Nonparametric Estimation in Case of Endogenous Selection

Christoph Breunig (Humboldt-Universität zu Berlin)
Enno Mammen (Universität Heidelberg)
Anna Simoni (CREST)

Discussion Paper No. 58

December 20, 2017

Collaborative Research Center Transregio 190 | www.rationality-and-competition.de
Ludwig-Maximilians-Universität München | Humboldt-Universität zu Berlin
Spokesperson: Prof. Dr. Klaus M. Schmidt, University of Munich, 80539 Munich, Germany
+49 (89) 2180 3405 | info@rationality-and-competition.de

Nonparametric Estimation in case of Endogenous Selection *

Christoph Breunig ⋆ Enno Mammen [1] Anna Simoni [2]

Humboldt-Universität zu Berlin *Universität Heidelberg* *CREST, CNRS*

December 20, 2017

This paper addresses the problem of estimation of a nonparametric regression function from selectively observed data when selection is endogenous. Our approach relies on independence between covariates and selection conditionally on potential outcomes. Endogeneity of regressors is also allowed for. In the exogenous and endogenous case, consistent two-step estimation procedures are proposed and their rates of convergence are derived. Pointwise asymptotic distribution of the estimators is established. In addition, bootstrap uniform confidence bands are obtained. Finite sample properties are illustrated in a Monte Carlo simulation study and an empirical illustration.

Keywords: Endogenous selection, instrumental variable, sieve minimum distance, regression estimation, inverse problem, inverse probability weighting, convergence rate, asymptotic normality, bootstrap uniform confidence bands.

JEL classification: C14, C26

*Financial support by Deutsche Forschungsgemeinschaft through the Research Training Group RTG 1953 is gratefully acknowledged. The authors gratefully thank the Co-Editor Jianqing Fan, an Associate Editor, and two anonymous referees for their many constructive comments on the previous version of the paper. The authors are also grateful to Timothy Armstrong, Xiaohong Chen, Victor Chernozhukov, Elise Coudin, Laurent Davezies, Kirill Evdokimov, Xavier D'Haultfoeuille, Michael Lebacher, Aureo de Paula, Christoph Rothe, Bernard Salanié and seminar participants at Bristol, Columbia University, CREST, Mannheim University, MIT and Yale University for useful comments and to the GIP team at SFB 884-Mannheim for providing data of the German Internet Panel. Christoph Breunig was supported by the DFG postdoctoral fellowship BR 4874/1-1. The author is also grateful for support and hospitality of the Cowles Foundation. Financial support by Deutsche Forschungsgemeinschaft through CRC TRR 190 is gratefully acknowledged. Anna Simoni gratefully acknowledges financial support from ANR-13-BSH1-0004 (IPANEMA), Labex ECODEC (ANR-11-LABEX-0047) and hospitality from University of Mannheim and Boston College.

⋆School of Business and Economics, Humboldt-Universität zu Berlin, Spandauer Straße 1, 10178 Berlin, Germany, e-mail: christoph.breunig@hu-berlin.de

[1]Institute for Applied Mathematics, Universität Heidelberg, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany, e-mail: mammen@math.uni-heidelberg.de

[2]CREST, CNRS, ENSAE, École Polytechnique - 5, avenue Henry Le Chatelier, 91120 Palaiseau, France, e-mail: anna.simoni@ensae.fr

1. Introduction

This paper addresses the problem of estimation of a regression function from selectively observed data. To explain the problem at stake, consider a partially observed dependent variable Y^*, a vector of covariates X and a binary indicator Δ. The econometrician observes a realization of Δ and X for each individual in the random sample but only observes a realization of Y^* when $\Delta = 1$. In many applications it is important to learn about $\mathbf{E}[Y^*|X]$ which, by the law of total expectation, can be written as

$$\mathbf{E}[Y^*|X] = \mathbf{E}[Y^*|X, \Delta = 1]\mathbb{P}(\Delta = 1|X) + \mathbf{E}[Y^*|X, \Delta = 0]\mathbb{P}(\Delta = 0|X).$$

The difficulty arises because the data available cannot identify $\mathbf{E}[Y^*|X, \Delta = 0]$ nor $\mathbf{E}[Y^*|X]$. In this paper, we address this lack of identification by assuming independence between the regressors X and the selection mechanism Δ conditionally on the selectively observed outcome Y^*. Relying on this assumption we propose a new methodology to consistently estimate the regression function $\varphi(\cdot) = \mathbf{E}[Y^*|X = \cdot]$.

We also allow for endogeneity of covariates. More precisely, we consider the nonparametric instrumental variable model of Newey and Powell [2003], Ai and Chen [2003] and Darolles et al. [2011] but where the dependent variable is only observed selectively. That is, we propose a method to estimate a structural function ψ which satisfies

$$Y^* = \psi(Z) + U$$

for some unobservables U, where Z is endogenous in the sense that $\mathbf{E}[U|Z] \neq 0$ and an additional instrumental variable X is available such that $\mathbf{E}[U|X] = 0$. If the instrument X is independent of the selection given potential outcome Y^*, we show that ψ is identified and can be consistently estimated under commonly imposed assumptions. The model considered in Ai and Chen [2003] is more general than the nonparametric instrumental variable model and, among others, it includes the nonparametric instrumental variable model with selectively observed Y^* implied by our assumptions. However, their analysis, while being an alternative to ours, is not explicitly tailored for this type of model.

Previous literature has proposed different solutions to overcome the problem of lack of identification of $\mathbf{E}[Y^*|X]$. One solution consists in assuming *missing-at-random* (MAR), namely, independence between the selection variable and the outcome conditional on the observed covariates, see Rubin [1976]. MAR implies $\mathbf{E}[Y^*|X] = \mathbf{E}[Y^*|X, \Delta = 1] = \mathbf{E}[Y^*|X, \Delta = 0]$. Unfortunately, the plausibility of this assumption may be questioned in many economic examples where missing observations arise due to self-selection, nonresponse or because counterfactual variables are unobservable (see the examples given in Heckman [1979]).

In his seminal work, Heckman [1974, 1979] relies on instruments that determine selection but not the outcome and proposes a consistent parametric estimation method. Point-identification comes from parametric restrictions. Ahn and Powell [1993] and Das et al. [2003] extend Heckman's approach to a semiparametric and nonparametric framework, respectively.

An alternative strategy relies on "identification at infinity", namely, on the fact that the selection problem becomes negligible for large values of the covariates. This strategy requires the existence of a covariate with a large support, see Chamberlain [1986]. Based on this idea Lewbel [2007] and D'Haultfoeuille and Maurel [2013] propose alternative identification strategies.

A completely different approach was proposed by Manski [1989] who poses, as the only restriction, a bound on the support of Y^* conditional on X. This implies a bound on $\mathbf{E}[Y^*|X]$. While such a weak restriction has the advantage of ensuring robust inference, only partial identification of $\mathbf{E}[Y^*|X]$ can be achieved. Following Manski [1989], an extensive literature on bounds and partial identification in econometrics has flourished (see *e.g.* Chernozhukov et al. [2013] and Tamer [2010] for a review).

In this paper, we solve the problem of endogenous selection by using the following instrumental variable assumption. We assume independence between selection Δ and instruments X, conditional on the outcome Y^* (and possibly additional covariates), namely

$$\Delta \perp\!\!\!\perp X \,|\, Y^*. \tag{1.1}$$

This assumption is suitable when selection is driven by the outcome Y^* and, once Y^* is present, X does not contain additional information on the missing data mechanism. For example, if Y^* denotes income and X expenditure then, typically in survey data, whether people report their income or not is primarily determined by the level of their income. Assumption (1.1) has been used in the previous literature on missing data (see *e.g.* Chen [2001], Tang et al. [2003], Zhao and Shao [2015] in the statistics literature and D'Haultfoeuille [2010], Davezies and D'Haultfoeuille [2013], Ramalho and Smith [2013] in the econometrics literature). This type of assumption is common in the finite mixtures literature (*e.g.* Henry et al. [2014]) as well as in the nonclassical measurement error literature (see *e.g.* Hu and Schennach [2008]). Moreover, it is a particular case of Assumption (41) in Manski [1994]. Assumption (1.1) alone is not sufficient for nonparametric identification of the regression function φ or the selection probability $\mathbb{P}(\Delta = 1|Y^*)$. In this paper, however, we show that the function φ is identified under additional assumptions. Moreover, under a completeness condition, Assumption (1.1) can be tested, see Theorem 2.4 in D'Haultfoeuille [2010].

There are many situations where the outcome Y^* is only selectively observed but there is an instrument X available satisfying condition (1.1). In survey data, for instance, selective missingness of Y^* can be caused by a respondent's reservation to answer a question that appears too sensitive, affects confidentiality, or is too complex. Examples of instruments in these cases, are past health status when the outcome is a symptom of breathlessness (see Zhao and Shao [2015]) or past income information when the outcome is selectively observed current labor income (see Breunig [2016]). To overcome the difficulty of selective nonresponse in financial investment questions, we may also follow the instrumental variable strategy of Huck et al. [2015], by providing information on randomized stock market returns before asking financial questions. Our approach does not only allow to estimate the causal effect ψ of beliefs about future stock market returns on investment decision (as in Huck et al. [2015]) but also to account for a selective response behavior.

Our paper goes beyond what has been proposed in the literature so far. While the nonparametric inverse selection probability function $g(\cdot) = 1/\mathbb{P}(\Delta = 1|Y^* = \cdot)$ is a nuisance function, the regression function φ or the structural function ψ are the functions of interest. We allow for settings where the rate of convergence for estimating the inverse selection probability function g is much slower as that of estimating the functions of interest. Our first main result is that we achieve the same rates for estimating the regression function φ or the structural function ψ as in the respective model where Y is fully observed. The second main result is that the sieve variances of estimating these functions of interest are not affected by the slow rate of convergence of g. Our estimator is based on a two-step procedure where an estimator of g is plugged in when estimating φ or ψ in the second step. This

corresponds to inverse probability weighting estimators but where g is obtained through a nonparametric instrumental variables estimator. The main difficulty in the proofs lies in showing that the error contributions coming from estimating g in the first step are canceled out when the estimator of g is smoothed in the second step. We go beyond the existing literature by providing inference on the regression function φ without imposing parametric restrictions on the probability function g (for a semiparametric approach see Zhao and Shao [2015] and the references therein). Moreover, we use the identification assumption (1.1) to establish inference on the structural function ψ in an instrumental regression model, which has not been considered in the literature yet.

The reason for slow convergence of the inverse conditional probability function g is due to nonparametric instrumental estimation which leads in general to an ill-posed inverse problem (see *e.g.* Newey and Powell [2003], Ai and Chen [2003]). In contrast to the rate of estimating the nonparametric regression function φ or the structural function ψ with fully observed Y, we get an additional bias due to estimation of the selection probability in the first step. This is in line with Das et al. [2003] who also obtain an additional bias term in their convergence rate but which is due to estimation of a propensity score. Under additional smoothness conditions, however, the additional bias for estimating the nuisance function g is asymptotically negligible and the usual nonparametric (instrumental) regression rate is obtained.

We establish pointwise asymptotic normality for our estimators of the regression function φ and the structural function ψ. While the derived sieve variance does not suffer from the potential ill-posed problem of estimating the inverse selection probability g, we see that the sieve variance is enlarged due to multiplication by g. We also extend these pointwise results by providing a bootstrap procedure to construct uniform confidence bands.

The remainder of this paper is organized as follows. In Section 2, we present the setup and discuss identification. In Section 3, we present our two-step estimator for φ, we give rates of convergence of the integrated squared error of our estimator, establish pointwise asymptotic normality of our estimator and construct uniform confidence bands. In section 4 we extend this analysis to the structural function ψ. The finite sample properties of our procedure are investigated through Monte Carlo experiments whose results are reported in Section 5. Section 6 presents an empirical application of our method to estimate the propensity to work in the German speaking population by using "German Internet Panel" data. All proofs are postponed to the appendix and Supplementary Material. Moreover, in the supplementary material we propose a nonparametric testing procedure to test the identifying assumption (1.1) based on a maintained completeness assumption. The large and small sample properties of the test are studied in appendix B in the Supplementary Material.

2. Identification

In this section, we provide assumptions under which the selection probability function $\mathbb{P}(\Delta = 1|Y^* = \cdot)$ and the regression function $\mathbf{E}[Y^*|X = \cdot]$ are identified. We further motivate our estimation procedure.

2.1. Setup and Main Assumptions

Let (Δ, Y^*, X^t) be a jointly distributed random vector where (Y^*, X^t) is a random vector which takes values in $\mathbb{R}^{1+d_x}$ and Δ is a random variable which takes values in $\{0,1\}$. A realization of (Δ, X^t) is observed for each individual in the random sample while a realization of the dependent variable Y^* is observed when $\Delta = 1$ and missing when $\Delta = 0$. We write $Y = \Delta Y^*$.[1] We assume that the marginal distribution of Y^* (resp. X) admits a probability density function p_Y (resp. p_X) with respect to the Lebesgue measure. The following three assumptions are sufficient to identify the joint distribution of (Δ, Y^*, X^t).

Assumption 1. *It holds that*

$$\Delta \perp\!\!\!\perp X \,|\, Y^*.$$

Assumption 1 states an exclusion restriction of the random vector X with respect to the selection variable Δ given potential outcomes Y^*. The vector X is referred to as the vector of instruments. This assumption can be justified in many settings. An example is provided by measurement error models where Y^* is observed with error for some individuals. Then, for some error ε, $X = Y^* + \varepsilon$ can be interpreted as a proxy for Y^* and satisfies Assumption 1 if $\varepsilon \perp\!\!\!\perp \Delta$, see *e.g.* Chen et al. [2011]. Other examples are given by data with nonresponse. For instance, consider the case where Y^* is income and X is expenditure. It could be that people with high income are less likely to report it. Examples of such type of incomplete data sets are the French "Enquête Budget de famille" of INSEE or the British "Family expenditure Survey". For further illustrations of Assumption 1 we refer to Ramalho and Smith [2013]. In particular, Ramalho and Smith [2013] justify the conditional independence condition in Assumption 1 in a standard crossing ordered choice model where Y^* denotes latent utility and X represents individual characteristics and other variables that affect utility (here Y is the censored discrete outcome).

Assumption 2. *For every function ϕ that is bounded from below almost surely and satisfies $\mathbf{E}\,|\phi(Y^*)| < \infty$ it holds that $\mathbf{E}[\phi(Y^*)|X = \cdot] = 0$ implies $\phi(Y^*) = 0$.*[2]

Assumption 2 is weaker than L^1–completeness (which requires completeness for all $\phi \in L^1$) but stronger than bounded–completeness (which requires completeness only for those functions that are bounded from above and from below). Completeness conditions have been largely used in econometrics as identification assumptions, see *e.g.* Darolles et al. [2011], Newey and Powell [2003], Blundell et al. [2007], Hu and Schennach [2008], D'Haultfoeuille [2011] and Hoderlein et al. [2016].

Assumption 3. *It holds* $\mathbf{E}[1/\mathbb{P}(\Delta = 1|Y^*)] < \infty$.

Assumption 3 restricts the selection probability $\mathbb{P}(\Delta = 1|Y^*)$ to be strictly positive on $\mathcal{Y}^*$ almost surely, where $\mathcal{Y}^*$ denotes the support of Y^*. On the other hand, if $\mathcal{Y}^*$ is compact then $\mathbb{P}(\Delta = 1|Y^* = y^*) > 0$ for every $y^* \in \mathcal{Y}^*$ implies Assumption 3. This assumption can rule out a selection when it is a deterministic function of Y^* such as $\mathbb{P}(\Delta = 1|Y^* = \cdot) = \mathbb{1}\{\cdot \geqslant c\}$ for some constant c belonging to the interval $(\min(\mathcal{Y}^*), \infty)$. Here $\mathbb{1}$ denotes the indicator

[1] In our setting, Y^* is assumed to be a scalar. Our results would still hold if we extended this framework to allow for a p-dimensional vector Y^* of selectively observed variables. In this case $\Delta = (\Delta^{(j)})_{1 \leqslant j \leqslant p}$ and the j-th component of Y^* would be observed when $\Delta^{(j)} = 1$ and missing when $\Delta^{(j)} = 0$. This extension would require little modifications of our method but would burden the notation and the presentation. For this reason we do not consider it.

[2] Since conditional expectations are defined only up to equality a.s., all (in)equalities with conditional expectations and/or random variables are understood as (in)equalities a.s., even if we do not say so explicitly.

function. To understand Assumption 3, consider the example where $\Delta = \xi(Y^*, \eta)$ for some function $\xi(\cdot)$ and a random variable η. Then, Assumption 3 is verified if the distribution of η is such that the set $\{\eta; \xi(y^*, \eta) = 1\}$ has positive probability for every $y^* \in \mathcal{Y}^*$ and $\mathcal{Y}^*$ is compact.

2.2. Identification and idea of the estimator

Our object of interest is $\mathbf{E}[Y^*|X]$ while the selection probability $\mathbb{P}(\Delta = 1|Y^*)$ is a nuisance (functional) parameter.[3] However, knowledge of the latter allows us to identify and estimate $\mathbf{E}[Y^*|X]$ in a way that we now explain. Let us introduce the inverse selection probability function $g(\cdot) := 1/\mathbb{P}(\Delta = 1|Y^* = \cdot)$. Under Assumptions 1–3, the function g is identified through the conditional moment restriction

$$\mathbf{E}\Big[\Delta g(Y^*)\Big| X\Big] = 1, \tag{2.1}$$

see Theorem 2.3 in D'Haultfoeuille [2010]. In the first step of our two-step procedure, we make use of (2.1) to estimate the inverse selection probability function g.

Since the function g is identified by equation (2.1), identification of the conditional expectation $\mathbf{E}[Y^*|X]$ follows from

$$\mathbf{E}[Y^*|X] = \mathbf{E}[Y^*\,\mathbb{P}(\Delta = 1|Y^*)g(Y^*)|X] = \mathbf{E}\Big[\,\mathbf{E}[Y^*\Delta g(Y^*)|Y^*]\,\Big|X\Big] = \mathbf{E}[Y^*\Delta g(Y^*)|X] = \mathbf{E}[Yg(Y)|X] \tag{2.2}$$

where the first equality follows from Assumption 3 and second to last equality follows from Assumption 1 and the fact that $g(Y^*) = g(Y)$ whenever $Y := Y^*\Delta$ differs from zero. This result shows that $\mathbf{E}[Y^*|X]$ can be written as a weighted average of the observed Y where the weight is equal to the inverse selection probability function. We use equation (2.2) to construct an estimator for $\mathbf{E}[Y^*|X]$ in the second step of our estimation procedure.

Remark 2.1 (Including additional covariates). In empirical applications, only a subset of the covariates might be independent of selection given potential outcome. We can cover this case by slightly extending Assumption 1. More precisely, suppose that $X = (X_1, X_2)$ and that Assumption 1 is modified as $\Delta \perp\!\!\!\perp X_1|(Y^*, X_2)$ and hence, X_2 can be correlated to Δ. Under this assumption and Assumption 3 modified as indicated below, $\mathbf{E}[Y^*|X] = \mathbf{E}[Yg(Y, X_2)|X]$ where $g(y, x) = 1/\mathbb{P}(\Delta = 1|Y^* = y, X_2 = x)$. Moreover, if Assumptions 2 and 3 are modified with $\phi(Y^*)$ replaced by $\phi(Y^*, X_2)$ and $\mathbb{P}(\Delta = 1|Y^*)$ replaced by $\mathbb{P}(\Delta = 1|Y^*, X_2) > 0$, respectively, then g is identified by $\mathbf{E}[\Delta g(Y^*, X_2)|X] = 1$. □

2.3. Notation

For a random vector V we use the corresponding calligraphic capital letter $\mathcal{V}$ to denote its support. Let $L^2_V = \{\phi : \|\phi\|^2_V := \mathbf{E}\,|\phi(V)|^2 < \infty\}$ denote the space of square integrable functions of V with respect to the distribution of V. We denote by $\langle\cdot,\cdot\rangle_V$ the inner product in L^2_V that induces $\|\cdot\|^2_V$. Moreover, $\|\phi\|_\infty := \sup_{v\in\mathcal{V}}|\phi(v)|$ denotes the sup norm and $\|\cdot\|$ is the usual Euclidean norm. We introduce the following Hilbert space

$$\mathcal{G} = \Big\{\phi : \|\phi\|^2_{\mathcal{G}} := \mathbf{E}[\Delta\phi^2(Y)] < \infty\Big\}$$

[3] In the following, we simplify the notation and for two random variables V, W we may use $\mathbf{E}[V|W]$ as a shorthand for $\mathbf{E}[V|W = \cdot]$ and $\mathbb{P}(V = v|W)$ as a shorthand for $\mathbb{P}(V = v|W = \cdot)$.

with associated inner product $\langle \phi_1, \phi_2 \rangle_{\mathcal{G}} = \mathbf{E}[\Delta \phi_1(Y)\phi_2(Y)]$ for $\phi_1, \phi_2 \in \mathcal{G}$. Note that L^2_Y is contained in $\mathcal{G}$. Further, we have $g \in \mathcal{G}$ since by Assumption 3 it holds $\mathbf{E}[g(Y^*)] < \infty$ and thus $\mathbf{E}[\Delta g^2(Y)] = \mathbf{E}[\mathbb{P}(\Delta = 1|Y^*)g^2(Y^*)] < \infty$. Now equation (2.1) can be written in a more compact form by using the following notation. Let $T : \mathcal{G} \to L^2_X$ be the linear operator $(T\phi)(\cdot) = \mathbf{E}[\Delta\phi(Y)|X = \cdot]$. This mapping is well defined since, for all $\phi \in \mathcal{G}$, the Jensen's inequality implies $\mathbf{E}[|(T\phi)(X)|^2] \leqslant \mathbf{E}[\Delta\phi^2(Y)] < \infty$. Thereby, equation (2.1) can be equivalently written as the operator equation

$$Tg = 1 \tag{2.3}$$

where the function g is identified under Assumptions 1–3.

Let $\{f_j\}_{j\geqslant 1}$ (resp. $\{e_j\}_{j\geqslant 1}$) be a sequence of approximating functions in L^2_X (resp. $\mathcal{G}$). Then, we denote by $f_{\underline{m_n}}(X) = (f_1(X), \ldots, f_{m_n}(X))^t$ (resp. $e_{\underline{k_n}}(Y) = (e_1(Y), \ldots, e_{k_n}(Y))^t$) a vector of functions which are used to approximate the conditional expectation $\mathbf{E}[Y^*|X]$ (resp. the inverse selection probability $g(Y)$) and by $\mathbf{X}_{m_n} = \left(f_{\underline{m_n}}(X_1), \ldots, f_{\underline{m_n}}(X_n)\right)^t$ (resp. $\mathbf{Y}_{k_n} = (\Delta_1 e_{\underline{k_n}}(Y_1), \ldots, \Delta_n e_{\underline{k_n}}(Y_n))^t$) the $n \times m_n$ (resp. $n \times k_n$) matrix obtained by putting together the n vectors $f_{\underline{m_n}}(X_i)$, $i = 1, \ldots, n$ (resp. $\Delta_i e_{\underline{k_n}}(Y_i)$, $i = 1, \ldots, n$, where $\Delta_i e_{\underline{k_n}}(Y_i)$ denotes the product of Δ_i and the vector $e_{\underline{k_n}}(Y_i)$). We denote by $\mathcal{F}_{m_n} = \left\{\phi(\cdot) = \sum_{j=1}^{m_n} \beta_j f_j(\cdot) : \beta \in \mathbb{R}^{m_n}\right\}$ the linear sieve space of dimension $m_n < \infty$ that becomes dense in L^2_X as n tends to infinity. For a matrix A we denote by A^- its generalized inverse and by $\lambda_{\min}(A)$ its minimum eigenvalue. For a function ϕ defined on $\mathcal{Y}$, we denote by $M_\phi : L^2_Y \to L^2_Y$ the multiplication operator $M_\phi \varsigma = \phi\varsigma$ which is bounded if ϕ is bounded on $\mathcal{Y}$. Then $(M_{id}\varsigma)(y) = y\varsigma(y)$ for all $y \in \mathcal{Y}$ and $\varsigma \in L^2_Y$, where id denotes the identity function.

3. Nonparametric Regression with Sample Selection

In this section, we consider estimation of the regression function φ. The first step estimation procedure for the inverse selection probability g is based on constrained sieve minimum distance. In the second step, we use a plug-in series estimator of the conditional expectation $\varphi(\cdot) = \mathbf{E}[Yg(Y)|X = \cdot]$. We also propose an alternative estimator based on normalization of the inverse probability weights. For both estimators, we derive the rate of convergence in mean square error and their asymptotic distribution. Finally, we propose a bootstrap procedure to obtain uniform confidence bands.

3.1. The Estimators and their Rates of Convergence

For every $\phi \in \mathcal{G}$, denote $\chi(\cdot, \phi) = \mathbf{E}[\Delta\phi(Y) - 1|X = \cdot]$. The least squares estimator of $\chi(\cdot, \phi)$ is given by

$$\widehat{\chi}_n(\cdot, \phi) = f_{\underline{m_n}}(\cdot)^t (\mathbf{X}^t_{m_n}\mathbf{X}_{m_n})^- \sum_{i=1}^{n} \Big(\Delta_i\phi(Y_i) - 1\Big) f_{\underline{m_n}}(X_i) \tag{3.1}$$

for some integer m_n which increases with the sample size n. Under conditions given below, $\mathbf{X}^t_{m_n}\mathbf{X}_{m_n}$ will be nonsingular with probability approaching one and hence its generalized inverse will be the standard inverse. We now introduce some assumptions.

Assumption 4. *(i) We observe a sample* $((\Delta_1, X_1, Y_1), \ldots, (\Delta_n, X_n, Y_n))$ *of independent and identical distributed (iid.) copies of* (Δ, X, Y) *where* $Y = \Delta Y^*$ *and* $\mathbf{E}[(Y^* - \varphi(X))^2 | X] < \infty$. *(ii) There exists a constant* $C > 0$ *and a sequence of positive integers* $(m_n)_{n \geqslant 1}$ *satisfying* $\sup_{x \in \mathcal{X}} \|f_{\underline{m_n}}(x)\|^2 \leqslant C m_n$ *such that* $m_n \log(m_n)/n = o(1)$. *(iii) The smallest eigenvalue of* $\mathbf{E}[f_{\underline{m}}(X) f_{\underline{m}}(X)^t]$ *is bounded away from zero uniformly in* m. *(iv) Let* $\varphi \in L^2_X$ *and there is* $F_{m_n} \varphi \in \mathcal{F}_{m_n}$ *such that* $\|F_{m_n}\varphi - \varphi\|_\infty = O(m_n^{-\alpha/d_x})$ *for some constant* $\alpha > 0$.

Assumption 4 $(ii) - (iii)$ restricts the magnitude of the approximating functions $\{f_j\}_{j \geqslant 1}$ and impose nonsingularity of their second moment matrix. Assumption 4 (ii) relaxes the classical assumption $m_n^2/n = o(1)$ and has been introduced in the recent econometric literature which employs either the Rudelson's inequality or the Bernstein inequality for random matrices, see Belloni et al. [2015] and Chen and Christensen [2015b] respectively. Here, the upper bound on the vector of basis functions holds for instance for polynomial splines, Fourier series and wavelet bases but rules out orthogonal polynomials and power series sieves. Assumption 4 (iv) determines the sieve approximation error which in turn characterizes the bias of the estimated regression function φ (see also Belloni et al. [2015] for a discussion of L^2 and L^∞ type approximation errors). For further discussion and examples of sieve bases, we refer to Chen [2007].

In the following, we consider the linear sieve space $\mathcal{G}_n = \left\{\phi(\cdot) = \sum_{j=1}^{k_n} \beta_j e_j(\cdot) : \beta \in \mathbb{R}^{k_n}\right\}$ of dimension $k_n < \infty$ that becomes dense in the function space $\mathcal{G}$ as n tends to infinity. We propose the following sieve minimum distance estimator

$$\widehat{g}_n = \underset{\{\phi \in \mathcal{G}_n : \phi(\cdot) \geqslant 1\}}{\arg\min} \sum_{i=1}^{n} \widehat{\chi}_n^2(X_i, \phi). \tag{3.2}$$

The constraint $\phi(\cdot) \geqslant 1$ imposed on the sieve space $\mathcal{G}_n$ ensures that the estimated conditional probability of observing Y^* belongs to the unit interval. This estimator of g corresponds to the penalized sieve minimum distance estimator suggested by Chen and Pouzo [2012].

If no constraint is imposed then the sieve estimator $\widehat{g}_n$ has an explicit solution given by

$$\widehat{g}_n(\cdot) = e_{\underline{k_n}}(\cdot)^t \widehat{\beta}_{k_n} \quad \text{and} \quad \widehat{\beta}_{k_n} = \left(\mathbf{Y}_{k_n}^t \mathbf{X}_{m_n} (\mathbf{X}_{m_n}^t \mathbf{X}_{m_n})^- \mathbf{X}_{m_n}^t \mathbf{Y}_{k_n}\right)^- \mathbf{Y}_{k_n}^t \mathbf{X}_{m_n} (\mathbf{X}_{m_n}^t \mathbf{X}_{m_n})^- \mathbf{X}_{m_n}^t \mathbf{1}_n \tag{3.3}$$

where $\mathbf{1}_n$ is a n-dimensional vector of ones and k_n is some integer such that $k_n \leqslant m_n$ and which increases with the sample size n.

The second step of our estimation procedure consists in using the estimator $\widehat{g}_n$ in (3.2) to construct an estimator for φ. Let $\mathbf{G}_n = \left(Y_1 \widehat{g}_n(Y_1), \ldots, Y_n \widehat{g}_n(Y_n)\right)^t$. Then, our estimator of the nonparametric regression function $\varphi(\cdot)$ is given by

$$\widehat{\varphi}_n(\cdot) = f_{\underline{m_n}}(\cdot)^t (\mathbf{X}_{m_n}^t \mathbf{X}_{m_n})^- \mathbf{X}_{m_n}^t \mathbf{G}_n. \tag{3.4}$$

When using inverse probability weights estimators, to account for missing data, the weights are typically normalized to sum to one (see, e.g., p. 823 in Wooldridge [2010]). We also pursue a similar strategy by constructing an alternative estimator for φ that involves an additional weighting of the empirical Gram matrix $\mathbf{X}_{m_n}^t \mathbf{X}_{m_n}$ by the estimated inverse selection probabilities. To do so, observe that $\mathbf{E}[f_{\underline{m_n}}(X) f_{\underline{m_n}}(X)^t] = \mathbf{E}[f_{\underline{m_n}}(X) f_{\underline{m_n}}(X)^t \Delta g(Y)]$. Let $\mathbf{D}_n$ denote a diagonal matrix with diagonal entries $\Delta_1 \widehat{g}_n(Y_1), \ldots, \Delta_n \widehat{g}_n(Y_n)$. Then, we consider the reweighted estimator of the nonparametric regression function $\varphi(\cdot)$ given by

$$\widetilde{\varphi}_n(\cdot) = f_{\underline{m_n}}(\cdot)^t (\mathbf{X}_{m_n}^t \mathbf{D}_n \mathbf{X}_{m_n})^- \mathbf{X}_{m_n}^t \mathbf{G}_n. \tag{3.5}$$

We see below that both estimators attain the same rate of convergence but require a different normalization factor to be asymptotically Gaussian.

Remark 3.1 (Relation to Ai and Chen [2007] and Chen and Pouzo [2012]). One may alternatively use the very general estimation approach proposed by Ai and Chen [2007], Chen and Pouzo [2012], and Chen and Pouzo [2015]. Indeed, the unknown nonparametric functions g and φ also satisfy the conditional moment restrictions $\mathbf{E}[\rho(Y, X, \Delta; g(\cdot), \varphi(\cdot))|X] = 0$, where

$$\rho(Y, X, \Delta; g(\cdot), \varphi(\cdot)) = \begin{pmatrix} \Delta g(Y) - 1 \\ Y g(Y) - \varphi(X) \end{pmatrix}.$$

Their resulting estimator, when there is no penalization and a linear sieve space is used, coincides with our estimators $(\widehat{g}_n, \widehat{\varphi}_n)$. There are, however, two main differences between Chen and Pouzo's fundamental work and ours. First, for our estimator the rates of convergence and the asymptotic distribution results are in general not affected by the slow rate of estimating g. This is because in our particular model only the nuisance function g but not the regression function φ stems from a potentially ill-posed inverse problem. In contrast, Chen and Pouzo [2012] introduce a scalar sieve measure of (local) ill-posedness to relate estimation in a strong norm relative to the weak norm induced by the conditional expectation given X. This scalar parameter depends on the ill-posedness of the inverse problem for g and influences the rate of convergence of $\widehat{\varphi}_n$. Thus, we use in this paper a different analysis tailored to our particular model to obtain the rates of convergence for our estimators.

Second, our estimator $\widetilde{\varphi}_n$ given in (3.5) differs from Chen and Pouzo [2012]'s proposed method as it normalizes the weights via reweighting. Normalizing the weights is common in the inverse probability weighting as it often stabilizes the estimators. In the treatment effect literature, it was also shown that normalization of the propensity score improves the finite sample results (see Frölich [2004]). In a similar way our Monte Carlo simulations show that normalizing the weights stabilizes and thus improves the finite sample behavior of our estimator (see Section 5). Therefore, we see our procedure more tailored to the selection model at hand and we make explicit use of this special structure to get asymptotic results. □

In the following, the sequence $(\mathcal{R}_n)_{n\geqslant 1}$ denotes the rate of convergence of the estimator $\widehat{g}_n$ w.r.t. to the norm $\|\cdot\|^2_{\mathcal{G}}$. The next assumption is used to recover the rate of convergence of $\widehat{\varphi}_n$.

Assumption 5. *(i) There exists a projection $F_{m_n} : L^2_X \to \mathcal{F}_{m_n}$ such that for all $\phi \in \mathcal{G}_n$, $\|F_{m_n} T\phi - T\phi\|_\infty = O(m_n^{-\alpha/d_x})$ for some constant $\alpha > 0$. (ii) There exists a sequence of positive integers $(\xi_n)_{n\geqslant 1}$ satisfying $\sup_{y\in\mathcal{Y}} \|e_{k_n}(y)\|^2 \leqslant \xi_n^2$ such that $k_n\xi_n^2/n = o(1)$. (iii) It holds $k_n^2\mathcal{R}_n = O(1)$ and $\mathbb{V}\mathrm{ar}(Y^*) < \infty$. (iv) $\|TM_{id}\phi\|_X/\|T\phi\|_X$ is bounded uniformly over all $\phi \in \mathcal{G}$ with $\|T\phi\|_X \neq 0$.*

Assumption 5 *(i)* holds true, for example, for splines or power series if the family of functions $\{T\phi : \phi \in \mathcal{G}_n\}$ contains only functions which are at least α–times continuously differentiable (see also Assumption 3 of Blundell et al. [2007]). Assumption 5 *(ii)* is satisfied with $\xi_n = \sqrt{k_n}$ when the approximating functions are for instance B-splines, Fourier series and wavelet bases. For Legendre polynomials this assumption is satisfied with $\xi_n = k_n$. Assumption 5 *(iii)* is a mild restriction on the rate of convergence of $\widehat{g}_n$ which we illustrate below. Instead of a bound on $\mathcal{Y}$, Assumption 5 *(iv)* restricts the size of the multiplication operator M_{id} in the norm induced by T. Otherwise stated, it requires that the norms of the

operators TM_{id} and T are equivalent. Assumption 5 *(iv)* is satisfied for instance under an additional link condition, like Assumption 6 *(ii)* below, if the basis functions coincide with Legendre polynomials or cardinal B-splines (see Example 3.1 below).

Estimation of g requires to "solve" a conditional moment restriction that is different from (2.2), namely $\mathbf{E}[\Delta g(Y^*)|X] = 1$. From Blundell et al. [2007] and Chen and Pouzo [2012] we obtain the rate of convergence of $\|T(\widehat{g}_n - g)\|_X$. Their result, however, is not enough to obtain the rate of convergence of $\widehat{\varphi}_n$ as our case requires to determine the rate of $\|\widehat{T}M_{id}\widehat{g}_n - TM_{id}g\|_X$ where $\widehat{T}$ is a series least square estimator of T. Thereby, we need to control for the estimation of T as well as to take into account the smoothing of this operator which allows to get rid of the ill-posedness in the estimation of g. In the proof we decompose this error in three parts: one that accounts for the estimation of T, one that accounts for the variance of $Y\widehat{g}_n(Y)$ and one that account for the sieve approximation error of φ. In the following, for any ϕ in $\mathcal{G}$ let $E_{k_n}\phi \in \mathcal{G}_n$ be such that $\|E_{k_n}\phi - \phi\|_\infty = o(1)$.

THEOREM 3.1. *Let Assumptions 1 – 5 hold true. Then* $\|\widehat{\varphi}_n - \varphi\|_X^2$ *and* $\|\widetilde{\varphi}_n - \varphi\|_X^2$ *are of the order*

$$O_p\Big(\max\Big(m_n^{-2\alpha/d_x}, \frac{m_n}{n}, \|T(E_{k_n}g - g)\|_X^2\Big)\Big).$$

As we see from Theorem 3.1, the rate of convergence of our estimators $\widehat{\varphi}_n$ and $\widetilde{\varphi}_n$ depends on both parameters k_n and m_n which correspond to the first and second estimation step, respectively. In addition to the usual nonparametric rate we obtain an additional bias term $\|T(E_{k_n}g - g)\|_X^2$ which is due to the sieve approximation of the inverse selection probability function g. An additional bias occurs also in the convergence rate for estimating regression functions in Theorem 4.1 of Das et al. [2003]. In their case, however, the additional bias arises from nonparametric estimation of a propensity score. From Theorem 3.1 we see that both estimators, $\widehat{\varphi}_n$ and $\widetilde{\varphi}_n$, attain the optimal nonparametric rate of convergence under the assumptions of the theorem if $\|T(E_{k_n}g - g)\|_X = O(m_n^{-\alpha/d_x})$. Also note that if the inverse selection probability g is sufficiently smooth in the sense that $\|E_{k_n}g - g\|_{\mathcal{G}} = O(m_n^{-\alpha/d_x})$ then, by the Jensen's inequality, the optimal nonparametric rate is obtained. In the following, we provide a rate of convergence under common smoothness assumptions.

ASSUMPTION 6. *(i) Assume* $\|E_{k_n}g - g\|_\infty = O(k_n^{-\beta})$ *for some constant* $\beta > 0$. *(ii) There exists a sequence of non-increasing positive real numbers* $(\tau_j)_{j\geqslant 1}$ *such that* $\|T\phi\|_X^2 \geqslant c\sum_{j=1}^\infty \tau_j\langle\phi, e_j\rangle_{\mathcal{G}}^2$ *and* $\|T\phi\|_X^2 \leqslant C\sum_{j=1}^\infty \tau_j\langle\phi, e_j\rangle_{\mathcal{G}}^2$ *for some constants* $c, C > 0$ *and all* $\phi \in \mathcal{G}$. *(iii) The largest eigenvalue of* $\big(\tau_j^{1/2}\tau_l^{-1/2}\langle M_{id}e_j, e_l\rangle_{\mathcal{G}}\big)_{j,l\geqslant 1}$ *is bounded away from infinity.*

Assumption 6 (*i*) determines the sieve approximation error for estimating the function g. Assumption 6 (*ii*) is also known as a link condition and commonly used in the analysis of inverse problems (see, *e.g.* Chen and Reiß [2011]). Under this link condition, we show in the proof of the following corollary that Assumption 6 (*iii*) implies Assumption 5 (*iv*).

COROLLARY 3.2. *Let Assumptions 1– 4, 5 (i)–(iii), and 6 hold true. Then* $\|\widehat{\varphi}_n - \varphi\|_X^2$ *and* $\|\widetilde{\varphi}_n - \varphi\|_X^2$ *are of the order*

$$O_p\Big(\max\Big(m_n^{-2\alpha/d_x}, \frac{m_n}{n}, \tau_{k_n}k_n^{-2\beta}\Big)\Big). \tag{3.6}$$

REMARK 3.2. In the *mildly ill-posed case* where $\tau_j \sim j^{-2t}$, $t \geqslant 0$, let $k_n \sim n^{1/(2t+2\beta+1)}$ and

$m_n \sim n^{d_x/(2\alpha+d_x)}$.[4] Hence, the rate in (3.6) coincides with

$$O_p\Big(\max\Big(n^{-(2t+2\beta)/(2t+2\beta+1)}, n^{-2\alpha/(2\alpha+d_x)}\Big)\Big)$$

which is $O_p(n^{-2\alpha/(2\alpha+d_x)})$ if $\alpha \leqslant d_x(t+\beta)$. In case of trigonometric basis functions, the operator T acts like integrating t–times which then automatically implies $\alpha = d_x(t+\beta)$ (cf. page 12 in Breunig and Johannes [2011]). Also, it holds $\|\widehat{g}_n - g\|^2_{\mathcal{G}} = O_p(n^{-2\beta/(2t+2\beta+1)})$ and in particular we have $k_n^2\mathcal{R}_n = O(n^{(2-2\beta)/(2t+2\beta+1)}) = o(1)$ if $\beta > 1$. In the *severely ill-posed case* where $\tau_j \sim \exp(-j^{2t})$, $t > 0$, we let $k_n \sim (\log n)^{1/2t}$ and obtain the rate

$$O_p\Big(n^{-2\alpha/(2\alpha+d_x)}\Big).$$

In this case, $\|\widehat{g}_n - g\|^2_{\mathcal{G}} = O_p(\log(n)^{-2\beta/t})$ and in particular, $k_n^2\mathcal{R}_n = O(\log(n)^{(2-2\beta)/t}) = o(1)$ again if $\beta > 1$. We conclude that under mild conditions on the smoothness of φ the optimal nonparametric rate of regression in mean squared error is obtained. □

The following example illustrates that Assumption 6 (*iii*) can be justified when $\{e_j\}_{j\geqslant 1}$ coincides with Legendre polynomials. Similarly to this example, Assumption 6 (*iii*) can also be motivated for B-splines (cf. De Boor [1978]).

Example 3.1. Assume that $\mathcal{Y}$ is contained in $[-1,1]$ and consider the Hilbert space $\mathcal{G} = L^2_{[-1,1]}$ endowed with the usual norm $\|\phi\|^2_{\mathcal{G}} = \int_{-1}^{1} |\phi(y)|^2 dy/2$. Let $\{e_j\}_{j\geqslant 1}$ be the Legendre polynomials. That is, for $y \in [-1,1]$ we define $e_1 = 1$, $e_2(y) = y$, and

$$j\, e_{j+1}(y) = (2j-1)y\, e_j(y) - (j-1)e_{j-1}(y)$$

for $j \geqslant 2$. This recursion formula is equivalent to

$$y\, e_j(y) = \frac{j\, e_{j+1}(y) + (j-1)e_{j-1}(y)}{2j-1}$$

We also have $\langle e_j, e_l\rangle_{\mathcal{G}} = 2/(2j-1)$ whenever $j = l$ and zero otherwise which implies

$$\begin{pmatrix} \langle M_{id}e_1, e_1\rangle_{\mathcal{G}} & \tau_1^{1/2}\tau_2^{-1/2}\langle M_{id}e_1, e_2\rangle_{\mathcal{G}} & \dots \\ \tau_2^{1/2}\tau_1^{-1/2}\langle M_{id}e_2, e_1\rangle_{\mathcal{G}} & \langle M_{id}e_2, e_2\rangle_{\mathcal{G}} & \dots \\ \vdots & \vdots & \ddots \end{pmatrix} = \begin{pmatrix} 0 & \frac{2\tau_1^{1/2}}{3\tau_2^{1/2}} & 0 & 0 & \dots \\ \frac{2\tau_2^{1/2}}{3\tau_1^{1/2}} & 0 & \frac{4\tau_2^{1/2}}{15\tau_3^{1/2}} & 0 & \dots \\ 0 & \frac{4\tau_3^{1/2}}{15\tau_2^{1/2}} & 0 & \ddots & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

The norm of the right hand side matrix is bounded by its Frobenius norm which is

$$\begin{aligned} \Big\|\big(\tau_j^{1/2}\tau_l^{-1/2}\langle M_{id}e_j, e_l\rangle_{\mathcal{G}}\big)_{j,l\geqslant 1}\Big\|_F^2 &= \sum_{j\geqslant 1}\left(\frac{\tau_j}{\tau_{j+1}}\frac{j^2}{(2j-1)^2}\frac{4}{(2j+1)^2} + \frac{\tau_{j+1}}{\tau_j}\frac{(j-1)^2}{(2j-1)^2}\frac{4}{(2j-1)^2}\right) \\ &\leqslant 2\sum_{j\geqslant 1}\frac{\tau_j}{\tau_{j+1}}\frac{j^2}{(2j-1)^2}\frac{4}{(2j-1)^2} \end{aligned}$$

[4] For two sequences of positive integers $\{a_n\}_{n\geqslant 1}$ and $\{b_n\}_{n\geqslant 1}$, $a_n \sim b_n$ means there exist two constants $0 < c, C < \infty$ such that $ca_n \leqslant b_n \leqslant Cb_n$.

which is bounded provided $\sup_{j\geqslant 1}\left\{\tau_j/\tau_{j+1}\right\} = O(1)$. The condition $\sup_{j\geqslant 1}\left\{\tau_j/\tau_{j+1}\right\} = O(1)$ is satisfied in the *mildly ill-posed case* $\tau_j \sim j^{-2t}$ for all $t \geqslant 0$. In the *severely ill-posed case* where $\tau_j \sim \exp(-j^{2t})$, $t > 0$, this condition is satisfied for all $0 \leqslant t \leqslant 1/2$ as $\sup_{j\geqslant 1}\left\{\tau_j/\tau_{j+1}\right\} = \exp\left((j+1)^{2t} - j^{2t}\right) \leqslant \exp(1)$ for $t \leqslant 1/2$. In both cases, Assumption 6 (*iii*) holds true. As it is shown in the proof of Corollary 3.2, under the link condition as stated in Assumption 6 (*ii*), this implies that Assumption 5 (*iv*) is also satisfied. □

3.2. Pointwise and Uniform Inference for the Regression Function

This subsection is about inference on the regression function φ both pointwise, that is for φ evaluated at some point, and uniform over the support of X. We first analyze the pointwise asymptotic distribution of our estimators at some $x \in \mathcal{X}$. For the estimator $\widehat{\varphi}_n(x)$ we introduce the sieve variance formula

$$\mathcal{V}_{1n}(x) = f_{\underline{m_n}}(x)^t Q_n^{-1}\, \mathbf{E}\left[f_{\underline{m_n}}(X)\, \mathbb{V}\mathrm{ar}(Yg(Y)|X) f_{\underline{m_n}}(X)^t\right] Q_n^{-1} f_{\underline{m_n}}(x),$$

where, here and in the following, we use the notation $Q_n = \mathbf{E}[f_{\underline{m_n}}(X) f_{\underline{m_n}}(X)^t]$. As mentioned earlier, for the reweighted estimator $\widetilde{\varphi}_n(x)$ we require a different normalization factor, namely

$$\mathcal{V}_{2n}(x) = f_{\underline{m_n}}(x)^t Q_n^{-1}\, \mathbf{E}\left[f_{\underline{m_n}}(X)\, \mathbf{E}[(Y-\varphi(X))^2 \Delta g^2(Y)|X] f_{\underline{m_n}}(X)^t\right] Q_n^{-1} f_{\underline{m_n}}(x).$$

In both cases, the sieve variance is increased by the multiplication of the inverse selection probability g and hence, results in larger variances than in the usual series regression. In contrast to our additional multiplicative term, Das et al. [2003] obtain for their sample selection estimator an additive term in the sieve variance, which is due to the estimation of a propensity score. The relation between the terms $\mathcal{V}_{1n}(x)$ and $\mathcal{V}_{2n}(x)$ is not straightforward to investigate. On the other hand, note that $\mathcal{V}_{1n}(x) \leqslant \mathcal{V}_{2n}(x)$ if and only if

$$2\varphi(X)\, \mathbf{E}[Yg^2(Y)|X] \leqslant \varphi^2(X)\left(1 + \mathbf{E}[\Delta g^2(Y)|X]\right).$$

We also emphasize that both normalization factors $\mathcal{V}_{1n}(x)$ and $\mathcal{V}_{2n}(x)$ are not affected by the ill-posed inverse problem of estimating g in the first step. This is in analogy to the rate of convergence results we obtain for the estimators $\widehat{\varphi}_n$ and $\widetilde{\varphi}_n$. In the next result we replace the variance $\mathcal{V}_{1n}(x)$ by the estimator

$$\widehat{\mathcal{V}}_{1n}(x) = n f_{\underline{m_n}}(x)^t (\mathbf{X}_{m_n}^t \mathbf{X}_{m_n})^- \sum_{i=1}^n f_{\underline{m_n}}(X_i)\left(Y_i\, \widehat{g}_n(Y_i) - \widehat{\varphi}_n(X_i)\right)^2 f_{\underline{m_n}}(X_i)^t (\mathbf{X}_{m_n}^t \mathbf{X}_{m_n})^- f_{\underline{m_n}}(x)$$

Moreover, we estimate the variance term of the reweighted estimator $\widetilde{\varphi}_n(x)$ by

$$\widehat{\mathcal{V}}_{2n}(x) = n f_{\underline{m_n}}(x)^t (\mathbf{X}_{m_n}^t \mathbf{D}_n \mathbf{X}_{m_n})^- \sum_{i=1}^n f_{\underline{m_n}}(X_i)\Delta_i\left(Y_i - \widetilde{\varphi}_n(X_i)\right)^2 \widehat{g}_n^2(Y_i) f_{\underline{m_n}}(X_i)^t (\mathbf{X}_{m_n}^t \mathbf{D}_n \mathbf{X}_{m_n})^- f_{\underline{m_n}}(x).$$

To establish the asymptotic distribution of our estimator we require the following additional assumptions. We introduce the notation $\mathbf{T}_n = \mathbf{E}[\Delta f_{\underline{m_n}}(X) e_{\underline{k_n}}(Y)^t]$ and $\mathbf{T}_n^Y = \mathbf{E}[Y f_{\underline{m_n}}(X) e_{\underline{k_n}}(Y)^t]$.

Assumption 7. *(i) Either* $\mathbf{E}[|Yg(Y)-\varphi(X)|^4|X] \leqslant C$ *and* $\mathbb{V}\mathrm{ar}(Yg(Y)|X) \geqslant c$ *or, in case of reweighting,* $\mathbf{E}[|Y-\varphi(X)|^4\Delta g^4(Y)|X] \leqslant C$ *and* $\mathbf{E}[|Y-\varphi(X)|^2\Delta g^2(Y)|X] \geqslant c$ *for some constants* $c, C > 0$. *(ii) The function* g *is uniformly bounded away from one. (iii) It holds* $\lambda_{\min}(\mathbf{T}_n^t Q_n^{-1} \mathbf{T}_n) \sim \tau_{k_n}$ *where* τ_{k_n} *is uniformly bounded away from zero.*

A bounded fourth moment of the error was also assumed by Newey [1997] to establish asymptotic normality of series estimators in the regression context. Assumption 7 *(iii)* restricts the minimum eigenvalue of $\mathbf{T}_n^t Q_n^{-1} \mathbf{T}_n$ to be bounded away from zero uniformly in n. A stronger restriction on $\lambda_{\min}(\mathbf{T}_n^t Q_n^{-1} \mathbf{T}_n)$ is made when this assumption is considered together with Assumption 6 *(ii)*. The consistency result established in Theorem 3.1 together with Assumption 7 *(iii)* imply that the constraint in (3.2) is not binding asymptotically and hence the estimator $\widehat{g}_n$ given in (3.3) coincides with the one in (3.2).

The next result establishes the asymptotic distribution of the estimators $\widehat{\varphi}_n$ and $\widetilde{\varphi}_n$ evaluated at some point x in the support of X.

THEOREM 3.3. *Let Assumptions 1 – 5 and 7 be satisfied. If for some $x \in \mathcal{X}$ it holds*

$$\sqrt{n} \max\Big((F_{m_n} \varphi - \varphi)(x),\, F_{m_n} TM_{id}(E_{k_n} g - g)(x) \Big) = o\Big(\sqrt{\mathcal{V}_n(x)} \Big) \quad \text{and} \quad k_n = o\Big(\tau_{k_n} \mathcal{V}_n(x) \Big) \tag{3.7}$$

with $\mathcal{V}_n(x)$ being equal to $\mathcal{V}_{1n}(x)$ and $\mathcal{V}_{2n}(x)$, respectively, then we have

$$\sqrt{n/\mathcal{V}_{1n}(x)} \Big(\widehat{\varphi}_n(x) - \varphi(x) \Big) \xrightarrow{d} \mathcal{N}(0,1) \quad \text{and} \quad \sqrt{n/\mathcal{V}_{2n}(x)} \Big(\widetilde{\varphi}_n(x) - \varphi(x) \Big) \xrightarrow{d} \mathcal{N}(0,1).$$

If, in addition, Assumption 6 (i)-(ii) is satisfied, $\mathcal{Y}$ is bounded, $m_n^2 = o(n)$, $\tau_{k_n} k_n^{-\beta} m_n = o(1)$ and $k_n^2 = o(\tau_{k_n} n)$ then

$$\sqrt{n/\widehat{\mathcal{V}}_{1n}(x)} \Big(\widehat{\varphi}_n(x) - \varphi(x) \Big) \xrightarrow{d} \mathcal{N}(0,1) \quad \text{and} \quad \sqrt{n/\widehat{\mathcal{V}}_{2n}(x)} \Big(\widetilde{\varphi}_n(x) - \varphi(x) \Big) \xrightarrow{d} \mathcal{N}(0,1).$$

The first part of condition (3.7) is an undersmoothing requirement to ensure that the sieve approximation biases become asymptotically negligible. It does not necessarily require undersmoothing of the estimator of the inverse selection probability function g. In the setting of Corollary 3.2, the first part of condition (3.7) is less restrictive than imposing $\max\left(n m_n^{-2\alpha/d_x}, n \tau_{k_n} k_n^{-2\beta} \right) = o(1)$ (see also Comment 4.3 in Belloni et al. [2015] for a discussion on such undersmoothing conditions). The rate restriction $k_n = o(\tau_{k_n} \mathcal{V}_n(x))$ ensures that the variance of the first step estimation of g does not enter the asymptotic distribution. Without this assumption we get a larger normalization factor due to an additional variance term. Moreover, the second result of the theorem requires the additional rate restriction $k_n^2 = o(\tau_{k_n} n)$. In the setting of Corollary 3.2 this is equivalent to $k_n^{1+t} = o(\sqrt{n})$ in the *mildly ill-posed case* and to $k_n^{2t} = o\Big(\log n \Big)$ in the *severely ill-posed case*. The rate restrictions $m_n^2 = o(n)$, $\tau_{k_n} k_n^{-\beta} m_n = o(1)$ and $k_n^2 = o(\tau_{k_n} n)$ are used to control the rates of $\|\widehat{\varphi}_n - \varphi\|_\infty$ and $\|\widehat{g}_n - g\|_\infty$. An assumption to control the sup norm rate in the estimation of the variance is also made in Chen and Christensen [2015a, Theorem A.1].

We now show how we can use a bootstrap procedure to construct uniform confidence bands for $\varphi(\cdot)$. Let $(\varepsilon_1, \ldots, \varepsilon_n)$ be a bootstrap sequence of i.i.d. random variables drawn independently of the data $\{(\Delta_1, Y_1, X_1), \ldots, (\Delta_n, Y_n, X_n)\}$, with $\mathbf{E}[\varepsilon_i] = 0$, $\mathbf{E}[\varepsilon_i^2] = 1$, $\mathbf{E}[\varepsilon_i^{2+\delta}] < \infty$ for all $1 \leqslant i \leqslant n$ and some $\delta \geqslant 1$. Common choices of distributions for ε_i include the standard Normal, Rademacher, and the two-point distribution of Mammen [1993]. Let $\mathbb{P}^*$ denote the probability distribution of the bootstrap innovations $(\varepsilon_1, \ldots, \varepsilon_n)$ conditional on the data. Let us define the bootstrap process

$$\mathbb{X}_n^*(x) = \frac{f_{\underline{m_n}}(x)^t (\mathbf{X}_{m_n}^t \mathbf{X}_{m_n}/n)^-}{\sqrt{\widehat{\mathcal{V}}_{1n}(x)}} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n f_{\underline{m_n}}(X_i)(Y_i \widehat{g}_n(Y_i) - \widehat{\varphi}_n(X_i)) \varepsilon_i \right).$$

Let $\Sigma_n = \mathbf{E}[f_{\underline{m}_n}(X) f_{\underline{m}_n}(X)^t (Yg(Y) - \varphi(X))^2]$ and let d_n be the standard deviation semimetric on $\mathcal{X}$ of the Gaussian Process $\mathbb{X}_n(x) = f_{\underline{m}_n}(x)^t \mathfrak{X}_n / \sqrt{\mathcal{V}_{1n}(x)}$ with $\mathfrak{X}_n \sim \mathcal{N}(0, \Sigma_n)$ defined as $d_n(x_1, x_2) = (\mathbf{E}[(\mathbb{X}_n(x_1) - \mathbb{X}_n(x_2))^2])^{1/2}$, see *e.g.* van der Vaart and Wellner [2000, Appendix A.2].

Moreover, let $N(\mathcal{X}, d_n, \varepsilon)$ denote the ε-entropy of $\mathcal{X}$ with respect to d_n and denote $\eta_n = \max\left(n^{-1/2}(k_n \tau_{k_n}^{-1/2} + m_n), m_n^{-\alpha/d_x}, (1 + \sqrt{m_n \tau_{k_n}})\, k_n^{-\beta}\right)$. In the following we assume $\xi_n = O(k_n)$ to simplify notation. We introduce the following assumption, which is similar to the assumptions required by Chen and Christensen [2015a] to establish the important result of validity of their bootstrap uniform confidence bands.

Assumption 8. *(i) $\mathcal{X}$ is compact and $(\mathcal{X}, d_n)$ is separable for each n. (ii) There exists a sequence of finite positive integers c_n such that*

$$1 + \int_0^\infty \sqrt{\log N(\mathcal{X}, d_n, \varepsilon)}\, d\varepsilon = O(c_n).$$

(iii) There exist two sequences of positive integers $b_{1,n}, b_{2,n}$ with $b_{j,n} = o(c_n^{-1})$ for $j = 1, 2$ such that $\sup_{x \in \mathcal{X}} \sqrt{n / \mathcal{V}_{1n}(x)}\, |\varphi(x) - F_{m_n} \varphi(x)| = O(b_{1,n})$ and $\sqrt{n m_n} \|\varphi - F_{m_n} \varphi\|_X = O(b_{2,n})$; (iv) There exists a sequence of positive integers r_n with $r_n = o(1)$ such that $m_n^{5/2} = o(r_n^3 \sqrt{n})$ and

$$m_n \sqrt{\frac{\max\left(\log(m_n), k_n\right)}{n \tau_{k_n}}} + \eta_n c_n + \sqrt{\frac{k_n}{\tau_{k_n} m_n}} + k_n^{-\beta} \max\left(\sqrt{m_n}, \sqrt{n \tau_{k_n}}\right) = o(c_n^{-1}).$$

The next theorem establishes the validity of the bootstrap for constructing uniform confidence bands for $\varphi(\cdot)$. The proof of the theorem is a slight modification of the proof of Chen and Christensen [2015a, Theorem B.1], it is based on strong approximation of a series process by a Gaussian process, and uses an anti-concentration inequality for the supremum of the approximating Gaussian process obtained in Chernozhukov et al. [2014]. The differences in our proof with respect to the proof of Chen and Christensen [2015a, Theorem B.1] are due to the fact that our two-step estimation problem is different than the nonparametric instrumental regression, with endogeneity, considered in Chen and Christensen [2015a]. For a nonparametric instrumental regression, also Horowitz and Lee [2012] proposed a boostrap procedure to construct uniform confidence bands. Their procedure is based on interpolation of joint confidence intervals. In the setting of nonparametric regression without selection, important results on uniform inference and bootstrap uniform confidence bands can be found in Belloni et al. [2015].

Theorem 3.4. *Let the assumptions of Theorem 3.3, Assumptions 6 and 8 hold. Moreover, we assume $k_n = o(\tau_{k_n} m_n)$ and $\sup_{x \in \mathcal{X}} \left(\|f_{\underline{m}_n}(x)^t \mathbf{T}_n^Y\|^2 / \mathcal{V}_{1n}(x) \right) = O(k_n / m_n)$. Then,*

$$\sup_{s \in \mathbb{R}} \left| \mathbb{P}\left(\sup_{x \in \mathcal{X}} \left| \frac{\sqrt{n}(\widehat{\varphi}_n(x) - \varphi(x))}{\sqrt{\widehat{\mathcal{V}}_{1n}(x)}} \right| \leqslant s \right) - \mathbb{P}^*\left(\sup_{x \in \mathcal{X}} \left| \mathbb{X}_n^*(x) \right| \leqslant s \right) \right| = o_p(1).$$

This theorem requires $\sup_{x \in \mathcal{X}} \left(\|f_{\underline{m}_n}(x)^t \mathbf{T}_n^Y\|^2 / \mathcal{V}_{1n}(x) \right) = O(k_n / m_n)$ which is not a very strong assumption since $f_{\underline{m}_n}(x)^t \mathbf{T}_n^Y$ is a k_n-vector and, under Assumption 7 *(i)*, $\mathcal{V}_{1n}(x) \geqslant c \|f_{\underline{m}_n}(x)\|$.

REMARK 3.3 (Uniform Confidence Bands based on $\widetilde{\varphi}_n$). Alternatively, we can construct a bootstrap process based on the reweighted estimator $\widetilde{\varphi}_n(\cdot)$ as follows:

$$\mathbb{X}^*_{2,n}(x) = \frac{f_{\underline{m}_n}(x)^t(\mathbf{X}^t_{m_n}\mathbf{D}_n\mathbf{X}_{m_n}/n)^-}{\sqrt{\widehat{\mathcal{V}}_{2n}(x)}}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n f_{\underline{m}_n}(X_i)(Y_i - \widetilde{\varphi}_n(X_i))\Delta_i\widehat{g}_n(Y_i)\varepsilon_i\right),$$

with bootstrap innovations $(\varepsilon_1, \ldots, \varepsilon_n)$ having the same properties as above. This bootstrap process can be used to construct valid uniform confidence bands for $\varphi(\cdot)$ and a result as in Theorem 3.4 holds with $\widehat{\varphi}_n(x)$, $\mathbb{X}^*_n(x)$ and $\widehat{\mathcal{V}}_{1n}(x)$ replaced by $\widetilde{\varphi}_n(x)$, $\mathbb{X}^*_{2,n}(x)$ and $\widehat{\mathcal{V}}_{2n}(x)$, respectively. □

4. Sample Selection with Endogenous Covariates

In many economic applications, it is necessary to correct for both sample selection of the dependent variable and endogeneity of (some) covariates. In this section, we show that, under the assumptions of Section 2, identification of the corresponding reduced form equation can be achieved. Under further conditions, which are common in the nonparametric instrumental variable literature, identification of the structural function is also obtained. An estimator of the nonparametric structural function is proposed, we establish its rate of convergence as well as its asymptotic distribution and we propose a bootstrap procedure to construct uniform confidence bands.

4.1. Model and Identification

In this section, we consider the instrumental variable model under selectively observed outcomes given by

$$Y^* = \psi(Z) + U \quad \text{where} \quad \mathbf{E}[U|X] = 0, \quad Y = \Delta Y^*, \tag{4.1}$$

Z is a d_z-vector of possibly endogenous regressors in the sense that $\mathbf{E}[U|Z] \neq 0$ and hence $\psi(Z)$ need not to coincide with $\mathbf{E}[Y^*|Z]$. Here, X is a vector of instruments used to identify the structural function ψ. The instrument X is also assumed to satisfy Assumption 1; that is, $\Delta \perp\!\!\!\perp X|Y^*$. An example is the estimation of Engel curves, where Y^* denotes budget share allocated to alcohol which is often not reported (see for instance the British "Family expenditure Survey") and Z is total expenditure. Expenditure is commonly thought of as endogenous and typically instrumented for with labor income X. In this case, the instrument certainly influences Y^* through Z but is unlikely to directly influence survey nonresponse. The reduced form equation of the structural model (4.1) is given by

$$\mathbf{E}[Y^*|X] = \mathbf{E}[\psi(Z)|X]$$

where the left hand side is not identified. By making use of equation (2.2), we obtain the reduced form

$$\mathbf{E}[Yg(Y)|X] = \mathbf{E}[\psi(Z)|X] \tag{4.2}$$

where the left hand side is identified under Assumptions 1–3. Thereby, L^2_Z completeness of the conditional distribution of Z given X ensures identification of the structural function ψ. In the following example, we see that Assumptions 1 and 2 are satisfied in a triangular model.

EXAMPLE 4.1. Let us rewrite model (4.1) in reduced form and additionally specify a selection equation. Then the assumption $\Delta \perp\!\!\!\perp X \,|\, Y^*$ is satisfied in the triangular model

$$Y^* = \mathbf{E}[\psi(Z)|X] + \varepsilon \quad \text{where } \mathbf{E}[\varepsilon|X] = 0$$
$$\Delta = \xi(Y^*, \eta)$$

and $\eta \perp\!\!\!\perp (X, \varepsilon)$. As in D'Haultfoeuille [2011] it can be argued that, under regularity conditions imposed on the distribution of (X, ε), Y^* is complete for X. □

4.2. The Estimator and its Rate of Convergence

In this section, we propose an estimator for the structural function ψ and derive its rate of convergence. For any $\phi \in L^2_Z$ we introduce the function $\varrho(\cdot, g, \phi) = \mathbf{E}[Yg(Y) - \phi(Z)|X = \cdot]$. The least squares estimator of $\varrho(\cdot, g, \phi)$ is given by

$$\widehat{\varrho_n}(\cdot, g, \phi) = f_{\underline{m_n}}(\cdot)^t \, (\mathbf{X}^t_{m_n} \mathbf{X}_{m_n})^- \sum_{i=1}^n \Big(Y_i \, g(Y_i) - \phi(Z_i)\Big) f_{\underline{m_n}}(X_i).$$

Let us now propose a plug-in minimum distance estimator of ψ which involves the estimator $\widehat{g_n}$ given in (3.2) of the inverse selection probability g. That is, we estimate ψ by

$$\widehat{\psi_n} = \arg\min_{\phi \in \Psi_n} \sum_{i=1}^n \widehat{\varrho_n}^2(X_i, \widehat{g_n}, \phi). \tag{4.3}$$

Here, we consider the linear sieve space $\Psi_n = \left\{\phi(\cdot) = \sum_{j=1}^{k_n} \beta_j p_j(\cdot) : \beta \in \mathbb{R}^{k_n}\right\}$ of dimension $k_n < \infty$ for some basis functions $\{p_j\}_{j \geqslant 1}$ in L^2_Z. In particular, we have the least squares solution

$$\widehat{\psi_n}(\cdot) = p_{\underline{k_n}}(\cdot)^t \, \widehat{\vartheta}_{k_n} \quad \text{and} \quad \widehat{\vartheta}_{k_n} = \left(\mathbf{Z}^t_{k_n} \mathbf{X}_{m_n} (\mathbf{X}^t_{m_n} \mathbf{X}_{m_n})^- \mathbf{X}^t_{m_n} \mathbf{Z}_{k_n}\right)^- \mathbf{Z}^t_{k_n} \mathbf{X}_{m_n} (\mathbf{X}^t_{m_n} \mathbf{X}_{m_n})^- \mathbf{X}^t_{m_n} \mathbf{G}_n \tag{4.4}$$

where $k_n \leqslant m_n$, $\mathbf{G}_n = \left(Y_1 \, \widehat{g_n}(Y_1), \ldots, Y_n \, \widehat{g_n}(Y_n)\right)^t$ and $\mathbf{Z}_{k_n} = (p_{\underline{k}}(Z_1), \ldots, p_{\underline{k}}(Z_n))^t$. Similarly to the previous section, we might consider a reweighted version of the estimator $\widehat{\psi}$ by replacing the empirical counterpart of $\mathbf{E}[f_{\underline{m_n}}(X) f_{\underline{m_n}}(X)^t]$ by the empirical counterpart of $\mathbf{E}[f_{\underline{m_n}}(X) f_{\underline{m_n}}(X)^t \Delta g(Y)]$. We do not consider such a reweighted version of $\widehat{\psi_n}$ explicitly in this paper due to the length of the paper and since such results can be derived from the previous analysis of $\widehat{\varphi_n}$ and the subsequent study of $\widehat{\psi_n}$. An alternative estimator is based on the observation that $\mathbf{E}[\rho(Y, Z, \Delta; g(\cdot), \psi(\cdot))|X] = 0$ with $\rho(Y, Z, \Delta; g(\cdot), \psi(\cdot)) = (\Delta g(Y) - 1, Yg(Y) - \psi(Z))^t$ (see also Remark 3.1 in the exogenous case). One might thus use the sieve minimum distance estimator of Chen and Pouzo [2012], which does not coincide with our proposed estimator $\widehat{\psi_n}$ in this case. Our proposed estimator $\widehat{\psi_n}$ is tailored to our selection model and does not suffer from the ill-posedness of recovering the inverse selection probability g, as we see below.

ASSUMPTION 9. *(i) We observe a sample $((\Delta_1, Y_1, Z_1, X_1), \ldots, (\Delta_n, Y_n, Z_n, X_n))$ of iid. copies of (Δ, Y, Z, X) where $Y = \Delta Y^*$ and $\mathbf{E}[U^2|X] < \infty$. (ii) There exists a constant $C \geqslant 1$ and a sequence of positive integers $(k_n)_{n \geqslant 1}$ satisfying $\sup_{z \in \mathcal{Z}} \|p_{\underline{k_n}}(z)\|^2 \leqslant C k_n$ such that $k_n^2/n = o(1)$. (iii) The smallest eigenvalue of $\mathbf{E}[p_{\underline{k}}(Z) p_{\underline{k}}(Z)^t]$ is bounded away from zero uniformly in k. (iv) For every $\psi \in L^2_Z$ there exists $\Pi_{k_n} \psi \in \Psi_n$ such that $\|\Pi_{k_n} \psi - \psi\|_\infty = O(k_n^{-\gamma/d_z})$ for some constant $\gamma > 0$.*

Let us introduce the linear conditional expectation operator $K : L_Z^2 \to L_X^2$ with $(K\phi)(\cdot) = \mathbf{E}[\phi(Z)|X = \cdot]$ for all $\phi \in L_Z^2$. We introduce the following assumption which ensures identification of the structural function ψ in model (4.1).

ASSUMPTION 10. *(i) For some $\alpha > 0$ and for every $\phi \in \Psi_n$ there exists $F_{m_n} K\phi \in \mathcal{F}_{m_n}$ such that $\|F_{m_n} K\phi - K\phi\|_\infty = O(m_n^{-\alpha/d_x})$. (ii) For every function $\phi \in L_Z^2$, $\mathbf{E}[\phi(Z)|X] = 0$ implies $\phi(Z) = 0$. (iii) There exists a sequence of non-increasing positive real numbers $(\kappa_j)_{j\geqslant 1}$ such that $\|K\phi\|_X^2 \leqslant C\sum_{j=1}^\infty \kappa_j \langle \phi, p_j\rangle_Z^2$ and $\|K\phi\|_X^2 \geqslant c\sum_{j=1}^\infty \kappa_j \langle \phi, p_j\rangle_Z^2$ for all $\phi \in L_Z^2$ and some constants $c, C > 0$.*

The next result establishes the rate of convergence of the estimator $\widehat{\psi}_n$.

THEOREM 4.1. *Let Assumptions 1–5, 9 and 10 hold true. Then we have*

$$\|\widehat{\psi}_n - \psi\|_Z^2 = O_p\Big(\max\Big(k_n^{-2\gamma/d_z}, \frac{k_n}{n\kappa_{k_n}}, \kappa_{k_n}^{-1}\|T(E_{k_n} g - g)\|_X^2\Big)\Big).$$

In contrast to Theorem 3.1, the additional bias due to sample selection is also effected by the potential ill-posedness coming from endogeneity of covarites Z. Under the conditions of Corollary 3.2, the bias $\kappa_{k_n}^{-1}\|T(E_{k_n} g - g)\|_X^2$ can be bounded by $\kappa_{k_n}^{-1}\tau_{k_n} k_n^{-2\beta}$. Thereby, the usual rate in nonparametric instrumental regression (see Chen and Reiß [2011]) can be only obtained if $\tau_{k_n} k_n^{2\gamma/d_z} \leqslant const.\ \kappa_{k_n} k_n^{2\beta}$ for all n sufficiently large. In particular, we see that the rate of convergence derived in Theorem 4.1 does not suffer from estimation of g in an inverse problem.

REMARK 4.1. To conclude this section it is worth to mention that with our estimation method we can deal with another type of endogeneity, different from the one just considered. Suppose that the random vector X that satisfies Assumption 1 is endogenous, in the sense that the relationship of interest is the structural function ψ satisfying

$$Y^* = \psi(X) + U \quad \text{where} \quad \mathbf{E}[U|X] \neq 0 \text{ and } Y = \Delta Y^*. \tag{4.5}$$

This situation can be easily dealt with by assuming that there exists another vector W of instruments such that $\mathbf{E}[U|W] = 0$ and $\Delta \perp\!\!\!\perp X|(Y^*, W)$. The latter assumption replaces Assumption 1 and corresponds to the one in Remark 2.1. For simplicity, we assume that W is observed for all the individuals so that the conditional distribution of $X|W$ is identified from the data. Moreover, we have to assume that $\mathbb{P}(\Delta = 1|Y^*, W) > 0$ and that Assumption 2 holds with $\phi(Y^*)$ replaced by $\phi(Y^*, W)$. Then ψ is identified through (4.5) and

$$\mathbf{E}[Y^*|W] = \mathbf{E}[Y^*\mathbb{P}(\Delta = 1|Y^*, W)g(Y^*, W)|W] = \mathbf{E}[Y^*\Delta g(Y^*, W)|W] = \mathbf{E}[Y g(Y, W)|W].$$

Consequently, we obtain the identified reduced form equation

$$\mathbf{E}[Y g(Y, W)|W] = \mathbf{E}[\psi(X)|W]$$

and identification of ψ follows as above. □

4.3. Pointwise and Uniform Inference for the Structural Function

This subsection is about inference on the structural function ψ evaluated at some point and also uniform over the support of Z. We first analyze the pointwise asymptotic distribution of our estimator at some $x \in \mathcal{X}$. For the estimator $\widehat{\psi}_n(z)$ we introduce the sieve variance

$$\mathcal{W}_n(z) = p_{\underline{k_n}}(z)^t A_n \mathbf{E}\Big[f_{\underline{m_n}}(X)\,\mathbb{V}\mathrm{ar}(Y g(Y) - \psi(Z)|X) f_{\underline{m_n}}(X)^t\Big] A_n^t p_{\underline{k_n}}(z)$$

where $A_n = (\mathbf{K}_n^t Q_n^{-1} \mathbf{K}_n)^{-1} \mathbf{K}_n^t Q_n^{-1}$ with $\mathbf{K}_n = \mathbf{E}[f_{\underline{m_n}}(X) p_{\underline{k_n}}(Z)^t]$. If there is no endogenous selection (that is, $g = 1$) then the sieve variance coincides with the one obtained by Chen and Pouzo [2015] in nonparametric instrumental regression. Under endogenous selection, however, we see that the sieve variance increases relative to the inverse probability function g. Again we observe that the variance formula is not affected by the potential ill-posedness of the inverse problem to estimate the function g, i.e., $Tg = 1$. We replace the variance $\mathcal{W}_n(z)$ by the estimator

$$\widehat{\mathcal{W}}_n(z) = p_{\underline{k_n}}(z)^t \widehat{A}_n \frac{1}{n} \sum_{i=1}^n f_{\underline{m_n}}(X_i) \Big(Y_i \widehat{g}_n(Y_i) - \widehat{\psi}_n(Z_i) \Big)^2 f_{\underline{m_n}}(X_i)^t \widehat{A}_n^t p_{\underline{k_n}}(z),$$

where $\widehat{A}_n = \Big(\mathbf{Z}_{k_n}^t \mathbf{X}_{m_n} (\mathbf{X}_{m_n}^t \mathbf{X}_{m_n})^- \mathbf{X}_{m_n}^t \mathbf{Z}_{k_n} / n \Big)^- \mathbf{Z}_{k_n}^t \mathbf{X}_{m_n} (\mathbf{X}_{m_n}^t \mathbf{X}_{m_n})^-$. To establish the asymptotic distribution of our estimator we require the following additional assumptions.

Assumption 11. *(i) Let* $\mathbf{E}[|Yg(Y) - \psi(Z)|^4 | X] \leqslant C$ *and* $\mathbb{V}\mathrm{ar}(Yg(Y) - \psi(Z)|X) \geqslant c$ *for some constants* $c, C > 0$. *(ii) It holds* $\lambda_{\min}(\mathbf{K}_n^t Q_n^{-1} \mathbf{K}_n) \sim \kappa_{k_n}$ *where* κ_{k_n} *is uniformly bounded away from zero.*

Since $\mathbf{E}[|Y^*|^2 | X] = \mathbf{E}[|Y|^2 g(Y) | X] \leqslant \mathbf{E}[|Yg(Y)|^2 | X]$, the Cauchy-Schwarz inequality and using twice the basic inequality $(a-b)^2 \geqslant a^2/2 - b^2$ yields

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(Yg(Y) - \psi(Z)|X) &\geqslant \mathbf{E}[|Yg(Y)|^2|X] - 2\sqrt{\mathbf{E}[|Yg(Y)|^2|X]}\sqrt{\mathbf{E}[\psi^2(Z)|X]} + \mathbf{E}[\psi^2(Z)|X] \\
&= \Big(\sqrt{\mathbf{E}[|Yg(Y)|^2|X]} - \sqrt{\mathbf{E}[\psi^2(Z)|X]} \Big)^2 \\
&\geqslant \mathbf{E}[|Y^*|^2|X]/2 - \mathbf{E}[\psi^2(Z)|X] \\
&\geqslant \mathbb{V}\mathrm{ar}(U|X)/4 - 3\,\mathbf{E}[\psi^2(Z)|X]/2,
\end{aligned}$$

where we also used model equation (4.1) for the last inequality. Thus, $\mathbb{V}\mathrm{ar}(Yg(Y) - \psi(Z)|X)$ is bounded from below by some constant $c > 0$ if $\mathbb{V}\mathrm{ar}(U|X)$ is sufficiently large, more precisely, if $\mathbb{V}\mathrm{ar}(U|X) \geqslant 2(2c + 3\,\mathbf{E}[\psi^2(Z)|X])$. On the other hand, assuming a bounded conditional fourth moment of the structural disturbance U ($\mathbf{E}[U^4|X] \leqslant const.$) implies $\mathbf{E}[|Yg(Y) - \psi(Z)|^4|X] \leqslant const.$ if the support of Y is bounded.

The next result establishes the asymptotic distribution of the estimator $\widehat{\psi}_n$ evaluated at some point z in the support $\mathcal{Z}$ of Z.

Theorem 4.2. *Let Assumptions 1 – 5, 7 (ii), (iii), and 9 – 11 be satisfied. If for some* $z \in \mathcal{Z}$ *it holds*

$$\begin{aligned}
\sqrt{n} \max\Big((\Pi_{k_n}\psi - \psi)(z),\, p_{\underline{k_n}}(z)^t A_n\, \mathbf{E}[f_{\underline{m_n}}(X) T M_{id}(E_{k_n} g - g)(X)] \Big) &= o\Big(\sqrt{\mathcal{W}_n(z)} \Big) \\
\text{and} \quad p_{\underline{k_n}}(z)^t A_n \mathbf{T}_n^Y \Big(\mathbf{T}_n^t \mathbf{T}_n \Big)^{-1} \Big(A_n \mathbf{T}_n^Y \Big)^t p_{\underline{k_n}}(z) &= o\Big(\mathcal{W}_n(z) \Big), \quad (4.6)
\end{aligned}$$

then we have

$$\sqrt{n/\mathcal{W}_n(z)} \Big(\widehat{\psi}_n(z) - \psi(z) \Big) \xrightarrow{d} \mathcal{N}(0,1).$$

If, in addition, Assumption 6 (i)-(ii) is satisfied, $\mathcal{Y}$ *is bounded,* $k_n^2 = o\Big(n \min(\tau_{k_n}, \kappa_{k_n}^2) \Big)$, $\tau_{k_n} k_n^{1-\beta} = o(\kappa_{k_n})$ *and* $m_n \log k_n = o(n\kappa_{k_n})$ *then*

$$\sqrt{n/\widehat{\mathcal{W}}_n(z)} \Big(\widehat{\psi}_n(z) - \psi(z) \Big) \xrightarrow{d} \mathcal{N}(0,1).$$

The first part of (4.6) is an undersmoothing condition. It is less restrictive than imposing the assumption $\max\left(n\,\kappa_{k_n}k_n^{-2\gamma/d_z}, n\tau_{k_n}k_n^{-2\beta}\right) = o(1)$, which can be seen by using the lower bound $\mathcal{W}_n(z) \geqslant const.\kappa_{k_n}^{-1}\|p_{\underline{k}_n}(z)\|^2$ (see also the proof of Theorem 4.2). The second part of the condition (4.6) ensures that the variance of the first step estimation of g does not enter the asymptotic distribution. It can be always ensured by choosing the dimension parameter for $\widehat{g}_n$ appropriately smaller than the one for $\widehat{\psi}_n$.

We now show how we can use a bootstrap procedure to construct uniform confidence bands for $\psi(\cdot)$. Let $(\varepsilon_1,\ldots,\varepsilon_n)$ be a bootstrap sequence of i.i.d. random variables drawn independently of the data $\{(\Delta_1,Y_1,X_1,Z_1),\ldots,(\Delta_n,Y_n,X_n,Z_n)\}$ and satisfying the same moment conditions as in Section 3.2, and let $\mathbb{P}^*$ be defined similarly as in Section 3.2. Let us define the bootstrap process

$$\mathcal{Z}_n^*(z) = \frac{p_{\underline{k}_n}(z)^t\widehat{A}_n}{\sqrt{\widehat{\mathcal{W}}_n(z)}}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n f_{\underline{m}_n}(X_i)(Y_i\widehat{g}_n(Y_i)-\widehat{\psi}_n(Z_i))\varepsilon_i\right). \tag{4.7}$$

Let $\Sigma_n^{\psi} = \mathbf{E}[f_{\underline{m}_n}(X)f_{\underline{m}_n}(X)^t(Yg(Y)-\psi(Z))^2]$ and let $\widetilde{d}_n$ be the standard deviation semimetric on $\mathcal{Z}$ of the Gaussian Process $\mathbb{Z}_n(z) = p_{k_n}(z)^tA_n\mathfrak{Z}_n/\sqrt{\mathcal{W}_n(z)}$ with $\mathfrak{Z}_n \sim \mathcal{N}(0,\Sigma_n^{\psi})$ defined as $\widetilde{d}_n(z_1,z_2) = (\mathbf{E}[(\mathbb{Z}_n(z_1)-\mathbb{Z}_n(z_2))^2])^{1/2}$. Further, we define

$$\widetilde{\eta}_n = m_n\sqrt{\frac{\log k_n}{n\kappa_{k_n}}} + \frac{k_n}{\sqrt{n\min(\tau_{k_n},\kappa_{k_n}^2)}} + k_n^{-\gamma/d_z} + k_n^{-\beta}\left(1+\sqrt{k_n\tau_{k_n}/\kappa_{k_n}}\right).$$

We introduce the following assumption which is similar to Chen and Christensen [2015a, Assumption 6].

Assumption 12. *(i) $\mathcal{Z}$ is compact and $(\mathcal{Z},\widetilde{d}_n)$ is separable for each n. (ii) There exists a sequence of finite positive constants $\widetilde{c}_n$ such that*

$$1+\int_0^{\infty}\sqrt{\log N(\mathcal{Z},\widetilde{d}_n,\varepsilon)}d\varepsilon = O(\widetilde{c}_n).$$

(iii) There exist two sequences of positive integers $\widetilde{b}_{1,n},\widetilde{b}_{2,n}$ with $\widetilde{b}_{j,n} = o(\widetilde{c}_n^{-1})$ for $j=1,2$ such that $\sup_{z\in\mathcal{Z}}\sqrt{n/\mathcal{W}_n(z)}|(\psi(z)-(\Pi_{k_n}\psi)(z))| = O(\widetilde{b}_{1,n})$ and $\sqrt{nm_n}\|\psi-\Pi_{k_n}\psi\|_{\mathcal{Z}} = O(\widetilde{b}_{2,n})$. (iv) There exists a sequence of positive constants r_n with $r_n = o(1)$ such that $m_n^2\sqrt{m_n} = o(r_n^3\sqrt{n})$ and

$$m_n\sqrt{\frac{\log k_n}{n\kappa_{k_n}\tau_{k_n}}} + \widetilde{\eta}_n c_n + \frac{k_n}{\kappa_{k_n}\inf_z\mathcal{W}_n(z)} + \left(\sqrt{m_n}+\sqrt{n\tau_{k_n}}\right)k_n^{-\beta} = o(\widetilde{c}_n^{-1}).$$

The next theorem establishes the validity of the bootstrap procedure for constructing uniform confidence bands for $\psi(\cdot)$. The proof of the theorem is a slight modification of the proof of Theorem 3.4 only because the bootstrap process we use here implies different rates of convergence to account for.

Theorem 4.3. *Let the assumptions of Theorem 4.2 and Assumption 12 hold. If $m_n\sqrt{\log k_n} = o(\sqrt{n\kappa_{k_n}})$ and $k_n\max(m_n,\xi_n^2) = o(n\tau_{k_n})$, then*

$$\sup_{s\in\mathbb{R}}\left|\mathbb{P}\left(\sup_{z\in\mathcal{Z}}\left|\frac{\sqrt{n}(\widehat{\psi}_n(z)-\psi(z))}{\sqrt{\widehat{\mathcal{W}}_n(z)}}\right|\leqslant s\right) - \mathbb{P}^*\left(\sup_{z\in\mathcal{Z}}\left|\mathcal{Z}_n^*(z)\right|\leqslant s\right)\right| = o_p(1).$$

5. Monte Carlo simulation

In this section, we study the finite sample performance of our estimators by presenting the results of a Monte Carlo simulation. There are 1000 Monte Carlo replications in each experiment and the sample size is $n = 1000$.

Regression with exogenous covariates. We consider estimation of the conditional expectation of Y^* given X. Let $X = \Phi(\chi)$ where $\chi \sim \mathcal{N}(0,1)$. Further, generate Y^* from the model

$$Y^* = \varphi(X) + c_V V$$

where $\varphi(x) = \Phi\big(8(x-0.5)\big)$ with standard normal distribution function Φ, $c_V = 0.4$, and $V \sim \mathcal{N}(0,1)$. We generate realizations of the selection variable Δ from

$$\Delta \sim \text{Binomial}(1, h(Y^*)), \tag{5.1}$$

where $h(y) = 0.4 * \mathbb{1}\{y \leqslant 0.4\} + \mathbb{1}\{y > 0.4\}$. Our estimators of the regression function φ are based on realizations of (Δ, Y, X) where $Y = \Delta Y^*$.

We estimate the function φ by using the series least squares plug-in estimator $\widehat{\varphi}_n$ given in (3.4) and the reweighted series least squares plug-in estimator $\widetilde{\varphi}_n$ given in (3.5). In both cases, we use B-splines as basis functions, either of order 3 with 2 knots (hence $k_n = 6$) or of order 3 with 4 knots (hence $k_n = 8$) and for the criterion function we use B-splines of order 3 with 6 knots (hence $m_n = 10$).

For the bootstrap uniform confidence bands, we consider one representative sample and generate the bootstrap innovations ε according to the two-point distribution suggested by Mammen [1993], i.e., ε equals $(1-\sqrt{5})/2$ with probability $(1+\sqrt{5})/(2\sqrt{5})$ and $(1+\sqrt{5})/2$ with probability $1-(1+\sqrt{5})/(2\sqrt{5})$. Based on the estimator $\widehat{\varphi}_n$ we generate the bootstrap process $\mathbb{X}_n^*$ as described in Subsection 3.2, while for $\widetilde{\varphi}_n$ we use the specific process described in Remark 3.3. The results are based on 1000 bootstrap iterations.

The first column of Figure 1 depicts the median of the estimators $\widehat{\varphi}_n$ and $\widetilde{\varphi}_n$ together with their 95% pointwise confidence bands and the median of the least squares series estimator under the missing at random (MAR) assumption based on listwise deletion. In the second column of Figure 1, we depict both estimators for a representative sample together with their bootstrap uniform confidence bands. As we see from Figure 1, MAR estimator becomes more biased for small values of x, as we expect, and, at least for small x, lies outside of the pointwise confidence intervals of $\widehat{\varphi}_n$.

From Figure 1 we also see that both estimators, $\widehat{\varphi}_n$ and $\widetilde{\varphi}_n$, have a similar finite sample performance in this setting. The conclusion changes, however, when the dimension parameter k_n is increased from 6 to 8, as depicted in Figure 2. As we see from this figure, the median of the estimators is accurate for $\widehat{\varphi}_n$ and $\widetilde{\varphi}_n$ but the variance of $\widehat{\varphi}_n$ is much larger for any x above 0.5. Hence in this particular setting, the estimator $\widetilde{\varphi}_n$ is much less sensitive than $\widehat{\varphi}_n$ to an accurate choice of k_n. This can be explained by the normalization of the inverse probability weights that is only performed for the estimator $\widetilde{\varphi}_n$.

Regression with endogenous covariates. In the following, we allow for endogeneity of covariates and aim to analyze the finite sample performance of our estimator of the

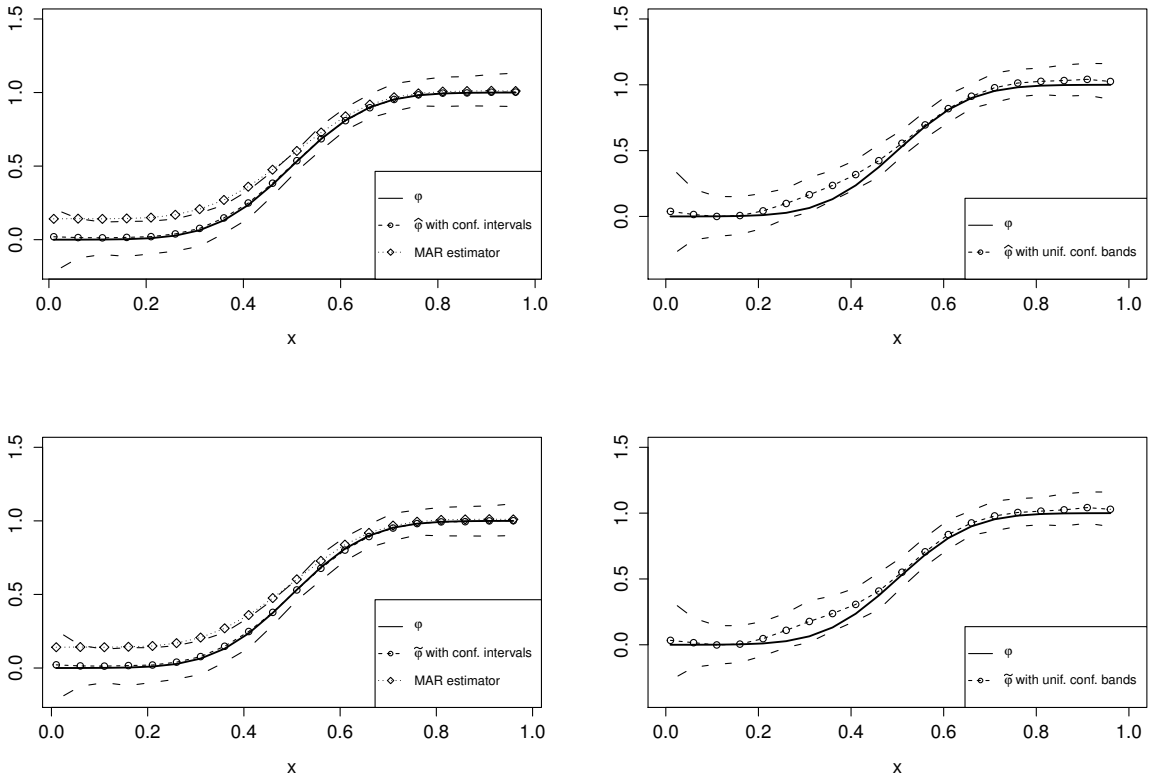

Figure 1: The first column shows the median of the estimators $\widehat{\varphi}_n$ and $\widetilde{\varphi}_n$ using $k_n = 6$, with their pointwise 95% confidence intervals and the median of an estimator under MAR assumption. The second column shows the estimators $\widehat{\varphi}_n$ and $\widetilde{\varphi}_n$ for a representative sample together with their uniform 95% confidence bands.

structural function ψ. Let $X = \Phi(\chi)$ and $Z = \Phi\big(\rho\chi + 0.4\varepsilon\big)$ where $\chi, \varepsilon \sim \mathcal{N}(0,1)$ independently and ρ is varied in the experiments. Further, generate Y^* from the model

$$Y^* = \psi(Z) + c_U U,$$

where $U = 0.3\,\varepsilon + 0.7\,\nu$ with $\nu \sim \mathcal{N}(0,1)$, $c_U = 0.4$, and $\psi(z) = \Phi\big(8\,(z - 0.5)\big)$. We generate the selection variable Δ as in equation (5.1). Our estimator of ψ is based on realizations of (Δ, Y, Z, X) where $Y = \Delta Y^*$.

We estimate the function ψ by using the nonparametric instrumental variable plug-in estimator $\widehat{\psi}_n$ proposed in (4.4). As basis functions we use B-splines of order 3 with 2 knots (hence $k_n = 6$) for the estimation of g and ψ. For the criterion function we use B-splines of order 3 with 6 knots (hence $m_n = 10$). For the bootstrap uniform confidence bands, we consider one representative sample and generate the bootstrap innovations ε according to the two-point distribution as described above. Based on the estimator $\widehat{\psi}_n$ we generate the bootstrap process $\mathbb{Z}_n^*$ given in (4.7). Again we use 1000 bootstrap iterations.

The first column of Figure 3 depicts the median of the estimator $\widehat{\psi}_n$ together with its 95% pointwise confidence intervals and the nonparametric instrumental variable estimator under the missing at random (MAR) assumption based on listwise deletion. In the second

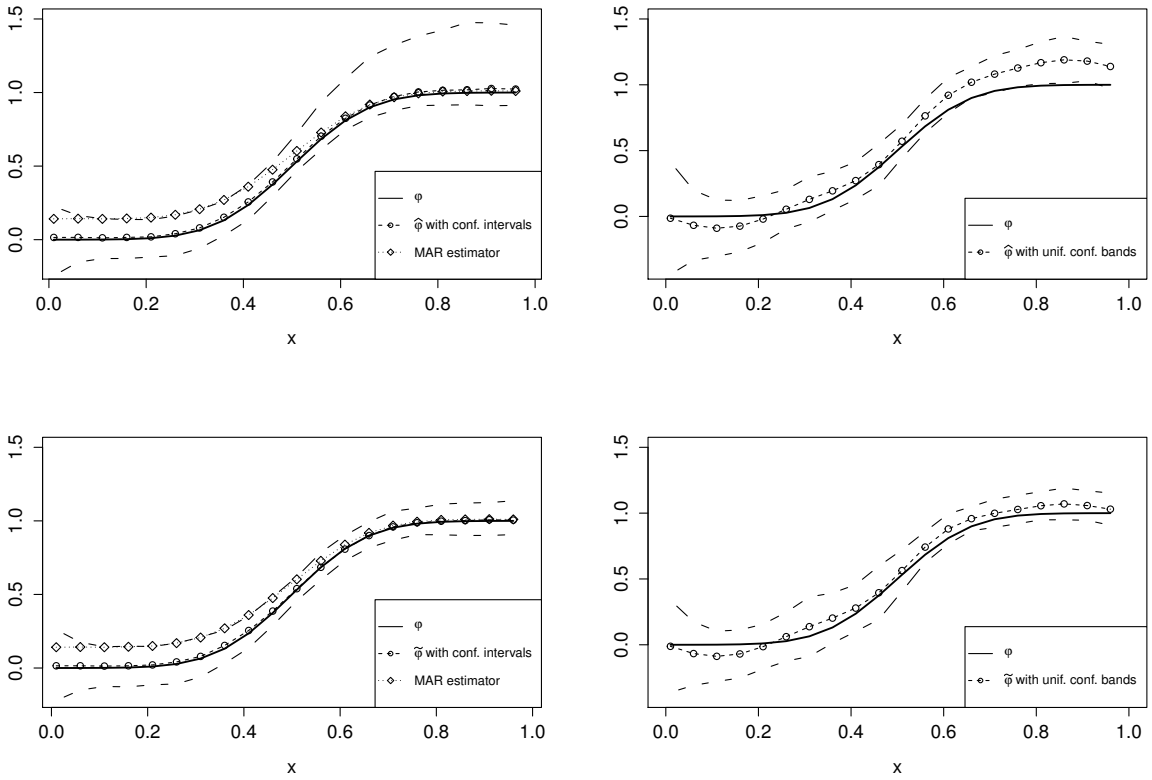

Figure 2: The first column shows the median of the estimators $\widehat{\varphi}_n$ and $\widetilde{\varphi}_n$ using $k_n = 8$, with their pointwise 95% confidence intervals and the median of an estimator under MAR assumption. The second column shows the estimators $\widehat{\varphi}_n$ and $\widetilde{\varphi}_n$ for a representative sample together with their uniform 95% confidence bands.

column of Figure 3, we depict $\widehat{\psi}_n$ together with its bootstrap uniform confidence bands for varying values of ρ. As we see from Figure 3, not surprisingly the confidence bands become wider as ρ decreases. In both cases, the confidence bands are also wider than in Figure 1 as we expect. In particular, the MAR does not lie outside of the 95% pointwise confidence intervals of ψ for any value of z.

6. Empirical Illustration

In this section, we apply our estimation procedure to study the way in which the level of expenditure of an individual affects his/her expected "propensity to work". We use data from the German Internet Panel (GIP)[5]. This data set contains data about individual attitudes and preferences which are relevant for political and economic decision-making processes. The survey represents the German speaking population aged 16 to 75 in Germany.

[5]This paper uses data from the German Internet Panel wave 4 (DOI: 10.4232/1.12610), (Blom et al. [2016]). A study description can be found in Blom et al. [2015]. The German Internet Panel is funded by the German Research Foundation through the Collaborative Research Center 884 "Political Economy of Reforms" (SFB 884).

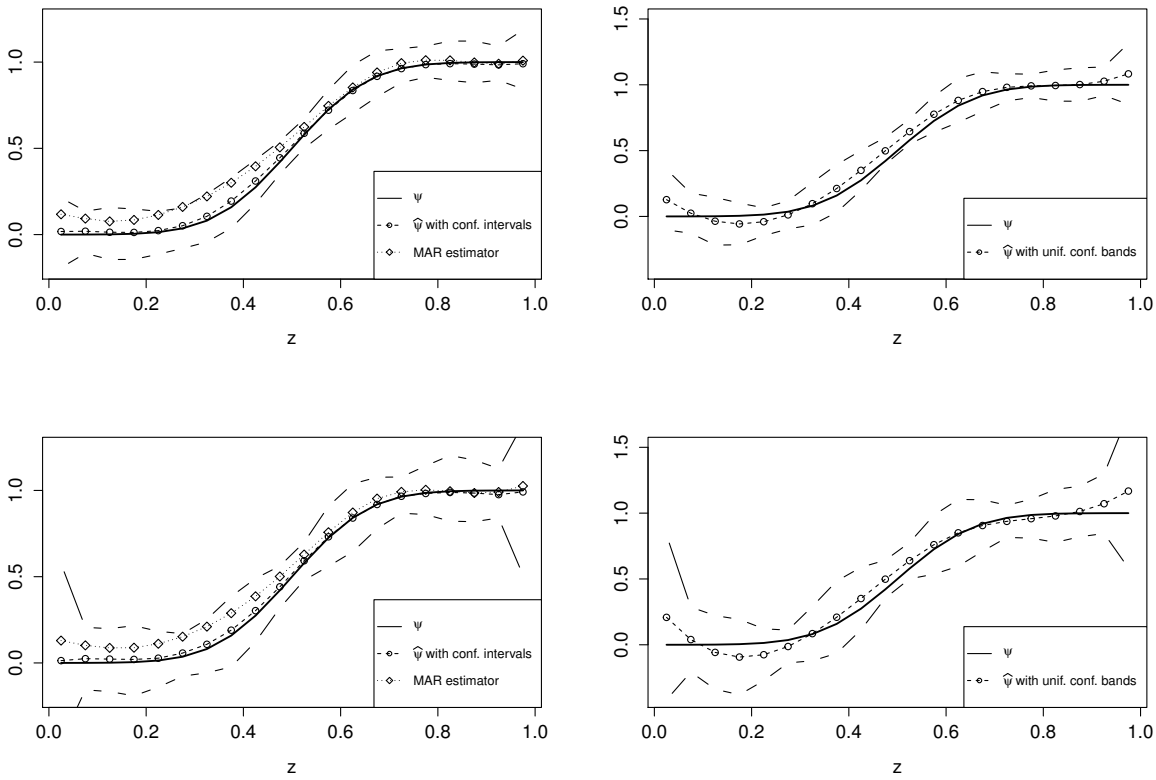

Figure 3: The first column shows the median of the estimator $\widehat{\psi}_n$ with their pointwise 95% confidence intervals and the median of an estimator under MAR assumption. In the first row, we choose $\rho = 0.8$ and in the second row, $\rho = 0.6$. The second column shows the estimator $\widehat{\psi}_n$ for a representative sample together with their uniform 95% confidence bands.

In our application we measure the "propensity to work" by the "number of desired hours" which is present in our data set. The latter variable is the number of weekly hours a person would like to work by taking into account that the income would change according to the hours of work. Let Y^* denote this variable and X denote the variable "expenditure". The latter measures the total average expenditure in one month of a person.

The object of interest in our study is the regression function of Y^* given X, that is, the expected number of desired hours given a level of monthly total expenditure.

We also restrict our sample to individuals that report a positive value of labor income. In our data set, we thus have 932 observations, a small number of missing values in the variable "expenditure" (33 observations) and a large number of missingness in the variable "number of desired hours" (332 missing observations). As the number of missing values in "expenditure" is small we eliminate these observations from our data set (since the bias is going to be negligible) so that the sample size we work with becomes $n = 899$ and the missing values in the variable "number of desired hours" are now 318.

The fact that Y^* is not observed is likely to be endogenous since one could think that in "extreme" situations, where Y^* is either excessively low or excessively high, a person

	Mean	*Median*	*Min*	*Max*	*Std.*	*missing*
$Y = \Delta Y^*$	21.91	30	0	60	18.089	318
X	1416	1200	0	6000	852.368	0

Table 1: Descriptive Statistics for $Y = \Delta Y^*$ and X.

would be more likely to not provide this information. While the variable "expenditure" is statistically related to Y^*, it is reasonable to assume that it is independent of the selection mechanism once Y^* is accounted for. In particular, as we restrict our sample to those individuals with positive labor income, we exclude individuals with high wealth that prefer not to work or unemployed individuals whose willingness to disclose information on their "propensity to work" might be driven by unobservables such as social pressure. We have also implemented our specification test proposed in the online supplementary material. We have computed the test statistics for a grid of values for m_n and the maximum value of the (standardized) test statistics is obtained for $m_n = 42$ and is 0.700. Therefore, our test fails to reject H_0 at the level 5%.

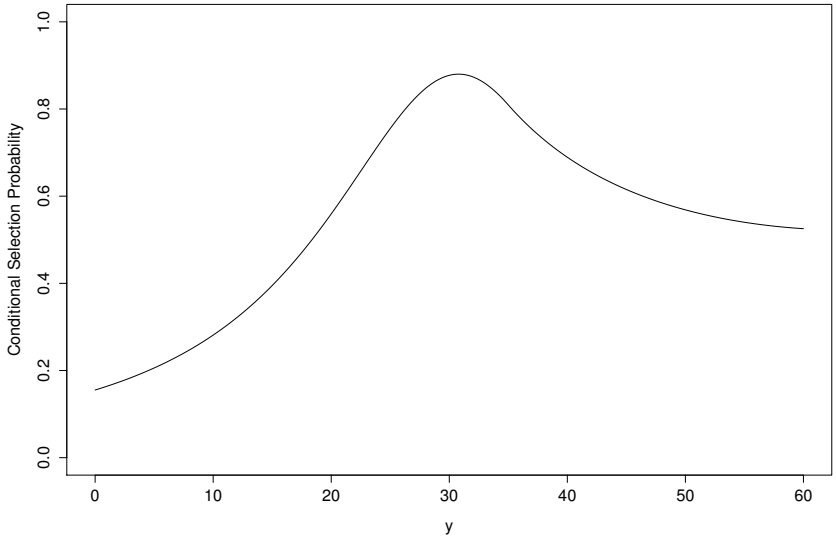

Figure 4: Graph of the estimator of $\mathbb{P}(\Delta = 1|Y^* = y) = 1/g(y)$.

Figure 4 depicts our estimator for the conditional probability $\mathbb{P}(\Delta = 1|Y^*)$, which is the inverse of the estimator $\widehat{g}_n$ as introduced in (3.2). We observe that this estimated probability of reporting increases with potential desired hours of working up to some point and decreases thereafter. We do not report uniform confidence bands here as they are too wide to draw any conclusion. As we emphasized in our theoretical analysis, the estimator of $\widehat{g}_n$ can be imprecise which we also see in finite samples.

Figure 5 shows the graphs of the estimators $\widehat{\varphi}_n$ and $\widetilde{\varphi}_n$ for the regression function of "number of desired hours" on "expenditure" together with the 90% percent bootstrap uniform confidence bands. The estimator is based on the nonparametric methodology

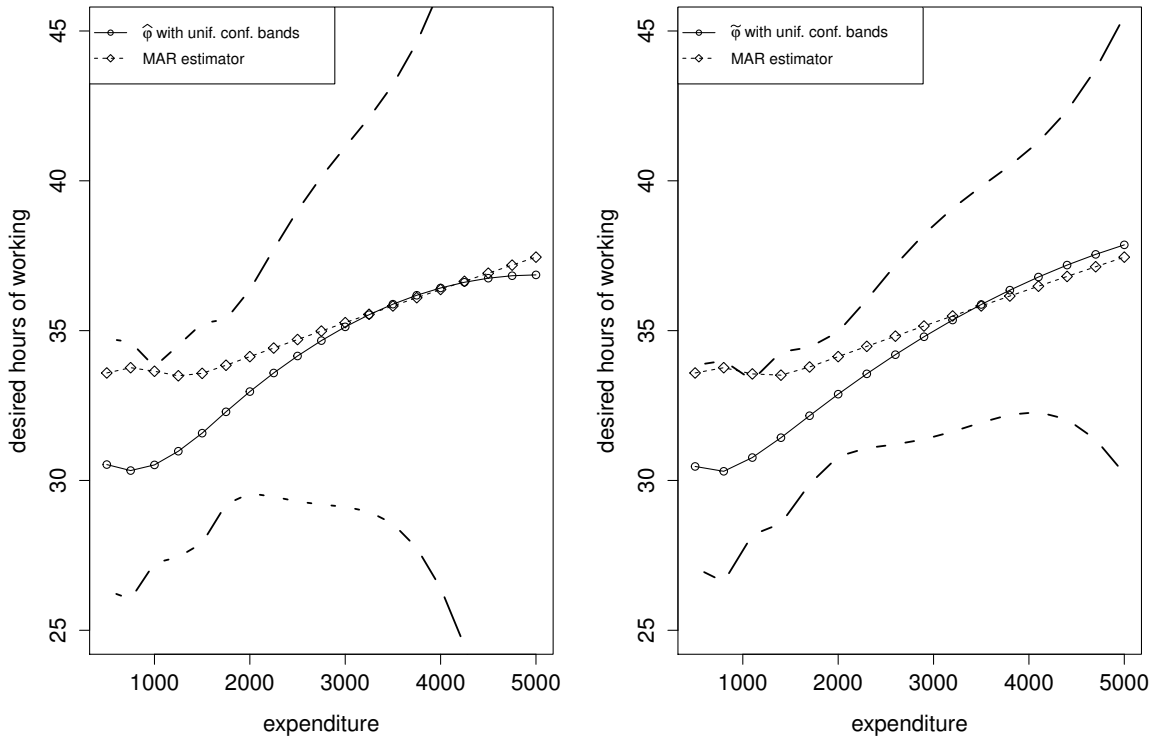

Figure 5: Regression curve of "number of desired hours" on "expenditure" using the estimators $\widehat{\varphi}_n$ and $\widetilde{\varphi}_n$ with 90% uniform confidence bands and series least squares estimator under the MAR assumption.

described in Section 3. The B-splines used here are of order 2 with 2 knots, hence $k_n = m_n = 5$. The uniform boostrap confidence bands are as described in Subsection 3.2 using bootstrap innovations ε generated by the two-point distribution as in the previous section and using 1000 bootstrap iterations. We also report the estimated regression function of "number of desired hours" on "expenditure" estimated under the MAR assumption. From Figure 5 we see that the reweighted estimator $\widetilde{\varphi}_n$ has smaller uniform confidence bands than the estimator $\widehat{\varphi}_n$. This is similar to our Monte Carlo simulation study in the previous section. One explanation is that estimation of the inverse probability function g is associated to a high level of variance as described above. We also see from Figure 5 that the MAR estimator is slightly beyond the uniform confidence bands of $\widehat{\varphi}_n$ when expenditure is around 1000 Euros.

A. Appendix

Throughout the proofs, we will use $C > 0$ to denote a generic finite constant that may be different in different uses. I_{m_n} denotes the $m_n \times m_n$ identity matrix. Further, for ease of notation we write $\sum_i$ for $\sum_{i=1}^n$. For a matrix A, we denote by $\|A\|$ its operator norm. In the following, we denote $\widehat{Q}_n = n^{-1}\sum_i f_{\underline{m_n}}(X_i)f_{\underline{m_n}}(X_i)^t$ and $\widetilde{Q}_n = n^{-1}\sum_i f_{\underline{m_n}}(X_i)f_{\underline{m_n}}(X_i)^t\Delta_i\widehat{g}_n(Y_i)$. By Assumption 4, the eigenvalues of $\mathbf{E}[f_{\underline{m_n}}(X)f_{\underline{m_n}}(X)^t]$ are bounded away from zero and hence, it may be assumed that $\mathbf{E}[f_{\underline{m_n}}(X)f_{\underline{m_n}}(X)^t] = I_{m_n}$. We also denote $\beta_{k_n} = (\mathbf{T}_n^t\mathbf{T}_n)^{-1}\mathbf{T}_n^t\,\mathbf{E}[f_{\underline{m_n}}(X)]$ and $E_{k_n}g(\cdot) = \mathbf{E}[g(Y)e_{\underline{k_n}}(Y)^t]e_{\underline{k_n}}(\cdot)$ where $\mathbf{T}_n = \mathbf{E}[\Delta f_{\underline{m_n}}(X)e_{\underline{k_n}}(Y)^t]$ and $\mathbf{T}_n^Y = \mathbf{E}[Yf_{\underline{m_n}}(X)e_{\underline{k_n}}(Y)^t]$.

A.1. Proofs of Section 3.

Proof of Theorem 3.1. The proof is based on the following decomposition

$$\begin{aligned}\|\widehat{\varphi}_n - \varphi\|_X^2 \leqslant 2\left\|\widehat{Q}_n^{-1}(I_{m_n} - \widehat{Q}_n)\Big(\mathbf{X}_{m_n}^t\mathbf{G}_n/n - \widehat{Q}_n\,\mathbf{E}[\varphi(X)f_{\underline{m_n}}(X)]\Big)\right\|^2 \\ + 4\left\|\mathbf{X}_{m_n}^t\mathbf{G}_n/n - \widehat{Q}_n\,\mathbf{E}[\varphi(X)f_{\underline{m_n}}(X)]\right\|^2 + 4\|F_{m_n}\varphi - \varphi\|_X^2 \\ = 2I_n + 4II_n + 4III_n \quad \text{(say)}. \quad \text{(A.1)}\end{aligned}$$

First observe that

$$I_n \leqslant \|\widehat{Q}_n^{-1}\|^2\|\widehat{Q}_n - I_{m_n}\|^2\|\mathbf{X}_{m_n}^t\mathbf{G}_n/n - \widehat{Q}_n\,\mathbf{E}[\varphi(X)f_{\underline{m_n}}(X)]\|^2.$$

We have $\|\widehat{Q}_n - I_{m_n}\|^2 = O_p\Big(n^{-1}m_n\log(m_n)\Big)$ see Lemma 2.1 of Chen and Christensen [2015b] and also $\|\widehat{Q}_n^{-1}\|^2 = 1 + o_p(1)$. Further, we observe

$$\begin{aligned}II_n &\leqslant 2\sum_{j=1}^{m_n}\Big|n^{-1}\sum_i\Big(Y_i\,g(Y_i) - (F_{m_n}\varphi(X_i)\Big)f_j(X_i)\Big|^2 \\ &+ 8\sum_{j=1}^{m_n}\Big|n^{-1}\sum_i Y_i\Big(\widehat{g}_n(Y_i) - (E_{k_n}g)(Y_i)\Big)f_j(X_i) - \langle TM_{id}(\widehat{g}_n - E_{k_n}g), f_j\rangle_X\Big|^2 \\ &+ 8\sum_{j=1}^{m_n}\Big|n^{-1}\sum_i Y_i\,(E_{k_n}g - g)(Y_i)f_j(X_i) - \langle TM_{id}(E_{k_n}g - g), f_j\rangle_X\Big|^2 \\ &+ 4\sum_{j=1}^{m_n}\langle TM_{id}(\widehat{g}_n - E_{k_n}g), f_j\rangle_X^2 + 4\sum_{j=1}^{m_n}\langle TM_{id}(E_{k_n}g - g), f_j\rangle_X^2 \\ &= 2A_1 + 8A_2 + 8A_3 + 4A_4 + 4A_5 \quad \text{(say)}.\end{aligned}$$

Consider A_1. Since $\mathbf{E}[(Y\,g(Y) - (F_{m_n}\varphi(X))f_j(X)] = 0$, $1 \leqslant j \leqslant m_n$, it holds

$$\mathbf{E}\,A_1 = n^{-1}\sum_{j=1}^{m_n}\mathbf{E}\,|f_j(X)Y\,g(Y)|^2 \leqslant 2n^{-1}\sup_{x\in\mathcal{X}}\|f_{\underline{m_n}}(x)\|^2\sup_{x\in\mathcal{X}}|\varphi(x)|\,\mathbf{E}\,|Y\,g(Y)| = O(m_n/n),$$

where we used $\mathbf{E}|Y\, g(Y)| = \mathbf{E}[\Delta g(Y^*)|Y^*|] = \mathbf{E}|Y^*| < \infty$. Consider A_2. The Cauchy Schwarz inequality implies

$$\begin{aligned} A_2 &\leqslant \left\|\widehat{\beta}_n - \mathbf{E}[g(Y)e_{\underline{k_n}}(Y)]\right\|^2 \sum_{j=1}^{m_n} \left\|n^{-1}\sum_i Y_i e_{\underline{k_n}}(Y_i) f_j(X_i) - \mathbf{E}[Y e_{\underline{k_n}}(Y) f_j(X)]\right\|^2 \\ &= \left\|\widehat{\beta}_n - \mathbf{E}[g(Y)e_{\underline{k_n}}(Y)]\right\|^2 O_p(k_n m_n/n) \\ &= O_p(m_n/n) \end{aligned}$$

where we used that $k_n \left\|\widehat{\beta}_n - \mathbf{E}[g(Y)e_{\underline{k_n}}(Y)]\right\|^2 = O_p(1)$. Consider A_3. Since $\|E_{k_n}g - g\|_\infty = o(1)$ we have

$$\begin{aligned} \mathbf{E}\, A_3 &\leqslant n^{-1}\sum_{j=1}^{m_n} \mathbf{E}\left|Y\,(E_{k_n}g - g)(Y) f_j(X)\right|^2 \\ &\leqslant n^{-1} \sup_{x\in\mathcal{X}} \|f_{\underline{m_n}}(x)\|^2 \|E_{k_n}g - g\|_\infty^2\, \mathbf{E}\, Y^2 \\ &= o(m_n/n), \end{aligned}$$

where $\mathbf{E}\, Y^2 < \infty$ holds due to assumption $\mathbf{E}[(Y^* - \varphi(X))^2|X] < \infty$ and since $g \in \mathcal{G}$. Consider A_4. We observe

$$\begin{aligned} A_4 = \|F_{m_n} T M_{id}(\widehat{g}_n - E_{k_n}g)\|_X^2 &\leqslant \|T M_{id}(\widehat{g}_n - E_{k_n}g)\|_X^2 \\ &\leqslant \sup_{\{\phi\in\mathcal{G}_n : \phi(\cdot)\geqslant 1\}} \left\{\frac{\|TM_{id}(\phi - E_{k_n}g)\|_X^2}{\|T(\phi - E_{k_n}g)\|_X^2}\right\} \|T(\widehat{g}_n - E_{k_n}g)\|_X^2 \end{aligned}$$

Due to Assumption 5 (*iv*) we have

$$\sup_{\{\phi\in\mathcal{G}_n : \phi(\cdot)\geqslant 1\}} \left\{\frac{\|TM_{id}(\phi - E_{k_n}g)\|_X^2}{\|T(\phi - E_{k_n}g)\|_X^2}\right\} < \infty.$$

From Blundell et al. [2007] page 1659 we deduce

$$\|T(\widehat{g}_n - E_{k_n}g)\|_X^2 = O_p\Big(m_n^{-2\alpha/d_x} + m_n/n + \|T(E_{k_n}g - g)\|_X^2\Big).$$

Consequently, we have

$$A_4 = O_p\Big(m_n^{-2\alpha/d_x} + m_n/n + \|T(E_{k_n}g - g)\|_X^2\Big).$$

Further,

$$A_5 \leqslant \|F_{m_n} T(E_{k_n}g - g)\|_X^2 = O\Big(\|T(E_{k_n}g - g)\|_X^2\Big),$$

which proves the rate of convergence of the estimator $\widehat{\varphi}_n$. To prove the rate result for $\widetilde{\varphi}_n$, we consider

$$\begin{aligned}\|\widetilde{Q}_n - I_{m_n}\|^2 \leqslant\, & 2\Big\|n^{-1}\sum_i f_{\underline{m_n}}(X_i)f_{\underline{m_n}}(X_i)^t\Delta_i\, g(Y_i) - \mathbf{E}[f_{\underline{m_n}}(X)f_{\underline{m_n}}(X)^t]\Big\|^2 \\ & + 4\Big\|n^{-1}\sum_i f_{\underline{m_n}}(X_i)f_{\underline{m_n}}(X_i)^t\Delta_i(g - E_{k_n}\, g)(Y_i)\Big\|^2 \\ & + 4\sum_{l=1}^{k_n}\Big\|n^{-1}\sum_i f_{\underline{m_n}}(X_i)f_{\underline{m_n}}(X_i)^t e_l(Y_i)\Delta_i\Big\|^2\, \big\|\widehat{\beta}_{k_n} - \mathbf{E}[g(Y)e_{\underline{k_n}}(Y)]\big\|^2 \\ & = O_p\Big(n^{-1}m_n\log(m_n) + \big\|\,\mathbf{E}[f_{\underline{m_n}}(X)f_{\underline{m_n}}(X)^t(T(g - E_{k_n}\, g))(X)]\big\|^2 \\ & \qquad + k_n^{-1}\max_{1\leqslant l\leqslant k_n}\big\|\,\mathbf{E}[f_{\underline{m_n}}(X)f_{\underline{m_n}}(X)^t e_l(Y)\Delta]\big\|^2\Big)\end{aligned}$$

where in the last line, we used that $k_n^2\,\|\widehat{\beta}_{k_n} - \mathbf{E}[g(Y)e_{\underline{k_n}}(Y)]\|^2 = O_p(1)$. It further holds $\max_{1\leqslant l\leqslant k_n}\|\,\mathbf{E}[f_{\underline{m_n}}(X)f_{\underline{m_n}}(X)^t e_l(Y)\Delta]\|^2 = O(1)$. Moreover, for $a_n \in \mathbb{R}^{m_n}$ with $a_n^t a_n = 1$ we obtain

$$\begin{aligned}\big\|\,\mathbf{E}[f_{\underline{m_n}}(X)f_{\underline{m_n}}(X)^t(T(g - E_{k_n}\, g))(X)]\big\|^2 &\leqslant \mathbf{E}[|a_n^t f_{\underline{m_n}}(X)(T(g - E_{k_n}\, g))(X)|^2] \\ &= O\big(\|g - E_{k_n}\, g\|_\infty^2\big) \\ &= o(1),\end{aligned}$$

and thereby the rate result for $\widetilde{\varphi}_n$ follows as above using the decomposition

$$\begin{aligned}\|\mathbf{X}_{m_n}^t\mathbf{G}_n/n - \widetilde{Q}_n\,\mathbf{E}[\varphi(X)f_{\underline{m_n}}(X)]\|^2 \leqslant & 2\Big\|n^{-1}\sum_i\Big(Y_i - F_{m_n}\varphi(X_i)\Big)\Delta_i g(Y_i)f_{\underline{m_n}}(X_i)\Big\|^2 \\ & + 2\Big\|n^{-1}\sum_i\Big(Y_i - F_{m_n}\varphi(X_i)\Big)\Big(\widehat{g}(Y_i) - g(Y_i)\Big)f_{\underline{m_n}}(X_i)\Big\|^2 \\ = & O_p\Big(m_n^{-2\alpha/d_x} + m_n/n + \|T(E_{k_n}g - g)\|_X^2\Big),\end{aligned}$$

where we used that $\mathbf{E}[(Y - (F_{m_n}\varphi(X))\Delta g(Y)f_j(X)] = 0$ for all $1 \leqslant j \leqslant m_n$. □

Proof of Corollary 3.2. It is sufficient to check that

$$\sup_{\phi\in\mathcal{G}_{k_n}}\frac{\|TM_{id}\phi\|_X}{\|T\phi\|_X} < \infty. \tag{A.2}$$

Let T^* denote the adjoint operator of T which is given by $(T^*\phi)(\cdot) = \mathbf{E}[\Delta\phi(X)|Y^* = \cdot]$. Since the multiplication operator M_{id} is a selfadjoint operator we obtain

$$\|TM_{id}\phi\|_X^2 = \langle TM_{id}\phi, TM_{id}\phi\rangle_X = \langle T\phi, (T^*)^{-1}M_{id}T^*TM_{id}\phi\rangle_X \leqslant \|T\phi\|_X\,\|(T^*)^{-1}M_{id}T^*TM_{id}\phi\|_X.$$

From the link condition $\|T\phi\|_X^2 \geqslant c\sum_{j=1}^\infty \tau_j\langle\phi, e_j\rangle_{\mathcal{G}}^2$ we infer by a duality argument $\|(T^*)^{-1}\phi\|_X^2 \leqslant c^{-1}\sum_{j=1}^\infty \tau_j^{-1}\langle\phi, e_j\rangle_{\mathcal{G}}^2$. Let L be a selfadjoint operator acting on $\mathcal{G}$ with eigenvalue decomposition $\{\tau_j^{1/2}, e_j\}_{j\geqslant 1}$. Then we conclude

$$\|TM_{id}\phi\|_X^2 \leqslant c^{-1}\|T\phi\|_X\,\|L^{-1}M_{id}T^*TM_{id}\phi\|_X \leqslant c^{-1}\|T\phi\|_X\,\|TM_{id}L^{-1}\|_X\,\|TM_{id}\phi\|_X$$

which gives

$$\frac{\|TM_{id}\phi\|_X}{\|T\phi\|_X} \leqslant c^{-1}\|TM_{id}L^{-1}\|_X.$$

Consequently, (A.2) follows since $\|TM_{id}L^{-1}\|_X$ is bounded. □

For the next proof, we recall and define the notations $\mathbf{T}_n = \mathbf{E}[\Delta f_{\underline{m_n}}(X)e_{\underline{k_n}}(Y)^t]$ and $\mathbf{T}_n^Y = \mathbf{E}[Yf_{\underline{m_n}}(X)e_{\underline{k_n}}(Y)^t]$. Further, we denote $\beta_{k_n} = \left(\mathbf{T}_n^t\mathbf{T}_n\right)^{-1}\mathbf{T}_n^t\,\mathbf{E}[f_{\underline{m_n}}(X)]$.

Proof of Theorem 3.3. To bound the sieve variance $\mathcal{V}_{1n}(x)$ from below we observe that assumption $\mathbb{V}\mathrm{ar}(Yg(Y)|X) \geqslant C$ yields

$$\mathcal{V}_{1n}(x) \geqslant f_{\underline{m_n}}(x)^t\,\mathbf{E}\left[f_{\underline{m_n}}(X)\,\mathbb{V}\mathrm{ar}(Yg(Y)|X)f_{\underline{m_n}}(X)^t\right]f_{\underline{m_n}}(x) \geqslant C\,\|f_{\underline{m_n}}(x)\|^2.$$

In the following, we also make use of

$$\begin{aligned}\mathbb{V}\mathrm{ar}(Yg(Y)-\varphi(X)|X) &= \mathbf{E}(|Yg(Y)-\varphi(X)|^2|X)\\ &= \mathbf{E}(|Yg(Y)|^2|X)-\varphi^2(X)\\ &= \mathbb{V}\mathrm{ar}(Yg(Y)|X).\end{aligned}$$

The proof is based on the decomposition

$$\begin{aligned}\widehat{\varphi}_n(x)-\varphi(x) &= f_{\underline{m_n}}(x)^t(n\widehat{Q}_n)^{-1}\sum_i f_{\underline{m_n}}(X_i)\Big(Y_ig(Y_i)-\varphi(X_i)\Big)\\ &+ f_{\underline{m_n}}(x)^t(n\widehat{Q}_n)^{-1}\sum_i f_{\underline{m_n}}(X_i)Y_i\Big(\widehat{g_n}(Y_i)-E_{k_n}g(Y_i)\Big)\\ &+ f_{\underline{m_n}}(x)^t(n\widehat{Q}_n)^{-1}\sum_i f_{\underline{m_n}}(X_i)Y_i\Big(E_{k_n}g(Y_i)-g(Y_i)\Big)\\ &+ f_{\underline{m_n}}(x)^t(n\widehat{Q}_n)^{-1}\sum_i f_{\underline{m_n}}(X_i)\varphi(X_i)-\varphi(x)\\ &= I_n+II_n+III_n+IV_n \quad \text{(say)}.\end{aligned} \tag{A.3}$$

Consider I_n. We obtain

$$\begin{aligned}\sqrt{n/\mathcal{V}_{1n}(x)}\,I_n &= \sum_i\Big(n\mathcal{V}_{1n}(x)\Big)^{-1/2}f_{\underline{m_n}}(x)^tf_{\underline{m_n}}(X_i)(Y_ig(Y_i)-\varphi(X_i))+o_p(1)\\ &= \sum_i s_{in}+o_p(1).\end{aligned}$$

Moreover, s_{in}, $1 \leqslant i \leqslant n$ satisfy the Lindeberg condition which can be seen as follows. It holds $\mathbf{E}[s_{in}] = 0$ and $n\,\mathbf{E}[s_{in}^2] = 1$. For all $\varepsilon > 0$ due to $\mathbf{E}\,|Yg(Y)-\varphi(X)|^4 \leqslant C$ we observe

$$\sum_i \mathbf{E}[s_{in}^2\,\mathbb{1}_{\{|s_{in}|>\varepsilon\}}] \leqslant n\varepsilon^2\,\mathbf{E}\,|s_{in}/\varepsilon|^4 \leqslant Cn^{-1}\varepsilon^{-2}m_n^2 = o(1).$$

Consider II_n. We observe

$$\begin{aligned}\sqrt{n}II_n &= \sqrt{n}\,f_{\underline{m_n}}(x)^t\mathbf{T}_n^Y(\widehat{\beta}_{k_n}-\beta_{k_n})+o_p(1)\\ &= n^{-1/2}\sum_i f_{\underline{m_n}}(x)^t\mathbf{T}_n^Y\Big(\mathbf{T}_n^t\mathbf{T}_n\Big)^{-1}\mathbf{T}_n^t\Big(f_{\underline{m_n}}(X_i)-\mathbf{E}[f_{\underline{m_n}}(X)]\Big)+o_p(1).\end{aligned}$$

Further, we have

$$\mathbf{E}\Big|n^{-1/2}\sum_i f_{\underline{m_n}}(x)^t\mathbf{T}_n^{\mathcal{Y}}\Big(\mathbf{T}_n^t\mathbf{T}_n\Big)^{-1}\mathbf{T}_n^t\Big(f_{\underline{m_n}}(X_i)-\mathbf{E}[f_{\underline{m_n}}(X)]\Big)\Big|^2 \leqslant f_{\underline{m_n}}(x)^t\mathbf{T}_n^{\mathcal{Y}}\Big(\mathbf{T}_n^t\mathbf{T}_n\Big)^{-1}(\mathbf{T}_n^{\mathcal{Y}})^t f_{\underline{m_n}}(x).$$

We obtain

$$\begin{aligned}
f_{\underline{m_n}}(x)^t\mathbf{T}_n^{\mathcal{Y}}\Big(\mathbf{T}_n^t\mathbf{T}_n\Big)^{-1}(\mathbf{T}_n^{\mathcal{Y}})^t f_{\underline{m_n}}(x) &\leqslant \Big\|\Big(\mathbf{T}_n^t\mathbf{T}_n\Big)^{-1}\Big\|\,\|f_{\underline{m_n}}(x)^t\mathbf{T}_n^{\mathcal{Y}}\|^2\\
&\leqslant C\tau_{k_n}^{-1}\sum_{l=1}^{k_n}\Big|\sum_{j=1}^{m_n} f_j(x)\,\mathbf{E}\Big[Y\Delta e_l(Y)f_j(X)\Big]\Big|^2\\
&= C\tau_{k_n}^{-1}\sum_{l=1}^{k_n}\Big|\sum_{j=1}^{m_n} f_j(x)\,\mathbf{E}\Big[\nu_l(X)f_j(X)\Big]\Big|^2\\
&\leqslant Ck_n\tau_{k_n}^{-1}\max_{1\leqslant l\leqslant k_n}\Big|(F_{m_n}\nu_l)(x)\Big|^2\\
&= O(k_n\tau_{k_n}^{-1}) \qquad\qquad (A.4)
\end{aligned}$$

where $\nu_l(x) = \mathbf{E}[Y\Delta e_l(Y)|X = x]$. Thus, the condition $k_n = o\Big(\tau_{k_n}\mathcal{V}_{1n}(x)\Big)$ yields $\sqrt{n}II_n = o_p\Big(\sqrt{\mathcal{V}_{1n}(x)}\Big)$. Consider III_n. We have

$$\begin{aligned}
&\mathbf{E}\Big|\,f_{\underline{m_n}}(x)^t n^{-1/2}\sum_i f_{\underline{m_n}}(X_i)Y_i\Big(E_{k_n}g(Y_i)-g(Y_i)\Big)\Big|^2\\
&\qquad\leqslant \mathbf{E}\Big[\Big|\,f_{\underline{m_n}}(x)^t f_{\underline{m_n}}(X)Y\Big(E_{k_n}g(Y)-g(Y)\Big)\Big|^2\Big] + n\Big(\,\mathbf{E}\;f_{\underline{m_n}}(x)^t f_{\underline{m_n}}(X)Y\Big(E_{k_n}g(Y)-g(Y)\Big)\Big)^2\\
&\qquad\leqslant \|E_{k_n}g-g\|_\infty^2\,\|\,f_{\underline{m_n}}(x)\|^2 + n|F_{m_n}TM_{id}(E_{k_n}g-g)(x)|^2
\end{aligned}$$

and thus, $\sqrt{n}III_n = o_p\Big(\sqrt{\mathcal{V}_{1n}(x)}\Big)$. Finally, $\sqrt{n}IV_n = o_p\Big(\sqrt{\mathcal{V}_{1n}(x)}\Big)$ follows from the condition $\sqrt{n}(F_{m_n}\varphi-\varphi)(x) = o\Big(\sqrt{\mathcal{V}_{1n}(x)}\Big)$, which completes the proof of the first statement in the theorem.

To prove $\sqrt{n/\mathcal{V}_{2n}(x)}\Big(\widetilde{\varphi}_n(x)-\varphi(x)\Big)\xrightarrow{d}\mathcal{N}(0,1)$ we make use of the decomposition

$$\begin{aligned}
\widetilde{\varphi}_n(x)-\varphi(x) &= f_{\underline{m_n}}(x)^t(n\widetilde{Q}_n)^{-1}\sum_i f_{\underline{m_n}}(X_i)\Big(Y_i-\varphi(X_i)\Big)\Delta_i g(Y_i)\\
&\quad + f_{\underline{m_n}}(x)^t(n\widetilde{Q}_n)^{-1}\sum_i f_{\underline{m_n}}(X_i)\Big(Y_i-\varphi(X_i)\Big)\Delta_i\Big(\widehat{g}_n(Y_i)-E_{k_n}g(Y_i)\Big)\\
&\quad + f_{\underline{m_n}}(x)^t(n\widetilde{Q}_n)^{-1}\sum_i f_{\underline{m_n}}(X_i)(Y_i-\varphi(X_i))\Delta_i\Big(E_{k_n}g(Y_i)-g(Y_i)\Big)\\
&\quad + (F_{m_n}\varphi)(x)-\varphi(x).
\end{aligned}$$

Hence, it is easily seen that the asymptotic normal distribution follows as above. Finally, by Lemma A.1 we see that the asymptotic distribution results remain valid if we replace $\mathcal{V}_{1n}(x)$ and $\mathcal{V}_{2n}(x)$ by their estimators. □

Lemma A.1. *Let Assumptions 1 – 5, 6 (i)-(ii) and 7 be satisfied. Moreover, assume that $\mathcal{Y}$ is bounded, $m_n^2 = o(n)$, $\tau_{k_n}k_n^{-\beta}m_n = o(1)$ and $k_n^2 = o(\tau_{k_n}n)$. Then, we have*

$$\mathcal{V}_{1n}(x)^{-1}\widehat{\mathcal{V}}_{1n}(x) = 1+o_p(1) \qquad \textit{and} \qquad \mathcal{V}_{2n}(x)^{-1}\widehat{\mathcal{V}}_{2n}(x) = 1+o_p(1)$$

uniformly in $x\in\mathcal{X}$.

Proof. We start by proving the first result. We denote $\Sigma_n = \mathbf{E}\left[f_{\underline{m_n}}(X)\,\mathbb{V}\mathrm{ar}(Yg(Y)|X)f_{\underline{m_n}}(X)^t\right]$, $\widehat{\Sigma}_n = n^{-1}\sum_i f_{\underline{m_n}}(X_i)f_{\underline{m_n}}(X_i)^t\Big(Y_i\,\widehat{g}_n(Y_i) - \widehat{\varphi}_n(X_i)\Big)^2$, and $\widetilde{\Sigma}_n = n^{-1}\sum_i f_{\underline{m_n}}(X_i)f_{\underline{m_n}}(X_i)^t\Big(Y_i\,g(Y_i) - \varphi(X_i)\Big)^2$. Moreover, let $\widehat{h}(x)^t = (\mathcal{V}_{1n}(x))^{-1/2}f_{\underline{m_n}}(x)^t\widehat{Q}_n^{-1}$, $h(x)^t = (\mathcal{V}_{1n}(x))^{-1/2}f_{\underline{m_n}}(x)^t$. Hence, $\mathcal{V}_{1n}(x)\widehat{h}(x)^t\widehat{\Sigma}_n\widehat{h}(x) = \widehat{\mathcal{V}}_{1n}(x)$ and by noticing that $h(x)^t\Sigma_n h(x) = 1$ since $Q_n = I_{m_n}$, the triangle inequality gives

$$\left|(\mathcal{V}_{1n}(x))^{-1/2}\widehat{\mathcal{V}}_{1n}(x)(\mathcal{V}_{1n}(x))^{-1/2} - 1\right| \leqslant \left|\widehat{h}(x)^t(\widehat{\Sigma}_n - \widetilde{\Sigma}_n)\widehat{h}(x)\right| + \left|\widehat{h}(x)^t(\widetilde{\Sigma}_n - \Sigma_n)\widehat{h}(x)\right| + \left|\widehat{h}(x)^t\Sigma_n\widehat{h}(x) - h(x)^t\Sigma_n h(x)\right|. \quad \text{(A.5)}$$

Remark that, by the Rudelson's inequality (see *e.g.* Belloni et al. [2015, Lemma 6.2]) $\sup_{x\in\mathcal{X}}\|\widehat{h}(x) - h(x)\| = O_p(\sqrt{m_n\log(m_n)/n})$ and, in particular, $\sup_{x\in\mathcal{X}}\|\widehat{h}(x)\| = O_p(1)$. Under Assumptions 4 *(ii)-(iii)* and 7 *(i)*, we can show similarly as in Newey [1997] page 165 – 166 that

$$\sup_{x\in\mathcal{X}}\left|\widehat{h}(x)^t\Sigma_n\widehat{h}(x) - h(x)^t\Sigma_n h(x)\right| = O_p\Big(\sqrt{m_n\log(m_n)/n}\Big) \quad \text{and} \quad \sup_{x\in\mathcal{X}}\left|\widehat{h}(x)^t(\widetilde{\Sigma}_n - \Sigma_n)\widehat{h}(x)\right| = O_p\Big(\sqrt{m_n\log(m_n)/n}\Big). \quad \text{(A.6)}$$

Moreover, denote $\widehat{\Im}_n = n^{-1}\sum_i f_{\underline{m_n}}(X_i)f_{\underline{m_n}}(X_i)^t|Y_i\,g(Y_i) - \varphi(X_i)|$, $\Im_n = \mathbf{E}[f_{\underline{m_n}}(X)f_{\underline{m_n}}(X)^t|Y\,g(Y) - \varphi(X)|]$, $D_{\widehat{g}_n}(\cdot) = \widehat{g}_n(\cdot) - g(\cdot)$ and $D_{\widehat{\varphi}_n}(\cdot) = \widehat{\varphi}_n(\cdot) - \varphi(\cdot)$ and remark that $\mathbf{E}\|\widehat{\Im}_n - \Im_n\|^2 = O(m_n\log(m_n)/n)$ under Assumption 7 (i). Also denote the rates $\eta_{n,1} = \max\left(k_n^2/(\tau_{k_n}n), k_n^{-2\beta}\right)$ and $\eta_{n,2} = \max\left(m_n^{-2\alpha/d_x}, n^{-1}m_n^2, m_n\tau_{k_n}k_n^{-2\beta}\right)$. Hence,

$$\begin{aligned}
&\sup_{x\in\mathcal{X}}\left|\widehat{h}(x)^t(\widehat{\Sigma}_n - \widetilde{\Sigma}_n)\widehat{h}(x)\right| \\
&= \sup_{x\in\mathcal{X}}\left|n^{-1}\sum_i\left(\widehat{h}(x)^t f_{\underline{m_n}}(X_i)\right)^2\left\{\left(Y_i\,\widehat{g}_n(Y_i) - \widehat{\varphi}_n(X_i)\right)^2 - \left(Y_i\,g(Y_i) - \varphi(X_i)\right)^2\right\}\right| \\
&\leqslant \sup_{x\in\mathcal{X}}\left|n^{-1}\sum_i\left(\widehat{h}(x)^t f_{\underline{m_n}}(X_i)\right)^2\left(Y_i D_{\widehat{g}_n}(Y_i) - D_{\widehat{\varphi}_n}(X_i)\right)^2\right| \\
&\qquad + 2\sup_{x\in\mathcal{X}}\left|n^{-1}\sum_i\left(\widehat{h}(x)^t f_{\underline{m_n}}(X_i)\right)^2\Big(Y_i\,g(Y_i) - \varphi(X_i)\Big)\left(Y_i D_{\widehat{g}_n}(Y_i) - D_{\widehat{\varphi}_n}(X_i)\right)\right| \\
&\leqslant 2\left(\|M_{id}D_{\widehat{g}_n}\|_\infty^2 + \|D_{\widehat{\varphi}_n}\|_\infty^2\right)\sup_{x\in\mathcal{X}}\left|(\mathcal{V}_n(x))^{-1}f_{\underline{m_n}}(x)^t\widehat{Q}_n^{-1}f_{\underline{m_n}}(x)\right| \\
&\qquad + 2\left(\|M_{id}D_{\widehat{g}_n}\|_\infty + \|D_{\widehat{\varphi}_n}\|_\infty\right)\sup_{x\in\mathcal{X}}\frac{1}{n}\sum_i\left(\widehat{h}_1(x)^t f_{\underline{m_n}}(X_i)\right)^2\left|Y_i\,g(Y_i) - \varphi(X_i)\right| \\
&\leqslant O_p\left(\eta_{n,1} + \eta_{n,2}\right) + O_p\left((\eta_{n,1} + \eta_{n,2})^{1/2}\right)\sup_{x\in\mathcal{X}}\left|\widehat{h}(x)^t[\widehat{\Im}_n - \Im_n]\widehat{h}(x) + \widehat{h}(x)^t\Im_n\widehat{h}_1(x)\right| \\
&= O_p\left((\eta_{n,1} + \eta_{n,2})^{1/2}\Big((\eta_{n,1} + \eta_{n,2})^{1/2} + \sqrt{\frac{m_n\log m_n}{n}} + 1\Big)\right) = O_p\left((\eta_{n,1} + \eta_{n,2})^{1/2}\right) \quad \text{(A.7)}
\end{aligned}$$

where in the third inequality we have used the facts that $|(\mathcal{V}_{1n}(x))^{-1} f_{\underline{m_n}}(x)^t \widehat{Q}_n^{-1} f_{\underline{m_n}}(x)| \leqslant (\mathcal{V}_{1n}(x))^{-1} \|f_{\underline{m_n}}(x)\|^2 (1 + o_p(1)) = O_p(1)$, $\|M_{id} D_{\widehat{g}_n}\|_\infty^2 = O_p(\eta_{n,1})$ and $\|D_{\widehat{\varphi}_n}\|_\infty^2 = O_p(\eta_{n,2})$, and the Cauchy Schwartz inequality. Then, by (A.5), (A.6), (A.7) and the assumptions of the lemma the result of the lemma follows. The proof of the second result of the Lemma proceeds similarly and we omit it. □

A.2. Proofs of Section 4.

Proof of Theorem 4.1. The proof is based on the inequality

$$\|\widehat{\psi}_n - \psi\|_Z \leqslant \|\Pi_{k_n}\psi - \psi\|_Z + \|\widehat{\psi}_n - \Pi_{k_n}\psi\|_Z.$$

By assumption, we have $\|\Pi_{k_n}\psi - \psi\|_Z = O(k_n^{-\gamma/d_z})$ and thus, it is sufficient to bound $\|\widehat{\psi}_n - \Pi_{k_n}\psi\|_Z$. By Lemma B.2 of Chen and Pouzo [2012] it holds $\|\widehat{\psi}_n - \Pi_{k_n}\psi\|_Z^2 \leqslant C\kappa_{k_n}^{-1}\|K(\widehat{\psi}_n - \Pi_{k_n}\psi)\|_X^2$. From the proof of Theorem 3.1 we have

$$\sum_{j=1}^{m_n} \left| n^{-1} \sum_i Y_i \widehat{g}_n(Y_i) f_j(X_i) - \mathbf{E}[Y g(Y) f_j(X)] \right|^2 = O_p(r_n) \tag{A.8}$$

where we denote $r_n = \max\left(m_n^{-2\alpha/d_x}, n^{-1}m_n, \|T(g - E_{k_n}g)\|_X^2\right)$. Consequently, we observe

$$\begin{aligned}\left\| n^{-1} \sum_i \left(Y_i \widehat{g}_n(Y_i) - (\Pi_{k_n}\psi)(Z_i)\right) f_{\underline{m_n}}(X_i) \right\|^2 &\leqslant 2 \left\| \mathbf{E}\left[\left(Y g(Y) - (\Pi_{k_n}\psi)(Z)\right) f_{\underline{m_n}}(X)\right] \right\|^2 + O_p(r_n) \\ &\leqslant 2\|K(\Pi_{k_n}\psi - \psi)\|_X^2 + O_p\Big(r_n + \|F_{m_n}^\perp K(\Pi_{k_n}\psi - \psi)\|_X^2\Big).\end{aligned}$$

Further, using the elementary inequality $(a-b)^2 \geqslant a^2/2 - b^2$ and again applying relation (A.8) gives uniformly in ϕ

$$\begin{aligned}\left\| n^{-1} \sum_i \left(Y_i \widehat{g}_n(Y_i) - \phi(Z_i)\right) f_{\underline{m_n}}(X_i) \right\|^2 &\geqslant \left\| \mathbf{E}\left[\left(Y^* - \phi(Z)\right) f_{\underline{m_n}}(X)\right] \right\|^2 / 2 \\ &\quad - \sum_{j=1}^{k_n} \max_{\phi \in \Psi_n} \left| n^{-1} \sum_i \left(Y_i \widehat{g}_n(Y_i) - \phi(Z_i)\right) f_j(X_i) - \mathbf{E}\left[\left(Y^* - \phi(Z)\right) f_j(X)\right] \right|^2 \\ &\geqslant C\|K(\Pi_{k_n}\psi - \psi)\|_X^2 - O_p\Big(r_n + \|F_{m_n}^\perp K(\Pi_{k_n}\psi - \psi)\|_X^2\Big).\end{aligned}$$

For some $\varepsilon > 0$ let us denote $\widetilde{\Psi}_n = \{\phi \in \Psi_n : \|K(\phi - \psi)\|_X^2 \geqslant \varepsilon \widetilde{r}_n\}$ where $\widetilde{r}_n = r_n + \|F_{m_n}^\perp K(\Pi_{k_n}\psi - \psi)\|_X^2$. Therefore, following the proof of Lemma B.1 of Chen and Pouzo [2012] we obtain

$$\begin{aligned}&\mathbb{P}\Big(\|K(\widehat{\psi}_n - \psi)\|_X^2 \geqslant \varepsilon\tau_n\Big) \\ &\quad\leqslant \mathbb{P}\Big(\min_{\phi \in \widetilde{\Psi}_n} \left\| \sum_i \left(Y_i \widehat{g}_n(Y_i) - \phi(Z_i)\right) f_{\underline{m_n}}(X_i) \right\|^2 \leqslant \left\| \sum_i \left(Y_i \widehat{g}_n(Y_i) - (\Pi_{k_n}\psi)(Z_i)\right) f_{\underline{m_n}}(X_i) \right\|^2 \Big) \\ &\quad\leqslant \mathbb{P}\Big(\min_{\phi \in \widetilde{\Psi}_n} \|K(\phi - \psi)\|_X^2 \leqslant \|K(\Pi_{k_n}\psi - \psi)\|_X^2 + O_p(\widetilde{r}_n)\Big)\end{aligned}$$

which goes to zero for all $n \geqslant 1$ as $\varepsilon \to \infty$. This shows $\|K(\widehat{\psi}_n - \Pi_{k_n}\psi)\|_X = O_p(\widetilde{r}_n)$ and thus proves the result. □

For the following proofs, recall the notation $\widehat{A}_n = \left(\mathbf{Z}_{k_n}^t \mathbf{X}_{m_n} \widehat{Q}_n^{-1} \mathbf{X}_{m_n}^t \mathbf{Z}_{k_n}/n\right)^{-1} \mathbf{Z}_{k_n}^t \mathbf{X}_{m_n} \widehat{Q}_n^{-1}$ and $A_n = (\mathbf{K}_n^t \mathbf{K}_n)^{-1} \mathbf{K}_n^t$ where $\mathbf{K}_n = \mathbf{E}[f_{\underline{m_n}}(X) p_{\underline{k_n}}(Z)^t]$.

Proof of Theorem 4.2. Observe that the sieve variance $\mathcal{W}_n(z)$ is bounded from below. Indeed, from the condition $\mathbb{V}\mathrm{ar}(Yg(Y) - \psi(Z)|X) \geqslant C$ we infer

$$\begin{aligned}\mathcal{W}_n(z) &\geqslant p_{\underline{k_n}}(z)^t A_n \, \mathbf{E}\left[f_{\underline{m_n}}(X) \, \mathbb{V}\mathrm{ar}(Yg(Y) - \psi(Z)|X) f_{\underline{m_n}}(X)^t\right] A_n^t p_{\underline{k_n}}(z) \\ &\geqslant C \, p_{\underline{k_n}}(z)^t \left(\mathbf{K}_n^t \mathbf{K}_n\right)^{-1} p_{\underline{k_n}}(z),\end{aligned}$$

which we use in the following. Then, we make use of the decomposition

$$\begin{aligned}\widehat{\psi}_n(z) - \psi(z) &= p_{\underline{k_n}}(z)^t n^{-1} \widehat{A}_n \sum_i f_{\underline{m_n}}(X_i)\left(Y_i g(Y_i) - \psi(Z_i)\right) \\ &\quad + p_{\underline{k_n}}(z)^t n^{-1} \widehat{A}_n \sum_i f_{\underline{m_n}}(X_i) Y_i \left(\widehat{g}_n(Y_i) - E_{k_n} g(Y_i)\right) \\ &\quad + p_{\underline{k_n}}(z)^t n^{-1} \widehat{A}_n \sum_i f_{\underline{m_n}}(X_i) Y_i \left(E_{k_n} g(Y_i) - g(Y_i)\right) \\ &\quad + p_{\underline{k_n}}(z)^t n^{-1} \widehat{A}_n \sum_i f_{\underline{m_n}}(X_i) \psi(Z_i) - \psi(z) \\ &= I_n + II_n + III_n + IV_n \quad \text{(say)}. \end{aligned} \tag{A.9}$$

Consider I_n. We observe

$$\begin{aligned}\sqrt{n/\mathcal{W}_n(z)}\, I_n &= \sum_i \left(n\mathcal{W}_n(z)\right)^{-1/2} p_{\underline{k_n}}(z)^t A_n f_{\underline{m_n}}(X_i)\left(Y_i g(Y_i) - \psi(Z_i)\right) + o_p(1) \\ &= \sum_i s_{in} + o_p(1).\end{aligned}$$

Again, s_{in}, $1 \leqslant i \leqslant n$ satisfy the Lindeberg condition which can be seen as follows. It holds $\mathbf{E}[s_{in}] = 0$ and $n\,\mathbf{E}[s_{in}^2] = 1$. From Assumption 5 (iv) we infer $\|\mathbf{T}_n^Y \left(\mathbf{T}_n^t \mathbf{T}_n\right)^{-1} \mathbf{T}_n^t\| \leqslant C\|\mathbf{T}_n \left(\mathbf{T}_n^t \mathbf{T}_n\right)^{-1} \mathbf{T}_n^t\| = C$ and due to $\mathbf{E}\,|Yg(Y) - \psi(X)|^4 \leqslant C$ we observe

$$\sum_i \mathbf{E}[s_{in}^2 \, \mathbb{1}_{\{|s_{in}|>\varepsilon\}}] \leqslant n\varepsilon^2 \, \mathbf{E}\,|s_{in}/\varepsilon|^4 \leqslant Cn^{-1}\varepsilon^{-2} m_n^2 = o(1).$$

Consider II_n. We have

$$\begin{aligned}\sqrt{n} II_n &= \sqrt{n}\, p_{\underline{k_n}}(z)^t A_n \mathbf{T}_n^Y (\widehat{\beta}_{k_n} - \beta_{k_n}) + o_p(1) \\ &= n^{-1/2} \sum_i p_{\underline{k_n}}(z)^t A_n \mathbf{T}_n^Y \left(\mathbf{T}_n^t \mathbf{T}_n\right)^{-1} \mathbf{T}_n^t \left(f_{\underline{m_n}}(X_i) - \mathbf{E}[f_{\underline{m_n}}(X)]\right) + o_p(1)\end{aligned}$$

and, further

$$\begin{aligned}&\mathbf{E}\left|n^{-1/2} \sum_i p_{\underline{k_n}}(z)^t A_n \mathbf{T}_n^Y \left(\mathbf{T}_n^t \mathbf{T}_n\right)^{-1} \mathbf{T}_n^t \left(f_{\underline{m_n}}(X_i) - \mathbf{E}[f_{\underline{m_n}}(X)]\right)\right|^2 \\ &\leqslant p_{\underline{k_n}}(z)^t A_n \mathbf{T}_n^Y \left(\mathbf{T}_n^t \mathbf{T}_n\right)^{-1} \left(A_n \mathbf{T}_n^Y\right)^t p_{\underline{k_n}}(z) \\ &= o\left(\mathcal{W}_n(z)\right),\end{aligned}$$

by the second part of assumption (4.6), which yields $\sqrt{n}III_n = o_p\Big(\sqrt{\mathcal{W}_n(z)}\Big)$. Consider III_n. We have

$$\begin{aligned}
&\mathbf{E}\Big| p_{\underline{k_n}}(z)^t A_n n^{-1/2} \sum_i f_{\underline{m_n}}(X_i) Y_i \Big(E_{k_n} g(Y_i) - g(Y_i)\Big)\Big|^2 \\
&\leqslant \mathbf{E}\left| p_{\underline{k_n}}(z)^t A_n f_{\underline{m_n}}(X) Y \Big(E_{k_n} g(Y) - g(Y)\Big)\right|^2 + n\left(\mathbf{E}\, p_{\underline{k_n}}(z)^t A_n f_{\underline{m_n}}(X) Y \Big(E_{k_n} g(Y) - g(Y)\Big)\right)^2 \\
&\leqslant \|E_{k_n} g - g\|_\infty^2 \left\|\Big(\mathbf{K}_n^t \mathbf{K}_n\Big)^{-1}\right\| \|p_{\underline{k_n}}(z)\|^2 + n\Big(p_{\underline{k_n}}(z)^t A_n\, \mathbf{E}\Big[f_{\underline{m_n}}(X) T M_{id}(E_{k_n} g - g)(X)\Big]\Big)^2,
\end{aligned}$$

using that $\mathbf{E}[Y^2|X] \leqslant C$. Thus, by the first part of assumption (4.6) we obtain $\sqrt{n}III_n = o_p\Big(\sqrt{\mathcal{W}_n(z)}\Big)$. Finally, $\sqrt{n}IV_n = o_p\Big(\sqrt{\mathcal{W}_n(z)}\Big)$ follows from condition $\sqrt{n}(\Pi_{k_n}\psi - \psi)(z) = o\Big(\sqrt{\mathcal{W}_n(z)}\Big)$, which completes the proof of the first statement in the theorem. Finally, by Lemma A.2 below we see that the asymptotic distribution result remains valid if we replace $\mathcal{W}_n(z)$ by its estimator, which completes the proof. □

Lemma A.2. *Let Assumptions 1 – 5, 6 (i)-(ii), 7 (ii), (iii) and 9 – 11 be satisfied. Moreover, assume that: $\mathcal{Y}$ is bounded, $k_n^2 = o\Big(n \min(\tau_{k_n}, \kappa_{k_n}^2)\Big)$, $\tau_{k_n} k_n^{1-\beta} = o(\kappa_{k_n})$ and $m_n \log k_n = o(n\kappa_{k_n})$. Then, we have*

$$\mathcal{W}_n(z)^{-1}\widehat{\mathcal{W}}_n(z) = 1 + o_p(1)$$

uniformly in $z \in \mathcal{Z}$.

Proof. In this proof, we use the notation $\Sigma_n^\psi = \mathbf{E}\Big[f_{\underline{m_n}}(X)\, \mathbb{V}\mathrm{ar}(Yg(Y) - \psi(Z)|X) f_{\underline{m_n}}(X)^t\Big]$, $\widehat{\Sigma}_n^\psi = n^{-1}\sum_i f_{\underline{m_n}}(X_i) f_{\underline{m_n}}(X_i)^t \Big(Y_i\, \widehat{g}_n(Y_i) - \widehat{\psi}_n(Z_i)\Big)^2$ and $\widetilde{\Sigma}_n^\psi = n^{-1}\sum_i f_{\underline{m_n}}(X_i) f_{\underline{m_n}}(X_i)^t \Big(Y_i\, g(Y_i) - \psi(X_i)\Big)^2$. Moreover, let

$$\eta_{n,3} = \sqrt{\frac{m_n \log(k_n)}{n\kappa_{k_n}}},$$

$\widehat{h}(z)^t = (\mathcal{W}_n(z))^{-1/2} p_{\underline{k_n}}(z)^t \widehat{A}_n$, $h(z)^t = (\mathcal{W}_n(z))^{-1/2} p_{\underline{k_n}}(z)^t A_n$. Hence, $\mathcal{W}_n(z)\widehat{h}(z)^t \widehat{\Sigma}_n^\psi \widehat{h}(z) = \widehat{\mathcal{W}}_n(z)$ and by noticing that $h(z)^t \Sigma_n^\psi h(z) = 1$, the triangle inequality gives

$$\begin{aligned}
\Big|(\mathcal{W}_n(z))^{-1/2}\widehat{\mathcal{W}}_n(z)(\mathcal{W}_n(z))^{-1/2} - 1\Big| \leqslant{}& \Big|\widehat{h}(z)^t(\widehat{\Sigma}_n^\psi - \widetilde{\Sigma}_n^\psi)\widehat{h}(z)\Big| + \Big|\widehat{h}(z)^t(\widetilde{\Sigma}_n^\psi - \Sigma_n^\psi)\widehat{h}(z)\Big| \\
&+ \Big|\widehat{h}(z)^t \Sigma_n^\psi \widehat{h}(z) - h(z)^t \Sigma_n^\psi h(z)\Big|. \quad \text{(A.10)}
\end{aligned}$$

Remark that, under Assumptions 4 *(ii) – (iii)* , 9 *(i) – (ii)* and 11 *(i)*: $\sup_{z\in\mathcal{Z}} \|\widehat{h}(z) - h(z)\| = O_p(\eta_{n,3})$ and $\|\widehat{h}(z)\| = O_p(1)$ by using the fact that

$$\begin{aligned}
\|(\mathcal{W}_n(z))^{-1/2} p_{\underline{k_n}}(z)^t (\widehat{A}_n - A_n)\| &= \|(\mathcal{W}_n(z))^{-1/2} p_{\underline{k_n}}(z)^t A_n \mathbf{K}_n (\widehat{A}_n - A_n)\| \\
&\leqslant C^{-1} \|\mathbf{K}_n(\widehat{A}_n - A_n)\| = O_p(\eta_{n,3})
\end{aligned}$$

where we have used $A_n \mathbf{K}_n = I_{k_n}$. Moreover, under the same assumptions, the Rudelson's inequality (see *e.g.* Belloni et al. [2015]) implies that $\|\widetilde{\Sigma}_n^\psi - \Sigma_n^\psi\| = O_p\Big(\sqrt{m_n \log m_n / n}\Big)$.

Therefore, we can show similarly as in Newey [1997] page 165 – 166 that:

$$\sup_{z\in\mathcal{Z}}\left|\widehat{h}(z)^t\Sigma_n^{\psi}\widehat{h}(z)-h(z)^t\Sigma_n^{\psi}h(z)\right|=O_p(\eta_{n,3})$$

$$\text{and}\qquad \sup_{z\in\mathcal{Z}}\left|\widehat{h}(z)^t(\widetilde{\Sigma}_n^{\psi}-\Sigma_n^{\psi})\widehat{h}(z)\right|=O_p\left(\sqrt{\frac{m_n\log m_n}{n}}\right). \quad \text{(A.11)}$$

Moreover, we denote in the following $\widehat{S}_n^{\psi}=n^{-1}\sum_i f_{\underline{m_n}}(X_i)f_{\underline{m_n}}(X_i)^t\left|\left(Y_i\,g(Y_i)-\psi(Z_i)\right)\right|$, $S_n^{\psi}=\mathbf{E}[f_{\underline{m_n}}(X)f_{\underline{m_n}}(X)^t\left|\left(Y_i\,g(Y_i)-\psi(Z_i)\right)\right|]$, $D_{\widehat{g}_n}(\cdot)=\widehat{g}_n(\cdot)-g(\cdot)$ and $D_{\widehat{\psi}_n}(\cdot)=\widehat{\psi}_n(\cdot)-\psi(\cdot)$ and remark that $\|\widehat{S}_n^{\psi}-S_n^{\psi}\|=O_p(\sqrt{m_n\log m_n/n})$ under Assumption 11 (i). Hence,

$$\begin{aligned}
&\sup_{z\in\mathcal{Z}}\left|\widehat{h}(z)^t(\widehat{\Sigma}_n^{\psi}-\widetilde{\Sigma}_n^{\psi})\widehat{h}(z)\right|=\\
&\qquad\sup_{z\in\mathcal{Z}}\left|n^{-1}\sum_{i=1}^n\left(\widehat{h}(z)^t f_{m_n}(X_i)\right)^2\left(\left(Y_i\,\widehat{g}_n(Y_i)-\widehat{\psi}_n(Z_i)\right)^2-\left(Y_i\,g(Y_i)-\psi(Z_i)\right)^2\right)\right|\\
&\leqslant\sup_{z\in\mathcal{Z}}\left|n^{-1}\sum_{i=1}^n\left(\widehat{h}(z)^t f_{m_n}(X_i)\right)^2(Y_iD_{\widehat{g}_n}(Y_i)-D_{\widehat{\psi}_n}(Z_i))^2\right|\\
&\qquad+2\sup_{z\in\mathcal{Z}}\left|n^{-1}\sum_{i=1}^n\left(\widehat{h}(z)^t f_{m_n}(X_i)\right)^2\left(Y_i\,g(Y_i)-\psi(Z_i)\right)\left(Y_iD_{\widehat{g}_n}(Y_i)-D_{\widehat{\psi}_n}(Z_i)\right)\right|\\
&\leqslant 2\left(\|M_{id}D_{\widehat{g}_n}\|_\infty^2+\|D_{\widehat{\psi}}\|_\infty^2\right)\sup_{z\in\mathcal{Z}}\left|\widehat{h}(z)^t\widehat{Q}_n^{-1}\widehat{h}(z)\right|\\
&\qquad+2\left(\|M_{id}D_{\widehat{g}_n}\|_\infty+\|D_{\widehat{\psi}}\|_\infty\right)\sup_{z\in\mathcal{Z}}\left|\widehat{h}(z)^t(\widehat{S}_n-S_n)\widehat{h}(z)+\widehat{h}(z)^tS_n\widehat{h}(z)\right|\\
&\qquad=O_p\left((\eta_{n,4}+\eta_{n,5})(\eta_{n,3}+1)(\sqrt{m_n\log m_n/n}+1)\right)\\
&\quad+O_p\left((\eta_{n,4}+\eta_{n,5})^{1/2}(\eta_{n,3}+1)(\sqrt{m_n\log m_n/n}+1)\right)=O_p\left((\eta_{n,4}+\eta_{n,5})^{1/2}\right) \quad \text{(A.12)}
\end{aligned}$$

where $\eta_{n,4}=\max\left(k_n^2/(n\tau_{k_n}),k_n^{-2\beta}\right)$, $\|M_{id}D_{\widehat{g}_n}\|_\infty^2=O_p\left(\eta_{n,4}\right)$,

$$\eta_{n,5}=\max\left(k_n^{-2\gamma/d_z},\frac{k_n^2}{n\kappa_{k_n}},k_n^{1-\beta}\tau_{k_n}/\kappa_{k_n}\right)$$

and $\|D_{\widehat{\psi}_n}\|_\infty^2=O_p\left(\eta_{n,5}\right)$ which converge to zero under the assumptions of the lemma. To see this, remark that $k_n\kappa_{k_n}^{-1}\|T(E_{k_n}g-g)\|_X^2=O(k_n^{1-\beta}\tau_{k_n}/\kappa_{k_n})$. Then, by (A.10)–(A.12) the result of the lemma follows. □

References

H. Ahn and J. L. Powell. Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics*, 58(1):3–29, 1993.

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

C. Ai and X. Chen. Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables. *Journal of Econometrics*, 141:5–43, 2007.

A. Belloni, V. Chernozhukov, D. Chetverikov, and K. Kato. Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2): 345 – 366, 2015.

A. G. Blom, C. Gathmann, and U. Krieger. Setting up an online panel representative of the general population. *Field Methods*, 27(4):391–408, 2015.

A. G. Blom, D. Bossert, F. Funke, F. Gebhard, A. Holthausen, and U. Krieger. German internet panel, wave 4 (march 2013), 2016. URL http://dx.doi.org/10.4232/1.12610.

R. Blundell, X. Chen, and D. Kristensen. Semi-nonparametric iv estimation of shape-invariant engel curves. 75(6):1613–1669, 2007.

C. Breunig. Goodness-of-fit tests based on series estimators in nonparametric instrumental regression. *Journal of Econometrics*, 184(2):328–346, 2015a.

C. Breunig. Specification testing in nonparametric instrumental quantile regression. Technical report, 2015b.

C. Breunig. Testing missing at random using instrumental variables. Technical report, 2016.

C. Breunig and J. Johannes. Adaptive estimation of functionals in nonparametric instrumental regression. *Econometric Theory*, pages 1–43, 2011.

G. Chamberlain. Asymptotic efficiency in semiparametric models with censoring. *Journal of Econometrics*, 32:189–218, 1986.

K. Chen. Parametric models for response-biased sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(4):775–789, 2001.

X. Chen. Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, 2007.

X. Chen and T. Christensen. Optimal sup-norm rates, adaptivity and inference in nonparametric instrumental variables estimation. Technical report, Cowles Foundation Discussion Paper no 1923R, 2015a.

X. Chen and T. M. Christensen. Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics*, 188(2):447–465, 2015b.

X. Chen and D. Pouzo. Estimation of nonparametric conditional moment models with possibly nonsmooth moments. *Econometrica*, 80(1):277–322, 2012.

X. Chen and D. Pouzo. Sieve wald and qlr inferences on semi/nonparametric conditional moment models. *Econometrica*, 83(3), 2015.

X. Chen and M. Reiß. On rate optimality for ill-posed inverse problems in econometrics. *Econometric Theory*, 27:497–521, 2011.

X. Chen, H. Hong, and D. Nekipelov. Nonlinear models of measurement errors. *Journal of Economic Literature*, 49(4):901–937, 2011.

V. Chernozhukov, S. Lee, and A. M. Rosen. Intersection bounds: Estimation and inference. *Econometrica*, 81:667–737, 2013.

V. Chernozhukov, D. Chetverikov, and K. Kato. Anti-concentration and honest, adaptive confidence bands. *Annals of Statistics*, 42(5):1787–1818, 2014.

S. Darolles, Y. Fan, J.-P. Florens, and E. Renault. Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565, 2011.

M. Das, W. K. Newey, and F. Vella. Nonparametric estimation of sample selection models. *The Review of Economic Studies*, 70(1):33–58, 2003.

L. Davezies and X. D'Haultfoeuille. Partial identification with missing data. *mimeo*, 2013.

C. De Boor. *A practical guide to splines*. Springer, 1978.

X. D'Haultfoeuille. A new instrumental method for dealing with endogenous selection. *Journal of Econometrics*, 154(1):1–15, 2010.

X. D'Haultfoeuille. On the completeness condition in nonparametric instrumental problems. *Econometric Theory*, 27(3):460, 2011.

X. D'Haultfoeuille and A. Maurel. Another look at identification at infinity of sample selection models. *Econometric Theory*, 29, 2013.

S. G. Donald, G. W. Imbens, and W. K. Newey. Empirical likelihood estimation and consistent tests with conditional moment restrictions. *Journal of Econometrics*, 117(1): 55–93, 2003.

M. Frölich. Finite-sample properties of propensity-score matching and weighting estimators. *The Review of Economics and Statistics*, 86(1):77–90, 2004.

J. Heckman. Shadow prices, market wages, and labor supply. *Econometrica*, 70:679–694, 1974.

J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47:153–161, 1979.

M. Henry, Y. Kitamura, and B. Salanié. Partial identification of finite mixtures in econometric models. *Quantitative Economics*, 5:123–144, 2014.

S. Hoderlein, L. Nesheim, and A. Simoni. Semiparametric estimation of random coefficients in structural economic models. *Econometric Theory*, 2016.

Y. Hong and H. White. Consistent specification testing via nonparametric series regression. *Econometrica*, 63:1133–1159, 1995.

J. L. Horowitz and S. Lee. Uniform confidence bands for functions estimated nonparametrically with instrumental variables. *Journal of Econometrics*, 168(2):175–188, 2012.

Y. Hu and S. Schennach. Instrumental variable treatment of nonclassical measurement error models. *Econometrica*, 76(1):195–216, 2008.

S. Huck, T. Schmidt, and G. Weizsacker. The standard portfolio choice problem in germany. Technical report, 2015.

A. Lewbel. Endogenous selection or treatment model estimation. *Journal of Econometrics*, 141:777–806, 2007.

E. Mammen. Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics*, pages 255–285, 1993.

C. F. Manski. Anatomy of the selection problem. *The Journal of Human Resources*, 24:343–360, 1989.

C. F. Manski. The selection problem. In C. E. Sims, editor, *Advances in Econometrics, Sixth World Congress*. Cambridge University Press., 1994.

W. K. Newey. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79(1):147 – 168, 1997.

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

D. Pollard. *A User's Guide to Measure Theoretic Probability*. Cambridge University Press, 2001.

E. A. Ramalho and R. J. Smith. Discrete choice non-response. *The Review of Economic Studies*, 80(1):343–364, 2013.

D. B. Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.

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

G. Tang, R. J. Little, and T. E. Raghunathan. Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika*, 90(4):747–764, 2003.

J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics (Springer Series in Statistics)*. Springer, corrected edition, Nov. 2000. ISBN 0387946403.

J. M. Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.

J. Zhao and J. Shao. Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. *Journal of the American Statistical Association*, 110(512): 1577–1590, 2015.

B. Supplementary Material: A Model Specification Test

Our estimation procedure crucially relies on the conditional independence between selection and covariates given potential outcomes (see Assumption 1). Hence, it would be desirable to test the validity of this assumption before conducting the estimation procedure. An attractive feature of Assumption 1 is that it is indeed testable (cf. Theorem 2.4 in D'Haultfoeuille [2010]) under a maintained completeness assumption of the distribution of Y^* conditional on X. In this section we construct a test for this assumption. As seen in Section 2, given Assumptions 2 and 3, Assumption 1 is equivalent to the operator equation $Tg = 1$. Let us consider a reasonable class of functions for g namely $\mathcal{F} = \left\{\phi \in \mathcal{G} : \phi(\cdot) \geqslant 1 \text{ and } \|\phi - E_{k_n}\phi\|_Y \leqslant Ck_n^{-\beta} \text{ for any } n \geqslant 1\right\}$ for some $\beta > 0$. The null hypothesis under consideration is

$$H_0 : \text{there exists a function } g \in \mathcal{F} \text{ such that } Tg = 1.$$

The test statistic. Our testing procedure is based on the criterion in (3.2). We verify whether $\sum_{i=1}^n \widehat{\chi}_n^2(X_i, \widehat{g}_n)$ does not become too large, which is the case if the true inverse conditional probability function g does not satisfy the minimal smoothness conditions imposed by H_0. By reformulating the quantity $\sum_{i=1}^n \widehat{\chi}_n^2(X_i, \widehat{g}_n)$ we obtain our test statistic

$$S_n = \Big(\sum_{i=1}^n (\widehat{g}_n(Y_i)\Delta_i - 1) f_{\underline{m_n}}(X_i)\Big)^t (\mathbf{X}_{m_n}^t \mathbf{X}_{m_n})^- \Big(\sum_{i=1}^n (\widehat{g}_n(Y_i)\Delta_i - 1) f_{\underline{m_n}}(X_i)\Big) \tag{B.1}$$

where the dimension m_n coincides with the second step dimension used for the estimator $\widehat{\varphi}_n$. Our testing procedure builds on Breunig [2015a]. Also related is the test proposed by Donald et al. [2003] but who consider a parametric function under the maintained hypothesis. However, as we consider a constraint estimation procedure we cannot apply the method of Breunig [2015a] directly. A constraint sieve testing procedure was proposed by Breunig [2015b] but for the specific situation of quantile versions of instrumental variable models. In addition, note that in these two papers the basis functions used to construct the test statistics are assumed to be orthonormal, which is not required in the following.

Asymptotic distribution of the statistic. Our test statistic S_n is asymptotically normally distributed if it is standardized by appropriate mean and variance, which are introduced in the next definition.

Definition B.1. *Let us introduce the matrix*

$$\Sigma_{m_n} = \mathbf{E}\Big[\big(g(Y)\Delta - 1\big)^2 Q_n^{-1/2} f_{\underline{m_n}}(X) f_{\underline{m_n}}(X)^t Q_n^{-1/2}\Big]$$

Then the trace and the Frobenius norm of Σ_{m_n} are respectively denoted by μ_{m_n} and ς_{m_n}.

Assumption 13. *There exist constants $c, C > 0$ such that $\mathbb{V}\mathrm{ar}(g(Y)\Delta|X) \geqslant c$ and $\mathbf{E}[(g(Y)\Delta - 1)^4|X] \leqslant C$.*

Due to Assumption 13 it holds $\varsigma_{m_n} \geqslant C\sqrt{m_n}$ for some constant $C > 0$. The next result establishes asymptotic normality of S_n after standardization.

Theorem B.1. *Let Assumptions 2–6 and 13 be satisfied. If*

$$m_n^3 = o(n), \quad k_n m_n^2 = O(n\tau_{k_n}), \quad \text{and} \quad \max\left(k_n, n\tau_{k_n} k_n^{-2\beta}\right) = o(\sqrt{m_n}) \tag{B.2}$$

then it holds under H_0

$$(\sqrt{2}\varsigma_{m_n})^{-1}\Big(nS_n - \mu_{m_n}\Big) \xrightarrow{d} \mathcal{N}(0,1).$$

Estimation of Critical Values. For the estimation of the critical values of Theorem B.1, let us define $\mathbf{U}_n = \Big(\Delta_1\widehat{g}_n(Y_1) - 1, \ldots, \Delta_n\widehat{g}_n(Y_n) - 1\Big)^t$. We estimate the matrix Σ_{m_n} by $\widehat{\Sigma}_{m_n} \equiv (\mathbf{X}_{m_n}^t\mathbf{X}_{m_n})^{-1/2}\mathbf{X}_{m_n}^t \operatorname{diag}(\mathbf{U}_n)^2\,\mathbf{X}_{m_n}(\mathbf{X}_{m_n}^t\mathbf{X}_{m_n})^{-1/2}$. The asymptotic result of Theorem B.1 continues to hold if we replace ς_{m_n} by the Frobenius norm of $\widehat{\Sigma}_{m_n}$, denoted by $\widehat{\varsigma}_{m_n}$, and μ_{m_n} by the trace of $\widehat{\Sigma}_{m_n}$, denoted by $\widehat{\mu}_{m_n}$.

Theorem B.2. *Let the assumptions of Theorem B.1 be satisfied. Then it holds under* H_0

$$(\sqrt{2}\widehat{\varsigma}_{m_n})^{-1}\Big(nS_n - \widehat{\mu}_{m_n}\Big) \xrightarrow{d} \mathcal{N}(0,1).$$

Limiting behavior under local alternatives. In the following, we study the power of the test, that is, the probability to reject a false hypothesis against a sequence of linear local alternatives that tends to zero as the sample size tends to infinity. We consider alternative models defined through a sequence of functions g_n that satisfy

$$\|Tg_n - 1 - m_n^{1/4}n^{-1/2}\delta\|_X = o(m_n^{1/4}n^{-1/2}) \tag{B.3}$$

for some function $\delta \in L_X^4$. Due to (B.3), for any $n \geqslant 1$ the function g_n does not satisfy $Tg_n = 1$. For our analysis of local alternatives we follow Hong and White [1995] and assume that the model approaches the data generating process rather than vice versa. Indeed, the sequence of alternative models converges to the model with $Tg = 1$, that is, Assumption 1 holds in the limit. The next result is a direct consequence of Proposition 2.4 of Breunig [2015b] and thus, its proof is omitted.

Proposition B.3. *Let Assumptions 2–6, 13, and the rate condition* (B.2) *be satisfied. Then, under* (B.3) *we have*

$$(\sqrt{2}\widehat{\varsigma}_{m_n})^{-1}\Big(n\,S_n - \widehat{\mu}_{m_n}\Big) \xrightarrow{d} \mathcal{N}\Big(2^{-1/2}\sum_{j=1}^{\infty}\Big(\mathbf{E}[\delta(X)f_j(X)]\Big)^2, 1\Big).$$

Monte Carlo Simulations Let us now study the finite sample behavior of our nonparametric specification test. There are 1000 Monte Carlo replications in each experiment and the sample size is $n = 1000$. Results are presented for the nominal level 0.05. We generate Y^* and X as described in the exogenous case of Section 5. We construct the observations of Δ via the function $h(y) = (10y + 7/2)^{-1} + 5/7$ for $y \in [0,1]$. Note that the normalization constant for h ensures that $h(0) = 1$. If H_0 holds true we generate $\Delta \sim \text{Binomial}(1, h(Y^*))$.

In the experiments where H_0 fails, X is not a valid instrument in the sense that it influences the endogenous selection. In this case, we generate realizations of Δ from

$$\Delta \sim \text{Binomial}\Big(1, \min(1, (1-\nu)h(Y^*) + \nu\rho_j(X))\Big)$$

for some constant $\nu > 0$ and where $j = 1,2,3$ is varied in the experiments. We consider the functions $\rho_1(x) = 1 - (2x-1)^2/6$, $\rho_2(x) = 1 - (2x-1)^2/4$, and $\rho_3(x) = 1 - (2x-1)^2/2$. Clearly, if $\nu = 0$ then the null hypothesis H_0 is true. We estimate the regression function φ

Model		*Empirical rejection probabilities*
ρ	ν	$(\sqrt{2}\widehat{\varsigma}_{m_n})^{-1}(nS_n - \widehat{\mu}_{m_n})$
H_0 true		0.037
ρ_1	0.5	0.139
	0.7	0.235
	0.9	0.569
ρ_2	0.5	0.137
	0.7	0.369
	0.9	0.754
ρ_3	0.5	0.425
	0.7	0.793
	0.9	0.951

Table 2: Empirical rejection probabilities for the nonparametric specification test under the nominal level 0.05.

by $\widehat{\varphi}_n$ as given in (3.4) and implement it as described in Section 5, that is, we use B-splines as basis functions of order 3 with 4 knots (hence $k_n = 8$) and for the criterion function we use B-splines of order 3 with 6 knots (hence $m_n = 10$). In Table (B), we report the empirical rejection probabilities of our test statistic with significance level 0.05. The critical values of these statistics are estimated as described in Theorem B.2. We see from this table that our test becomes more powerful as the parameter ν increases.

B.1. Proofs of the Appendix B.

Proof of Theorem B.1. Since we have $\|\widehat{Q}_n - I_{m_n}\|^2 = o_p(m_n^2/n)$ it is sufficient to prove that $(\sqrt{2}\varsigma_{m_n})^{-1}\big(\sum_{j=1}^{m_n}|n^{-1/2}\sum_i(\Delta_i\widehat{g}_n(Y_i)-1)f_j(X_i)|^2 - \mu_{m_n}\big) \xrightarrow{d} \mathcal{N}(0,1)$. The proof of this statement is based on the decomposition

$$\sum_{j=1}^{m_n}|n^{-1}\sum_i(\Delta_i\widehat{g}_n(Y_i)-1)f_j(X_i)|^2 = \sum_{j=1}^{m_n}\Big|n^{-1}\sum_i\Big(\Delta_i g(Y_i)-1\Big)f_j(X_i)\Big|^2$$
$$-\frac{2}{n^2}\sum_{j=1}^{m_n}\Big(\sum_i\Big(\Delta_i g(Y_i)-1\Big)f_j(X_i)\Big)\Big(\sum_i\Delta_i\Big(\widehat{g}_n(Y_i)-g(Y_i)\Big)f_j(X_i)\Big)$$
$$+\sum_{j=1}^{m_n}\Big|n^{-1}\sum_i\Delta_i\Big(\widehat{g}_n(Y_i)-g(Y_i)\Big)f_j(X_i)\Big|^2 = I_n - 2II_n + III_n. \quad \text{(B.4)}$$

Consider I_n. We calculate

$$\varsigma_{m_n}^{-1}\Big(nI_n - \mu_{m_n}\Big) = \frac{1}{\varsigma_{m_n}n}\sum_i\sum_{j=1}^{m_n}\Big(\Big|\Big(\Delta_i g(Y_i)-1\Big)f_j(X_i)\Big|^2 - \mathbf{E}\Big[\Big(\Delta g(Y)-1\Big)^2 f_j^2(X)\Big]\Big)$$
$$+\frac{1}{\varsigma_{m_n}n}\sum_{i\neq i'}\sum_{j=1}^{m_n}\Big(\Delta_i g(Y_i)-1\Big)\Big(\Delta_{i'} g(Y_{i'})-1\Big)f_j(X_i)f_j(X_{i'})$$

where the first summand tends in probability to zero as $n \to \infty$. Indeed,we have

$$\begin{aligned}
\mathbf{E}\Big|\frac{1}{\varsigma_{m_n} n}\sum_i\sum_{j=1}^{m_n}\Big(\Big|\big(\Delta_i g(Y_i)-1\big)f_j(X_i)\Big|^2-\mathbf{E}\Big[\big(\Delta g(Y)-1\big)^2 f_j^2(X)\Big]\Big)\Big|^2 \\
\leqslant \frac{1}{\varsigma_{m_n}^2 n}\,\mathbf{E}\Big|\sum_{j=1}^{m_n}\Big|\big(\Delta g(Y)-1\big)f_j(X)\Big|^2-\mathbf{E}\Big[\big(\Delta g(Y)-1\big)^2 f_j^2(X)\Big]\Big|^2 \\
\leqslant \frac{1}{\varsigma_{m_n}^2 n}\sup_{x\in\mathcal{X}}\|f_{\underline{m_n}}(x)\|^4\,\mathbf{E}\Big|\Delta g(Y)-1\Big|^4 \\
\leqslant \frac{Cm_n^2}{\varsigma_{m_n}^2 n}=o(1).
\end{aligned}$$

Therefore, to establish $(\sqrt{2}\varsigma_{m_n})^{-1}(nI_n-\mu_{m_n})\xrightarrow{d}\mathcal{N}(0,1)$ it is sufficient to show

$$\frac{\sqrt{2}}{\varsigma_{m_n} n}\sum_{i\neq i'}\sum_{j=1}^{m_n}\big(\Delta_i g(Y_i)-1\big)\big(\Delta_{i'} g(Y_{i'})-1\big)f_j(X_i)f_j(X_{i'})\xrightarrow{d}\mathcal{N}(0,1).$$

This follows from Lemma A.2 of Breunig [2015a].

Consider II_n. We observe

$$\begin{aligned}
nII_n &= \sum_{j=1}^{m_n}\Big(\sum_i\big(\Delta_i g(Y_i)-1\big)f_j(X_i)\Big)\Big(n^{-1}\sum_i\Delta_i\big(\widehat{g}_n(Y_i)-g(Y_i)\big)f_j(X_i)\Big) \\
&= \sum_{j=1}^{m_n}\Big(\sum_i\big(\Delta_i g(Y_i)-1\big)f_j(X_i)\Big)\Big(n^{-1}\sum_i\Delta_i\big(\widehat{g}_n(Y_i)-E_{k_n}g(Y_i)\big)f_j(X_i)\Big) \\
&\quad+\sum_{j=1}^{m_n}\Big(\sum_i\big(\Delta_i g(Y_i)-1\big)f_j(X_i)\Big)\Big(n^{-1}\sum_i\Delta_i\big(E_{k_n}g_n(Y_i)-g(Y_i)\big)f_j(X_i)\Big) \\
&= C_{n1}+C_{n2}.
\end{aligned}$$

Consider C_{n1}. We have

$$\begin{aligned}
C_{n1} =& \|Diag(\tau_1,\ldots,\tau_{k_n})^{1/2}(\widehat{\beta}_{k_n}-\beta_{k_n})\| \\
&\times\sum_{j=1}^{m_n}\Big(\sum_i\big(\Delta_i g(Y_i)-1\big)f_j(X_i)\Big)\|Diag(\tau_1,\ldots,\tau_{k_n})^{-1/2}\,\mathbf{E}[\Delta e_{\underline{k_n}}(Y)f_j(X)]\|+o_p(1).
\end{aligned}$$

Using

$$\begin{aligned}
&\mathbf{E}\Big|\sum_{j=1}^{m_n}\Big(\sum_i\big(\Delta_i g(Y_i)-1\big)f_j(X_i)\Big)\|Diag(\tau_1,\ldots,\tau_{k_n})^{-1/2}\,\mathbf{E}[\Delta e_{\underline{k_n}}(Y)f_j(X)]\|\Big|^2 \\
&\quad\leqslant n\sum_{j=1}^{m_n}\mathbf{E}\Big[\big((\Delta g(Y)-1)f_j(X)\big)^2\Big]\|Diag(\tau_1,\ldots,\tau_{k_n})^{-1/2}\,\mathbf{E}[\Delta e_{\underline{k_n}}(Y)f_j(X)]\|^2 \\
&\quad\leqslant Cn\|Diag(\tau_1,\ldots,\tau_{k_n})^{-1/2}\,\mathbf{E}[\Delta e_{\underline{k_n}}(Y)f_{\underline{m_n}}(X)^t]\|^2.
\end{aligned}$$

Since $\|Diag(\tau_1,\dots,\tau_{k_n})^{-1/2}\,\mathbf{E}[\Delta e_{\underline{k_n}}(Y) f_{\underline{m_n}}(X)^t]\|^2 = O(k_n)$ it holds

$$C_{n1} = \|Diag(\tau_1,\dots,\tau_{k_n})^{1/2}(\widehat{\beta}_{k_n} - \beta_{k_n})\|\, O_p(\sqrt{k_n n}) = O_p(k_n) = o_p(\sqrt{m_n}).$$

Further, we have

$$\begin{aligned}
\mathbf{E}\,|C_{n2}| &\leqslant \sum_{j=1}^{m_n} \sqrt{\mathbf{E}\,|(E_{k_n}g - g)(Y) f_j(X)|^2}\,\sqrt{\mathbf{E}\left|(\Delta_i g(Y) - 1) f_j(X)\right|^2} \\
&\quad + Cn^{1/2}\,\mathbf{E}\Big|\sum_{j=1}^{m_n}\langle T(E_{k_n}g - g), f_j\rangle_X f_j(X)\Big| \\
&\leqslant C\Big(m_n\|E_{k_n}g - g\|_\infty + \sqrt{n}\|T(E_{k_n}g - g)\|_X\Big) \\
&= O(m_n k_n^{1/2-\beta} + \sqrt{n\tau_{k_n}}\,k_n^{-\beta}) = o(\sqrt{m_n})
\end{aligned}$$

where we used that $m_n k_n^{1-2\beta} = o(1)$. Consider III_n. It holds true that

$$\begin{aligned}
III_n &\leqslant Cn\|(\widehat{\beta}_{k_n} - \beta_{k_n})^t\,\mathbf{E}[\Delta e_{\underline{k_n}}(Y) f_{\underline{m_n}}(X)^t]\|^2 \\
&\quad + Cn\|T(E_{k_n}g - g)\|_X^2\Big) \\
&= O_p\Big(k_n + n\tau_{k_n}k_n^{-2\beta}\Big) \\
&= o_p(\sqrt{m_n})
\end{aligned}$$

which completes the proof of the first result in the theorem. □

Proof of Theorem B.2. Remark that $(\sqrt{2}\widehat{\varsigma}_{m_n})^{-1}\Big(nS_n - \widehat{\mu}_{m_n}\Big) = (\sqrt{2}\varsigma_{m_n})^{-1}\Big(nS_n - \mu_{m_n}\Big)\frac{\varsigma_{m_n}}{\widehat{\varsigma}_{m_n}} + (\sqrt{2}\varsigma_{m_n})^{-1}(\mu_{m_n} - \widehat{\mu}_{m_n})\frac{\varsigma_{m_n}}{\widehat{\varsigma}_{m_n}}$. The statement of the theorem follows from the results of Theorem B.1, Lemma B.4 and Lemma B.5. □

Lemma B.4. *Let Assumptions 1 – 6 be satisfied. Then,*

$$\varsigma_{m_n}^{-1}\widehat{\varsigma}_{m_n} = 1 + o_p(1)$$

where ς_{m_n} and $\widehat{\varsigma}_{m_n}$ are as defined in Theorem B.1.

Proof. Let $\|\cdot\|_F$ denote the Frobenius norm of a matrix. Then $\varsigma_{m_n} = \|\Sigma_{m_n}\|_F$ and $\widehat{\varsigma}_{m_n} = \|\widehat{\Sigma}_{m_n}\|_F$. Let us denote $\widetilde{\Sigma}_{m_n} = n^{-1}\sum_i f_{m_n}(X_i) f_{m_n}(X_i)^t (E_{k_n}g(Y_i)\Delta_i - 1)^2$. Observe that

$$\begin{aligned}
&\|\widehat{\Sigma}_{m_n} - \widetilde{\Sigma}_{m_n}\|_F^2 \\
&= \Big\|n^{-1}\sum_i f_{\underline{m_n}}(X_i) f_{\underline{m_n}}(X_i)^t\left[|(\widehat{g_n} - E_{k_n}g)(Y_i)|^2\Delta_i + 2(E_{k_n}g(Y_i)\Delta_i - 1)\Delta_i(\widehat{g_n} - E_{k_n}g)(Y_i)\right]\Big\|_F^2 \\
&\leqslant 2\Big\|n^{-1}\sum_i f_{\underline{m_n}}(X_i) f_{\underline{m_n}}(X_i)^t|(\widehat{g_n} - E_{k_n}g)(Y_i)|^2\Delta_i\Big\|_F^2 \\
&+ 2\Big\|n^{-1}\sum_i f_{\underline{m_n}}(X_i) f_{\underline{m_n}}(X_i)^t(E_{k_n}g(Y_i)\Delta_i - 1)\Delta_i(\widehat{g_n} - E_{k_n}g)(Y_i)\Big\|_F^2 \\
&= 2I_n + 2II_n.
\end{aligned}$$

We further calculate

$$
\begin{aligned}
I_n &\leqslant \left\| \frac{1}{n} \sum_i \Delta_i (\widehat{\beta}_{k_n} - \beta_{k_n})^t e_{\underline{k_n}}(Y_i) f_{\underline{m_n}}(X_i) f_{\underline{m_n}}(X_i)^t e_{\underline{k_n}}(Y_i)^t | (\widehat{\beta}_{k_n} - \beta_{k_n}) \right\|_F^2 \\
&\leqslant \left\| (\widehat{\beta}_{k_n} - \beta_{k_n})^t \, \mathbf{E}[\Delta e_{\underline{k_n}}(Y) f_{\underline{m_n}}(X) f_{\underline{m_n}}(X)^t e_{\underline{k_n}}(Y)^t] (\widehat{\beta}_{k_n} - \beta_{k_n}) \right\|_F^2 + o_p(1) \\
&\leqslant \|\widehat{\beta}_{k_n} - \beta_{k_n}\|^4 \sum_{j,l=1}^{m_n} \mathbf{E}[\|\Delta e_{\underline{k_n}}(Y)\|^2 | f_j(X) f_l(X)|]^2 + o_p(1) \\
&\leqslant C m_n^2 \|\widehat{\beta}_{k_n} - \beta_{k_n}\|^4 \, \mathbf{E}\left(\|\Delta e_{\underline{k_n}}(Y)\|^2 \right)^2 + o_p(1) \\
&= O_p\left(m_n^2 k_n^4 / (\tau_{k_n} n)^2 \right) = o_p(1)
\end{aligned}
$$

by using $k_n = o(\sqrt{m_n})$. Similarly, we conclude

$$
\begin{aligned}
II_n &\leqslant \left\| \frac{1}{n} \sum_i (E_{k_n} g(Y_i) \Delta_i - 1) f_{\underline{m_n}}(X_i) f_{\underline{m_n}}(X_i)^t e_{\underline{k_n}}(Y_i)^t | (\widehat{\beta}_{k_n} - \beta_{k_n}) \right\|_F^2 \\
&\leqslant \left\| \mathbf{E}[(E_{k_n} g(Y) \Delta - 1) \Delta f_{\underline{m_n}}(X) f_{\underline{m_n}}(X)^t e_{\underline{k_n}}(Y)^t] (\widehat{\beta}_{k_n} - \beta_{k_n}) \right\|_F^2 + o_p(1) \\
&\leqslant \|\widehat{\beta}_{k_n} - \beta_{k_n}\|^2 \sum_{j,l=1}^{m_n} \sum_{l'=1}^{k_n} \left(\mathbf{E}[(E_{k_n} g(Y) \Delta - 1) \Delta e_{l'}(Y) f_j(X) f_l(X)] \right)^2 + o_p(1) \\
&= \|\widehat{\beta}_{k_n} - \beta_{k_n}\|^2 \sum_{j=1}^{m_n} \sum_{l'=1}^{k_n} \|F_{m_n} T((E_{k_n} g - 1) \cdot e_{l'}) \cdot f_j\|_X^2 + o_p(1) \\
&= O_p\left(m_n k_n^2 / (\tau_{k_n} n) \right) = o_p(1)
\end{aligned}
$$

by using $k_n = o(\sqrt{m_n})$. Next,

$$
\begin{aligned}
\mathbf{E} \|\widetilde{\Sigma}_{m_n} - \Sigma_{m_n}\|_F^2 &= \sum_{j,l=1}^{m_n} \mathbf{E} \left(\frac{1}{n} \sum_i \left(f_j(X_i) f_l(X_i) (E_{k_n} g(Y_i) \Delta_i - 1)^2 - \mathbf{E}\left[f_j(X) f_l(X) (g(Y)\Delta - 1)^2 \right] \right) \right)^2 \\
&\leqslant \frac{1}{n} \sum_{j,l=1}^{m_n} \mathbf{E}\left[f_j^2(X) f_l^2(X) (E_{k_n} g(Y) \Delta - 1)^4 \right] + \sum_{j,l=1}^{m_n} \left(\mathbf{E}\left[f_j(X) f_l(X) \Delta (E_{k_n} g(Y) - g(Y))^2 \right] \right)^2 \\
&\leqslant C \frac{m_n^2}{n} + \sum_{j=1}^{m_n} \|F_{m_n} T(E_{k_n} g - g) \cdot f_j\|_X^2 = O(m_n^2 n^{-1} + m_n k_n^{1-2\beta}) = o(1)
\end{aligned}
$$

Finally, by these results and the reverse triangle inequality we conclude that

$$
\left| \varsigma_{m_n}^{-1} \widehat{\varsigma}_{m_n} - 1 \right| = \varsigma_{m_n}^{-1} \left| \|\widehat{\Sigma}_{m_n}\|_F - \|\Sigma_{m_n}\|_F \right| \leqslant \varsigma_{m_n}^{-1} \|\widehat{\Sigma}_{m_n} - \widetilde{\Sigma}_{m_n}\|_F + \varsigma_{m_n}^{-1} \|\widetilde{\Sigma}_{m_n} - \Sigma_{m_n}\|_F = o_p(1)
$$

which proves the result. □

LEMMA B.5. *Let Assumptions 1–6 be satisfied. Then,*

$$
\widehat{\mu}_{m_n} = \mu_{m_n} + o_p(\varsigma_{m_n})
$$

where μ_{m_n} and $\widehat{\mu}_{m_n}$ are as defined in Theorem B.1.

Proof. The proof of Lemma B.4 establishes $\|\widehat{\Sigma}_{m_n} - \Sigma_{m_n}\|_F = o_p(1)$. In particular, convergence of the trace of $\widehat{\Sigma}_{m_n}$ to the trace of Σ_{m_n} follows by using $|\widehat{\mu}_{m_n} - \mu_{m_n}| \leqslant \sqrt{m_n}\,\|\widehat{\Sigma}_{m_n} - \Sigma_{m_n}\|_F = o_p(\varsigma_{m_n})$. □

C. Supplementary Material: Proofs of Theorems 3.4 and 4.3

In this appendix we present the proofs of Theorems 3.4 and 4.3. The construction of these proofs follows the proof of Chen and Christensen [2015a, Theorem B.1] with minor modifications. In order to make the paper self-contained we report here the main steps and refer to Chen and Christensen [2015a] for the full description of the proof strategy. The modifications with respect to the proof of Chen and Christensen [2015a] are due to the fact that we are estimating a different model and the type of estimator is different, that is, two-step instead of one step.

Proof of Theorem 3.4. The proof is made of four steps and we use all along the proof the inequality $\mathcal{V}_{1,n}(x) \geqslant c\,\|f_{\underline{m}_n}(x)\|^2$ which is valid under Assumption 7 *(i)*. Let $D_n = \{(\Delta_1, Y_1, X_1), \ldots, (\Delta_n, Y_n, X_n)\}$.

Step 1. We start by showing that $\sqrt{n/\widehat{\mathcal{V}}_{1n}(x)}\big(\widehat{\varphi}_n(x) - \varphi(x)\big)$ can be uniformly approximated by the process

$$\widehat{\mathbb{X}}_n(x) = \frac{f_{\underline{m}_n}(x)^t}{\sqrt{\mathcal{V}_{1n}(x)}}\frac{1}{\sqrt{n}}\sum_i f_{\underline{m}_n}(X_i)(Y_i g(Y_i) - \varphi(X_i)).$$

From the decomposition (A.3) we can write

$$\begin{aligned}\left|\sqrt{n/\widehat{\mathcal{V}}_{1n}(x)}\big(\widehat{\varphi}_n(x) - \varphi(x)\big) - \widehat{\mathbb{X}}_n(x)\right| &\leqslant \left|\frac{\sqrt{n}I_n}{\sqrt{\mathcal{V}_{1n}(x)}} - \widehat{\mathbb{X}}_n(x)\right| \\ &\quad + \frac{\sqrt{n}|I_n|}{\sqrt{\mathcal{V}_{1n}(x)}}\left|\frac{\sqrt{\mathcal{V}_{1n}(x)}}{\sqrt{\widehat{\mathcal{V}}_{1n}(x)}} - 1\right| + \frac{\sqrt{n}\,|II_n + III_n + IV_n|}{\sqrt{\widehat{\mathcal{V}}_{1n}(x)}} \\ &= \mathfrak{I}_{n,1}(x) + \mathfrak{I}_{n,2}(x) + \mathfrak{I}_{n,3}(x) \qquad \text{(say)} \quad \text{(C.1)}\end{aligned}$$

Because, under Assumption 4 *(i)-(iii)*, $\|\widehat{Q}_n^{-1} - I_{m_n}\| = O_p(\sqrt{m_n \log(m_n)/n})$ by the Rudelson's inequality and $\|\sum_i f_{\underline{m}_n}(X_i)(Y_i g(Y_i) - \varphi(X_i))/\sqrt{n}\| = O_p(\sqrt{m_n})$, then $\sup_{x\in\mathcal{X}}\mathfrak{I}_{n,1}(x) = O_p(m_n\sqrt{\log(m_n)/n})$.

Next, we consider term $\mathfrak{I}_{n,2}(x)$. Lemma A.1 implies that $\sup_{x\in\mathcal{X}}\left|\frac{\sqrt{\mathcal{V}_{1n}(x)}}{\sqrt{\widehat{\mathcal{V}}_{1n}(x)}} - 1\right| = O_p(\eta_n)$,

where $\eta_n = \max\left(k_n/(\tau_{k_n} n)^{1/2}, k_n^{-\beta}\right) + \max\left(m_n^{-\alpha/d_x}, \frac{m_n}{\sqrt{n}}, \sqrt{m_n}\|T(E_{k_n} g - g)\|_X\right)$. Therefore,

$$\begin{aligned}
\sup_{x\in\mathcal{X}} \mathfrak{I}_{n,2}(x) &= \sup_{x\in\mathcal{X}} \mathfrak{I}_{n,1}(x) \left| \frac{\sqrt{\mathcal{V}_{1n}(x)}}{\sqrt{\widehat{\mathcal{V}}_{1n}(x)}} - 1 \right| + \sup_{x\in\mathcal{X}} |\widehat{\mathbb{X}}_n(x)| \left| \frac{\sqrt{\mathcal{V}_{1n}(x)}}{\sqrt{\widehat{\mathcal{V}}_{1n}(x)}} - 1 \right| \\
&= O_p(\eta_n)\left(O_p(m_n \sqrt{\log(m_n)/n}) + \sup_{x\in\mathcal{X}} |\widehat{\mathbb{X}}_n(x) - \mathbb{X}_n(x)| + \sup_{x\in\mathcal{X}} |\mathbb{X}_n(x)| \right) \\
&= O_p(\eta_n)\left(O_p(m_n \sqrt{\log(m_n)/n}) + o_p(r_n) + \sup_{x\in\mathcal{X}} |\mathbb{X}_n(x)| \right) \\
&= O_p(\eta_n)\left(o_p(r_n) + O_p(c_n) \right). \qquad \text{(C.2)}
\end{aligned}$$

where the third equality is due to step 2 below and the last equality is because of the condition $m_n^{3/2}/(r_n^3 \sqrt{n}) = o(1)$ and by Chen and Christensen [2015a, LemmaD.7], which is valid under Assumptions 4 *(i)*, 7 *(i)* and 8 *(i)* - *(ii)* and which implies $\sup_{x\in\mathcal{X}} |\mathbb{X}_n(x)| = O_p(c_n)$. Finally, let us analyze the three terms in $\mathfrak{I}_{n,3}(x)$ separately. By denoting $\widehat{\mathbf{T}}_n^Y = \sum_{i=1}^n f_{\underline{m_n}}(X_i) Y_i e_{\underline{k_n}}(Y_i)^t/n$ we have

$$\begin{aligned}
\sup_{x\in\mathcal{X}} \frac{\sqrt{n}\,|II_n|}{\sqrt{\mathcal{V}_{1n}(x)}} &= \sup_{x\in\mathcal{X}} \frac{\sqrt{n}\left| f_{\underline{m}}(x)^t (\widehat{Q}_n^{-1} - I_{m_n}) \widehat{\mathbf{T}}_n^Y (\widehat{\beta}_{k_n} - \mathbf{E}[g(Y) e_{\underline{k_n}}(Y)]) \right|}{\sqrt{\mathcal{V}_{1n}(x)}} \\
&\quad + \sup_{x\in\mathcal{X}} \frac{\sqrt{n}\left| f_{\underline{m}}(x)^t (\widehat{\mathbf{T}}_n^Y - \mathbf{T}_n^Y)(\widehat{\beta}_{k_n} - \mathbf{E}[g(Y) e_{\underline{k_n}}(Y)]) \right|}{\sqrt{\mathcal{V}_{1n}(x)}} + \sup_{x\in\mathcal{X}} \frac{\sqrt{n}\left| f_{\underline{m}}(x)^t \mathbf{T}_n^Y (\widehat{\beta}_{k_n} - \beta_{k_n}) \right|}{\sqrt{\mathcal{V}_{1n}(x)}} \\
&\quad + \sup_{x\in\mathcal{X}} \frac{\sqrt{n}\left| f_{\underline{m}}(x)^t \mathbf{T}_n^Y (\mathbf{T}_n^t \mathbf{T}_n)^{-1} \mathbf{T}_n^t \, \mathbf{E}[f_{\underline{m_n}}(X) \Delta (g(Y) - E_{k_n} g(Y))] \right|}{\sqrt{\mathcal{V}_{1n}(x)}} = \mathcal{A}_1 + \mathcal{A}_2 + \mathcal{A}_3 + \mathcal{A}_4 \qquad \text{(say).}
\end{aligned} \qquad \text{(C.3)}$$

Terms $\mathcal{A}_i$, $i = 1, 2, 4$ are easily controlled by using the Cauchy Schwartz-inequality so that:

$$\begin{aligned}
\mathcal{A}_1 &= O_p(\sqrt{m_n \log(m_n)/n}(\sqrt{m_n \log(k_n)/n} \max\{\sqrt{m_n}, \xi_n\}/\tau_{k_n} + \sqrt{m_n/\tau_{k_n}} \\
&\qquad + \sqrt{k_n/(m_n \tau_{k_n})}\, \sqrt{n}\|T(g - E_{k_n} g)\|_X), \\
\mathcal{A}_2 &= O_p(\sqrt{m_n k_n/n}(\sqrt{m_n \log(k_n)/n} \max\{\sqrt{m_n}, \xi_n\}/\tau_{k_n} + \sqrt{m_n/\tau_{k_n}} + \sqrt{n}/\tau_{k_n}\, \|T(g - E_{k_n} g)\|_X)), \\
\mathcal{A}_4 &= O_p(\sqrt{k_n/(m_n \tau_{k_n})}\, \sqrt{n}\|T(g - E_{k_n} g)\|_X)
\end{aligned}$$

where, to get $\mathcal{A}_4$, we have used the same argument as in (A.4) and the assumption $\sup_{x\in\mathcal{X}} \|f_{\underline{m_n}}(x)^t \mathbf{T}_n^Y\|/\|\sqrt{\mathcal{V}_{1n}(x)}\| \leqslant C\sqrt{k_n/m_n}$. Term $\mathcal{A}_3$ can be decomposed as

$$\begin{aligned}
\mathcal{A}_3 = \sup_{x\in\mathcal{X}} &\frac{\left| f_{\underline{m}}(x)^t \mathbf{T}_n^Y [(\widehat{\mathbf{T}}_n^t \widehat{Q}_n^{-1} \widehat{\mathbf{T}}_n)^{-1} \widehat{\mathbf{T}}_n^t \widehat{Q}_n^{-1} - (\mathbf{T}_n^t \mathbf{T}_n)^{-1} \mathbf{T}_n^t] \sum_i f_{\underline{m_n}}(X_i)/\sqrt{n} \right|}{\sqrt{\mathcal{V}_{1n}(x)}} \\
&+ \sup_{x\in\mathcal{X}} \frac{\left| f_{\underline{m}}(x)^t \mathbf{T}_n^Y (\mathbf{T}_n^t \mathbf{T}_n)^{-1} \mathbf{T}_n^t \sum_i (f_{\underline{m_n}}(X_i) - \mathbf{E}[f_{\underline{m_n}}(X)])/\sqrt{n} \right|}{\sqrt{\mathcal{V}_{1n}(x)}} = \mathcal{A}_{3,1} + \mathcal{A}_{3,2}
\end{aligned}$$

where $\mathcal{A}_{3,1} = O_p(\sqrt{k_n/m_n}\max\{\sqrt{m_n}, \xi_n\}\tau_{k_n}^{-1}\sqrt{m_n \log k_n/n})$. To control term $\mathcal{A}_{3,2}$, let $s(x)^t = f_{\underline{m_n}}(x)^t \mathbf{T}_n^Y \left(\mathbf{T}_n^t \mathbf{T}_n\right)^{-1} \mathbf{T}_n^t / \| f_{\underline{m_n}}(x)^t \mathbf{T}_n^Y \left(\mathbf{T}_n^t \mathbf{T}_n\right)^{-1} \mathbf{T}_n^t \|$ then

$$
\begin{aligned}
&\mathbf{E} \sup_{x\in\mathcal{X}} \left| n^{-1/2} \sum_i f_{\underline{m_n}}(x)^t \mathbf{T}_n^Y \left(\mathbf{T}_n^t \mathbf{T}_n\right)^{-1} \mathbf{T}_n^t \left(f_{\underline{m_n}}(X_i) - \mathbf{E}[f_{\underline{m_n}}(X)] \right) \right|^2 \\
&\quad = \sup_{x\in\mathcal{X}} \| f_{\underline{m_n}}(x)^t \mathbf{T}_n^Y \left(\mathbf{T}_n^t \mathbf{T}_n\right)^{-1} \mathbf{T}_n^t \|^2 \, \mathbf{E} \left| n^{-1/2} \sum_i s(x)^t \left(f_{\underline{m_n}}(X_i) - \mathbf{E}[f_{\underline{m_n}}(X)] \right) \right|^2 \\
&\quad \leqslant \sup_{x\in\mathcal{X}} \| f_{\underline{m_n}}(x)^t \mathbf{T}_n^Y \|^2 \left\| \left(\mathbf{T}_n^t \mathbf{T}_n\right)^{-1} \right\| \mathbf{E} \max_{\|a\|=1} \left| a^t f_{\underline{m_n}}(X) \right|^2 \\
&\quad \leqslant C \tau_{k_n}^{-1} \sum_{l=1}^{k_n} \sup_{x\in\mathcal{X}} \left| \sum_{j=1}^{m_n} f_j(x) \, \mathbf{E}\left[Y e_l(Y) f_j(X) \right] \right|^2 \\
&\quad \leqslant C k_n \tau_{k_n}^{-1} \max_{1\leqslant l\leqslant k_n} \sup_{x\in\mathcal{X}} \left| (F_{m_n} v_l)(x) \right|^2
\end{aligned}
$$

where $v_l(x) = \mathbf{E}[Y e_l(Y)|X = x]$. Thus, $\mathcal{A}_{3,2} = O_p\left(\sqrt{k_n/(\tau_{k_n} \inf_x \mathcal{V}_{1n}(x))} \right) = O_p(\sqrt{k_n/(\tau_{k_n} m_n)})$ where the second equality holds under the assumption $\sup_{x\in\mathcal{X}} \| f_{\underline{m_n}}(x)^t \mathbf{T}_n^Y \| / \| \sqrt{\mathcal{V}_{1n}(x)} \| = O_p(\sqrt{k_n/m_n})$.

Finally, under Assumption 6 (i) we get

$$
\sup_{x\in\mathcal{X}} \frac{\sqrt{n}\,|III_n|}{\sqrt{\mathcal{V}_{1n}(x)}} = O_p\left(\sqrt{m_n} k_n^{-\beta} + \sqrt{n} \| T(E_{k_n} g - g) \|_X \right) = O_p\left((\sqrt{m_n} + \sqrt{n\tau_{k_n}}) k_n^{-\beta} \right)
$$

and

$$
\sup_{x\in\mathcal{X}} \frac{\sqrt{n}\,|IV_n|}{\sqrt{\mathcal{V}_{1n}(x)}} = O_p\left(\sqrt{n m_n} \| \varphi - F_{m_n}\varphi \|_X + \sqrt{n} \left\| \frac{\varphi - F_{m_n}\varphi}{\sqrt{\mathcal{V}_{1n}}} \right\|_\infty \right).
$$

By using the assumptions $k_n = o(\tau_k m_n)$, $m_n^2/n = o(1)$, $k_n \leqslant m_n$ and by eliminating the negligible rates, we can simplify the rate as

$$
\begin{aligned}
\sup_{x\in\mathcal{X}} \left| \sqrt{n/\widehat{\mathcal{V}}_{1n}(x)} \left(\widehat{\varphi}_n(x) - \varphi(x) \right) - \widehat{\mathbb{X}}_n(x) \right| &= O_p\Big(\frac{m_n}{\sqrt{n}} \sqrt{\frac{\max\{\log(m_n), k_n\}}{\tau_{k_n}}} + \eta_n c_n \\
&\quad + \sqrt{\frac{k_n}{\tau_{k_n} m_n}} + (\sqrt{m_n} + \sqrt{n\tau_{k_n}}) k_n^{-\beta} + b_{1,n} + b_{2,n} \Big) = o_p(c_n^{-1}) \quad \text{(C.4)}
\end{aligned}
$$

by using: $\sqrt{n m_n} \| \varphi - F_{m_n}\varphi \|_X = O(b_{2,n})$ and Assumption 8 *(iii)-(iv)* in the final line.

Step 2. Let $\Sigma_n = \mathbf{E}[f_{\underline{m_n}}(X) f_{\underline{m_n}}(X)^t (Yg(Y) - \varphi(X))^2]$. Because, under Assumptions 4 *(ii)* and 7 *(i)*,

$$
\sum_i \mathbf{E} \left\| \frac{1}{\sqrt{n}} f_{\underline{m_n}}(X_i)(Y_i g(Y_i) - \varphi(X_i)) \right\|^3 \leqslant \frac{1}{\sqrt{n}} m_n^{3/2} C
$$

then, by Pollard [2001, Theorem 10], if $\frac{m_n^2 \sqrt{m_n}}{r_n^3 \sqrt{n}} = o(1)$ for some sequence $r_n = o(1)$, there exists a sequence of $\mathcal{N}(0, \Sigma_n)$ random vectors $\mathfrak{X}_n$ such that

$$
\left\| \frac{1}{\sqrt{n}} \sum_i f_{\underline{m_n}}(X_i)(Y_i g(Y_i) - \varphi(X_i)) - \mathfrak{X}_n \right\| = o_p(r_n). \quad \text{(C.5)}
$$

Define the process $\mathbb{X}_n(x) = f_{\underline{m_n}}(x)^t \mathfrak{X}_n / \sqrt{\mathcal{V}_{1n}(x)}$, which is a centered Gaussian process with covariance function $\mathbf{E}[\mathbb{X}_n(x_1)\mathbb{X}_n(x_2)] = f_{\underline{m_n}}(x_1)^t \Sigma_n f_{\underline{m_n}}(x_2) / \sqrt{\mathcal{V}_{1n}(x_1)\mathcal{V}_{1n}(x_2)}$. Hence, by (C.5):

$$\sup_{x\in\mathcal{X}} \left| \widehat{\mathbb{X}}_n(x) - \mathbb{X}_n(x) \right| = o_p(r_n). \tag{C.6}$$

Step 3. In this step we approximate the bootstrap process by a Gaussian process. Let $\mathcal{U}_i = Y_i g(Y_i) - \varphi(X_i)$ and $\widehat{\mathcal{U}}_i = Y_i \widehat{g}(Y_i) - \widehat{\varphi}(X_i)$. Under the bootstrap distribution $\mathbb{P}^*$ each term $f_{\underline{m_n}}(X_i)\widehat{\mathcal{U}}_i \varepsilon_i / \sqrt{n}$ has mean zero $\forall i = 1, \ldots, n$. Moreover, define the matrix $\widehat{\Sigma}_n$ as

$$\sum_i \mathbf{E}^* \left[\frac{1}{n} f_{\underline{m_n}}(X_i) \widehat{\mathcal{U}}_i^2 \varepsilon_i^2 f_{\underline{m_n}}(X_i)^t \middle| D_n \right] = \frac{1}{n} \sum_i f_{\underline{m_n}}(X_i) f_{\underline{m_n}}(X_i)^t \widehat{\mathcal{U}}_i^2 = \widehat{\Sigma}_n$$

where $\mathbf{E}^*[\cdot|D_n]$ denotes the expectation under $\mathbb{P}^*$. Since $\mathbf{E}^*[\varepsilon_i|D_n] < \infty$ uniformly in i, we have, under Assumptions 4 *(ii)* and 7 *(i)*, for some generic constant C

$$\sum_i \mathbf{E}^* \left[\left\| \frac{1}{\sqrt{n}} f_{\underline{m_n}}(X_i) \widehat{\mathcal{U}}_i \varepsilon_i \right\|^3 \middle| D_n \right] \leqslant C \frac{m_n \sqrt{m_n}}{\sqrt{n}} \left(\|\widehat{g} - g\|_\infty^3 \frac{1}{n} \sum_i |Y_i|^3 + C + \|\widehat{\varphi} - \varphi\|_\infty^3 \right) = O\left(\frac{m_n \sqrt{m_n}}{\sqrt{n}} \right)$$

under the assumptions of the theorem. Then, an application of Pollard [2001, Theorem 10], conditional on the data, yields that, if $\frac{\sqrt{m_n} m_n^2}{r_n^3 \sqrt{n}} = o(1)$ with $r_n = o(1)$, then there exists a sequence of $\mathcal{N}(0, \widehat{\Sigma}_n)$ random vectors $\mathfrak{X}_n^*$ such that

$$\left\| \frac{1}{\sqrt{n}} \sum_i f_{\underline{m_n}}(X_i) \widehat{\mathcal{U}}_i - \mathfrak{X}_n^* \right\| = o_{p^*}(r_n) \tag{C.7}$$

with probability approaching 1. Therefore,

$$\sup_{x\in\mathcal{X}} \left| \mathbb{X}_n^*(x) - \frac{f_{\underline{m_n}}(x)^t \widehat{Q}_n^{-1}}{\sqrt{\widehat{\mathcal{V}}_{1n}(x)}} \mathfrak{X}_n^* \right| = o_{p^*}(r_n)$$

with probability approaching 1. Define a centered Gaussian process $\widetilde{\mathbb{X}}_n(\cdot)$ under $\mathbb{P}^*$ as

$$\widetilde{\mathbb{X}}_n(x) = \frac{f_{\underline{m_n}}(x)^t}{\sqrt{\mathcal{V}_{1n}(x)}} \Sigma_n^{1/2} \widehat{\Sigma}_n^{-1/2} \mathfrak{X}_n^*$$

which has the same covariance function as $\mathbb{X}_n(x)$ (since $\widehat{\Sigma}_n$ is invertible with probability one). By Lemma C.1 below we have:

$$\sup_{x\in\mathcal{X}} \left| \frac{f_{\underline{m_n}}(x)^t \widehat{Q}_n^{-1}}{\sqrt{\widehat{\mathcal{V}}_{1n}(x)}} \mathfrak{X}_n^* - \widetilde{\mathbb{X}}_n(x) \right| = o_{p^*}(c_n^{-1})$$

with probability approaching one. This and the previous convergence imply that

$$\sup_{x\in\mathcal{X}} \left| \mathbb{X}_n^*(x) - \widetilde{\mathbb{X}}_n(x) \right| = o_{p^*}(r_n) + o_{p^*}(c_n^{-1}). \tag{C.8}$$

Step 4. Given the results (C.4), (C.6) and (C.8), this last step proceeds exactly as Part 4 in the proof of Theorem B.1 in Chen and Christensen [2015a] and thus we omit it. □

LEMMA C.1. *Let Assumptions 1 – 5 and 7-8 be satisfied. Moreover, assume that $\mathcal{Y}$ is bounded. Then,*

$$\sup_{x\in\mathcal{X}}\left|\frac{f_{\underline{m}_n}(x)^t\widehat{Q}_n^{-1}}{\sqrt{\widehat{\mathcal{V}}_{1n}(x)}}\mathfrak{X}_n^* - \widetilde{\mathbb{X}}_n(x)\right| = o_{p^*}(c_n^{-1})$$

with probability approaching one, where $\mathfrak{X}_n^*$ and $\widetilde{\mathbb{X}}_n(x)$ are as defined in the proof of Theorem 3.4.

Proof. The proof of this lemma follows the proof of Chen and Christensen [2015a, Lemma D.8] and so we provide only the main parts where the two proofs differ. Let Σ_n and $\widehat{\Sigma}$ be as defined in steps 2 and 3 of the proof of Theorem 3.4. We make the decomposition

$$\begin{aligned}\sup_{x\in\mathcal{X}}\left|\frac{f_{\underline{m}_n}(x)^t\widehat{Q}_n^{-1}}{\sqrt{\widehat{\mathcal{V}}_{1n}(x)}}\mathfrak{X}_n^* - \widetilde{\mathbb{X}}_n(x)\right| &\leqslant \sup_{x\in\mathcal{X}}\left|\frac{f_{\underline{m}_n}(x)^t\left(\widehat{Q}_n^{-1}-\Sigma_n^{1/2}\widehat{\Sigma}_n^{-1/2}\right)}{\sqrt{\mathcal{V}_{1n}(x)}}\mathfrak{X}_n^*\right|\sup_{x\in\mathcal{X}}\frac{\sqrt{\mathcal{V}_{1n}(x)}}{\sqrt{\widehat{\mathcal{V}}_{1n}(x)}}\\ &+\sup_{x\in\mathcal{X}}\left|\frac{\sqrt{\mathcal{V}_{1n}(x)}}{\sqrt{\widehat{\mathcal{V}}_{1n}(x)}}-1\right|\sup_{x\in\mathcal{X}}\left|\frac{f_{\underline{m}_n}(x)^t}{\sqrt{\mathcal{V}_{1n}(x)}}\Sigma_n^{1/2}\widehat{\Sigma}_n^{-1/2}\mathfrak{X}_n^*\right| = \mathcal{B}_1+\mathcal{B}_2 \qquad \text{(say).} \quad \text{(C.9)}\end{aligned}$$

We start with the analysis of term $\mathcal{B}_1$. Denote $\mathfrak{D}_n(x_1,x_2)$ the standard deviation semimetric on $\mathcal{X}$ associated with the Gaussian process (under $\mathbb{P}^*$) $\frac{f_{\underline{m}_n}(x)^t\left(\widehat{Q}_n^{-1}-\Sigma_n^{1/2}\widehat{\Sigma}_n^{-1/2}\right)}{\sqrt{\mathcal{V}_{1n}(x)}}\mathfrak{X}_n^*$ and defined as

$$\mathfrak{D}_n(x_1,x_2)^2 = \mathbf{E}^*\Big[\Big(\Big(\frac{f_{\underline{m}_n}(x_1)^t}{\sqrt{\mathcal{V}_{1n}(x_1)}}-\frac{f_{\underline{m}_n}(x_2)^t}{\sqrt{\mathcal{V}_{1n}(x_2)}}\Big)(\widehat{Q}_n^{-1}-\Sigma_n^{1/2}\widehat{\Sigma}_n^{-1/2})\mathfrak{X}_n^*\Big)^2\Big].$$

Therefore, $\mathfrak{D}_n(x_1,x_2)\leqslant d_n(x_1,x_2)\|(\widehat{Q}_n^{-1}-\Sigma_n^{1/2}\widehat{\Sigma}_n^{-1/2})\|$ and

$$\begin{aligned}\|(\widehat{Q}_n^{-1}-\Sigma_n^{1/2}\widehat{\Sigma}_n^{-1/2})\| &\leqslant \|\widehat{Q}_n^{-1}-I_{m_n}\|+\|\widehat{\Sigma}_n^{1/2}-\Sigma_n^{1/2}\|\,\|\widehat{\Sigma}_n^{-1/2}\|\\ &\leqslant \|\widehat{Q}_n^{-1}\|\,\|\widehat{Q}_n-I_{m_n}\|+(\lambda_{\min}^{1/2}(\Sigma_n)+\lambda_{\min}^{1/2}(\widehat{\Sigma}_n))^{-1}\|\widehat{\Sigma}_n-\Sigma_n\|\\ &= O_p(\sqrt{m_n\log(m_n)/n}+\eta_n) \quad \text{(C.10)}\end{aligned}$$

where η_n is as defined in the proof of Theorem 3.4, to get the inequality in the second line we have used Lemma E.3 in Chen and Christensen [2015a] and to get the rates in the last line we have used: $\|\widehat{Q}_n-I_{m_n}\| = O_p(\sqrt{m_n\log(m_n)/n})$ and $\|\widehat{\Sigma}_n-\Sigma_n\| = O_p(\sqrt{m_n\log(m_n)/n}+\eta_n)$, obtained by using twice Theorem 1.6 in Tropp [2012]. By using similar arguments as in the proof of Lemma D.8 in Chen and Christensen [2015a, page 37] we obtain that $\mathcal{B}_1 = O_{p^*}(\eta_n c_n) = o_{p^*}(c_n^{-1})$ under Assumption 8 *(iii) - (iv)*.
Next, let us consider term $\mathcal{B}_2$ which is the supremum of a Gaussian process with the same distribution (under $\mathbb{P}^*$) as $\mathbb{X}_n(x)$ (under the data distribution). Therefore, by applying Lemma A.1 and Chen and Christensen [2015a, LemmaD.7], which is valid under Assumptions 4 *(i)*, 7 *(i)* and 8 *(i) - (ii)*, we obtain that $\mathcal{B}_2 = O_p(\eta_n)O_{p^*}(c_n)$. □

PROOF OF THEOREM 4.3. The proof is made of four steps and we use all along the proof the inequality $\mathcal{W}_n(x)\geqslant c\,\|\left(\mathbf{K}_n^t\mathbf{K}_n\right)^{-1/2}p_{\underline{k}_n}(z)\|^2$ which is valid under Assumption 11 *(i)*. As in the

proof of Theorem 4.2, denote $\widehat{A}_n = \left(\mathbf{Z}_{k_n}^t \mathbf{X}_{m_n}(\mathbf{X}_{m_n}^t \mathbf{X}_{m_n})^{-1} \mathbf{X}_{m_n}^t \mathbf{Z}_{k_n}/n\right)^{-1} \mathbf{Z}_{k_n}^t \mathbf{X}_{m_n}(\mathbf{X}_{m_n}^t \mathbf{X}_{m_n})^{-1}$ and $A_n = (\mathbf{K}_n^t \mathbf{K}_n)^{-1}\mathbf{K}_n^t$ where $\mathbf{K}_n = \mathbf{E}[f_{\underline{m_n}}(X) p_{\underline{k_n}}(Z)^t]$. Moreover, let $\mathcal{D}_n = \{(\Delta_1, Y_1, X_1, Z_1), \ldots, (\Delta_n, Y_n, X_n, Z_n)\}$.

Step 1. We start by showing that $\sqrt{n/\widehat{\mathcal{W}}_n(z)}\left(\widehat{\psi}_n(z) - \psi(z)\right)$ can be uniformly approximated by the process

$$\widehat{\mathbb{Z}}_n(z) = \frac{p_{k_n}(z)^t A_n}{\sqrt{\mathcal{W}_n(z)}} \frac{1}{\sqrt{n}} \sum_i f_{\underline{m_n}}(X_i)(Y_i g(Y_i) - \psi(Z_i)).$$

Let I_n, II_n, III_n, IV_n be as defined in (A.9), then from the decomposition (A.9) we can write

$$\begin{aligned}\left|\sqrt{n/\widehat{\mathcal{W}}_n(z)}\left(\widehat{\psi}_n(z) - \psi(z)\right) - \widehat{\mathbb{Z}}_n(z)\right| &\leqslant \left|\frac{\sqrt{n} I_n}{\sqrt{\mathcal{W}_n(z)}} - \widehat{\mathbb{Z}}_n(z)\right| + \frac{\left|\sqrt{n} I_n\right|}{\sqrt{\mathcal{W}_n(z)}}\left|\frac{\sqrt{\mathcal{W}_n(z)}}{\sqrt{\widehat{\mathcal{W}}_n(z)}} - 1\right| \\ &+ \sqrt{n}\frac{|II_n + III_n + IV_n|}{\sqrt{\widehat{\mathcal{W}}_n(z)}} = \mathfrak{T}_{n,1}(z) + \mathfrak{T}_{n,2}(z) + \mathfrak{T}_{n,3}(z) \qquad \text{(say).} \quad \text{(C.11)}\end{aligned}$$

Let us consider term $\mathfrak{T}_{n,1}(z)$ and denote $\eta_{n,3} = m_n\sqrt{\log k_n/(n\kappa_{k_n})}$,

$$\begin{aligned}\sup_{z\in\mathcal{Z}} \mathfrak{T}_{n,1}(z) &\leqslant \sup_{z\in\mathcal{Z}} \frac{\|p_{\underline{k_n}}(z)^t(\widehat{A}_n - A_n)Q_n^{1/2}\|}{\sqrt{\mathcal{W}_n(z)}} \left\|\frac{Q_n^{-1/2}}{\sqrt{n}} \sum_{i=1}^n f_{\underline{m_n}}(X_i)(Y_i g(Y_i) - \psi(Z_i))\right\| \\ &= O_p\left(\kappa_{k_n}^{-1/2} \max\left(\sqrt{m_n}, \sqrt{k_n}\right)\sqrt{\frac{m_n \log k_n}{n}}\right) = O_p\left(\eta_{n,3}\right). \quad \text{(C.12)}\end{aligned}$$

Next, we consider term $\mathfrak{T}_{n,2}(z)$. Lemma A.2 implies that $\sup_{z\in\mathcal{Z}}\left|\frac{\sqrt{\mathcal{W}_n(z)}}{\sqrt{\widehat{\mathcal{W}}_n(z)}} - 1\right| = O_p(\widetilde{\eta}_n)$, where

$$\widetilde{\eta}_n = \eta_{n,3} + \frac{k_n}{\sqrt{n\min(\tau_{k_n}, \kappa_{k_n}^2)}} + k_n^{-\beta} + k_n^{-\gamma/d_z} + \frac{\sqrt{k_n}\|T(E_{k_n}g - g)\|_X}{\sqrt{\kappa_{k_n}}}.$$

Therefore,

$$\begin{aligned}\sup_{z\in\mathcal{Z}} \mathfrak{T}_{n,2}(z) &= \sup_{z\in\mathcal{Z}} \mathfrak{T}_{n,1}(z)\left|\frac{\sqrt{\mathcal{W}_n(z)}}{\sqrt{\widehat{\mathcal{W}}_n(z)}} - 1\right| + \sup_{z\in\mathcal{Z}}|\widehat{\mathbb{Z}}_n(z)|\left|\frac{\sqrt{\mathcal{W}_n(z)}}{\sqrt{\widehat{\mathcal{W}}_n(z)}} - 1\right| \\ &= O_p(\widetilde{\eta}_n)\left(O_p(\eta_{n,3}) + o_p(r_n) + \sup_{z\in\mathcal{Z}}|\mathbb{Z}_n(z)|\right) \\ &= O_p(\widetilde{\eta}_n)\left(O_p(\eta_{n,3}) + o_p(r_n) + O_p(c_n)\right). \quad \text{(C.13)}\end{aligned}$$

where the second line is due to step 2 below and the last line is due to Chen and Christensen [2015a, LemmaD.7], which is valid under Assumptions 9 *(i)*, 11 *(i)* and 12 *(i)* – *(ii)*, and that implies $\sup_{z\in\mathcal{Z}}|\mathbb{Z}_n(z)| = O_p(c_n)$.

Let us analyze term $\mathfrak{T}_{n,3}(z)$. Denote $\widehat{\mathbf{T}}_n^Y = \sum_{i=1}^n f_{\underline{m_n}}(X_i) Y_i e_{\underline{k_n}}(Y_i)^t/n$ and $\beta_{k_n} = (\mathbf{T}_n^t \mathbf{T}_n)^{-1}\mathbf{T}_n\,\mathbf{E}[f_{\underline{m_n}}(X)]$.

First, we use the decomposition

$$\sup_{z\in\mathcal{Z}}\frac{\sqrt{n}\,|II_n|}{\sqrt{\mathcal{W}_n(z)}}=\sup_{z\in\mathcal{Z}}\frac{\sqrt{n}\left|p_{\underline{k_n}}(z)^t(\widehat{A}_n-A_n)\widehat{\mathbf{T}}_n^Y(\widehat{\beta}_{k_n}-\mathbf{E}[g(Y)e_{\underline{k_n}}(Y)])\right|}{\sqrt{\mathcal{W}_n(z)}}$$
$$+\sup_{z\in\mathcal{Z}}\frac{\sqrt{n}\left|p_{\underline{k_n}}(z)^tA_n(\widehat{\mathbf{T}}_n^Y-\mathbf{T}_n^Y)(\widehat{\beta}_{k_n}-\mathbf{E}[g(Y)e_{\underline{k_n}}(Y)])\right|}{\sqrt{\mathcal{W}_n(z)}}$$
$$+\sup_{z\in\mathcal{Z}}\frac{\sqrt{n}\left|p_{\underline{k_n}}(z)^tA_n\mathbf{T}_n^Y(\widehat{\beta}_{k_n}-\beta_{k_n})\right|}{\sqrt{\mathcal{W}_n(z)}}$$
$$+\sup_{z\in\mathcal{Z}}\frac{\sqrt{n}\left|p_{\underline{k_n}}(z)^tA_n\mathbf{T}_n^Y(\mathbf{T}_n^t\mathbf{T}_n)^{-1}\mathbf{T}_n^t\,\mathbf{E}[f_{\underline{m_n}}(X)\Delta(g(Y)-E_{k_n}g(Y))]\right|}{\sqrt{\mathcal{W}_n(z)}}=\mathcal{A}_1+\mathcal{A}_2+\mathcal{A}_3+\mathcal{A}_4\qquad\text{(say)}.\tag{C.14}$$

Terms $\mathcal{A}_i$, $i=1,2,4$ are easily controlled by using the Cauchy Schwartz-inequality so that:

$$\mathcal{A}_1 = O_p(\sqrt{m_n}\sqrt{\frac{\log k_n}{\kappa_{k_n}n}}\left(\sqrt{\frac{m_n\log(k_n)}{n}}\frac{\max\{\sqrt{m_n},\xi_n\}}{\tau_{k_n}}+\sqrt{\frac{m_n}{\tau_{k_n}}}+\sqrt{\frac{k_nn}{m_n\tau_{k_n}}}\|T(g-E_{k_n}g)\|_X\right)),$$
$$\mathcal{A}_2 = O_p(\sqrt{m_nk_n/n}(\sqrt{m_n\log(k_n)/n}\max\{\sqrt{m_n},\xi_n\}/\tau_{k_n}+\sqrt{m_n/\tau_{k_n}}+\sqrt{n}/\tau_{k_n}\|T(g-E_{k_n}g)\|_X)),$$
$$\mathcal{A}_4 = O_p(\sqrt{n}\|T(g-E_{k_n}g)\|_X)$$

where, to get $\mathcal{A}_4$, we have used the Cauchy Schwartz inequality and the fact that the largest eigenvalue of $\mathbf{T}_n^Y(\mathbf{T}_n^t\mathbf{T}_n)^{-1}\mathbf{T}_n^t$ is bounded. Term $\mathcal{A}_3$ can be decomposed as

$$\mathcal{A}_3=\sup_{z\in\mathcal{Z}}\frac{\left|p_{\underline{k_n}}(z)^tA_n\mathbf{T}_n^Y[(\widehat{\mathbf{T}}_n^t\widehat{Q}_n^{-1}\widehat{\mathbf{T}}_n)^{-1}\widehat{\mathbf{T}}_n^t\widehat{Q}_n^{-1}-(\mathbf{T}_n^t\mathbf{T}_n)^{-1}\mathbf{T}_n^t]\sum_i f_{\underline{m_n}}(X_i)/\sqrt{n}\right|}{\sqrt{\mathcal{W}_n(z)}}$$
$$+\sup_{z\in\mathcal{Z}}\frac{\left|p_{\underline{k_n}}(z)^tA_n\mathbf{T}_n^Y(\mathbf{T}_n^t\mathbf{T}_n)^{-1}\mathbf{T}_n^t\sum_i(f_{\underline{m_n}}(X_i)-\mathbf{E}[f_{\underline{m_n}}(X)])/\sqrt{n}\right|}{\sqrt{\mathcal{W}_n(z)}}=\mathcal{A}_{3,1}+\mathcal{A}_{3,2}$$

where

$$\mathcal{A}_{3,1}\leqslant\sup_{z\in\mathcal{Z}}\frac{\left\|p_{\underline{k_n}}(z)^tA_n\right\|}{\sqrt{\mathcal{W}_n(z)}}\|\mathbf{T}_n^Y(\mathbf{T}_n^t\mathbf{T}_n)^{-1}\mathbf{T}_n^t\|\left\|\mathbf{T}_n[(\widehat{\mathbf{T}}_n^t\widehat{Q}_n^{-1}\widehat{\mathbf{T}}_n)^{-1}\widehat{\mathbf{T}}_n^t\widehat{Q}_n^{-1}-(\mathbf{T}_n^t\mathbf{T}_n)^{-1}\mathbf{T}_n^t]\sum_i f_{\underline{m_n}}(X_i)/\sqrt{n}\right\|$$
$$=O_p\left(\max\{\sqrt{m_n},\xi_n\}\sqrt{\frac{m_n\log k_n}{n\tau_{k_n}}}\right).$$

To control term $\mathcal{A}_{3,2}$, let $s(z)^t = p_{\underline{k_n}}(z)^t A_n \mathbf{T}_n^Y \left(\mathbf{T}_n^t \mathbf{T}_n\right)^{-1} \mathbf{T}_n^t / \|p_{\underline{k_n}}(z)^t A_n \mathbf{T}_n^Y \left(\mathbf{T}_n^t \mathbf{T}_n\right)^{-1} \mathbf{T}_n^t\|$ then

$$\begin{aligned}
&\mathbf{E} \sup_{z\in\mathcal{Z}} \left| n^{-1/2} \sum_i p_{\underline{k_n}}(z)^t A_n \mathbf{T}_n^Y \left(\mathbf{T}_n^t \mathbf{T}_n\right)^{-1} \mathbf{T}_n^t \left(f_{\underline{m_n}}(X_i) - \mathbf{E}[f_{\underline{m_n}}(X)]\right)\right|^2 \\
&\quad = \sup_{z\in\mathcal{Z}} \|p_{\underline{k_n}}(z)^t A_n \mathbf{T}_n^Y \left(\mathbf{T}_n^t \mathbf{T}_n\right)^{-1} \mathbf{T}_n^t\|^2 \, \mathbf{E} \left| n^{-1/2} \sum_i s(z)^t \left(f_{\underline{m_n}}(X_i) - \mathbf{E}[f_{\underline{m_n}}(X)]\right)\right|^2 \\
&\quad \leqslant C\kappa_{k_n}^{-1} \sup_{z\in\mathcal{Z}} \|p_{\underline{k_n}}(z)^t\|^2 \, \mathbf{E} \max_{\|a\|=1} \left| a^t f_{\underline{m_n}}(X)\right|^2 \\
&\quad = O(k_n \kappa_{k_n}^{-1})
\end{aligned}$$

where we have used the fact that the largest eigenvalue of $\mathbf{T}_n^Y (\mathbf{T}_n^t \mathbf{T}_n)^{-1} \mathbf{T}_n^t$ is bounded and $\|A_n\|^2 \leqslant c\kappa_{k_n}^{-1}$. Thus, $\mathcal{A}_{3,2} = O\Big(\sqrt{k_n/(\kappa_{k_n} \inf_z \mathcal{W}_n(z))}\Big)$.

Finally,

$$\begin{aligned}
\sup_{z\in\mathcal{Z}} \frac{\sqrt{n}\,|III_n|}{\sqrt{\mathcal{W}_n(z)}} &= O_p\left(\left(\sqrt{\frac{m_n \log k_n}{n\kappa_{k_n}}} + 1 \right) \left(\sqrt{m_n} k_n^{-\beta} + \sqrt{n} \|T(E_{k_n} g - g)\|_X \right) \right) \\
&= O_p\left(\left(\sqrt{\frac{m_n \log k_n}{n\kappa_{k_n}}} + 1 \right) \left(\sqrt{m_n} + \sqrt{n\tau_{k_n}} \right) k_n^{-\beta} \right)
\end{aligned}$$

and (by using the fact that $\widehat{A}_n \mathbf{X}_{m_n}^t \mathbf{Z}_{k_n} / n = I_{k_n}$)

$$\begin{aligned}
&\sup_{z\in\mathcal{Z}} \sqrt{n} \frac{|IV_n|}{\sqrt{\mathcal{W}_n(z)}} \\
&= \sup_{z\in\mathcal{Z}} \frac{|p_{\underline{k}}(z)^t \widehat{A}_n \left(\sum_{i=1}^n f_{\underline{m}}(X_i)\psi(Z_i)/\sqrt{n} - \sqrt{n} \mathbf{X}_{m_n}^t \mathbf{Z}_{k_n}/n \, \mathbf{E}[p_{\underline{k}}(Z)\psi(Z)] \right) - \sqrt{n}(\psi(Z) - (\Pi_{k_n}\psi)(Z))|}{\sqrt{\mathcal{W}_n(z)}} \\
&\leqslant \sup_{z\in\mathcal{Z}} \frac{\|p_{\underline{k}}(z)^t \widehat{A}_n\| \left\| \sum_{i=1}^n f_{\underline{m}}(X_i)[\psi(Z_i) - (\Pi_{k_n}\psi)(Z_i)]/\sqrt{n} \right\|}{\sqrt{\mathcal{W}_n(z)}} + \sup_{z\in\mathcal{Z}} \sqrt{n} \frac{|(\psi(z) - (\Pi_{k_n}\psi)(z))|}{\sqrt{\mathcal{W}_n(z)}} \\
&= O_p\left(\left(\sqrt{\frac{m_n \log k_n}{n\kappa_{k_n}}} + 1 \right) \sqrt{n} \|\psi - \Pi_{k_n}\psi\|_Z \right) + \sup_{z\in\mathcal{Z}} \sqrt{n} \left| \frac{(\psi(z) - (\Pi_{k_n}\psi)(z))}{\sqrt{\mathcal{W}_n(z)}} \right|.
\end{aligned}$$

By using the assumptions $m_n \sqrt{\log k_n} = o(\sqrt{n\kappa_{k_n}})$, $k_n \max(m_n, \xi_n^2) = o(n\tau_{k_n})$ and by eliminating the negligible rates, we can simplify the rate as

$$\begin{aligned}
\left| \sqrt{n/\widehat{\mathcal{W}}_n(z)} \left(\widehat{\psi}_n(z) - \psi(z) \right) - \widehat{\mathbb{Z}}_n(z) \right| &= O_p\Big(m_n \sqrt{\frac{\log k_n}{n\kappa_{k_n}\tau_{k_n}}} + \widetilde{\eta}_n c_n + \frac{k_n}{\kappa_{k_n} \inf_z \mathcal{W}_n(z)} + \widetilde{b}_{1,n} \\
&\quad + \left(\sqrt{m_n} + \sqrt{\frac{n}{\tau_{k_n}}} \right) k_n^{-\beta} + \widetilde{b}_{2,n} \Big) = o_p(\widetilde{c}_n^{-1}) \quad \text{(C.15)}
\end{aligned}$$

by using $\sqrt{nm_n}\|\psi - \Pi_{k_n}\psi\|_Z = O(\widetilde{b}_{2,n})$ and Assumption 12 *(iii)* – *(iv)* in the final line.

Step 2. Let $\Sigma_n^\psi = \mathbf{E}[f_{\underline{m_n}}(X) f_{\underline{m_n}}(X)^t (Yg(Y) - \psi(Z))^2]$. Because, under Assumptions 4 *(ii)* and

11 *(i)*,

$$\sum_i \mathbf{E}\left\| \frac{1}{\sqrt{n}} f_{\underline{m_n}}(X_i)(Y_i g(Y_i) - \psi(Z_i)) \right\|^3 \leqslant \frac{1}{\sqrt{n}} m_n^{3/2} C$$

then, by Pollard [2001, Theorem 10], if $\frac{m_n^2 \sqrt{m_n}}{r_n^3 \sqrt{n}} = o(1)$ for some sequence $r_n = o(1)$, there exists a sequence of $\mathcal{N}(0, \Sigma_n^{\psi})$ random vectors $\mathfrak{Z}_n$ such that

$$\left\| \frac{1}{\sqrt{n}} \sum_i f_{\underline{m_n}}(X_i)(Y_i g(Y_i) - \psi(Z_i)) - \mathfrak{Z}_n \right\| = o_p(r_n). \tag{C.16}$$

Define the process $\mathbb{Z}_n(z) = p_{\underline{k_n}}(z)^t A_n \mathfrak{Z}_n / \sqrt{\mathcal{W}_n(z)}$, which is a centered Gaussian process with covariance function $\mathbf{E}[\mathbb{Z}_n(z_1)\mathbb{Z}_n(z_2)] = p_{\underline{k_n}}(z_1)^t A_n \Sigma_n A_n^t p_{\underline{k_n}}(z_2) / \sqrt{\mathcal{W}_n(z_1)\mathcal{W}_n(z_2)}$. Hence, by (C.16):

$$\sup_{z \in \mathcal{Z}} \left| \widehat{\mathbb{Z}}_n(z) - \mathbb{Z}_n(z) \right| = o_p(r_n). \tag{C.17}$$

Step 3. In this step we approximate the bootstrap process by a Gaussian process. Let $\mathcal{U}_i = Y_i g(Y_i) - \psi(Z_i)$ and $\widehat{\mathcal{U}}_i = Y_i \widehat{g}(Y_i) - \widehat{\psi}(Z_i)$. Under the bootstrap distribution $\mathbb{P}^*$ each term $f_{\underline{m_n}}(X_i)\widehat{\mathcal{U}}_i \varepsilon_i / \sqrt{n}$ has mean zero $\forall i = 1, \ldots, n$. Moreover, define the matrix $\widehat{\Sigma}_n^{\psi}$ as

$$\sum_i \mathbf{E}^*\left[\frac{1}{n} f_{\underline{m_n}}(X_i) \widehat{\mathcal{U}}_i^2 \varepsilon_i^2 f_{\underline{m_n}}(X_i)^t \middle| \mathcal{D}_n \right] = \frac{1}{n} \sum_i f_{\underline{m_n}}(X_i) f_{\underline{m_n}}(X_i)^t \widehat{\mathcal{U}}_i^2 = \widehat{\Sigma}_n^{\psi}$$

where $\mathbf{E}^*[\cdot|\mathcal{D}_n]$ denotes the expectation taken with respect to $\mathbb{P}^*$. Since $\mathbf{E}^*[\varepsilon_i|\mathcal{D}_n] < \infty$ uniformly in i, we have, under Assumptions 4 *(ii)* and 11 *(i)*, for some generic constant C

$$\sum_i \mathbf{E}^*\left[\left\| \frac{1}{\sqrt{n}} f_{\underline{m_n}}(X_i) \widehat{\mathcal{U}}_i \varepsilon_i \right\|^3 \middle| \mathcal{D}_n \right] \leqslant C \frac{m_n \sqrt{m_n}}{\sqrt{n}} \left(\|\widehat{g} - g\|_\infty^3 \frac{1}{n} \sum_i |Y_i|^3 + C + \|\widehat{\psi} - \psi\|_\infty^3 \right) = O\left(\frac{m_n \sqrt{m_n}}{\sqrt{n}} \right)$$

under the assumptions of the theorem. Then, an application of Pollard [2001, Theorem 10], conditional on the data, yields that, if $\frac{\sqrt{m_n} m_n^2}{r_n^3 \sqrt{n}} = o(1)$ with $r_n = o(1)$, then there exists a sequence of $\mathcal{N}(0, \widehat{\Sigma}_n^{\psi})$ random vectors $\mathfrak{Z}_n^*$ such that

$$\left\| \frac{1}{\sqrt{n}} \sum_i f_{\underline{m_n}}(X_i) \widehat{\mathcal{U}}_i - \mathfrak{Z}_n^* \right\| = o_{p^*}(r_n) \tag{C.18}$$

with probability approaching 1. Therefore,

$$\sup_{z \in \mathcal{Z}} \left| \mathbb{Z}_n^*(z) - \frac{p_{\underline{k_n}}(z)^t \widehat{A}_n}{\sqrt{\widehat{\mathcal{W}}_n(z)}} \mathfrak{Z}_n^* \right| = o_{p^*}(r_n)$$

with probability approaching 1. Define a centered Gaussian process $\widetilde{\mathbb{Z}}_n$ under $\mathbb{P}^*$ as

$$\widetilde{\mathbb{Z}}_n(z) = \frac{p_{\underline{k_n}}(z)^t A_n}{\sqrt{\mathcal{W}_n(z)}} (\Sigma_n^{\psi})^{1/2} (\widehat{\Sigma}_n^{\psi})^{-1/2} \mathfrak{Z}_n^*$$

which has the same covariance function as $\mathbb{Z}_n(z)$ (since $\widehat{\Sigma}_n^{\psi}$ is invertible with probability one). By Lemma C.2 below we have:

$$\sup_{z\in\mathcal{Z}}\left|\frac{p_{\underline{k}_n}(z)^t\widehat{A}_n}{\sqrt{\widehat{\mathcal{W}}_n(z)}}\mathfrak{Z}_n^* - \widetilde{\mathbb{Z}}_n(z)\right| = o_{p^*}(\widetilde{c}_n^{-1})$$

with probability approaching one. This and the previous convergence imply that

$$\sup_{z\in\mathcal{Z}}\left|\mathbb{Z}_n^*(z) - \widetilde{\mathbb{Z}}_n(z)\right| = o_{p^*}(r_n) + o_{p^*}(\widetilde{c}_n^{-1}). \tag{C.19}$$

Step 4. Given the results (C.15), (C.17) and (C.19), this last step proceeds exactly as Part 4 in the proof of Theorem B.1 in Chen and Christensen [2015a] and thus we omit it. □

LEMMA C.2. *Let the assumptions of Theorem 4.2 and Assumption 12 hold. If* $m_n\sqrt{\log k_n} = o(\sqrt{n\kappa_{k_n}})$, $k_n\max(m_n,\xi_n^2) = o(n\tau_{k_n})$. *Then,*

$$\sup_{z\in\mathcal{Z}}\left|\frac{p_{\underline{k}_n}(z)^t\widehat{A}_n}{\sqrt{\widehat{\mathcal{W}}_n(z)}}\mathfrak{Z}_n^* - \widetilde{\mathbb{Z}}_n(z)\right| = o_{p^*}(c_n^{-1})$$

with probability approaching one, where $\mathfrak{Z}_n^*$ *and* $\widetilde{\mathbb{Z}}_n(z)$ *are as defined in the proof of Theorem 4.3.*

Proof. The proof of this lemma follows the proof of Chen and Christensen [2015a, Lemma D.8] and so we provide only the main parts where the two proofs differ. Let Σ_n^{ψ} and $\widehat{\Sigma}_n^{\psi}$ be as defined in steps 2 and 3 of the proof of Theorem 4.3. We make the decomposition

$$\begin{aligned}\sup_{z\in\mathcal{Z}}\left|\frac{p_{\underline{k}_n}(z)^t\widehat{A}_n}{\sqrt{\widehat{\mathcal{W}}_n(z)}}\mathfrak{Z}_n^* - \widetilde{\mathbb{Z}}_n(z)\right| &\leqslant \sup_{z\in\mathcal{Z}}\left|\frac{p_{\underline{k}_n}(z)^t\left(\widehat{A}_n - A_n(\Sigma_n^{\psi})^{1/2}(\widehat{\Sigma}_n^{\psi})^{-1/2}\right)}{\sqrt{\mathcal{W}_n(z)}}\mathfrak{Z}_n^*\right|\sup_{z\in\mathcal{Z}}\sqrt{\frac{\mathcal{W}_n(z)}{\widehat{\mathcal{W}}_n(z)}}\\ &+\sup_{z\in\mathcal{Z}}\left|\sqrt{\frac{\mathcal{W}_n(z)}{\widehat{\mathcal{W}}_n(z)}}-1\right|\sup_{z\in\mathcal{Z}}\left|\frac{p_{\underline{k}_n}(z)^tA_n}{\sqrt{\mathcal{W}_n(z)}}(\Sigma_n^{\psi})^{1/2}(\widehat{\Sigma}_n^{\psi})^{-1/2}\mathfrak{Z}_n^*\right| = \mathcal{B}_1+\mathcal{B}_2. \end{aligned}\tag{C.20}$$

We start with the analysis of term $\mathcal{B}_1$. Denote $\widetilde{\mathfrak{D}}_n(z_1,z_2)$ the standard deviation norm on $\mathcal{Z}$ associated with the Gaussian process (under $\mathbb{P}^*$) $\frac{p_{\underline{k}_n}(z)^t\left(\widehat{A}_n - A_n(\Sigma_n^{\psi})^{1/2}(\widehat{\Sigma}_n^{\psi})^{-1/2}\right)}{\sqrt{\mathcal{W}_n(z)}}\mathfrak{Z}_n^*$ and defined as

$$\widetilde{\mathfrak{D}}_n(z_1,z_2)^2 = \mathbf{E}^*[\left(\left(\frac{p_{\underline{k}_n}(z_1)^tA_n}{\sqrt{\mathcal{W}_n(z_1)}} - \frac{p_{\underline{k}_n}(z_2)^tA_n}{\sqrt{\mathcal{W}_n(z_2)}}\right)\mathbf{K}_n(\widehat{A}_n - A_n(\Sigma_n^{\psi})^{1/2}(\widehat{\Sigma}_n^{\psi})^{-1/2})\mathcal{Z}_n^*\right)^2].$$

Therefore, $\widetilde{\mathfrak{D}}_n(z_1,z_2) \leqslant \widetilde{d}_n(z_1,z_2)\|\mathbf{K}_n(\widehat{A}_n - A_n(\Sigma_n^{\psi})^{1/2}(\widehat{\Sigma}_n^{\psi})^{-1/2})\|$ and

$$\begin{aligned}\|\mathbf{K}_n(\widehat{A}_n - A_n(\Sigma_n^{\psi})^{1/2}(\widehat{\Sigma}_n^{\psi})^{-1/2})\| &\leqslant \|\mathbf{K}_n(\widehat{A}_n - A_n)\| + \|\mathbf{K}_nA_n\|\|(\widehat{\Sigma}_n^{\psi})^{1/2} - (\Sigma_n^{\psi})^{1/2}\|\,\|(\widehat{\Sigma}_n^{\psi})^{-1/2}\|\\ &\leqslant O_p\left(\sqrt{\frac{m_n\log k_n}{n\kappa_{k_n}}}\right) + (\lambda_{\min}^{1/2}(\Sigma_n^{\psi}) + \lambda_{\min}^{1/2}(\widehat{\Sigma}_n)^{\psi})^{-1}\|\widehat{\Sigma}_n^{\psi} - \Sigma_n^{\psi}\|\\ &= O_p\left(\sqrt{\frac{m_n\log k_n}{n\kappa_{k_n}}}\right) + O_p(\eta_{n,4}^{1/2} + \eta_{n,5}^{1/2}) \end{aligned}\tag{C.21}$$

where, $\eta_{n,4} = \max\left(k_n^2/(n\tau_{k_n}), k_n^{-2\beta}\right)$ and $\eta_{n,5} = \max\left(k_n^{-2\gamma/d_z}, \frac{k_n^2}{n\kappa_{k_n}}, k_n\kappa_{k_n}^{-1}\|T(E_{k_n}g - g)\|_X^2\right)$, to get the second inequality, we have used Lemma E.3 in Chen and Christensen [2015a] and to get the rate of $\|\widehat{\Sigma}_n^{\psi} - \Sigma_n^{\psi}\|$ we have used the same argument as in the proof of Lemma A.2. By using similar arguments as in the proof of Lemma D.8 in Chen and Christensen [2015a, page 37] we obtain that $\mathcal{B}_1 = O_p(\widetilde{\eta}_n\widetilde{c}_n) = o_{p^*}(\widetilde{c}_n^{-1})$ under Assumption 12 *(iii) – (iv)*.
Next, let us consider term $\mathcal{B}_2$ which is the supremum of a Gaussian process with the same distribution (under $\mathbb{P}^*$) as $\mathbb{Z}_n(z)$ (under the data distribution). Therefore, by applying Lemma A.2 and Chen and Christensen [2015a, LemmaD.7], which is valid under Assumptions 5 *(i)*, 7 *(i)* and 12 *(i) – (ii)*, we obtain that $\mathcal{B}_2 = O_p(\widetilde{\eta}_n)O_{p^*}(\widetilde{c}_n)$. □