Nonparametric estimation of triangular simultaneous equations models under weak identification

Sukjin Han
Department of Economics, University of Texas at Austin

This paper analyzes the problem of weak instruments on identification, estimation, and inference in a simple nonparametric model of a triangular system. The paper derives a necessary and sufficient rank condition for identification, based on which weak identification is established. Then *nonparametric weak instruments* are defined as a sequence of reduced-form functions where the associated rank shrinks to zero. The problem of weak instruments is characterized as *concurvity*, which motivates the introduction of a regularization scheme. The paper proposes a penalized series estimation method to alleviate the effects of weak instruments and shows that it achieves desirable asymptotic properties. A data-driven procedure is proposed for the choice of the penalization parameter. The findings of this paper provide useful implications for empirical work. To illustrate them, Monte Carlo results are presented and an empirical example is given in which the effect of class size on test scores is estimated nonparametrically.

Keywords. Triangular models, nonparametric identification, weak identification, weak instruments, series estimation, regularization, concurvity.

JEL classification. C13, C14, C36.

1. Introduction

Instrumental variables (IVs) are widely used in empirical research to identify and estimate models with endogenous explanatory variables. In linear simultaneous equations models, it is well known that standard asymptotic approximations break down when instruments are weak in the sense that (partial) correlation between the instruments and endogenous variables is weak. The consequences of and solutions for weak instruments in linear settings have been extensively studied in the literature over the past two decades; see, for example, Bound, Jaeger, and Baker (1995), Staiger and Stock (1997), Dufour (1997), Kleibergen (2002, 2005), Moreira (2003), Stock and Yogo (2005), and Andrews

Sukjin Han: sukjin.han@austin.utexas.edu

I am very grateful to my advisors, Donald Andrews and Edward Vytlacil, and committee members, Xiaohong Chen and Yuichi Kitamura, for their inspiration, guidance, and support. I am deeply indebted to Donald Andrews for his thoughtful advice throughout the project. I also appreciate the helpful suggestions by the two anonymous referees. This paper has benefited from discussions with Joseph Altonji, Ivan Canay, Tim Christensen, Philip Haile, Keisuke Hirano, Han Hong, Joel Horowitz, Seokbae Simon Lee, Oliver Linton, Whitney Newey, Byoung Park, Peter Phillips, and Alex Torgovitsky. I gratefully acknowledge financial support from a Carl Arvid Anderson Prize from the Cowles Foundation. I also thank the seminar participants at Yale, UT Austin, Chicago Booth, Notre Dame, SUNY Albany, Duke, Sogang, SKKU, and Yonsei, as well as the participants at NASM and Cowles Summer Conference.

Available at http://qeconomics.org. https://doi.org/10.3982/QE975

and Stock (2007), among others. Weak instruments in nonlinear parametric models have recently been studied in the literature in the context of weak identification by, for example, Stock and Wright (2000), Kleibergen (2005), Andrews and Cheng (2012), Andrews and Mikusheva (2016a,b), Andrews and Guggenberger (forthcoming), and Han and McCloskey (forthcoming).

One might expect that nonparametric models with endogenous explanatory variables will generally require strong identification power as there is an infinite number of unknown parameters to identify, and hence, strong instruments may be crucial for a reasonable performance of estimation.[1] Despite the problem's importance and the growing popularity of nonparametric models, weak instruments in nonparametric settings have not received much attention.[2] Furthermore, surprisingly little attention has been paid to the consequences of weak instruments in empirical research using nonparametric models; see below for references. Part of the neglect is due to the existing complications embedded in nonparametric models.

In a simple nonparametric framework, this paper analyzes the problem of weak instruments on identification, estimation, and inference, and proposes an estimation strategy to mitigate the effect. Identification results are obtained so that the concept of weak identification can subsequently be introduced via localization. The problem of weak instruments is characterized as concurvity. An estimation method is proposed through regularization and the resulting estimators are shown to have desirable asymptotic properties even when instruments are possibly weak.

As a nonparametric framework, we consider a triangular simultaneous equations model. The specification of weak instruments is intuitive in the triangular model because it has an explicit reduced-form relationship. Additionally, clear interpretation of the effect of weak instruments can be made through a specific structure produced by the control function approach. To make our analysis succinct, we specify additive errors in the model. This particular model is considered in Newey, Powell, and Vella (1999) (NPV) and Pinkse (2000) in a situation without weak instruments. Although relatively recent developments in nonparametric triangular models contribute to models with nonseparable errors (e.g., Imbens and Newey (2009)), such flexibility complicates the exposition of the main results of this paper.[3] Also, having a form analogous to its popular parametric counterpart, the model with additive errors is broadly used in applied research such as Blundell and Duncan (1998), Yatchew and No (2001), Lyssiotou, Pashardes, and Stengos (2004), Dustmann and Meghir (2005), Skinner, Fisher, and Wennberg (2005),

[1] This conjecture is shown to be true in the setting considered in this paper; see Theorem 5.1 and Corollary 5.2.

[2] Chesher (2003, 2007) mentioned the issue of weak instruments in applying his key identification condition in the empirical example of Angrist and Keueger (1991). Blundell, Chen, and Kristensen (2007) determined whether weak instruments are present in the Engel curve dataset of their empirical section. They do this by applying the Stock and Yogo (2005) test developed in linear models to their reduced form, which is linearized by sieve approximation. Darolles, Fan, Florens, and Renault (2011) briefly discussed weak instruments that are indirectly characterized within their source condition.

[3] For instance, the control function employed in Imbens and Newey (2009) requires large variation in instruments, and hence discussing weak instruments (i.e., weak association between endogenous variables and instruments or little variation in instruments) in such a context requires more care.

Blundell, Browning, and Crawford (2008), Del Bono and Weber (2008), Frazer (2008), Mazzocco (2012), Coe, von Gaudecker, Lindeboom, and Maurer (2012), Breza (2013), Henderson, Papageorgiou, and Parmeter (2013), Chay and Munshi (2015), and Koster, Ommeren, and Rietveld (2014).

One of the contributions of this paper is that it derives novel identification results in nonparametric triangular models that complement the existing results in the literature. With a mild support condition, we show that a particular rank condition is necessary and sufficient for the identification of the structural relationship. This rank condition is substantially weaker than what is established in NPV. Deriving such a rank condition is key to establishing the notion of weak identification. Since the condition is minimal, a "slight violation" of it has a binding effect on identification, hence resulting in weak identification.

To characterize weak identification, we consider a drifting sequence of reduced-form functions that converges to a nonidentification region, namely, a space of reduced-form functions that violate the rank condition for identification. A particular rate is designated relative to the sample size, which effectively measures the strength of the instruments, so that it appears in asymptotic results for the estimator of the structural function. The concept of *nonparametric weak instruments* generalizes the concept of weak instruments in linear models such as in Staiger and Stock (1997).

In the nonparametric control function framework, the problem of weak instruments becomes a nonparametric analogue of a multicollinearity problem known as *concurvity* (Hastie and Tibshirani (1986)). Once the endogeneity is controlled by a control function, the model can be rewritten as an additive nonparametric regression, where the endogenous variables and reduced-form errors comprise two regressors, and weak instruments result in the variation of the former regressor being mainly driven by the variation of the latter. Therefore, the regularization methods used in the literature to solve inverse problems can be introduced to our problem. Among the regularization methods, only penalization (i.e., Tikhonov-type regularization) alleviates the effect of weak instruments, while truncation does not.

This paper proposes a *penalized series estimator* for the structural function and establishes its asymptotic properties. We use L_2-type penalization to control the penalty bias. Our results on the rate of convergence of the estimator suggest that, without penalization, weak instruments characterized as concurvity slow down the overall convergence rate, exacerbating bias, and variance "symmetrically." We show that a faster convergence rate can be achieved with penalization than without, while the penalty bias can be dominated by the standard approximation bias. We propose a data-driven procedure of choosing the penalization parameter, and derive the adaptive convergence rate. We also derive consistency and asymptotic normality with mildly weak instruments.

The problem of concurvity in additive nonparametric models is also recognized in the literature where different estimation methods are proposed to address the problem, for example, the backfitting methods (Linton (1997), Nielsen and Sperlich (2005)) and the integration method (Jiang, Fan, and Fan (2010)); also see Sperlich, Linton, and Härdle (1999). In particular, as closely related work to the asymptotic results of this paper, Jiang, Fan, and Fan (2010) established pointwise asymptotic normality for local linear and integral estimators in an additive nonparametric model with highly correlated covariates. In the present paper, where an additive model results from a triangular model accompanied with the control function approach, the problem of concurvity is addressed in a more direct manner via penalization. Although the main conclusions of this paper do not depend on the choice of the nonparametric estimation method, using series estimation is justified in our triangular model setup, where the joint density of the endogenous variable and control variable is singular near the boundary of the support (Imbens and Newey (2009)).

Another possible nonparametric framework in which to examine the problem of weak instruments is a nonparametric IV (NPIV) model (Newey and Powell (2003), Hall and Horowitz (2005), and Blundell, Chen, and Kristensen (2007), among others). Unlike in a triangular model, the absence of an explicit reduced-form relationship forces weak instruments in this setting to be characterized as a part of the ill-posed inverse problem. Therefore, in this model, the performance of the estimator can be severely deteriorated as the problem is "doubly ill-posed."[4] Further, it may also be hard to separate the effects of the two in asymptotic theory. As a related recent work, Freyberger (2017) provided a framework by which the size of the identification can be learned even though the completeness condition is not testable in a NPIV model (Canay, Santos, and Shaikh (2013)). Instead of using a drifting sequence of distributions, he indirectly defines weak instruments as a failure of a restricted version of the completeness condition. While he applies his framework to test weak instruments, our focus is on estimation and inference of the function of interest in a different nonparametric model with a more explicit definition of weak instruments.

The findings of this paper provide useful implications for empirical work. First, when estimating a nonparametric structural function, the results of IV estimation and subsequent inference can be misleading even when the instruments are strong in terms of conventional criteria for linear models.[5] Second, the symmetric effect of weak instruments on bias and variance implies that the bias–variance trade-off is the same across different strengths of instruments, and hence, weak instruments cannot be alleviated by exploiting the trade-off. Third, penalization on the other hand can alleviate weak instruments by significantly reducing variance and sometimes bias as well. Fourth, there is a trade-off between the smoothness of the structural function (or the dimensionality of its argument) and the requirement of strong instruments. Fifth, if a triangular model along with its assumptions is considered to be reasonable, it provides an estimator that is more precise than that with a NPIV model, which is an attractive feature especially in the presence of weak instruments. Sixth, although a linear first-stage reduced form is commonly used in applied research (e.g., in NPV, Blundell and Duncan (1998), Blundell, Duncan, and Pendakur (1998), Dustmann and Meghir (2005), Coe et al. (2012), and Henderson, Papageorgiou, and Parmeter (2013)), the strength of instruments can be improved by

[4] In Section 8, we illustrate this point in an empirical application by comparing estimates calculated from the triangular and NPIV models.

[5] For instance, in Coe et al. (2012), the first-stage F-statistic value that is reported is (sometimes barely) in favor of strong instruments, but the judgement is based on the criterion for linear models. The majority of empirical works referenced above do not report first-stage results.

having a nonparametric reduced form so that the nonlinear relationship between the endogenous variable and instruments can be fully exploited. The last point is related to the identification results of this paper. In Section 8, we apply the findings of this paper to an empirical example, where we nonparametrically estimate the effect of class size on students' test scores.

The rest of the paper is organized as follows. Section 2 introduces the model and obtains new identification results. Section 3 discusses weak identification and Section 4 relates the weak instrument problem to the concurvity problem and defines our penalized series estimator. Section 5 establishes the rate of convergence and consistency of the penalized series estimator. It also provides the adaptive rate with the data-driven choice of the penalization parameter. Section 6 establishes the asymptotic normality of some functionals of the estimator. Section 7 presents the Monte Carlo simulation results. Section 8 discusses the empirical application. Finally, Section 9 concludes.

2. Identification

We consider a nonparametric triangular simultaneous equations model

$$y = g_0(x, z_1) + \varepsilon, \quad x = \Pi_0(z) + v, \tag{2.1a}$$

$$E[\varepsilon|v, z] = E[\varepsilon|v] \quad \text{a.s.}, \qquad E[v|z] = 0 \quad \text{a.s.}, \tag{2.1b}$$

where $g_0(\cdot,\cdot)$ is an unknown structural function of interest, $\Pi_0(\cdot)$ is an unknown reduced-form function, x is a d_x-vector of endogenous variables, $z = (z_1, z_2)$ is a $(d_{z_1} + d_{z_2})$-vector of exogenous variables, and z_2 is a vector of excluded instruments. The stochastic assumptions (2.1b) are more general than the assumption of full independence between (ε, v) and z and $E[v] = 0$. Following the control function approach,

$$E[y|x, z] = g_0(x, z_1) + E\big[\varepsilon|\Pi_0(z) + v, z\big] = g_0(x, z_1) + E[\varepsilon|v] = g_0(x, z_1) + \lambda_0(v), \tag{2.2}$$

where $\lambda_0(v) = E[\varepsilon|v]$ and the second equality is from the first part of (2.1b). In effect, we capture endogeneity ($E[\varepsilon|x, z] \neq 0$) by an unknown function $\lambda_0(v)$, which serves as a control function. Once v is controlled for, the only variation of x comes from the exogenous variation of z. Based on equation (2.2), we establish identification, weak identification, and estimation results.

First, we obtain identification results that complement the results of NPV. Given (2.2), the identification of $g_0(x, z_1)$ is achieved if one can separately vary (x, z_1) and v in $g(x, z_1) + \lambda(v)$. Since $x = \Pi_0(z) + v$, a suitable condition on $\Pi_0(\cdot)$ will guarantee this via the separate variation of z and v. In light of this intuition, NPV proposes the following identification condition, and show that $g_0(x, z_1)$ is identified up to an additive constant:

$$\Pr\left[\operatorname{rank}\left(\frac{\partial \Pi_0(z)}{\partial z_2'}\right) = d_x\right] = 1. \tag{2.3}$$

Note that this condition is only a sufficient condition, which suggests that the model can possibly be identified with a relaxed rank condition. This observation motivates our identification analysis. The identification analysis of this section is also important for our later purpose of defining the notion of *weak identification*. Henceforth, in order to keep our presentation succinct, we drop z_1 from model (2.1) and let $z = z_2$. With z_1 included, all the results of this paper follow conditional on z_1.

Assumption ID1. *The functions $g_0(\cdot)$, $\lambda_0(\cdot)$, and $\Pi_0(\cdot)$ are continuously differentiable in their arguments.*

This condition is also assumed in NPV. Let $\mathcal{X} \subset \mathbb{R}^{d_x}$ and $\mathcal{Z} \subset \mathbb{R}^{d_z}$ be the marginal supports of x and z, respectively. Also, let $\mathcal{X}_z$ be the conditional support of x given $z \in \mathcal{Z}$. We partition $\mathcal{Z}$ into two regions where the rank condition is satisfied, that is, where z is relevant, and otherwise.

Definition 2.1 (Relevant Set). Let $\mathcal{Z}^r$ be the subset of $\mathcal{Z}$ defined by

$$\mathcal{Z}^r = \mathcal{Z}^r\big(\Pi_0(\cdot)\big) = \left\{z \in \mathcal{Z} : \operatorname{rank}\left(\frac{\partial \Pi_0(z)}{\partial z'}\right) = d_x\right\}.$$

Let $\mathcal{Z}^0 = \mathcal{Z} \backslash \mathcal{Z}^r$ be the complement of the relevant set. Let $\mathcal{X}^r$ be the subset of $\mathcal{X}$ defined by $\mathcal{X}^r = \{x \in \mathcal{X}_z : z \in \mathcal{Z}^r\}$. Given the definitions, we introduce an additional support condition.

Assumption ID2. *The supports $\mathcal{X}$ and $\mathcal{X}^r$ differ only on a set of probability zero, that is,* $\Pr[x \in \mathcal{X} \backslash \mathcal{X}^r] = 0$.

Intuitively, when z is in the relevant set, $x = \Pi_0(z) + v$ varies as z varies and, therefore, the support of x corresponding to the relevant set is large. Assumption ID2 assures that the corresponding support is large enough to almost surely cover the entire support of x. ID2 is not as strong as it may appear to be. Below, we show this by providing mild sufficient conditions for ID2.

If we identify $g_0(x)$ for any $x \in \mathcal{X}^r$, then we achieve identification of $g_0(x)$ by Assumption ID2. Now, in order to identify $g_0(x)$ for $x \in \mathcal{X}^r$, we need a rank condition, which will be minimal. The following is the identification result.

Theorem 2.2. *In model* (2.1), *suppose Assumptions* ID1 *and* ID2 *hold. Then $g_0(x)$ is identified on $\mathcal{X}$ up to an additive constant if and only if*

$$\Pr\big[z \in \mathcal{Z}^r\big] > 0. \tag{2.4}$$

The proof of this theorem and all subsequent proofs can be found in Appendix A and in Appendix B in the Online Supplemental Material (Han (2020)).

The rank condition (2.4) is necessary and sufficient. The condition is substantially weaker than (2.3), which is $\Pr[z \in \mathcal{Z}^r] = 1$ (with $z = z_2$). That is, Theorem 2.2 extends the result of NPV in the sense that when $\mathcal{Z}^r = \mathcal{Z}$, ID2 is trivially satisfied with $\mathcal{X} = \mathcal{X}^r$. Theorem 2.2 shows that it is enough for identification of $g_0(x)$ to have any fixed positive probability with which the rank condition is satisfied.[6] This condition can be seen as the local rank condition as in Chesher (2003). We achieve *global* identification with a *local* rank condition. Although this gain comes from having the additional support condition, the trade-off is appealing given the later purpose of building a weak identification notion. Even without Assumption ID2, maintaining the assumptions of Theorem 2.2, we still achieve identification of $g_0(x)$, but on the set $\{x \in \mathcal{X}^r\}$.

Lastly, in order to identify the level of $g_0(x)$, we need to introduce some normalization as in NPV. Either $E[\varepsilon]=0$ or $\lambda_0(\bar{v})=\bar{\lambda}$ suffices to pin down $g_0(x)$. With the latter normalization, it follows that $g_0(x)=E[y|x,v=\bar{v}]-\bar{\lambda}$, which we apply in estimation as it is convenient to implement.

The following is a set of sufficient conditions for Assumption ID2. Let $\mathcal{V}_z$ be the conditional support of v given $z \in \mathcal{Z}$.

Assumption ID2$'$. *Either* (a) *or* (b) *holds.* **(a)** (i) *x is univariate and x and v are continuously distributed,* (ii) *$\mathcal{Z}$ is a cartesian product of connected intervals, and* (iii) *$\mathcal{V}_z=\mathcal{V}_{\tilde{z}}$ for all $z,\tilde{z}\in\mathcal{Z}^0$*; **(b)** *$\mathcal{V}_z=\mathbb{R}^{d_x}$ for all $z\in\mathcal{Z}$.*

Lemma 2.1. *Under Assumption* ID1, *Assumption* ID2$'$ *implies Assumption* ID2.

In Assumption ID2$'$, the continuity of the r.v. is implied by the support condition imposed in NPV that the boundary of support of (z,v) has probability zero. Assumption ID2$'$(a)(i) assumes that the endogenous variable is univariate, which is most empirically relevant in nonparametric models.[7] Still, the exogenous covariate z_1 in $g(x,z_1)$, which is omitted in the discussion, can be a vector. ID2$'$(a)(ii) and (iii) are rather mild. ID2$'$(a)(ii) assumes that z has a connected support, that is, z varies smoothly.[8] The assumptions on the continuity of the r.v. and the connectedness of $\mathcal{Z}$ are also useful in deriving the asymptotic theory of the series estimator; see Assumption B below. ID2$'$(a)(iii) means that the conditional support of v given z is invariant when z is in $\mathcal{Z}^0$. This *support invariance* condition is the key to obtaining a rank condition that is considerably weaker than that of NPV. Our support invariance condition imposes no extra restriction on the support of z, and thus is different from the support invariance condition introduced in Imbens and Newey (2009), which typically requires large support. Also, the conditional support does not have to equal the marginal support of v here. ID2$'$(a)(iii), along with the control function assumptions (2.1b), is a weaker orthogonality condition for z than the full independence condition $z\perp v$. Note that $\mathcal{V}_z=\{x-\Pi_0(z):x\in\mathcal{X}_z\}$. Therefore, ID2$'$(a)(iii) equivalently means that $\mathcal{X}_z$ is invariant for those z satisfying $\partial\Pi_0(z)/\partial z=0$. Moreover, one can introduce a condition that is weaker than ID2$'$(a)(iii): $\mathcal{X}_z\subset\mathcal{X}^r$ for those z satisfying $\partial\Pi_0(z)/\partial z=0$.[9] These conditions in terms of $\mathcal{X}_z$ can be checked from

[6] A similar condition appears in the identification analysis of Hoderlein (2009), where endogenous semiparametric binary choice models are considered in the presence of heteroskedasticity.

[7] An additional condition is required with multivariate x, which is omitted in this paper.

[8] This is in contrast to, for example, Torgovitsky (2015) and D'Haultfœuille and Février (2015), which allow discrete instruments for identification in nonseparable models. The trade-off is that they require full independence for instruments, which is stronger than our mean independence of (2.1b).

[9] Therefore, heteroskedasticity of v may in general violate ID2, and thus ID2$'$(a)(iii), although some types of heteroskedasticity can still be allowed (e.g., heteroskedasticity only when z is relevant).

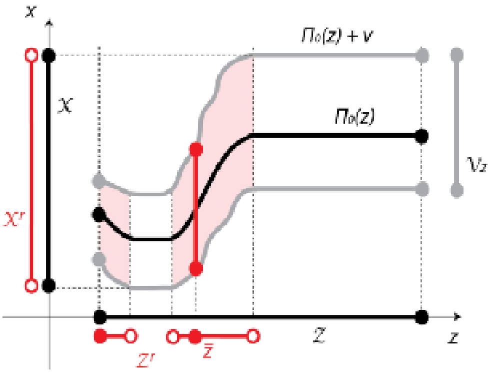


FIGURE 1. Identification under Assumption ID2′(a), univariate z and no z_1.

the data. Given ID2′(b) that v (and thus x) has a full conditional support, ID2 is trivially satisfied without additional restrictions. This assumption on $\mathcal{V}_z$ is satisfied with, for example, a normally distributed error term (conditional on regressors).

Figure 1 illustrates the intuition of the identification proof under ID2′(a) in a simple case where z is univariate. First, by $\partial E[y|v,z]/\partial z = (\partial g_0(x)/\partial x)\cdot(\partial \Pi_0(z)/\partial z)$ and the rank condition, $g_0(x)$ is locally identified on x corresponding to a point of z in the relevant set $\mathcal{Z}^r$. As such a point of z varies within $\mathcal{Z}^r$, the x corresponding to it also varies enough to cover almost the entire support of x. At the same time, for any x corresponding to an irrelevant z (i.e., z outside of $\mathcal{Z}^r$), one can always find z inside of $\mathcal{Z}^r$ that gives the same value of such an x. The probability $\Pr[z \in \mathcal{Z}^r]$ being small is related to the weak identification concept discussed later. Note that $g_0(x)$ is overidentified on a subset of $\mathcal{X}$ that corresponds to *multiple* subsets of $\mathcal{Z}$ where $\Pi_0(\cdot)$ has a nonzero slope, since each association of x and z contributes to identification. This discussion implies that the shape of $\Pi_0(\cdot)$ provides useful information on the strength of identification in different parts of the domain of $g_0(x)$.

3. WEAK IDENTIFICATION

The previous section discusses the structure of the joint distribution of x and z that contributes to the identification of $g_0(\cdot)$. Specifically, (2.4) imposes a minimal restriction on the shape of the conditional mean function $E[x|z] = \Pi_0(z)$. This necessity result suggests that "slight violation" of (2.4) will result in weak identification of $g_0(\cdot)$. Note that this approach will *not* be successful with (2.3) of NPV, since violating the condition, that is, $\Pr[\text{rank}(\partial \Pi_0(z)/\partial z') = d_x] < 1$, can still result in identification.

In this section, we formally construct the notion of weak identification via localization. We define *nonparametric weak instruments* as a drifting sequence of reduced-form functions that are localized around a function with no identification power. Such a sequence of models or drifting data-generating process (Davidson and MacKinnon (1993)) is introduced to define weak instruments relative to the sample size n. As a result, the strength of instruments is represented in terms of the rate of localization, and hence, it can eventually be reflected in the local asymptotics of the estimator of $g_0(\cdot)$.

Let $\mathcal{C}(\mathcal{Z})$ be the class of conditional mean functions $\Pi(\cdot)$ on $\mathcal{Z}$ that are bounded and continuously differentiable. Define a *nonidentification region* $\mathcal{C}_0(\mathcal{Z})$ as a class of functions that satisfy the lack-of-identification condition motivated by (2.4):[10] $\mathcal{C}_0(\mathcal{Z}) = \{\Pi(\cdot) \in \mathcal{C}(\mathcal{Z}) : \Pr[\text{rank}(\partial \Pi(z)/\partial z') < d_x] = 1\}$. Define an *identification region* as $\mathcal{C}_1(\mathcal{Z}) = \mathcal{C}(\mathcal{Z}) \backslash \mathcal{C}_0(\mathcal{Z})$. We consider a sequence of triangular models $y = g_0(x) + \varepsilon$ and $x = \Pi_n(z) + v$ with corresponding stochastic assumptions. Although $g(x)$ is identified with $\Pi_n(\cdot) \in \mathcal{C}_1(\mathcal{Z})$ for any fixed n by Theorem 2.2, $g(x)$ is only weakly identified as $\Pi_n(\cdot)$ drifts toward a function $\bar{\Pi}(\cdot)$ in $\mathcal{C}_0(\mathcal{Z})$. Namely, the noise (i.e., v) contributes more than the signal (i.e., $\Pi_n(z)$) to the total variation of $x \in \{\Pi_n(z) + v : z \in \mathcal{Z}, v \in \mathcal{V}\}$ as $n \to \infty$. In order to facilitate a meaningful asymptotic theory in which the effect of weak instruments is reflected, we further proceed by considering a specific sequence of $\Pi_n(\cdot)$.

Assumption L (Localization). *For some $\delta > 0$, the true reduced-form function $\Pi_n(\cdot)$ satisfies the following. For some $\tilde{\Pi}(\cdot) \in \mathcal{C}_1(\mathcal{Z})$ that does not depend on n and for $z \in \mathcal{Z}$,*

$$\frac{\partial \Pi_n(z)}{\partial z'} = n^{-\delta} \cdot \frac{\partial \tilde{\Pi}(z)}{\partial z'} + o(n^{-\delta}).$$

Assumption L is equivalent to

$$\Pi_n(z) = n^{-\delta} \cdot \tilde{\Pi}(z) + c + o(n^{-\delta}) \tag{3.1}$$

for some constant vector c. This specification of a uniform convergent sequence over $\mathcal{Z}$ can be justified by our identification analysis. The "local nesting" device in (3.1) is also used in Stock and Wright (2000) and Jun and Pinkse (2012) among others. In contrast to these papers, the value of δ measures the strength of identification here and is not specified to be $1/2$.[11] Unlike a linear reduced form, to characterize weak instruments in a more general nonparametric reduced form, we need to control the complete behavior of the reduced-form function, and the derivation of local asymptotic theory seems to be more demanding. Nevertheless, the particular sequence considered in Assumption L makes the weak instrument asymptotic theory straightforward while embracing the most interesting local alternatives against nonidentification.[12]

4. Estimation

Once the endogeneity is controlled by the control function in (2.2), the problem becomes one of estimating the additive nonparametric regression function $E[y|x, z] =$

[10] The lack of identification condition is satisfied either when the order condition fails ($d_z < d_x$), or when z are jointly irrelevant for one or more of x, almost everywhere in their support.

[11] It would be interesting to have different rates across columns or rows of $\frac{\partial \Pi_n(\cdot)}{\partial z'}$. One can also consider different rates for different elements of the matrix. The analyses in these cases can analogously be done by slight modifications of the arguments.

[12] In defining weak instruments in Assumption L, one can consider an intermediate case where $\frac{\partial \Pi_n(\cdot)}{\partial z'}$ converges to a matrix with reduced-rank rather than that with zero rank. Extending the analysis in this case can follow analogously but omitted in the paper for succinctness.

$g_0(x)+\lambda_0(v)$. In a weak instrument environment, however, we face a nonstandard problem called *concurvity*: $x = \mathit{\Pi}_n(z) + v \to v$ as $n \to \infty$ under the weak instrument specification (3.1) of Assumption L (with $c = 0$ for simplicity). With a series representation $g_0(x)+\lambda_0(v) = \sum_{j=1}^{\infty}\{\beta_{1j}p_j(x)+\beta_{2j}p_j(v)\}$, where the $p_j(\cdot)$'s are the approximating functions, it becomes a familiar problem of multicollinearity as $p_j(x) \to p_j(v)$ for all j. More precisely, $p_j(x) - p_j(v) = O(n^{-\delta})$ by mean value expansion $p_j(v) = p_j(x - n^{-\delta}\tilde{\mathit{\Pi}}(z)) = p_j(x) - n^{-\delta}\tilde{\mathit{\Pi}}(z)\partial p_j(\tilde{x})/\partial x$ with an intermediate value $\tilde{x}$. Alternatively, by plugging this expression of $p_j(v)$ back into the series, we can see that the variation of the regressor shrinks as $n \to \infty$.

Given the connection between the weak instruments and concurvity, the regularization methods used in the realm of research concerning inverse problems are suitable for use with weak instruments. There are two types of regularization methods used in the literature: *the truncation method* and *the penalization method*.[13] In this paper, we introduce the penalization scheme. The nature of our problem is such that the truncation method does not work properly, since $p_j(x) \to p_j(v)$ even for $j \le J < \infty$ as $n \to \infty$, where J is a truncation point. On the other hand, the penalization directly controls the behavior of the β_{1j}'s and β_{2j}'s, and hence, it successfully regularizes the weak instrument problem.

We propose a penalized series estimation procedure for $h_0(w) = g_0(x) + \lambda_0(v)$ where $w = (x, v)$. We choose to use series estimation rather than other nonparametric methods as it is more suitable in our particular framework. Because $x \to v$ the joint density of w becomes concentrated along a lower dimensional manifold as n tends to infinity. With series estimation, it is easy to impose the additivity of $h_0(\cdot)$ and to characterize the problem of weak instruments as the concurvity problem.

The estimation procedure takes two steps. In the first stage, we estimate the reduced form $\mathit{\Pi}_n(\cdot)$ using a standard series estimation method and obtain the residual $\hat{v}$. In the second stage, we estimate the structural function $h_0(\cdot)$ using a penalized series estimation method with $\hat{w} = (x, \hat{v})$ as the regressors. The theory that follows uses orthogonal wavelets or B-splines as approximating functions; see Chen (2007, Section 2.3.1) for the formal definitions. Let $\{(y_i, x_i, z_i)\}_{i=1}^{n}$ be the data with n observations, and let $r^L(z_i) = (r_{1L}(z_i), \dots, r_{LL}(z_i))'$ be a vector of approximating functions of order L for the first stage. Define a matrix $\underset{n\times L}{R} = (r^L(z_1), \dots, r^L(z_n))'$. Then, regressing x_i on $r^L(z_i)$ gives $\hat{\mathit{\Pi}}(\cdot) = r^L(\cdot)'\hat{\gamma}$ where $\hat{\gamma} = (R'R)^{-1}R'(x_1, \dots, x_n)'$, and we obtain $\hat{v}_i = x_i - \hat{\mathit{\Pi}}(z_i)$. Define a vector of approximating functions of order K for the second stage as $p^K(w) = (p_{1K}(w), \dots, p_{KK}(w))'$. To reflect the additive structure of $h_0(\cdot)$, there are no interaction terms between the approximating functions for $g_0(\cdot)$ and those for $\lambda_0(\cdot)$ in this vector; see Appendix A for the explicit expression. Denote a matrix of approximating functions as $\underset{n\times K}{\hat{P}} = (p^K(\hat{w}_1), \dots, p^K(\hat{w}_n))'$ where $\hat{w}_i = (x_i, \hat{v}_i)$. Note that $L = L(n)$ and $K = K(n)$ grow with n.

[13] In Chen and Pouzo (2012), closely related concepts are used in different terminologies: minimizing a criterion over finite sieve space and minimizing a criterion over infinite sieve space with a Tikhonov-type penalty.

We define a *penalized series estimator*:

$$\hat{h}_\tau(w) = p^K(w)'\hat{\beta}_\tau, \tag{4.1}$$

where the "interim" estimator $\hat{\beta}_\tau$ optimizes a penalizing criterion function

$$\hat{\beta}_\tau = \arg\min_{\tilde{\beta}\in\mathbb{R}^K} (y - \hat{P}\tilde{\beta})'(y - \hat{P}\tilde{\beta})/n + \tau_n\tilde{\beta}'D_n\tilde{\beta}, \tag{4.2}$$

where $y = (y_1, \dots, y_n)'$, D_n is some diagonal matrix (that may depend on n), and $\tau_n \geq 0$ the penalization parameter. Note that the penalty term $\tau_n\tilde{\beta}'D_n\tilde{\beta}$ penalizes the coefficients of the series, which effectively imposes smoothness restrictions on $h_0(\cdot)$.[14] In order to control the bias, τ_n is assumed to converge to zero. The optimization problem (4.2) yields a closed-form solution:

$$\hat{\beta}_\tau = \big(\hat{P}'\hat{P} + n\tau_n D_n\big)^{-1}\hat{P}'y.$$

The concurvity feature discussed above is manifested here by the fact that the matrix $\hat{P}'\hat{P}$ is nearly singular under Assumption L, since the two columns of $\hat{P}$ become nearly identical. In terms of the population second moment matrix $Q = E[p^K(w_i)p^K(w_i)']$, the challenge is that the minimum eigenvalue of Q is not bounded away from zero, which is manifested as $\lambda_{\max}(Q^{-1}) = O(n^{2\delta})$ (shown in Lemma A.1 in Appendix A) where $\lambda_{\max}$ denotes the maximum eigenvalue. The term $n\tau_n D_n$ mitigates such singularity, without which the performance of the estimator of $h_0(\cdot)$ would deteriorate severely.[15] The relative effects of weak instruments ($n^{2\delta}$) and penalization (τ_n) will determine the asymptotic performance of $\hat{h}_\tau(\cdot)$. Given $\hat{h}_\tau(\cdot)$, with the normalization that $\lambda_0(\bar{v}) = \bar{\lambda}$, we have $\hat{g}_\tau(x) = \hat{h}_\tau(x, \bar{v}) - \bar{\lambda}$.

5. Consistency and rate of convergence

First, we state the regularity conditions and key preliminary results under which we find the rate of convergence of the penalized series estimator introduced in the previous section. Let $\tilde{x} = (x, z)$.

Assumption A. $\{(y_i, x_i, z_i) : i = 1, 2, \dots\}$ *are i.i.d. and* $\operatorname{var}(x|z)$ *and* $\operatorname{var}(y|\tilde{x})$ *are bounded functions of z and $\tilde{x}$, respectively.*

Assumption B. (z, v) *is continuously distributed with density that is bounded away from zero on $\mathcal{Z} \times \mathcal{V}$, and $\mathcal{Z} \times \mathcal{V}$ is a cartesian product of compact, connected intervals.*

[14] Our main theory may still follow with a more general form of penalization, but we use this L_2-type as it is one of the penalty specifications that ensure a closed-form solution for $\hat{\beta}_\tau$. Within the L_2-type penalty, we flexibly allow D_n to depend on n to help control the penalty bias; for example, we can penalize higher order terms or equally-spaced terms.

[15] In linear settings, the introduction of a regularization method is less appealing as it creates the well-known biased estimator of ridge regression. In contrast, we do not directly interpret $\hat{\beta}_\tau$ in the current nonparametric setting, since it is only an interim estimator calculated to obtain $\hat{h}_\tau(\cdot)$. More importantly, the overall bias of $\hat{h}_\tau(\cdot)$ is unlikely to be worsened in the sense that the additional bias introduced by penalization can be dominated by the existing series estimation bias.

Assumption B is useful to bound below and above the eigenvalues of the "transformed" second moment matrix of approximating functions. This condition is worthy of discussion in the context of identification and weak identification. Let f_u and f_w denote the density functions of $u=(z,v)$ and $w=(x,v)$, respectively. An identification condition like Assumption ID2′ in Section 2 is embodied in Assumption B. To see this, note that f_u being bounded away from zero means that there is no functional relationship between z and v, which in turn implies Assumption ID2′(a)(iii).[16] On the other hand, an assumption written in terms of f_w like Assumption 2 in NPV (p. 574) cannot be imposed here. Observe that $w=(\Pi_n(z)+v,v)$ depends on the behavior of $\Pi_n(\cdot)$, and hence f_w is not bounded away from zero uniformly over n under Assumption L and approaches a singular density. Technically, making use of a transformation matrix (see Appendix A), an assumption is made in terms of f_u, which is not affected by weak instruments, and the effect of weak instruments can be handled separately in the asymptotics proof. Note that the assumption for the Cartesian products of supports, namely $\mathcal{Z}\times\mathcal{V}$ and its compactness can be replaced by introducing a trimming function as in NPV, that ensures bounded rectangular supports.[17] Assumption B can be weakened to hold only for some component of the distribution of z; some components of z can be allowed to be discrete as long as they have finite supports.

Next, Assumption C is a smoothness assumption on the structural and reduced-form functions. Let $\mathcal{W}$ be the support of $w=(x,v)$.

Assumption C. *$g_0(\cdot)$ and $\lambda_0(\cdot)$ are Lipschitz and continuously differentiable of order s on $\mathcal{W}$. $\Pi_n(\cdot)$ is bounded and continuously differentiable of order s_π on $\mathcal{Z}$.*

This assumption ensures that the series approximation error shrinks as the number of approximating functions increases.

Assumption D. (i) *$n^\delta K^2(\sqrt{L/n}+L^{-s_\pi/d_z})\to 0$ and $n^{-\delta}K^3\to 0$.* (ii) *Also, $\tau_n\to 0$ and $\beta' D_n\beta=O(\lambda_n^2)=O(1)$.*

Assumption D(i) restricts the rate of growth of the numbers K and L of the approximating functions. The conditions on K and L are more restrictive than the corresponding assumption for splines in NPV (Assumption 4, p. 575) where weak instruments are not considered. Assumption D(ii) ensures the bias from the penalization shrinks to zero. The assumption that $\lambda_n=O(1)$ is naturally satisfied when the series has decaying coefficients (e.g., Trefethen (2008, Theorem 4.2)) provided that the diagonal elements of D_n are bounded by a fixed constant. Now, we provide the upper bounds of the rates of convergence in probability of the penalized series estimator $\hat{h}_\tau(w)$ in terms of L_2 and uniform distance.[18] Let $\|h\|_{L_2}=\{\int[h(w)]^2\,dF(w)\}^{\frac{1}{2}}$ and $\|h\|_\infty=\sup_{w\in\mathcal{W}}|h(w)|$.

[16] The definition of a functional relationship can be found, for example, in NPV (p. 568).

[17] Assumption ID2′(b) is then viewed to hold for $h(w)$ multiplied by a trimming function, thus identification is still achieved over the trimmed support.

[18] In the Appendix, we also provide the rate results for the structural estimator $\hat{g}_\tau(\cdot)$ after subtracting the constant term which is not identified; see Theorem A.1.

Theorem 5.1. *Suppose Assumptions* A–D *and* L *are satisfied. Let* $R_n = \min\{n^\delta, \tau_n^{-1/2}\}$. *Then*

$$\|\hat{h}_\tau - h_0\|_{L_2} = O_p\big(R_n\big(\sqrt{K/n} + K^{-\frac{s}{d_x}} + \tau_n R_n \lambda_n + \sqrt{L/n} + L^{-\frac{s_\pi}{d_z}}\big)\big).$$

Also,

$$\|\hat{h}_\tau - h_0\|_\infty = O_p\big(R_n\sqrt{K}\big(\sqrt{K/n} + K^{-\frac{s}{d_x}} + \tau_n R_n \lambda_n + \sqrt{L/n} + L^{-\frac{s_\pi}{d_z}}\big)\big).$$

Suppose there is no penalization ($\tau_n = 0$). Then with $R_n = O_p(n^\delta)$, Theorem 5.1 provides the rates of convergence of the unpenalized series estimator $\hat{h}(\cdot)$. For example, with $\|\cdot\|_{L_2}$,

$$\|\hat{h} - h_0\|_{L^2} = O_p\big(n^\delta\big(\sqrt{K/n} + K^{-\frac{s}{d_x}} + \sqrt{L/n} + L^{-\frac{s_\pi}{d_z}}\big)\big). \tag{5.1}$$

Compared to the strong instrument case of NPV (Lemma 4.1, p. 575), the rate deteriorates by the leading n^δ rate, the weak instrument rate. Note that the terms $\sqrt{K/n}$ and K^{-s/d_x} correspond to the variance and bias of the second stage estimator, respectively,[19] and $\sqrt{L/n}$ and L^{-s_π/d_z} are those of the first stage estimator. The latter rates appear here due to the fact that the residuals $\hat{v}_i$ are generated regressors obtained from the first-stage nonparametric estimation. The way that n^δ enters into the rate implies that the effect of weak instruments (hence concurvity) not only exacerbates the variance but also the bias.[20] Moreover, the symmetric effect of weak instruments on bias and variance implies that the effect of weak instruments *cannot* be alleviated by the choice of the number of terms in the series estimator. This is also related to the discussion in Section 4 that the truncation method does not work as a regularization method for weak instruments.

More importantly, in the case where penalization is in operation ($\tau_n > 0$), the way that R_n enters into the convergence rates implies that penalization can reduce both bias and variance by the same mechanism working in an opposite direction to the effect of weak instruments. Penalization introduces additional bias τ_n, but it can possibly be controlled in the context of the current nonparametric estimation, for example, by assuming that $\tau_n R_n \lambda_n \leq CK^{-s/d_x}$ for some $C > 0$. Then the rate becomes

$$\|\hat{h}_\tau - h_0\|_{L_2} = O_p\big(\tau_n^{-\frac{1}{2}}\big(\sqrt{K/n} + K^{-\frac{s}{d_x}} + \sqrt{L/n} + L^{-\frac{s_\pi}{d_z}}\big)\big). \tag{5.2}$$

Here, the overall rate is improved since the multiplying rate $\tau_n^{-1/2}$ is of smaller order than the multiplying rate n^δ of the previous case. The faster convergence rate is achieved in this upper bound at the expense of introducing a tuning parameter τ_n. Assumption D implicitly assumes that $\delta > 0$. Using penalization when instruments are rather strong ($\delta = 0$) may cause a substantial bias problem. This motivates a data-driven choice of τ_n that is adaptive to the strength of instruments; see Theorem 5.4 below.

[19] The dimension of w is reduced to the dimension of x as the additive structure of $h_0(w)$ is exploited; see, for example, Andrews and Whang (1990).

[20] This is different from a linear case where multicollinearity only results in imprecise estimates but does not introduce bias. This is also different from the ill-posed inverse problem where the degree of ill-posedness only affects variance.

Next, we find the balanced L^2 convergence rate. For a more concrete comparison between the rates n^{δ} and $\tau_n^{-1/2}$, let $\tau_n = n^{-2\delta_\tau}$ for some $\delta_\tau > 0$. For example, the larger δ_τ is, the faster the penalization parameter converges to zero, and hence, the smaller the effect of penalization is.

Corollary 5.2. *Suppose the Assumptions of Theorem 5.1 are satisfied and suppose* $\tau_n R_n \lambda_n = \lambda_n \min\{n^{\delta-2\delta_\tau}, n^{-\delta_\tau}\} \le CK^{-s/d_x}$ *for some* $C > 0$*. Let* $K = O(n^{1/(1+2s/d_x)})$ *and* $L = O(n^{1/(1+2s_\pi/d_z)})$*. Then* $\|\hat{h}_\tau - h_0\|_{L^2} = O_p(n^{-q}) = o_p(1)$*, where* $q = \min\{\frac{s}{d_x+2s}, \frac{s_\pi}{d_z+2s_\pi}\} - \min\{\delta, \delta_\tau\}$.

In order to facilitate discussions on the balanced convergence rate, suppose the reduced form $\Pi_n(\cdot)$ is known. Then $q = \frac{s}{d_x+2s} - \min\{\delta, \delta_\tau\}$. Corollary 5.2 has several implications. First, consider a weak instruments-prevailing case of $\delta < \delta_\tau$. When the structural function is less smooth or has a high dimensional argument (i.e., small s or large d_x, and hence, small $\frac{s}{d_x+2s}$), instruments should not be too weak (i.e., small δ) to achieve the same rate (i.e., holding q fixed).[21] This implies a trade-off between the smoothness of the structural function (or the dimensionality) and the required strength of instruments. This, in turn, implies that the weak instrument problem can be mitigated with some smoothness restrictions, which is in fact one of our justifications for introducing the penalization method.[22] When the effect of weak instruments is prevailing, the balanced convergence rate of the unpenalized estimator $\hat{h}$ becomes

$$\|\hat{h} - h_0\|_{L^2} = O_p\big(n^{-\frac{s}{d_x+2s}+\delta}\big). \tag{5.3}$$

Even in the best scenario of $s \to \infty$, it requires that $0 < \delta < 1/2$ for consistency, which implies that instruments need to be mildly weak compared to the $n^{-1/2}$ rate typically introduced in parametric settings.

Once the penalization effect is prevailing ($\delta_\tau < \delta$), Corollary 5.2 suggests that q increases and the penalized estimator $\hat{h}_\tau$ can achieve a faster rate. As mentioned earlier, however, this rate is only an upper bound. To complete the argument that $\hat{h}_\tau$ can outperform $\hat{h}$ in terms of the convergence rate, we show the rate of $\hat{h}$ in (5.3) is the best achievable rate in the sense that the minimax lower bound rate of $\hat{h}$ coincides its upper bound $n^{-\frac{s}{d_x+2s}+\delta}$ in (5.3). Then, when the penalization effect is prevailing, the best possible rate of $\hat{h}$ is no faster than the upper bound of the rate of $\hat{h}_\tau$, namely $O_p(n^{-\frac{s}{d_x+2s}+\delta_\tau})$. This in turn implies that the minimax risk of the penalized estimator *cannot* be larger than that of the unpenalized estimator. The following theorem states the result under the setting considered in the preceding discussion. Let $B(s, L)$ be a Sobolev ball with smoothness s and radius L; see the proof of the theorem in the Appendix for the detailed definition.

[21] This relationship between the "degree of weak instruments" and the smoothness is analogous to the relationship between the measure of ill-posedness and the smoothness of functions in the NPIV literature (Blundell, Chen, and Kristensen (2007), Chen and Christensen (2018b)).

[22] With weak instruments, optimal rates in the sense of Stone (1982) (which is the rate achieved in NPV when the first stage is known) are not attainable. Also the uniform convergence rate does not attain Stone's (1982) bound even without the weak instrument factor (Newey (1997, p. 151)), and hence is not discussed here and in the minimax bound below.

Theorem 5.3 (Lower Bound Without Penalization). *Suppose Assumptions A, B, and L hold in model (2.1a)–(2.1b), and* $E[(y_i - h_0(w_i))^2|w_i] \geq \underline{\sigma}^2 > 0$ *uniformly for* w_i*. Also, suppose* $h_0(\cdot) \in B(s, L)$ *and* $\Pi_n(\cdot)$ *is known. Then*

$$\inf_{\hat{h}} \sup_{h \in B(s,L)} \Pr{}_h\big(\|\hat{h} - h\|_{L^2} \geq Cn^{-\frac{s}{d_x+2s}+\delta}\big) \geq C' > 0,$$

where $\inf_{\hat{h}}$ *denotes the infimum over all estimators of* h *with the sample size* n*, and* C *and* C' *are constants that do not depend on* n*.*

The same minimax rate is achieved for $g_0(x)$ as well, which can be found within the proof of this theorem.

When implementing the penalized series estimator in practice, there remains the issue of choosing tuning parameters, namely, the penalization parameter $\tau = \tau_n$ and the orders K and L of the series.[23] In particular, the choice of τ is important in the context of this paper, since we want τ to be adaptive to the strength of instruments, among other things. For the choice of τ, we propose the following data-driven method.[24] Recall that $\hat{\beta}_\tau = \hat{Q}_\tau^{-1}\hat{P}'y/n$ where $\hat{Q}_\tau = \hat{Q} + \tau D_n$ with $\hat{Q} = \hat{P}'\hat{P}/n$. We motivate our data-driven method by the balancing principle in the decomposition of $\|\hat{\beta}_\tau - \beta\|$:

$$\|\hat{\beta}_\tau - \beta\| \leq \big\|\hat{\beta}_\tau - \hat{Q}_\tau^{-1}\hat{Q}\beta\big\| + \big\|\hat{Q}_\tau^{-1}\hat{Q}\beta - \beta\big\|,$$

where, on the right-hand side, the first term is the stability bound that is decreasing in τ and the second term is the approximation error that is increasing in τ. This decomposition is motivated by the proof of Theorem 5.1. To apply a Lepskii (1991)-type method, we discretize the support of τ and define $\mathcal{T} = \{\tau_j : 0 < \tau_0 < \tau_1 < \cdots < \tau_N\}$. The data-driven τ would be

$$\tau^\dagger = \max \mathcal{T}_0 = \max\big\{\tau \in \mathcal{T} : \big\|\hat{Q}_\tau^{-1}\hat{Q}\beta - \beta\big\| \leq \big\|\hat{\beta}_\tau - \hat{Q}_\tau^{-1}\hat{Q}\beta\big\|\big\}.$$

Based on $\mathcal{T}_0$, we can introduce a feasible choice $\hat{\tau}$ by defining a set $\hat{\mathcal{T}}$:

$$\begin{aligned}\hat{\tau} = \max \hat{\mathcal{T}} = \max\big\{&\tau_j \in \mathcal{T} : \|\hat{\beta}_{\tau_j} - \hat{\beta}_{\tau_k}\| \leq 2\big\{\big\|\big(I - \hat{Q}_{\tau_j}^{-1}\hat{Q}\big)\hat{\beta}_{\tau_j}\big\| + \big\|\big(I - \hat{Q}_{\tau_k}^{-1}\hat{Q}\big)\hat{\beta}_{\tau_k}\big\|\big\},\\ &k = 0, 1, \ldots, j\big\}.\end{aligned}$$

This approach is related to Pereverzev and Schock (2005), who develop a Lepskii-type procedure of choosing a Tikhonov regularization parameter in ill-posed inverse problems.[25] Given $\hat{\tau}$, we can show that the data-driven penalized estimator $\hat{h}_{\hat{\tau}}$ achieves the following adaptive rate.

[23] In the simulations, we present results with a few chosen values of τ, K, and L. The cross-validation method (Arlot and Celisse (2010)) may also work here.

[24] For the data-driven choice of K and L, a similar approach can be used. Since such a problem is studied in the literature (e.g., Chen and Christensen (2018a)) in a related setting, we only focus on the choice of τ in the current paper.

[25] For references for similar approaches that use the balancing principle, see also Chen and Christensen (2018a), Breunig and Johannes (2016), Pouzo (2016), and Jansson and Pouzo (2019), among others.

Theorem 5.4 (Rate-Adaptivity). *Suppose Assumptions A–D and L are satisfied. Let $R_n^\dagger = \min\{n^\delta, (\tau^\dagger)^{-1/2}\}$. Then*

$$\|\hat{h}_{\hat{\tau}} - h_0\|_{L_2} = O_p\big(R_n^\dagger\big(\sqrt{K/n} + K^{-\frac{s}{d_x}} + \sqrt{L/n} + L^{-\frac{s_\pi}{d_z}}\big)\big).$$

By the construction of $\tau^\dagger$, the penalty bias ($\tau^\dagger R_n^\dagger \lambda_n$) is dominated by the variance term ($\sqrt{K/n}$), which is shown within the proof of this theorem, and thus the former is omitted in the rate expression. When instruments are strong (i.e., $\delta = 0$, violating Assumption D), $\tau^\dagger$ is chosen to be a negligible value by construction, since the upper bound in $\mathcal{T}_0$ satisfies $\|\hat{\beta}_{\tau_j} - \hat{Q}_{\tau_j}^{-1}\hat{Q}\beta\| = O_p(\sqrt{K/n} + K^{-s/d_x} + \Delta_\pi)$ for any τ_j in this case.

Before closing this section, we discuss one of the practical implications of the identification and asymptotic results thus far. In applied research that uses nonparametric triangular models, a linear specification of the reduced form is largely prevalent; see, for example, NPV, Blundell and Duncan (1998), Blundell, Duncan, and Pendakur (1998), Yatchew and No (2001), Lyssiotou, Pashardes, and Stengos (2004), Dustmann and Meghir (2005), and Del Bono and Weber (2008). While a linear reduced-form relationship is rarely justified by economic theory, linear specification is introduced to avoid the curse of dimensionality with many covariates, or for an ad hoc reason that it is easy to implement and that the nonparametric structural equation is of primary interest. When the reduced form is linearly specified, however, any true nonlinear relationship is "flattened out," and the situation is more likely to have the problem of weak instruments, let alone the problem of misspecification. On the other hand, one can achieve a significant gain in the performance of the estimator by nonparametrically estimating the relationship of x and z. According to (2.4), identification power can be enhanced by exploiting the entire nonlinear relationship between x and z. This phenomenon may be interpreted in terms of the "optimal instruments" in the GMM settings of Amemiya (1977); see also Newey (1990) and Jun and Pinkse (2012). The nonparametric first stage estimation is not likely to worsen the overall convergence rate of the estimator, since the nonparametric rate from the second stage is already present.

6. Asymptotic distributions

We establish the asymptotic normality of the functionals of the penalized series estimator $\hat{h}_\tau(\cdot)$. We consider linear functionals of $h_0(\cdot)$ that include $h_0(\cdot)$ at a certain value (i.e., $h_0(\bar{w})$) and the weighted average derivative of $h_0(\cdot)$ (i.e., $\int \vartheta(w)[\partial h_0(w)/\partial x]\,dw$). The linear functionals of $h = h_0(\cdot)$ are denoted as $a(h)$. Then the estimator $\hat{\theta}_\tau = a(\hat{h}_\tau)$ of $\theta_0 = a(h)$ is the natural "plug-in" estimator. Let $A = (a(p_{1K}), a(p_{2K}), \dots, a(p_{KK}))$, where $p_{jK}(\cdot)$ is an element of $p^K(\cdot)$. Then

$$\hat{\theta}_\tau = a(\hat{h}_\tau) = a\big(p^K(x)'\hat{\beta}_\tau\big) = A\hat{\beta}_\tau.$$

Then the following variance estimator of $\hat{\theta}_\tau$ can naturally be defined:

$$\hat{V}_\tau = A\hat{Q}_\tau^{-1}\big(\hat{\Sigma}_\tau + \hat{H}_\tau\hat{Q}_1^{-1}\hat{\Sigma}_1\hat{Q}_1^{-1}\hat{H}_\tau'\big)\hat{Q}_\tau^{-1}A',$$

$$\hat{\Sigma}_\tau = \sum_{i=1}^{n} p^K(\hat{w}_i)p^K(\hat{w}_i)'\big[y_i - \hat{h}_\tau(\hat{w}_i)\big]^2/n, \qquad \hat{\Sigma}_1 = \sum_{i=1}^{n} \hat{v}_i^2 r^L(z_i)r^L(z_i)'/n,$$

$$\hat{H}_\tau = \sum_{i=1}^{n} p^K(\hat{w}_i)\big\{\big[\partial\hat{h}_\tau(\hat{w}_i)/\partial w\big]'\partial\omega\big(\tilde{x}_i, \hat{\Pi}(z_i)\big)/\partial\pi\big\}r^L(z_i)'/n, \qquad \hat{Q}_1 = R'R/n,$$

where $\tilde{x}$ is a vector of variables that includes x and z and $\omega(\tilde{x}, \pi)$ is a vector of functions of $\tilde{x}$ and π where π is a possible value of $\Pi(z)$. The following are additional regularity conditions for the asymptotic normality of $\hat{\theta}_\tau$. Let $\eta = y - h$.

Assumption E. $\sigma^2(\tilde{x}) = \mathrm{var}(y|\tilde{x})$ *is bounded away from zero,* $E[\eta^4|\tilde{x}]$ *is bounded, and* $E[\|v\|^4|\tilde{x}]$ *is bounded. Also,* $h_0(w)$ *is twice continuously differentiable in* v *with bounded first and second derivatives.*

This assumption strengthens the boundedness of conditional second moments in Assumption A. For the next assumption, let $|h|_r = \max_{|\mu|\le r}\sup_{w\in\mathcal{W}}|\partial^\mu h(w)|$. Also let $\tilde{p}^K(w)$ be a generic vector of approximating functions and let $\tilde{p}^{*K}(z, v)$ be a "transformation" of $\tilde{p}^K(w)$ purged of the weak instruments effect (see Appendix A).

Assumption F. *Either* (a) *or* (b) *hold*: (a) $a(h)$ *is a scalar and is continuous under* $|h|_{\tilde{r}}$ *for some* $\tilde{r} \ge 0$*, and there exists* β_K *such that as* $K \to \infty$*,* $a(p^{K\prime}\beta_K)$ *is bounded away from zero while* $E[(\tilde{p}^K(w)'\beta_K)^2] \to 0$; (b) *There exists* $\nu(w)$ *and* α_K *such that* $E[\|\nu(w)\|^2] < \infty$*,* $a(h) = E[\nu(w)h_0(w)]$*,* $a(p_j) = E[\nu(w)p_j(w)]$*, and* $E[\|\nu(\Pi_n(z) + v, v) - \tilde{p}^{*K}(z, v)'\alpha_K\|^2] \to 0$ *as* $K \to \infty$.

Assumption F(a) includes the case of h at a certain value and F(b) includes the case of the weighted average derivative of h, in which case $\nu(w) = -f_w(w)^{-1}\partial\vartheta(w)/\partial w$. The next condition restricts the rate of growth of K and L and the rate of convergence of τ_n.

Assumption G. *The following terms converge to zero as* $n \to \infty$: $\sqrt{n}K^{-s/d_x}$, $\sqrt{n}L^{-s_\pi/d_z}$, $\sqrt{L\log(L)/n}$, $R_n\sqrt{K^3L^3/n}$, $R_n^3K^{1/2}(K^3L/n + K^2\sqrt{L/n} + \sqrt{K\log(K)/n})$, $R_nK^2L/\sqrt{n}$, $R_n^2(K+L)/\sqrt{n}$. *Also,* $\tau_nR_n\lambda_n \le CK^{-s/d_x}$ *for some* $C > 0$.

Assumption G imposes more restrictions on the behavior of weak instruments and τ_n (and of K and L) than Assumption D. The conditions $\sqrt{n}K^{-s/d_x} \to 0$ and $\sqrt{n}L^{-s_\pi/d_z} \to 0$ introduce overfitting in that the bias (K^{-s/d_x}) shrinks faster than $1/\sqrt{n}$, the usual rate of standard deviation of the estimator. The same feature is found in the corresponding assumption in NPV (Assumption 8, p. 582). While the overall rate conditions on K and L in Assumption G may be stronger than that in NPV due to weak instruments, we relax the rate required to approximate the sample second moment matrices to their population counterparts, by applying recent development in Chen and Christensen (2015) and Belloni, Chernozhukov, Chetverikov, and Kato (2015). As before, the last part of Assumption G assumes that the penalty bias is no larger than the approximation bias.

Theorem 6.1. *If Assumptions A–G and L are satisfied, then*

$$\sqrt{n}\hat{V}_{\tau}^{-1/2}(\hat{\theta}_{\tau} - \theta_0) \to_d N(0, 1).$$

There still remain issues when the result of Theorem 6.1 is used for inference, for example, for constructing pointwise asymptotic confidence intervals. As nuisance parameters are present (i.e., δ or τ), an inferential procedure may depend on the strength of instruments or on the choice of the penalization parameter. The data-driven method developed in the previous section for the choice of τ may not be applicable to Theorem 6.1. Developing a procedure robust to the strength of instruments in nonparametric models or a data-driven method of choosing τ for inference is beyond the scope of our paper, and we leave it to future research.

7. Monte Carlo simulations

In this section, we document the problems of weak instruments in nonparametric estimation and investigate the finite sample performance of the penalized estimator. We are particularly interested in the finite sample gain in terms of the bias, variance, and mean squared errors (MSE) of the penalized series estimators defined in Section 4 ("penalized IV (PIV) estimators") relative to those of the unpenalized series estimators ("IV estimators") for a wide range of strength of instruments.

We consider the following data generating process:

$$y = \Phi\left(\frac{x - \mu_x}{\sigma_x}\right) + \varepsilon, \quad x = \pi_1 + z\pi + v,$$

where y, x, and z are univariate, $z \sim N(\mu_z, \sigma_z^2)$ with $\mu_z = 0$ and $\sigma_z^2 = 1$, and $(\varepsilon, v)' \sim N(0, \Sigma)$ with $\Sigma = \left[\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right]$. Note that $|\rho|$ measures the degree of endogeneity, and we consider $\rho \in \{0.2, 0.5, 0.95\}$. The sample $\{z_i, \varepsilon_i, v_i\}$ is i.i.d. with size $n = 1000$. The number of simulation repetitions is $s \in \{500, 1000\}$. We consider different strengths of the instrument by considering different values of π. Let the intercept $\pi_1 = \mu_x - \pi\mu_z$ with $\mu_x = 2$ so that $E[x] = \mu_x$ does not depend on the choice of π. Note that $\sigma_x^2 = \pi^2\sigma_z^2 + 1$ still depends on π, which is reasonable since the signal contributed to the total variation of x is a function of π. More specifically, to measure the strength of the instrument, we define the concentration parameter (Stock and Yogo (2005)): $\mu^2 = \pi^2 \sum_{i=1}^{n} z_i^2/\sigma_v^2$. Note that since the dimension of z is one, the concentration parameter value and the first-stage F-statistic are similar to each other. For example, in Staiger and Stock (1997), for $F = 30.53$ (strong instrument), a 97.5% confidence interval for μ^2 is $[17.3, 45.8]$, and for $F = 4.747$ (weak instrument), a confidence interval for μ^2 is $[2.26, 5.64]$. The candidate values of μ^2 are $\{4, 8, 16, 32, 64, 128, 256\}$, which range from a weak to a strong instrument in the conventional sense.[26] Also, with $\pi = n^{-\delta}\tilde{\pi}$ under Assumption L and $\sigma_z^2 = 1$, the concentration parameter is related to δ by $\mu^2 \approx n^{1-2\delta}\tilde{\pi}$. Suppose $\tilde{\pi} = 1$, then the range of δ that corresponds to the chosen range of μ^2 is approximately

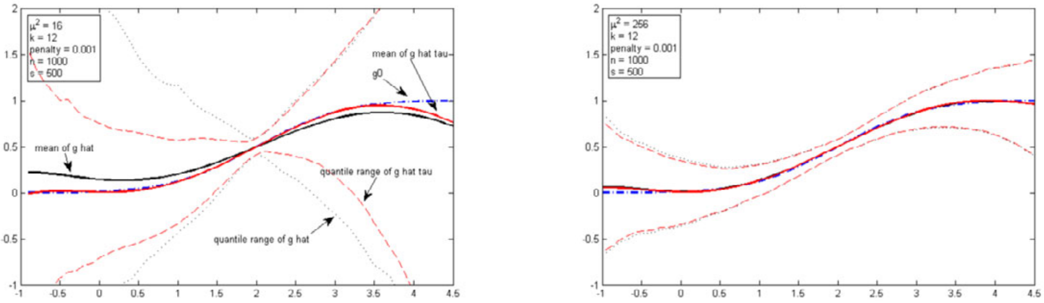
[26] The simulation results seem to be unstable when $\mu^2 = 4$ (presumably because instruments in this range are severely weak in nonparametric settings), and hence need caution when interpreting them.

$\{0.4, 0.35, 0.3, 0.25, 0.2, 0.15, 0.1\}$. As for the penalization parameter τ, we consider candidate values of $\{0.001, 0.005, 0.01, 0.05, 0.1\}$. As a benchmark in relation to Theorem 5.4, $\hat{\tau} \approx 0.0126$ for $\mu^2 = 64$ in one instance of our simulation.

The approximating functions used for $g_0(x)$ and $\lambda_0(v)$ are polynomials with different choices of (K_1, K_2), where K_1 is the number of terms for $g_0(\cdot)$, K_2 for $\lambda_0(\cdot)$, and $K = K_1 + K_2$. We introduce the normalization $\lambda_0(1) = \rho$, where ρ is chosen because of the joint normality of (ε, v). Then $g_0(x) = h_0(x, 1) - \rho$, where $h(x, v) = g(x) + \lambda(v)$.

In the first part of the simulation, we calculate $\hat{g}_\tau(\cdot)$ and $\hat{g}_0(\cdot)$, the penalized and unpenalized IV estimates, respectively, and compare their performances. For different strengths of the instrument, we compute estimates with different values of the penalization parameter. We choose $K_1 = K_2 = 6$, and $\rho = 0.5$.[27] As one might expect, the choice of orders of the series is not significant as long as we are only interested in comparing $\hat{g}_\tau(\cdot)$ and $\hat{g}(\cdot)$.

Figures 2 and 3 present some representative results. Results with different values of μ^2 and τ are similar, and hence are omitted to save space. In Figure 2, we plot the mean of $\hat{g}_\tau(\cdot)$ and $\hat{g}(\cdot)$ with concentration parameter $\mu^2 = 16$ and penalization parameter $\tau = 0.001$. In Figure 2(a), the plot for the unpenalized estimate indicates that with the given strength of the instrument, the variance is very large, which implies that functions with any trends can fit within the 0.025–0.975 quantile ranges; it indicates that the bias is also large. The graph for the penalized estimate shows that the penalization significantly reduces the variance so that the quantile range implies the upward trend of the true $g_0(\cdot)$. Note that the bias of $\hat{g}_\tau(\cdot)$ is no larger than that of $\hat{g}(\cdot)$. Although $\mu^2 = 16$ is considered to be strong according to the conventional criterion, this range of the concentration parameter value can be seen as the case where the instrument is "nonparametrically" weak in the sense that the penalization induces a significant difference between $\hat{g}_\tau(\cdot)$ and $\hat{g}(\cdot)$. Figure 2(b) is drawn with $\mu^2 = 256$, while all else remains the same.

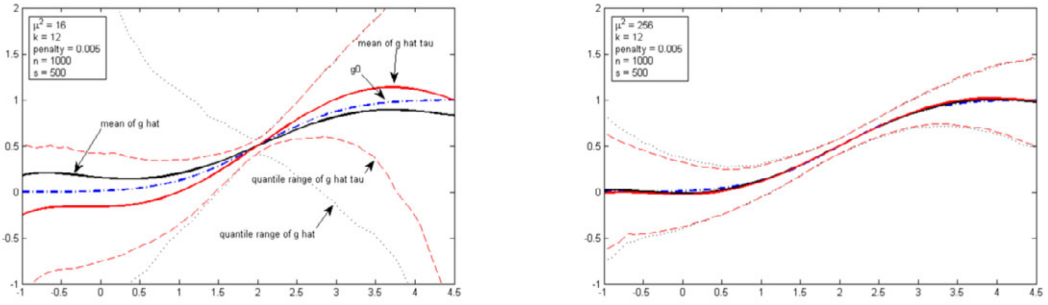


(a) with a weak instrument

(b) with a strong instrument

FIGURE 2. Penalized versus unpenalized estimators ($\hat{g}_\tau(\cdot)$ versus $\hat{g}(\cdot)$), $\tau = 0.001$.

[27] Because of the bivariate normal assumption for $(\varepsilon, v)'$, we implicitly impose linearity in the function $E[\varepsilon|v] = \lambda(v)$. Although K_2 being smaller than K_1 would better reflect the fact that $\lambda_0(\cdot)$ is smoother than $g_0(\cdot)$, we assume that we are agnostic about such knowledge.



(a) with a weak instrument

(b) with a strong instrument

FIGURE 3. Penalized versus unpenalized estimators ($\hat{g}_\tau(\cdot)$ versus $\hat{g}(\cdot)$), $\tau = 0.005$.

In this case, the penalization induces no significant difference between $\hat{g}_\tau(\cdot)$ and $\hat{g}(\cdot)$. This can be seen as the case where the instrument is "nonparametrically" strong. It is noteworthy that the bias of the penalized estimate is no larger than the unpenalized one even in this case.

Figure 3 presents similar plots but with penalization parameter $\tau = 0.005$. Figure 3(a) shows that with a larger value of τ than the previous case, the variance is significantly reduced, while the biases of the two estimates are comparable to each other. The change in the patterns of the graphs from Figure 3(a) to 3(b) is similar to those in the previous case. Furthermore, the comparison between Figure 2 and Figure 3 shows that the results are more sensitive to the change of τ in the weak instrument case than in the strong instrument case.

The fact that the penalized and unpenalized estimates differ significantly when the instrument is weak has a practical implication: Practitioners can be informed about whether the instrument they are using is worryingly weak by comparing penalized series estimates with unpenalized estimates. A similar approach can be found in the linear weak instruments literature; for example, the biased TSLS estimates and the approximately median-unbiased LIML estimates of Staiger and Stock (1997) can be compared to detect weak instruments.

Table 1 reports the integrated squared bias, integrated variance, and integrated MSE of the penalized and unpenalized IV estimators and least squares (LS) estimators of $g_0(\cdot)$. The LS estimates are calculated by series estimation of the outcome equation (with order K_1), ignoring the endogeneity. We also calculate the relative integrated MSE for comparisons. We use $K_1 = K_2 = 6$, and $\rho = 0.5$ as before. Results with different choices of orders K_1 and K_2 between 3 and 10 and a different degree of endogeneity ρ in $\{0.2, 0.95\}$ show similar patterns. Note that the usual bias and variance trade-offs are present as the order of the series changes. In the table, as the instrument becomes weaker, the bias and variance of the unpenalized IV ($\tau = 0$) increase with a greater proportion in variance. The integrated MSE ratios between the IV and LS estimators ($MSE_{\text{IV}}/MSE_{\text{LS}}$) indicate the relative performance of the IV estimator compared to the LS estimator. A ratio *larger than unity* implies that IV performs *worse* than LS. In the table, the IV estimator

TABLE 1. Integrated squared bias, integrated variance, and integrated MSE of the penalized and unpenalized IV estimators $\hat{g}_\tau(\cdot)$ and $\hat{g}(\cdot)$.

		μ^2						
		4	8	16	32	64	128	256
$\tau = 0$	$Bias^2$	0.0377	0.0335	0.0054	0.0008	0.0000	0.0003	0.0000
	Var	99.0147	3.8019	0.9395	0.1419	0.0711	0.0310	0.0186
	MSE	99.0524	3.8354	0.9449	0.1426	0.0711	0.0313	0.0186
	$MSE_{\text{IV}}/MSE_{\text{LS}}$	374.7291	15.9165	3.3232	0.5875	0.2790	0.1472	0.0901
$\tau = 0.001$	$Bias^2$	0.0328	0.0131	0.0030	0.0010	0.0002	0.0000	0.0000
	Var	0.3727	0.2557	0.1497	0.0829	0.0427	0.0349	0.0174
	MSE	0.4055	0.2688	0.1527	0.0839	0.0429	0.0349	0.0174
	$MSE_{\text{PIV}}/MSE_{\text{IV}}$	0.0035	0.1203	0.5365	0.6888	0.8297	0.9074	0.9452
$\tau = 0.005$	$Bias^2$	0.1145	0.0682	0.0305	0.0150	0.0042	0.0017	0.0010
	Var	0.4727	0.1332	0.0732	0.0894	0.0345	0.0248	0.0354
	MSE	0.5872	0.2014	0.1037	0.1045	0.0387	0.0265	0.0364
	$MSE_{\text{PIV}}/MSE_{\text{IV}}$	0.0024	0.0894	0.3501	0.5594	0.6462	0.7795	0.7464
$\tau = 0.01$	$Bias^2$	0.1566	0.1068	0.0685	0.0346	0.0158	0.0047	0.0022
	Var	0.2117	0.1981	0.2965	0.0318	0.0265	0.0183	0.0132
	MSE	0.3684	0.3049	0.3649	0.0664	0.0423	0.0230	0.0154
	$MSE_{\text{PIV}}/MSE_{\text{IV}}$	0.0037	0.0795	0.3862	0.4655	0.5942	0.7345	0.8238

does poorly in terms of MSE even when $\mu^2 = 16$, which is in the range of conventionally strong instruments; therefore, this can be considered as the case where the instrument is nonparametrically weak.

The rest of the results in Table 1 are for the penalized IV (PIV) estimator $\hat{g}_\tau(\cdot)$. Overall the variance is reduced significantly compared to that of IV without sacrificing much bias. In the case of $\tau = 0.001$, the variance is reduced for the entire range of instrument strength (compared to the unpenalized estimator). Remarkably, the bias is no larger even though penalization is in operation and is reduced when the instrument is weak. This provides evidence for the theoretical discussion in Section 5 that the penalty bias can be dominated by the existing series estimation bias. This feature diminishes as the increased value of τ introduces more bias. The integrated MSE ratios between PIV and IV ($MSE_{\text{PIV}}/MSE_{\text{IV}}$) in Table 1 suggest that PIV outperforms IV in terms of MSE for all the values of τ considered here. For example, when $\mu^2 = 8$, the MSE of PIV with $\tau = 0.001$ is only about 12% of that of IV, while the bias (squared) of PIV is only about 39% of that of IV. These results imply that PIV performs substantially better than LS unlike the previous case of IV.

The simulation results can be summarized as follows. Even with a strong instrument in a conventional sense, unpenalized IV estimators do poorly in terms of mean squared errors compared to LS estimators. Variance seems to be a bigger problem, but bias is also worrisome. Penalization alleviates much of the variance problem induced by the weak instrument, and it also works well in terms of bias for relatively weak instruments and for some values of the penalization parameter.

8. Application: Effect of class size

To illustrate our approach and apply the theoretical findings, we nonparametrically estimate the effect of class size on students' test scores. Estimating the effect of class size has been an interesting topic in the schooling literature, since among school inputs that affect students' performance, class size is thought to be easier to manipulate. Angrist and Lavy (1999) analyzed the effect of class size on students' reading and math scores in Israeli primary schools. With linear models, they find that the estimated effect is negative in most of the specifications they consider. This specific empirical application is chosen for the following reasons: (i) Although Angrist and Lavy (1999) used an instrument that is considered to be strong for their parametric model, it may not be sufficiently strong when applied in a nonparametric specification of the relationship; see below for details. (ii) The instrument is continuous in this example and presents a nonlinear relationship with the endogenous variable; see Figure 1 in Angrist and Lavy (1999). (iii) We also compare estimates calculated from our triangular model and the NPIV model in Horowitz (2011), where the same example is considered.

In this section, we investigate whether the results of Angrist and Lavy (1999) are driven by their parametric assumptions. It is also more reasonable to allow a nonlinear effect of class size, since it is unlikely that the marginal effect is constant across class-size levels. We nonparametrically extend their linear model by considering

$$score_{sc} = g(classize_{sc}, disadv_{sc}) + \alpha_s + \varepsilon_{sc}$$

for school s and class c, where $score_{sc}$ is the average test score within class, $classize_{sc}$ the class size, $disadv_{sc}$ the fraction of disadvantaged students, and α_s an unobserved school-specific effect. Note that this model allows for different patterns for different subgroups of school/class characteristics (here, $disadv_{sc}$).

Class size is endogenous because it results from choices made by parents, schooling providers or legislatures, and hence is correlated with other determinants of student achievement. Angrist and Lavy (1999) used Maimonides' rule on maximum class size in Israeli schools to construct an IV. According to the rule, class size increases one-for-one with enrollment until 40 students are enrolled, but when 41 students are enrolled, the class size is dropped to an average of 20.5 students. Similarly, classes are split when enrollment reaches 80, 120, 160, and so on, so that each class does not exceed 40. With e_s being the beginning-of-the-year enrollment count, this rule can be expressed as $f_{sc} = e_s/\{\text{int}((e_s - 1)/40) + 1\}$, which produces the IV. This rule generates discontinuity in the enrollment/class-size relationship, which serves as exogenous variation. Note that with the sample around the discontinuity points, IV exogeneity is more credible in addressing the endogeneity issue.

The dataset we use is the 1991 Israel Central Bureau of Statistics survey of Israeli public schools from Angrist and Lavy (1999). We only consider fourth graders. The sample size is $n = 2019$ for the full sample and 650 for the discontinuity sample. Given a linear reduced form, first stage tests have $F = 191.66$ with the discontinuity sample (± 7 students around the discontinuity points) and $F = 2150.4$ with the full sample. Lessons from the theoretical analyses of the present paper suggest that an instrument that is strong in a conventional sense ($F = 191.66$) can still be weak in nonparametric estimation of the class-size effect, and a nonparametric reduced form can enhance identification power. We consider the following nonparametric reduced form:

$$classize_{sc} = \Pi(f_{sc}, disadv_{sc}) + v_{sc}.$$

The sample is clustered, an aspect which is reflected in α_s of the outcome equation. Hence, we use the block bootstrap when computing standard errors and take schools as bootstrap sampling units to preserve within-cluster (school) correlation.[28] This produces cluster-robust standard errors. We use $b = 100$ bootstrap repetitions.

With the same example and dataset (only the full sample), Horowitz (2011, Section 5.2) uses the model and assumptions of the NPIV approach to nonparametrically estimate the effect of class size. To address the ill-posed inverse problem, he conducts regularization by replacing the operator with a finite-dimensional approximation. First, we compare the NPIV estimate of Horowitz (2011) with the IV estimate obtained by the control function approach of this paper. Figure 3 in Horowitz (2011) is the NPIV estimate of the function of class size ($g(\cdot, \cdot)$) for $disadv = 1.5(\%)$ with the full sample. The solid line is the estimate of g and the dots show the cluster-robust 95% confidence band. As noted in his paper (p. 374), "the result suggests that the data and the instrumental variable assumption, by themselves, are uninformative about the form of any dependence of test scores on class size." Using the same scales in the axes for comparison, Figure 4 in the present paper depicts the (unpenalized) IV estimate calculated with the full sample using the triangular model (2.1) and the control function approach. Although not entirely flexible, the nonparametric reduced form above is justified for use in the comparison with the NPIV estimate, since the NPIV approach does not specify any reduced-form

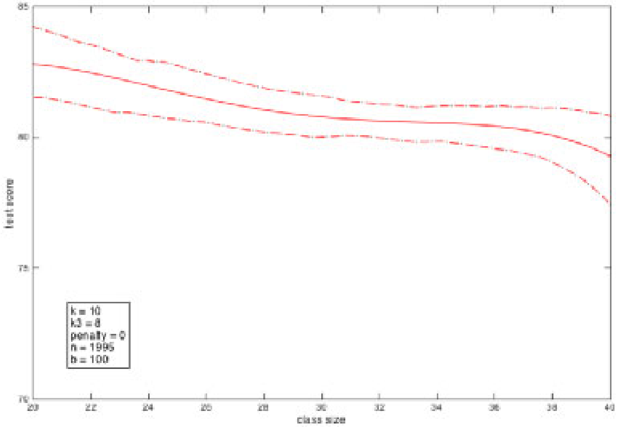


FIGURE 4. Unpenalized IV estimates with nonparametric first-stage equations, the full sample ($n = 2019$), 95% confidence band.

[28] It is worth caution that the validity of the bootstrap has not been established here. Nonetheless, the confidence bands provide a measure of the variability of the estimates.

relationship. The sample, the orders of the series, and the value of *disadv* are identical to those for the NPIV estimate. The dashed lines in the figure indicate the cluster-robust 95% confidence band. The result suggests a nonlinear shape of the effect of class size and that the marginal effect diminishes as class size increases. The overall trend seems to be negative, which is consistent with the results of Angrist and Lavy (1999).

It is important to note that the control function and NPIV approaches maintain different sets of assumptions. For example in terms of orthogonality conditions for IV, assumptions (2.1b) are not stronger or weaker than $E[\varepsilon|z]=0$, the orthogonality condition introduced in the NPIV model; only if $v \perp z$ is assumed, then $E[\varepsilon|v,z]=E[\varepsilon|v]$ with $E[\varepsilon]=0$ implies $E[\varepsilon|z]=0$. Therefore, this comparison does not imply that one estimate performs better than the other. It does, however, imply that if the triangular model and control function assumptions are considered to be reasonable, they make the data to be informative about the relationship of interest. Moreover, since the NPIV approach suffers from the ill-posed inverse problem even without the problem of weak instruments, the control function approach may be a more appealing framework than the NPIV approach in the possible presence of weak instruments.

We proceed by calculating the penalized IV estimates from the proposed estimation method of this paper. For all cases below, we find estimates for $disadv = 1.5(\%)$ as before. To better justify the usage of our method in this part, we use the discontinuity sample and a linear reduced-form where the instrument is possibly weak in this nonparametric setting. For the penalization parameter τ, we use the data-driven method suggested in Section 5, which yields $\hat{\tau}=2.2$. Figure 5 depicts the penalized and unpenalized IV estimates. There is a certain difference in the estimates, and the trend in the penalized estimate is relatively less negative. Still, note that the penalized estimate is within the 95% band of the unpenalized estimate and vice versa. Overall, the results suggest a nonlinear effect of class size with the negative trend.

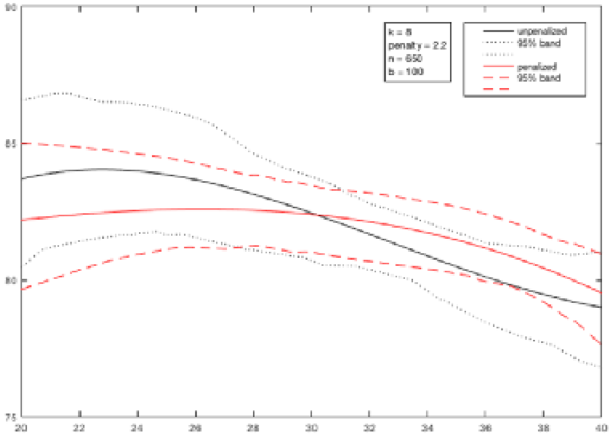


FIGURE 5. Penalized IV estimates, the discontinuity sample ($n=650$, $F=191.66$), 95% confidence band.

9. Conclusions

This paper analyzes identification, estimation, and inference in a nonparametric triangular model in the presence of weak instruments and proposes an estimation strategy to mitigate the effect. The findings and implications of this paper can be adapted to other nonparametric models, such as nonparametric limited dependent variable models of Das, Newey, and Vella (2003) and Blundell and Powell (2004), IV quantile regression models of Chernozhukov and Hansen (2005) and Lee (2007), and the marginal treatment effect (MTE) framework by Heckman and Vytlacil (2005).[29] The results can also directly be applicable in semiparametric versions of the model of this paper. As more structure is imposed on the model, the identification condition of Section 2 and the regularity conditions of Sections 5 and 6 can be weakened.

Subsequent research can consider two specification tests: a test for the relevance of the instruments and a test for endogeneity. These tests can be conducted by adapting the existing literature on specification tests where the test statistics can be constructed using the series estimators of this paper; see, for example, Hong and White (1995). Testing whether instruments are relevant can be conducted with the nonparametric reduced-form estimate $\hat{\Pi}(\cdot)$. A possible null hypothesis is $H_0 : \Pr\{\Pi_0(z) = const.\} = 1$, which is motivated by our rank condition for identification. Constructing a test for instrument weakness would be more demanding. Developing inference procedures that are robust to identification of arbitrary strength is also an important research question.

Appendix A

A.1 *Key technical steps for asymptotic theory (Section 5)*

For asymptotic theory, we require a key preliminary step to separate out the weak instrument factor from the second moment matrices of interest. Define a vector of approximating functions of orders $K = K_1 + K_2 + 1$ for the second stage,

$$p^K(w) = \big(1, p_{1K_1}(x), \ldots, p_{K_1K_1}(x), p_{1K_2}(v), \ldots, p_{K_2K_2}(v)\big)' = \big[1 \vdots p^{K_1}(x)' \vdots p^{K_2}(v)'\big]',$$

where $p^{K_1}(x)$ and $p^{K_2}(v)$ are vectors of approximating functions for $g_0(\cdot)$ and $\lambda_0(\cdot)$ of orders K_1 and K_2, respectively. Note that this rewrites $p^K(w) = (p_{1K}(w), \ldots, p_{KK}(w))'$ of the main body for expositional convenience. Since $g_0(\cdot)$ and $\lambda_0(\cdot)$ can only be separately identified up to a constant, when estimating $h_0(\cdot)$, we include only one constant

[29] For example, the MTE framework may present a similar inverse problem when instruments are weak. Suppose $\Pr[d = 1|z] = P(z)$ is the propensity score with the endogenous treatment d and a scalar instrument z. Under index sufficiency, we have $E[y|z] = E[y|P(z)]$ where y is the outcome variable. Then

$$\frac{\partial E[y|z]}{\partial z} = \frac{\partial E\big[y|P(z)\big]}{\partial z} = \frac{\partial E\big[y|P(z)\big]}{\partial p}\frac{\partial P(z)}{\partial z} = \text{MTE}\big(P(z)\big)\frac{\partial P(z)}{\partial z}$$

and, therefore, in recovering the MTE, the function $P(z)$ being close to a constant function (as in Assumption L) results in an inverse problem. Analogous to the approach in the current paper, a regularization method may be used to address this problem.

function. Define a $K \times K$ sample second moment matrix:

$$\hat{Q} = \frac{\hat{P}'\hat{P}}{n} = \frac{\sum_{i=1}^{n} p^K(\hat{w}_i) p^K(\hat{w}_i)'}{n}. \tag{A.1}$$

Then $\hat{\beta}_\tau = (\hat{Q} + \tau_n D)^{-1}\hat{P}'y/n$.

For the rest of this section, we consider univariate x for simplicity. This corresponds to Assumption ID2$'$(a). The analysis can also be generalized to the case of a vector x by using multivariate mean value expansion, but omitted for succinctness. Note that z is still a vector. Under Assumption L, $\Pi_n(\cdot) = n^{-\delta}\tilde{\Pi}(\cdot)$ after applying a normalization $c = 0$ and suppressing $o(n^{-\delta})$ for simplicity in (3.1). Omitting $o(n^{-\delta})$ does not affect the asymptotic results developed in the paper. For $r \in \{1, 2\}$, define its rth derivative as $\partial^r p_j(x) = d^r p_j(x)/dx^r$. By mean value expanding each element of $p^{K_1}(x_i)$ around v_i, we have, for $j \leq K_1$ (with the second subscript suppressed),

$$p_j(x_i) = p_j\bigl(n^{-\delta}\tilde{\Pi}(z_i) + v_i\bigr) = p_j(v_i) + n^{-\delta}\tilde{\Pi}(z_i)\partial p_j(\tilde{v}_i), \tag{A.2}$$

where $\tilde{v}_i$ is a value between x_i and v_i. Define $\partial^r p^{K_1}(x) = [\partial^r p_{1K_1}(x), \partial^r p_{3K_1}(x), \dots, \partial^r p_{K_1K_1}(x)]'$. Then, by (A.2) the vector of regressors $p^K(w_i)$ for estimating $h(\cdot)$ can be written as

$$p^K(w_i)' = \bigl[1 \vdots p^{K_1}(x_i)' \vdots p^{K_2}(v_i)'\bigr] = \bigl[1 \vdots p^{K_1}(v_i)' + n^{-\delta}\tilde{\Pi}(z_i)\partial p^{K_1}(\tilde{v}_i)' \vdots p^{K_2}(v_i)'\bigr]. \tag{A.3}$$

Let $\kappa = K_1 = K_2 = (K-1)/2$. Again, K_1, K_2, L, K, and κ all depends on n. Note that $K \asymp K_1 \asymp K_2 \asymp \kappa$, where $a_n \asymp b_n$ denote that a_n/b_n is bounded below and above by constants that are independent of n. This setting can be justified by $g_0(\cdot)$ and $\lambda_0(\cdot)$ with the same smoothness, which is imposed in Assumption C. Extending the analysis to a general case of $K_1 \neq K_2$ can follow by a slight modification of the argument with $\kappa = \min\{K_1, K_2\}$, which we omit for succinctness. Now we choose a transformation matrix T_n to be

$$T_n = \begin{bmatrix} 1 & 0_{1\times\kappa} & 0_{1\times\kappa} \\ 0_{\kappa\times 1} & n^{\delta} I_\kappa & 0_{\kappa\times\kappa} \\ 0_{\kappa\times 1} & -n^{\delta} I_\kappa & I_\kappa \end{bmatrix}.$$

After multiplying T_n on both sides of (A.3), the weak instrument factor is separated from $p^K(w_i)'$: with $u_i = (z_i, v_i)$,

$$\begin{aligned} p^K(w_i)'T_n &= \bigl[1 \vdots p^{\kappa}(v_i)' + n^{-\delta}\tilde{\Pi}(z_i)\partial p^{\kappa}(\tilde{v}_i)' \vdots p^{\kappa}(v_i)'\bigr] \cdot T_n \\ &= \bigl[1 \vdots \tilde{\Pi}(z_i)\partial p^{\kappa}(\tilde{v}_i)' \vdots p^{\kappa}(v_i)'\bigr] = p^{*K}(u_i)' + m_i^{K\prime}, \end{aligned} \tag{A.4}$$

where $p^{*K}(u_i)' = [1 \vdots \tilde{\Pi}(z_i)\partial p^{\kappa}(v_i)' \vdots p^{\kappa}(v_i)']$ and $m_i^{K\prime} = [0 \vdots \tilde{\Pi}(z_i)(\partial p^{\kappa}(\tilde{v}_i)' - \partial p^{\kappa}(v_i)') \vdots (0_{\kappa\times 1})']$. To illustrate the role of this linear transformation, rewrite the original vector of regressors in (A.3) as

$$p^K(w_i)' = p^K(w_i)'T_n T_n^{-1} = \bigl\{p^{*K}(u_i) + m_i^K\bigr\}' T_n^{-1}. \tag{A.5}$$

Ignoring the remainder vector m_i^K which is shown to be asymptotically negligible below, the original vector $p^K(w_i)$ is separated into $p^{*K}(u_i)$ and T_n^{-1}. Note that $p^{*K}(u_i)$ is not affected by the weak instruments and can be seen as a new set of regressors.[30]

Now, consider

$$Q = E\big[p^K(w_i)p^K(w_i)'\big]. \tag{A.6}$$

By equations (A.6) and (A.4), it follows

$$T_n' Q T_n = Q^* + E\big[m_i^K p^{*K}(u_i)'\big] + E\big[p^{*K}(u_i)m_i^{K\prime}\big] + E\big[m_i^K m_i^{K\prime}\big], \tag{A.7}$$

where the newly defined $Q^* = E[p^{*K}(u_i)p^{*K}(u_i)']$ is the population second moment matrix with the new regressors. Furthermore, since $\tilde{\Pi}(\cdot) \in \mathcal{C}_1(\mathcal{Z})$ can have nonempty $\mathcal{Z}_0$ as a subset of its domain, we define $Q^{r*} = E[p^{*K}(u_i)p^{*K}(u_i)'|z_i \in \mathcal{Z}^r]$ and $Q^{0*} = E[p^{*K}(u_i)p^{*K}(u_i)'|z_i \in \mathcal{Z}^0]$. Also define the second moment matrix for the first-stage estimation as $Q_1 = E[r^L(z_i)r^L(z_i)']$.

Assumptions B, C, D, and L of the main text serve as sufficient conditions for high-level assumptions stated here. Section B.1.1 in Appendix B in the Online Supplemental Material proves that the latter are implied by the former. For a symmetric matrix B, let $\lambda_{\min}(B)$ and $\lambda_{\max}(B)$ denote the minimum and maximum eigenvalues of B, respectively, and $\det(B)$ the determinant of B.

Assumption B†. (i) $\lambda_{\min}(Q^{r*})$ *is bounded away from zero for all* $K(n)$ *and* $\lambda_{\min}(Q_1)$ *is bounded away from zero for all* $L(n)$; (ii) $\lambda_{\max}(Q)$ *is bounded by a fixed constant, for all* $K(n)$, *and* $\lambda_{\max}(Q_1)$ *bounded by a fixed constant, for all* $L(n)$.

Assumption C†. *There exist* $\beta = (\beta_1, \dots, \beta_{\tilde{K}})$ *and* $\gamma = (\gamma_1, \dots, \gamma_{\tilde{L}})$ *such that* $\sup_{w \in \mathcal{W}} |h_0(w) - p^{\tilde{K}}(w)'\beta| \le C\tilde{K}^{-s/d_x}$ *as* $\tilde{K} \to \infty$ *and* $\sup_{z \in \mathcal{Z}} \|\Pi_0(z) - p^{\tilde{L}}(z)'\gamma\| \le C\tilde{L}^{-s_\pi/d_z}$ *as* $\tilde{L} \to \infty$.

For a generic dimension d and a d-vector μ of nonnegative integers, let $|\mu| = \sum_{l=1}^d \mu_l$. Define the derivative $\partial^\mu g(x) = \partial^{|\mu|} g(x)/\partial x_1^{\mu_1} \partial x_2^{\mu_2} \cdots \partial x_{d_x}^{\mu_{d_x}}$ of order $|\mu|$. For the next assumption, let $\zeta_r^v(\kappa)$ and $\xi_r^v(L)$ satisfy

$$\max_{|\mu| \le r} \sup_{v \in \mathcal{V}} \big\|\partial^\mu p^\kappa(v)\big\| \le \zeta_r^v(\kappa), \qquad \max_{|\mu| \le r} \sup_{z \in \mathcal{Z}} \big\|\partial^\mu r^L(z)\big\| \le \xi_r(L),$$

which impose nonstochastic uniform bounds on the vectors of approximating functions. Let $\Delta_\pi = \sqrt{L/n} + L^{-s_\pi/d_z}$.

Assumption D†. (i) $n^\delta \kappa^{1/2} \zeta_1^v(\kappa)\Delta_\pi \to 0$; (ii) $n^{-\delta}\kappa^{1/2}\zeta_2^v(\kappa) \to 0$; (iii) $\tau_n \to 0$ *and* $\beta' D \beta = O(\lambda_n^2) = O(1)$.

[30] For justification that $p^{*K}(u_i)$ can be regarded as regressors, see Assumption B in Section 5 and Assumption B† below.

Under these assumptions, Lemma A.1 below obtains the orders of magnitudes of eigenvalues of the second moment matrices in term of the weak instrument. In proving this lemma, we frequently apply two useful mathematical lemmas (Lemmas B.1 and B.2 in Appendix B in the Online Supplemental Material) that are stated and proved in Section B.1.2. For any matrix A, let the matrix norm be the Euclidean norm $\|A\| = \sqrt{\operatorname{tr}(A'A)}$.

Lemma A.1. *Suppose Assumptions* A, B†, D†, *and* L *are satisfied. Then* (a) $\lambda_{\max}(Q^{-1}) = O(n^{2\delta})$ *and* (b) $\lambda_{\max}(\hat{Q}^{-1}) = O_p(n^{2\delta})$.

In all proofs, let C denote a generic positive constant that may be different in different use. TR, CS, MK, LIE are the triangular inequality, Cauchy–Schwarz inequality and Markov inequality, the law of iterated expectation, respectively.

Proof of Lemma A.1. Consider (a) first. Let $p_i^* = p^{*K}(u_i)$ and $m_i = m_i^K$ for brevity. Recall (A.7) that $T_n'QT_n = Q^* + E[m_i p_i^{*\prime}] + E[p_i^* m_i'] + E[m_i m_i']$. Then

$$\|T_n'QT_n - Q^*\| \le 2E\|m_i\| \|p_i^*\| + E\|m_i\|^2 \le 2\big(E\|m_i\|^2\big)^{1/2}\big(E\|p_i^*\|^2\big)^{1/2} + E\|m_i\|^2$$

by CS. But $m_i = m_i^K = [0 \vdots \tilde{\Pi}(z_i)(\partial p^\kappa(\tilde{v}_i)' - \partial p^\kappa(v_i)') \vdots (0_{\kappa\times 1})']'$ where $\tilde{v}$ is the intermediate value between x and v. Then, by mean value expanding $\partial p^\kappa(\tilde{v}_i)$ around v_i and $|\tilde{v}_i - v_i| \le |x_i - v_i|$, we have

$$\begin{aligned}\|m_i\|^2 &= \big\|\tilde{\Pi}(z_i)\partial^2 p^\kappa(\bar{v}_i)(\tilde{v}_i - v_i)\big\|^2 \le \big|\tilde{\Pi}(z_i)\big|^2 \zeta_2^v(\kappa)^2 |x_i - v_i|^2 \\ &= n^{-2\delta}\big|\tilde{\Pi}(z_i)\big|^4 \zeta_2^v(\kappa)^2 \le Cn^{-2\delta}\zeta_2^v(\kappa)^2, \qquad \text{(A.8)}\end{aligned}$$

where $\bar{v}$ is the intermediate value between v and $\tilde{v}$, and by Assumption L that $\sup_z |\tilde{\Pi}(z_i)| < \infty$. Therefore,

$$E\|m_i\|^2 \le Cn^{-2\delta}\zeta_2^v(\kappa)^2. \qquad \text{(A.9)}$$

Then

$$E\big[p_i^{*\prime} p_i^*\big] = \operatorname{tr}\big(Q^*\big) \le \operatorname{tr}(I_K)\lambda_{\max}\big(Q^*\big) \le C\cdot K = O_p(\kappa), \qquad \text{(A.10)}$$

where $\lambda_{\max}(Q^*) \le C$ is by the fact that the polynomials are defined on bounded sets and by Assumption L that $\tilde{\Pi}(\cdot) \in \mathcal{C}_1(\mathcal{Z})$. Therefore, by combining (A.9) and (A.10) it follows

$$\|T_n'QT_n - Q^*\| \le O\big(\kappa^{1/2}n^{-\delta}\zeta_2^v(\kappa)\big) + O\big(n^{-2\delta}\zeta_2^v(\kappa)^2\big) = o(1) \qquad \text{(A.11)}$$

by Assumption D†(ii), which shows that all the remainder terms are negligible.

Now, by Lemma B.2, we have

$$\big|\lambda_{\min}\big(T_n'QT_n\big) - \lambda_{\min}\big(Q^*\big)\big| \le \|T_n'QT_n - Q^*\|. \qquad \text{(A.12)}$$

Combine the results (A.11) and (A.12) to have $\lambda_{\min}(T_n'QT_n) = \lambda_{\min}(Q^*) + o(1)$. But note that, with simpler notation $p_1 = \Pr[z \in \mathcal{Z}^r]$ and $p_0 = \Pr[z \in \mathcal{Z}^0]$, we have $Q^* = p_1 Q^{r*} +$

$p_0 Q^{0*}$. Then, by a variant of Lemma B.2 (with the fact that $\lambda_1(-B) = -\lambda_k(B)$ for any symmetric matrix B), it follows that $\lambda_{\min}(Q^*) \geq p_1 \cdot \lambda_{\min}(Q^{r*}) + p_0 \cdot \lambda_{\min}(Q^{0*}) = p_1 \cdot \lambda_{\min}(Q^{r*})$, because $\lambda_{\min}(Q^{0*}) = 0$. Since $p_1 > 0$, it holds that $\lambda_{\min}(Q^*) \geq p_1 \cdot \lambda_{\min}(Q^{r*}) \geq c > 0$ for all $K(n)$ by Assumption B†(i). Therefore,

$$\frac{1}{\lambda_{\min}\big(T_n' Q T_n\big)} = \frac{1}{\lambda_{\min}\big(Q^*\big) + o(1)} \leq \frac{1}{c + o(1)} = O(1). \tag{A.13}$$

Let

$$T_{0n} = \begin{bmatrix} n^\delta & 0 \\ -n^\delta & 1 \end{bmatrix} \otimes I_\kappa, \quad \text{so that} \quad T_n = \begin{bmatrix} 1 & 0_{1\times 2\kappa} \\ 0_{2\kappa\times 1} & T_{0n} \end{bmatrix}.$$

Then, by solving $\left| \begin{smallmatrix} n^\delta - \tilde{\lambda} & 0 \\ -n^\delta & 1-\tilde{\lambda} \end{smallmatrix} \right| = 0$, we have $\tilde{\lambda} = n^\delta$ or 1 for eigenvalues of T_{0n}, and since $\lambda_{\max}(I_\kappa) = 1$, it follows

$$\lambda_{\max}(T_n) = \lambda_{\max}(T_{0n}) = n^\delta. \tag{A.14}$$

Note that $\lambda_{\max}(T_n T_n') \leq n^{2\delta}$ by Lemma B.3. Since (A.13) implies $\lambda_{\max}((T_n' Q T_n)^{-1}) = O(1)$, it follows

$$\lambda_{\max}\big(Q^{-1}\big) = \lambda_{\max}\big(T_n \big(T_n' Q T_n\big)^{-1} T_n'\big) \leq O(1) \lambda_{\max}\big(T_n T_n'\big) = O\big(n^{2\delta}\big)$$

by applying Lemma B.3 again.

The proof of part (b) proceeds similarly as above. Using (A.3), $p^K(\hat{w}_i)' = [1 \vdots p^{K_1}(x_i)' \vdots p^{K_2}(\hat{v}_i)'] = [1 \vdots p^{K_1}(v_i)' + n^{-\delta} \tilde{\Pi}(z_i) \partial p^{K_1}(\tilde{v}_i)' \vdots p^{K_2}(\hat{v}_i)']$ and

$$\begin{aligned} p^K(\hat{w}_i)' T_n &= \big[1 \vdots p^\kappa(v_i)' + n^{-\delta} \tilde{\Pi}(z_i) \partial p^\kappa(\tilde{v}_i)' \vdots p^\kappa(\hat{v}_i)'\big] \cdot T_n \\ &= \big[1 \vdots n^\delta\big(p^\kappa(v_i)' - p^\kappa(\hat{v}_i)'\big) + \tilde{\Pi}(z_i) \partial p^\kappa(\tilde{v}_i)' \vdots p^\kappa(\hat{v}_i)'\big] = p^{*K}(\hat{u}_i)' + \hat{r}_i', \end{aligned}$$

where $p^{*K}(\hat{u}_i)' = [1 \vdots \tilde{\Pi}(z_i) \partial p^\kappa(v_i)' \vdots p^\kappa(\hat{v}_i)']$ with $\hat{u}_i = (z_i, v_i, \hat{v}_i)$ and $\hat{r}_i' = [0 \vdots n^\delta(p^\kappa(v_i)' - p^\kappa(\hat{v}_i)') \vdots (0_{\kappa\times 1})']$. Let $\hat{p}_i^* = p^{*K}(\hat{u}_i)$. For a random matrix X_i, denote $\sum_i X_i/n$ as $E_n X_i$ for simplicity. Then by (A.5),

$$T_n' \hat{Q} T_n = \hat{Q}^* + E_n\big[\hat{r}_i \hat{p}_i^{*\prime}\big] + E_n\big[\hat{p}_i^* \hat{r}_i'\big] + E_n\big[\hat{r}_i \hat{r}_i'\big],$$

and thus

$$\big\| T_n' \hat{Q} T_n - \hat{Q}^* \big\| \leq 2 E_n \|\hat{r}_i\| \big\|\hat{p}_i^*\big\| + E_n \|\hat{r}_i\|^2 \leq 2\big(E_n \|\hat{r}_i\|^2\big)^{1/2} \big(E_n \big\|\hat{p}_i^*\big\|^2\big)^{1/2} + E_n \|\hat{r}_i\|^2.$$

Similarly as (A.10), the bound on $E_n \|\hat{p}_i^*\|^2$ can be derived as

$$E_n\big[\hat{p}_i^{*\prime} \hat{p}_i^*\big] = \operatorname{tr}\big(E_n\big[\hat{p}_i^* \hat{p}_i^{*\prime}\big]\big) \leq \operatorname{tr}(I_K) \lambda_{\max}\big(E_n\big[\hat{p}_i^* \hat{p}_i^{*\prime}\big]\big) \leq O_p(K) = O_p(\kappa), \tag{A.15}$$

where $\lambda_{\max}(E_n[\hat{p}_i^* \hat{p}_i^{*\prime}]) \leq O_p(1)$ is by the fact that the polynomials are defined on bounded sets and by Assumption L that $\tilde{\Pi}(\cdot) \in \mathcal{C}_1(\mathcal{Z})$. For $E_n \|\hat{r}_i\|^2$, note that

$$\|\hat{r}_i\|^2 = \big\| n^\delta\big(p^\kappa(v_i)' - p^\kappa(\hat{v}_i)'\big) \big\|^2 \leq n^{2\delta} \zeta_1^v(\kappa)^2 |v_i - \hat{v}_i|^2 = O_p\big(n^{2\delta} \zeta_1^v(\kappa)^2 \Delta_\pi^2\big)$$

and, therefore, $E_n\|\hat{r}_i\|^2 = O_p(n^{2\delta}\zeta_1^v(\kappa)^2\Delta_\pi^2)$. Combining this with (A.15) and (A.9) yields

$$\|T_n'\hat{Q}T_n - \hat{Q}^*\| \le O_p\big(n^\delta\kappa^{1/2}\zeta_1^v(\kappa)\Delta_\pi\big) + O_p\big(n^{2\delta}\zeta_1^v(\kappa)^2\Delta_\pi^2\big) = o_p(1)$$

by Assumption D†(i). Also, by Lemma B.2, we have $|\lambda_{\min}(T_n'\hat{Q}T_n) - \lambda_{\min}(Q^*)| \le \|T_n'\hat{Q}T_n - Q^*\|$. Combining the two results yields $\lambda_{\min}(T_n'\hat{Q}T_n) = \lambda_{\min}(Q^*) + o_p(1)$. Similar as before

$$\frac{1}{\lambda_{\min}(T_n'\hat{Q}T_n)} = \frac{1}{\lambda_{\min}(Q^*) + o_p(1)} \le \frac{1}{c + o_p(1)} = O_p(1). \tag{A.16}$$

Therefore, we have $\lambda_{\max}(\hat{Q}^{-1}) = \lambda_{\max}(T_n(T_n'\hat{Q}T_n)^{-1}T_n') \le O_p(1)\lambda_{\max}(T_nT_n') = O_p(n^{2\delta})$. □

A.2 *Proofs of rate of convergence (Section 5)*

We first derive the rate of convergence of the unpenalized series estimator $\hat{h}(\cdot)$ defined as $\hat{h}(w) = p^K(w)'\hat{\beta}$ where $\hat{\beta} = (\hat{P}'\hat{P})^{-1}\hat{P}'y$. Then we prove Theorem 5.1 with the penalized estimator $\hat{h}_\tau(\cdot)$ defined in Section 4. Next to the proof, we provide a theorem for the rate of $\hat{g}_\tau(\cdot)$ and prove it.

Lemma A.2. *Suppose Assumptions A–D and L are satisfied. Then*

$$\|\hat{h} - h_0\|_{L_2} = O_p\big(n^\delta\big(\sqrt{K/n} + K^{-s/d_x} + \sqrt{L/n} + L^{-s_\pi/d_z}\big)\big).$$

Proof of Lemma A.2. Let $\beta = (\beta_1, \dots, \beta_K)$. By TR of L_2 norm (first inequality),

$$\begin{aligned}\|\hat{h} - h_0\|_{L_2} &= \left\{\int\big[\hat{h}(w) - h_0(w)\big]^2 dF(w)\right\}^{1/2}\\ &\le \left\{\int\big[p^K(w)'(\hat{\beta} - \beta)\big]^2 dF(w)\right\}^{1/2} + \left\{\int\big[p^K(w)'\beta - h_0(w)\big]^2 dF(w)\right\}^{1/2}\\ &= \big\{(\hat{\beta} - \beta)'Ep^K(w)p^K(w)'(\hat{\beta} - \beta)\big\}^{1/2} + O\big(K^{-s/d_x}\big)\\ &\le C\|\hat{\beta} - \beta\| + O\big(K^{-s/d_x}\big)\end{aligned}$$

by Assumption B†(ii) and using Lemma B.3 (last equation). As $\hat{\beta} - \beta = (\hat{P}'\hat{P})^{-1}\hat{P}'(y - \hat{P}\beta)$, it follows that

$$\begin{aligned}\|\hat{\beta} - \beta\|^2 &= (y - \hat{P}\beta)'\hat{P}\big(\hat{P}'\hat{P}\big)^{-1}\big(\hat{P}'\hat{P}\big)^{-1}\hat{P}'(y - \hat{P}\beta)\\ &= (y - \hat{P}\beta)'\hat{P}\hat{Q}^{-1/2}\hat{Q}^{-1}\hat{Q}^{-1/2}\hat{P}'(y - \hat{P}\beta)/n^2\\ &\le O_p\big(n^{2\delta}\big)(y - \hat{P}\beta)'\hat{P}\big(\hat{P}'\hat{P}\big)^{-1}\hat{P}'(y - \hat{P}\beta)/n\end{aligned}$$

by Lemma B.3 and Lemma A.1(b) (last inequality).

Let $h = (h(w_1), \dots, h(w_n))'$ and $\tilde{h} = (h(\hat{w}_1), \dots, h(\hat{w}_n))'$. Also let $\eta_i = y_i - h_0(w_i)$ and $\eta = (\eta_1, \dots, \eta_n)'$. Let $W = (w_1, \dots, w_n)'$, then $E[y_i|W] = h_0(w_i)$ which implies $E[\eta_i|W] =$

0. Also similar to the proof of Lemma A1 in NPV (p. 594), by Assumption A, we have $E[\eta_i^2|W]$ being bounded and $E[\eta_i\eta_j|W]=0$ for $i\neq j$, where the expectation is taken for y. Then, given that $y-\hat{P}\beta=(y-h)+(h-\tilde{h})+(\tilde{h}-\hat{P}\beta)$, we have, by TR,

$$\begin{aligned}\|\hat{\beta}-\beta\| &= O_p\big(n^{\delta}\big)\big\|\hat{Q}^{-1/2}\hat{P}'(y-\hat{P}\beta)/n\big\| \\ &\le O_p\big(n^{\delta}\big)\big\{\big\|\hat{Q}^{-1/2}\hat{P}'\eta/n\big\|+\big\|\hat{Q}^{-1/2}\hat{P}'(h-\tilde{h})/n\big\| \\ &\quad +\big\|\hat{Q}^{-1/2}\hat{P}'(\tilde{h}-\hat{P}\beta)/n\big\|\big\}. \end{aligned} \tag{A.17}$$

For the first term of equation (A.17), consider

$$\begin{aligned}E\big[\big\|\big(PT_n-P^*\big)'\eta/n\big\|^2|W\big] &= E\big[\big\|M'\eta/n\big\|^2|W\big]\le C\frac{1}{n^2}\sum_i\|m_i\|^2 \\ &= O_p\big(n^{-2\delta-1}\zeta_2^v(\kappa)^2\big)=o_p(1)\end{aligned}$$

by (A.8) and $o_p(1)$ is implied by Assumption D†(ii). Therefore, by MK,

$$\big\|\big(PT_n-P^*\big)'\eta/n\big\|=o_p(1). \tag{A.18}$$

Also,

$$\begin{aligned}E\big[\big\|(\hat{P}T_n-PT_n)'\eta/n\big\|^2|W\big] &\le C\frac{1}{n^2}\sum_i\big\|(\hat{p}_i-p_i)'T_n\big\|^2\le C\frac{1}{n^2}\sum_i\lambda_{\max}(T_n)^2\|\hat{p}_i-p_i\|^2 \\ &\le \frac{1}{n}O\big(n^{2\delta}\big)O_p\big(\zeta_1^v(\kappa)^2\Delta_\pi^2\big)=O_p\big(n^{2\delta}\zeta_1^v(\kappa)^2\Delta_\pi^2/n\big)\end{aligned} \tag{A.19}$$

by (A.14) and

$$\begin{aligned}\|\hat{p}_i-p_i\|^2 &= \big\|p^\kappa(x_i)-p^\kappa(x_i)\big\|^2+\big\|\partial p^\kappa(\bar{v}_i)(\hat{v}_i-v_i)\big\|^2 \\ &\le C\zeta_1^v(\kappa)^2\frac{1}{n}\sum_i|\hat{v}_i-v_i|^2\le O_p\big(\zeta_1^v(\kappa)^2\Delta_\pi^2\big).\end{aligned} \tag{A.20}$$

Therefore,

$$\big\|(\hat{P}T_n-PT_n)'\eta/n\big\|=o_p(1) \tag{A.21}$$

by Assumption D†(i) and MK. Also

$$\begin{aligned}E\big\|P^{*\prime}\eta/n\big\|^2 &= E\big[E\big[\big\|P^{*\prime}\eta/n\big\|^2|W\big]\big]=E\Big[\sum_i p_i^{*\prime}p_i^*E\big[\eta_i^2|W\big]/n^2\Big] \\ &\le C\frac{1}{n^2}\sum_i E\big[p_i^{*\prime}p_i^*\big]=C\operatorname{tr}\big(Q^*\big)/n=O(\kappa/n)\end{aligned}$$

by Assumption A (first inequality) and equation (A.10) (last equation). By MK, this implies

$$\big\|P^{*\prime}\eta/n\big\|\le O_p(\sqrt{\kappa/n}). \tag{A.22}$$

Hence by TR with (A.18), (A.21), and (A.22),

$$\|T_n'\hat{P}'\eta/n\| \le \|(\hat{P}T_n - PT_n)'\eta/n\| + \|(PT_n - P^*)'\eta/n\| + \|P^{*\prime}\eta/n\| \le O_p(\sqrt{\kappa/n}).$$

Therefore, the first term of (A.17) becomes

$$\|\hat{Q}^{-1/2}\hat{P}'\eta/n\|^2 = \left(\frac{\eta'\hat{P}T_n}{n}\right)(T_n'\hat{Q}T_n)^{-1}\left(\frac{T_n'\hat{P}'\eta}{n}\right) \le O_p(1)\|T_n'\hat{P}'\eta/n\|^2 = O_p(\kappa/n) \tag{A.23}$$

by Lemma B.3 and (A.16).

Because $I - \hat{P}(\hat{P}'\hat{P})^{-1}\hat{P}'$ is a projection matrix, hence is p.s.d., the second term of (A.17) becomes

$$\begin{aligned}\|\hat{Q}^{-1/2}\hat{P}'(h-\tilde{h})/n\|^2 &= (h-\tilde{h})'\hat{P}(\hat{P}'\hat{P})^{-1}\hat{P}'(h-\tilde{h})/n \le (h-\tilde{h})'(h-\tilde{h})/n \\ &= \sum_i \big(h(w_i) - h(\hat{w}_i)\big)^2/n = \sum_i \big(\lambda(v_i) - \lambda(\hat{v}_i)\big)^2/n \\ &\le C\sum_i |v_i - \hat{v}_i|^2/n = \sum_i \big|\Pi_n(z_i) - \hat{\Pi}(z_i)\big|^2/n \\ &= O_p(\Delta_\pi^2) \end{aligned} \tag{A.24}$$

by Assumption C (Lipschitz continuity of $\lambda(v)$) (last inequality). Similarly, the last term is

$$\begin{aligned}\|\hat{Q}^{-1/2}\hat{P}'(\tilde{h} - \hat{P}\beta)/n\|^2 &= (\tilde{h} - \hat{P}\beta)'\hat{P}(\hat{P}'\hat{P})^{-1}\hat{P}'(\tilde{h} - \hat{P}\beta)/n \\ &\le (\tilde{h} - \hat{P}\beta)'(\tilde{h} - \hat{P}\beta)/n \\ &= \sum_i \big(h(\hat{w}_i) - p^K(\hat{w}_i)'\beta\big)^2/n = O_p(K^{-2s/d_x}) \end{aligned} \tag{A.25}$$

by Assumption C†. Therefore, by combining (A.23), (A.24), and (A.25),

$$\|\hat{\beta} - \beta\| \le O_p(n^\delta)\big[O_p(\sqrt{\kappa/n}) + O_p(\Delta_\pi) + O_p(K^{-s/d_x})\big].$$

Consequently, since $\kappa \asymp K$,

$$\|\hat{h} - h_0\|_{L_2} \le O_p(n^\delta)\big[O_p(\sqrt{K/n}) + O_p(K^{-s/d_x}) + O_p(\Delta_\pi)\big] + O(K^{-s/d_x})$$

and we have the conclusion of the lemma. □

Proof of Theorem 5.1. Let $C_n = n\tau_n D$ for simplicity, and consider

$$\|\hat{\beta}_\tau - \beta\| \le \|(\hat{P}'\hat{P} + C_n)^{-1}\hat{P}'(y - \hat{P}\beta)\| + \|(\hat{P}'\hat{P} + C_n)^{-1}\hat{P}'\hat{P}\beta - \beta\|.$$

Recalling that $\hat{Q}_\tau = (\hat{P}'\hat{P} + C_n)/n$, we have

$$\begin{aligned}&\|(\hat{P}'\hat{P} + C_n)^{-1}\hat{P}'(y - \hat{P}\beta)\|^2 \\ &\quad = (y - \hat{P}\beta)'\hat{P}(\hat{P}'\hat{P} + C_n)^{-1}(\hat{P}'\hat{P} + C_n)^{-1}\hat{P}'(y - \hat{P}\beta)\end{aligned}$$

$$
\begin{aligned}
&= (y - \hat{P}\beta)'\hat{P}\big(\hat{P}'\hat{P} + C_n\big)^{-1/2}\big(\hat{P}'\hat{P} + C_n\big)^{-1}\big(\hat{P}'\hat{P} + C_n\big)^{-1/2}\hat{P}'(y - \hat{P}\beta) \\
&= (y - \hat{P}\beta)'\hat{P}\hat{Q}_\tau^{-1/2}\hat{Q}_\tau^{-1}\hat{Q}_\tau^{-1/2}\hat{P}'(y - \hat{P}\beta)/n^2 \\
&\le \lambda_{\max}\big(\hat{Q}_\tau^{-1}\big)\big\|\hat{Q}_\tau^{-1/2}\hat{P}'(y - \hat{P}\beta)/n\big\|^2.
\end{aligned}
$$

First, note that

$$
\begin{aligned}
\lambda_{\max}\big(\hat{Q}_\tau^{-1}\big) &= \frac{1}{\lambda_{\min}(\hat{Q} + \tau_n I)} \le \frac{1}{\lambda_{\min}(\hat{Q}) + \lambda_{\min}(\tau_n I)} = \frac{1}{\lambda_{\min}(\hat{Q}) + \tau_n} \\
&\le \min\left\{\frac{1}{\lambda_{\min}(\hat{Q})}, \frac{1}{\tau_n}\right\} = \min\big\{O_p\big(n^{2\delta}\big), \tau_n^{-1}\big\},
\end{aligned}
\tag{A.26}
$$

where the last equation is by Lemma A.1(b). Also, note that $c'\hat{P}'\hat{Q}_\tau^{-1}\hat{P}c \le c'\hat{P}'\hat{Q}^{-1}\hat{P}c$ for any vector c, since $(\hat{Q}^{-1} - \hat{Q}_\tau^{-1})$ is p.s.d. Therefore, by (A.23), (A.24), and (A.25) in Lemma A.2, we have

$$
\begin{aligned}
\big\|\hat{Q}_\tau^{-1/2}\hat{P}'(y - \hat{P}\beta)/n\big\| &\le \big\|\hat{Q}_\tau^{-1/2}\hat{P}'(y - h)/n\big\| + \big\|\hat{Q}_\tau^{-1/2}\hat{P}'(h - \hat{P}\beta)/n\big\| \\
&= O_p(\sqrt{K/n}) + O_p\big(K^{-s/d_x} + \Delta_\pi\big).
\end{aligned}
\tag{A.27}
$$

Now, consider

$$
\begin{aligned}
\big\|\big(\hat{P}'\hat{P} + C_n\big)^{-1}\hat{P}'\hat{P}\beta - \beta\big\|^2 &= \big\|\big(\hat{P}'\hat{P} + C_n\big)^{-1}\big\{\hat{P}'\hat{P} - \big(\hat{P}'\hat{P} + C_n\big)\big\}\beta\big\|^2 \\
&\le \lambda_{\max}\big(\hat{Q}_\tau^{-1}\big)\big\|\hat{Q}_\tau^{-1/2}\{-\tau_n D\}\beta\big\|^2 \\
&= \lambda_{\max}\big(\hat{Q}_\tau^{-1}\big)\big\|\tau_n\hat{Q}_\tau^{-1/2}D\beta\big\|^2,
\end{aligned}
$$

but we have $\|\tau_n\hat{Q}_\tau^{-1/2}D\beta\| = O(\tau_n R_n \lambda_n)$ by Assumption D that $\sqrt{\beta' D\beta} = O(\lambda_n)$. Consequently, analogous to the proof of Lemma A.2, and by letting $R_n = \min\{n^\delta, \tau_n^{-1/2}\}$,

$$
\|\hat{h}_\tau - h_0\|_{L_2} = O_p\big(R_n\big(\sqrt{K/n} + K^{-s/d_x} + \tau_n R_n \lambda_n + \sqrt{L/n} + L^{-s_\pi/d_z}\big)\big).
$$

This proves the first part of the theorem. The conclusion of the second part follows from

$$
\begin{aligned}
\big\|\hat{h}(w) - h_0(w)\big\|_\infty &\le \big\|p^K(w)'\beta - h_0(w)\big\|_\infty + \big\|p^K(w)'(\hat{\beta}_\tau - \beta)\big\|_\infty \\
&\le O\big(K^{-s/d_x}\big) + \zeta_0^v(K)\|\hat{\beta}_\tau - \beta\|. \qquad \square
\end{aligned}
$$

Theorem 5.1 leads to the following theorem, which focuses on the rate of convergence of the structural estimator $\hat{g}_\tau(\cdot)$.

Theorem A.1. *Suppose Assumptions A–D and L are satisfied. Let* $R_n = \min\{n^\delta, \tau_n^{-1/2}\}$. *For* $\hat{\Delta}(x) = \hat{g}_\tau(x) - g_0(x)$,

$$
\left\|\hat{\Delta}(x) - \int \hat{\Delta}(x)\, dF_w\right\|_{L_2} = O_p\big(R_n\big(\sqrt{K/n} + K^{-\frac{s}{d_x}} + \tau_n R_n \lambda_n + \sqrt{L/n} + L^{-\frac{s_\pi}{d_z}}\big)\big).
$$

Also, if $\hat{g}_\tau(x) = \hat{h}_\tau(x, \bar{v}) - \bar{\lambda}$ *and* $\bar{\lambda} = \lambda_0(\bar{v})$, *then*

$$\sup_{x\in\mathcal{X}} \bigl|\hat{g}_\tau(x) - g_0(x)\bigr| = O_p\bigl(R_n\sqrt{K}\bigl(\sqrt{K/n} + K^{-\frac{s}{d_x}} + \tau_n R_n \lambda_n + \sqrt{L/n} + L^{-\frac{s_\pi}{d_z}}\bigr)\bigr).$$

The balanced rate results for $\hat{g}_\tau(\cdot)$ and the related analyses can be followed analogously, and we omit them here. The convergence rate is net of the constant term which is not identified. We can further assume $E[\varepsilon] = 0$ to identify the constant.

Proof of Theorem A.1. The proof follows directly from the proofs of Theorems 4.2 and 4.3 of NPV. As for notation, we use v instead of u of NPV and the other notation are identical. □

Proof of Theorem 5.3. By (2.2), we have

$$E[y|x, z] = h_0(x, v) = g_0(x) + \lambda_0(v) = g_0(v) + \Pi_n(z)'\nabla g_0(\tilde{x}) + \lambda_0(v),$$

where the second equation is by the elementwise mean value expansion around x with $\tilde{x}$ between x and v, and ∇g_0 is the $d_x \times 1$ gradient of g_0. Since $\Pi_n(z)$ is assumed to be known, $\tilde{x}$ is also known as a function of x and z. Consider a nonparametric additive regression with $k_0(v) = g_0(v) + \lambda_0(v)$:

$$y = \Pi_n(z)'\nabla g_0(\tilde{x}) + k_0(v) + \eta. \tag{A.28}$$

By Horowitz et al. (2006, p. 272) with different normalization, there exists an estimator for $g_0(\tilde{x})$ in (A.28) that achieves the same asymptotic minimax risk as for a nonparametric regression model where k_0 is known. Therefore, we alternatively consider a model with $k_0(x)$ being known. To this end, let $\tilde{y} = y - k_0(v)$. Consider a Gaussian model

$$\tilde{y} = \Pi_n(z)'\nabla g_0(\tilde{x}) + \eta, \tag{A.29}$$

where $\eta|x, z \sim N(0, \sigma^2(x, z))$. By Chen and Reiss (2011, Lemma 1), the risk for the original model (A.28) with known $k_0(x)$ is at least as large as the risk for this Gaussian model. Therefore, we calculate the lower bound for the Gaussian model (A.29). Let P_0 be the joint distribution of (y_i, x_i, z_i) in this model.

Since $h_0(x, v) = g_0(x) + \lambda_0(v)$, it suffices to prove the minimax rate for g_0. As before, C is a generic positive constant. Consider the wavelets as defined in Daubechies (1992), Chen (2007), and Chen and Christensen (2018b). We define a family of submodels where the function g_0 is perturbed by using elements of the wavelet space W_j where j is a function of n. For given j, W_j consists of 2^j functions $\{\psi_{j,k}\}_{0\leq k\leq 2^j-1}$, such that $\{\psi_{j,k}\}_{r\leq k\leq 2^j-N-1}$ are interior univariate wavelets for which $\psi_{j,k}(\cdot) = 2^{j/2}\psi(2^j(\cdot) - k)$. Then $\tilde{\psi}_{j,m,G}$ is an orthonormal tensor-product of d_x interior univariate wavelets at resolution j with $G = (w_\psi)^{d_x}$; see, for example, Chen and Christensen (2018b, Appendix E) for details of these definitions. Since the support of each univariate wavelet is an interval of length $2^{-j}(2r - 1)$, we may choose a set $M \subset \{r, \ldots, 2^j - N - 1\}^{d_x}$ of interior wavelets with $\#(M) \asymp 2^{jd_x}$ such that the supports of $\psi_{j,m}$ and $\psi_{j,m'}$ do not overlap for $m, m' \in M$

and $m \neq m'$. Let $g_0 \in B(s, L)$ where $B(s, L)$ is a Sobolev ball of smoothness $s > 0$ with the norm $\|\cdot\|_{B^s_{2,2}}$ and radius $0 < L < \infty$ (Triebel (2006, Chapter 4)). For each $m \in M$, define $\theta = \{\theta_m\}_{m\in M}$ where $\theta_m \in \{0, 1\}$, and let

$$g_\theta = g_0 + C_0 2^{-j(s+d_x/2)} \sum_{m\in M} \theta_m \tilde{\psi}_{j,m,G}, \tag{A.30}$$

with $C_0 > 0$ subsequently chosen. We first prove the closedness:

$$\begin{aligned}\|g_\theta\|_{B^s_{2,2}} &\le L/2 + \left\| C_0 2^{-j(s+d_x/2)} \sum_{m\in M} \theta_m \tilde{\psi}_{j,m,G} \right\|_{B^s_{2,2}} \\ &\le L/2 + C \cdot C_0 2^{-j(s+d_x/2)} \left(\sum_{m\in M} \theta_m^2 2^{2js} \right)^{1/2} \le L/2 + C \cdot C_0,\end{aligned}$$

and thus $g_\theta \in B(s, L)$ for sufficiently small C_0. We now prove the well-separatedness:

$$\begin{aligned}\left\|\partial^\mu g_\theta - \partial^\mu g_{\theta'}\right\|_{L^2} &= C_0 2^{-j(s-|\mu|+d_x/2)} \left(\sum_{m\in M} \left(\theta_m - \theta'_m\right) \left\| \left\{2^{j/2} \psi^{(|\mu|)}\left(2^j(\cdot) - m\right)\right\}^{d_x} \right\|^2_{L^2} \right)^{1/2} \\ &\ge C \cdot C_0 2^{-j(s-|\mu|+d_x/2)} \sqrt{\rho\left(\theta, \theta'\right)},\end{aligned}$$

where $\|\{2^{j/2}\psi^{(|\mu|)}(2^j(\cdot) - m)\}^{d_x}\|^2_{L^2} \asymp 1$ since $\psi_{j,m} \in C^\gamma$ with $\gamma > |\mu|$ has compact support and the density of x_i is bounded away from zero and ∞, and $\rho(\theta, \theta')$ is the Hamming distance. By Tsybakov (2009, Chapter 2.6), choose a subset $\theta^{(0)}, \theta^{(1)}, \dots, \theta^{(M^*)}$ such that θ^0 is a zero vector, $\rho(\theta^{(a)}, \theta^{(b)}) \ge \#(M)/8 \ge C2^{jd_x}$ (for large j that $\#(M) \ge 8$) for all $0 \le a < b \le M^*$ and $M^* \ge 2^{\#(M)/8}$. For each $m \in \{0, 1, \dots, m^*\}$ let $h_m = h_{\theta^{(m)}}$. Consequently, we have

$$\left\|\partial^\mu g_m - \partial^\mu g_{m'}\right\|_{L^2} \ge C \cdot C_0 2^{-j(s-|\mu|)}$$

for each $0 \le m < m' \le M^*$.

Next, for $0 \le m \le M^*$, let P_m be the joint distribution of $\{(y_i, x_i, z_i)\}^n_{i=1}$ with $\tilde{y}_i = \Pi_n(z)'\nabla g_0(\tilde{x}) + \eta_i$ where $\eta_i | x_i, z_i \sim N(0, \sigma^2(x_i, z_i))$. Given (A.30), the Kullbeck–Leibler divergence $K(P_m, P_0)$ satisfies

$$\begin{aligned}K(P_m, P_0) &\le \frac{1}{2} \sum_{i=1}^n \left(C_0 2^{-j(s+d_x/2)}\right)^2 E\left[\frac{\left(\sum_{k\in M} \theta_k^{(m)} \Pi_n(z_i)' \nabla \tilde{\psi}_{j,k,G}(\tilde{x}_i) \right)^2}{\sigma^2(x_i, z_i)} \right] \\ &\le \frac{1}{2} n \left(C_0 2^{-j(s+d_x/2)}\right)^2 E\left[\frac{n^{-2\delta} \left(\sum_{k\in M} \theta_k^{(m)} \tilde{\Pi}(z_i)' \nabla \tilde{\psi}_{j,k,G}(\tilde{x}_i) \right)^2}{\underline{\sigma}^2} \right]\end{aligned}$$

$$\leq C \cdot C_0^2 n^{1-2\delta} 2^{-2j(s+d_x/2)} \|\tilde{\Pi}_k(z_i)\|_{L^2}^2 \|\partial^1 \psi_{j,k}\|_{L^2}^2 \leq \tilde{C} \cdot C_0^2 n^{1-2\delta} 2^{-2j(s+d_x/2)}$$

by the assumption that $E[\eta_i^2|x_i, z_i] \geq \underline{\sigma}^2 > 0$ uniformly for (x_i, z_i) (the second inequality), $\sum_{k\in M}(\theta_k^{(m)})^2 \leq \#(M) \asymp 2^{jd_x}$ (the third inequality), and the assumption that Π_n (and thus each element $\tilde{\Pi}_k$ of $\tilde{\Pi}$) is bounded (the last inequality).

Now, we choose $2^j \asymp n^{\frac{1}{d_x+2s}-\delta/s}$. Then $K(P_m, P_0) \leq \tilde{C} \cdot C_0^2 n^{\frac{d_x\delta}{s}}$ and $\log M^* \geq C2^{jd_x} \asymp n^{\frac{d_x}{d_x+2s}-\frac{d_x\delta}{s}}$. But since $\frac{s}{2(1+2s)} \geq \delta$, we have $\frac{d_x}{d_x+2s} - \frac{d_x\delta}{s} - \frac{d_x\delta}{s} \geq 0$ and, therefore,

$$K(P_m, P_0) \leq a \log M^*$$

for $0 < a < 1/8$ by choosing C_0 sufficiently small. Therefore, by Theorem 2.5 of Tsybakov (2009), we obtain the lower bound for g_0:

$$\inf_{\hat{g}} \sup_{g\in B(s,L)} \Pr{}_g\big(\|\hat{g} - g\|_{L^2} \geq Cn^{-\frac{s}{d_x+2s}+\delta}\big) \geq C' > 0,$$

which proves the lower bound for h_0 of the theorem. □

Proof of Theorem 5.4. We focus on $\|\hat{\beta}_{\hat{\tau}} - \beta\|$, which suffices to derive the adaptive rate for $\|\hat{h}_{\hat{\tau}} - h_0\|$. Note that

$$\|\hat{\beta}_{\hat{\tau}} - \beta\| \leq \|\hat{\beta}_{\tau^\dagger} - \beta\| + \|\hat{\beta}_{\hat{\tau}} - \hat{\beta}_{\tau^\dagger}\| \leq O_p\big(\Delta_h^\dagger\big) + O_p\big(\sqrt{K/n} + K^{-s/d_x} + \Delta_\pi\big), \quad \text{(A.31)}$$

where $\|\hat{\beta}_{\tau^\dagger} - \beta\| \leq O_p(\Delta_h^\dagger)$ is by Theorem 5.1 with $\Delta_h^\dagger = R_n^\dagger(\sqrt{K/n} + K^{-s/d_x} + \tau^\dagger R_n\lambda_n + \sqrt{L/n} + L^{-s_\pi/d_z})$. We show $\|\hat{\beta}_{\hat{\tau}} - \hat{\beta}_{\tau^\dagger}\| \leq O_p(\sqrt{K/n} + K^{-s/d_x} + \Delta_\pi)$ below. Therefore, by (A.31),

$$\|\hat{\beta}_{\hat{\tau}} - \beta\| = O_p\big(\Delta_h^\dagger\big) + O_p\big(\sqrt{K/n} + K^{-s/d_x} + \Delta_\pi\big) = O_p\big(\Delta_h^\dagger\big),$$

which proves the theorem. Note that with the choice $\tau^\dagger$, the penalty bias $\|\hat{Q}_{\tau^\dagger}^{-1}\hat{Q}\beta - \beta\| = O_p(R_n^\dagger \cdot \tau^\dagger R_n^\dagger \lambda_n)$ is dominated by the variance term in $\|\hat{\beta}_\tau - \hat{Q}_\tau^{-1}\hat{Q}\beta\| = O_p(R_n^\dagger(\sqrt{K/n} + K^{-s/d_x} + \Delta_\pi))$ (this equation is by (A.34) below) by the definition of $\mathcal{T}_0$.

The remaining part proves $\|\hat{\beta}_{\hat{\tau}} - \hat{\beta}_{\tau^\dagger}\| \leq O_p(\sqrt{K/n} + K^{-s/d_x} + \Delta_\pi)$. To show this, we first show that $\|\hat{\beta}_{\hat{\tau}} - \hat{\beta}_{\tau^\dagger}\| \leq O_p\{\|(I - \hat{Q}_{\tau_j}^{-1}\hat{Q})\hat{\beta}_{\tau_j})\| + \|(I - \hat{Q}_{\tau_k}^{-1}\hat{Q})\hat{\beta}_{\tau_k})\|\}$, for which we need to show that $\tau^\dagger \in \hat{\mathcal{T}}$ with probability approaching 1, so that the result follows by the definition of $\hat{\mathcal{T}}$. First, choose any $\tau_j \in \mathcal{T}_0 \subset \hat{\mathcal{T}}$ with $\tau_j \geq \tau^\dagger$. Then

$$\begin{aligned}\|\hat{\beta}_{\tau_j} - \hat{\beta}_{\tau^\dagger}\| &\leq \|\hat{\beta}_{\tau_j} - \beta\| + \|\hat{\beta}_{\tau^\dagger} - \beta\| \\ &\leq \big\|\hat{\beta}_{\tau_j} - \hat{Q}_{\tau_j}^{-1}\hat{Q}\beta\big\| + \big\|\hat{Q}_{\tau_j}^{-1}\hat{Q}\beta - \beta\big\| + \big\|\hat{\beta}_{\tau^\dagger} - \hat{Q}_{\tau^\dagger}^{-1}\hat{Q}\beta\big\| + \big\|\hat{Q}_{\tau^\dagger}^{-1}\hat{Q}\beta - \beta\big\| \\ &\leq 2\big\|\hat{\beta}_{\tau_j} - \hat{Q}_{\tau_j}^{-1}\hat{Q}\beta\big\| + 2\big\|\hat{\beta}_{\tau^\dagger} - \hat{Q}_{\tau^\dagger}^{-1}\hat{Q}\beta\big\|. \qquad \text{(A.32)}\end{aligned}$$

For each $\tau \in \{\tau_j, \tau^\dagger\}$,

$$\big\|\hat{\beta}_\tau - \hat{Q}_\tau^{-1}\hat{Q}\beta\big\| \leq \big\|\hat{\beta}_\tau - \hat{Q}_\tau^{-1}\hat{Q}\hat{\beta}_\tau\big\| + \big\|\hat{Q}_\tau^{-1}\hat{Q}(\hat{\beta}_\tau - \beta)\big\|.$$

But, for small $0 < \eta < 1/2$,

$$\begin{aligned}\|\hat{Q}_\tau^{-1}\hat{Q}(\hat{\beta}_\tau - \beta)\| &\leq \eta\|(I - \hat{Q}_\tau^{-1}\hat{Q})\hat{\beta}_\tau\| \\ &\leq \eta\{\|\hat{\beta}_\tau - \hat{Q}_\tau^{-1}\hat{Q}\beta\| + \|\hat{Q}_\tau^{-1}\hat{Q}(\hat{\beta}_\tau - \beta)\|\},\end{aligned} \tag{A.33}$$

where the first inequality is by the convergence of $\|\hat{\beta}_\tau - \beta\|$ as in the proof of Theorem 5.1 and the fact that $I - \hat{Q}_\tau^{-1}\hat{Q}$ is p.d. for $\tau \in \mathcal{T}$, and thus $\|\hat{Q}_\tau^{-1}\hat{Q}(\hat{\beta}_\tau - \beta)\| < \|\hat{\beta}_\tau - \beta\|$. Then, (A.33) can be rearranged to have

$$\|\hat{Q}_\tau^{-1}\hat{Q}(\hat{\beta}_\tau - \beta)\| \leq \frac{\eta}{1-\eta}\|\hat{\beta}_\tau - \hat{Q}_\tau^{-1}\hat{Q}\beta\|.$$

Therefore,

$$\|\hat{\beta}_\tau - \hat{Q}_\tau^{-1}\hat{Q}\beta\| \leq \|\hat{\beta}_\tau - \hat{Q}_\tau^{-1}\hat{Q}\hat{\beta}_\tau\| + \frac{\eta}{1-\eta}\|\hat{\beta}_\tau - \hat{Q}_\tau^{-1}\hat{Q}\beta\|$$

or

$$\|\hat{\beta}_\tau - \hat{Q}_\tau^{-1}\hat{Q}\beta\| \leq \frac{1-\eta}{1-2\eta}\|\hat{\beta}_\tau - \hat{Q}_\tau^{-1}\hat{Q}\hat{\beta}_\tau\|.$$

Consequently,

$$\|\hat{\beta}_{\tau_j} - \hat{\beta}_{\tau^\dagger}\| \leq 2C_\eta\{\|(I - \hat{Q}_{\tau_j}^{-1}\hat{Q})\hat{\beta}_{\tau_j})\| + \|(I - \hat{Q}_{\tau_k}^{-1}\hat{Q})\hat{\beta}_{\tau_k})\|\},$$

where $C_\eta = \frac{1-\eta}{1-2\eta} = 1 + o(1)$ and, therefore, $\tau^\dagger \in \hat{\mathcal{T}}$ with probability approaching 1. Lastly, note that

$$\|\hat{\beta}_{\tau_j} - \hat{Q}_{\tau_j}^{-1}\hat{Q}\beta\| \leq O_p(\sqrt{K/n} + K^{-s/d_x} + \Delta_\pi) \tag{A.34}$$

by (A.27). Therefore, using (A.32), we have

$$\|\hat{\beta}_{\tau_j} - \hat{\beta}_{\tau^\dagger}\| \leq O_p(\sqrt{K/n} + K^{-s/d_x} + \Delta_\pi),$$

which completes the proof of the theorem. □

References

Amemiya, T. (1977), "The maximum likelihood and the nonlinear three-stage least squares estimator in the general nonlinear simultaneous equation model." *Econometrica*, 45, 955–968. [176]

Andrews, D. W. K. and X. Cheng (2012), "Estimation and inference with weak, semi-strong, and strong identification." *Econometrica*, 80, 2153–2211. [162]

Andrews, D. W. K. and P. Guggenberger (forthcoming), "Identification- and singularity-robust inference for moment condition models." *Quantitative Economics*. [162]

Andrews, D. W. K. and J. H. Stock (2007), "Inference with weak instruments." In *Advances in Econometrics: Proceedings of the Ninth World Congress of the Econometric Society*. [161, 162]

Andrews, D. W. K. and Y.-J. Whang (1990), "Additive interactive regression models: Circumvention of the curse of dimensionality." *Econometric Theory*, 6, 466–479. [173]

Andrews, I. and A. Mikusheva (2016a), "Conditional inference with a functional nuisance parameter." *Econometrica*, 84, 1571–1612. [162]

Andrews, I. and A. Mikusheva (2016b), "A geometric approach to nonlinear econometric models." *Econometrica*, 84, 1249–1264. [162]

Angrist, J. D. and A. B. Keueger (1991), "Does compulsory school attendance affect schooling and earnings?" *The Quarterly Journal of Economics*, 106, 979–1014. [162]

Angrist, J. D. and V. Lavy (1999), "Using Maimonides' rule to estimate the effect of class size on scholastic achievement." *The Quarterly Journal of Economics*, 114, 533–575. [182, 184]

Arlot, S. and A. Celisse (2010), "A survey of cross-validation procedures for model selection." *Statistics Surveys*, 4, 40–79. [175]

Belloni, A., V. Chernozhukov, D. Chetverikov, and K. Kato (2015), "Some new asymptotic theory for least squares series: Pointwise and uniform results." *Journal of Econometrics*, 186, 345–366. [177]

Blundell, R., M. Browning, and I. Crawford (2008), "Best nonparametric bounds on demand responses." *Econometrica*, 76, 1227–1262. [163]

Blundell, R., X. Chen, and D. Kristensen (2007), "Semi-nonparametric IV estimation of shape-invariant Engel curves." *Econometrica*, 75, 1613–1669. [162, 164, 174]

Blundell, R. and A. Duncan (1998), "Kernel regression in empirical microeconomics." *Journal of Human Resources*, 33, 62–87. [162, 164, 176]

Blundell, R., A. Duncan, and K. Pendakur (1998), "Semiparametric estimation and consumer demand." *Journal of Applied Econometrics*, 13, 435–461. [164, 176]

Blundell, R. W. and J. L. Powell (2004), "Endogeneity in semiparametric binary response models." *The Review of Economic Studies*, 71, 655–679. [185]

Bound, J., D. A. Jaeger, and R. M. Baker (1995), "Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak." *Journal of the American Statistical Association*, 90, 443–450. [161]

Breunig, C. and J. Johannes (2016), "Adaptive estimation of functionals in nonparametric instrumental regression." *Econometric Theory*, 32, 612–654. [175]

Breza, E. (2013), "Peer effects and loan repayment: Evidence from the Krishna default crisis." Unpublished working paper. [163]

Canay, I. A., A. Santos, and A. M. Shaikh (2013), "On the testability of identification in some nonparametric models with endogeneity." *Econometrica*, 81, 2535–2559. [164]

Chay, K. and K. Munshi (2015), "Black networks after emancipation: Evidence from Reconstruction and the Great Migration." Unpublished working paper. [163]

Chen, X. (2007), “Large sample sieve estimation of semi-nonparametric models.” *Handbook of Econometrics*, 6, 5549–5632. [170, 194]

Chen, X. and T. M. Christensen (2015), “Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions.” *Journal of Econometrics*, 188, 447–465. [177]

Chen, X. and T. M. Christensen (2018a), “Optimal sup-norm rates, adaptivity and inference in nonparametric instrumental variables estimation.” *Quantitative Economics*, 9, 39–84. [175]

Chen, X. and T. M. Christensen (2018b), “Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression.” *Quantitative Economics*, 9, 39–84. [174, 194]

Chen, X. and D. Pouzo (2012), “Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals.” *Econometrica*, 80, 277–321. [170]

Chen, X. and M. Reiss (2011), “On rate optimality for ill-posed inverse problems in econometrics.” *Econometric Theory*, 27, 497–521. [194]

Chernozhukov, V. and C. Hansen (2005), “An IV model of quantile treatment effects.” *Econometrica*, 73, 245–261. [185]

Chesher, A. (2003), “Identification in nonseparable models.” *Econometrica*, 71, 1405–1441. [162, 167]

Chesher, A. (2007), “Instrumental values.” *Journal of Econometrics*, 139, 15–34. [162]

Coe, N. B., H.-M. von Gaudecker, M. Lindeboom, and J. Maurer (2012), “The effect of retirement on cognitive functioning.” *Health Economics*, 21, 913–927. [163, 164]

Darolles, S., Y. Fan, J.-P. Florens, and E. Renault (2011), “Nonparametric instrumental regression.” *Econometrica*, 79, 1541–1565. [162]

Das, M., W. K. Newey, and F. Vella (2003), “Nonparametric estimation of sample selection models.” *The Review of Economic Studies*, 70, 33–58. [185]

Daubechies, I. (1992), *Ten Lectures on Wavelets*, Vol. 61. SIAM. [194]

Davidson, R. and J. G. MacKinnon (1993), *Estimation and Inference in Econometrics*. OUP Catalogue. [168]

Del Bono, E. and A. Weber (2008), “Do wages compensate for anticipated working time restrictions? Evidence from seasonal employment in Austria.” *Journal of Labor Economics*, 26, 181–221. [163, 176]

D’Haultfœuille, X. and P. Février (2015), “Identification of nonseparable triangular models with discrete instruments.” *Econometrica*, 83, 1199–1210. [167]

Dufour, J.-M. (1997), “Some impossibility theorems in econometrics with applications to structural and dynamic models.” *Econometrica*, 65, 1365–1387. [161]

Dustmann, C. and C. Meghir (2005), “Wages, experience and seniority.” *The Review of Economic Studies*, 72, 77–108. [162, 164, 176]

Frazer, G. (2008), “Used-clothing donations and apparel production in Africa.” *The Economic Journal*, 118, 1764–1784. [163]

Freyberger, J. (2017), “On completeness and consistency in nonparametric instrumental variable models.” *Econometrica*, 85, 1629–1644. [164]

Hall, P. and J. L. Horowitz (2005), “Nonparametric methods for inference in the presence of instrumental variables.” *The Annals of Statistics*, 33, 2904–2929. [164]

Han, S. (2020), “Supplement to ‘Nonparametric estimation of triangular simultaneous equations models under weak identification’.” *Quantitative Economics Supplemental Material*, 11, https://doi.org/10.3982/QE975. [166]

Han, S. and A. McCloskey (forthcoming), “Estimation and inference with a (nearly) singular Jacobian.” *Quantitative Economics*. [162]

Hastie, T. and R. Tibshirani (1986), “Generalized additive models.” *Statistical Science*, 1, 297–310. [163]

Heckman, J. J. and E. Vytlacil (2005), “Structural equations, treatment effects, and econometric policy evaluation 1.” *Econometrica*, 73, 669–738. [185]

Henderson, D. J., C. Papageorgiou, and C. F. Parmeter (2013), “Who benefits from financial development? New methods, new evidence.” *European Economic Review*, 63, 47–67. [163, 164]

Hoderlein, S. (2009), “Endogenous semiparametric binary choice models with heteroscedasticity.” Cemmap working paper. [167]

Hong, Y. and H. White (1995), “Consistent specification testing via nonparametric series regression.” *Econometrica*, 63, 1133–1159. [185]

Horowitz, J., J. Klemelä, E. Mammen et al. (2006), “Optimal estimation in additive regression models.” *Bernoulli*, 12, 271–298. [194]

Horowitz, J. L. (2011), “Applied nonparametric instrumental variables estimation.” *Econometrica*, 79, 347–394. [182, 183]

Imbens, G. W. and W. K. Newey (2009), “Identification and estimation of triangular simultaneous equations models without additivity.” *Econometrica*, 77, 1481–1512. [162, 164, 167]

Jansson, M. and D. Pouzo (2019), “Towards a general large sample theory for regularized estimators.” arXiv:1712.07248v2. [175]

Jiang, J., Y. Fan, and J. Fan (2010), “Estimation in additive models with highly or nonhighly correlated covariates.” *The Annals of Statistics*, 38, 1403–1432. [163]

Jun, S. J. and J. Pinkse (2012), “Testing under weak identification with conditional moment restrictions.” *Econometric Theory*, 28, 1229. [169, 176]

Kleibergen, F. (2002), “Pivotal statistics for testing structural parameters in instrumental variables regression.” *Econometrica*, 70, 1781–1803. [161]

Kleibergen, F. (2005), “Testing parameters in GMM without assuming that they are identified.” *Econometrica*, 73, 1103–1123. [161, 162]

Koster, H. R., J. Ommeren, and P. Rietveld (2014), “Agglomeration economies and productivity: A structural estimation approach using commercial rents.” *Economica*, 81, 63–85. [163]

Lee, S. (2007), “Endogeneity in quantile regression models: A control function approach.” *Journal of Econometrics*, 141, 1131–1158. [185]

Lepskii, O. (1991), “On a problem of adaptive estimation in Gaussian white noise.” *Theory of Probability & Its Applications*, 35, 454–466. [175]

Linton, O. B. (1997), “Miscellanea: Efficient estimation of additive nonparametric regression models.” *Biometrika*, 84, 469–473. [163]

Lyssiotou, P., P. Pashardes, and T. Stengos (2004), “Estimates of the black economy based on consumer demand approaches.” *The Economic Journal*, 114, 622–640. [162, 176]

Mazzocco, M. (2012), “Testing efficient risk sharing with heterogeneous risk preferences.” *The American Economic Review*, 102, 428–468. [163]

Moreira, M. J. (2003), “A conditional likelihood ratio test for structural models.” *Econometrica*, 71, 1027–1048. [161]

Newey, W. K. (1990), “Efficient instrumental variables estimation of nonlinear models.” *Econometrica*, 58, 809–837. [176]

Newey, W. K. (1997), “Convergence rates and asymptotic normality for series estimators.” *Journal of Econometrics*, 79, 147–168. [174]

Newey, W. K., J. L. Powell, and F. Vella (1999), “Nonparametric estimation of triangular simultaneous equations models.” *Econometrica*, 67, 565–603. [162]

Newey, W. K. and J. L. Powell (2003), “Instrumental variable estimation of nonparametric models.” *Econometrica*, 71, 1565–1578. [164]

Nielsen, J. P. and S. Sperlich (2005), “Smooth backfitting in practice.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 43–61. [163]

Pereverzev, S. and E. Schock (2005), “On the adaptive selection of the parameter in regularization of ill-posed problems.” *SIAM Journal on Numerical Analysis*, 43, 2060–2076. [175]

Pinkse, J. (2000), “Nonparametric two-step regression estimation when regressors and error are dependent.” *Canadian Journal of Statistics*, 28, 289–300. [162]

Pouzo, D. (2016), “On the non-asymptotic properties of regularized M-estimators.” arXiv:1512.06290v3. [175]

Skinner, J. S., E. S. Fisher, and J. Wennberg (2005), “The efficiency of medicare.” In *Analyses in the Economics of Aging*, 129–160, University of Chicago Press. [162]

Sperlich, S., O. B. Linton, and W. Härdle (1999), “Integration and backfitting methods in additive models-finite sample properties and comparison.” *Test*, 8, 419–458. [163]

Staiger, D. and J. H. Stock (1997), “Instrumental variables regression with weak instruments.” *Econometrica*, 65, 557–586. [161, 163, 178, 180]

Stock, J. H. and J. H. Wright (2000), “GMM with weak identification.” *Econometrica*, 68, 1055–1096. [162, 169]

Stock, J. H. and M. Yogo (2005), “Testing for weak instruments in linear IV regression.” In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, 80–108. [161, 162, 178]

Stone, C. J. (1982), “Optimal global rates of convergence for nonparametric regression.” *The Annals of Statistics*, 10, 1040–1053. [174]

Torgovitsky, A. (2015), “Identification of nonseparable models using instruments with small support.” *Econometrica*, 83, 1185–1197. [167]

Trefethen, L. N. (2008), “Is Gauss quadrature better than Clenshaw–Curtis?” *SIAM (Society for Industrial and Applied Mathematics) Review*, 50, 67–87. [172]

Triebel, H. (2006), *Theory of Function Spaces. III*. Monographs in Mathematics, Vol. 100. Birkhauser Verlag, Basel. [195]

Tsybakov, A. B. (2009), *Introduction to Nonparametric Estimation*. Springer. [195, 196]

Yatchew, A. and J. A. No (2001), “Household gasoline demand in Canada.” *Econometrica*, 69, 1697–1709. [162, 176]

Co-editor Andres Santos handled this manuscript.

Manuscript received 10 September, 2017; final version accepted 11 June, 2019; available online 1 July, 2019.