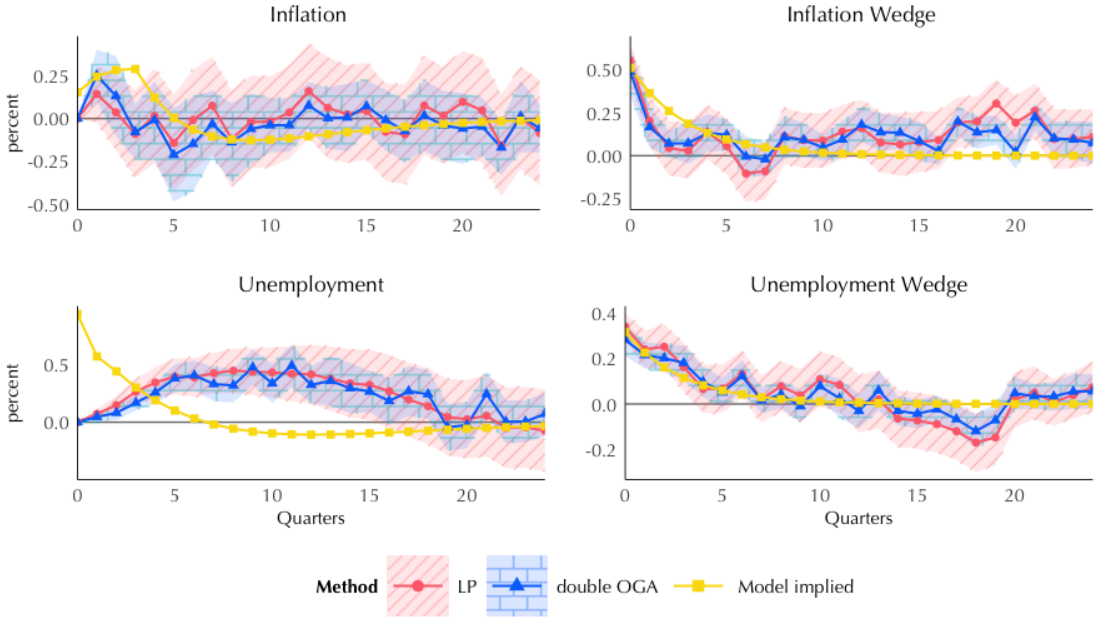
These results reveal important insights about the robustness of different estimation methods across various model specifications. As the number of included variables and lags increases, the conventional LP method shows increasing variability in its results, producing wiggly estimates with wider standard errors. In contrast, the proposed method demonstrates remarkable robustness, maintaining consistent patterns and narrower confidence intervals across different specifications.

This robustness becomes particularly evident when the model includes more variables and higher lag orders. Importantly, the proposed method performs consistently in both simple and more complex settings, suggesting there's no disadvantage to using it when the underlying model structure is uncertain—a common scenario in empirical analysis.

For comparison, I also ran debiased LASSO[4] from Adamek et al. (2024) in Figure 1.9. The results with other model

[4]For the debiased lasso estimation, I use the R package *desla* provided by Adamek et al. (2024).

(a) Baseline with 4 lags



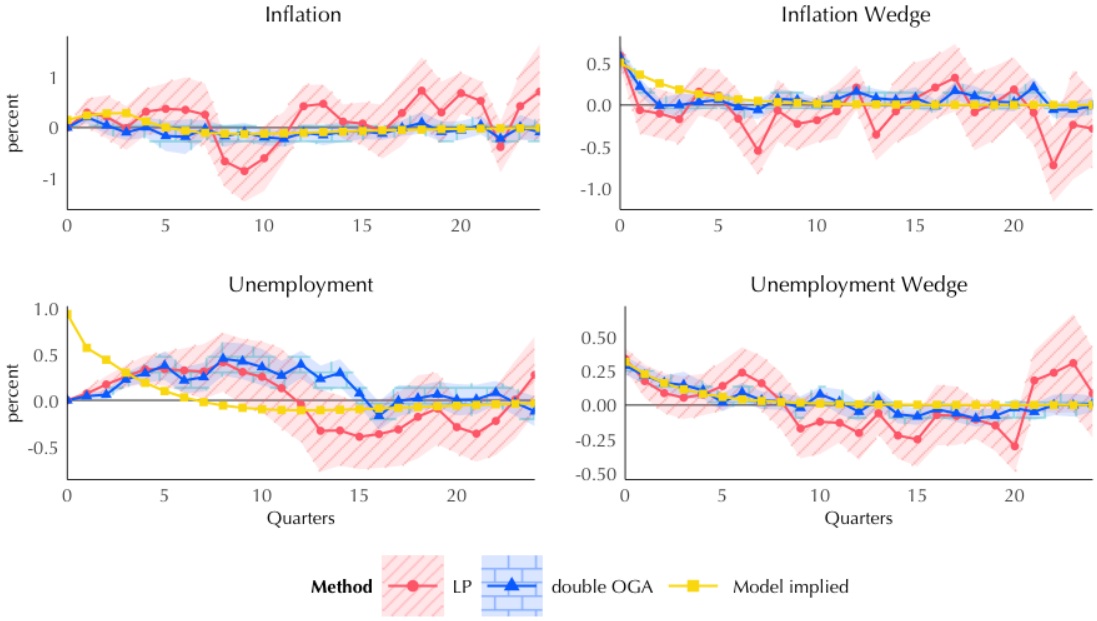
(b) VAR controls with 4 lags

Figure 1.7: Baseline and VAR controls added models with 4 lags

Notes. This figure compares the baseline model with 4 lags (Panel (a)) and the VAR controls added model with 4 lags (Panel (b)) for inflation, inflation wedge, unemployment, and unemployment wedge. Point estimates are shown as solid lines with circle markers for LP (red), triangle markers for double OGA (blue), and square markers for model-implied (yellow). Shaded areas represent 90% confidence intervals, with diagonal shading for LP and brick-shaped shading for double OGA.

specifications are in Appendix A.4, where debiased LASSO shows similar patterns across different model specifications. In particular, the method yields highly persistent responses for inflation and inflation wedges. The unemployment

(a) VAR controls with 8 lags



(b) VAR controls with 10 lags

Figure 1.8: VAR controls added models with longer lags

Notes. This figure compares the VAR controls added model with 8 lags (Panel (a)) and with 10 lags (Panel (b)) for inflation, inflation wedge, unemployment, and unemployment wedge. Point estimates are shown as solid lines with circle markers for LP (red), triangle markers for double OGA (blue), and square markers for model-implied (yellow). Shaded areas represent 90% confidence intervals, with diagonal shading for LP and brick-shaped shading for double OGA.

response, while significant, tends to be overestimated, and the unemployment wedge exhibits a persistent negative bias over longer horizons.

These results are challenging to interpret both intuitively and within the theoretical perspective proposed by Bhandari et al. (2024). One possible explanation is that these models are relatively low-dimensional, which is not an ideal setting for LASSO. Alternatively, as suggested by the simulations, the high persistence in the data might be driving the estimates away from expected values.

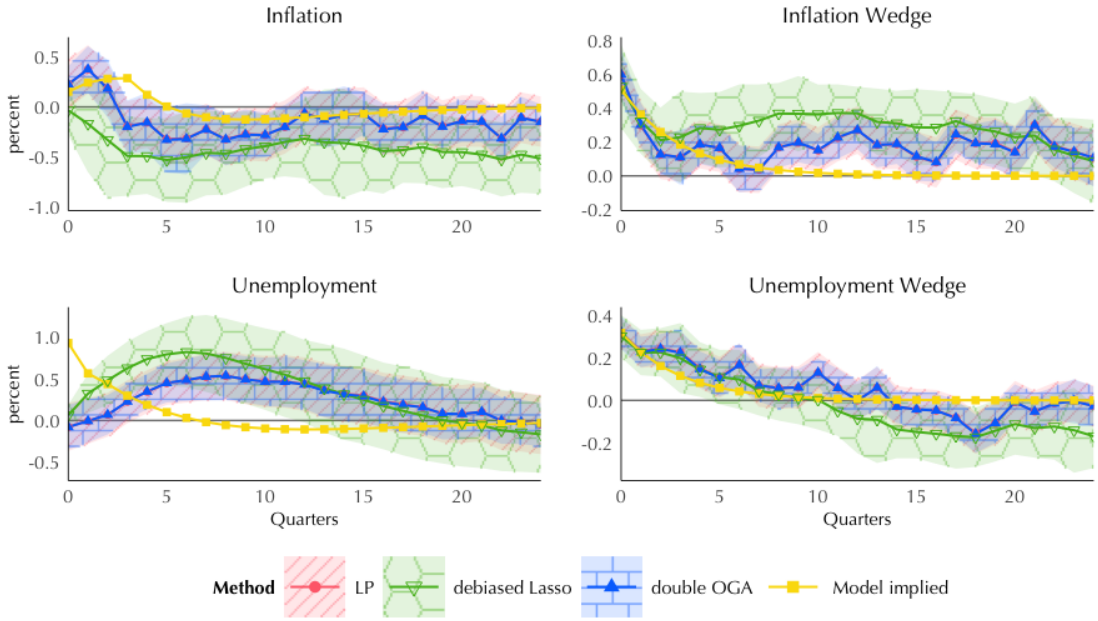


Figure 1.9: Baseline model with debiased LASSO

Notes. This figure compares the baseline model for inflation, inflation wedge, unemployment, and unemployment wedge. Point estimates are shown as solid lines with circle markers for LP (red), inverted triangle markers for debiased LASSO (green), triangle markers for double OGA (blue), and square markers for model-implied (yellow). Shaded areas represent 90% confidence intervals, with diagonal shading for LP, hexagonal shading for debiased LASSO, and brick-shaped shading for double OGA.

In summary, the empirical analysis largely supports the findings of Bhandari et al. (2024), showing that positive shocks to belief wedges lead to increases in both inflation and unemployment, consistent with the hypothesized effects of heightened pessimism. The proposed method demonstrates robustness across different model specifications, producing stable results with added controls and varying lag lengths.

1.6 Concluding Remarks

Local projection has emerged as the preferred alternative to VARs in impulse response analysis due to its robustness to model misspecification and ease of implementation (Montiel Olea and Plagborg-Møller, 2021; Olea et al., 2024). As the use of local projections has gained traction, recent literature has addressed the challenges of incorporating high-dimensional covariates, with a particular focus on methods like LASSO. However, the reliance on strong sparsity assumptions in these methods limits their applicability in many empirical settings, especially when dense DGPs are

present.

This paper aims to address the gap in the literature by introducing a high-dimensional local projection approach that caters to both sparse and dense settings and takes into account the uncertainty of sparseness in the DGP. Building on the OGA with HDAIC method proposed by Ing (2020), the proposed method relaxes the need for strict sparsity by allowing parameters to decay toward zero as dimensionality increases, making it especially suitable for economic time series with autoregressive properties.

The proposed framework demonstrates significant advantages in both DiD estimation and impulse response analysis. In addition, the proposed approach has the advantage of interpretability through the use of OGA, which orders covariates based on their explanatory power. Through simulations, I show strong performance in handling persistent data and longer horizons estimations.

By utilizing the OGA with HDAIC and the NED assumptions, the local projection estimator achieves $\sqrt{T}$ asymptotic normality with HAC standard errors. The theoretical foundation of this method is based on the error bounds of Ing (2020), the triplex inequality of Jiang (2009), and double selection arguments of Belloni et al. (2013a). These advances strengthen the framework's capacity to handle complex time series data while maintaining reliable inference.

In conclusion, this method advances econometric modeling by offering a practical, robust solution for high-dimensional datasets. Its applicability to both sparse and dense scenarios makes it a practical tool for researchers working on a range of empirical applications, from macroeconomic analysis to event studies, providing a framework that ensures consistency and reliability across a broad range of empirical applications.

CHAPTER 2

Inference in High-Dimensional Regression Models without the Exact or L^p sparsity

This chapter is coauthored with Harold D. Chiang and Yuya Sasaki.

2.1 Introduction

The advent of modern machine learning[1] techniques has significantly widened the class of analyzable regression models that include models with high-dimensional controls and/or models with flexible nonlinearity. Perhaps the most popular and important machine learning approaches to estimation and inference in high-dimensional regression models today are those based on shrinkage and regularization, such as the least absolute shrinkage and selection operation (LASSO). While they have practically appealing properties, these popular machine learning methods yet rely on a list of assumptions that may not be necessarily mild under certain applications. In particular, the assumptions of the exact sparsity and the L^p sparsity required for these methods are sometimes controversial and perceived to be strong for some applications. As such, there still remains room in the literature for further widening the class of analyzable high-dimensional regression models if these assumption can be relaxed by a new machine learning method.

This paper proposes a method of inference in high-dimensional regression models without requiring the exact sparsity or the L^p sparsity. We set a low-dimensional parameter vector as the object of interest, and treat the remaining high-dimensional parameter vector as a nuisance component. Under this common setting, our proposed method of inference works as follows. First, use the orthogonal greedy algorithm (OGA; Temlyakov, 2000) to order the high-dimensional regressors in a descending order of explanatory power. Second, use the high-dimensional Akaike information criterion (HDAIC; Ing, 2020) to select a model among the ordered list of models constructed in the first step. Third, estimate the models selected in the second step. Fourth, plug the estimated selected models in a Neyman orthogonal score and estimate the low-dimensional parameter vector of interest.

We take advantage of a number of recent methodological and theoretical developments, namely OGA, HDAIC, and DML, to derive asymptotic statistical properties of this new method of inference. Ing (2020) investigates convergence rate properties of an estimator of high-dimensional regression models based on OGA and HDAIC (hereafter referred to as OGA+HDAIC). Importantly, the setting imposes assumptions on the high-dimensional parameter vector that are weaker than the exact sparsity and the L^p sparsity. Furthermore, this approach does not require to impose high-level conditions on the sample Gram matrix, such as a restricted eigenvalue type condition, that are often required in the

[1] We use the phrase "machine learning" following the recent related literature, but it is worthy to remark that it is a synonym of semiparametric or high-dimensional 'estimation.'

literature. Even under these weaker conditions, it is still possible to obtain similar rates of convergence to those based on existing machine learning techniques such as the LASSO. Given the adequately fast convergence rates of the preliminary nuisance parameter estimators based on OGA+HDAIC, we can apply the post-double-selection approach (Belloni et al., 2013b) or the double/debiased machine learning (DML) framework (Chernozhukov et al., 2018) to in turn obtain a root-N convergence of the low-dimensional parameter vector of interest with a limit normal distribution.

Simulation studies demonstrate that this proposed method performs significantly better than those based on the LASSO or the random forest, especially when the data generating model becomes less sparse. We apply the proposed method to an analysis of production functions using a panel of Chilean firms.

Relation to the literature: This paper is related to several branches of the econometrics and statistics literature. First, it is closely related to the literature on inference in high-dimensional regression models and high-dimensional IV regression models, e.g., Belloni et al. (2012), Belloni et al. (2013b), Javanmard and Montanari (2014), Van de Geer et al. (2014), Zhang and Zhang (2014), Caner and Kock (2018a), Caner and Kock (2018b), Belloni et al. (2018a), Galbraith and Zinde-Walsh (2020), Gold et al. (2020), Kueck et al. (2021) to list a few. The majority of these papers focus on utilizing LASSO (Tibshirani, 1996) and its variants for estimation, and thus rely crucially on the exact sparsity, approximate sparsity[2] or L^p sparsity. We contribute to this literature, as emphasized above, by relaxing the conventional assumptions of these notions of sparsity. Second, also related is the literature on model selection methods for high-dimensional models. For extensive reviews of this vast literature, we refer readers to the monographs of Bühlmann and van de Geer (2011), Giraud (2015), and Hastie et al. (2019). In particular, we take advantage of the theoretical results of OGA+HDAIC by Ing (2020) as one of the main auxiliary steps to our goal as emphasized earlier. Methodologically, our paper benefits from the theoretical studies of various greedy algorithms in Temlyakov (2000), Tropp (2004), Tropp and Gilbert (2007), Ing and Lai (2011), and share ties with other iterated model selection methods such as the least absolute angle regression of Efron et al. (2004), the L_2-boosting of Bühlmann and Yu (2003), the test-based forward model selection of Kozbur (2017, 2020), and so forth. Third, this paper is related to the literature on Neyman orthogonal scores or locally robust scores, e.g., Belloni et al. (2015), Chernozhukov et al. (2016), Belloni et al. (2018b), and, in particular, the DML (Chernozhukov et al., 2018). We contribute to this literature by proposing to add OGA+HDAIC to the library of the list of preliminary estimators. Fourth, the conditions that we impose in place of the exact sparsity and the L^p sparsity concern the speed at which the absolute size of the regression parameters decays when descendingly ordered. These conditions are analogous to those that are used for model selection problems in autoregressive time series models (e.g., Shibata, 1980; Ing, 2007) as well as the ordinary- and super-smoothness on probability density functions that are used in the deconvolution literature (e.g., Fan, 1991; Fan and Truong, 1993).

[2]The approximate sparsity is closely related to the exact sparsity.

2.2 High-Dimensional Linear Regression Models

2.2.1 The Model

Consider the linear regression model

$$Y = D\theta_0 + X'\Lambda_0 + U, \qquad E[U|X,D] = 0, \tag{2.2.1}$$

where Y denotes an outcome variable, D denotes a treatment variable, X denotes a p-dimensional vector of controls, and U denotes unobserved factors. We allow for a high dimensionality in the sense that p can be increasing in N and may be even larger than N – more details will follow. In this framework, we are interested in the partial effect θ_0 of D on Y. Also write the linear projection of D on X:

$$D = X'\beta_0 + V, \qquad E[V|X] = 0, \tag{2.2.2}$$

In Section 2.4.1, we consider an extended model in which we introduce approximation errors in (2.2.1)–(2.2.2).

To construct a moment restriction under (2.2.1)–(2.2.2), consider the orthogonal score function from Robinson (1988):

$$\psi(Y,D,X;\theta,\eta) := \left\{Y - X'\gamma - \theta(D - X'\beta)\right\}(D - X'\beta), \tag{2.2.3}$$

where $X'\gamma_0 = E[Y|X]$ and $\eta = (\gamma,\beta)$. Note also that $X'\beta_0 = E[D|X]$ follows by construction from (2.2.2).

Notations. To proceed, we first fix basic notations. We use subscripts i and j to denote indices of observations and coordinates, respectively. Define $X_{Ij} = (X_{ij},\, i \in I)$ as a $|I| \times 1$ vector, $X_{iJ} = (X_{ij},\, j \in J)$ as a $|J| \times 1$ vector, and $X_{IJ} = (X_{iJ},\, i \in I)'$ as a $|I| \times |J|$ matrix, where I is a subset of observation indices $\{1,2,...,N\}$, J is a subset of coordinate indices $\mathfrak{P} \equiv \{1,2,\ldots,p\}$, and $|.|$ denotes the set cardinality. For any vector, $\|.\|$ refers to the Euclidean norm. The L^q norm is defined by $\|\xi\|_q = (\sum_{j=1}^p \xi_j^q)^{1/q}$ for $q < \infty$ and $\|\xi\|_\infty = \max_{1\le j\le p} |\xi_j|$.

2.2.2 The Method

This section provides an overview of the method. We propose the following procedure for a root-N consistent estimation and inference about the partial effect θ_0 without assuming sparsity on the high-dimensional parameters, β_0 or γ_0.

We highlight three notable elements of this algorithm. First, the overall procedure (Steps 1–4) uses the cross fitting to remove an over-fitting bias. Specifically, by using complementary sub-sample I_k^c to estimate the nuisance parameters $\widehat{\eta}_k = (\widehat{\gamma}_k, \widehat{\beta}_k)$ that are in turn evaluated in the I_k-mean of the score, we can circumvent a bias that arises from products

Algorithm 2 OGA+HDAIC with DML for high-dimensional linear models

S1. Randomly split the sample indices $\{1,...,N\}$ into K folds $(I_k)_{k=1}^K$. For simplicity, let the size of each fold be $n = N/K$ and the size of I_k^c be n^c.

S2. For each fold $k \in \{1,...,K\}$, perform following procedure using $\{(X_i', D_i)'\}_{i \in I_k^c}$ to get $\widehat{\beta}_k$.

(a) Compute $\widehat{\mu}_{0,j} = X_{I_k^c j}' D_{I_k^c} / \sqrt{n^c} \| X_{I_k^c j} \|$. Select the coordinate $\widehat{j}_1 = \text{argmax}_{1 \le j \le p} \left| \widehat{\mu}_{0,j} \right|$. Define $\widehat{J}_1 = \{\widehat{j}_1\}$.

(b) Compute $\widehat{\mu}_{1,j} = X_{I_k^c j}' (I_{n^c} - H_1) D_{I_k^c} / \sqrt{n^c} \| X_{I_k^c j} \|$, where $H_1 = X_{I_k^c \widehat{j}_1} (X_{I_k^c \widehat{j}_1}' X_{I_k^c, \widehat{j}_1})^{-1} X_{I_k^c, \widehat{j}_1}'$. Select the coordinate $\widehat{j}_2 = \text{argmax}_{1 \le j \le p, j \notin \widehat{J}_1} \left| \widehat{\mu}_{1,j} \right|$. Update $\widehat{J}_2 = \widehat{J}_1 \cup \{\widehat{j}_2\}$.

(c) Given $m-1$ coordinates $\widehat{J}_{m-1}$ that have been obtained, compute $\widehat{\mu}_{m-1,j} = X_{I_k^c j}' (I_{n^c} - H_{m-1}) D_{I_k^c} / \sqrt{n^c} \| X_{I_k^c j} \|$, where $H_{m-1} = X_{I_k^c \widehat{J}_{m-1}} (X_{I_k^c \widehat{J}_{m-1}}' X_{I_k^c \widehat{J}_{m-1}})^{-1} X_{I_k^c \widehat{J}_{m-1}}'$. Select the coordinate $\widehat{j}_m = \text{argmax}_{1 \le j \le p, j \notin \widehat{J}_{m-1}} \left| \widehat{\mu}_{m,j} \right|$. Iteractively update $\widehat{J}_m = \widehat{J}_{m-1} \cup \{\widehat{j}_m\}$.

(d) Compute HDAIC $(\widehat{J}_m) = (1 + C^* |\widehat{J}_m| \log p / n^c) \widehat{\sigma}_m^2$ for each m, where C^* is from (2.2.5) in Section 2.2.3 and $\widehat{\sigma}_m^2 = 1/n^c D_{I_k^c}' (I - H_m) D_{I_k^c}$. Choose $\widehat{m} = \text{argmin}_{1 \le m \le M_n^*} \text{HDAIC}(\widehat{J}_m)$, where M_n^* is defined in (2.2.4) in Section 2.2.3.

(e) With coordinates $\widehat{J}_{\widehat{m}}$, run OLS of D_i on $X_{i \widehat{J}_{\widehat{m}}}$ to get $\widehat{\beta}_k$.

S3. Repeat S2 with $\{(X_i', Y_i)'\}_{i \in I_k^c}$ instead of $\{(X_i', D_i)'\}_{i \in I_k^c}$, to get $\widehat{\gamma}_k$ for each fold $k \in \{1,...,K\}$.

S4. Obtain $\breve{\theta}$ as a solution to $1/K \sum_{k=1}^K 1/n \sum_{i \in I_k} \psi(Y_i, D_i, X_i; \breve{\theta}, \widehat{\eta}_k) = 0$ where $\widehat{\eta}_k = (\widehat{\gamma}_k, \widehat{\beta}_k)$ and ψ is defined in (2.2.3).

S5. Compute $\widehat{M} = -1/K \sum_{k=1}^K 1/n \sum_{i \in I_k} (D_i - X_i' \widehat{\beta})^2$. Obtain a variance estimator of $\breve{\theta}$ as $\widehat{\Omega} = \widehat{M}^{-1} \frac{1}{K} \sum_{k=1}^K \frac{1}{n} \sum_{i \in I_k} [\psi(Y, D, X; \breve{\theta}, \widehat{\eta}_k) \psi(Y, D, X; \breve{\theta}, \widehat{\eta}_k)'] (\widehat{M}^{-1})'$.

of dependent factors in the score. Our combined use of the orthogonal score (2.2.3) and this cross-fitting method allows for the high-level theory of the double/debiased machine learning (DML, Chernozhukov et al., 2018) to be applicable. Section 2.4.2 discusses an alternative algorithm that does not rely on the cross fitting at the cost of an additional assumption. Second, the coordinates $\{\widehat{j}_1,...,\widehat{j}_p\}$ are ranked in Step 2 (a)–(c) in the order of decreasing importance after successive orthogonalization using OGA as in Ing (2020). Third, a subset $\widehat{J}_{\widehat{m}} = \{\widehat{j}_1,...,\widehat{j}_{\widehat{m}}\}$ of the ordered set $\{\widehat{j}_1,...,\widehat{j}_p\}$ is selected in Step 2 (d) using HDAIC as in Ing (2020). Our combined use of these three elements (DML, OGA, and HDAIC) together allows for a novel root N consistent estimation of θ_0 without assuming traditional functional class restrictions (e.g., the sparsity) required by existing popular estimators (e.g., LASSO). In Section 2.2.4, we formally present theoretical arguments in support of this claim.

We remark that the proposed method is not scale-invariant like LASSO and unlike OLS. This feature of the method causes a practical issue that estimates change as a one changes scales and units of covariates. As in Ing (2020), we suggest standardizing the covariates to mitigate this drawback. In case where the high-dimensional modeling is used for polynomial approximation of nonparametric functions, we recommend normalizing each basis element by the maximum absolute value. We implement this normalization for our empirical application in Section 2.5.

2.2.3 Tuning Parameters

While Algorithm 2 provides nearly full details of the proposed method, it omits a couple of details. Specifically, Step 2 (d) on HDAIC uses two tuning parameters, C^* and M_n^*. This section provides details about these two tuning parameters. Let c_1, c_2 be sufficiently large positive constants that satisfy

$$P\left(\max_{1\le j\le p}\left|\frac{1}{N}\sum_{i=1}^{N}X_{ij}\varepsilon_i\right| \ge c_1\sqrt{\frac{\log p}{N}}\right) = o(1)$$

and

$$P\left(\max_{1\le j,\ell\le p}\left|\frac{1}{N}\sum_{i=1}^{N}X_{ij}X_{i\ell} - E[X_{1j}X_{1\ell}]\right| \ge c_2\sqrt{\frac{\log p}{N}}\right) = o(1),$$

for each of $\varepsilon = V$ and $\varepsilon = Y - X'\gamma_0$. We require c_1, c_2 to satisfy the above restrictions with $\varepsilon = V$ for the estimation of the part β_0 of the nuisance parameters and with $\varepsilon = Y - X'\gamma_0$ for the estimation of the part γ_0 of the nuisance parameters.

Define $\Gamma(J) = E[X_{iJ}X_{iJ}']$,

$$M_N^* = \min\left\{\left(\frac{N}{\log p}\right)^{1/2\alpha}, \overline{\delta}\left(\frac{N}{\log p}\right)^{1/2}\right\}, \tag{2.2.4}$$

and

$$\overline{\tau} = \sup\left\{\tau : \tau > 0, \limsup_{N\to\infty} \frac{\tau c_2}{\min_{|J|\leq \tau(N/\log p)^{1/2}} \lambda_{\min}(\Gamma(J))} \leq 1\right\},$$

where $0 < \overline{\delta} < \min\{\overline{\tau}, \overline{C}\}$ with an arbitrary strictly positive constant $\overline{C}$ restricted in Assumption 2 (b). In our simulation and real data analysis, we set $\overline{\delta} = 5$ following Ing (2020). Let $\overline{B}$ be a positive constant satisfying

$$\overline{B} > \frac{1}{\liminf_{N\to\infty} \min_{|J|\leq \overline{\delta}(N/\log p)^{1/2}} \lambda_{\min}(\Gamma(J)) - c_2\overline{\delta}}.$$

Define a sufficiently large positive constant C^* satisfying

$$C^* > \frac{2\overline{B}(c_1^2 + c_2^2)}{\sigma_\varepsilon^2}, \tag{2.2.5}$$

for each of $\sigma_\varepsilon^2 = E[V^2]$ and $\sigma_\varepsilon^2 = E[(Y - X'\gamma_0)^2]$. In our main simulations and real data analysis, we set $C^* = 2$ following Ing (2020). We require C^* to satisfy this restriction with $\sigma_\varepsilon^2 = E[V^2]$ for estimation of the part β_0 of the nuisance parameters and with $\sigma_\varepsilon^2 = E[(Y - X'\gamma_0)^2]$ for estimation of the part γ_0 of the nuisance parameters.

In fact, $C^* = 2$ can be considered as an intrinsic number associated with the conventional AIC, rather than a tuning parameter. The conventional information criteria are written in terms of our notations as $\log(\widehat{\sigma}^2) + C^{IC}|\widehat{J}_m|/n$, where the penalty term C^{IC} is 2 for the AIC and it is $\log n$ for the BIC. In this regard, one can think of $C^* = 2$ as that intrinsic number'2' of the AIC. Since it strengthens the penalty on the number of covariates, a larger value of C^* selects a smaller number of covariates.

In Online Supplementary Appendix B.5.1, we run simulations across various values of C^*. The results are optimal and robust around $C^* = 2$ – see Table B.1. In this table, we can also see consequences of choosing too small or too large values of C^*. One may also choose C^* in a data-driven manner. For instance, C^* can be chosen to minimize the prediction loss – see Ing (2020).

2.2.4 The Theory

This section presents and discusses assumptions under which one can conduct an inference about $\breve{\theta}$ based on root-N asymptotic normality using the method described in Algorithm 2. We use the notations $c, C, \overline{C}, \overline{\tau}$ and q for strictly positive constants such that their values can differ depending on the location. Let $q > 4$ be a positive integer, c_q, C_q, λ_1 be some positive constants and $K_{N,q}$ be a positive sequence of constants such that $K_{N,q} \geq E[\max_{1\leq j\leq p} |X_{ij}|^q]$. Wherever there is no risk of confusion, we also use the generic notation ξ to refer to both β and γ to avoid repetitions. All the

random variables and parameter vectors are N-dependent unless otherwise specified. We abbreviate the N index for brevity.

Assumption 1. For each $N \in \mathbb{N}$, it holds that

(a) $(Y_i, D_i, X_i')_{i=1}^N$ are i.i.d. copies of (Y, D, X').

(b) (2.2.1) and (2.2.2) hold.

(c) $E[|Y|^q] + E[|D|^q] \leq C_q$.

(d) $E[|UV|^2] \geq c_q^2$ and $E[V^2] \geq c_q$.

(e) $\max_{1 \leq j \leq p} E[|X_{ij}|^q] \leq C_q$, $E[|V|^q] \leq C_q$, and $E[|U|^q] \leq C_q$.

Furthermore, it holds asymptotically that (f) $K_{N,q}^2 \log p / N^{1-2/q} = o(1)$.

Assumption 1 (a) requires a random sampling of data. Assumption 1 (b) requires that the correct model is given by (2.2.1) and (2.2.2). Assumption 1 (c)–(e) requires bounded moments of various variables. Assumption 1 (f) requires constraints on the speed at which the dimensionality as well as the maximal of the covariate vector can grow. Vectors consist of independent subgaussian random variables with bounded variances, for example, are special cases satisfying this restriction, as the expectation of the maximum of X_{ij} is bounded by a factor of $\sqrt{\log p}$. We emphasize that these assumptions are mild in comparison with the counterpart assumptions made in the high-dimensional regression literature.

Assumption 2. It holds over $N \in \mathbb{N}$ that

(a) $\lambda_{\min}(\Gamma) \geq \lambda_1 > 0$ and $\lambda_{\max}(\Gamma) \leq C_q$, where $\Gamma = E[XX']$.

(b) Define $\Gamma(J) = E[X_{iJ}X_{iJ}']$ and $d_\ell(J) = E[X_{i\ell}X_{iJ}]$ for a set of coordinate indices $J \subseteq \mathfrak{P}$. Then

$$\max_{1 \leq |J| \leq \overline{C}(N/\log p)^{1/2}, \ell \notin J} \left|\Gamma^{-1}(J) d_\ell(J)\right| < C_q.$$

Assumption 2 (a) requires that the minimum eigenvalue $\lambda_{\min}(\Gamma)$ of the Gram matrix Γ to be positive, and it is a standard restriction in the high dimensional literature. Although it permits a wide range of correlations, we note that it has its limitations in high-dimensional settings as it rules out sequences of increasing number of covariates whose correlations approach one. Assumption 2 (b) is a restriction on the covariance structure of X_{iJ}. Observe that $\Gamma^{-1}(J)d_\ell(J)$ takes the form of regression coefficient of $X_{i\ell}$ on X_{iJ}, and so Assumption 2 (b) means that $X_{i\ell}$ cannot be

strongly correlated with X_{iJ} for $\ell \notin J$. We remark that these conditions are imposed at the population level; unlike in the LASSO or Dantzig selector (Candes and Tao, 2007), a restricted eigenvalue type condition for the sample Gram matrix (see e.g., Bickel et al., 2009) is not required here – see also the discussion in Ing (2020, Sec. 3.2).

The following assumption imposes restrictions on the function classes in terms the parameters β_0 and γ_0. We will use the generic notation ξ_0 to refer to β_0 and γ_0. Note that $\xi_0 \in \mathbb{R}^p$ in both cases. Define $\xi(J) = (\xi_j)_{j\in J}$ to be a $|J| \times 1$ vector, where recall that ξ is a generic notation to refer to β_0 and γ_0.

Assumption 3. It holds over $N \in \mathbb{N}$ that for each of $\xi_0 = \beta_0$ and γ_0, ξ_0 follows either (a) or (b) described below.

(a) Polynomial decay: $\log p = o(N^{1-2/q})$. Each ξ_0 is such that $\|\xi_0\|_2^2 \leq C_0$ for some $C_0 > 0$ and there exist $\alpha > 1$ such that for any $J \subseteq \mathfrak{P}$,

$$\|\xi_0(J)\|_1 \leq C\left(\|\xi_0(J)\|_2^2\right)^{(\alpha-1)/(2\alpha-1)}.$$

(b) Exponential decay: $\log p = o(N^{1/4})$. Each ξ_0 is such that $\|\xi_0\|_\infty \leq C_0$ for some $C_0 > 0$ and there exists $C_1 > 1$ such that for any $J \subseteq \mathfrak{P}$,

$$\|\xi_0(J)\|_1 \leq C_1 \|\xi_0(J)\|_\infty.$$

This is a key assumption in this paper, and defines admissible function classes for the high-dimensional linear models. While the literature on LASSO requires the exact sparsity and the L^p sparsity (including approximate sparsity) conditions, Assumption 3 does not impose such conditions. We remark that, if we rearrange the components of the parameter vector ξ_0 by their absolute values in a descending order (denote it again as ξ_0 with an abuse of notation), then Condition (a) contains special cases such as the conventional polynomial decay condition

$$Lj^{-\alpha} \leq |\xi_{0j}| \leq Uj^{-\alpha}, \quad 0 < L \leq U < \infty,$$

as well as the polynomial summability condition

$$\sum_{j=1}^{p} |\xi_{0j}|^{1/\alpha} < M, \quad M \in (0,\infty),$$

following the discussion in Ing (2020, pp. 1962). On the other hand, Condition (b) implies the conventional exponential decay condition that, for some $\alpha' > 0$,

$$L'\exp(-\alpha' j) \leq |\xi_{0j}| \leq U'\exp(-\alpha' j), \quad 0 < L' \leq U' < \infty,$$

so long as the regressors have bounded second moments. Hence throughout the paper, Conditions (a) and (b) are referred to as the polynomial decay condition and the exponential decay condition, respectively, albeit their extra generality. Clearly, the case of polynomial decay accommodates a larger function class, but we remark that there is a tradeoff in terms of how fast the dimension p can diverge as the sample size N increases.

Following Ing (2020, pp. 1960), we now present a concrete example where our Assumption 3 holds but the sparsity does not. Suppose that

$$Lj^{-\alpha} \leq \left|\Lambda_{0(j)}\sigma_{(j)}\right| \leq Uj^{-\alpha}, \qquad\qquad j = 1,\ldots,p,$$

for some $\alpha > 1$, where $0 < L \leq U < \infty$, and $|\Lambda_{0(1)}\sigma_{(1)}| \geq |\Lambda_{0(2)}\sigma_{(2)}| \geq \cdots \geq |\Lambda_{0(p)}\sigma_{(p)}|$ is a descending reordering of $\{\Lambda_{0,j}\sigma_j\}$ with $\sigma_j^2 = E[X_{ij}^2]$. In this setting, our Assumption 3 holds for the same α, but $\sum_{j=1}^p \left|\Lambda_{0,j}\sigma_j\right|^{1/\gamma}$ is now unbounded as $L(1+\log p) \leq \sum_{j=1}^p \left|\Lambda_{0,j}\sigma_j\right|^{1/\gamma} \leq U(1+\log p)$.

Theorem 1. Let $(\mathscr{P}_N)_{N\in\mathbb{N}}$ be a sequence of sets of DGPs such that Assumptions 1–3 are satisfied on the model (2.2.1)–(2.2.2). Then, the estimator $\check{\theta}$ satisfies

$$\sqrt{N}\left(\check{\theta} - \theta_0\right) \xrightarrow{d} N(0,\Omega),$$

where $\Omega = (E[V^2])^{-1}E[V^2U^2](E[V^2])^{-1}$. Define $\widehat{M} := -1/K\sum_{k=1}^K 1/n\sum_{i\in I_k}(D_i - X_i'\widehat{\beta})^2$. Then, we can define the variance estimator

$$\widehat{\Omega} = \widehat{M}^{-1}\frac{1}{K}\sum_{k=1}^K \frac{1}{n}\sum_{i\in I_k}[\psi(Y,D,X;\check{\theta},\widehat{\eta}_k)\psi(Y,D,X;\check{\theta},\widehat{\eta}_k)'](\widehat{M}^{-1})'$$

and the confidence regions with significance level $a \in (0,1)$ have uniform asymptotic validity:

$$\sup_{P\in\mathscr{P}_N}\left|P\left(\theta_0 \in \left[\check{\theta} \pm \Phi^{-1}(1-a/2)\sqrt{\widehat{\Omega}/N}\right]\right) - (1-a)\right| = o(1).$$

A proof is provided in Online Supplementary Appendix B.1.1. This theorem guarantees that the estimator $\check{\theta}$ of θ_0 provided by Algorithm 2 converges at the rate of $\sqrt{N}$ and asymptotically follows the normal distribution under Assumption 1–3. Furthermore, the sample-counterpart asymptotic variance estimator constructs an asymptotically valid confidence interval. We emphasize that this result does not rely on the sparsity assumption which is used in the literature on high-dimensional linear models.

2.3 Simulation Studies

In this section, we investigate the finite sample properties of our proposed estimator $\breve{\theta}$ and compare them with those of two existing estimators, namely the LASSO-based DML and random-forest-based DML.[3]

We follow Belloni et al. (2013b) in developing baseline data generating processes (DGPs). The linear regression model is specified by

$$Y = D\theta_0 + X'\Lambda_0 + U,$$

where $\theta_0 = 0.5$ and $p = dim(X) = 500$. Consistently with this specification, data are generated by the system

$$\begin{aligned} Y &= \theta_0(D - X'\beta_0) + X'\gamma_0 + U, & U &\sim N(0,1), \\ D &= X'\beta_0 + V, & V &\sim N(0,1), \end{aligned}$$

where the covariates are in turn generated by $X \sim N(0,\Sigma)$ with $\Sigma_{jk} = (0.5)^{|k-j|}$.

For the high-dimensional nuisance parameters, $\eta_0 = (\gamma_0, \beta_0)$, we set $p = 500$ throughout and consider a couple of alternative designs. In the first design, each of β_0 and γ_0 has ten coordinates taking the value of 1 and $p - 10$ coordinates taking the value of zero, i.e., sparse deign. In the second design, both β_0 and γ_0 decay exponentially. Specifically, the j-th coordinate of each of β_0 and γ_0 is set to e^{-j}. The third design has both β_0 and γ_0 decaying at polynomial rates. Specifically, the j-th coordinate of each of β_0 and γ_0 is set to j^{-2}, $j^{-1.75}$, $j^{-1.5}$, $j^{-1.25}$ and j^{-1} for five sets of simulations. For each of these sets of simulations, we experiment with the two sample sizes $N \in \{500, 1000\}$ with 1000 Monte Carlo iterations.

Table 2.1 summarize simulation results. Displayed are four Monte Carlo simulation statistics for each set of simulations, including the bias, standard deviation (SD), root mean square error (RMSE), and 95% coverage frequency. In the first row group of the table displaying the results the sparse design, both LASSO-based method and our proposed method based on the OGA and HDAIC work well, while that based on Random Forest significantly underperforms. In the second row group of the table displaying the results under the exponential decay, all the three machine learning methods yield desired results both in terms of all the displayed statistics. There are no significant differences across the three methods under this sparse model. In the subsequent row groups of the table displaying the results for the cases of the polynomial decays, however, observe that the performance varies across the three machine learning methods. While our proposed method based on the OGA and HDAIC continues to perform well in terms of all the displayed statistics, the LASSO-based method slightly underperforms and the random-forest-based method significantly underperforms. In

[3] For these two existing methods, we use the R package "DoubleML : Double Machine Learning in R."

$\beta_{0,j}, \gamma_{0,j}$	N	p	Method of Preliminary Estimation	Bias	SD	RMSE	95%
Sparse	500	500	LASSO	0.020	0.044	0.049	0.917
			Random Forest	0.411	0.020	0.412	0.000
			OGA+HDAIC	-0.003	0.045	0.045	0.950
	1000	500	LASSO	0.009	0.031	0.033	0.928
			Random Forest	0.400	0.015	0.400	0.000
			OGA+HDAIC	-0.003	0.032	0.031	0.956
e^{-j}	500	500	LASSO	0.009	0.044	0.045	0.935
			Random Forest	0.009	0.044	0.045	0.952
			OGA+HDAIC	0.001	0.045	0.044	0.948
	1000	500	LASSO	0.004	0.032	0.032	0.933
			Random Forest	0.007	0.031	0.033	0.927
			OGA+HDAIC	-0.003	0.032	0.031	0.958
j^{-2}	500	500	LASSO	0.014	0.044	0.046	0.932
			Random Forest	0.035	0.043	0.056	0.881
			OGA+HDAIC	-0.001	0.045	0.046	0.948
	1000	500	LASSO	0.006	0.031	0.033	0.938
			Random Forest	0.027	0.031	0.042	0.845
			OGA+HDAIC	-0.002	0.032	0.032	0.955
$j^{-1.75}$	500	500	LASSO	0.016	0.044	0.047	0.928
			Random Forest	0.047	0.043	0.063	0.802
			OGA+HDAIC	0.000	0.045	0.046	0.945
	1000	500	LASSO	0.008	0.031	0.033	0.935
			Random Forest	0.037	0.031	0.049	0.773
			OGA+HDAIC	-0.002	0.032	0.032	0.949
$j^{-1.5}$	500	500	LASSO	0.020	0.044	0.049	0.916
			Random Forest	0.068	0.042	0.080	0.635
			OGA+HDAIC	0.002	0.045	0.046	0.941
	1000	500	LASSO	0.011	0.031	0.034	0.920
			Random Forest	0.055	0.030	0.063	0.568
			OGA+HDAIC	-0.001	0.032	0.032	0.945
$j^{-1.25}$	500	500	LASSO	0.028	0.044	0.053	0.892
			Random Forest	0.111	0.041	0.118	0.217
			OGA+HDAIC	0.007	0.044	0.047	0.926
	1000	500	LASSO	0.016	0.031	0.036	0.897
			Random Forest	0.094	0.029	0.099	0.100
			OGA+HDAIC	0.001	0.031	0.032	0.944
j^{-1}	500	500	LASSO	0.042	0.043	0.063	0.826
			Random Forest	0.195	0.037	0.198	0.000
			OGA+HDAIC	0.022	0.044	0.053	0.885
	1000	500	LASSO	0.024	0.031	0.041	0.855
			Random Forest	0.180	0.026	0.182	0.000
			OGA+HDAIC	0.012	0.031	0.035	0.925

Table 2.1: Monte Carlo simulation results. Displayed are Monte Carlo simulation statistics including the bias, standard deviation (SD), root mean square error (RMSE), and 95% coverage frequency.

particular, these differences in the finite-sample performance widen as the degree of polynomial decay becomes smaller, i.e., as the model becomes less sparse. These results demonstrate the relative robustness of the method proposed in this paper under less sparse high-dimensional regression models.

We ran many other sets of simulations and present their results in Online Supplementary Appendix B.5. In particular, Online Supplementary Appendix B.5.1 presents simulation results with various values of the tuning parameters, and demonstrate the robustness of the qualitative patterns observed above. From these results, we recommend to use the method based on the OGA and HDAIC over the two alternative methods for its robust performance across various designs even including the sparse design.

2.4 Extensions

2.4.1 Models with Approximation Errors

Extending the baseline model (2.2.1)–(2.2.2), consider the following partially linear model motivated by Belloni et al. (2013b):

$$Y = D\theta_0 + f(X) + U, \qquad E[U|X,D] = 0, \tag{2.4.1}$$

$$D = g(X) + V, \qquad E[V|X] = 0, \tag{2.4.2}$$

where Y denotes an outcome variable, D denotes a treatment variable, X denotes a p-dimensional vector of controls, and U and V denotes unobserved factors. We do not directly impose any parametric restriction on f or g unlike the baseline model presented in Section 2.2. This extension is useful in certain applications, such as the one we present in Section 2.5. In this semi-parametric framework, we are interested in the partial effect θ_0 of D on Y.

Now consider the followng reduced form regressions for (2.4.1)–(2.4.2):

$$Y = \underbrace{X'\gamma_0 + r_Y(X)}_{f(X)+\theta_0 g(X)} + \mathscr{E}, \qquad E[\mathscr{E}|X] = 0, \tag{2.4.3}$$

$$D = \underbrace{X'\beta_0 + r_D(X)}_{g(X)} + V, \qquad E[V|X] = 0, \tag{2.4.4}$$

where $X'\gamma_0$ and $X'\beta_0$ are approximations to $E[Y|X]$ and $E[D|X]$, and $r_Y(X)$ and $r_D(X)$ are approximation errors. The functions r_Y and r_D are nonparametric as are f and g. We will impose conditions on the magnitudes of r_Y and r_D below. Models under these conditions, along with certain sparsity conditions imposed on β_0 and γ_0, are said to be "approximate sparse" in Belloni et al. (2012, 2013b).

Recall the orthogonal score $\psi(Y,D,X;\theta,\eta)$ defined in (2.2.3). With this orthogonal score, we propose to obtain $\check{\theta}$ and $\widehat{\Omega}$ via Algorithm 2 presented in Section 2.2 even under the current extended setting with approximation errors.

With the extended model (2.4.3)–(2.4.4), a different set of assumptions are imposed from those in the baseline model. First, we slightly modify Assumption 1 as follows.

Assumption 4. For each $N \in \mathbb{N}$, it holds that

(a) $(Y_i, D_i, X_i')_{i=1}^N$ are i.i.d. copies of (Y,D,X').

(b) (2.4.3) and (2.4.4) hold.

(c) $E[|Y|^q] + E[|D|^q] \leq C_q$.

(d) $E[|UV|^2] \geq c_q^2$ and $E[V^2|(Y,D,X')] \geq c_q$.

(e) $\max_{1\leq j\leq p} E[|X_{ij}|^q] \leq C_q$, $E[|V|^q] \leq C_q$, and $E[|\mathscr{E}|^q] \leq C_q$.

Furthermore, it holds asymptotically that (f) $K_{N,q}^2 C \log p / N^{1-2/q} = o(1)$.

In part (d), we require the conditional variance of V given (Y,D,X') to be bounded away from zero whereas the counterpart in the baseline model assumed the unconditional variance to be bounded away from zero.

We continue to use Assumption 2 from the baseline model. However, it should be stressed that we now impose Assumption 2 on (2.4.3)–(2.4.4) rather than (2.2.1)–(2.2.2). With the approximation errors introduced in the current extended model, we make the following assumption on the approximation error functions r_Y and r_D.

Assumption 5. For $r(X) = r_Y(X)$ and $r_D(X)$, it holds that

(a) $E[r^4(X)] \leq C$.

(b) $E[r^2(X)] \leq C \log p / N$.

(c) $\max_{1\leq j\leq p} \left|E[r(X)X_{ij}]\right| \leq C_{p,1}\sqrt{\log p}/N^{1/4}$.

Assumption 5 (a) requires the fourth moment of the approximation error to be bounded, (b) assumes the second moment to be of order $\log p/N$, and (c) bounds the maximum cross moment of the approximation and the covariates.

Finally, we focus on the more difficult case, namely the polynomial decay case, for brevity in this section.

Assumption 6. It holds over $N \in \mathbb{N}$ that for each of $\xi_0 = \beta_0$ and γ_0, ξ_0 follows polynomial decay, i.e., $\log p = o(N^{1-2/q})$. Each ξ_0 is such that $\|\xi_0\|_2^2 \leq C_0$ for some $C_0 > 0$ and there exist $\alpha > 1$ and $C_\alpha > 0$ such that for any $J \subseteq \mathfrak{P}$,

$$\|\xi_0(J)\|_1 \leq C_\alpha \left(\|\xi_0(J)\|_2^2\right)^{(\alpha-1)/(2\alpha-1)}.$$

The following theorem establishes the asymptotic normality of $\check{\theta}$ along with the asymptotic validity of inference under the extended model with approximation errors.

Theorem 2. Let $(\mathscr{P}_N)_{N\in\mathbb{N}}$ be a sequence of sets of DGPs such that Assumptions 2 and 4–6 are satisfied on the model (2.4.1)–(2.4.2) entailing the reduced forms (2.4.3)–(2.4.4). Then, the estimator $\check{\theta}$ defined in Algorithm 2 satisfies

$$\sqrt{N}\left(\check{\theta}-\theta_0\right) \xrightarrow{d} N(0,\Omega),$$

where $\Omega = (E[V^2])^{-1}E[V^2U^2](E[V^2])^{-1}$. The confidence regions with significance level $a \in (0,1)$ have uniform asymptotic validity:

$$\sup_{P\in\mathscr{P}_N}\left|P\left(\theta_0 \in \left[\check{\theta} \pm \Phi^{-1}(1-a/2)\sqrt{\widehat{\Omega}/N}\right]\right)-(1-a)\right| = o(1),$$

where $\widehat{\Omega}$ is defined in Section 2.2.

A proof is presented in Online Supplementary Appendix B.3.1. The same remarks as those presented below the statement of Theorem 1 apply here.

We want to stress that Theorem 2 is not an immediate consequence given Theorem 1, because the original proofs for convergence rates of OGA+HDAIC in Ing (2020) do not permit approximately sparse models. In order to show Theorem 2, we establish convergence rates for the OGA+HDAIC in approximately sparse regression models.

2.4.2 Estimation and Inference without Cross Fitting

Thus far, our proposed procedures of estimation and inference are based on cross fitting. A drawback of using the cross fitting is the randomness of estimates given the data. To overcome this drawback, we provide an alternative procedure of estimation and inference without relying on cross fitting in this section. However, we stress that this benefit comes with costs in some assumptions as discussed below.

We continue from Section 2.4.1 to consider the partial linear model (2.4.1)–(2.4.2) entailing the reduced forms (2.4.3)–(2.4.4). Furthermore, we continue to use the same orthogonal score $\psi(Y,D,X;\theta,\eta)$ defined in (2.2.3). However, we now replace Algorithm 2 by the following algorithm which does not involve the cross-fitting procedure. Let $[N] = \{1,\ldots,N\}$, so that $X_{[N]j} = \{X_{ij}, i \in [N]\}$ and $D_{[N]} = (D_1,\ldots,D_N)'$.

To establish asymptotic properties for this new estimator $\widetilde{\theta}$, we continue to impose Assumptions 2, 4, and 5. As in Section 2.4.1, we focus on the more difficult case, namely the polynomial decay case, for brevity.

Algorithm 3 OGA+HDAIC with DML for high-dimensional linear models without cross fitting

S1. Perform following procedure using $\{(X_i', D_i)'\}_{i=1}^N$ to get $\widehat{\beta}$.

(a) Compute $\widehat{\mu}_{0,j} = X_{[N]j}' D_{[N]} / \sqrt{N}\|X_{[N]j}\|$. Select the coordinate $\widehat{j}_1 = \text{argmax}_{1\le j\le p} |\widehat{\mu}_{0,j}|$. Define $\widehat{J}_1 = \{\widehat{j}_1\}$.

(b) Compute $\widehat{\mu}_{1,j} = X_{[N]j}'(I_N - H_1)D_{[N]} / \sqrt{N}\|X_{[N]j}\|$, where $H_1 = X_{[N]\widehat{j}_1}(X_{[N]\widehat{j}_1}' X_{[N]\widehat{j}_1})^{-1} X_{[N]\widehat{j}_1}'$. Select the coordinate $\widehat{j}_2 = \text{argmax}_{1\le j\le p, j\notin\widehat{J}_1} |\widehat{\mu}_{1,j}|$. Update $\widehat{J}_2 = \widehat{J}_1 \cup \{\widehat{j}_2\}$.

(c) Given $m-1$ coordinates $\widehat{J}_{m-1}$ that have been obtained, compute $\widehat{\mu}_{m-1,j} = X_{[N]j}'(I_N - H_{m-1})D_{[N]} / \sqrt{N}\|X_{[N]j}\|$, where $H_{m-1} = X_{[N]\widehat{J}_{m-1}}(X_{[N]\widehat{J}_{m-1}}' X_{[N]\widehat{J}_{m-1}})^{-1} X_{[N]\widehat{J}_{m-1}}'$. Select the coordinate $\widehat{j}_m = \text{argmax}_{1\le j\le p, j\notin\widehat{J}_{m-1}} |\widehat{\mu}_{m,j}|$. Iteractively update $\widehat{J}_m = \widehat{J}_{m-1} \cup \{\widehat{j}_m\}$.

(d) Compute HDAIC $(\widehat{J}_m) = (1 + C^* |\widehat{J}_m| \log p / N)\widehat{\sigma}_m^2$ for each m and $\widehat{\sigma}_m^2 = 1/ND'(I - H_m)D$. Choose $\widehat{m} = \text{argmin}_{1\le m\le M_n^*} \text{HDAIC}(\widehat{J}_m)$, where C^* and M_n^* are defined in (2.2.5) and (2.2.4) in Section 2.2.3.

(e) With coordinates $\widehat{J}_{\widehat{m}}$, run OLS of D_i on $X_{i\widehat{J}_{\widehat{m}}}$ to get $\widehat{\beta}$.

S2. Repeat S1 with $\{(X_i', Y_i)'\}_{i=1}^N$ instead of $\{(X_i', D_i)'\}_{i=1}^N$, to get $\widehat{\gamma}$.

S3. Obtain $\widetilde{\theta}$ as a solution to $1/N\sum_{i=1}^N \psi(Y_i, D_i, X_i; \widetilde{\theta}, \widehat{\eta}) = 0$ where $\widehat{\eta} = (\widehat{\gamma}, \widehat{\beta})$ and ψ is defined in (2.2.3).

Assumption 7. It holds over $N \in \mathbb{N}$ that $|\theta_0| \le C$, and for each of $\xi_0 = \beta_0$ and γ_0, ξ_0 follows polynomial decay, i.e., each ξ_0 is such that $\|\xi_0\|_2^2 \le C_0$ for some $C_0 > 0$ and there exist $\alpha > 1$ such that $\log p = o(N^{(\alpha-1)/(3\alpha-1)})$ and for any $J \subseteq \mathfrak{P}$,

$$\|\xi_0(J)\|_1 \le C\left(\|\xi_0(J)\|_2^2\right)^{(\alpha-1)/(2\alpha-1)}.$$

Unlike the previous sections, however, we now require $\log p = o(N^{(\alpha-1)/(3\alpha-1)})$ for $\alpha > 1$.

The following theorem establishes the asymptotic normality of $\widetilde{\theta}$ defined without the cross-fitting procedure.

Theorem 3. Let $(\mathscr{P}_N)_{N\in\mathbb{N}}$ be a sequence of sets of DGPs such that Assumptions 2, 4, 5, and 7 are satisfied on the model (2.4.1)–(2.4.2) entailing the reduced forms (2.4.3)–(2.4.4). Then, the estimator $\widetilde{\theta}$ satisfies

$$\sqrt{N}\left(\widetilde{\theta} - \theta_0\right) \xrightarrow{d} N(0, \Omega),$$

where $\Omega = (E[V^2])^{-1} E[V^2U^2](E[V^2])^{-1}$.

A proof is given in Online Supplementary Appendix B.3.2.

We stress that, although the proof builds on that of Theorem 1 in Belloni et al. (2013b), it is far from being trivial as the lack of cross-fitting and L^p-sparsity creates extra challenges. Specifically, a key intermediate step is to control the L^1 distances between β_0 and $\widetilde{\beta}(\widetilde{J})$, an oracle regression estimator defined in the proof of Theorem 3. Due to the lack of

exact or approximate sparsity of β_0, this is shown via different strategies from those employed in Belloni et al. (2013b).

As emphasized at the beginning of the current subsection, the main advantage of the estimation procedure without cross fitting is that the estimate is now non-random given data. Besides, this framework without cross fitting offers an additional advantage. Recall that our main motivation to use the OGA+HDAIC is to weaken the sparsity assumptions required for conventional high-dimensional methods such as the LASSO. In the current framework without cross fitting, there is another motivation to use the OGA+HDAIC. Namely, it selects regressors based on their strength in explanatory power by the algorithm. Hence, we have better interpretations of the model selected by the OGA+HDAIC than the conventional high-dimensional methods under the current framework without cross-fitting. We highlight this additional advantage of our proposed method.

Online Supplementary Appendix B.5.2 presents simulation results with this modified method without cross fitting. The results are similar to those obtained for the baseline model presented in Section 2.3.

2.5 An Empirical Application

In this section, we demonstrate an application of the proposed method to estimation of production functions. The main challenge in the econometrics of production functions is the simultaneity in the choice of input firms (Marschak and Andrews, 1944). While early studies of production functions address this simultaneity problem by explicitly modeling rational choice structures of firms, Olley and Pakes (1996) more recently propose a novel idea to use the inverse of the reduced-form investment choice function as a control function. Levinsohn and Petrin (2003) propose to use intermediate input, instead of investment, as a control variable for a number of advantages.

The use of the control function *a la* Levinsohn and Petrin (2003) entails the partial linear estimating equation for the labor elasticity of the form

$$y_{it} = \ell_{it}\theta + f(k_{it}, m_{it}) + u_{it}, \tag{2.5.1}$$

where y_{it} denotes the logarithm of output, ℓ_{it} denotes the logarithm of labor input, k_{it} denotes the logarithm of capital input, m_{it} denotes the logarithm of intermediate input, g is a nonparametric function that subsumes a part of the production function and the control function, and u_{it} denotes a mean-orthogonal reduced-form composite error. See Olley and Pakes (1996) and Levinsohn and Petrin (2003) for details.

In light of the partial linear form (2.5.1), Olley and Pakes (1996) and Levinsohn and Petrin (2003) propose to use the estimator of Robinson (1988) which is semiparametric root-n consistent for θ. Following these seminal papers, numerous researchers have estimated production functions. That said, many of these subsequent studies follow the Stata command (Petrin et al., 2004) which implements estimation of (2.5.1) via the parametric third-degree polynomial

approximation

$$y_{it} = \ell_{it}\theta + \sum_{\rho_1=0}^{3}\sum_{\rho_2=0}^{3-\rho_1} \delta_{\rho_1\rho_2} k_{it}^{\rho_1} m_{it}^{\rho_2} + u_{it}. \tag{2.5.2}$$

See Petrin et al. (2004, pages 116–118).

To mitigate the approximation bias asymptotically, we consider a higher-dimensional approximation

$$y_{it} = \ell_{it}\theta + \underbrace{\sum_{j=1}^{p} \delta_j \phi_j(k_{it}, m_{it}) + r_p(k_{it}, m_{it})}_{f(k_{it}, m_{it})} + u_{it} \tag{2.5.3}$$

with an error $r_p(k_{it}, m_{it})$ in approximation, where $\phi = (\phi_1, \phi_2, \phi_3, \cdots)$ is a basis and p can be large and increasing with the sample size. The basis ϕ could be defined as the Cartesian product of polynomials, i.e., $(\phi_1(k,m), \phi_2(k,m), \phi_3(k,m), \cdots) = (1, k, m, k^2, m^2, km, \cdots)$, as a generalization of the popular estimating equation (2.5.2) in the Stata command. More generally, we can define the basis ϕ as the tensor product of orthonomal bases. We employ the tensor product of polynomial bases for our basis ϕ, and apply our proposed method to (2.5.3) to get an estimate of θ and its standard error.

Following Levinsohn and Petrin (2003), we use a plant-level panel of Chilean firms from 1979 to 1986. See Liu (1991) for details about the construction of the data. Among others, we focus on the 3-digit level industry of food products (311) because of its large sample size compared to other industries. We are interested in the elasticity with respect to unskilled labor input ℓ_{it}^u and skilled labor input ℓ_{it}^s. The intermediate input variables include electricity m_{it}^e, fuels m_{it}^f, and materials m_{it}^m. To estimate the elasticity with respect to unskilled labor input ℓ_{it}^u using m_{it}^m as a proxy following Levinsohn and Petrin (2003), we consider the estimating equation of the form

$$\underbrace{y_{it}}_{Y} = \underbrace{\ell_{it}^u \theta^u}_{D\theta_0} + \underbrace{\ell_{it}^s \theta^s + m_{it}^e \theta^e + m_{it}^f \theta^f + \sum_{j=1}^{p} \delta_j \phi_j(k_{it}, m_{it}^m) + \tau_t + r_p(k_{it}, m_{it}^m)}_{f(X)} + \underbrace{u_{it}}_{U} \tag{2.5.4}$$

as in (2.4.1). To estimate the elasticity with respect to skilled labor input ℓ_{it}^s, we swap $\ell_{it}^u \theta^u$ and $\ell_{it}^s \theta^s$ in the above estimating equation:

$$\underbrace{y_{it}}_{Y} = \underbrace{\ell_{it}^s \theta^s}_{D\theta_0} + \underbrace{\ell_{it}^u \theta^u + m_{it}^e \theta^e + m_{it}^f \theta^f + \sum_{j=1}^{p} \delta_j \phi_j(k_{it}, m_{it}^m) + \tau_t + r_p(k_{it}, m_{it}^m)}_{f(X)} + \underbrace{u_{it}}_{U} \tag{2.5.5}$$

as in (2.4.1).

The term τ_t represents time effects. Following Levinsohn and Petrin (2003), we include the indicator for year groups

		Unskilled Labor	Skilled Labor
(I)	Levinsohn and Petrin (2003)	0.139 (0.010)	0.051 (0.009)
(II)	Double Machine Learning with LASSO Preliminary Estimation	0.170 (0.011)	0.063 (0.008)
(III)	Double Machine Learning with Random Forest Preliminary Estimation	0.185 (0.013)	0.062 (0.010)
(IV)	Double Machine Learning with OGA+HDAIC Preliminary Estimation	0.168 (0.011)	0.060 (0.010)

Table 2.2: Estimates of labor elasticities in the 3-digit level industry of food products (311) in Chile.

1979–1981, 1982–1983, and 1984–1986.

For estimation of (2.5.4) using a polynomial basis, we let X consist of (i) ℓ_{it}^s (ii) m_{it}^e, (iii) m_{it}^f, (iv) $k_{it}, \ldots, k_{it}^{10}$, (v) $m_{it}^m, \ldots, (m_{it}^m)^{10}$, (vi) dummy for 1979–1981, (vii) dummy for 1982–1983, and (viii) interactions of the terms in (iv) and (v). We use a finite-sample adjusted version of the DML estimates following Chernozhukov et al. (2018, Sec. 3.4) – see Online Supplementary Appendix B.4 for details. We repeat an analogous estimation procedure for (2.5.4).

Table 2.2 summarizes estimation results. Row (I) copies estimates from Levinsohn and Petrin (2003). Rows (II), (III), and (IV) report results based on the DML with LASSO, DML with random forest, and DML with the OGA and HDAIC (the estimator proposed in this paper), respectively.[4]

First, observe that all the three machine learning estimates, (II), (III), and (IV), yield larger point estimates than the low-dimensional estimates (I). This may indicate a potential bias of the conventional estimator based on a low-dimensional polynomial approximation. For the unskilled labor coefficient, the estimate based on random forest (III) is even higher than the other two machine learning estimates, (II) and (IV). This may be imputed to the larger bias of the random forest we observed through the simulation studies in Section 2.3. For the skilled labor coefficient, the three machine learning methods (II)–(IV) yield similar values. Our proposed method based on OGA+HDAIC yields slightly smaller estimates than (II) and (III).

We ran several other estimates for robustness checks. Online Supplementary Appendix B.6 presents estimation results based on alternative values of the tuning parameter. Online Supplementary Appendix B.6 also presents results based on the method without cross fitting introduced in Section 2.4.2.

Finally, we conclude this section with a few remarks about the validity of the estimation approach employed in this empirical application following Olley and Pakes (1996) and Levinsohn and Petrin (2003). It is well known today that the estimating equation (2.5.1) fails to identify the parameter θ in general, as first pointed out by Ackerberg et al.

[4]For (II) and (III), we use the R package "DoubleML : Double Machine Learning in R." We set the parameters as folds = 10, num.trees = 100, min.node.size = 2, max.depth = 5, and the number of repetitions = 20

(2015). That said, they also suggest that θ can be correctly identified by (2.5.1) under certain DGPs. They include DGPs with: (1) i.i.d. optimization error in ℓ_{it} and not in m_{it}; or (2) i.i.d. shocks to the price of labor or output after m_{it}; for instance (Ackerberg et al., 2015, Sec. 3.1). As such, we stress that the validity of the estimation method present above is contingent on these assumptions about the underlying DGPs.

2.6 Summary and Discussions

In this paper, we propose a new method of inference in high-dimensional regression models. The estimation procedure is based on a combined use of the OGA, HDAIC, and DML. The method of inference about any low-dimensional subvector of high-dimensional parameters is based on a root-N asymptotic normality, which does not require the exact sparsity condition or the L^p sparsity condition. In stead imposed are conditions on the rate at which the absolute size of parameters decays when descendingly ordered. We demonstrate through simulation studies superior finite sample performance of this proposed method over those based on two popular alternatives, namely the LASSO and the random forest. The extent of this outperformance is more prominent under less sparse models characterized by slower polynomial decays. Finally, we illustrate an application of the method to production analysis using a panel of Chilean firms.

We close this paper with discussions of limitations, omitted extensions and potential directions for future research. First, unlike regressions and like the LASSO, the method is not invariant to invertible linear transformation of the regressors X. Practitioners should be aware of this drawback in our proposed method. Second, as is the case with other DML methods, our proposed method based on DML is subject to random estimates. To overcome this problem, we present an alternative procedure without cross fitting in Section 2.4.2, but this comes at the expense of an alternative set of assumptions. Again, practitioners should be aware of these tradeoffs in choosing an appropriate method. Third, we focus on high-dimensional linear regression models with exogenous regressors throughout the main text. We provide an extension to high-dimensional linear IV models in Section B.2. Extensions to other important models are left for future research.

CHAPTER 3

Bounds for Standard Errors from Interdependent Data

This chapter is coauthored with Yuya Sasaki.

3.1 Introduction

Researchers often combine empirical moments from multiple data sources that are interdependent. Such practices include combining survey data with administrative data, cross-sectional data with time-series data, and time series with different frequencies. In the context of meta-analysis and calibration, there is often a greater focus on parameter estimation than inference. This tendency may be imputed to the impossibility of conventional inference, as the covariance among the empirical moments from multiple data sets may not be available, even though the marginal variances are available.

While the covariance is not available, its bounds are feasible to compute from marginal variances. In recent work, Cocci and Plagborg-Møller (2024) propose an upper bound of the covariance computed from marginal variances based on Cauchy-Schwarz inequality. We show that this idea can be generalized based on the best-possible distributional bounds of Frank et al. (1987). Consequently, we propose a lower bound, as well as an upper bound, of the standard error for the parameter of interest. Furthermore, we formally account for the finite-sample randomness of estimated marginal variances in developing our bounds.

The problem of inference under data combination often boils down to identifying the distribution of the sum of dependent variables, given the marginal distributions of the variables to be summed. This question was first raised by A.N. Kolmogorov and answered by his student, G.D. Makarov (1982), for the case of the sum of two random variables. The solution is later generalized by Frank et al. (1987) to deal with more than two variables. We take advantage of these best-possible distributional bounds derived by Frank et al. (1987).

There exists an extensive body of literature on data combinations. To our knowledge, the existing literature mostly focuses on identification and estimation – see the handbook chapter by Ridder and Moffitt (2007). The study of asymptotic variance and standard errors appears a new topic in this literature, and we are only aware of Cocci and Plagborg-Møller (2024) for reference. Although this paper is not about identification, also related in terms of bounds is the broad literature on partial identification – see the review by Tamer (2010).

The rest of this paper is organized as follows. Section 3.2 introduces the framework. Section 3.3 presents our proposed method. Section 3.4 provides supporting theories. Section 3.5 presents numerical illustrations both with

simulated data and real data. Section 3.6 summarizes the paper. Mathematical proofs are collected in the appendix.

3.2 The Framework

Suppose that we are interested in a scalar parameter $\varphi(\theta)$ for a known function $\varphi : \mathbb{R}^k \to \mathbb{R}$ of the k-dimensional vector θ of structural parameters, which in turn produces a p-dimensional vector $\mu = h(\theta)$ of reduced-form parameters.

If estimates $\widehat{\mu} = (\widehat{\mu}_1, \ldots, \widehat{\mu}_p)'$ of the reduced-form parameters μ are available, then the standard procedure of inference about $\varphi(\theta)$ is based on the linear approximation

$$\varphi(\widehat{\theta}) - \varphi(\theta) = \ell'(\widehat{\mu} - \mu) + o_p(n^{-1/2}), \qquad \ell = \frac{\partial \phi}{\partial \theta'}(\theta)\frac{\partial h^{-1}}{\partial \mu}(\mu), \tag{3.2.1}$$

where $\widehat{\theta} = h^{-1}(\widehat{\mu})$, $\partial\varphi(\theta)/\partial\theta'$ is a $1 \times k$ vector, and $\partial h^{-1}(\mu)/\partial\mu$ is a $k \times p$ matrix with its (i,j)-th element given by $\partial h_i^{-1}(\mu)/\partial\mu_j$.

In particular, if we have the standard convergence in distribution

$$\sqrt{n}(\widehat{\mu} - \mu) \xrightarrow{d} N(0, \Sigma),$$

then we usually apply the delta method to obtain

$$\sqrt{n}(\varphi(\widehat{\theta}) - \varphi(\theta)) \xrightarrow{d} N(0, \sigma^2), \qquad \text{where} \quad \sigma^2 = \ell'\Sigma\ell. \tag{3.2.2}$$

This convergence in distribution paves the way for inference about $\varphi(\theta)$ provided that the asymptotic variance σ^2 were estimable.

However, if the reduced-form estimates, $\widehat{\mu}_1, \ldots, \widehat{\mu}_p$, stem from different samples, then not all of the elements of the asymptotic covariance matrix Σ are estimable. To fix ideas, suppose that $\widehat{\mu}_j$ is generated by the sample $X_j = (X_{j,1}, \ldots, X_{j,n_j})$, where X_j is separately observed across the indices $j = 1, \ldots, p$. In this case, we can only estimate the diagonal elements

$$s_j^2 := \Sigma_{jj}$$

for $j = 1, \ldots, p$. With the off-diagonal elements of Σ unknown, however, it is not feasible to estimate the asymptotic variance $\sigma^2 = \ell'\Sigma\ell$ for the parameter $\varphi(\theta)$ of interest.

Example. Many economic questions can be stated in this framework of data combination. For example, consider the

moment-matching estimator

$$\widehat{\theta} = \arg\min_{\theta \in \Theta} (\widehat{\mu} - h(\theta))' \widehat{W} (\widehat{\mu} - h(\theta))$$

for a compact parameter set Θ. We can estimate the loadings ℓ by $\widehat{\ell} = \widehat{W}\widehat{G}(\widehat{G}'\widehat{W}\widehat{G})^{-1}$ with $\widehat{G} \equiv \partial h(\widehat{\theta})/\partial \theta'$, but we cannot estimate Σ if the moments $\widehat{\mu}_1, \ldots, \widehat{\mu}_p$ come from separate samples $X_1, \ldots, X_p$. Consequently, the asymptotic variance $\sigma^2 = \ell' \Sigma \ell$ is not estimable. ▲

In the rest of the paper, we propose a method to obtain both upper and lower bounds of the asymptotic variance σ^2 of interest, given only marginal variances s_j^2 for $j = 1, \ldots, p$. We can accommodate heterogeneous sample sizes $n_1, \ldots, n_p$ as far as they are asymptotically proportional. For simplicity of writing and clear exposition, however, we consider homogeneous sample sizes $n := n_1 = \cdots = n_p$ throughout.

3.3 The Method

We overview our proposed method in this section. Formal theoretical justifications will follow in Section 3.4. Our proposal takes advantage of the following useful lemma, which we cite from probability theory.

Lemma 3.3.1 (Frank et al. (1987))**.** *Let $F_j(t)$ denote the CDF of a random variable ξ_j for $j = 1, \ldots, p$. The CDF, F, of the sum $\xi = \sum_{j=1}^{p} \xi_j$ is bounded as*

$$\underline{F}(x) \leq F(x) \leq \overline{F}(x) \qquad \text{for all } x,$$

where the lower and upper bound functions are given by

$$\underline{F}(x) = \max\left\{ \sum_{j=1}^{p-1} F_j(x_j) + F_p\left(x - \sum_{j=1}^{p-1} x_j \right) - (p-1),\ 0 \right\} \qquad \text{and}$$
$$\overline{F}(x) = \min\left\{ \sum_{j=1}^{p-1} F_j(x_j) + F_p\left(x - \sum_{j=1}^{p-1} x_j \right),\ 1 \right\}.$$

They propose the best-possible distributional bounds as the title of their article suggests. For any distributional bounds, there will always be a copula such that the true distribution of the sum of the variables meets the bound at a given point. In other words, one cannot construct bounds any tighter (Williamson and Downs, 1990).

In our framework introduced in Section 3.2, we let ξ_j denote the random variable drawn from the limit marginal distribution F_j of $\sqrt{n}\ell_j(\widehat{\mu}_j - \mu_j)$ for each $j = 1, \ldots, p$, and let ξ denote the random variable drawn from the limit distribution F of $\sqrt{n}(\varphi(\widehat{\theta}) - \varphi(\theta))$. With these definitions, the linear approximation (3.2.1) yields $\xi = \sum_{j=1}^{p} \xi_j$ as in the statement of Lemma 3.3.1.

Note that we can explicitly write F as

$$F(t) = \Phi\big(t/\sigma\big)$$

by (3.2.2), where $\Phi(t)$ denotes the CDF of the standard normal distribution. Similarly, we can write F_j as

$$F_j(t) = \Phi\big(t/(\ell_j s_j)\big)$$

for each $j = 1,\ldots,p$.

Motivated by Lemma 3.3.1, we define the lower and upper bounds of the limit distribution F of interest in terms of the marginal distributions, $F_1,\ldots,F_p$, by

$$\begin{aligned}
\underline{F}(v) &= \max\left\{\sum_{j=1}^{p-1} F_j(x_j) + F_p\left(v - \sum_{j=1}^{p-1} x_j\right) - (p-1),\ 0\right\} \qquad \text{and} \\
\overline{F}(v) &= \min\left\{\sum_{j=1}^{p-1} F_j(x_j) + F_p\left(v - \sum_{j=1}^{p-1} x_j\right),\ 1\right\},
\end{aligned}$$

respectively. With these definitions, Lemma 3.3.1 implies

$$\underline{F}(v) \le F(v) \le \overline{F}(v) \tag{3.3.1}$$

for all $v \in \mathbb{R}$.

These bounds are infeasible to construct because we do not know the limit distributions $F_1,\ldots,F_p$. In other words, we do not know $\ell_1 s_1,\ldots,\ell_p s_p$. Instead, suppose that we have estimates $\widehat{\ell}_1\widehat{s}_1,\ldots,\widehat{\ell}_p\widehat{s}_p$ of them. We can then construct a feasible counterpart of $F_j(t)$ by

$$F_{n,j}(t) = \Phi\big(t/(\widehat{\ell}_j\widehat{s}_j)\big)$$

for each $j = 1,\ldots,p$.

Using these estimates of the marginal distributions, we in turn define the feasible counterparts of the bounds, $\underline{F}(v)$ and $\overline{F}(v)$, by

$$\underline{F}_n(v) = \max\left\{\sum_{j=1}^{p-1} F_{n,j}(x_j) + F_{n,p}\left(v - \sum_{j=1}^{p-1} x_j\right) - (p-1), 0\right\} \qquad \text{and}$$

$$\overline{F}_n(v) = \min\left\{\sum_{j=1}^{p-1} F_{n,j}(x_j) + F_{n,p}\left(v - \sum_{j=1}^{p-1} x_j\right), 1\right\},$$

respectively.

Because $F_{n,1}, \ldots, F_{n,p}$ are random, however, these bounds may fail to bound F in a finite sample with high probability. We, therefore, augment these feasible bounds by

$$\underline{F}_n^{\delta}(v) = \underline{F}_n(v) - \underline{\delta}_n(v) \qquad \text{and} \qquad \overline{F}_n^{\delta}(v) = \overline{F}_n(v) + \overline{\delta}_n(v), \tag{3.3.2}$$

for some buffer functions $\underline{\delta}_n(v)$ and $\overline{\delta}_n(v)$. In practice, we may use

$$\begin{aligned}
\underline{\delta}_n(v) &= \underline{\delta}'_n(v) \cdot n^{-1/2.1} \quad \text{where} \quad \underline{\delta}'_n(v) = 0.001 \cdot \left(0.001 + \sqrt{\underline{F}_n(v)(1-\underline{F}_n(v))}\right), \\
\overline{\delta}_n(v) &= \overline{\delta}'_n(v) \cdot n^{-1/2.1} \quad \text{where} \quad \overline{\delta}'_n(v) = 0.001 \cdot \left(0.001 + \sqrt{\overline{F}_n(v)(1-\overline{F}_n(v))}\right).
\end{aligned}$$

The formal theory presented in Section 3.4 supports this construction of the bounds.

Consequently, the lower and upper bounds of the asymptotic standard deviation $\sigma = \sqrt{\ell' \Sigma \ell}$ for $\varphi(\widehat{\theta})$ of interest are given by

$$\underline{\sigma} = \inf\left\{\sigma \in \mathbb{R}_+ : \underline{F}_n^{\delta}(v) \leq \Phi(v/\sigma) \leq \overline{F}_n^{\delta}(v)\right\} \qquad \text{and} \tag{3.3.3}$$

$$\overline{\sigma} = \sup\left\{\sigma \in \mathbb{R}_+ : \underline{F}_n^{\delta}(v) \leq \Phi(v/\sigma) \leq \overline{F}_n^{\delta}(v)\right\}, \tag{3.3.4}$$

respectively.

Example. To illustrate, suppose that we are interested in $\varphi(\theta) = \theta = \mu_1 + \mu_2$. Let $F_{n,1}(t) = \Phi(t)$ and $F_{n,2}(t) = \Phi(t/2)$ be estimated CDFs of the limit distributions of $\sqrt{n}(\widehat{\mu}_1 - \mu_1)$ and $\sqrt{n}(\widehat{\mu}_2 - \mu_2)$, respectively. The lower and upper bounds for the CDF, F, of $\sqrt{n}(\varphi(\widehat{\theta}) - \varphi(\theta))$ are illustrated by $\underline{F}_n^{\delta}(v)$ and $\overline{F}_n^{\delta}(v)$, respectively, in the left panel of Figure 3.1. The lower and upper bounds of the asymptotic standard deviations can be then computed by (3.3.3) and (3.3.4), respectively. The CDFs of $F_{\underline{\sigma}} = \Phi(t/\underline{\sigma})$ and $F_{\overline{\sigma}} = \Phi(t/\overline{\sigma})$ are shown in the right panel of Figure 3.1. Observe that $\underline{\sigma}$ is characterized by the normal CDF with the sharpest slope between the bounds. In general, we cannot rule out the possibility of a non-informative lower bound $\underline{\sigma} = 0$. ▲

Discussion. In a closely related framework, Cocci and Plagborg-Møller (2024) propose the worst case standard error

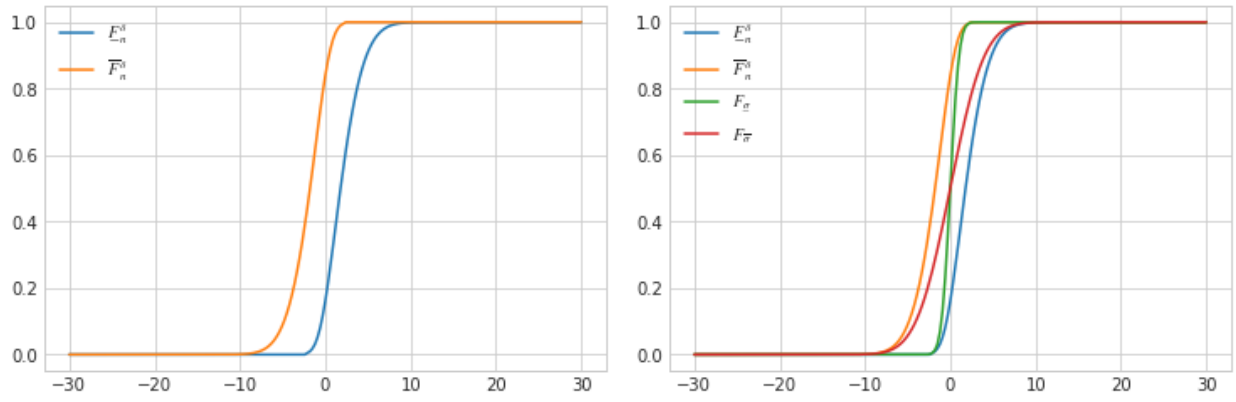


Figure 3.1: Distributional Bounds: Illustration

Note: The left panel depicts the upper bound ($\overline{F}_n^{\delta}$) and the lower bound ($\underline{F}_n^{\delta}$) of the distribution function F computed from (3.3.2) as a function of $F_{n,1}(t) = \Phi(t)$ and $F_{n,2}(t) = \Phi(t/2)$. The right panel incorporates the CDFs with the smallest standard deviation ($\underline{\sigma}$) and the largest standard deviation ($\overline{\sigma}$) defined in (3.3.3) and (3.3.4), respectively, which are labeled as $F_{\underline{\sigma}}$ and $F_{\overline{\sigma}}$, respectively.

($\overline{\sigma}^{\mathrm{WC}}/\sqrt{n}$) for σ, defined by

$$\overline{\sigma}^{\mathrm{WC}}/\sqrt{n} \equiv \max SE(\varphi(\widehat{\theta})) = \max SE(\widehat{\ell}'\widehat{\mu}) = \sum_{j=1}^{p} |\widehat{\ell}_j|\widehat{s}_j/\sqrt{n}, \tag{3.3.5}$$

Whereas they only propose the upper bound, we in addition propose the lower bound. In this sense, our method complements that of Cocci and Plagborg-Møller (2024). ▲

3.4 Theory

3.4.1 The Main Result

To establish a theoretical guarantee for the method proposed in Section 3.3, we state the following assumption.

Assumption 5. *(i)* $\left\|F_{n,j} - F_j\right\|_{\infty} = O_p(n^{-1/q_j})$ *holds for all* $j = 1,\ldots,p$, *where* $\|f(v)\|_{\infty} = \sup_{v\in\mathbb{R}} |f(v)|$. *(ii)* $\underline{\delta}_n(v) = \underline{\delta}'_n(v)\cdot n^{-1/q}$ *and* $\overline{\delta}_n(v) = \overline{\delta}'_n(v)\cdot n^{-1/q}$, *where* $\inf_v \underline{\delta}'_n(v) > 0$, $\inf_v \overline{\delta}'_n(v) > 0$, *and* $q > \max\{q_1,\ldots,q_p\}$.

Part (i) of this assumption is a high-level statement, but is satisfied in most applications, if not all. We provide a low-level sufficient condition in Section 3.4.2. Part (ii) of this assumption concerns the augmentation parameters, which are under the researcher's control. Hence, Assumption 5 is quite mild.

The following theorem and the subsequent corollary support the validity of the procedure outlined in Section 3.3.

Theorem 3.4.1. *Under Assumption 5,*

$$\underline{F}_n^{\delta}(v) \leq F(v) \leq \overline{F}_n^{\delta}(v) \tag{3.4.1}$$

holds uniformly for all $v \in \mathbb{R}$ *with probability approaching* 1 *as* $n \to \infty$.

Proof of this statement is found in Appendix C.1.1. As an immediate consequence, we obtain the bounds (3.3.3) and (3.3.4) for the asymptotic standard deviation σ as formally stated below.

Corollary 3.4.1. *Under Assumption 5,* $\underline{\sigma} \leq \sigma \leq \overline{\sigma}$ *holds with probability approaching* 1 *as* $n \to \infty$.

3.4.2 A Sufficient Condition for Assumption 5 (i)

Although it is mild, Part (i) of Assumption 5 in the previous section is a high-level statement. In the current section, we argue that the following low-level condition is sufficient for the high-level statement.

Assumption 6. $\ell_j s_j > 0$ *and* $\widehat{\ell}_j \widehat{s}_j - \ell_j s_j = O_p(n^{-1/2})$ *for all* $j = 1, \ldots, p$.

This sufficient condition is fairly mild, as it accommodates standard root-n-consistent estimators for the loadings and the asymptotic marginal variances. The following proposition argues that Assumption 6 is sufficient for Assumption 5 (i).

Proposition 3.4.1. *Assumption 6 implies Assumption 5 (i) with* $q_j = 2$ *for all* $j = 1, \ldots, p$.

Proof of this statement is found in Appendix C.1.2.

While Assumption 6 focuses on the most popular case of root-n convergence, it can be generalized. Specifically, if $\widehat{\ell}_j \widehat{s}_j - \ell_j s_j = O_p(n^{-1/q_j})$ for all $j = 1, \ldots, p$, then Assumption 5 (i) is satisfied with the corresponding q_j for all $j = 1, \ldots, p$. Essentially the same proof strategy as the proof of Proposition 3.4.1 can be used for this argument, although we omit it for brevity.

3.5 Numerical Illustrations

In this section, we present two sets of numerical illustrations. The first illustration is based on simulated data. The second illustration is based on real data.

3.5.1 Simulated Data

We first consider a data-generating process following the settings in Section 4.2 of Cocci and Plagborg-Møller (2024), where they introduce a simple model for the purpose of illustrating analytical calculations.

Define the reduced-form moment vector μ, corresponding asymptotic standard deviations s, and the structural parameters θ as follows:

$$\mu = (\mu_1, \mu_2, \mu_3)', \qquad s = (s_1, s_2, s_3)', \qquad \theta = (\theta_1, \theta_2)'.$$

The parameter of interest is the subvector $\theta_1 = \varphi(\theta)$ of θ, which reduces to μ as

$$h(\theta) = \begin{pmatrix} a & 0 \\ b & c \\ 0 & d \end{pmatrix} \theta = \begin{pmatrix} a\theta_1 \\ b\theta_1 + c\theta_2 \\ d\theta_2 \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} =: \mu.$$

The structural parameters are estimated by minimum distance:

$$\widehat{\theta} = \arg\min_{\theta} (\widehat{\mu} - h(\theta))' \widehat{W} (\widehat{\mu} - h(\theta)),$$

where the weight matrix is set to $\widehat{W} = \text{diag}(\widehat{s_1}^{-2}, \widehat{s_2}^{-2}, \widehat{s_3}^{-2})$. Following the discussion in Section 3.2, under the standard regularity conditions, we can write

$$\sqrt{n}(\widehat{\theta} - \theta_0) = \sqrt{n}\widehat{\ell}'(\widehat{\mu} - \mu) + o_p(1),$$

where $\widehat{\ell} = \widehat{W}\widehat{G}(\widehat{G}'\widehat{W}\widehat{G})^{-1}$ with $\widehat{G} = \partial h(\widehat{\theta}) / \partial \theta'$.

The data are generated from the following distributions.

$$X_1 \sim N(\mu_1, s_1^2) \qquad X_2 \sim \text{Gumbel}(\mu_2, s_2) \qquad X_3 \sim \text{Laplace}(\mu_3, s_3).$$

We set $\mu = (-1, -5, -4)'$ and $s = (0.322, 0.890, 0.588)'$.[1] We use these location and scale parameters to generate 100 observations for each sample X_j. The resultant empirical moments are

$$\widehat{\mu} = (-1.015, -4.642, -3.900)', \quad \widehat{s} = (0.331, 0.983, 0.721)',$$
$$\text{and} \qquad \widehat{V} = \begin{pmatrix} 0.111 & 0.020 & -0.020 \\ 0.020 & 0.975 & -0.695 \\ -0.020 & -0.695 & 0.525 \end{pmatrix}.$$

[1] We obtain these data-generating parameters from random draws. The locations μ_j are randomly generated from integers of range $(-5, 5)$ and scales σ_j are generated from the uniform distribution.

Table 3.1 presents the results across different model specifications. The first two columns show infeasible estimates based on the joint sample of (X_1, X_2, X_3). The last four columns show infeasible estimates based on the three separate samples. Here, we present our bounds, $\underline{\sigma}_{\theta_1}$ and $\overline{\sigma}_{\theta_1}$. Besides, we report the worst-case standard deviations $\overline{\sigma}^{WC}_{\theta_1}$ proposed by Cocci and Plagborg-Møller (2024) in the last column.

	Infeasible		Feasible			
(a,b,c,d)	$\widehat{\theta}_1$	$\widehat{\sigma}_{\theta_1}/\sqrt{n}$	$\widetilde{\theta}_1$	$\underline{\sigma}_{\theta_1}/\sqrt{n}$	$\overline{\sigma}_{\theta_1}/\sqrt{n}$	$\overline{\sigma}^{WC}_{\theta_1}/\sqrt{n}$
(1.0, 1.0, 2.0, 1.0)	-0.870	0.034	-0.870	0.011	0.042	0.040
(1.1, 0.9, 2.2, 0.9)	-0.810	0.031	-0.810	0.010	0.036	0.035
(1.2, 0.8, 2.4, 0.8)	-0.763	0.028	-0.763	0.007	0.031	0.031
(0.9, 1.1, 1.8, 1.1)	-0.952	0.038	-0.952	0.012	0.049	0.047
(0.8, 1.2, 1.6, 1.2)	-1.069	0.042	-1.069	0.007	0.058	0.055

Table 3.1: Parameter Estimates and Standard Error Bounds

Note: The table shows the parameter estimates, the corresponding standard errors, and their bounds under different model specifications. The infeasible standard errors are calculated with the covariance information given in $\widehat{V}$, whereas the feasible bounds only use marginal standard deviations. The feasible $\underline{\sigma}_{\theta_1}$ and $\overline{\sigma}_{\theta_1}$ are calculated by (3.3.3) and (3.3.4), and $\overline{\sigma}^{\mathrm{WC}}_{\theta_1}$ are from Cocci and Plagborg-Møller (2024).

Observe that our bounds contain the infeasible estimate $\widehat{\sigma}_{\theta_1}$. Our upper bounds $\overline{\sigma}_{\theta_1}$ are slightly larger than $\overline{\sigma}^{WC}_{\theta_1}$ of Cocci and Plagborg-Møller (2024). We remark that this difference is primarily due to the augmentation (3.3.2) to formally account for the finite-sample randomness of the marginal standard errors.

3.5.2 Real Data: Neoclassical Growth Model

We next replicate the example code from the Github page of Cocci and Plagborg-Møller (2024) for an analysis of the neoclassical growth model.

We focus on a simplified version of the neoclassical growth model that assumes no population growth and technological growth. The goal of this calibration exercise is to match the empirical moments to the parameters of the neoclassical growth model, such that the equilibrium conditions align with the moments.

The empirical moments and structural parameters that we consider in this study are as follows:

$$\widehat{\mu} = \left(\widehat{r}, \frac{\widehat{I}}{K}, \frac{\widehat{K}}{Y}\right)', \qquad \theta = (\rho, \delta, \alpha)',$$

where r denotes the real interest rate, I/K denotes the investment-to-capital ratio, K/Y denotes the capital-to-output ratio, ρ denotes the household discount rate, δ denotes the capital depreciation rate, and α denotes the capital elasticity

in the production function. The moments should match the three equilibrium relationships

$$r = \rho, \qquad \frac{I}{K} = \delta, \qquad \frac{K}{Y} = \frac{\alpha}{\rho + \delta},$$

where each condition is derived by the Euler equation, the capital accumulation equation, and the equilibrium of the rental rate, respectively.

The parameter of interest is $\alpha = \varphi(\theta)$, where we can express it as a linear combination of the empirical moments given the explicit formula: $\alpha = (K/Y)(\rho + \delta)$. By the linear approximation, it follows that

$$\widehat{\alpha} = \frac{\widehat{K}}{Y}\left(\widehat{r} + \frac{\widehat{I}}{K}\right) \approx \alpha + \frac{K}{Y}(\widehat{r} - r) + \frac{K}{Y}\left(\frac{\widehat{I}}{K} - \frac{I}{K}\right) + (r + I/K)\left(\frac{\widehat{K}}{Y} - \frac{K}{Y}\right).$$

We can rewrite this linear approximation as

$$\sqrt{n}(\widehat{\alpha} - \alpha) = \sqrt{n}\widehat{\ell}'(\widehat{\mu} - \mu) + o_p(1)$$

with $\widehat{\ell} = (\widehat{\mu}_3, \widehat{\mu}_3, \widehat{\mu}_1 + \widehat{\mu}_2)'$.

We construct the data from the Federal Reserve Bank of St. Louis. We set the time period to quarterly intervals and use the effective federal funds rate as the real interest rate to obtain the covariance matrix of the empirical moments. The data spans from 2003Q3 to 2022Q1. The following empirical moments are obtained from the data, where we compute the heteroskedasticity-and-autocorrelation-consistent (HAC) estimator for the variance terms. To compute the covariance terms, we assume that the empirical moments are jointly stationary.

$$\widehat{\mu} = (0.013, 0.060, 2.905)', \qquad \widehat{V} = \begin{pmatrix} 0.00134 & 0.00011 & -0.00136 \\ 0.00011 & 0.00007 & 0.00268 \\ -0.00136 & 0.00268 & 0.73159 \end{pmatrix}.$$

Table 3.2 shows the results using the above moments. Again, we present both the infeasible estimate $\widehat{\sigma}_\alpha$ and our feasible bounds $(\underline{\sigma}_\alpha, \overline{\sigma}_\alpha)$ as well as $\overline{\sigma}_\alpha^{WC}$ of Cocci and Plagborg-Møller (2024).

Observe that the infeasible estimate $\widehat{\sigma}_\alpha$ is around the middle of our bounds $\underline{\sigma}_\alpha$ and $\overline{\sigma}_\alpha$. Our upper bounds $\overline{\sigma}_\alpha$ are slightly larger than $\overline{\sigma}_\alpha^{WC}$ of Cocci and Plagborg-Møller (2024). Again, we remark that this difference is primarily due to the augmentation (3.3.2) to formally account for the finite-sample randomness of the marginal standard errors.

Figure 3.2 displays the CDFs corresponding to these bounds. The CDF associated with $\underline{\sigma}_\alpha$ exhibits a sharp slope around zero. In addition, we observe that the CDFs associated with $\overline{\sigma}_\alpha$ and $\overline{\sigma}_\alpha^{\mathrm{WC}}$ are similar.

Infeasible		Feasible			
$\widehat{\alpha}$	$\widehat{\sigma}_\alpha/\sqrt{n}$	$\widetilde{\alpha}$	$\underline{\sigma}_\alpha/\sqrt{n}$	$\overline{\sigma}_\alpha/\sqrt{n}$	$\overline{\sigma}_\alpha^{WC}/\sqrt{n}$
0.422	0.016	0.422	0.001	0.023	0.022

Table 3.2: Parameter Estimates and Standard Error Bounds

Note: The table shows the parameter estimates, the corresponding standard errors, and their bounds of the neoclassical growth model example. The infeasible standard error is calculated with the covariance information given in $\widehat{V}$, whereas the feasible bounds only use marginal standard deviations. The feasible $\underline{\sigma}_\alpha$ and $\overline{\sigma}_\alpha$ are calculated by (3.3.3) and (3.3.4), and $\overline{\sigma}_\alpha^{\text{WC}}$ are from Cocci and Plagborg-Møller (2024).

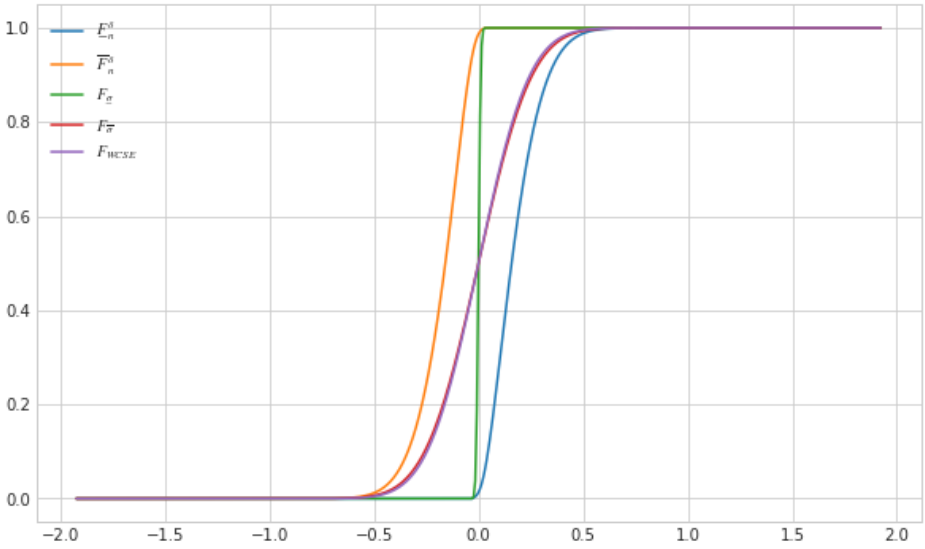


Figure 3.2: Distributional Bounds

Note: The figure displays the bounds for the asymptotic distribution of $\sqrt{n}(\widehat{\alpha}-\alpha)$ by Theorem 3.4.1, which are the $\underline{F}_n^\delta$ and $\overline{F}_n^\delta$. The CDFs with $\underline{\sigma}_\alpha$ and $\overline{\sigma}_\alpha$ are labeled as $F_{\underline{\sigma}}$ and $F_{\overline{\sigma}}$. The CDF with $\overline{\sigma}_\alpha^{\text{WC}}$ is labeled as F_{WCSE}.

3.6 Summary

This paper proposes a method to construct upper and lower bounds of the standard errors for parameters that are constructed by moments from interdependent data, where a researcher may not observe the covariance among combined data sets. Using the best-possible distributional bounds of Frank et al. (1987), we construct bounds for the asymptotic distribution for a parameter of interest, and then define the bounds for its standard error based on them. We provide a theoretical guarantee that this proposed method works in large samples. While Cocci and Plagborg-Møller (2024) propose an upper bound, we provide a lower bound in addition and thus complement their work. Furthermore, we formally account for the finite-sample randomness of the marginal standard errors.

References

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The review of economic studies*, 72(1):1–19.

Acemoglu, D., Naidu, S., Restrepo, P., and Robinson, J. A. (2019). Democracy does cause growth. *Journal of Political Economy*, 127(1):47–100.

Ackerberg, D. A., Caves, K., and Frazer, G. (2015). Identification properties of recent production function estimators. *Econometrica*, 83(6):2411–2451.

Adamek, R., Smeekes, S., and Wilms, I. (2023). Lasso inference for high-dimensional time series. *Journal of Econometrics*, 235(2):1114–1143.

Adamek, R., Smeekes, S., and Wilms, I. (2024). Local projection inference in high dimensions. *The Econometrics Journal*, page utae012.

Andrews, D. W. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, pages 817–858.

Angrist, J. D., Jordà, Ò., and Kuersteiner, G. M. (2018). Semiparametric estimates of monetary policy effects: String theory revisited. *Journal of Business & Economic Statistics*, 36(3):371–387.

Babii, A., Ghysels, E., and Striaukas, J. (2022). Machine learning time series regressions with an application to nowcasting. *Journal of Business & Economic Statistics*, 40(3):1094–1106.

Barnichon, R. and Brownlees, C. (2019). Impulse response estimation by smooth local projections. *Review of Economics and Statistics*, 101(3):522–530.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.

Belloni, A., Chernozhukov, V., Chetverikov, D., Hansen, C., and Kato, K. (2018a). High-dimensional econometrics and regularized gmm. *arXiv preprint arXiv:1806.01888*.

Belloni, A., Chernozhukov, V., Chetverikov, D., and Wei, Y. (2018b). Uniformly valid post-regularization confidence regions for many functional parameters in z-estimation framework. *Annals of statistics*, 46(6B):3643.

Belloni, A., Chernozhukov, V., and Hansen, C. (2013a). Inference on Treatment Effects after Selection among High-Dimensional Controls†. *Review of Economic Studies*, 81(2):608–650.

Belloni, A., Chernozhukov, V., and Hansen, C. (2013b). Inference on Treatment Effects after Selection among High-Dimensional Controls†. *The Review of Economic Studies*, 81(2):608–650.

Belloni, A., Chernozhukov, V., and Kato, K. (2015). Uniform post-selection inference for least absolute deviation regression and other z-estimation problems. *Biometrika*, 102(1):77–94.

Bhandari, A., Borovička, J., and Ho, P. (2024). Survey data and subjective beliefs in business cycle models. *Review of Economic Studies*, page rdae054.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732.

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

Bühlmann, P. and Yu, B. (2003). Boosting with the l 2 loss: regression and classification. *Journal of the American Statistical Association*, 98(462):324–339.

Callaway, B. and Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of econometrics*, 225(2):200–230.

Candes, E. and Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *Annals of statistics*, 35(6):2313–2351.

Caner, M. and Kock, A. B. (2018a). Asymptotically honest confidence regions for high dimensional parameters by the desparsified conservative lasso. *Journal of Econometrics*, 203(1):143–168.

Caner, M. and Kock, A. B. (2018b). High dimensional linear gmm. *arXiv preprint arXiv:1811.08779*.

Cha, J., Chiang, H. D., and Sasaki, Y. (2023). Inference in high-dimensional regression models without the exact or l^p sparsity. *Review of Economics and Statistics*, pages 1–32.

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., Chetverikov, D., and Kato, K. (2015). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Probability Theory and Related Fields*, 162:47–70.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. (2016). Locally robust semiparametric estimation. *arXiv preprint arXiv:1608.00033*.

Chernozhukov, V., Härdle, W. K., Huang, C., and Wang, W. (2021). LASSO-driven inference in time and space. *The Annals of Statistics*, 49(3):1702 – 1735.

Cocci, M. D. and Plagborg-Møller, M. (2024). Standard errors for calibrated parameters. *Review of Economic Studies*, page rdae099.

Davidson, J. (1994). *Stochastic limit theory: An introduction for econometricians*. Oxford University Press.

Davidson, J. (2021). *Stochastic limit theory: An introduction for econometricians (2nd ed.)*. Oxford University Press.

Dinh, V. H., Nibbering, D., and Wong, B. (2024). Random subspace local projections. *arXiv preprint arXiv:2406.01002*.

Dube, A., Girardi, D., Jorda, O., and Taylor, A. M. (2023). A local projections approach to difference-in-differences event studies. Technical report, National Bureau of Economic Research.

D'Amour, A., Ding, P., Feller, A., Lei, L., and Sekhon, J. (2021). Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *Annals of statistics*, 32(2):407–499.

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Annals of Statistics*, pages 1257–1272.

Fan, J. and Truong, Y. K. (1993). Nonparametric regression with errors in variables. *The Annals of Statistics*, pages 1900–1925.

Feng, G., Giglio, S., and Xiu, D. (2020). Taming the factor zoo: A test of new factors. *The Journal of Finance*, 75(3):1327–1370.

Frank, M. J., Nelsen, R. B., and Schweizer, B. (1987). Best-possible bounds for the distribution of a sum - a problem of kolmogorov. *Probability Theory and Related Fields*, 74(2):199–211.

Galbraith, J. W. and Zinde-Walsh, V. (2020). Simple and reliable estimators of coefficients of interest in a model with high-dimensional confounding effects. *Journal of Econometrics*, 218(2):609–632.

Gao, F., Ing, C.-K., and Yang, Y. (2013). Metric entropy and sparse linear approximation of ℓ_q-hulls for $0 < q \leq 1$. *Journal of Approximation Theory*, 166.

Giannone, D., Lenza, M., and Primiceri, G. E. (2021). Economic predictions with big data: The illusion of sparsity. *Econometrica*, 89(5):2409–2437.

Giraud, C. (2015). Introduction to high-dimensional statistics. *Monographs on Statistics and Applied Probability*, 139:139.

Gold, D., Lederer, J., and Tao, J. (2020). Inference for high-dimensional instrumental variables regression. *Journal of Econometrics*, 217(1):79–111.

Hansen, B. E. (1991). Strong laws for dependent heterogeneous processes. *Econometric theory*, 7(2):213–221.

Hastie, T., Tibshirani, R., and Wainwright, M. (2019). *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC.

Heckman, J. J., Ichimura, H., Smith, J. A., and Todd, P. E. (1998). Characterizing selection bias using experimental data.

Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The review of economic studies*, 64(4):605–654.

Ing, C.-K. (2007). Accumulated prediction errors, information criteria and optimal forecasting for autoregressive time series. *Annals of Statistics*, 35(3):1238–1277.

Ing, C.-K. (2020). Model selection for high-dimensional linear regression with dependent observations. *The Annals of Statistics*, 48(4):1959–1980.

Ing, C.-K. and Lai, T. L. (2011). A stepwise regression method and consistent model selection for high-dimensional sparse linear models. *Statistica Sinica*, pages 1473–1513.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909.

Jiang, W. (2009). On uniform deviations of general empirical risks with unboundedness, dependence, and high dimensionality. *Journal of Machine Learning Research*, 10(4).

Jordà, Ò. (2005). Estimation and inference of impulse responses by local projections. *American Economic Review*, 95(1):161–182.

Känzig, D. R. (2021). The macroeconomic effects of oil supply news: Evidence from opec announcements. *American Economic Review*, 111(4):1092–1125.

Kolesár, M., Müller, U. K., and Roelsgaard, S. T. (2023). The fragility of sparsity. *arXiv preprint arXiv:2311.02299*.

Kozbur, D. (2017). Testing-based forward model selection. *American Economic Review*, 107(5):266–69.

Kozbur, D. (2020). Analysis of testing-based forward model selection. *Econometrica*, 88(5):2147–2173.

Kueck, J., Luo, Y., Spindler, M., and Wang, Z. (2021). Estimation and inference of treatment effects with l_2-boosting in high-dimensional settings.

Levinsohn, J. and Petrin, A. (2003). Estimating production functions using inputs to control for unobservables. *Review of Economic Studies*, 70(2):317–341.

Liu, L. (1991). *Entry-exit and productivity change: An empirical analysis of efficiency frontiers*. PhD thesis, University of Michigan.

Makarov, G. (1982). Estimates for the distribution function of a sum of two random variables when the marginal distributions are fixed. *Theory of Probability & its Applications*, 26(4):803–806.

Marschak, J. and Andrews, W. H. (1944). Random simultaneous equations and the theory of production. *Econometrica*, pages 143–205.

Masini, R. P., Medeiros, M. C., and Mendes, E. F. (2023). Machine learning advances for time series forecasting. *Journal of economic surveys*, 37(1):76–111.

Montiel Olea, J. L. and Plagborg-Møller, M. (2021). Local projection inference is simpler and more robust than you think. *Econometrica*, 89(4):1789–1823.

Olea, J. L. M., Plagborg-Møller, M., Qian, E., and Wolf, C. K. (2024). Double robustness of local projections and some unpleasant varithmetic. Technical report, National Bureau of Economic Research.

Olley, G. S. and Pakes, A. (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica*, 64(6):1263–1297.

Petrin, A., Poi, B. P., and Levinsohn, J. (2004). Production function estimation in stata using inputs to control for unobservables. *Stata Journal*, 4(2):113–123.

Plagborg-Møller, M. and Wolf, C. K. (2021). Local projections and vars estimate the same impulse responses. *Econometrica*, 89(2):955–980.

Ridder, G. and Moffitt, R. (2007). The econometrics of data combination. volume 6 of *Handbook of Econometrics*, pages 5469–5547. Elsevier.

Robinson, P. (1988). Root- n-consistent semiparametric regression. *Econometrica*, 56(4):931–54.

Rosenbaum, P. R. and Rosenbaum, P. R. (2002). *Overt bias in observational studies*. Springer.

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

Semenova, V., Goldman, M., Chernozhukov, V., and Taddy, M. (2023). Inference on heterogeneous treatment effects in high-dimensional dynamic panels under weak dependence. *Quantitative Economics*, 14(2):471–510.

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Annals of Statistics*, pages 147–164.

Stock, J. and Watson, M. (2016). Chapter 8 - dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. volume 2 of *Handbook of Macroeconomics*, pages 415–525. Elsevier.

Stock, J. H. and Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179.

Tamer, E. (2010). Partial identification in econometrics. *Annual Review of Economics*, 2(1):167–195.

Temlyakov, V. N. (2000). Weak greedy algorithms. *Advances in Computational Mathematics*, 12(2):213–227.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tropp, J. A. (2004). Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory*, 10:2231–2242.

Tropp, J. A. and Gilbert, A. C. (2007). Signal recovery from random measurements. *Adaptif Komşuluk Seçimi ve Ağırlık Atama Yöntemleri ile Hiperspektral Görüntülerin Sınıflandırılması*.

Van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

Wang, Z., Paterlini, S., Gao, F., and Yang, Y. (2014). Adaptive minimax regression estimation over sparse lq-hulls. *The Journal of Machine Learning Research*, 15(1):1675–1711.

Williamson, R. C. and Downs, T. (1990). Probabilistic arithmetic. i. numerical methods for calculating convolutions and dependency bounds. *International Journal of Approximate Reasoning*, 4(2):89–158.

Xu, K.-L. (2023). Local projection based inference under general conditions. *Available at SSRN 4372388*.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):217–242.

Appendix A

Appendix to Chapter 1

A.1 Proof of Main Theorems

Notations. For a vector $X = (X_1, \ldots, X_p)'$, $||X||_q = (\sum_{j=1}^{p} X_j^q)^{1/q}$ denotes the L-q norm for $q < \infty$ and $||X||_\infty = \max_{1 \le j \le p} |X_j|$.

A.1.1 Proof of Theorem 1.4.1

Proof. The main purpose of this proof is to show that the assumptions **(2.33)** and **(2.34)** in Ing (2020) hold under Assumptions 1 – 4. Since the estimating equations are (1.2.3) and (1.2.4), the assumptions should hold on the two equations.

I start with condition **(2.34)** in Ing (2020) because it is a stringent condition than **(2.33)**. It states that there exists a constant $c > 0$ such that

$$P\left(\max_{1 \le j,k \le p, j \ne k} \left| \frac{1}{T} \sum_{t=1}^{T} w_{t,j} w_{t,k} - E[w_{t,j} w_{t,k}] \right| \ge c \sqrt{\frac{(\log p)^3}{T}} \right) = o(1). \tag{A.1.1}$$

For any $x \ge 0$, it follows from the union bound that

$$\begin{aligned} &P\left(\max_{1 \le j,k \le p, j \ne k} \left| \frac{1}{T} \sum_{t=1}^{T} w_{t,j} w_{t,k} - E[w_{t,j} w_{t,k}] \right| \ge x \right) \\ &\quad \le \sum_{j=1}^{p} \sum_{k=1}^{p} P\left(\left| \frac{1}{T} \sum_{t=1}^{T} w_{t,j} w_{t,k} - E[w_{t,j} w_{t,k}] \right| \ge x \right) \\ &\quad \le p^2 P\left(\left| \frac{1}{T} \sum_{t=1}^{T} w_{t,j} w_{t,k} - E[w_{t,j} w_{t,k}] \right| \ge x \right). \end{aligned}$$

Notice that by Lemma 1.4.3 (c), we can apply the triplex inequality Lemma 1.4.2,

$$\begin{aligned} &p^2 P\left(\left| \frac{1}{T} \sum_{t=1}^{T} w_{t,j} w_{t,k} - E[w_{t,j} w_{t,k}] \right| \ge x \right) \\ &\le 2p^2 m \exp\left(-\frac{Tx^2}{288 m^2 M^2} \right) && \text{(BD1)} \\ &\quad + \frac{6\bar{c}_T}{x} p^2 \rho_m && \text{(BD2)} \\ &\quad + \frac{15C}{x} p^2 \exp(-M(\bar{q} - 1)), && \text{(BD3)} \end{aligned}$$

where m is a positive integer number and $M > 0$ is some constant. I proceed by showing that there is a sequence $\eta_T \to 0$ as $T \to \infty$ that bounds all the components (BD1) – (BD3). Since Lemma 1.4.2 holds for all positive integer m and a constant $M > 0$, the bounds can be further simplified by defining m and M in terms of T and p (Jiang (2009), Remark 2). Let $M = c_M \log p$ and $m = 1$, where $c_M > 2/(\overline{q} - 1)$. To match the convergence rate in (A.1.1), let $x = c_1((\log p)^3/T)^{1/2}$, where $c_1 > 24c_M$. By Assumption 2, $\rho_m \leq \exp(-c_\kappa \log p)$, and the bounds become

$$
\begin{aligned}
&P\left(\max_{1\leq j,k\leq p, j\neq k}\left|\frac{1}{T}\sum_{t=1}^{T} w_{t,j}w_{t,k} - E[w_{t,j}w_{t,k}]\right| \geq x\right) \\
&\leq Cp^2 \exp\left(-\frac{c_1^2}{288c_M^2}\log p\right) \\
&\quad + Cx^{-1}p^2 \exp\left(-c_\kappa \log p\right) \\
&\quad + Cx^{-1}p^2 \exp(-c_M(\overline{q}-1)\log p),
\end{aligned}
$$

where I abuse the notation C for a generic constant, because only the constants inside the exponential terms matter. The first term is $o(1)$ if $c_1^2/(288c_M^2) > 2$, which is satisfied by setting $c_1 > 24c_M$. Similarly, the second and the third term are $o(1)$ if $c_\kappa > 2$ and $c_M(\overline{q}-1) > 2$, which is satisfied by definitions of the constants. Therefore, all terms are $o(1)$ and hence equation A.1.1 is satisfied.

Next we turn to condition **(2.32)** in Ing (2020). It states that there exists a constant $c > 0$ such that

$$
P\left(\max_{1\leq j\leq p}\left|\frac{1}{T}\sum_{t=1}^{T} w_{t,j}e_{t,h}\right| \geq c\sqrt{\frac{(\log p)^3}{T}}\right) = o(1) \tag{A.1.2}
$$

and

$$
P\left(\max_{1\leq j\leq p}\left|\frac{1}{T}\sum_{t=1}^{T} w_{t,j}v_{t,h}\right| \geq c\sqrt{\frac{(\log p)^3}{T}}\right) = o(1), \tag{A.1.3}
$$

for all $h = 1, \ldots, H_{\max}$.

Consider the case with $\{w_{t,j}e_{t,h}\}$. Applying the union bound, for any $x \geq 0$ it holds that

$$
\begin{aligned}
&P\left(\max_{1\leq j\leq p}\left|\frac{1}{T}\sum_{t=1}^{T} w_{t,j}e_{t,h} - E[w_{t,j}e_{t,h}]\right| \geq x\right) \\
&\qquad \leq \sum_{j=1}^{p} P\left(\left|\frac{1}{T}\sum_{t=1}^{T} w_{t,j}e_{t,h} - E[w_{t,j}e_{t,h}]\right| \geq x\right) \\
&\qquad \leq pP\left(\left|\frac{1}{T}\sum_{t=1}^{T} w_{t,j}e_{t,h} - E[w_{t,j}e_{t,h}]\right| \geq x\right).
\end{aligned}
$$

By Assumption 2 2 and Lemma 1.4.3 (b), we can apply Lemma 1.4.2, and

$$\begin{aligned} & pP\left(\left|\frac{1}{T}\sum_{t=1}^{T} w_{t,j}e_{t,h} - E[w_{t,j}e_{t,h}]\right| \geq x\right) \\ & \leq Cp\exp\left(-\frac{c_1^2}{288c_M^2}\log p\right) \\ & \quad + Cx^{-1}p\exp\left(-c_\kappa \log p\right) \\ & \quad + Cx^{-1}p\exp(-c_M(\overline{q}-1)\log p), \end{aligned}$$

where all three terms are bounded by $\eta_T \to 0$ used to bound (BD1) – (BD3) since the above right hand side terms grow at a slower rate, by p^{-1}. Therefore, (A.1.2) holds.

The same argument applies for $\{w_{t,j}v_{t,h}\}$ case, because Assumption 2 2 and Lemma 1.4.3 applies analogously with same constants $\{c_t\}$ and ρ_m. By Lemma 1.4.2, all three terms are bounded by the same $\eta_T \to 0$, and hence (A.1.3) holds.

Assumptions (A3) in Ing (2020) are assumed in Assumption 3 and Assumption (A5) in Ing (2020) is assumed in Assumption 4 (a). This set of assumptions collectively satisfies the necessary conditions outlined in Theorem 3.1 of Ing (2020), thereby yielding the desired results. □

A.1.2 Proof of Theorem 1.4.2

Proof. In this proof, I use notations in (1.4.9). Further, write (1.4.9) as

$$\begin{aligned} \boldsymbol{y}_h =& \beta_h \boldsymbol{x}_h + W_h\beta_{-h} + \boldsymbol{u}_h =: \beta_h\boldsymbol{x}_h + \mathfrak{g} + \boldsymbol{u}_h, \\ \boldsymbol{x}_h =& W_h\gamma_h + \boldsymbol{v}_h =: \mathfrak{m} + \boldsymbol{v}_h, \end{aligned} \tag{A.1.4}$$

where $\mathfrak{g}$ and $\mathfrak{m}$ are functions of W_h. In the following proof, I first show the probability limit of $\sqrt{T}(\widehat{\beta}_h - \beta)$ using similar arguments in Belloni et al. (2013a) and then derive asymptotic distribution of it using Theorem **25.12** in Davidson (2021).

Recall the subset of the covariates notation $W_h(J) = (W_{h,j})_{j\in J}$ and denote $\widehat{\Sigma}(J) = W_h(J)'W_h(J)/T$. Define

$$\begin{aligned} \sqrt{T}(\widehat{\beta}_h - \beta) &= (\boldsymbol{x}_h' M \boldsymbol{x}_h/T)^{-1}(\boldsymbol{x}_h' M(\mathfrak{g} + \boldsymbol{u}_h)/\sqrt{T}) \\ &=: A^{-1}B, \end{aligned}$$

where $M_J := I - P_J$ and $P_J = W_h(J)(W_h(J)'W_h(J))^{-1}W_h(J)'$. Define $\widetilde{J} = \widehat{J}_x \cup \widehat{J}_y$, where $\widehat{J}_x$ and $\widehat{J}_y$ are the chosen

covariates from the steps 1 and 2 in Algorithm 1. Also, let $\widetilde{\gamma}_h$ the projection coefficient of $\boldsymbol{x}_h$ onto span $W_h(\widetilde{J})$ and let $\widetilde{\alpha}_h$ be the projection coefficient of $\boldsymbol{v}_h$ onto span $W_h(\widetilde{J})$, respectively.

I will proceed by showing that

$$A = \boldsymbol{v}_h'\boldsymbol{v}_h/T + o_p(1), \qquad\qquad B = \boldsymbol{v}_h'\boldsymbol{u}_h/\sqrt{T} + o_p(1).$$

Decompose A and B into pieces using (A.1.4).

$$\begin{aligned}
A =& (\boldsymbol{x}_h' M_{\tilde{J}} \boldsymbol{x}_h)/T \\
=& (\mathfrak{m} + \boldsymbol{v}_h)' M_{\tilde{J}} (\mathfrak{m} + \boldsymbol{v}_h)/T \\
=& \mathfrak{m}' M_{\tilde{J}} \mathfrak{m}/T + 2\mathfrak{m}' M_{\tilde{J}} \boldsymbol{v}_h/T + \boldsymbol{v}_h'\boldsymbol{v}_h/T - \boldsymbol{v}_h' P_{\tilde{J}} \boldsymbol{v}_h/T, \\
B =& \boldsymbol{x}_h' M_{\tilde{J}} (\mathfrak{g} + \boldsymbol{u}_h)/\sqrt{T} \\
=& (\mathfrak{m} + \boldsymbol{v}_h)' M_{\tilde{J}} (\mathfrak{g} + \boldsymbol{u}_h)/\sqrt{T} \\
=& \mathfrak{m}' M_{\tilde{J}} \mathfrak{g}/\sqrt{T} + \mathfrak{m}' M_{\tilde{J}} \boldsymbol{u}_h/\sqrt{T} + \boldsymbol{v}_h' M_{\tilde{J}} \mathfrak{g}/\sqrt{T} + \boldsymbol{v}_h'\boldsymbol{u}_h/\sqrt{T} - \boldsymbol{v}_h' P_{\tilde{J}} \boldsymbol{u}_h/\sqrt{T}.
\end{aligned}$$

Notice that the pieces in A and B can be bounded similarly, except the bounds in B should be more restrictive. Each piece in A and B can be then decomposed into components, using idempotent property of the orthogonal projection matrix M. Below are the components used as ingredients for bounding A and B.

i) Bounds for $\left\|W_h'\boldsymbol{v}_h/T\right\|_\infty$ and $\left\|W_h'\boldsymbol{u}_h/T\right\|_\infty$

From the proof of Theorem 1.4.1, (A.1.2) and (A.1.3) holds, and hence

$$\left\|W_h'\boldsymbol{v}_h/T\right\|_\infty = O_p\left(\left((\log p)^3/T\right)^{1/2}\right), \tag{A.1.5}$$

and the same argument applies to $\left\|W_h'\boldsymbol{u}_h/T\right\|_\infty$ because $e_{t,h}$ and $u_{t,h}$ share the same assumption in Assumption 1.

ii) Bounds for $\|\widetilde{\gamma}_h - \gamma_h\|$ and $\left\|\widehat{\beta}_{-h} - \beta_{-h}\right\|$

Recall from Assumption 4 (b) and Theorem 1.4.1 we have

$$\begin{aligned}
\|\widehat{\gamma}_h - \gamma_h\| &\leq \left(\Lambda_{\min}^{-1} E_T\left[\frac{1}{T}\sum_{t=1}^{T}\left((\widehat{\gamma}_h - \gamma_h)'\boldsymbol{w}_t\right)^2\right]\right)^{1/2} \\
&= O_p\left(\left(\frac{(\log p)^3}{T}\right)^{(2\delta-1)/(4\delta)}\right).
\end{aligned}$$

Similarly,

$$\begin{aligned}\left\|\widehat{\beta}_{-h}-\beta_{-h}\right\| &\leq \left(\Lambda_{\min}^{-1}E_T\left[\frac{1}{T}\sum_{t=1}^{T}\left((\widehat{\beta}_{-h}-\beta_{-h})'\boldsymbol{w}_t\right)^2\right]\right)^{1/2}\\ &\leq \left(\Lambda_{\min}^{-1}E_T\left[\frac{1}{T}\sum_{t=1}^{T}\left((\widehat{\lambda}_{h}-\lambda_{h})'\boldsymbol{w}_t\right)^2\right]\right)\\ &= O_p\left(\left(\frac{(\log p)^3}{T}\right)^{(2\delta-1)/(4\delta)}\right),\end{aligned}$$

where the first inequality comes from Assumption 4 (b), the second from $\widetilde{J}\subset\widehat{J}_y$, and the third from Theorem 1.4.1.

iii) Bounds for $\left\|M_{\tilde{J}}\mathfrak{m}/\sqrt{T}\right\|$ and $\left\|M_{\tilde{J}}\mathfrak{g}/\sqrt{T}\right\|$

It follows from the definition of $M_{\tilde{J}}$ and $\widetilde{\gamma}_h$ that $M_{\tilde{J}}\mathfrak{m}=(I-P_{\tilde{J}})W_h\gamma_h=W_h(\gamma_h-\widetilde{\gamma}_h)$ and hence

$$\begin{aligned}\left\|M_{\tilde{J}}\mathfrak{m}/\sqrt{T}\right\| &\leq \left\|W_h(\widetilde{\gamma}_h-\gamma_h)/\sqrt{T}\right\|\\ &= O_p\left(\left(\frac{(\log p)^3}{T}\right)^{(2\delta-1)/(4\delta)}\right)\end{aligned}\tag{A.1.6}$$

from Theorem 1.4.1. For $\left\|M_{\tilde{J}}\mathfrak{g}/\sqrt{T}\right\|$, it holds by equation (A.1.4) and triangle inequality that

$$\begin{aligned}\left\|M_{\widehat{J}_y}\boldsymbol{y}_h/\sqrt{T}\right\| &\geq \left\|M_{\tilde{J}}(\beta_h\boldsymbol{x}_h+\mathfrak{g})/\sqrt{T}\right\|\\ &\geq \|\beta_h\|\left\|M_{\tilde{J}}\boldsymbol{x}_h/\sqrt{T}\right\|-\left\|M_{\tilde{J}}\mathfrak{g}/\sqrt{T}\right\|,\end{aligned}$$

where $\|\beta_h\|\leq C$ by Assumption 3. Since $\left\|M_{\widehat{J}_y}\boldsymbol{y}_h/\sqrt{T}\right\|$ and $\left\|M_{\tilde{J}}\right\|\boldsymbol{x}_h/\sqrt{T}$ share the same error bound by Theorem 1.4.1,

$$\left\|M_{\tilde{J}}\mathfrak{g}/\sqrt{T}\right\|=O_p\left(\left(\frac{(\log p)^3}{T}\right)^{(2\delta-1)/(4\delta)}\right).$$

iv) Bounds for $\|\widetilde{\gamma}_h-\gamma_h\|_1$ and $\left\|\widehat{\beta}_{-h}-\beta_{-h}\right\|_1$

Denote $\gamma_h(J)=(\gamma_{j,h})_{j=1}^{p}$, where $\gamma_{k,h}=0$ for all $k\notin J$. By triangle inequality it follows that

$$\begin{aligned}\|\widetilde{\gamma}_h-\gamma_h\|_1 &\leq \left\|\widetilde{\gamma}_h-\gamma_h(\widetilde{J})\right\|_1+\left\|\gamma_h(\widetilde{J}^c)\right\|\\ &\leq \left\|\widetilde{\gamma}_h-\gamma_h(\widetilde{J})\right\|_1+\left\|\gamma_h-\gamma_h(\widetilde{J})\right\|_1.\end{aligned}$$

The first term is bounded by

$$\begin{aligned}\left\|\widetilde{\gamma}_h - \gamma_h(\widetilde{J})\right\|_1 &\leq \sqrt{|\widetilde{J}|}\left\|\widetilde{\gamma}_h - \gamma_h(\widetilde{J})\right\| \\ &\leq \sqrt{\widehat{m}^y + \widehat{m}^x}\left\|\widehat{\gamma}_h - \gamma_h(\widehat{J}_x)\right\| \\ &\leq C\sqrt{M_T^*}\left\|\widehat{\gamma}_h - \gamma_h(\widehat{J}_x)\right\| \\ &= O_p\left(\left(\frac{(\log p)^3}{T}\right)^{(\delta-1)/(2\delta)}\right),\end{aligned}$$

where the second inequality comes from $\widetilde{J} = \widehat{J}^{[1]} \cup \widehat{J}^{[2]}$ and the third from $P(\widehat{m} \geq CM_T^*) = 0$, as proved in Section S2 of the Supplementary Material of Ing (2020). The fourth comes from the definition of M_T^* in equation (A.3.1) and the error bounds in Theroem 1.4.1. The second term can be bounded by

$$\begin{aligned}\left\|\gamma_h - \gamma_h(\widetilde{J})\right\|_1 &\leq C\sum_{j\notin\widetilde{J}}|\gamma_{j,h}| \\ &\leq CC_\delta\left(\sum_{j\notin\widetilde{J}}\gamma_{j,h}^2\right)^{(\delta-1)/(2\delta-1)} \\ &\leq CC_\delta\Lambda_{\min}^{(-\delta+1)/(2\delta-1)}\left(E_T\left[\frac{1}{T}\sum_{t=1}^{T}(x_t - \widetilde{\gamma}_h'\boldsymbol{w}_t)^2\right]\right)^{(\delta-1)/(2\delta-1)} \\ &= O_p\left(\left(\frac{(\log p)^3}{T}\right)^{(\delta-1)/(2\delta)}\right),\end{aligned}$$

where the first inequality comes from Assumption 4 (a): for all $J \subseteq \mathfrak{P}$ such that $|J| \leq C(T/(\log p)^3)^{1/2}$, the first inequality holds as shown in Ing (2020) equation **(2.16)** and the following equation. The second inequality follows from Assumption 3 3, the third is implied by Assumption 4 (b) and the error bounds from Theorem 1.4.1. Combining the two bounds, we have

$$\begin{aligned}\|\widetilde{\gamma}_h - \gamma_h\|_1 &\leq \left\|\widetilde{\gamma}_h - \gamma_h(\widetilde{J})\right\|_1 + \left\|\gamma_h - \gamma_h(\widetilde{J})\right\|_1 \\ &= O_p\left(\left(\frac{(\log p)^3}{T}\right)^{(\delta-1)/(2\delta)}\right).\end{aligned} \tag{A.1.7}$$

Similar argument applies to $\left\|\widehat{\beta}_{-h} - \beta_{-h}\right\|_1$, where $\left\|\widehat{\beta}_{-h} - \beta_{-h}\right\|_1 \leq \left\|\widehat{\beta}_{-h} - \beta_h(\widetilde{J})\right\|_1 + \left\|\beta_h - \beta_h(\widetilde{J})\right\|_1$ by triangle

inequality. The two terms are bounded by

$$\begin{aligned}\left\|\widehat{\beta}_h - \beta_h(\widetilde{J})\right\|_1 &\leq \sqrt{|\widetilde{J}|}\left\|\widehat{\beta}_h - \beta_h(\widetilde{J})\right\| \\ &\leq \sqrt{M_T^*}\left\|\widehat{\lambda}_h - \lambda_h(\widehat{J}_y)\right\| \\ &= O_p\left(\left(\frac{(\log p)^3}{T}\right)^{(\delta-1)/(2\delta)}\right)\end{aligned}$$

and

$$\begin{aligned}\left\|\beta_h(\widehat{J}) - \beta_h\right\|_1 &\leq C\sum_{j\notin\widetilde{J}}|\beta_{j,-h}| \\ &\leq CC_\delta\left(\sum_{j\notin\widetilde{J}}\beta_{j,-h}^2\right)^{(\delta-1)/(2\delta-1)} \\ &\leq CC_\delta\Lambda_{\min}^{(-\delta+1)/(2\delta-1)}\left(E_T\left[\frac{1}{T}\sum_{t=1}^{T}(y_{t+h} - \widehat{\beta}_{-h}'\boldsymbol{w}_t)^2\right]\right)^{(\delta-1)/(2\delta-1)} \\ &\leq CC_\delta\Lambda_{\min}^{(-\delta+1)/(2\delta-1)}\left(E_T\left[\frac{1}{T}\sum_{t=1}^{T}(y_{t+h} - \widehat{\lambda}_h'\boldsymbol{w}_t)^2\right]\right)^{(\delta-1)/(2\delta-1)} \\ &= O_p\left(\left(\frac{(\log p)^3}{T}\right)^{(\delta-1)/(2\delta-1)}\right),\end{aligned}$$

where all the steps are analogous to deriving the bounds of $\|\widetilde{\gamma}_h - \gamma_h\|_1$ except for the fourth inequality here, which comes from $\widehat{J}_y \subset \widetilde{J}$. Hence it follows that

$$\left\|\widehat{\beta}_{-h} - \beta_h\right\|_1 = O_p\left(\left(\frac{(\log p)^3}{T}\right)^{(\delta-1)/(2\delta-1)}\right).$$

v) Bounds for $\|\widetilde{\alpha}_h\|_1$

Similarly to the previous bound,

$$\begin{aligned}\|\widetilde{\alpha}_h\|_1 &\leq \sqrt{\widetilde{J}}\|\widetilde{\alpha}_h\| \\ &\leq \sqrt{|\widetilde{J}|}\left\|\widehat{\Sigma}^{-1}(\widetilde{J})\right\|\left\|W_h(\widetilde{J})'\boldsymbol{v}_h/T\right\| \\ &\leq |\widetilde{J}|\left\|\widehat{\Sigma}^{-1}(\widetilde{J})\right\|\left\|W_h(\widetilde{J})'\boldsymbol{v}_h/T\right\|_\infty \\ &= O_p\left(\left(\frac{(\log p)^3}{T}\right)^{(\delta-1)/(2\delta)}\right)\end{aligned}$$

Now back to A and B,

$$A - \boldsymbol{v}_h'\boldsymbol{v}_h/T \leq \left|\mathfrak{m}'M\mathfrak{m}/T\right| + 2\left|\mathfrak{m}'M\boldsymbol{v}_h/T\right| + \left|\boldsymbol{v}_h'P\boldsymbol{v}_h/T\right|$$

and

$$B - \boldsymbol{v}_h'\boldsymbol{u}_h/\sqrt{T} \leq \left|\mathfrak{m}'M\mathfrak{g}/\sqrt{T}\right| + \left|\mathfrak{m}'M\boldsymbol{u}_h/\sqrt{T}\right| + \left|\boldsymbol{v}_h'M\mathfrak{g}/\sqrt{T}\right| + \left|\boldsymbol{v}_h'P\boldsymbol{u}_h/\sqrt{T}\right|,$$

where the second and third pieces in $B - \boldsymbol{v}_h'\boldsymbol{u}_h/\sqrt{T}$ share the same bounds and $A - \boldsymbol{v}_h'\boldsymbol{v}_h/T = o_p(1)$ if $B - \boldsymbol{v}_h'\boldsymbol{u}_h/\sqrt{T} = o_p(1)$. The pieces in $B - \boldsymbol{v}_h'\boldsymbol{u}_h/\sqrt{T}$ can be bounded by

$$\begin{aligned}
\left|\mathfrak{m}'M\mathfrak{g}/\sqrt{T}\right| &\leq \sqrt{T}\left\|M\mathfrak{m}/\sqrt{T}\right\|\left\|M\mathfrak{g}/\sqrt{T}\right\| \\
&= O_p\left(\left((\log p)^3\right)^{\frac{2\delta-1}{2\delta}} T^{\frac{1-\delta}{2\delta}}\right), \\
\left|\mathfrak{m}'M\boldsymbol{u}_h/\sqrt{T}\right| &\leq \sqrt{T}\left|(\widetilde{\gamma}_h - \gamma_h)'W_h'\boldsymbol{u}_h/T\right| \\
&\leq \sqrt{T}\left\|\widetilde{\gamma}_h - \gamma_h\right\|_1 \left\|W_h'\boldsymbol{u}_h/T\right\|_\infty \\
&= O_p\left(\left((\log p)^3\right)^{\frac{2\delta-1}{2\delta}} T^{\frac{1-\delta}{2\delta}}\right), \\
\left|\boldsymbol{v}_h'P\boldsymbol{u}_h/\sqrt{T}\right| &\leq \left|\widetilde{\alpha}_h'W_h'\boldsymbol{u}_h/\sqrt{T}\right| \\
&\leq \left\|\widetilde{\alpha}_h\right\|_1 \sqrt{T}\left\|W_h'\boldsymbol{u}_h/T\right\|_\infty \\
&= O_p\left(\left((\log p)^3\right)^{\frac{2\delta-1}{2\delta}} T^{\frac{1-\delta}{2\delta}}\right),
\end{aligned}$$

where all the pieces become $o(1)$ if $\log p = o(T^{\frac{\delta-1}{3(2\delta-1)}})$, which is satisfied by Assumption 4 (c). Therefore we have

$$\sqrt{T}(\widehat{\beta}_h - \beta_h) = \left(\boldsymbol{v}_h'\boldsymbol{v}_h/T\right)^{-1}\left(\boldsymbol{v}_h'\boldsymbol{u}_h/\sqrt{T}\right) + o_p(1).$$

The remainder of the proof continues by deriving the asymptotic distribution of

$$\left(\boldsymbol{v}_h'\boldsymbol{v}_h/T\right)^{-1}\left(\boldsymbol{v}_h'\boldsymbol{u}_h/\sqrt{T}\right) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} v_{h,t}u_{h,t}/\tau_h^2.$$

I proceed by applying Theorem **25.12** in Davidson (2021), showing that conditions (a) – (c) hold. First, condition (a)

states that $E\left[(\sum_{t=1}^{T} X_t)^2\right] = 1$, where $X_t = v_{t,h}u_t/(\tau_h^2\sigma_h\sqrt{T})$.

$$E\left[\left(\sum_{t=1}^{T} X_t\right)^2\right] = \frac{1}{\tau_h^4\sigma_h^2} E\left[\frac{1}{T}\left(\sum_{t=1}^{T} v_{t,h}u_t\right)^2\right] = \frac{1}{\tau_h^4\sigma_h^2}\Omega_h = 1.$$

Next, consider the condition (b). By Lemma 1.4.3 (a) and Theorem **18.9** of Davidson (2021), $\{v_{t,h}u_t\}$ is a causal L_q-NED of size $-b$, and hence of size $-1/2$. By Assumption 1 (d), it is NED on an $\alpha-$mixing array $\{\mathrm{Y}_{Tt}\}$ of size $-b/(1/q-1/\overline{q}) < -\overline{q}/(\overline{q}-2)$, which is satisfied by $\overline{q} > q > 2$ in Assumption 1.

The last condition (c) is on $L_{\overline{q}}-$boundedness. From the definition of τ_h,

$$\begin{aligned}\tau_h^2 =& \min_{\gamma_j}\left\{E\left[\frac{1}{T}\sum_{t=1}^{T}(x_{h,t}-\gamma_h'\boldsymbol{w}_t)^2\right]\right\}\\ \leq& E\left[\frac{1}{T}\sum_{t=1}^{T}(x_{h,t}-0'\boldsymbol{w}_t)^2\right] = \frac{1}{T}\sum_{t=1}^{T}E[x_{h,t}^2] = \Sigma_{h,h} \leq C,\end{aligned}$$

where 0 denotes the 0 vector and the last inequality follows by Assumption 4 (b). Also, by Assumption 1 (b) and Cauchy-Schwarz inequality, it follows that $v_{t,h}u_t$ is $L_{\overline{q}}-$bounded. Consider the condition (c)

$$\sup_{T,t}\left(E\left[\left|v_{t,h}u_t/\tau_h^2\right|^{\overline{q}}\right]\right)^{1/\overline{q}} \leq \frac{1}{\tau_h^2}\sup_{T,t}\left(E\left[\left|v_{t,h}u_t\right|^{\overline{q}}\right]\right)^{1/\overline{q}} \leq C,$$

which follows from $\tau_h^2 \leq C$ and $L_{\overline{q}}-$boundedness of $v_{t,h}u_t$.

With conditions (a) – (c) satisfied, it holds that $1/\sqrt{T}\sum_{t=1}^{T} v_{t,h}u_t/(\tau_h^2\sigma_h) \xrightarrow{d} N(0,1)$, and we can obtain the desired results by applying Slutsky's theorem. □

A.1.3 Proof of Theorem 1.4.3

Proof. Decompose

$$\begin{aligned}\left|\widehat{\sigma}_h^2 - \sigma_h^2\right| &= \left|\frac{1}{\widehat{\tau}_h^4}\widehat{\Omega}_h - \frac{1}{\tau_h^4}\Omega_h\right|\\ &\leq \left|\left(\frac{1}{\widehat{\tau}_h^4}-\frac{1}{\tau_h^4}\right)\left(\widehat{\Omega}_h-\Omega_h\right)\right| + \left|\left(\frac{1}{\widehat{\tau}_h^4}-\frac{1}{\tau_h^4}\right)\Omega_h\right| + \left|\frac{1}{\tau_h^4}\left(\widehat{\Omega}_h-\Omega_h\right)\right|.\end{aligned} \tag{A.1.8}$$

From Assumption 4 (b), $1/\tau_h^4$ is bounded by a constant. I further show the probability bounds for the other components, namely $\left|\frac{1}{\widehat{\tau}_h^4}-\frac{1}{\tau_h^4}\right|$, Ω_h, and $\left|\widehat{\Omega}_h-\Omega_h\right|$.

First, I establish

$$\left| \frac{1}{\widehat{\tau}_h^4} - \frac{1}{\tau_h^4} \right| \xrightarrow{p} 0.$$

From the definition of $\widehat{\tau}_h^2$, it can be decomposed as

$$\begin{aligned}
\widehat{\tau}_h^2 &= \left\| \left(\boldsymbol{x}_h - W_h \widehat{\gamma}_h \right) / \sqrt{T} \right\|^2 \\
&= \frac{1}{T} \sum_{t=1}^{T} v_{t,h}^2 + \left\| W_h \left(\gamma_h - \widehat{\gamma}_h \right) / \sqrt{T} \right\|^2 + 2 \left| \boldsymbol{v}_h' W_h \left(\gamma_h - \widehat{\gamma}_h \right) / T \right|,
\end{aligned}$$

and we can write

$$\left| \widehat{\tau}_h^2 - \tau_h^2 \right| \leq \left| \frac{1}{T} \sum_{t=1}^{T} v_{t,h}^2 - \frac{1}{T} \sum_{t=1}^{T} E[v_{t,h}^2] \right| + \left\| W_h \left(\gamma_h - \widehat{\gamma}_h \right) / \sqrt{T} \right\|^2 + 2 \left| \boldsymbol{v}_h' W_h \left(\gamma_h - \widehat{\gamma}_h \right) / T \right|,$$

where

$$\begin{aligned}
\left\| W_h \left(\gamma_h - \widehat{\gamma}_h \right) / \sqrt{T} \right\|^2 &= O_p \left(\left(\frac{(\log p)^3}{T} \right)^{\frac{2\delta - 1}{2\delta}} \right), \\
\left| \boldsymbol{v}_h' W_h \left(\gamma_h - \widehat{\gamma}_h \right) / T \right| &\leq \left\| W_h' \boldsymbol{v}_h / T \right\|_\infty \left\| \gamma_h - \widehat{\gamma}_h \right\|_1 = O_p \left(\left(\frac{(\log p)^3}{T} \right)^{\frac{2\delta - 1}{2\delta}} \right)
\end{aligned}$$

from equations (A.1.6), (A.1.5), and (A.1.7). Now applying the triplex inequality to the first component by setting $x = c_1 ((\log p)^3 / T)^{1/2}$,

$$\begin{aligned}
&P \left(\left| \frac{1}{T} \sum_{t=1}^{T} v_{t,h}^2 - \frac{1}{T} \sum_{t=1}^{T} E[v_{t,h}^2] \right| < c_1 \left(\frac{(\log p)^3}{T} \right)^{\frac{1}{2}} \right) \\
&\leq C \exp \left(- \frac{c_1^2}{288 c_M^2} \log p \right) + C x^{-1} \exp(-c_\kappa \log p) + C x^{-1} \exp(-c_M (\overline{q} - 1) \log p)
\end{aligned} \tag{A.1.9}$$

can be bounded by the sequence $\eta_T \to 0$ which bounds (BD1) – (BD3), because the bounds (BD1) – (BD3) grow p^2 times faster than the above bounds. Combining all three parts, we have

$$\left| \widehat{\tau}_h^2 - \tau_h^2 \right| = O_p \left(\left(\frac{(\log p)^3}{T} \right)^{\frac{1}{2}} \right). \tag{A.1.10}$$

We can write

$$\begin{aligned}\left|\frac{1}{\widehat{\tau}_h^4}-\frac{1}{\tau_h^4}\right| &= \left|\frac{1}{\widehat{\tau}_h^2}-\frac{1}{\tau_h^2}\right|\left|\frac{1}{\widehat{\tau}_h^2}+\frac{1}{\tau_h^2}\right| \\ &= \left|\frac{\widehat{\tau}_h^2-\tau_h^2}{\tau_h^4-\tau_h^2(\widehat{\tau}_h^2-\tau_h^2)}\right|\left|\frac{\widehat{\tau}_h^2-\tau_h^2+2\tau_h^2}{\tau_h^4-\tau_h^2(\widehat{\tau}_h^2-\tau_h^2)}\right| \\ &\leq \left|\frac{|\widehat{\tau}_h^2-\tau_h^2|}{\tau_h^4-\tau_h^2|\widehat{\tau}_h^2-\tau_h^2|}\right|\left|\frac{|\widehat{\tau}_h^2-\tau_h^2|+2\tau_h^2}{\tau_h^4-\tau_h^2|\widehat{\tau}_h^2-\tau_h^2|}\right| = o_p(1),\end{aligned}$$

where the last bound follows from (A.1.10) and Assumption 4 (b).

Second, to show $\Omega_h \leq C$, I first establish that $\max_{t\leq T} E[\psi_t \psi_{t-\ell}] \leq C\xi_\ell$, where $\xi_\ell \to 0$ is a sequence of size $-b$. This proof follows similar reasoning found in the proof of Lemma B.2. (ii) of Adamek et al. (2023). First, by Assumption 1 (d) and Lemma 1.4.3 (a), both $\{u_{t,h}\}$ and $\{v_{t,h}\}$ are $L_{2q}-$NED of size $-b$, and hence by Theorem **18.9** in Davidson (2021), $\{v_{t,h}u_{t,h}\}$ is L_q-NED of size $-b$ for all $h = 1,\ldots,H_{\max}$. Also, by Assumption 1 (b) and Cauchy-Schwarz inequality, $\{v_{t,h}u_{t,h}\}$ is $L_{\overline{q}}-$bounded. Applying Theorem **18.6** of Davidson (2021), $\{v_{t,h}u_{t,h}\}$ is an L_q-mixingale of size $-b$.

Let $k = [\ell/2]$, and decompose $E[\psi_t \psi_{t-\ell}]$ by Minkowski's inequality.

$$\begin{aligned}E[\psi_t \psi_{t-\ell}] &\leq \left|E\left[\psi_t\left(\psi_{t-\ell} - E_{t-\ell-k}^{t-\ell+k}[\psi_{t-\ell}]\right)\right]\right| + \left|E\left[\psi_t E_{t-\ell-k}^{t-\ell+k}[\psi_{t-\ell}]\right]\right| \\ &=: A + B.\end{aligned}$$

The first term can be bounded by Hölder inequality,

$$A \leq \left|\left(E\left[\psi_t^{\frac{q}{q-1}}\right]\right)^{\frac{q-1}{q}}\left(E\left[\left|\psi_{t-\ell} - E_{t-\ell-k}^{t-\ell+k}[\psi_{t-\ell}]\right|^q\right]\right)^{\frac{1}{q}}\right|,$$

where the $q/(q-1)$-th moment of ψ_t is bounded by a constant from $L_{\overline{q}}-$boundedness, and the latter term is bounded by

$$\left(E\left[\left|\psi_{t-\ell} - E_{t-\ell-k}^{t-\ell+k}[\psi_{t-\ell}]\right|^q\right]\right)^{\frac{1}{q}} \leq C\zeta_k,$$

where ζ_k is the sequence from the Definition 4, since $\{\psi_{t-\ell}\}$ is L_q-NED of size $-b$. It follows that $\zeta_k = O(k^{-\widetilde{b}}) = O(\ell^{-\widetilde{b}})$ for $\widetilde{b} > b$. By LIE and Hölder's inequality it holds that

$$B = \left|E\left[E_{t-\ell-k}^{t-\ell+k}\psi_t E_{t-\ell-k}^{t-\ell+k}\psi_{t-\ell}\right]\right|$$

$$\leq \left| \left(E\left[\left| E_{t-\ell-k}^{t-\ell+k} \psi_t \right|^q \right] \right)^{\frac{1}{q}} \left(E\left[\left| E_{t-\ell-k}^{t-\ell+k} \psi_{t-\ell} \right|^{\frac{q}{q-1}} \right] \right)^{\frac{q-1}{q}} \right|,$$

where again the latter term can be bounded by a constant from $L_{\overline{q}}-$boundedness, and the first term can be bounded by

$$\left(E\left[\left| E_{t-\ell-k}^{t-\ell+k} \psi_t \right|^q \right] \right)^{\frac{1}{q}} \leq \left(E\left[\left| E_{-\infty}^{t-\ell+k} \psi_t \right|^q \right] \right)^{\frac{1}{q}} \leq C\rho_{\ell-k},$$

where the first inequality follows because conditioning is a contractionary projection in L_p spaces. The sequence $\rho_{\ell-k}$ is from the Definition 2, since $\{\psi_t\}$ is an L_q-mixingale of size $-b$. Similarly to ζ_k, it follows that $\rho_{\ell-k} = O((\ell-k)^{-\widetilde{b}}) = O(\ell^{-\widetilde{b}})$. Note also that ζ_k and $\rho_{\ell-k}$ are both independent of t and hence,

$$\max_{t\leq T} E[\psi_t \psi_{t-\ell}] \leq C\xi_\ell, \tag{A.1.11}$$

where $\xi_\ell = O(\ell^{-\widetilde{b}})$. This implies that the covariances are absolutely summable.

Now consider

$$\begin{aligned} |\Omega_h| &= \sum_{\ell=-T+1}^{T-1} \left| \frac{1}{T} \sum_{t=\ell+1}^{T} E[\psi_t \psi_{t-\ell}] \right| \\ &\leq 2 \sum_{\ell=0}^{T-1} \left| \max_{t\leq T} E[\psi_t \psi_{t-\ell}] \right| \leq C, \end{aligned}$$

where the last inequality follows from the absolute summability.

Finally, I show $\left|\widehat{\Omega}_h - \Omega_h\right| \to 0$. First, divide it into two terms,

$$\left|\widehat{\Omega}_h - \Omega_h\right| \leq \left|\widehat{\Omega}_h - \Omega_h^K\right| + \left|\Omega_h^K - \Omega_h\right|,$$

where $\Omega_h^K = \sum_{\ell=-K+1}^{K-1} \frac{1}{T} \sum_{t=\ell+1}^{T} E[\psi_t \psi_{t-\ell}]$. Note that by (A.1.11), the latter term can be bounded by

$$\left|\Omega_h^K - \Omega_h\right| \leq 2 \sum_{\ell=K}^{T} \left| \frac{1}{T} \sum_{t=\ell+1}^{T} E[\psi_t \psi_{t-\ell}] \right| \leq C \sum_{\ell=K}^{T} \xi_\ell \leq C \sum_{\ell=K}^{T} \ell^{-\widetilde{b}}, \tag{A.1.12}$$

where it converges to 0 by the following arguments. Let $\widetilde{b} = b + \varepsilon = 1 + (b-1) + \varepsilon$ for $\varepsilon > 0$ and let $\delta = \varepsilon/2$. Since $K \leq \ell$,

$$\sum_{\ell=K}^{T} \ell^{-\widetilde{b}} \leq K^{-b+1} \sum_{\ell=K}^{T} \ell^{-1-\varepsilon} \leq K^{1-b-\delta} \sum_{\ell=K}^{T} \ell^{-1-\delta},$$

and $K^{1-b-\delta} \to 0$ with $b \geq 1$ and $\sum_{\ell=K}^{T} \ell^{-1-\delta} \to 0$ with $K \to \infty$ and the property of $p-$series.

Now consider the first term. Using a telescopic sum argument, decompose

$$\begin{aligned}\left|\widehat{\Omega}_h - \Omega_h^K\right| &= \left|\sum_{\ell=-K+1}^{K-1}\left(1-\frac{\ell}{K}\right)\frac{1}{T}\sum_{t=\ell+1}^{T}\widehat{\psi}_t\widehat{\psi}_{t-\ell} - \frac{1}{T}\sum_{t=\ell+1}^{T}E[\psi_t\psi_{t-\ell}]\right| \\ &\leq 2\sum_{\ell=0}^{K-1}\left(\left|\frac{1}{T}\sum_{t=\ell+1}^{T}(\widehat{\psi}_t\widehat{\psi}_{t-\ell} - E[\psi_t\psi_{t-\ell}])\right| + \frac{\ell}{K}\left|\frac{1}{T}\sum_{t=\ell+1}^{T}E[\psi_t\psi_{t-\ell}]\right|\right). \end{aligned} \tag{A.1.13}$$

Note that from (A.1.11), the latter term can be bounded by

$$\sum_{\ell=0}^{K-1}\frac{\ell}{K}\left|\frac{1}{T}\sum_{t=\ell+1}^{T}E[\psi_t\psi_{t-\ell}]\right| \leq \frac{C}{K}\sum_{\ell=1}^{K-1}\ell^{1-\widetilde{b}} \leq CK^{-\widetilde{b}}\sum_{\ell=1}^{K-1}\left(\frac{\ell}{K}\right)^{1-\widetilde{b}} \leq CK^{1-\widetilde{b}},$$

where the last inequality comes from $\ell < K$ and $\sum_{\ell=1}^{K-1}\ell^{-1-\varepsilon} \leq C$ for $\varepsilon > 0$, and hence it converges to 0 since $K^{1-\widetilde{b}} \to 0$ for $b \geq 1$.

Next, we turn to the first term in (A.1.13). From triangle inequality we have

$$\begin{aligned}&\left|\frac{1}{T}\sum_{t=\ell+1}^{T}(\widehat{\psi}_t\widehat{\psi}_{t-\ell} - E[\psi_t\psi_{t-\ell}])\right| \\ &\leq \left|\frac{1}{T}\sum_{t=\ell+1}^{T}(\widehat{\psi}_t\widehat{\psi}_{t-\ell} - \psi_t\psi_{t-\ell})\right| + \left|\frac{1}{T}\sum_{t=\ell+1}^{T}(\psi_t\psi_{t-\ell} - E[\psi_t\psi_{t-\ell}])\right| \\ &=: A + B. \end{aligned} \tag{A.1.14}$$

We can further write A as

$$\begin{aligned}A \leq &\left|\frac{1}{T}\sum_{t=\ell+1}^{T}\underbrace{(\widehat{u}_t\widehat{u}_{t-\ell} - u_tu_{t-\ell})}_{A(a)}\underbrace{v_tv_{t-\ell}}_{A(b)}\right| + \left|\frac{1}{T}\sum_{t=\ell+1}^{T}\underbrace{(\widehat{v}_t\widehat{v}_{t-\ell} - v_tv_{t-\ell})}_{A(c)}\underbrace{u_tu_{t-\ell}}_{A(b')}\right| \\ &+ \left|\frac{1}{T}\sum_{t=\ell+1}^{T}\underbrace{(\widehat{u}_t\widehat{u}_{t-\ell} - u_tu_{t-\ell})}_{A(a)}\underbrace{(\widehat{v}_t\widehat{v}_{t-\ell} - v_tv_{t-\ell})}_{A(c)}\right|, \end{aligned} \tag{A.1.15}$$

where I omit the subscript for h for simplicity. Consider the component $A(a)$ in the first and the last term. Using the baseline model specification (1.2.2) and the triangle inequality, we can write

$$\begin{aligned}|\widehat{u}_t\widehat{u}_{t-\ell} - u_tu_{t-\ell}| \leq\ & 2\left|\left(\beta_h - \widehat{\beta}_h\right)x_tu_{t-\ell}\right| + 2\left|\left(\beta_{-h} - \widehat{\beta}_{-h}\right)'w_tu_{t-\ell}\right| + \left|\left(\beta_h - \widehat{\beta}_h\right)^2x_tx_{t-\ell}\right| \\ &+ 2\left|(\beta_h - \widehat{\beta}_h)x_{t-\ell}w_t'(\beta_{-h} - \widehat{\beta}_{-h})\right| + \left|\left(\beta_{-h} - \widehat{\beta}_{-h}\right)'w_t\left(\beta_{-h} - \widehat{\beta}_{-h}\right)'w_{t-\ell}\right| \end{aligned}$$

$$=: A(a)_i + A(a)_{ii} + A(a)_{iii} + A(a)_{iv} + A(a)_v.$$

Each term is then bounded by

$$\begin{aligned}
\frac{1}{T}\sum_{t=\ell+1}^{T} A(a)_i &\leq 2\frac{1}{T}\sum_{t=\ell+1}^{T}\left|\beta_h - \widehat{\beta}_h\right| \max_{t\leq T}|x_t u_{t-\ell}| = O_p\left(T^{-\frac{1}{2}}\left(\frac{(\log p)^3}{T}\right)^{\frac{1}{2}}\right), \\
\frac{1}{T}\sum_{t=\ell+1}^{T} A(a)_{ii} &\leq 2\left|(\beta_{-h} - \widehat{\beta}_{-h})W_h'\boldsymbol{u}_h/T\right| \\
&\leq 2\left\|\beta_{-h} - \widehat{\beta}_{-h}\right\|_1 \left\|W_h'\boldsymbol{u}_h/T\right\|_\infty = O_p\left(\left(\frac{(\log p)^3}{T}\right)^{\frac{2\delta-1}{2\delta}}\right) \\
\frac{1}{T}\sum_{t=\ell+1}^{T} A(a)_{iii} &\leq \left|(\beta_h - \widehat{\beta}_h)^2\right| \max_{t\leq T}|x_t x_{t-\ell}| = O_p(T^{-1+\frac{1}{\overline{q}}}) \\
\frac{1}{T}\sum_{t=\ell+1}^{T} A(a)_{iv} &\leq 2\left|\beta_h - \widehat{\beta}_h\right| \max_{t\leq T}|x_t| \left\|\boldsymbol{w}_t\right\|_\infty \left\|\beta_{-h} - \widehat{\beta}_{-h}\right\|_1 = O_p\left(T^{-\frac{1}{2}+\frac{1}{\overline{q}}}\left(\frac{(\log p)^3}{T}\right)^{\frac{2\delta-1}{2\delta}}\right) \\
\frac{1}{T}\sum_{t=\ell+1}^{T} A(a)_{v} &\leq \left\|M\mathfrak{g}/\sqrt{T}\right\|^2 = O_p\left(\left(\frac{(\log p)^3}{T}\right)^{\frac{2\delta-1}{2\delta}}\right),
\end{aligned}$$

where I use the $\overline{q}-$th moment boundedness from Assumption 1 (b), $\sqrt{T}-$rate for $\left|\beta_h - \widehat{\beta}_h\right|$ from Theorem 1.4.2, and the error bounds for $\left|\beta_{-h} - \widehat{\beta}_{-h}\right|$ from Theorem 1.4.1. Note that $|A(b)|$ can be bounded by $T^{1/\overline{q}}$ by Assumption 1 (b) and Cauchy-Schwarz inequality. Combining with $|A(b)|$, the first term in (A.1.15) can be bounded at the rate of $o(1)$ if $T^{-1/2+2/\overline{q}} = o(1)$, which is satisfied by Assumption 1 (b).

Turning to $A(c)$, it can be similarly expanded by (1.2.4) and the triangle inequality. We can write

$$|\widehat{v}_t\widehat{v}_{t-\ell} - v_t v_{t-\ell}| \leq \left|(\gamma_h - \widehat{\gamma}_h)'\boldsymbol{w}_t v_{t-\ell}\right| + \left|(\gamma_h - \widehat{\gamma}_h)'\boldsymbol{w}_{t-\ell} v_t\right| + \left|(\gamma_h - \widehat{\gamma}_h)'\boldsymbol{w}_t(\gamma_h - \widehat{\gamma}_h)'\boldsymbol{w}_{t-\ell}\right|,$$

and each term is bounded By

$$\begin{aligned}
\frac{1}{T}\sum_{t=\ell+1}^{T}\left|(\gamma_h - \widehat{\gamma}_h)'\boldsymbol{w}_t v_{t-\ell}\right| &\leq \|\gamma_h - \widehat{\gamma}_h\|_1 \left\|W_h'\boldsymbol{v}_h/T\right\|_\infty = O_p\left(\left(\frac{(\log p)^3}{T}\right)^{\frac{2\delta-1}{2\delta}}\right) \\
\frac{1}{T}\sum_{t=\ell+1}^{T}\left|(\gamma_h - \widehat{\gamma}_h)'\boldsymbol{w}_{t-\ell} v_t\right| &\leq \|\gamma_h - \widehat{\gamma}_h\|_1 \left\|W_h'\boldsymbol{v}_h/T\right\|_\infty = O_p\left(\left(\frac{(\log p)^3}{T}\right)^{\frac{2\delta-1}{2\delta}}\right) \\
\frac{1}{T}\sum_{t=\ell+1}^{T}\left|(\gamma_h - \widehat{\gamma}_h)'\boldsymbol{w}_t(\gamma_h - \widehat{\gamma}_h)'\boldsymbol{w}_{t-\ell}\right| &\leq \left\|M_{\widehat{J}_x}\mathfrak{m}/\sqrt{T}\right\|^2 = O_p\left(\left(\frac{(\log p)^3}{T}\right)^{\frac{2\delta-1}{2\delta}}\right).
\end{aligned}$$

Combining with $|A(b')|$, which can be bounded analogously to $|A(b)|$, the second term in (A.1.15) can be bounded at the rate of $o(1)$ if $\log p = o(T^{1-\frac{1}{q}\frac{2\delta}{2\delta-1}})$, which is satisfied by Assumption 1 (b) and 4 (c). The third term in (A.1.15) is bounded at the rate of $o(1)$ if both $A(a)$ and $A(b)$ are bounded at the rate of $o(1)$, and it trivially holds by the previous arguments. Therefore, $A = o_p(1)$.

For part B, we use the triplex inequality. By Theorem **18.11** of Davidson (2021), $\{\psi_t\psi_{t-\ell}\}$ is a $L_{\overline{q}/2}-$bounded $L_{q/2}-$NED, and by Theorem **18.6** of Davidson (2021), $\{\psi_t\psi_{t-\ell} - E[\psi_t\psi_{t-\ell}]\}$ is a $L_{\overline{q}/2}-$bounded $L_{q/2}-$mixingale. Applying Lemma 1.4.2 with $x = c_1((\log p)^3/T)^{1/2}$,

$$\begin{aligned} &P\left(\left|\frac{1}{T}\sum_{t=\ell+1}^{T}(\psi_t\psi_{t-\ell} - E[\psi_t\psi_{t-\ell}])\right| < c_1\left(\frac{(\log p)^3}{T}\right)^{1/2}\right) \\ &\leq\ C\exp\left(-\frac{c_1^2}{288c_M^2}\log p\right) + Cx^{-1}\exp\left(-c_\kappa \log p\right) + Cx^{-1}\exp(-c_M(\overline{q}/2-1)\log p), \end{aligned}$$

where all three terms are bounded by $\eta_T \to 0$ used to bound (BD1) – (BD3) since the above right hand side terms grow at a slower rate, by p^{-2}. Therefore, $B = o_p(1)$ and hence (A.1.14) is $o_p(1)$.

So far I've shown that (A.1.12) and (A.1.13) are $o_p(1)$, and therefore we have $\left|\widehat{\Omega}_h - \Omega_h\right| = o_p(1)$. Combining with other elements in (A.1.8), all the three terms are bounded by $o_p(1)$ and hence we accomplish $|\widehat{\sigma}_h^2 - \sigma_h^2| = o_p(1)$. □

A.2 Proof of Lemmas

A.2.0.1 Proof of Lemma 1.4.2

Proof. First, consider the dependence bound from (1.4.4). From Lyapunov inequality and the definition of mixingales,

$$E[E[X_t|\mathscr{F}_{t-m}] - EX_t] \leq (E[|E[X_t|\mathscr{F}_{t-m}] - EX_t|^q])^{1/q} \leq c_t\rho_m,$$

and it follows that

$$(6/\varepsilon)\frac{1}{T}\sum_{t=1}^{T}E[E[X_t|\mathscr{F}_{t-m}] - EX_t] \leq (6/\varepsilon)\frac{1}{T}\sum_{t=1}^{T}c_t\rho_m \leq (6\overline{c}_T/\varepsilon)\rho_m.$$

Now consider the tail bound in (1.4.4). From Hölder inequality,

$$\begin{aligned} E\left[|X_t|\,\mathbb{1}\{|X_t| > M\}\right] &\leq \left(E\left[|X_t|^{\overline{q}}\right]\right)^{1/\overline{q}}\left(E\left[\mathbb{1}\{|X_t| > M\}\right]\right)^{1-1/\overline{q}} \\ &\leq \left(E\left[|X_t|^{\overline{q}}\right]\right)^{1/\overline{q}}\left(P\left(|X_t| > M\right)\right)^{1-1/\overline{q}} \\ &\leq \left(E\left[|X_t|^{\overline{q}}\right]\right)^{1/\overline{q}}\left(e^{-\overline{q}M}M_X(t)\right)^{1-1/\overline{q}} \end{aligned}$$

$$\leq C^{1/\overline{q}} C^{1-1/\overline{q}} \exp(-M(\overline{q}-1)) = C\exp(-M(\overline{q}-1)),$$

where the second inequality comes from Lyapunov inequality and the third and fourth from Markov's inequality and Assumption 1 (b) – (c). It follows that

$$(15/\varepsilon)\frac{1}{T}\sum_{t=1}^{T} E\left[|X_t|\,\mathbb{1}\{|X_t|>M\}\right] \leq \frac{15}{\varepsilon} C\exp(-M(\overline{q}-1)),$$

and hence the desired result follows. □

A.2.1 Proof of Lemma 1.4.3

Proof. (a) Recall that $v_{h,t}$ is the projection error from equation (1.2.4). Since x_t and $w_{t,j}$ are $L_{2q}-$NED for all $j = 1,\ldots,p$, the proposed result follows from Theorem **18.8** in Davidson (2021).

(b) Recall that $\varepsilon_{t,h} = \{u_{t,h}, e_{t,h}, v_{t,h}\}$. I start by showing the results for $\varepsilon_{t,h} = u_{t,h}$. By Theorem **18.9** in Davidson (2021), $\{w_{t,j}u_t\}$ are causal L_q-NED of size $-b$ with NED constants $\widetilde{d_t}$ and sequence $\widetilde{\zeta}_m$, where

$$\widetilde{d_t} = \max\{C^{1/2\overline{q}} d_t, d_t^2\}, \qquad \widetilde{\zeta}_m = 2\zeta_m + \zeta_m^2, \tag{A.2.1}$$

with d_t and ζ_m defined in Assumption 1 (d).

I will proceed by showing that $\{w_{t,j}u_t - E[w_{t,j}u_t]\}$ is a causal mixingale for all $j = 1,\ldots,p$, following similar arguments in the proof of Theorem **18.6** in Davidson (2021). Though $E[w_{t,j}u_t] = 0$ by Assumption 1 (a), I maintain the mean extraction to ensure consistency across Lemma 1.4.3 (b) – (c). For simplicity, write $E_s^t[.] = E[.|\mathscr{F}_s^t]$ where $\mathscr{F}_s^t = \sigma(\varepsilon_s,\ldots,\varepsilon_t)$. Also let $k := [q/2]$, the largest integer less or equal to $q/2$. We can start from the left hand side of equation (1.4.1). By Minkowski's inequality,

$$\begin{aligned}\left(E\left[\left|E_{-\infty}^{t-m}[w_{t,j}u_t - E[w_{t,j}u_t]]\right|^q\right]\right)^{1/q} \leq{}& \left(E\left[\left|E_{-\infty}^{t-m}[w_{t,j}u_t - E_{t-k}^t[w_{t,j}u_t]]\right|^q\right]\right)^{1/q} \\ &+ \left(E\left[\left|E_{-\infty}^{t-m}[E_{t-k}^t[w_{t,j}u_t]] - E[w_{t,j}u_t]\right|^q\right]\right)^{1/q}\end{aligned} \tag{A.2.2}$$

holds for all $j = 1,\ldots,p$. Consider the first term.

$$\begin{aligned}\left(E\left[\left|E_{-\infty}^{t-m}[w_{t,j}u_t - E_{t-k}^t[w_{t,j}u_t]]\right|^q\right]\right)^{1/q} &\leq \left(E\left[E_{-\infty}^{t-m}\left[\left|w_{t,j}u_t - E_{t-k}^t[w_{t,j}u_t]\right|^q\right]\right]\right)^{1/q} \\ &= \left(E\left[\left|w_{t,j}u_t - E_{t-k}^t[w_{t,j}u_t]\right|^q\right]\right)^{1/q} \\ &\leq \widetilde{d_t}\widetilde{\zeta}_k,\end{aligned}$$

where the first inequality is the conditional Jensen's inequality and the following equality is the law of iterated expectations (LIE). The last follows from the Definition 4's equation (1.4.3) and (A.2.1). Now consider the second term in equation (A.2.2). Because $E^t_{t-k}[w_{t,j}u_t] - E[w_{t,j}u_t]$ is a finite lag measurable function of $\varepsilon_{t-k},\dots,\varepsilon_t$, it is $\alpha-$mixing of the same size as $\{\varepsilon_t\}$. By the mixing inequality in Theorem **15.2** in Davidson (2021),

$$\begin{aligned}
\left(E\left[\left|E^{t-m}_{-\infty}[E^t_{t-k}[w_{t,j}u_t]] - E[w_{t,j}u_t]]\right|^q\right]\right)^{1/q} &\leq 6\alpha_k^{1/q-1/\overline{q}}\left(E\left[\left|E^t_{t-k}[w_{t,j}u_t]\right|^{\overline{q}}\right]\right)^{1/\overline{q}} \\
&\leq 6\alpha_k^{1/q-1/\overline{q}}\left(E\left[E^t_{t-k}[\left|w_{t,j}u_t\right|^{\overline{q}}]\right]\right)^{1/\overline{q}} \\
&\leq 6\alpha_k^{1/q-1/\overline{q}}\left(E[|w_{t,j}u_t|^{\overline{q}}]\right)^{1/\overline{q}} \\
&\leq 6\alpha_k^{1/q-1/\overline{q}}C^{1/\overline{q}},
\end{aligned}$$

where α_k is the mixing coefficient, the first inequality comes from the conditional Jensen's inequality, the second from LIE, and the last from Assumption 1 (b) and Cauchy-Schwarz inequality. Combining both bounds, equation (A.2.2) is bounded by

$$\begin{aligned}
\left(E\left[\left|E^{t-m}_{-\infty}[w_{t,j}u_t - E[w_{t,j}u_t]]\right|^q\right]\right)^{1/q} &\leq \widetilde{d}_t\widetilde{\zeta}_k + 6\alpha_k^{1/q-1/\overline{q}}C^{1/\overline{q}} \\
&\leq c_t\rho_m,
\end{aligned}$$

where $c_t = \max\{\widetilde{d}_t, C^{1/\overline{q}}\}$ and $\rho_m = 6\alpha_k^{1/q-1/\overline{q}} + 2\widetilde{\zeta}_k$. The above proof applies to all $j = 1,\dots,p$ because, for each j, $w_{t,j}$ shares the same constant $\{d_t\}$ and ζ_m by Assumption 1 (d). Therefore, $\{w_{t,j}u_t\}$ is a causal L_q mixingale with a constant c_t and sequence ρ_m for all $j = 1,\dots,p$.

Because $u_{t,h}$ and $e_{t,h}$ share the same assumptions, the same arguments apply for $\varepsilon_{t,h} = e_{t,h}$. For $\varepsilon_{t,h} = v_{t,h}$, because we have the NED property by (a), the rest of the proof follows the previous arguments.

(d) Since both $w_{t,j}$ and $w_{t,k}$ are $L_{2q}-$NED for all $j \neq k$, the same argument in the proof of (b) applies. □

A.3 Further Details

A.3.1 Finite Lag Approximation in Impulse Response Analysis

This part elaborates on the finite lag approximation in the impulse response analysis application in Section 1.2. When using a finite number of lags L in LP as in (1.2.1), the impulse response estimand from the finite lag order model necessarily includes a bias term that diminishes as L grows. To effectively approximate the infinite lags with a finite number of lags, there should be a condition on how fast p should grow so that the bias term diminishes fast enough. I will fix the notations following Plagborg-Møller and Wolf (2021). Denote the impulse response parameter from LP

with infinite lag as β_h^*, and the finite-lag counterpart as $\beta_h^*(L)$. Define the projection residual $\widetilde{x}_t = x_t - \sum_{\ell=0}^{\infty} \gamma_\ell^{*\prime} w_{t-\ell}$ and the finite lag counterparat as $\widetilde{x}_t(L) = x_t - \sum_{\ell=0}^{L} \gamma_\ell^*(L)' w_{t-\ell}$. Then the following lemma holds.

Lemma 1. Assume the data $\{w_t\}$ are covariance stationary and non-deterministic, with an everywhere nonsingular spectral density matrix and absolutely summable Wold decomposition coefficients. Then, $\beta_h^* = \frac{E[\widetilde{x}_t(L)^2]}{E[\widetilde{x}_t]} \times \beta_h^*(L) + \frac{1}{E[\widetilde{x}_t^2]} \{\sum_{\ell=0}^{\infty} \operatorname{cov}(y_{t+h}, w_{t-\ell})(\gamma_\ell^* - \gamma_\ell^*(L))\}$.

Proof. By the Frisch-Waugh theorem, we can write

$$\begin{aligned}
\beta_h^* =& \frac{\operatorname{cov}(y_{t+h}, \widetilde{x}_t)}{E(\widetilde{x}_t)} \\
=& \frac{\operatorname{cov}(y_{t+h}, \widetilde{x}_t(L)) + (\operatorname{cov}(y_{t+h}, \widetilde{x}_t) - \operatorname{cov}(y_{t+h}, \widetilde{x}_t(L)))}{E(\widetilde{x}_t)} \\
=& \beta_h^*(L) \frac{E(\widetilde{x}_t(L))}{E(\widetilde{x}_t)} + \frac{1}{E(\widetilde{x}_t)} \left(\sum_{\ell=0}^{\infty} \operatorname{cov}(y_{t+h}, w_{t-\ell})(\gamma_\ell^* - \gamma_\ell^*(L)) \right).
\end{aligned}$$

□

Denote the bias term as $\phi_t(L) := \frac{1}{E[\widetilde{x}_t^2]} \{\sum_{\ell=0}^{\infty} \operatorname{cov}(y_{t+h}, w_{t-\ell})(\gamma_\ell^* - \gamma_\ell^*(L))\}$. The bias term shrinks as the projection coefficient $\gamma(L)^*$ get closer to γ^*. This term can be bounded by assuming that the projection coefficients for the lags later than L decays fast enough. We have

$$\begin{aligned}
\phi_t(L) =& \frac{1}{E[\widetilde{x}_t^2]} \{ \sum_{\ell=0}^{\infty} \operatorname{cov}(y_{t+h}, w_{t-\ell})(\gamma_\ell^* - \gamma_\ell^*(L)) \} \\
\leq& \frac{1}{E[\widetilde{x}_t^2]} \sup_\ell |\operatorname{cov}(y_{t+h}, w_{t-\ell})| \, \|\gamma_\ell^* - \gamma_\ell^*(L)\|_1 .
\end{aligned}$$

Now assume $\sup_\ell |\operatorname{cov}(y_{t+h}, w_{t-\ell})| < C$ for some constant $0 < C < \infty$. For bounding $\left\|\gamma_\ell^* - \gamma_\ell^*(L)\right\|_1$, by equation (2.14) of Ing (2020), if $n|L| \leq C(T/\log^3 p)^{1/2}$ where n is the dimension of w_t, then there is some constant $0 < C' < \infty$ such that $\left\|\gamma_\ell^* - \gamma_\ell^*(L)\right\|_1 \leq C' \sum_{\ell=L+1}^{\infty} |\gamma_\ell^*|$. By assuming $\sum_{\ell=L+1}^{\infty} \left\|\gamma_\ell^*\right\|_1 = O_p(p^{-\widetilde{c}})$, where $\underline{c} < \widetilde{c}$ is large enough so that $(T/(\log p)^3)^{1-1/(2\delta)} p^{-\widetilde{c}} = o(1)$, the bias term can be bounded: $\phi_t(L) = o_p(1)$. Note that by Assumption 4 (c), $(T/(\log p)^3)^{1/2} p^{-\underline{c}} = o(1)$. While $(T/(\log p)^3)^{1-1/(2\delta)}$ grows faster, a small increase in $\widetilde{c}$ can still bound it to be $o(1)$, as the growth rate of p is restricted at the exponential level.

A.3.2 Further Definitions

First, I introduce the constants defined in Section 1.2. This entails the assumptions (A1) and (A2) Ing (2020) used in the proof of Theorem 1.4.1, written in (A.1.2), (A.1.3), and (A.1.1). Also denote the constant c on the right hand side

bound of (A.1.2) and (A.1.3) as c_1, and the one in (A.1.1) as c_2. Further denote the constant c in Assumption 4 (a) as c_3. These constants are assumed to be some positive constants. In equation (1.2.9), the parameter M_T^* is defined as

$$M_T^* = \overline{c}\left(\frac{T}{(\log p)^3}\right)^{1/2\delta}, \tag{A.3.1}$$

with a constant $\overline{c}$, which is some small constant that satisfies $0 < \overline{c} < \min\{\overline{\tau}, c_3\}$. $\overline{\tau}$ is defined as

$$\begin{aligned}\overline{\tau} &= \sup \tau \\ &= \sup\left\{\tau \middle| \tau > 0, \limsup_{T\to\infty} \frac{\tau c_2}{\min_{|J|\leq \tau(T/(\log p)^3)^{1/2}} \lambda_{\min}(\Gamma(J))} \leq 1\right\},\end{aligned}$$

where $\Gamma(J) = E[\boldsymbol{w}_t(J)\boldsymbol{w}_t(J)']$ and $\lambda_{\min}(.)$ refers to the minimum eigenvalue of the matrix.

Second, in equation (1.2.8), the parameter C^* is defined as

$$C^* > V_0 := \frac{\overline{B}(c_1 + c_2)}{\sigma_\varepsilon^2}, \tag{A.3.2}$$

where $\sigma_\varepsilon^2 = \min\{\sigma_e^2, \sigma_v^2\}$ with $\sigma_e^2 = \lim_{T\to\infty} 1/T \sum_{t=1}^T e_{t,h}^2$ and $\sigma_v^2 = \lim_{T\to\infty} 1/T \sum_{t=1}^T v_{t,h}^2$. $\overline{B}$ is defined as

$$\overline{B} > \frac{1}{\liminf_{T\to\infty} \min_{|J|\leq M_T^*} \lambda_{\min}(\Gamma(J)) - c_2\overline{c}},$$

where $\overline{c}$ is the constant appeared in (A.3.1).

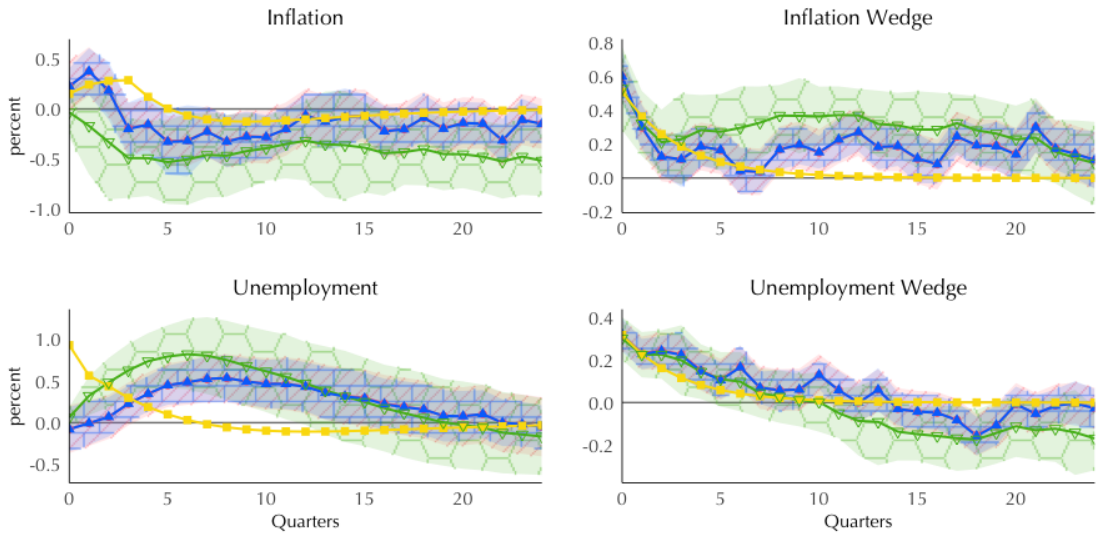
A.4 Additional Empirical Results in Section 1.5.2

Further Results on Debiased LASSO

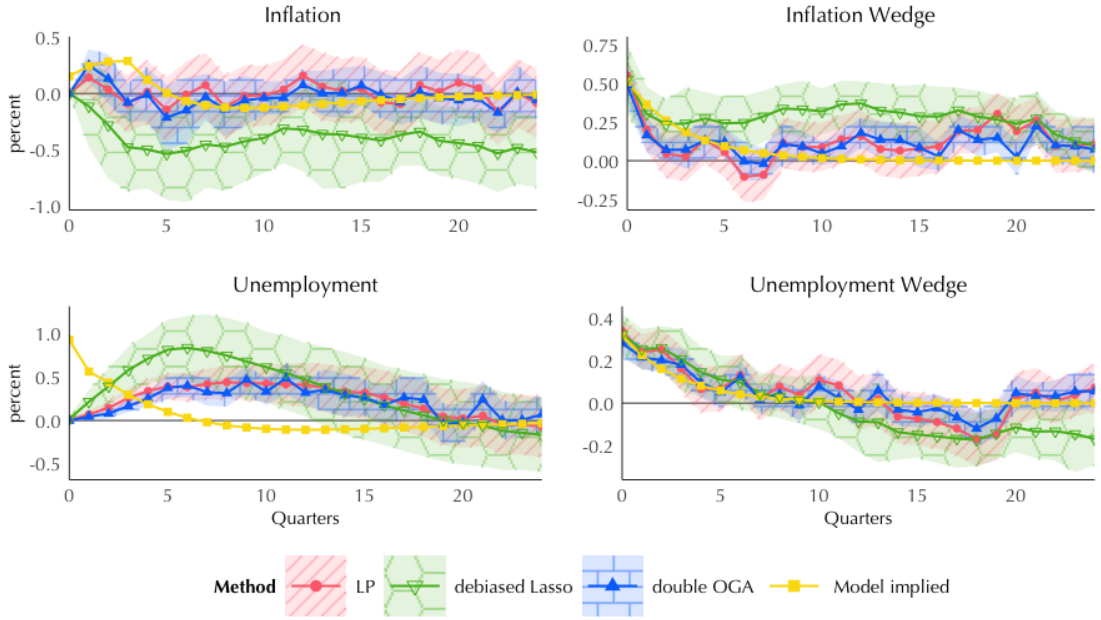
In this subsection, I present additional results on the performance of debiased LASSO in the empirical applications discussed in Section 1.5.2. For comparison purposes, the baseline model from the main text is also included. The patterns observed in the main text persist across these alternative specifications. Debiased LASSO consistently yields highly persistent responses, particularly for inflation and inflation wedges. The unemployment response tends to be overstated, while the unemployment wedge exhibits a persistent negative skew over extended time frames. These observations raise the concerns regarding straightforward theoretical interpretation.

For a comprehensive view, the following Figures present the full array of debiased LASSO results as well as the other methods' estimates. This expanded set of findings further corroborates the notion that debiased LASSO may not be ideally suited for this particular empirical context, likely due to the low-dimensional nature of the models and the

high persistence in the data.



(a) Baseline model with 4 lags

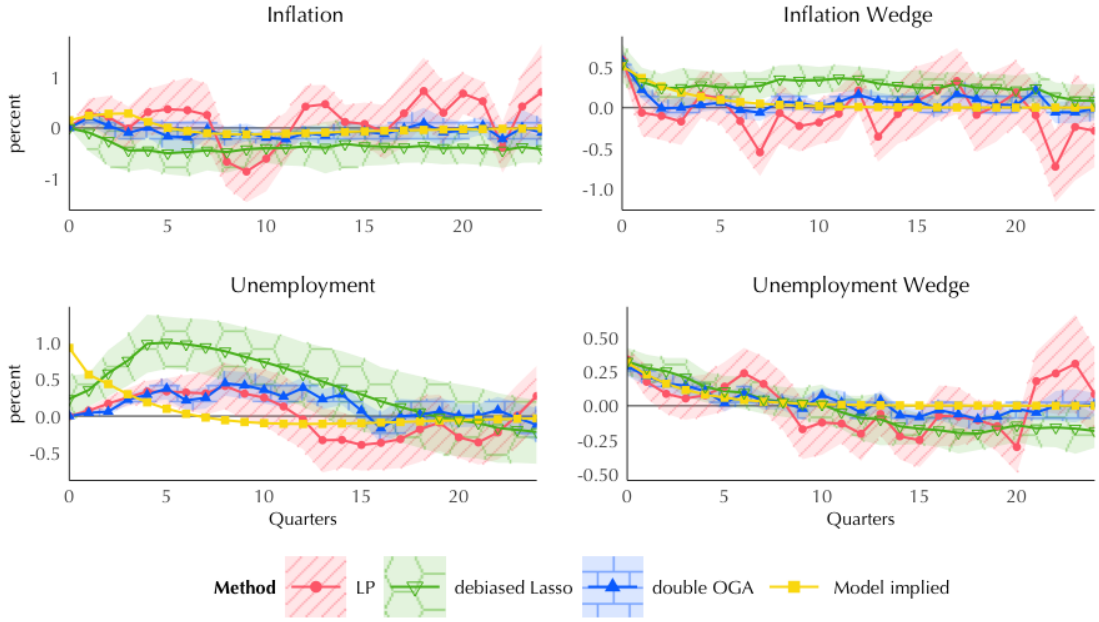


(b) VAR controls with 4 lags

Figure A.1: Baseline and VAR controls added models with 4 lags, all methods

Notes. This figure compares the baseline model with 4 lags (Panel (a)) and the VAR controls added model with 4 lags (Panel (b)) for inflation, inflation wedge, unemployment, and unemployment wedge. Point estimates are shown as solid lines with circle markers for LP (red), inverted triangle markers for debiased LASSO (green), triangle markers for double OGA (blue), and square markers for model-implied (yellow). Shaded areas represent 90% confidence intervals, with diagonal shading for LP, hexagonal shading for debiased LASSO, and brick-shaped shading for double OGA.

(a) VAR controls with 8 lags



(b) VAR controls with 10 lags

Figure A.2: VAR controls added models with longer lags, all methods

Notes. This figure compares the VAR controls added model with 8 lags (Panel (a)) and the VAR controls added model with 10 lags (Panel (b)) for inflation, inflation wedge, unemployment, and unemployment wedge. Point estimates are shown as solid lines with circle markers for LP (red), inverted triangle markers for debiased LASSO (green), triangle markers for double OGA (blue), and square markers for model-implied (yellow). Shaded areas represent 90% confidence intervals, with diagonal shading for LP, hexagonal shading for debiased LASSO, and brick-shaped shading for double OGA.

Appendix B

Appendix to Chapter 2

B.1 Proofs of the Main Result and Auxiliary Lemmas

B.1.1 Proof of Theorem 1

Proof. In this proof, we first show that the convergence rates from Theorem 3.1. of Ing (2020) can be guaranteed under our Assumptions 1-3. It then suffices to demonstrate that Theorem 4.1. in Chernozhukov et al. (2018) can be applied under our assumptions and with such convergence rates. Following Corollary 1 in Belloni et al. (2013b), we fix a sequence of DGPs $(P_N)_{N\in\mathbb{N}}$, $P_N \in \mathscr{P}_N$ and establish the asymptotic statements in order to show uniformity over the sequence of sets of DGPs.

Step 1. We shall verify all the conditions for Theorem 3.1. in Ing (2020). Let us first examine each of the regularity conditions (A1)-(A5) for Theorem 3 in Ing (2020). Note that Assumption 3 (a) and (b), correspond to the conditions (A3) and (A4) in Ing (2020), respectively; Assumption 2 (b) corresponds to the condition (A5) in Ing (2020). Assumption 2 (a) is the additional conditions listed in Theorem 2.1, equations (2.19)-(2.21) in Ing (2020).

Now, we show that Assumption 1(e) and Assumption 3 imply (A1) in Ing (2020). We shall verify that there exists a strictly positive constant c_1 such that

$$P\left(\max_{1\le j\le p}\left|\frac{1}{N}\sum_{i=1}^{N}X_{ij}V_i\right| \ge c_1\sqrt{\frac{\log p}{N}}\right) = o(1). \tag{B.1.1}$$

uniformly over $P \in \mathscr{P}_N$. Set to $Z_{ij} = X_{ij}V_i$, $F = \max_{1\le i\le N}\max_{1\le j\le p}X_{ij}V_i$, and

$$\sigma^2 = \max_{1\le j\le p} E[X_{ij}^2V_i^2] \le \max_{1\le j\le p}\sqrt{E[X_{ij}^4]E[V_i^4]} \le C_q$$

by Assumption 1(e). Then Jensen's inequality and some calculations yield

$$\begin{aligned}
\sqrt{E[F^2]} =& \sqrt{E\left[\max_{1\le i\le N}\max_{1\le j\le p}X_{ij}^2V_i^2\right]} \le \left(E\left[\max_{1\le i\le N}\max_{1\le j\le p}\left|X_{ij}^{q/2}V_i^{q/2}\right|\right]\right)^{2/q} \\
&\le \left(\sum_{i=1}^{N}E\left[\max_{1\le j\le p}\left|X_{ij}^{q/2}V_i^{q/2}\right|\right]\right)^{2/q} \le \left(\sum_{i=1}^{N}\left(E\left[\max_{1\le j\le p}\left|X_{ij}\right|^qE\left[|V_i|^q\right]\right]\right)^{1/2}\right)^{2/q} \\
&\le N^{2/q}\left(E\left[\max_{1\le j\le p}\left|X_{ij}\right|^q\right]E\left[|V_i|^q\right]\right)^{1/q}.
\end{aligned}$$

By applying Lemma 2, we have

$$E\left[\max_{1\leq j\leq p}\left|\frac{1}{N}\sum_{i=1}^{N}X_{ij}V_i\right|\right]\leq C\left\{\sqrt{\frac{\log p}{N}}+K_{N,q}^{1/q}\frac{\log p}{N^{1-2/q}}\right\}.$$

Note that the constant C is universal and the bound depends on the DGP $P\in\mathscr{P}_N$ only via q, N, p, C_q, and $K_{N,1/q}$. Thus the right hand side of the bound is $o(1)$ uniformly as $N\to\infty$ by Assumption 14. A similar argument applies to the case of estimating γ_0, where we replace V with $\mathscr{E}$. Therefore, Assumption 1(e), Assumption 3, and Markov's inequality imply (B.1.1).

Next, we show that Assumption 1(e) and Assumption 3 imply (A2) in Ing (2020). We shall illustrate that there exists a strictly positive constant c_2 such that

$$P\left(\max_{1\leq j,\ell\leq p}\left|\frac{1}{N}\sum_{i=1}^{N}X_{ij}X_{i\ell}-E[X_{1j}X_{1\ell}]\right|\geq c_2\sqrt{\frac{\log p}{N}}\right)=o(1). \tag{B.1.2}$$

We apply Lemma 2 with $Z_{ij\ell}=X_{ij}X_{i\ell}$, $F=\max_{1\leq i\leq N}\max_{1\leq j,\ell\leq p}\left|Z_{ij\ell}\right|$, and $\sigma^2=\max_{1\leq j,\ell\leq p}E[X_{ij}^2X_{i\ell}^2]\leq\max_{1\leq j\leq p}E[X_{ij}^4]\leq C_q$ by Assumption 1(e). Also note that

$$\begin{aligned}\sqrt{E[F^2]}=&\sqrt{E\left[\max_{1\leq i\leq N}\max_{1\leq j,\ell\leq p}X_{ij}^2X_{i\ell}^2\right]}\leq\left(E\left[\max_{1\leq i\leq N}\max_{1\leq j,\ell\leq p}\left|X_{ij}^{q/2}X_{i\ell}^{q/2}\right|\right]\right)^{2/q}\\ \leq&\left(\sum_{i=1}^{N}E\left[\max_{1\leq j,\ell\leq p}\left|X_{ij}^{q/2}X_{i\ell}^{q/2}\right|\right]\right)^{2/q}\leq\left(\sum_{i=1}^{N}\left(E\left[\max_{1\leq j\leq p}\left|X_{ij}\right|^q\right]E\left[\max_{1\leq\ell\leq p}\left|X_{i\ell}\right|^q\right]\right)^{1/2}\right)^{2/q}\\ \leq&N^{2/q}\left(E\left[\max_{1\leq j\leq p}\left|X_{ij}\right|^q\right]\right)^{2/q}.\end{aligned}$$

An application of Lemma 2 yields that

$$E\left[\max_{1\leq j,\ell\leq p}\left|\frac{1}{N}\sum_{i=1}^{N}X_{ij}X_{i\ell}-E[X_{1j}X_{1\ell}]\right|\right]\leq C\left\{\sqrt{\frac{\log p}{N}}+K_{N,q}^{2/q}\frac{\log p}{N^{1-2/q}}\right\},$$

where the last equality is $o(1)$ uniformly as $N\to\infty$ following Assumptions 14 and 3. By Markov's inequality, we conclude that Assumption 1(e) and Assumption 3 imply (B.1.2).

Now, by invoking Theorem 3.1. in Ing (2020) and Equation (3.16), we obtain the convergence rates

$$\|\widehat{\beta}-\beta_0\|^2=O_p\left(\left(\frac{\log p}{N}\right)^{1-1/2\alpha}\right) \tag{B.1.3}$$

for the polynomial decay case, and

$$\|\widehat{\beta}-\beta_0\|^2 = O_p\left(\frac{\log p \log N}{N}\right) \tag{B.1.4}$$

for the exponential decay case., Note that we use N instead of $\left|I_k^c\right|$ since the number of folds K is fixed and the fold size is proportional to N.

Step 2. Given the convergence rates obtained in Step 1, we shall next show that Theorem 4.1. in Chernozhukov et al. (2018) can be applied. To this end, it suffices to verify Assumption 4.1. in Chernozhukov et al. (2018). First, note that Assumption 1 (b)-(e) corresponds to Assumption 4.1 (a)-(d) in Chernozhukov et al. (2018).

Let us now vindicate that Assumption 4.1. (e) in Chernozhukov et al. (2018) is implied by our (B.1.3) and (B.1.4), obtained in Step 1. Observe that (B.1.3) implies that there exists a sequence of events $(A_N)_N$ with probability $P(A_N) = 1-o(1)$ such that conditionally on A_N, we have

$$\|\widehat{\beta}-\beta_0\|^2 \lesssim \left(\frac{\log p}{N}\right)^{1-1/2\alpha}.$$

Therefore, for X, an independent copy of the regressor vector X_i, we have

$$\begin{aligned}
E\left[\left\|X'(\widehat{\beta}-\beta_0)\right\|^2 \mid A_N\right] =& E[E[(X'(\widehat{\beta}-\beta_0))^2|\widehat{\beta},A_N] \mid A_N] && \text{(B.1.5)}\\
=& E[(\widehat{\beta}-\beta_0)'E[XX' \mid \widehat{\beta},A_N](\widehat{\beta}-\beta_0) \mid A_N] && \text{(B.1.6)}\\
=& E[(\widehat{\beta}-\beta_0)'E[XX'](\widehat{\beta}-\beta_0) \mid A_N] && \text{(B.1.7)}\\
=& \lambda_{\max}(E[XX'])\|\widehat{\beta}-\beta_0\|^2, && \text{(B.1.8)}
\end{aligned}$$

where the third equality follows from the fact that the sigma-algebra generated by X is independent of A_N. Observe that (B.1.8) is bounded by Assumption 2(a), (B.1.3), and (B.1.4). Thus the above bound holds with probability at least $1-\Delta_N$ for some $\Delta_N = o(1)$.

Define $\|\cdot\|_{P,q}$ as the $L^q(P)$ norm, where $\|f\|_{P,q} = (\int |f(w)|^q dP(w))^{1/q}$ with P being the law with respect to (Y,D,X). We shall now establish that with P-probability no less than $1-\Delta_N$, $\|\widehat{\eta}-\eta_0\|_{P,q} \leq C$, $\|\widehat{\eta}-\eta_0\|_{P,2} \leq \delta_N$, and $\|\widehat{\beta}-\beta_0\|_{P,2} \times (\|\widehat{\beta}-\beta_0\|_{P,2} + \|\widehat{\gamma}-\gamma_0\|_{P,2}) \leq \delta_N N^{-1/2}$ is satisfied for some Δ_N, δ_N such that both are sequences of strictly positive constants converging to zero.

From a similar argument as in (B.1.5)–(B.1.8),

$$\begin{aligned}\|\widehat{\eta}-\eta_0\|_{P,q} =& \|\widehat{\beta}-\beta_0\|_{P,q} \vee \|\widehat{\gamma}-\gamma_0\|_{P,q} = \left(E\left[\left\|X'(\widehat{\beta}-\beta_0)\right\|^q \mid A_N\right]\right)^{1/q} \\ &\leq \left(E[\left((\widehat{\beta}-\beta_0)'E[XX'](\widehat{\beta}-\beta_0)\right)^{q/2} \mid A_N]\right)^{1/q} = \left(\left(\lambda_{\max}(E[XX'])\|\widehat{\beta}-\beta_0\|^2\right)^{q/2}\right)^{1/q} \\ &= \left(\lambda_{\max}(E[XX'])\|\widehat{\beta}-\beta_0\|^2\right)^{1/2},\end{aligned}$$

where it is bounded by Assumption 2(a), (B.1.3), and (B.1.4). Thus, $\|\widehat{\eta}-\eta_0\|_{P,q} \leq C$ holds with probability at least $1-\Delta_N$ for some $\Delta_N = o(1)$.

Since we use the identical procedure, namely the OGA and HDAIC, in estimating both of β_0 and γ_0, the convergence rates apply to both. We consider two cases where both parameters follow the polynomial decay case or the exponential decay case.

Case 1. For the polynomial decay case, let $\delta_N = (\log p)^{1-1/2\alpha}N^{1/2\alpha-1/2}$. Then $\delta_N = o(1)$ since α is assumed to be strictly larger than 1 in Assumption 3 (a). We have

$$\|\widehat{\eta}-\eta_0\|_{P,2} = \|\widehat{\beta}-\beta_0\|_{P,2} \vee \|\widehat{\gamma}-\gamma_0\|_{P,2} \lesssim \left(\frac{\log p}{N}\right)^{1/2-1/4\alpha} \leq \delta_N,$$

where the last inequality holds since $(\log p/N)^{1/4\alpha} \leq \log p$ holds with $\alpha > 1$. Also,

$$\|\widehat{\beta}-\beta_0\|_{P,2} \times \left(\|\widehat{\beta}-\beta_0\|_{P,2} + \|\widehat{\gamma}-\gamma_0\|_{P,2}\right) \lesssim \left(\frac{\log p}{N}\right)^{1-1/2\alpha} = \delta_N N^{-1/2}$$

holds.

Case 2. For the exponential decay case, let $\delta_N = N^{-1/2}\log p \log N$. By the assumption $\log p = o(N^{1/4})$, $\delta_N = o(1)$. We have

$$\|\widehat{\eta}-\eta_0\|_{P,2} \lesssim \sqrt{\frac{\log p \log N}{N}} < \delta_N$$

and

$$\|\widehat{\beta}-\beta_0\|_{P,2} \times \left(\|\widehat{\beta}-\beta_0\|_{P,2} + \|\widehat{\gamma}-\gamma_0\|_{P,2}\right) \lesssim \frac{\log p \log N}{N} = \delta_N N^{-1/2}.$$

A similar argument applies to the cross cases.

We have shown that Assumption 4.1 in Chernozhukov et al. (2018) holds for both sparsity assumptions. Therefore, applying Theorem 4.1. in Chernozhukov et al. (2018), we get the desired results. □

B.1.2 Local Maximum Inequality

Lemma 2 (Chernozhukov et al. (2015), Lemma 8)**.** Let $(Z_i)_{i=1}^N$ be i.i.d. random vectors, where $Z_i \in \mathbb{R}^p$ with $p \geq 2$. Define $F = \max_{1\leq i\leq N} \max_{1\leq j\leq p} |Z_{ij}|$ and $\sigma^2 = \max_{1\leq j\leq p} E[Z_{ij}^2]$. Then there exists a universal constant $C > 0$ such that

$$E\left[\max_{1\leq j\leq p}\left|\frac{1}{N}\sum_{i=1}^{N} Z_{ij} - E[Z_{ij}]\right|\right] \leq C\left\{\sigma\sqrt{\frac{\log p}{N}} + \frac{\sqrt{E[F^2]}\log p}{N}\right\}.$$

B.2 High-Dimensional Linear IV Regression Models

Section 2.2 in the main text presented the method for high-dimensional linear regression models. In this section, we extend the method by accommodating high-dimensional linear IV regression models.

B.2.1 The Model

Consider the high-dimensional linear IV model

$$Y = D\theta_0 + X'\Lambda_0 + U, \qquad E[U|X,Z] = 0, \tag{B.2.1}$$

$$Z = X'\beta_0 + V, \qquad E[V|X] = 0, \tag{B.2.2}$$

where Z denotes an instrumental variable and the parameter of interest is the partial effect θ_0 of the endogenous treatment variable D on the outcome variable Y. To construct a moment restriction under (B.2.1)–(B.2.2), consider the orthogonal score function

$$\psi(Y,D,X,Z;\theta,\eta) := \{Y - X'\gamma - \theta(D - X'\zeta)\}(Z - X'\beta), \tag{B.2.3}$$

where $X'\gamma_0 = E[Y|X]$, $X'\zeta_0 = E[D|X]$ and $\eta = (\gamma,\zeta,\beta)$.

B.2.2 The Method

This section describes the algorithm for estimation and inference about θ_0 in the high-dimensional linear IV regression model (B.2.1)–(B.2.2).

Observe that this algorithm parallels Algorithm 2, and hence similar remarks are in order. First, the procedure (Steps 1–4) uses the cross fitting to remove an over-fitting bias. Second, the coordinates $\{\widehat{j}_1,...,\widehat{j}_p\}$ are ranked in Step 2 (a)–(c)

Algorithm 4 OGA+HDAIC with DML for high-dimensional linear IV models

S1. Randomly split the sample indices $\{1,...,N\}$ into K folds $(I_k)_{k=1}^K$. For simplicity, let the size of each fold be $n = N/K$ and the size of I_k^c be n^c.

S2. For each fold $k \in \{1,...,K\}$, perform following procedure using $\{(X_i', Z_i)'\}_{i \in I_k^c}$ to get $\widehat{\beta}_k$.

(a) Compute $\widehat{\mu}_{0,j} = X'_{I_k^c j} Z_{I_k^c} / \sqrt{n^c} \| X_{I_k^c j} \|$. Select the coordinate $\widehat{j}_1 = \text{argmax}_{1 \le j \le p} \left| \widehat{\mu}_{0,j} \right|$. Define $\widehat{J}_1 = \{\widehat{j}_1\}$.

(b) Compute $\widehat{\mu}_{1,j} = X'_{I_k^c j} (I_{n^c} - H_1) Z_{I_k^c} / \sqrt{n^c} \| X_{I_k^c j} \|$, where $H_1 = X_{I_k^c \widehat{j}_1} (X'_{I_k^c \widehat{j}_1} X_{I_k^c, \widehat{j}_1})^{-1} X'_{I_k^c, \widehat{j}_1}$. Select the coordinate $\widehat{j}_2 = \text{argmax}_{1 \le j \le p, j \notin \widehat{J}_1} \left| \widehat{\mu}_{1,j} \right|$. Update $\widehat{J}_2 = \widehat{J}_1 \cup \{\widehat{j}_2\}$.

(c) Given $m-1$ coordinates $\widehat{J}_{m-1}$ that have been obtained, compute $\widehat{\mu}_{m-1,j} = X'_{I_k^c j} (I_{n^c} - H_{m-1}) Z_{I_k^c} / \sqrt{n^c} \| X_{I_k^c j} \|$, where $H_{m-1} = X_{I_k^c \widehat{J}_{m-1}} (X'_{I_k^c \widehat{J}_{m-1}} X_{I_k^c \widehat{J}_{m-1}})^{-1} X'_{I_k^c \widehat{J}_{m-1}}$. Select the coordinate $\widehat{j}_m = \text{argmax}_{1 \le j \le p, j \notin \widehat{J}_{m-1}} \left| \widehat{\mu}_{m,j} \right|$. Iteractively update $\widehat{J}_m = \widehat{J}_{m-1} \cup \{\widehat{j}_m\}$.

(d) Compute $\text{HDAIC}(\widehat{J}_m) = (1 + C^* |\widehat{J}_m| \log p / n) \widehat{\sigma}_m^2$ for each m, where C^* is from (2.2.5) in Section 2.2.3 and $\widehat{\sigma}_m^2 = 1/n Z'_{I_k^c} (I - H_m) Z_{I_k^c}$. Choose $\widehat{m} = \text{argmin}_{1 \le m \le M_n^*} \text{HDAIC}(\widehat{J}_m)$, where M_n^* is defined in (2.2.4) in Section 2.2.3.

(e) With coordinates $\widehat{J}_{\widehat{m}}$, run OLS of Z_i on $X_{i\widehat{J}_{\widehat{m}}}$ to get $\widehat{\beta}_k$.

S3. Repeat S2 with $\{(X_i', D_i)'\}_{i \in I_k^c}$ in place of $\{(X_i', Z_i)'\}_{i \in I_k^c}$, to get $\widehat{\zeta}_k$ for each fold $k \in \{1,...,K\}$.

S4. Repeat S2 using $\{(X_i', Y_i)'\}$ to get $\widehat{\gamma}_k$ for each fold $k \in \{1,...,K\}$.

S5. Obtain $\check{\theta}$ as a solution to $1/K \sum_{k=1}^K 1/n \sum_{i \in I_k} \psi(Y_i, D_i, X_i, Z_i; \check{\theta}, \widehat{\eta}_k) = 0$ where $\widehat{\eta}_k = (\widehat{\gamma}_k, \widehat{\beta}_k, \widehat{\zeta}_k)$ and ψ is defined in (B.2.3).

S6. Compute $\widehat{M} = -1/K \sum_{k=1}^K 1/n \sum_{i \in I_k} (D_i - X_i' \widehat{\zeta})(Z_i - X_i' \widehat{\beta})$. Obtain a variance estimator of $\check{\theta}$ as $\widehat{\Omega} = \widehat{M}^{-1} \frac{1}{K} \sum_{k=1}^K \frac{1}{n} \sum_{i \in I_k} [\psi(Y, D, X, Z; \check{\theta}, \widehat{\eta}_k) \psi(Y, D, X, Z; \check{\theta}, \widehat{\eta}_k)'] (\widehat{M}^{-1})'$.

in the order of decreasing importance after successive orthogonalization using OGA as in Ing (2020). Third, a subset $\widehat{J}_{\widehat{m}} = \{\widehat{j}_1, ..., \widehat{j}_{\widehat{m}}\}$ of the ordered set $\{\widehat{j}_1, ..., \widehat{j}_p\}$ is selected in Step 2 (d) using HDAIC as in Ing (2020). The combined use of these three elements (DML, OGA, and HDAIC) together allows for a novel root N consistent estimation of θ_0 without assuming traditional functional class restrictions (e.g., the sparsity) required by existing popular estimators (e.g., LASSO). Section B.2.3 formally presents theoretical arguments in support of this claim.

B.2.3 The Theory

This section follows as a corollary to the main theory in section 2. Again, we use a generic notation $\mathscr{E}$ to refer to $\mathscr{E}_D = D - X'\zeta_0$ and $\mathscr{E}_Y = Y - X'\gamma_0$.

Assumption 8. For each $N \in \mathbb{N}$, it holds that

(a) $(Y_i, D_i, X_i', Z_i)_{i=1}^N$ are i.i.d. copies of (Y, D, X', Z).

(b) (B.2.1) and (B.2.2) hold.

(c) $E[|Y|^q] + E[|D|^q] + E[|Z|^q] \leq C_q$.

(d) $E[|UV|^2] \geq c_q^2$ and $E[DV] \geq c_q$.

(e) $\max_{1 \leq j \leq p} E[|X_{ij}|^q] \leq C_q$, $E[|V|^q] \leq C_q$, and $E[|\mathscr{E}|^q] \leq C_q$.

Furthermore, it holds asymptotically that (f) $K_{N,q}^2 \log p / N^{1-2/q} = o(1)$.

Assumption 9. It holds over $N \in \mathbb{N}$ that

(a) $\lambda_{\min}(\Gamma) \geq \lambda_1 > 0$ and $\lambda_{\max}(\Gamma) \leq C_q$, where $\Gamma = E[XX']$.

(b) Define $\Gamma(J) = E[X_{iJ}X_{iJ}']$ and $d_\ell(J) = E[X_{i\ell}X_{iJ}]$ for the set of coordinate indices J.

$$\max_{1 \leq |J| \leq \overline{C}(N/\log p)^{1/2}, \ell \notin J} \left|\Gamma^{-1}(J) d_\ell(J)\right| < C_q.$$

Assumption 10. For each of $\xi_0 = \beta_0$, ζ_0 and γ_0, ξ_0 follows either (a) or (b) described below.

(a) Polynomial decay: $\log p = o(N^{1-2/q})$. Each ξ_0 is such that $\|\xi_0\|_2^2 \leq C_0$ for some $C_0 > 0$, there exist $\alpha > 1$ such that for any $J \subseteq \mathfrak{P}$,

$$\|\xi_0(J)\|_1 \leq C \left(\|\xi_0(J)\|_2^2\right)^{(\alpha-1)/(2\alpha-1)}.$$

(b) Exponential decay: $\log p = o(N^{1/4})$. Each ξ_0 is such that $\|\xi_0\|_\infty \leq C_0$ for some $C_0 > 0$ and there exists $C_1 > 1$ such that for any $J \subseteq \mathfrak{P}$,

$$\|\xi_0(J)\|_1 \leq C_1 \|\xi_0(J)\|_\infty .$$

Assumptions 8–10 closely parallel Assumptions 1–3, and thus similar remarks apply here. The following theorem supports the estimation and inference procedure presented in Algorithm 4.

Theorem 4. Let $(\mathscr{P}_N)_{N\in\mathbb{N}}$ be a sequence of sets of DGPs such that Assumptions 8–10 are satisfied on the model (B.2.1)–(B.2.2). Then, the estimator $\check{\theta}$ follows

$$\sqrt{N}\left(\check{\theta} - \theta_0\right) \xrightarrow{d} N(0,\Omega),$$

where $\Omega = (E[DV])^{-1}E[V^2U^2](E[DV])^{-1}$. Define $\widehat{M} := -1/K\sum_{k=1}^{K} 1/n\sum_{i\in I_k}(D_i - X_i'\widehat{\zeta})(Z_i - X_i'\widehat{\beta})$. Then, we can define the variance estimator

$$\widehat{\Omega} = \widehat{M}^{-1}\frac{1}{K}\sum_{k=1}^{K}\frac{1}{n}\sum_{i\in I_k}[\psi(Y,D,X,Z;\check{\theta},\widehat{\eta}_k)\psi(Y,D,X,Z;\check{\theta},\widehat{\eta}_k)'](\widehat{M}^{-1})'$$

and the confidence regions with significance level $a \in (0,1)$ have uniform asymptotic validity:

$$\sup_{P\in\mathscr{P}_N}\left|P\left(\theta_0 \in \left[\check{\theta} \pm \Phi^{-1}(1-a/2)\sqrt{\widehat{\Omega}/N}\right]\right) - (1-a)\right| = o(1).$$

A proof is provided in Appendix B.3.3. As in the case of the baseline regression model, we once again emphasize that this result does not rely on the sparsity assumption which is used in the literature on high-dimensional linear models.

Appendix B.5.3 presents simulation designs and results for high-dimensional IV regression models studied in the current appendix section. The results are similar to those obtained for the baseline model presented in Section 2.3. Namely, while our proposed method based on the OGA and HDAIC perform well in terms of all the simulation statistics, the LASSO-based method slightly underperforms and the random-forest-based method significantly underperforms. These differences in the finite-sample performance widen as the degree of polynomial decay becomes smaller.

B.3 Proofs for the Extensions

B.3.1 Proof of Theorem 2

Proof. Throughout this proof, write $\|v\|_A^2 = v^\top Av$ for a vector v and nonnegative definite matrix A. In this proof, we show the prediction norm rates of the nuisance parameters, that are subsequently used in the proof of Theorem 1, to

be attainable with the reduced form models (2.4.3) and (2.4.4). The proof consists of two parts: first we establish the convergence rates of OGA under the setting with approximation errors and, in the second part, the convergence rates of OGA coupled with HDAIC. These two parts correspond to Theorems 2.1 and 3.1. in Ing (2020), respectively. Hence the proof strategies follow closely the proofs of these two results with appropriate modifications.

We consider the estimation of β_0 from (2.4.4) within a partition with sample size n. The same logic applies to γ_0 in (2.4.3).

Step 1. (OGA part 1: Definitions) In this step, we show how to modify the definitions of some objects and sets of events in Ing (2020) to accommodate the presence of the extra approximation errors. Define $X_{\cdot j}$ as the $j-$th coordinate of X. Consider the model (2.4.4) and define

$$\begin{aligned}
D(X) &= \sum_{j=1}^{p} \beta_j X_{\cdot j}, \quad D_J(X) = \sum_{j\in J} \beta_j X_{\cdot j}, \\
\widehat{D}_m(X) &\equiv \widehat{D}_{\widehat{J}_m}(X) = \sum_{j\in \widehat{J}_m} \widehat{\beta}_j X_{\cdot j}, \quad \widehat{D}_{i;\widehat{J}_m} = \sum_{j\in \widehat{J}_m} \widehat{\beta}_j X_{ij}, \\
\mu_{J,k} &= E[(D(X) - D_J(X))X_{\cdot k}]/\sigma_k, \quad \sigma_k = \sqrt{E[X_{ik}^2]}, \\
\widehat{\mu}_{J,k} &= \frac{1/n\sum_{i=1}^n (D_i - \widehat{D}_{i;J})X_{ik}}{(1/n\sum_{i=1}^n X_{ik}^2)^{1/2}},
\end{aligned}$$

and the collections of events

$$A_n(m) = \{\max_{(J,k):|J|\le m-1, k\notin J} |\widehat{\mu}_{J,k} - \mu_{J,k}| \le C(\log p/n)^{1/2}\}, \text{ and} \tag{B.3.1}$$

$$B_n(m) = \{\min_{0\le j\le m-1} \max_{1\le k\le p} \left|\mu_{\widehat{J}_j,k}\right| > \overline{\xi} C(\log p/n)^{1/2}\}, \tag{B.3.2}$$

where $\overline{\xi}, C > 0$ are some large constants.

Now, define the corresponding variables with the approximation errors:

$$\begin{aligned}
D_i^r &= \sum_{j=1}^{p} \beta_j X_{ij} + r_D(X_i) + V_i, \quad D^r(X) = \sum_{j=1}^{p} \beta_j X_{\cdot j} + r_D(X), \\
\mu_{J,k}^r &= E[(D^r(X) - D_J(X))X_{\cdot k}]/\sigma_k = \mu_{J,k} + E[r(X)X_{\cdot k}]/\sigma_k, \\
\widehat{\mu}_{J,k}^r &= \frac{1/n\sum_{i=1}^n (D_i^r - \widehat{D}_{i;J})X_{ik}}{(1/n\sum_{i=1}^n X_{ik}^2)^{1/2}} = \widehat{\mu}_{J,k} + \frac{1/n\sum_{i=1}^n r(X_i)X_{ik}}{(1/n\sum_{i=1}^n X_{ik}^2)^{1/2}},
\end{aligned}$$

and

$$A_n^r(m) = \{ \max_{(J,k):|J|\le m-1, k\notin J} \left|\widehat{\mu}_{J,k}^r - \mu_{J,k}^r\right| \le C(\log p/n)^{1/2}\}, \text{ and} \tag{B.3.3}$$

$$B_n^r(m) = \{ \min_{0\le j\le m-1} \max_{1\le k\le p} \left|\mu_{\widehat{J}_j,k}^r\right| > \widetilde{\xi} C(\log p/n)^{1/2}\}, \tag{B.3.4}$$

where $\widetilde{\xi} = 2/(1-\xi)$ for some $0 < \xi < 1$.

We will show that (B.3.1), (B.3.2) and Assumption 5 imply (B.3.3) and (B.3.4) with appropriate choices on the constants.

$$\begin{aligned}
\left|\widehat{\mu}_{J,k}^r - \mu_{J,k}^r\right| &= \left|\widehat{\mu}_{J,k} - \mu_{J,k} + \frac{1/n\sum_{i=1}^n r(X_i)X_{ik}}{(1/n\sum_{i=1}^n X_{ij}^2)^{1/2}} - E[r(X)X_{\cdot k}]/\sigma_k\right| \\
&\le |\widehat{\mu}_{J,i} - \mu_{J,i}| + \left|\frac{1/n\sum_{i=1}^n r(X_i)X_{ik}}{(1/n\sum_{i=1}^n X_{ij}^2)^{1/2}} - E[r(X)X_{\cdot k}]/\sigma_k\right| \\
&\le |\widehat{\mu}_{J,i} - \mu_{J,i}| + R_{p,3},
\end{aligned}$$

where $R_{p,3} \equiv \max_{1\le k\le p}\left|1/n\sum_{i=1}^n r(X_i)X_{ik}/(1/n\sum_{i=1}^n X_{ij}^2)^{1/2} - E[r(X)X_{\cdot k}]/\sigma_k\right|$ and $R_{p,3} = o_p(1)$ by Assumption 5 (a), (c), and Lemma 2 in Appendix A.2.

Conditional on the events $B_n(m)$,

$$\begin{aligned}
\left|\mu_{\widehat{J}_j,k}^r\right| &= \left|\mu_{\widehat{J}_j,k} + E[r(X)X_{\cdot k}]/\sigma_k\right| \\
&\le \left|\mu_{\widehat{J}_j,k}\right| + |E[r(X)X_{\cdot k}]/\sigma_k| \\
&\le \overline{\xi} C(\log p/n)^{1/2} + C(\log p/n)^{1/2},
\end{aligned}$$

where $C > 0$ and $\overline{\xi}$ is so that $\widetilde{\xi} \equiv \overline{\xi} + 1 = 2/(1-\xi)$.

Using the above definitions, it holds for all $1 \le q \le m$ on $A_n^r(m) \cap B_n^r(m)$,

$$\begin{aligned}
\left|\mu_{\widehat{J}_{q-1},j_q}^r\right| &\ge -\left|\widehat{\mu}_{\widehat{J}_{q-1},\widehat{j}_q}^r - \mu_{\widehat{J}_{q-1},j_q}^r\right| + \left|\widehat{\mu}_{\widehat{J}_{q-1},\widehat{j}_q}^r\right| \\
&\ge -\max_{(J,i):\#(J)\le m-1, i\notin J}\left|\widehat{\mu}_{\widehat{J}_{q-1},\widehat{j}_q}^r - \mu_{\widehat{J}_{q-1},j_q}^r\right| + \left|\widehat{\mu}_{\widehat{J}_{q-1},\widehat{j}_q}^r\right| \\
&\ge -C(\log p_n/n)^{1/2} + \max_{1\le j\le p_n}\left|\widehat{\mu}_{\widehat{J}_{q-1},j}^r\right| \\
&\ge -2C(\log p_n/n)^{1/2} + \max_{1\le j\le p_n}\left|\mu_{\widehat{J}_{q-1},j}^r\right|
\end{aligned}$$

$$> \xi \max_{1\le j\le p_n} \left|\mu^r_{\widehat{J}_{q-1},j}\right|,$$

where the first inequality comes from the triangle inequality, the second from taking the maximum, the third from (B.3.3) and since $\left|\widehat{\mu}^r_{\widehat{J}_{q-1},\widehat{j}_q}\right| = \max_{1\le j\le p_n}\left|\widehat{\mu}^r_{\widehat{J}_{q-1},j}\right|$, the fourth from the triangle inequality and (B.3.3), and the last from $2C(\log p_n/n)^{1/2} < (2/\widetilde{\xi})\max_{1\le j\le p_n}\left|\mu^r_{\widehat{J}_{q-1},j}\right|$ on $B^r_n(m)$ and $1-\xi = 2/\widetilde{\xi}$.

Hence, with Assumption 5 and $\widetilde{\xi} = \overline{\xi}+1$, $A_n(m)$ implies $A^r_n(m)$ and thus $\lim_{n\to\infty} P(A_n(m)) = 1$ implies $\lim_{n\to\infty} P(A^r_n(m)) = 1$. In *Step 3*, we derive the bounds on $A^r_n(m)\cap B^r_n(m)$ and $A^r_n(m)\cap (B^r_n(m))^c$, respectively, and use the fact that $\lim_{n\to\infty} P(A_n(m)) = 1$ to show that it is always the case that either $A^r_n(m)\cap B^r_n(m)$ or $A^r_n(m)\cap (B^r_n(m))^c$ holds.

Step 2. (OGA part 2: Lemma A.1. from Ing (2020)) We now establish error bounds for the population OGA under some high-level conditions (for polynomial decay). Recursively define $J_{\xi,m} = J_{\xi,m-1}\cup\{j_{\xi,m}\}$, with $J_{\xi,0} = \emptyset$ and $j_{\xi,m}$ any element $\ell\in\{1,\dots,p\}$ satisfying

$$|E[V_{m-1}X_{\cdot\ell}]| \ge \xi \max_{1\le j\le p}\left|E[V_{m-1}X_{\cdot j}]\right|. \tag{B.3.5}$$

Denote $V_m = D^r(X) - D_{J_{\xi,m}}(X)$, then

$$\begin{aligned} E[V_m^2] \le & E\left[(D^r(X) - D_{J_{\xi,m}}(X))\sum_{j=1}^p \beta_j X_{\cdot j}\right] \\ \le & \max_{1\le j\le p}\left|\mu^r_{J_{\xi,m},j}\right| \sum_{j=1, j\notin J_{\xi,m}}^p |\beta_j|. \end{aligned} \tag{B.3.6}$$

Recall it can be written that $V_m = X'\beta(J^c_{\xi,m})$. Then Assumption 2 (a) implies

$$E[V_m^2] \ge \lambda_1 \sum_{j=1, j\notin J_{\xi,m}}^p \beta_j^2. \tag{B.3.7}$$

Combining (B.3.6) and (B.3.7), we obtain

$$\begin{aligned} E[V_m^2] \le & C \max_{1\le j\le p}\left|\mu^r_{J_{\xi,m},j}\right| \left(\sum_{j=1, j\notin J_{\xi,m}}^p \beta_j^2\right)^{(\alpha-1)/(2\gamma-1)} \\ \le & C\lambda_1^{-(\alpha-1)/(2\alpha-1)} \max_{1\le j\le p}\left|\mu^r_{J_{\xi,m},j}\right| \{E[V_m^2]\}^{(\alpha-1)/(2\alpha-1)}. \end{aligned} \tag{B.3.8}$$

Note that $\beta(J) = \Gamma^{-1}(J)E[X(J)'D] = \text{argmin}_{c\in\mathbb{R}^{|J|}} E[(D - X'c)^2]$. We now have

$$\begin{aligned}
E[V_{m+1}^2] =& E\left[(D^r(X) - \sum_{j\in J_{\xi,m}} \beta_j(J_{\xi,m+1})X_{\cdot j} - \beta_{j_{m+1}}(J_{\xi,m+1})X_{\cdot j_{\xi,m+1}})^2\right] \\
\leq& E\left[(D^r(X) - \sum_{j\in J_{\xi,m}} \beta_j(J_{\xi,m+1})X_{\cdot j} - \mu^r_{J_{\xi,m},j_{\xi,m+1}}X_{\cdot j_{\xi,m+1}})^2\right] \\
=& E\left[(V_m - \mu^r_{J_{\xi,m},j_{\xi,m+1}}X_{\cdot j_{\xi,m+1}})\right]^2 \\
\leq& E[V_m^2] - \xi^2 \max_{1\leq j\leq p}(\mu^r_{J_{\xi,m},j})^2 \\
\leq& E[V_m^2] - \xi^2\lambda_1^{2(\gamma-1)/(2\gamma-1)}C_\gamma^{-2}\left[\mathrm{E}\left(V_m^2\right)\right]^{2\gamma/(2\gamma-1)} \\
=& \mathrm{E}\left(V_m^2\right)\left\{1 - \xi^2\lambda_1^{2(\gamma-1)/(2\gamma-1)}C_\gamma^{-2}\left[\mathrm{E}\left(V_m^2\right)\right]^{1/(2\gamma-1)}\right\},
\end{aligned} \tag{B.3.9}$$

where the second inequality comes from (B.3.5) and the third inequality comes from (B.3.8). Using (B.3.9) and Lemma 1 of Gao et al. (2013), we obtain the following bound for $G_1 > 0$:

$$E[V_m^2] \leq G_1 m^{-2\alpha+1}. \tag{B.3.10}$$

Step 3. (OGA part 3: Combining Steps 1 and 2) Define the shorthand notations $W_1^N = (Y_i, D_i, X_i')_{i=1}^N$ and $E_{W_1^N} = E[\cdot|W_1^N]$. Combining *Step 1* and *Step 2*, we obtain the bound

$$E_{W_1^N}[(D^r(X) - D_{\widehat{J}_m}(X))^2] \leq G_1 m^{-2\alpha+1} \quad \text{on } A_n^r(m)\cap B_n^r(m). \tag{B.3.11}$$

Using Assumption 6 (a) and (B.3.7), we have for any $0 \leq l \leq m-1$,

$$E_{W_1^N}[(D^r(X) - D_{\widehat{J}_l}(X))^2] \leq \left(C_\alpha \max_{1\leq j\leq p}\left|\mu^r_{\widehat{J}_l,j}\right|\right)^{2-1/\alpha}\lambda_1^{-1+1/\alpha}. \tag{B.3.12}$$

By (B.3.12), we have

$$\begin{aligned}
E_{W_1^N}\left[(D^r(X) - D_{\widehat{J}_m}(X))^2\right] \leq& \min_{0\leq l\leq m-1} E_{W_1^N}\left[(D^r(X) - D_{\widehat{J}_l}(X))^2\right] \\
\leq& C_\alpha^{2-1/\alpha}\lambda_1^{-1+1/\alpha}\left(\min_{0\leq l\leq m-1}\max_{1\leq j\leq p}\left|\mu^r_{\widehat{J}_l,j}\right|\right)^{2-1/\alpha} \\
\leq& C_\alpha^{2-1/\alpha}\lambda_1^{-1+1/\alpha}\left(\widetilde{\xi}C\right)^{2-1/\alpha}(\log p/n)^{1-1/2\alpha},
\end{aligned} \tag{B.3.13}$$

where the last inequality holds conditioning on $(B_n^r(m))^c$.

Combining (B.3.11) and (B.3.13), for all $1 \leq m \leq K_n$ and $C > 0$, we have

$$E_{W_1^N}\left[(D^r(X) - D_{\widehat{J}_m}(X))^2\right] I_{A_n^r(K_n)} \leq C \max\left\{m^{-2\alpha+1}, \{\log p/n\}^{1-1/2\alpha}\right\}. \tag{B.3.14}$$

Under Assumption 5, $\lim_{n\to\infty} P(A_n(m)) = 1$ as shown in Section S1 of supplementary material of Ing (2020), we then have $\lim_{n\to\infty} P(A_n^r(m)) = 1$ following the conclusion of *Step 1*. With (B.3.14) we achieve

$$\max_{1\leq m\leq K_n} \frac{E_{W_1^N}[(D^r(X) - D_{\widehat{J}_m}(X))^2]}{\max\left\{m^{-2\alpha+1}, \{\log p/n\}^{1-1/2\alpha}\right\}} \lesssim_P C. \tag{B.3.15}$$

Note that we are interested in the conditional mean squared prediction error, $E_{W_1^N}[(D^r(X) - \widehat{D}_m(X))^2] = E_{W_1^N}[(D^r(X) - D_{\widehat{J}_m}(X))^2] + E_{W_1^N}[(D_{\widehat{J}_m}(X) - \widehat{D}_m(X))^2]$. The convergence rate for the latter term is

$$\max_{1\leq m\leq K_n} \frac{E_{W_1^N}[(D_{\widehat{J}_m}(X) - \widehat{D}_m(X))^2]}{m \log p/n} \lesssim_P C, \tag{B.3.16}$$

where the proof follows exactly the same arguments as in Section S1 of supplementary material in Ing (2020) under our current setting. Combining (B.3.14) and (B.3.16), we obtain

$$\max_{1\leq m\leq K_n} \frac{E_{W_1^N}[(D^r(X) - \widehat{D}_m(X))^2]}{m^{-2\alpha+1} + m \log p/n} \lesssim_P C. \tag{B.3.17}$$

Step 4. (OGA+HDAIC) Using the results in the previous steps, we now replace m with $\widehat{k}_n$ obtained from HDAIC and establish the convergence rate under such setting. Define

$$\begin{aligned}
V(J) =& D(X) - X(J)'\beta(J), \\
V^r(J) =& V(J) + r(X) = D(X) - X(J)'\beta(J) + r(X), \\
V_i(J) =& D_i - V_i - X_i(J)'\beta(J), \text{ and} \\
V_i^r(J) =& V_i(J) + r(X_i) = D_i - V_i - \sum_{j\in J} \beta_j X_{ij} + r(X_i).
\end{aligned}$$

We will establish the following four inequalities for any $1 \leq m \leq K_n$

$$\left|\frac{1}{n}\sum_{i=1}^n (V_i^r(\widehat{J}_m))^2 - E_{W_1^N}[(V^r(\widehat{J}_m))^2]\right| \leq C R_{1,p}\{E_{W_1^N}[V^2(\widehat{J}_m)]\}^{(\alpha-1)/(2\alpha-1)}, \tag{B.3.18}$$

$$\left|\frac{1}{n}\sum_{i=1}^{n} V_i V_i^r(\widehat{J}_m)\right| \leq CR_{2,p}\{E_{W_1^N}[V^2(\widehat{J}_m)]\}^{(\alpha-1)/(2\alpha-1)}, \tag{B.3.19}$$

$$\max_{1\leq m\leq K_n} \frac{\left\|\frac{1}{n}\sum_{i=1}^{n} X_i(\widehat{J}_m)V_i^r(\widehat{J}_m)\right\|^2_{\widehat{\Gamma}^{-1}(\widehat{J}_m)}}{m\{E_{W_1^N}[V^2(\widehat{J}_m)]\}^{(2\alpha-2)/(2\alpha-1)}} \leq \left\|\widehat{\Gamma}^{-1}(K_n)\right\| CR_{1,p}^2, \tag{B.3.20}$$

$$\max_{1\leq m\leq K_n} \left\|\frac{1}{n}\sum_{i=1}^{n} X_i(\widehat{J}_m)V_i\right\|^2_{\widehat{\Gamma}^{-1}(\widehat{J}_m)} \leq \left\|\widehat{\Gamma}^{-1}(K_n)\right\| R_{2,p}^2, \tag{B.3.21}$$

where $R_{r,1} \equiv \left|1/n\sum_{i=1}^{n} r^2(X_i) - E[r^2(X)]\right| = o_p(1)$ by Assumption 5 (a) and the LLN, $R_{2,p} \equiv \max_{1\leq j\leq p}\left|1/n\sum_{i=1}^{n} X_{ij}V_i\right| \lesssim_P (\log p/n)^{1/2}$ from (B.1.1), and recall that $\|v\|_A^2 = v^\top Av$ for a vector v and nonnegative definite matrix A.

Among the above, we prove (B.3.18)–(B.3.20) since they include the variables with the approximation errors as (B.3.21) does not depend on the newly introduced approximation error in the current result.

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}(V_i^r(\widehat{J}_m))^2 - E_{W_1^N}[(V^r(\widehat{J}_m))^2]\right| \\
&= \left|\frac{1}{n}\sum_{i=1}^{n} V_i^2(\widehat{J}_m) + \frac{2}{n}\sum_{i=1}^{n} V_i(\widehat{J}_m)r(X_i) + \frac{1}{n}\sum_{i=1}^{n} r^2(X_i) - E_{W_1^N}[V^2(\widehat{J}_m)] - E_{W_1^N}[r^2(X)]\right| \\
&\leq \left|\frac{1}{n}\sum_{i=1}^{n} V_i^2(\widehat{J}_m) - E_{W_1^N}[V^2(\widehat{J}_m)]\right| + 2\left|\frac{1}{n}\sum_{i=1}^{n} V_i(\widehat{J}_m)r(X_i)\right| + \left|\frac{1}{n}\sum_{i=1}^{n} r^2(X_i) - E_{W_1^N}[r^2(X)]\right| \\
&\leq \underbrace{\left|\frac{1}{n}\sum_{i=1}^{n} V_i^2(\widehat{J}_m) - E_{W_1^N}[V^2(\widehat{J}_m)]\right|}_{(a)} + 2\underbrace{\sqrt{\frac{1}{n}\sum_{i=1}^{n} V_i^2(\widehat{J}_m)}\sqrt{\frac{1}{n}\sum_{i=1}^{n} r^2(X_i)}}_{(b)} + \underbrace{R_{r,1}}_{(c)},
\end{aligned}$$

where we want to show that $(b) \lesssim_P (a)$. Note that from Assumption 6 (a) and (B.3.7) as shown in Section S2 of supplementary material of Ing (2020), we have

$$\begin{aligned}
\left|\frac{1}{n}\sum_{i=1}^{n} V_i^2(\widehat{J}_m)\right| &\lesssim_P E_{W_1^N}[V^2(\widehat{J}_m)] + CR_{1,p}\{E_{W_1^N}[V^2(\widehat{J}_m)]\}^{(\alpha-1)/(2\alpha-1)}, \\
&\lesssim_P E_{W_1^N}[V^2(\widehat{J}_m)],
\end{aligned}$$

where the second inequality comes from Assumption 4 (d), and hence

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n} V_i^2(\widehat{J}_m)}\sqrt{\frac{1}{n}\sum_{i=1}^{n} r^2(X_i)} \lesssim_P E_{W_1^N}[V^2(\widehat{J}_m)]^{1/2}(\log p/n)^{1/2} \lesssim_P (a),$$

where the first bound comes from Assumption 5 (b) and the last comes from Assumption 4 (d). Since $(b) \lesssim_P (a)$ and $(c) = o_p(1)$ by Assumption 5 (a), we obtain (B.3.18).

Now, note that

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^{n}V_iV_i^r(\widehat{J}_m)\right| &= \left|\frac{1}{n}\sum_{i=1}^{n}V_iV_i(\widehat{J}_m)+\frac{1}{n}\sum_{i=1}^{n}V_ir(X_i)\right| \\ &\le \left|\frac{1}{n}\sum_{i=1}^{n}V_iV_i(\widehat{J}_m)\right|+\left|\frac{1}{n}\sum_{i=1}^{n}V_ir(X_i)\right| \\ &\le \underbrace{\left|\frac{1}{n}\sum_{i=1}^{n}V_iV_i(\widehat{J}_m)\right|}_{(d)}+\underbrace{\sqrt{\frac{1}{n}\sum_{i=1}^{n}V_i^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}r^2(X_i)}}_{(e)}.\end{aligned}$$

By an argument similar to deriving (B.3.18) and since term (e) above is bounded by a constant from Assumption 4 (e) and 5 (a), we obtain (B.3.19).

Next, observe that

$$\begin{aligned}&\left\|\frac{1}{n}\sum_{i=1}^{n}X_i(\widehat{J}_m)V_i^r(\widehat{J}_m)\right\|_{\widehat{\Gamma}^{-1}(m)} \\ &\le \underbrace{\left\|\widehat{\Gamma}^{-1}(K_n)\right\|^{1/2}\left\|\frac{1}{n}\sum_{i=1}^{n}X_i(\widehat{J}_m)V_i(\widehat{J}_m)\right\|}_{(f)}+\underbrace{\left\|\widehat{\Gamma}^{-1}(K_n)\right\|^{1/2}\left\|\frac{1}{n}\sum_{i=1}^{n}X_i(\widehat{J}_m)r(X_i)\right\|}_{(g)}.\end{aligned}$$

By a similar manipulations as in (B.3.18) and since

$$\left\|1/n\sum_{i=1}^{n}X_i(\widehat{J}_m)r(X_i)\right\| \le \left\|1/n\sum_{i=1}^{n}X_i(\widehat{J}_m)r(X_i)\right\|_1 \lesssim_P (\log p/n)^{1/2},$$

where the last inequality holds from Assumption 5 (a), (c), and Lemma 2 in Appendix A.2., we achieve (B.3.20).

Recall $M_n^* = \min\{(n/\log p)^{1/2\alpha}, \overline{\delta}(n/\log p)^{1/2}\}$ and let $\widetilde{k}_n = \min_{1\le k\le K_n}\{E_{W_1^N}(V^2(\widehat{J}_k)) \le GM_n^{*-2\alpha+1}\}$ where $G > C$ is an appropriate constant that is large enough, where C is defined in (B.3.14).

Using (B.3.18)–(B.3.21), it follows exactly the same the proof shown in Ing (2020) under our current setting that

$$\lim_{n\to\infty} P(\widehat{k}_n \le \widetilde{k}_n - 1) = 0, \tag{B.3.22}$$

$$\lim_{n\to\infty} P(\widehat{k}_n \ge CM_n^*) = 0, \tag{B.3.23}$$

hold and hence we have the following results:

$$E_{W_1^N}[(D^r(X) - \widehat{D}_{\widehat{k}_n}(X))^2]I_{\left\{\widetilde{k}_n \le \widehat{k}_n < CM_n^*\right\}} = O_p\left(M_n^{*-2\gamma+1}\right), \tag{B.3.24}$$

and the desired result follows:

$$E_{W_1^N}[(D^r(X) - \widehat{D}_{\widehat{k}_n}(X))^2] \lesssim_P (\log p/N)^{1-1/2\alpha}. \tag{B.3.25}$$

Following the same argument as in the proof of Theorem 1, the conditions for Assumption 4.1. in Chernozhukov et al. (2018) is satisfied and thus by applying Theorem 4.1. in Chernozhukov et al. (2018) we obtain the desired asymptotic normality results. □

B.3.2 Proof of Theorem 3

Proof. For notational simplicity, we write $\mathscr{Y} = [Y_1,\ldots,Y_N]'$, $\mathscr{X} = [X_1,\ldots,X_N]'$, $\mathscr{D} = [D_1,\ldots,D_N]'$, $\mathscr{V} = [V_1,\ldots,V_N]'$, $\mathscr{U} = [U_1,\ldots,U_N]'$, $\mathscr{R}_{Y_\theta} = [r_{Y_\theta}(X_1),\ldots,r_{Y_\theta}(X_N)]'$, $\mathscr{R}_D = [r_D(X_1),\ldots,r_D(X_N)]'$, and $f = [f(X_1),\ldots,f(X_N)]'$ and $g = [g(X_1),\ldots,g(X_N)]'$ with

$$f(X) = X'\Lambda_0 + r_{Y_\theta}(X), \qquad g(X) = X'\beta_0 + r_D(X).$$

Define for a non-empty set of coordinate indices $J \subseteq \mathfrak{P}$ that $X_{[N]J} = \{X_{ij}, i \in \{1,\ldots,N\}, j \in J\}$, $P_J = X_{[N]J}(X'_{[N]J}X_{[N]J})^{-}X'_{[N]J}$, and $M_J = I_N - P_J$, where I_N is an N-dimensional identity matrix. Define the indices of chosen coordinates from Algorithm 3 using $(\mathscr{X},\mathscr{D})$ as $\widetilde{J}^{\mathscr{D}}_{\widehat{m}}$, $(\mathscr{X},\mathscr{Y})$ as $\widetilde{J}^{\mathscr{Y}}_{\widehat{m}}$, and let $\widetilde{J} = \widetilde{J}^{\mathscr{D}}_{\widehat{m}} \cup \widetilde{J}^{\mathscr{Y}}_{\widehat{m}}$.

First note that

$$\widetilde{\theta} = \left(\frac{1}{N}\mathscr{D}'M_{\widetilde{J}}\mathscr{D}\right)^{-1}\left(\frac{1}{N}\mathscr{D}'M_{\widetilde{J}}Y\right), \text{ and}$$
$$\sqrt{N}\left(\widetilde{\theta} - \theta_0\right) = \underbrace{\left(\frac{1}{N}\mathscr{D}'M_{\widetilde{J}}\mathscr{D}\right)^{-1}}_{=\mathbf{A}^{-1}}\underbrace{\left(\frac{1}{\sqrt{N}}\mathscr{D}'M_{\widetilde{J}}(f+\mathscr{U})\right)}_{=\mathbf{B}}.$$

Thus, if we show

$$\mathbf{A} = \frac{1}{N}\mathscr{V}'\mathscr{V} + o_p(1) \quad \text{and} \quad \mathbf{B} = \frac{1}{\sqrt{N}}\mathscr{V}'\mathscr{U} + o_p(1), \tag{B.3.26}$$

then an application of CLT yields the desired conclusion. In *Step 1* we derive component-wise bounds that will be used in the following steps. We show (B.3.26) in *Step 2* and in *Step 3* we conclude.

Step 1. (Component-wise bounds) Note we have $R_{2,p} = \max_{1\le j\le p}\left|1/N\sum_{i=1}^N X_{ij}V_i\right| \lesssim_P (\log p/N)^{1/2}$ from (B.1.1).

First we derive bounds for $\left\|\frac{1}{N}\mathscr{X}'\mathscr{U}\right\|_\infty$:

$$\left\|\frac{1}{N}\mathscr{X}'\mathscr{V}\right\|_\infty \leq R_{2,p} \lesssim_P (\log p/N)^{1/2}, \tag{B.3.27}$$

and similarly $\|\mathscr{X}'\mathscr{U}/N\|_\infty \lesssim_P (\log p/N)^{1/2}$.

Note $\widetilde{\beta}(J) = (X'_{[N]J}X_{[N]J})^{-1}X'_{[N]J}\mathscr{D}$. From the convergence rates from Ing (2020), we have

$$\begin{aligned}\left\|\widetilde{\beta}(\widetilde{J}) - \beta_0\right\| &\lesssim_P \left\|\widetilde{\beta}(\widehat{J}^D_{\widehat{m}}) - \beta_0\right\| \\ &\leq \left(\lambda_1^{-1}E_{W_1^N}[(D^r(X) - \widehat{D}_m(X))^2]\right)^{1/2} \\ &\lesssim_P (\log p/N)^{(2\alpha-1)/4\alpha},\end{aligned} \tag{B.3.28}$$

where the first inequality comes from $\widehat{J}^D_{\widehat{m}} \subseteq \widetilde{J}$, the second from Assumption 2 (a) and Equation (3.16) of Ing (2020), and the last from the results in Theorem 3.1. of Ing (2020).

In what followings, we establish a bound for $\|\widetilde{\beta}(\widetilde{J}) - \beta_0\|_1$. Recall $\beta_0(J) = (\beta_{0j}, j = 1;\ldots,p)$, where $\beta_{0j} = 0$ for $j \notin J$. Notice that

$$\left\|\widetilde{\beta}(\widetilde{J}) - \beta_0\right\|_1 = \left\|\widetilde{\beta}(\widetilde{J}) - \beta_0(\widetilde{J})\right\|_1 + \left\|\beta_0(\widetilde{J^c})\right\|_1 = \left\|\widetilde{\beta}(\widetilde{J}) - \beta_0(\widetilde{J})\right\|_1 + \left\|\beta_0 - \beta_0(\widetilde{J})\right\|_1.$$

The first term on RHS is bounded by

$$\begin{aligned}\left\|\widetilde{\beta}(\widetilde{J}) - \beta_0(\widetilde{J})\right\|_1 &\leq \sqrt{\left|\widetilde{J}\right|}\left\|\widetilde{\beta}(\widetilde{J}) - \beta_0(\widetilde{J})\right\| \\ &\leq \sqrt{\widehat{m}^D + \widehat{m}^Y}\left\|\widetilde{\beta}(\widehat{J}^D_{\widehat{m}}) - \beta_0(\widehat{J}^D_{\widehat{m}})\right\| \\ &\lesssim_P (M_N^*)^{1/2}(\log p/N)^{(2\alpha-1)/4\alpha} \\ &\lesssim_P (\log p/N)^{-1/4\alpha}(\log p/N)^{(2\alpha-1)/4\alpha} \\ &\lesssim_P (\log p/N)^{(\alpha-1)/2\alpha},\end{aligned}$$

where the second inequality comes from $\widetilde{J} = \widehat{J}^D_{\widehat{m}} \cup \widehat{J}^Y_{\widehat{m}}$, and the third from (B.3.23). The fourth comes from the definition of M_N^* defined in (2.2.4), and the last follows. On the other hand, the second term on RHS can be controlled by

$$\left\|\beta_0 - \beta_0(\widetilde{J})\right\|_1 \leq C\sum_{j\notin\widetilde{J}}\left|\beta_{0j}\right|$$

$$
\begin{aligned}
&\leq CC_\alpha \left(\sum_{j \notin \widetilde{J}} {\beta_0}_j^2 \right)^{(\alpha-1)/(2\alpha-1)} \\
&\leq CC_\alpha \lambda_1^{(-\alpha+1)/(2\alpha-1)} \left(E[V(\widetilde{J})^2] \right)^{(\alpha-1)/(2\alpha-1)} \\
&\leq CC_\alpha \lambda_1^{(-\alpha+1)/(2\alpha-1)} G_1 |\widetilde{J}|^{-2\alpha+1} \\
&\leq CC_\alpha \lambda_1^{(-\alpha+1)/(2\alpha-1)} G_1 (2M_n^*)^{-2\alpha+1} \\
&\leq CC_\alpha \lambda_1^{(-\alpha+1)/(2\alpha-1)} G_1 (2)^{-2\alpha+1} (\log p/N)^{(2\alpha-1)/2\alpha} \\
&\lesssim_P (\log p/N)^{(2\alpha-1)/2\alpha}
\end{aligned}
$$

The first inequality comes from Assumption 2 (b), where it holds for all $J \subseteq \mathfrak{P}$ such that $|J| \leq C(N/\log p)^{1/2}$, as shown in Ing (2020) Equation (2.16) and the following equation. The second inequality comes from Assumption 7. The third comes from (B.3.7), which holds under Assumption 2 (a). The fourth comes from (B.3.10) in the previous section's proof. The fifth comes from $|\widetilde{J}| \leq |\widehat{J}_{\widehat{m}}^{\mathcal{D}}| + |\widehat{J}_{\widehat{m}}^{\mathcal{Y}}|$, where $|\widehat{J}_{\widehat{m}}| \leq M_N^*$. The sixth comes from the definition of M_N^* given in (2.2.4), and the last follows. By combining the bounds, we conclude that

$$
\left\| \widetilde{\beta}(\widetilde{J}) - \beta_0 \right\|_1 \lesssim_P (\log p/N)^{(\alpha-1)/4\alpha} + (\log p/N)^{(2\alpha-1)/2\alpha} \lesssim_P (\log p/N)^{(\alpha-1)/4\alpha}. \tag{B.3.29}
$$

It also holds that

$$
\begin{aligned}
\left\| \frac{1}{\sqrt{N}} M_{\widetilde{J}} g \right\| &\leq \left\| \frac{1}{\sqrt{N}} M_{\widehat{J}_{\widehat{m}}^{\mathcal{D}}} g \right\| \\
&\leq \left\| \frac{1}{\sqrt{N}} \left(\mathscr{X} \widetilde{\beta}(\widehat{J}_{\widehat{m}}^{\mathcal{D}}) - g \right) \right\| \\
&\leq \left\| \frac{1}{\sqrt{N}} \mathscr{X} \left(\widetilde{\beta}(\widehat{J}_{\widehat{m}}^{\mathcal{D}}) - \beta_0 \right) \right\| + \left\| \frac{1}{\sqrt{N}} \mathscr{R}_D \right\| \\
&\lesssim_P (\log p/N)^{(2\alpha-1)/4\alpha} + \sqrt{\frac{1}{N} \sum_{i=1}^{N} r_D^2(X_i)} \\
&\lesssim_P (\log p/N)^{(2\alpha-1)/4\alpha} + \sqrt{\log p/N} \\
&\lesssim_P (\log p/N)^{(2\alpha-1)/4\alpha},
\end{aligned} \tag{B.3.30}
$$

where the fourth comes from the convergence rates from (B.3.25), and the fifth from Assumption 5 (b), and the remainder follows.

Similarly, note the convergence rate for $\|\widetilde{\gamma}(\widehat{J}^Y_{\widehat{m}}) - \gamma_0\|$ is of the same order as in (B.3.28), hence we have

$$\begin{aligned}
\left\| \frac{1}{\sqrt{N}} M_{\widetilde{J}}(\theta_0 g + f) \right\| &\leq \left\| \frac{1}{\sqrt{N}} \left(\mathscr{X} \widetilde{\gamma}(\widetilde{J}) - (\theta_0 g + f) \right) \right\| \\
&\leq \left\| \frac{1}{\sqrt{N}} \left(\mathscr{X} \widetilde{\gamma}(\widehat{J}^Y_{\widehat{m}}) - (\theta_0 g + f) \right) \right\| \\
&\leq \left\| \frac{1}{\sqrt{N}} \mathscr{X} \left(\widetilde{\gamma}(\widehat{J}^Y_{\widehat{m}}) - \gamma_0 \right) \right\| + \left\| \frac{1}{\sqrt{N}} \mathscr{R}_Y \right\| \\
&\lesssim_P (\log p/N)^{(2\alpha-1)/4\alpha},
\end{aligned}$$

where the second inequality comes from $\widehat{J}^Y_{\widehat{m}} \subseteq \widetilde{J}$, the third from triangle inequality, and the last follows Assumption 5 (b) and the convergence rate of $\|\widetilde{\gamma}(\widehat{J}^Y_{\widehat{m}}) - \gamma_0\|$. Using triangle inequality,

$$\left| \left\| M_{\widetilde{J}} \theta_0 g \right\| - \left\| M_{\widetilde{J}} f \right\| \right| \leq \left\| M_{\widetilde{J}} (\theta_0 g + f) \right\|,$$

where $\|M_{\widetilde{J}} \theta_0 g / \sqrt{N}\| = \|\theta_0\| \|\mathscr{X}(\widetilde{\beta}(\widetilde{J}) - \beta_0)/\sqrt{N}\| \lesssim_P (\log p/N)^{(2\alpha-1)/4\alpha}$ by the assumption on bounded $\|\theta_0\|$ in Assumption 7. Therefore, $\left\| M_{\widetilde{J}} f \right\|$ is bounded by the same bound and

$$\left\| \widetilde{\Lambda}(\widetilde{J}) - \Lambda_0 \right\| \lesssim_P \left\| \frac{1}{\sqrt{N}} \mathscr{X} \left(\widetilde{\Lambda}(\widetilde{J}) - \Lambda_0 \right) \right\| \lesssim_P \left\| \frac{1}{\sqrt{N}} M_{\widetilde{J}} f \right\| \lesssim_P (\log p/N)^{(2\alpha-1)/4\alpha} \tag{B.3.31}$$

by a similar argument as in (B.3.30).

Let $\widetilde{\beta}_V(J) = (\mathscr{X}(J)' \mathscr{X}(J))^{-1} \mathscr{X}(J)' \mathscr{V}$.

$$\begin{aligned}
\left\| \widetilde{\beta}_V(\widetilde{J}) \right\|_1 &\leq \sqrt{|\widetilde{J}|} \left\| \widetilde{\beta}_V(\widetilde{J}) \right\|_2 \\
&\leq \sqrt{|\widetilde{J}|} \left\| \widehat{\Gamma}^{-1}(\widetilde{J}) \right\|_2 \left\| \frac{1}{N} \mathscr{X}(\widetilde{J})' \mathscr{V} \right\|_2 \\
&\leq |\widetilde{J}| \left\| \widehat{\Gamma}^{-1}(\widetilde{J}) \right\|_2 \left\| \frac{1}{N} \mathscr{X}(\widetilde{J})' \mathscr{V} \right\|_\infty \\
&\lesssim_P (\log p/N)^{(\alpha-1)/2\alpha}.
\end{aligned} \tag{B.3.32}$$

The last inequality comes from $\lim_{N\to\infty} P(\|\widehat{\Gamma}^{-1}(\widehat{J}_{K_N})\| \leq \overline{B}) = 1$ as shown in Section S1 of the Supplementary Material of Ing (2020), which holds under current setting and $\overline{B}$ is some large constant defined in Theorem 2.1. of Ing (2020).

The last component is

$$\left| \frac{1}{\sqrt{N}} \mathscr{R}_D' \mathscr{U} \right| \lesssim_P \sqrt{E\left[\frac{1}{N} \sum_{i=1}^N r_D^2(X_i) \right]} \lesssim_P (\log p/N)^{1/2}, \tag{B.3.33}$$

where the first inequality comes from Chebyshev and the last from 5 (b). The same logic applies to $\left|\mathscr{R}'_{Y_\theta}\mathscr{V}/\sqrt{N}\right|$.

Step 2. (Bounding **A** and **B**) Decompose the two objects in (B.3.26) into

$$\mathbf{A} = \frac{1}{N}\mathscr{V}'\mathscr{V} + \underbrace{\frac{1}{N}f'M_{\widetilde{J}}f}_{(a)} + \underbrace{\frac{2}{N}f'M_{\widetilde{J}}\mathscr{V}}_{(b)} - \underbrace{\frac{1}{N}\mathscr{V}'P_{\widetilde{J}}\mathscr{V}}_{(c)},$$
$$\mathbf{B} = \frac{1}{\sqrt{N}}\mathscr{V}'\mathscr{U} + \underbrace{\frac{1}{\sqrt{N}}g'M_{\widetilde{J}}f}_{(d)} + \underbrace{\frac{1}{\sqrt{N}}g'M_{\widetilde{J}}\mathscr{U}}_{(e)} + \underbrace{\frac{1}{\sqrt{N}}\mathscr{V}'M_{\widetilde{J}}f}_{(f)} - \underbrace{\frac{1}{\sqrt{N}}\mathscr{V}'P_{\widetilde{J}}\mathscr{U}}_{(g)},$$

where the components (a)-(g) can be further controlled by

$$\begin{aligned}
|(a)| &\le \left\|\frac{1}{\sqrt{N}}M_{\widetilde{J}}f\right\|^2 \lesssim_P (\log p/N)^{(2\alpha-1)/2\alpha},\\
|(b)| &\le 2\left|\frac{1}{N}\mathscr{R}'_{Y_\theta}\mathscr{V}\right| + 2\left|\left(\widetilde{\Lambda}(\widetilde{J})-\Lambda_0\right)'\frac{1}{N}\mathscr{X}'\mathscr{V}\right|\\
&\le 2\left|\frac{1}{N}\mathscr{R}'_{Y_\theta}\mathscr{V}\right| + 2\left\|\widetilde{\Lambda}(\widetilde{J})-\Lambda_0\right\|_1\left\|\frac{1}{N}\mathscr{X}'\mathscr{V}\right\|_\infty\\
&\lesssim_P \sqrt{N}^{-1}(\log p/N)^{1/2} + (\log p/N)^{(\alpha-1)/4\alpha}(\log p/N)^{1/2},\\
|(c)| &\le \left|\widetilde{\beta}_V(\widetilde{J})'\frac{1}{N}\mathscr{X}'\mathscr{V}\right| \le \left\|\widetilde{\beta}_V(\widetilde{J})\right\|_1\left\|\frac{1}{N}\mathscr{X}'\mathscr{V}\right\|_\infty\\
&\lesssim_P (\log p/N)^{(\alpha-1)/2\alpha}(\log p/N)^{1/2},
\end{aligned} \tag{B.3.34}$$

and

$$\begin{aligned}
|(d)| &\le \sqrt{N}\left\|\frac{1}{\sqrt{N}}M_{\widetilde{J}}f\right\|\left\|\frac{1}{\sqrt{N}}M_{\widetilde{J}}g\right\| \lesssim_P \sqrt{N}(\log p/N)^{(2\alpha-1)/2\alpha},\\
|(e)| &\le \left|\frac{1}{\sqrt{N}}\mathscr{R}'_D\mathscr{U}\right| + \left|\left(\widetilde{\beta}(\widetilde{J})-\beta_0\right)'\frac{1}{\sqrt{N}}\mathscr{X}'\mathscr{U}\right|\\
&\le \left|\frac{1}{\sqrt{N}}\mathscr{R}'_D\mathscr{U}\right| + \left\|\widetilde{\beta}(\widetilde{J})-\beta_0\right\|_1\left\|\frac{1}{\sqrt{N}}\mathscr{X}'\mathscr{U}\right\|_\infty\\
&\lesssim_P (\log p/N)^{1/2} + \sqrt{N}(\log p/N)^{(\alpha-1)/4\alpha}(\log p/N)^{1/2},\\
|(f)| &\le \left|\frac{1}{\sqrt{N}}\mathscr{R}'_{Y_\theta}\mathscr{V}\right| + \left|\left(\widetilde{\Lambda}(\widetilde{J})-\Lambda_0\right)'\frac{1}{\sqrt{N}}\mathscr{X}'\mathscr{V}\right|\\
&\le \left|\frac{1}{\sqrt{N}}\mathscr{R}'_{Y_\theta}\mathscr{V}\right| + \left\|\widetilde{\Lambda}(\widetilde{J})-\Lambda_0\right\|_1\left\|\frac{1}{\sqrt{N}}\mathscr{X}'\mathscr{V}\right\|_\infty\\
&\lesssim_P (\log p/N)^{1/2} + \sqrt{N}(\log p/N)^{(\alpha-1)/4\alpha}(\log p/N)^{1/2},\\
|(g)| &\le \left|\widetilde{\beta}_V(\widetilde{J})'\frac{1}{\sqrt{N}}\mathscr{X}'\mathscr{U}\right| \le \left\|\widetilde{\beta}_V(\widetilde{J})\right\|_1\left\|\frac{1}{\sqrt{N}}\mathscr{X}'\mathscr{U}\right\|_\infty\\
&\lesssim_P (\log p/N)^{(\alpha-1)/2\alpha}\sqrt{N}(\log p/N)^{1/2}.
\end{aligned} \tag{B.3.35}$$

Now, we show that each part in (B.3.34) and (B.3.35) is $o_p(1)$. Note that $|(a)|$, $|(b)|$, and $|(c)|$ are $o(1)$ if $|(d)|$, $|(e)|$, and $|(g)|$ are all $o(1)$. Since

$$
\begin{aligned}
|(d)| &\lesssim_P \sqrt{N}(\log p/N) = o(1), \\
|(e)| &\lesssim_P \sqrt{N}(\log p/N)^{(\alpha-1)/4\alpha}(\log p/N)^{1/2} = o(1), \\
|(g)| &\lesssim_P (\log p/N)^{(\alpha-1)/2\alpha}(\log p)^{1/2} = o(1),
\end{aligned}
$$

and as (e) and (f) both share the same upper bound, $\log p = o(N^{(\alpha-1)/(3\alpha-1)})$ is a sufficient condition for all the components in (B.3.34) and (B.3.35) to be $o_p(1)$, given any $\alpha > 1$ for Assumption 7 (a). Therefore we achieve (B.3.26).

Step 3. (CLT) From Assumption 4 on the error terms, (B.3.26) implies

$$
\begin{aligned}
\sqrt{N}\left(\widetilde{\theta} - \theta_0\right) &= \left(E[V^2]^{-1} + o_p(1)\right)\left(\frac{1}{\sqrt{N}}\mathscr{V}'\mathscr{U} + o_p(1)\right) \\
&\xrightarrow{d} E[V^2]^{-1} N(0, E[V^2U^2])
\end{aligned}
$$

following Lindeberg–Lévy CLT, which concludes the proof. □

B.3.3 Proof of Theorem 4

Proof. In proof of Theorem 1 we have shown that the convergence rates from Theorem 3.1. of Ing hold under our assumptions. Hence it is enough to show that Assumption 4.2. in CCDDHNR holds. Note again that Assumption 8 (b)–(e) corresponds to Assumption 4.2. (a)–(d) in Chernozhukov et al. (2018).

We shall thus verify condition (e) in Chernozhukov et al. (2018) is implied by the convergence rates (B.1.3) and (B.1.4). Recall that by (B.1.3) and an argument similar to (B.1.5)–(B.1.8),

$$
\left\|\widehat{\xi} - \xi_0\right\|^2 \lesssim \left(\frac{\log p}{N}\right)^{1-1/2\alpha} \tag{B.3.36}
$$

holds with probability at least $1 - \Delta_N$ for some $\Delta_N = o(1)$ and $\xi = \gamma, \zeta$, and β.

Now we will show that with $P-$probability no less than $1 - \Delta_N$, $\|\widehat{\eta} - \eta_0\|_{P,q} \leq C$, $\|\widehat{\eta} - \eta_0\|_{P,2} \leq \delta_N$, and $\|\widehat{\beta} - \beta_0\|_{P,2} \times (\|\widehat{\gamma} - \gamma_0\|_{P,2} + \|\widehat{\zeta} - \zeta_0\|_{P,2})$ holds that for Δ_N, δ_N such that both are sequences of strictly positive constants converging to zero, where $\eta = (\beta, \gamma, \zeta)$. From (B.3.36), we have

$$
\begin{aligned}
\|\widehat{\eta} - \eta_0\|_{P,q} &= \|\widehat{\beta} - \beta_0\|_{P,q} \vee \|\widehat{\gamma} - \gamma_0\|_{P,q} \vee \|\widehat{\zeta} - \zeta_0\|_{P,q} = \left(E\left[\left\|X'(\widehat{\beta} - \beta_0)\right\|^q \mid A_N\right]\right)^{1/q} \\
&\leq \left(E[\left((\widehat{\beta} - \beta_0)'E[XX'](\widehat{\beta} - \beta_0)\right)^{q/2} \mid A_N]\right)^{1/q} = \left(\left(\lambda_{\max}(E[XX'])\|\widehat{\beta} - \beta_0\|^2\right)^{q/2}\right)^{1/q}
\end{aligned}
$$

$$= \left(\lambda_{\max}(E[XX'])\|\widehat{\beta} - \beta_0\|^2\right)^{1/2},$$

where it is bounded by Assumption 9(a), (B.1.3), and (B.1.4).

Thus, $\|\widehat{\eta} - \eta_0\|_{P,q} \leq C$ holds with probability at least $1 - \Delta_N$ for some $\Delta_N = o(1)$.

Since we use the identical procedure, namely the OGA and HDAIC, in estimating β_0, γ_0, and η_0, the convergence rates apply to all the nuisance parameters. Like we did in the proof of Theorem 1, We consider two cases where all the parameters follow the polynomial decay case or the exponential decay case.

Case 1. For the polynomial decay case, let $\delta_N = (\log p)^{1-1/2\alpha} N^{1/2\alpha-1/2}$. Then $\delta_N = o(1)$ since α is assumed to be strictly larger than 1 in Assumption 10 (a). We have

$$\|\widehat{\eta} - \eta_0\|_{P,2} = \|\widehat{\beta} - \beta_0\|_{P,q} \vee \|\widehat{\gamma} - \gamma_0\|_{P,q} \vee \|\widehat{\zeta} - \zeta_0\|_{P,q} \lesssim \left(\frac{\log p}{N}\right)^{1/2-1/4\alpha} \leq \delta_N,$$

where the last inequality holds since $(\log p/N)^{1/4\alpha} \leq \log p$ holds with $\alpha > 1$. Also,

$$\|\widehat{\beta} - \beta_0\|_{P,2} \times \left(\|\widehat{\zeta} - \zeta_0\|_{P,2} + \|\widehat{\gamma} - \gamma_0\|_{P,2}\right) \lesssim \left(\frac{\log p}{N}\right)^{1-1/2\alpha} = \delta_N N^{-1/2}$$

holds.

Case 2. For the exponential decay case, let $\delta_N = N^{-1/2} \log p \log N$. By the assumption $\log p = o(N^{1/4})$, $\delta_N = o(1)$. We have

$$\|\widehat{\eta} - \eta_0\|_{P,2} \lesssim \sqrt{\frac{\log p \log N}{N}} < \delta_N$$

and

$$\|\widehat{\beta} - \beta_0\|_{P,2} \times \left(\|\widehat{\zeta} - \zeta_0\|_{P,2} + \|\widehat{\gamma} - \gamma_0\|_{P,2}\right) \lesssim \frac{\log p \log N}{N} = \delta_N N^{-1/2}.$$

A similar argument applies to the cross cases.

We have shown that Assumption 4.2 in Chernozhukov et al. (2018) holds for both sparsity assumptions. Therefore, applying Theorem 4.2. in Chernozhukov et al. (2018), we get the desired results.

□

B.4 Finite Sample Adjustment

The DML estimator is random even conditionally on data, because of the random splitting of the sample for cross fitting. To mitigate the effects of this randomness of the DML, Chernozhukov et al. (2018, Sec. 3.4) propose procedures of finite-sample adjustments. In this section, we present one of these procedures for completeness.

Suppose that we repeat the DML estimation S times to obtain $\{\check{\theta}^s\}_{s=1}^S$. A robust estimator that incorporates the impact of sample splitting is defined by

$$\check{\theta}^{Med} = \text{Median}\left\{\check{\theta}^s\right\}_{s=1}^S.$$

Chernozhukov et al. (cf. 2018, Definition 3.3). Its associated variance estimator is given by

$$\widehat{\Omega}^{Med} = \text{Median}\left\{\widehat{\Omega}^s + (\check{\theta}^s - \check{\theta}^{Med})(\check{\theta}^s - \check{\theta}^{Med})'\right\}_{s=1}^S.$$

Chernozhukov et al. (cf. 2018, Equation 3.14). We use these estimators with $S = 20$ in reporting the estimation results in Section 2.5.

B.5 Additional Simulations

B.5.1 Alternative Values of the Tuning Parameters

In Section 2.3 in the main text, we present simulation results using the choice $C^* = 2$ of the tuning parameter following Ing (2020) – see Section 2.2.3 for the implementation details. In the current appendix section, we show additional simulation results that we obtain by varying the value of C^* from 1 to 3 and report the sensitivity of the results to this variation.

Table B.1 shows the results under $C^* = 1.0, 1.2, \ldots, 2.8, 3.0$ along with the baseline value of $C^* = 2.0$. We focus on one DGP, namely the case of $\beta_{0,j} = \gamma_{0,j} = j^{-1.5}$. Observe that the values of C^* from 1.8 to 2.2 yield the best results both in terms of the RMSE and the 95% coverage. The RMSE is stable widely across $C^* \in \{1.6, 1.8, \cdots, 2.8, 3.0\}$.

B.5.2 Estimation and Inference without Cross Fitting

In Section 2.3 in the main text, we present simulation results for Algorithm 2 which uses sample splitting for cross fitting. In the current appendix section, we present simulation results for Algorithm 3 based on the full sample without cross fitting. We use the the same DGPs as in Section 2.3 in the main text.

The results presented in Table B.2 for Algorithm 3 are almost the same as those presented in Table 2.1 in the main text. In other words, Algorithm 3 behaves similarly to Algorithm 2 under DGPs, while the latter slightly outperforms especially under less sparse designs.

$\beta_{0,j}, \gamma_{0,j}$	C^*	N	p	Bias	SD	RMSE	95%
$j^{-1.5}$	1	500	500	-0.054	0.046	0.073	0.769
		1000	500	-0.024	0.032	0.043	0.847
	1.2	500	500	-0.033	0.045	0.058	0.874
		1000	500	-0.015	0.032	0.037	0.901
	1.4	500	500	-0.019	0.045	0.051	0.913
		1000	500	-0.008	0.032	0.034	0.935
	1.6	500	500	-0.009	0.045	0.047	0.923
		1000	500	-0.003	0.032	0.033	0.943
	1.8	500	500	-0.003	0.045	0.047	0.937
		1000	500	0.000	0.031	0.033	0.937
	2	500	500	0.001	0.045	0.046	0.936
	(Baseline)	1000	500	0.002	0.031	0.033	0.938
	2.2	500	500	0.004	0.044	0.046	0.930
		1000	500	0.003	0.031	0.032	0.941
	2.4	500	500	0.006	0.044	0.047	0.927
		1000	500	0.005	0.031	0.033	0.939
	2.6	500	500	0.009	0.044	0.047	0.926
		1000	500	0.006	0.031	0.033	0.932
	2.8	500	500	0.010	0.044	0.048	0.912
		1000	500	0.007	0.031	0.033	0.943
	3	500	500	0.012	0.044	0.048	0.912
		1000	500	0.008	0.031	0.033	0.937

Table B.1: Monte Carlo simulation results under various values of the tuning parameter C^*. Displayed are Monte Carlo simulation statistics including the bias, standard deviation (SD), root mean square error (RMSE), and 95% coverage frequency.

$\beta_{0,j}, \gamma_{0,j}$	N	p	Method of Preliminary Estimation	Bias	SD	RMSE	95%
Sparse	500	500	OGA+HDAIC	-0.002	0.045	0.045	0.943
	1000	500		0.000	0.032	0.032	0.943
e^{-j}	500	500	OGA+HDAIC	0.001	0.045	0.045	0.937
	1000	500		0.000	0.032	0.032	0.946
j^{-2}	500	500	OGA+HDAIC	-0.001	0.045	0.046	0.939
	1000	500		0.001	0.032	0.032	0.947
$j^{-1.75}$	500	500	OGA+HDAIC	0.000	0.045	0.046	0.942
	1000	500		0.001	0.032	0.032	0.946
$j^{-1.5}$	500	500	OGA+HDAIC	0.002	0.045	0.046	0.935
	1000	500		0.003	0.032	0.033	0.936
$j^{-1.25}$	500	500	OGA+HDAIC	0.007	0.044	0.048	0.924
	1000	500		0.005	0.031	0.034	0.923
j^{-1}	500	500	OGA+HDAIC	0.022	0.044	0.056	0.872
	1000	500		0.015	0.031	0.037	0.898

Table B.2: Monte Carlo simulation results without cross fitting. Displayed are Monte Carlo simulation statistics including the bias, standard deviation (SD), root mean square error (RMSE), and 95% coverage frequency.

B.5.3 High-Dimensional IV Regression

We consider the simple setting where the data are generated by the system

$$\begin{aligned} Y &= \theta_0(D - X'\zeta_0) + X'\gamma_0 + U, \\ D &= \mu Z + X'\zeta_0 + E, \\ Z &\sim N(0,1), \end{aligned}$$

where

$$(U,E) \sim N\left(0, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right).$$

For our method, we use Algorithm 4 for the estimation. For DML methods, we used the R package DoubleML's partially linear IV setting. The results are shown in Table B.3.

B.6 Additional Empirical Results

In the current appendix section, we provide additional empirical estimates following Section 2.5 in the main text. We provide the following two types of alternative estimates. First, we vary the value of the tuning parameter C^*. Second, we use Algorithm 3 without cross fitting, instead of Algorithm 2. Table B.4 gives the baseline results using our method

$\gamma_{0,j}, \zeta_{0,j}$	N	p	Method of Preliminary Estimation	Bias	SD	RMSE	95%
Sparse	500	500	OGA+HDAIC	-0.001	0.046	0.047	0.949
	1000	500		0.002	0.032	0.033	0.951
e^{-j}	500	500	OGA+HDAIC	-0.001	0.046	0.046	0.951
	1000	500		0.002	0.032	0.032	0.947
j^{-2}	500	500	OGA+HDAIC	-0.001	0.046	0.047	0.943
	1000	500		0.002	0.032	0.033	0.951
$j^{-1.75}$	500	500	OGA+HDAIC	-0.001	0.047	0.047	0.942
	1000	500		0.002	0.032	0.033	0.948
$j^{-1.5}$	500	500	OGA+HDAIC	-0.001	0.047	0.047	0.942
	1000	500		0.001	0.033	0.033	0.952
$j^{-1.25}$	500	500	OGA+HDAIC	-0.002	0.048	0.049	0.943
	1000	500		0.002	0.033	0.033	0.934
j^{-1}	500	500	OGA+HDAIC	-0.001	0.051	0.051	0.951
	1000	500		0.002	0.034	0.035	0.946

Table B.3: Monte Carlo simulation results for IV model. Displayed are Monte Carlo simulation statistics including the bias, standard deviation (SD), root mean square error (RMSE), and 95% coverage frequency.

		Unskilled Labor	SkilledLabor
(I)	Double Machine Learning with OGA+HDAIC (Baseline)	0.168 (0.011)	0.060 (0.010)
(II)	Double Machine Learning with OGA+HDAIC, $C^* = 1.8$	0.169 (0.011)	0.060 (0.009)
(III)	Double Machine Learning with OGA+HDAIC, $C^* = 2.2$	0.163 (0.011)	0.062 (0.010)
(IV)	Double Machine Learning with OGA+HDAIC without Cross Fitting	0.175 (0.010)	0.067 (0.008)

Table B.4: Estimates of labor elasticities in the 3-digit level industry of food products (311) in Chile based on four alternative methods.

in row (I). Rows (II) and (III) show the estimates under the choices $C^* = 1.8$ and 2.2, respectively. Row (IV) shows the estimates using the whole sample without cross fitting. Observe that the results are robust, and our empirical findings qualitatively remain the same as those discussed in the main text.

Appendix C

Appendix to Chapter 3

C.1 Proofs of the Main Results

This section collects mathematical proofs of the main results. Appendix C.1.1 presents a proof of Theorem 3.4.1. Appendix C.1.2 presents a proof of Proposition 3.4.1.

C.1.1 Proof of Theorem 3.4.1

Appendix C.1.1 presents a proof of

Proof. From (3.3.1), define

$$\underline{\Delta}(v) := F(v) - \underline{F}(v) \geq 0 \qquad \text{and} \qquad \overline{\Delta}(v) := \overline{F}(v) - F(v) \geq 0. \tag{C.1.1}$$

We are going to use Assumption 5 (i) to establish $\|\underline{F}_n - \underline{F}\|_\infty = O_p(n^{-1/\overline{q}})$ where $\overline{q} = \max_j q_j$. Recall from the definition of $\underline{F}_n$ that it consists of a summation of $F_{n,j}$ and a $\max\{.,0\}$ operator. Consider the summation part:

$$\begin{aligned}
&\sup_{x_1,\ldots,x_{p-1},v} \left| \sum_{j=1}^{p-1} F_{n,j}(x_j) + F_{n,p}\left(v - \sum_{j=1}^{p-1} x_j\right) - \sum_{j=1}^{p-1} F_j(x_j) - F_p\left(v - \sum_{j=1}^{p-1} x_j\right) \right| \\
&\leq \sup_{x_1} |F_{n,1}(x_1) - F_1(x_1)| + \cdots + \sup_{x_1,\ldots,x_{p-1},v} \left| F_{n,p}\left(v - \sum_{j=1}^{p-1} x_j\right) - F_p\left(v - \sum_{j=1}^{p-1} x_j\right) \right| \\
&= \sup_t |F_{n,1}(t) - F_1(t)| + \cdots + \sup_t \left|F_{n,p}(t) - F_p(t)\right| \\
&\lesssim_P n^{-1/q_1} + \cdots + n^{-1/q_p} \\
&\lesssim n^{-1/\overline{q}},
\end{aligned} \tag{C.1.2}$$

where the first inequality follows from the triangle inequality, the next equality reflects a simple change of notations, the next inequality follows from Assumption 5 (i), and the last inequality is due to the definition of $\overline{q} = \max_j q_j$.

For simplicity, define

$$\begin{aligned}
S_n(v) &= \sum_{j=1}^{p-1} F_{n,j}(x_j) + F_{n,p}\left(v - \sum_{j=1}^{p-1} x_j\right) - (p-1) \qquad \text{and} \\
S(v) &= \sum_{j=1}^{p-1} F_j(x_j) + F_p\left(v - \sum_{j=1}^{p-1} x_j\right) - (p-1),
\end{aligned}$$

so that $\underline{F}_n(v) = \max\{S_n(v),0\}$ and $\underline{F}(v) = \max\{S(v),0\}$. We can write

$$\begin{aligned}\sup_v |\underline{F}_n(v) - \underline{F}(v)| =& \sup_v |\max\{S_n(v),0\} - \max\{S(v),0\}| \\ \leq& \sup_v |S_n(v) - S(v)| \\ =& O_p(n^{-1/\overline{q}}),\end{aligned}$$

where the first inequality comes from the following reasoning and the last bound is from (C.1.2).

$$\begin{aligned}|\max\{S_n(v),0\} - \max\{S(v),0\}| =& \begin{cases} |S_n(v) - S(v)|, & \text{if } S_n(v) \geq \text{ and } S(v) \geq 0 \\ S_n(v), & \text{if } S_n(v) \geq 0 \text{ and } S(v) < 0 \\ S(v), & \text{if } S_n(v) < 0 \text{ and } S(v) \geq 0 \\ 0, & \text{otherwise,} \end{cases} \\ \leq& |S_n(v) - S(v)|.\end{aligned}$$

Therefore, we have

$$\|\underline{F}_n - \underline{F}\|_\infty = O_p(n^{-1/\overline{q}}). \tag{C.1.3}$$

Next, we are going to show that $\inf_v \left\{F(v) - \underline{F}_n^\delta(v)\right\} \geq 0$ w.p.a. 1. Decompose

$$\begin{aligned}F(v) - \underline{F}_n^\delta(v) =& F(v) - \underline{F}(v) + \underline{\delta}_n(v) + \underline{F}(v) - \underline{F}_n(v) \\ =& \underline{\Delta}(v) + \underline{\delta}_n(v) + \underline{F}(v) - \underline{F}_n(v) \\ \geq& \underline{\Delta}(v) + \underline{\delta}_n(v) - |\underline{F}_n(v) - \underline{F}(v)| \\ \geq& \underline{\Delta}(v) + \underline{\delta}_n(v) - \|\underline{F}_n - \underline{F}\|_\infty.\end{aligned}$$

Recall $\underline{\Delta}(\cdot) \geq 0$ from (C.1.1). Thus, by Assumption 5 (ii)

$$\inf_v \{F(v) - \underline{F}_n^\delta(v)\} \geq \inf_v \underline{\delta}'_n(v) \cdot n^{-1/q} - \|\underline{F}_n - \underline{F}\|_\infty.$$

Since $\inf_v \underline{\delta}_n'(v) > 0$ by Assumption 5 (ii), $\|\underline{F}_n - \underline{F}\|_\infty = O_p(n^{-1/\overline{q}})$ from (C.1.3), and $q > \overline{q}$, it follows that

$$\inf_v \{F(v) - \underline{F}_n^{\delta}(v)\} \geq 0$$

holds with probability approaching 1.

By similar lines of arguments,

$$\inf_v \{\overline{F}_n^{\delta}(v) - F(v)\} \geq 0$$

holds with probability approaching 1. □

C.1.2 Proof of Proposition 3.4.1

Proof. First, the mean value expansion yields

$$\begin{aligned}
\left|F_{n,j}(t) - F_j(t)\right| &= \left| \left.\frac{\partial \Phi(t/u)}{\partial u}\right|_{u=u(\ell_j s_j, \widehat{\ell}_j \widehat{s}_j)} \left(\widehat{\ell}_j \widehat{s}_j - \ell_j s_j\right) \right| \\
&\leq \frac{1}{u(\ell_j s_j, \widehat{\ell}_j \widehat{s}_j)} \cdot \sup_u |u\phi(u)| \cdot \left|\widehat{\ell}_j \widehat{s}_j - \ell_j s_j\right| \\
&\leq \frac{1}{u(\ell_j s_j, \widehat{\ell}_j \widehat{s}_j)} \cdot \left|\widehat{\ell}_j \widehat{s}_j - \ell_j s_j\right|,
\end{aligned}$$

where $\phi(.)$ denotes the standard normal PDF and $u(\ell_j s_j, \widehat{\ell}_j \widehat{s}_j)$ is between $\ell_j s_j$ and $\widehat{\ell}_j \widehat{s}_j$.

Therefore, we obtain

$$\begin{aligned}
&Pr\left(\sup_t \left|F_{n,j}(t) - F_j(t)\right| \geq \varepsilon \text{ and } \widehat{\ell}_j \widehat{s}_j \geq \ell_j s_j / 2\right) \\
\leq &Pr\left(\frac{1}{u(\ell_j s_j, \widehat{\ell}_j \widehat{s}_j)} \cdot \left|\widehat{\ell}_j \widehat{s}_j - \ell_j s_j\right| \geq \varepsilon \text{ and } \widehat{\ell}_j \widehat{s}_j \geq \ell_j s_j / 2\right) \\
\leq &Pr\left(\left|\widehat{\ell}_j \widehat{s}_j - \ell_j s_j\right| \geq \frac{\varepsilon \ell_j s_j}{2}\right) \\
\lesssim &n^{-1/2}
\end{aligned}$$

where the second inequality follows from the fact that the mean value satisfies $u(\ell_j s_j, \widehat{\ell}_j \widehat{s}_j) \geq \ell_j s_j / 2$ provided $\widehat{\ell}_j \widehat{s}_j \geq \ell_j s_j / 2$, and the last inequality is due to Assumption 6. Also,

$$Pr\left(\sup_t \left|F_{n,j}(t) - F_j(t)\right| \geq \varepsilon \text{ and } \widehat{\ell}_j \widehat{s}_j < \ell_j s_j / 2\right) \leq Pr\left(\widehat{\ell}_j \widehat{s}_j < \ell_j s_j / 2\right)$$

$$\leq Pr\left(\left|\widehat{\ell}_j\widehat{s}_j - \ell_j s_j\right| > \ell_j s_j/2\right)$$
$$\lesssim n^{-1/2}$$

holds by Assumption 6. Using the above results, we obtain

$$\begin{aligned} Pr\left(\sup_t \left|F_{n,j}(t) - F_j(t)\right| \geq \varepsilon\right) &= Pr\left(\sup_t \left|F_{n,j}(t) - F_j(t)\right| \geq \varepsilon \text{ and } \widehat{\ell}_j\widehat{s}_j \geq \ell_j s_j/2\right) \\ &\quad + Pr\left(\sup_t \left|F_{n,j}(t) - F_j(t)\right| \geq \varepsilon \text{ and } \widehat{\ell}_j\widehat{s}_j < \ell_j s_j/2\right) \\ &\lesssim n^{-1/2}. \end{aligned}$$

Therefore, Assumption 5 (i) is satisfied with $q_j = 2$ for all $j = 1, \ldots, p$. □