3-1-2016

Methods for Nonparametric and Semiparametric Regressions with Endogeneity: a Gentle Guide

Xiaohong Chen

Yin Jia Qiu

METHODS FOR NONPARAMETRIC AND SEMIPARAMETRIC REGRESSIONS WITH ENDOGENEITY: A GENTLE GUIDE

By

Xiaohong Chen and Yin Jia Qiu

March 2016

COWLES FOUNDATION DISCUSSION PAPER NO. 2032

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.yale.edu/

Methods for Nonparametric and Semiparametric Regressions with Endogeneity: a Gentle Guide

Xiaohong Chen*and Yin Jia Jeff Qiu†

First version: October 2015; Revised: March 2016

Abstract

This paper reviews recent advances in estimation and inference for nonparametric and semiparametric models with endogeneity. It first describes methods of sieves and penalization for estimating unknown functions identified via conditional moment restrictions. Examples include nonparametric instrumental variables regression (NPIV), nonparametric quantile IV regression and many more semi-nonparametric structural models. Asymptotic properties of the sieve estimators and the sieve Wald, quasi-likelihood ratio (QLR) hypothesis tests of functionals with nonparametric endogeneity are presented. For sieve NPIV estimation, the rate-adaptive data-driven choices of sieve regularization parameters and the sieve score bootstrap uniform confidence bands are described. Finally, simple sieve variance estimation and over-identification test for semiparametric two-step GMM are reviewed. Monte Carlo examples are included.

Keywords: Conditional moment restrictions containing unknown functions, (Quantile) Instrumental variables, Linear and nonlinear functionals, Sieve minimum distance, Sieve GMM, Sieve Wald, QLR, Bootstrap, Semiparametric two-step GMM, Numerical equivalence.

JEL Codes: C12, C14, C32.

*Corresponding author. Cowles Foundation for Research in Economics, Yale University, New Haven, CT 06520, USA; Email: xiaohong.chen@yale.edu

†Department of Economics, Yale University, New Haven, CT 06520, USA

1 Introduction

Models with endogeneity are arguably the most important feature that differentiates econometrics from statistics. There is a rapidly growing literature on semiparametric and nonparametric models with endogeneity. All existing results could be classified into either the ones via the instrumental variables (IV) approach (e.g., Newey and Powell 2003) or the ones via the control function (CF) approach (e.g., Blundell and Powell 2003). In linear models with endogeneous regressors and additive disturbances both approaches generate consistent and often analytically identical estimators (e.g., Hausman 1987). Both approaches are also closely related in some other parametric models with endogeneity (e.g., Wooldridge 2002). In nonparametric models with endogeneity, the IV and CF approaches have slightly different identification and estimation strategies with different advantages and weaknesses. See, e.g., Blundell, Kristensen and Matzkin (2013) and Horowitz (2013) for recent discussions. General identification results for various nonparametric models with endogeneity are available using either approaches. See, e.g., Newey, Powell and Vella (1999), Blundell and Powell (2003), Chesher (2003), Matzkin (2007, 2013), Florens, Heckman, Meghir and Vytlacil (2008), Imbens and Newey (2009), Chernozhukov and Hansen (2013), Blundell and Matzkin (2014), Berry and Haile (2014), Chen, Chernozhukov, Lee and Newey (2014) and references therein. This short review will thus focus on recent advances in estimation of and inference on semiparametric and nonparametric models with endogeneity.

Earlier nonparametric and semiparametric IV models are typically cast into the framework of conditional moment restrictions containing unknown functions of endogenous variables (e.g., Newey and Powell (2003), Ai and Chen (2003)). More complicated nonparametric and semiparametric IV models could be cast into the framework of several conditional moment restrictions with different conditioning information sets, some of the moment restrictions contain unknown functions of endogenous variables (e.g., Ai and Chen 2007). In comparison, nonparametric and semiparametric models using the CF approach are typically set into the framework of semiparametric two-step or multi-step GMM, where the first-step unknown functions are typically reduced form functions of exogenous variables such as conditional mean (or quantile) regressions or conditional choice probabilities without endogeneity (e.g., Olley and Pakes (1996), Newey, Powell and Vella (1999), Blundell and Powell (2003)). Within both approaches most of the existing estimation and inference results could be divided further into those using kernel or local polynomial smoothing methods (e.g., Fan and Gijbels 1996), and those using sieves or penalization methods (e.g., Grenander 1981). The folk knowledge is that the estimation and inferences for functionals of structural parameters in nonparametric and semiparametric models with endogeneity are much more difficult than those for the corresponding parametric models with endogeneity, regardless whether they are identified via the IV or the CF approach and whether they are estimated via kernel, sieve or penalization method. In particular, a nonparametric IV regression is typically an ill-posed inverse problem and hence the nonparametric convergence rate is slower than that for the corresponding nonparametric regression without endogeneity; see, e.g., Hall and Horowitz (2005), Darolles, Fan, Florens and Renault (2011), Chen and Reiss (2011), Chen and Christensen (2015). While a typical nonparametric CF approach does not suffer the ill-posed inverse problem, it involves multi-stage nonparametric and/or semiparametric estimation with the previously estimated functions as generated regressors in the next stage estimation, which makes it difficult to correctly characterize the asymptotic variance of the final-stage estimator of the functional of interest and hence difficult to conduct asymptotically valid inference; see, e.g., Pakes and Olley (1995), Hahn and Ridder (2013), Mammen et al. (2012, 2015).

In this review, we shall present (penalized) sieve based estimation and inference theories for functionals in nonparametric and semiparametric models with endogeneity, where the model parameters are assumed to be identified via either the IV or the CF approach. We shall mainly describe computationally attractive procedures via finite-dimensional linear sieve approximations to unknown functions. Linear sieves are also called series, which are typically linear combinations of known basis functions such as Bernstein polynomials, Chebychev polynomials, Hermite polynomials, Fourier series, polynomial splines, B-splines, wavelets. Among different linear sieves, splines and wavelets have nice theoretical properties in terms of achieving the nonparametric optimal convergence rates even for models with endogeneity (e.g., Chen and Christensen, 2015); B-splines, Bernstein polynomial and some other sieves have nice shape-preserving properties; see, e.g., DeVore and Lorentz (1993), Chen (2007) and references therein. However, linear sieves are not as flexible as nonlinear sieves (such as neural networks) in approximating a unknown function of a multivariate covariate. In empirical studies when lots of covariates are present, a flexible and computationally attractive approach is to combine linear sieve approximations with various dimension-reduction modeling tools (such as partially linear, single index, additive models, varying coefficients). See Chen and Pouzo (2012) and Chen (2007, 2013) for more general penalized, possibly infinite-dimensional linear or nonlinear sieve methods and various trade-offs. Although slightly less general, estimation and inference for nonparametric models with (or without) endogeneity via finite-dimensional linear sieves can be easily implemented using existing softwares for parametric models with (or without) endogeneity. For example, once after a nonparametric IV regression is approximated by a finite-dimensional linear sieve, its estimation and inference can be easily conducted via Hansen's (1982) GMM or minimum distance as if the sieve approximated model were a correctly-specified parametric model (e.g., Chen and Pouzo 2015). As another example, once after the unknown reduce form functions in a nonparametric CF problem are approximated by finite-dimensional linear sieves, the problem becomes a parametric CF two-step or multi-step, and the unknown asymptotic variance of the final-stage estimator in the original nonparametric CF problem is now consistently estimated by a variance estimator for the final-stage estimator in the sieved parametric CF model (e.g., Newey, Powell and Vella (1999), Ackerberg, Chen and Hahn (2012)). Moreover, for inference on a root-n estimable functional in a nonparametric model with (or without) endogeneity, the sieve dimension could be chosen optimally to balance the bias (sieve approximation error) and the standard deviation in the nonparametric part.[1] Although for inference on a slower than root-n estimable functional, the sieve dimension has to be slightly larger so that the sieve bias goes to zero faster.[2] In practice if an empirical researcher is unsure whether the functional of interest is root-n estimable or not, it is safer to choose sieve dimension slightly larger than the "optimal" one balancing the sieve bias and

[1] see, e.g., Newey (1994) and Chen and Shen (1998) for nonparametric models without endogeneity, Chen and Pouzo (2009) for nonparametric models with endogeneity.

[2] see, e.g., Newey (1997) and Chen, Liao and Sun (2014) for nonparametric models without endogeneity, Chen and Pouzo (2015), and Chen and Christensen (2015) for nonparametric models with endogeneity.

the standard deviation in the nonparametric part.[3]

This short review mainly describes implementation aspects of the (penalized) sieve estimators and tests for nonparametric models with endogeneity, and refers to the original papers for regularity conditions and technical details. The rest of the paper is organized as follows. Section 2 first presents a general class of conditional moment restrictions containing unknown functions of possibly endogenous variables. Examples include nonparametric instrumental variables regression (NPIV), nonparametric quantile IV regression, partially additive IV regression, single-index IV regression, quantile transformation IV model and numerous other semi-nonparametric structural models. Various (penalized) sieve extremum estimators, such as sieve minimum distance (MD), sieve GMM, sieve conditional empirical likelihood (EL), sieve unconditional EL and generalized EL are described. Some commonly used sieves, including shape-preserving sieves and simple ways to impose shape restrictions, are mentioned. The convergence rates of the penalized sieve estimators of the nonparametric part of general conditional moment restrictions are briefly summarized. Section 3 first reviews the asymptotic normality of sieve t statistics for functionals that are either root-n estimable (i.e., regular) or slower than root-n estimable (i.e., irregular). It then presents sieve Wald and quasi-likelihood ratio (QLR) inferences for regular or irregular functionals, and their bootstrap versions. Section 4 presents additional results for sieve NPIV estimation. It describes rate-adaptive data-driven choices of sieve regularization parameters, and score bootstrap uniform confidence bands based on sieve t statistics for irregular (nonlinear) functional processes of a NPIV. Section 5 describes simple sieve variance estimation and over-identification test for semiparametric two-step GMM with a sieve estimated nonparametric first step. It also mentions sieve multi-step estimation for semiparametric models via the CF approach. Section 6 contains Monte Carlo illustrations programmed in R. Section 7 concludes by briefly mentioning additional related results and open questions.

[3] In practice one could just use AIC to choose the linear sieve dimension if one mainly cares about asymptotic validity and is not too concerned with asymptotic optimality. See simulation Section 6 for examples.

2 Conditional Moment Restrictions Containing Unknown Functions

2.1 Models

Economic models often imply a set of semi-nonparametric conditional moment restrictions of the following form:

$$E[\rho(Y, X; \theta_0, h_0)|X] = 0 \quad a.s. - X, \tag{2.1}$$

where $\rho(\cdot; \theta_0, h_0)$ is a $d_\rho \times 1-$vector of generalized residual functions whose functional forms are known up to the true but unknown parameters value (θ_0', h_0), Y is a vector of endogenous variables and X is a vector of conditioning (or instrumental) variables. The conditional distribution of Y given X, $F_{Y|X}$, is not specified beyond that it satisfies (2.1). Let $\alpha \equiv (\theta', h) \in \mathcal{A} \equiv \Theta \times \mathcal{H}$ denote the parameters of interest, with $\theta \in \Theta$ being a $d_\theta \times 1-$vector of finite dimensional parameters and $h \equiv (h_1(\cdot), ..., h_q(\cdot)) \in \mathcal{H}$ being a $1 \times d_q-$vector valued function. The arguments of each unknown function $h_\ell(\cdot)$ may differ across $\ell = 1, ..., q$, may depend on θ, $h_{\ell'}(\cdot)$, $\ell' \neq \ell$, X and Y. The residual function $\rho(\cdot; \alpha)$ could be nonlinear and pointwise non-smooth in parameters $\alpha \equiv (\theta', h)$. This paper calls a model with at least one unknown function $h_l(\cdot)$ depending on the endogenous variable Y as a model with nonparametric endogeneity.

Model (2.1) nests many widely used semi/nonparametric generalized regression models. Examples include, *but are not limited to* nonparametric mean instrumental variables regression (NPIV):

$$E[Y_1 - h_0(Y_2)|X] = 0 \quad a.s. - X, \tag{2.2}$$

(Hall and Horowitz (2005), Carrasco et al. (2007), Blundell et al. (2007), Darolles et al. (2011)); nonparametric quantile IV regression (NPQIV):

$$E[1\{Y_1 \leq h_0(Y_2)\} - \gamma|X] = 0 \quad a.s. - X, \tag{2.3}$$

(Chernozhukov and Hansen (2005, 2013), Chernozhukov et al. (2007), Horowitz and Lee (2007), Chen and Pouzo (2012), Gagliardini and Scaillet (2012), Chen et al. (2014)); partially linear IV: $E[Y_1 - Y_2'\theta_0 - h_0(Y_3)|X] = 0$ (Florens et al. 2012) and partially linear quantile IV: $E[1\{Y_1 \leq Y_2'\theta_0 + h_0(Y_3)\} - \gamma|X] = 0$ (Chen and Pouzo, 2009); partially additive IV: $E[Y_1 - Y_0'\theta_0 - h_{01}(Y_2) - h_{02}(Y_3)|X] = 0$ and nonparametric additive quantile IV: $E[1\{Y_1 \leq h_{01}(Y_2) + h_{02}(Y_3)\} - \gamma|X] = 0$ (Chen and Pouzo, 2012); varying coefficient IV: $E[Y_1 - h_{01}(Y_2)X_1 - h_{02}(Y_3)X_2|X] = 0$ with $X = (X_1, X_2, X_3)$ and its quantile version $E[1\{Y_1 \leq h_{01}(Y_2)X_1 + h_{02}(Y_3)X_2\} - \gamma|X] = 0$; single-index IV: $E[Y_1 - h_0(Y_2'\theta_0)|X] = 0$ (Chen et al. 2014) and its quantile version $E[1\{Y_1 \leq h_0(Y_2'\theta_0)\} - \gamma|X] = 0$; transformation IV model: $E[h_0(Y_1) - Y_2'\theta_0|X] = 0$ and its quantile version $E[1\{h_0(Y_1) \leq Y_2'\theta_0\} - \gamma|X] = 0$ for $h_0(\cdot)$ being monotone. While the above examples are natural extensions of popular existing regression models in econometrics and statistics that allow for nonparametric endogeneity, model (2.1) also includes numerous complex economic structural models with endogeneity. Some real data economic applications include semi/nonparametric spatial models with endogeneity (Pinkse et al. (2002), Merlo and de Paula (2015)); systems of shape-invariant Engle curves with endogeneity (Blundell et al. 2007) and its quantile version (Chen and Pouzo, 2009); semi/nonparametric asset pricing models (e.g., Gallant and Tauchen (1989), Chen and Ludvigson (2009), Hansen (2014)); semi/nonparametric static and dynamic game models (e.g., Bajari et al., 2011); nonparametric optimal endogenous contract models (e.g., Bontemps and Martimort (2013)). Additional examples of the general model (2.1) can be found in Chamberlain (1992a), Newey and Powell (2003), Ai and Chen (2003), Chen and Pouzo (2012), Chen et al. (2014), Berry and Haile (2014) and the references therein.

Ai and Chen (2012) considers an extension of model (2.1) to a general semiparametric conditional moment restrictions with nested information set:

$$E[\rho_t(Y, X; \theta_0, h_0(\cdot))|X^{(t)}] = 0 \quad a.s. - X^{(t)} \quad \text{for } t = 1, ..., T < \infty, \tag{2.4}$$

$$\{1\} \subseteq \sigma\left(X^{(1)}\right) \subset \sigma\left(X^{(2)}\right) \subset \cdots \subset \sigma\left(X^{(T)}\right) \quad \text{with } X^{(T)} = X, \tag{2.5}$$

where $\sigma\left(X^{(t)}\right)$ denotes the sigma-field generated by $X^{(t)}$. When $X^{(1)}$ is the constant 1 (i.e., a degenerate random variable), the conditional expectation $E[\rho_1(\cdot)|X^{(1)}]$ is simply the unconditional expectation $E[\rho_1(\cdot)]$. Model (2.4-2.5) is a direct extension of Chamberlain's (1992b) sequential moment restrictions model $E[\rho_t(Y,X;\theta_0)|X^{(t)}]=0$ by inclusion of unknown functions $h_0(\cdot)$. It obviously includes model (2.1) and semi-nonparametric panel data models where the information set expands over time. With $T=2$, $\theta=(\theta_1',\theta_2')'$, $X^{(1)}=1$ and $X=X^{(2)}$, model (2.4-2.5) nests the following widely used semiparametric two-step GMM problem:

$$E[\rho_1(Y,X;\theta_{01},\theta_{02},h_0(\cdot))]=0 \text{ with } \dim(\rho_1)\geq\dim(\theta_1), \tag{2.6}$$

$$E[\rho_2(Y,X;\theta_{02},h_0(\cdot))|X]=0, \tag{2.7}$$

where the unknown parameter θ_{02} and the unknown function $h_0(\cdot)$ can be identified and estimated using the conditional moment restriction (2.7) in the first step, and can then be plugged into the unconditional moment restriction (2.6) to identify and estimate the unknown parameter θ_{01} in the second step. An example of the model (2.6-2.7) is the estimation of a weighted average derivative of a NPIV regression: $\theta_{01}=E[a(Y_2)\nabla h_0(Y_2)]$, where $a()$ is a known positive weight function and $\nabla h_0()$ is the first derivative of h_0 in the NPIV (2.2). See Section 5 for a review on the semiparametric two-step GMM problems. Many semiparametric program evaluation models, semiparametric missing data models, choice-based sampling problems, some nonclassical measurement error models and semiparametric control function models could also fit into framework (2.4-2.5).

There are further applications where different equations may require different sets of instruments. Ai and Chen (2007) studies a generalization of (2.4-2.5) (and hence (2.1)) to the semiparametric conditional moment restriction with a different information set:

$$E[\rho_t(Y,X,\theta_0,h_0())|X^{(t)}]=0 \quad a.s.-X^{(t)} \quad \text{for } t=1,2,...,T<\infty, \tag{2.8}$$

where $X^{(t)}$ is either equal to a subset of X or a degenerate random variable; but the sigma-field $\sigma\left(X^{(t)}\right)$ no longer needs to be nested as t increases. Examples of model (2.8) include, *but are* *not restricted to*, the triangular simultaneous equations system studied in Newey, Powell and Vella (1999); a semiparametric hedonic price system where some explanatory variables in some equations are correlated with the errors in other equations; the simultaneous equations with measurement error in some exogenous variables; a semi-nonparametric panel data model where some variables that are uncorrelated with the error in a given time period are correlated with the errors in previous periods; semi-nonparametric dynamic panel sample selection model; and semiparametric game models with incomplete information.

Of course one could consider further generalizations of model (2.8), say, to models with increasing T, or to models with parameter index sets in the conditioning set. However, it suffices to say that even model (2.1) already covers many economics applications. We shall review estimation and inference results for model (2.1) in the rest of the paper, and refer readers to Ai and Chen (2007, 2012) for results for models (2.4-2.5) and (2.8).

2.2 Penalized Sieve Extremum Estimation

Let $\{Z_i \equiv (Y_i', X_i')'\}_{i=1}^n$ be a random sample from the probability distribution P_0 of $Z \equiv (Y', X')'$ that satisfies the conditional moment restrictions (2.1). Let the infinite-dimensional parameter space $\mathcal{A} \equiv \Theta \times \mathcal{H}$ be endowed with a metric $||.||_s = ||.||_e + ||.||_H$, where $||.||_e$ is the Euclidean norm on Θ (a compact subset in $\mathbb{R}^{d_\theta}$), and $||.||_H$ denotes a norm on the infinite-dimensional function space $\mathcal{H}$ (typical choices of $||.||_H$ include $||.||_\infty$ and $||.||_{L^2}$). We call

$$\mathcal{A}_I(P_0) \equiv \left\{\alpha \equiv (\theta', h) \in (\mathcal{A}, ||.||_s) : E[\rho(Y, X; \alpha)|X] = 0 \quad a.s. - X\right\}$$

the set of parameters that are identified by the model (2.1) (or simply the identified set). When $\mathcal{A}_I(P_0) = \{\alpha_0 \equiv (\theta_0', h_0)\}$ is a singleton in $(\mathcal{A}, ||.||_s)$, the parameter α_0 is (point) identified by the model (2.1). One can consider estimation of the identified set $\mathcal{A}_I(P_0)$ by recasting it as a set of minimizers to a non-random criterion function $Q() : (\mathcal{A}, ||.||_s) \to \mathbb{R}$ such that $Q(\alpha) = 0$ whenever $\alpha \in \mathcal{A}_I(P_0)$, and $Q(\alpha) > 0$ for all $\alpha \in \mathcal{A} \backslash \mathcal{A}_I(P_0)$. There are many choices of $Q()$ that captures exactly the same identified set $\mathcal{A}_I(P_0)$ of the model (2.1) (see subsections 2.2.1 and 2.2.2

for examples). Let $\widehat{Q}_n$ be a random criterion function that converges to Q in probability uniformly over totally bounded subsets of $(\mathcal{A}, ||.||_s)$. Then one may want to estimate $\mathcal{A}_I(P_0)$ by an extremum estimator: $\arg\inf_{\alpha\in\mathcal{A}} \widehat{Q}_n(\alpha)$. Since the parameter space $\mathcal{A}$ is infinite dimensional and possibly non-compact in $||.||_s$, $\arg\inf_{\alpha\in\mathcal{A}} \widehat{Q}_n(\alpha)$ may be difficult to compute and not well-defined; or even if it exists, it may be inconsistent for $\mathcal{A}_I(P_0)$ under $||.||_s$ when $\mathcal{A}$ is not compact in $||.||_s$. See Chen (2007, p 5560-61) for discussions of well-posed vs ill-posed optimization problems over infinite-dimensional parameter spaces.[4]

Method of sieves and method of penalization are two general approaches to solve possibly ill-posed, infinite-dimensional optimization problems. The *method of sieves* replaces $\inf_{\alpha\in\mathcal{A}} \widehat{Q}_n(\alpha)$ by $\inf_{\alpha\in\mathcal{A}_{k(n)}} \widehat{Q}_n(\alpha)$, where the sieves $\mathcal{A}_{k(n)}$ is a sequence of approximating parameter spaces that are less complex but dense in $(\mathcal{A}, ||.||_s)$ (see Grenander (1981)). Popular sieves are typically compact, non-decreasing ($\mathcal{A}_k \subseteq \mathcal{A}_{k+1} \subseteq \cdots$) and are such that $\mathcal{A} \subseteq cl\left(\cup_k \mathcal{A}_k\right)$ (i.e., for any $\alpha \in \mathcal{A}$ there exists an element $\pi_{k(n)}\alpha$ in $\mathcal{A}_{k(n)}$ satisfying $||\alpha - \pi_{k(n)}\alpha||_s \to 0$ as $n \to \infty$). The *method of penalization* (or *regularization*) replaces $\inf_{\alpha\in\mathcal{A}} \widehat{Q}_n(\alpha)$ by $\inf_{\alpha\in\mathcal{A}} \left\{\widehat{Q}_n(\alpha) + \lambda_n Pen(\alpha)\right\}$, where $\lambda_n > 0$ is a penalization parameter such that $\lambda_n \to 0$ as $n \to \infty$ and the penalty $Pen() > 0$ is typically chosen such that $\{\alpha \in \mathcal{A} : Pen(\alpha) \leq M\}$ is compact in $||.||_s$ for all $M \in (0, \infty)$.

Chen and Pouzo (2012) and Chen (2013) introduced a class of *penalized sieve extremum* (PSE) estimators, $\widehat{\alpha}_n = (\widehat{\theta}_n, \widehat{h}_n) \in \mathcal{A}_{k(n)} = \Theta \times \mathcal{H}_{k(n)}$, defined by:

$$\left\{\widehat{Q}_n(\widehat{\alpha}_n) + \lambda_n \widehat{P}_n(\widehat{h}_n)\right\} \leq \inf_{\alpha\in\Theta\times\mathcal{H}_{k(n)}} \left\{\widehat{Q}_n(\alpha) + \lambda_n \widehat{P}_n(h)\right\}, \tag{2.9}$$

where $\mathcal{H}_{k(n)}$ is a sieve parameter space whose complexity, denoted as $k(n) \equiv \dim(\mathcal{H}_{k(n)})$, grows with sample size n and becomes dense in the original function space $\mathcal{H}$ under the metric $||.||_H$; $\lambda_n \geq 0$ is a penalization parameter such that $\lambda_n \to 0$ as $n \to \infty$; and the penalty $\widehat{P}_n() \geq 0$, which is an empirical analog of a non-random penalty function $Pen : \mathcal{H} \to [0, +\infty)$, is jointly measurable in h and the data $\{Z_t\}_{t=1}^n$.

[4] Also see Carrasco et al (2007) and Horowitz (2013) for reviews on linear ill-posed inverse problems that include the NPIV model as a leading case.

The definition of PSE (2.9) includes both the method of sieves and the method of penalization as special cases. In particular, when $\lambda_n \widehat{P}_n() = 0$, PSE (2.9) becomes the *sieve extremum estimator*, i.e., the solution to $\inf_{\alpha \in \Theta \times \mathcal{H}_{k(n)}} \widehat{Q}_n(\alpha)$. When $\lambda_n \widehat{P}_n() > 0$, $\widehat{P}_n() = Pen()$ and $\mathcal{H}_{k(n)} = \mathcal{H}$ (i.e., $k(n) = \infty$), PSE (2.9) becomes the *function space penalized extremum estimator*, i.e., the solution to $\inf_{\alpha \in \Theta \times \mathcal{H}} \left\{ \widehat{Q}_n(\alpha) + \lambda_n Pen(h) \right\}$.

The sieve space $\mathcal{H}_{k(n)}$ in the definition of PSE (2.9) could be finite dimensional ($k(n) < \infty$), infinite dimensional ($k(n) = \infty$), compact or non-compact (in $||.||_H$). Commonly used finite-dimensional linear sieves (also called *series*) take the form:

$$\mathcal{H}_{k(n)} = \left\{ h \in \mathcal{H} : h(\cdot) = \sum_{k=1}^{k(n)} \pi_k q_k(\cdot) \right\}, \quad k(n) < \infty, \ k(n) \to \infty \text{ slowly as } n \to \infty, \tag{2.10}$$

where $\{q_k\}_{k=1}^{\infty}$ is a sequence of known basis functions of a Banach space $(\mathbf{H}, ||.||_H)$ such as polynomial splines, B-splines, wavelets, Fourier series, Hermite polynomial series, Power series, Chebychev series, etc. Linear sieves with constraints, which are commonly used, can be expressed as:

$$\mathcal{H}_{k(n)} = \left\{ h \in \mathcal{H} : h(\cdot) = \sum_{k=1}^{k(n)} \pi_k q_k(\cdot), \ R_n(h) \le B_n \right\}, \quad k(n) \le \infty, \ B_n \to \infty \text{ slowly as } n \to \infty, \tag{2.11}$$

where the constraint $R_n(h) \le B_n$ reflects prior information about $h_0 \in \mathcal{H}$ such as smoothness properties. The sieve space $\mathcal{H}_{k(n)}$ in (2.11) is finite dimensional and compact (in $||.||_H$) if and only if $k(n) < \infty$ and $\mathcal{H}_{k(n)}$ is closed and bounded; it is infinite dimensional and compact (in $||.||_H$) if and only if $k(n) = \infty$ and $\mathcal{H}_{k(n)}$ is closed and totally bounded. For example, $\mathcal{H}_{k(n)} = \left\{ h \in \mathcal{H} : h(\cdot) = \sum_{k=1}^{k(n)} \pi_k q_k(\cdot), \ \|h\|_H \le \log(n) \right\}$ is compact if $k(n) < \infty$, but is not compact (in $||.||_H$) if $k(n) = \infty$. Linear sieves (or series) are widely used in empirical economics due to the computationally simplicity. See DeVore and Lorentz (1993), Chen (2007) and references therein for examples of nonlinear sieves and shape-preserving sieves.

The penalty function $Pen()$ is typically convex and/or *lower semicompact* (i.e., the set $\{h \in \mathcal{H} : Pen(h) \le M\}$ is compact in $(\mathbf{H}, ||.||_H)$ for all $M \in (0, \infty)$) and reflects prior information about

$h_0 \in \mathcal{H}$. For instance, when $\mathcal{H} \subseteq L^p(d\mu)$, $1 \leq p < \infty$, a commonly used penalty function is for a known measure $d\mu$, or $\widehat{P}_n(h) = ||h||^p_{L^p(d\widehat{\mu})}$ for an empirical measure $d\widehat{\mu}$ when $d\mu$ is unknown. When $\mathcal{H}$ is a mixed weighted Sobolev space $\{h : ||h||^2_{L^2(d\mu)} + ||\nabla^r h||^p_{L^p(leb)} < \infty\}$, $1 \leq p < \infty$, $r \geq 1$, we can let $||.||_H$ be the $L^2(d\mu)-$norm, and $\widehat{P}_n(h) = ||h||^2_{L^2(d\widehat{\mu})} + ||\nabla^k h||^p_{L^p(leb)}$ or $\widehat{P}_n(h) = ||\nabla^k h||^p_{L^p(leb)}$ for some $k \in [1, r]$, where $\nabla^k h$ denotes the $k-$th derivative of $h()$. When the sieve dimension $k(n)$ grows very fast in the sense of $k(n) \geq n$, the penalty $\widehat{P}_n(h) = ||h||_{L^1(d\mu)}$ is a LASSO type penalty on the sieve coefficients.

Model (2.1) is a natural extension of the unconditional moment restrictions $E[g(Y, X; \theta_0)] = 0$ studied in Hansen's (1982) seminal work on the generalized method of moment (GMM). Therefore, all different criterion functions and estimation procedures designed for estimating θ_0 of the original model $E[g(Y, X; \theta_0)] = 0$, such as GMM, minimum distance (MD), empirical likelihood (EL), generalized empirical likelihood (GEL) and others,[5] could be extended to estimate $\alpha_0 \equiv (\theta_0', h_0)$ of model (2.1) via our PSE (2.9) with different choices of criterion functions $\widehat{Q}_n$. See the next two subsections for examples.

2.2.1 Criteria based on nonparametrically estimated conditional moments

Let $m(X, \alpha) \equiv E[\rho(Y, X; \alpha)|X]$ be the $d_\rho \times 1$- conditional mean function of the residual function $\rho(Y, X; \alpha)$, and $\Sigma(X)$ be any $d_\rho \times d_\rho$ -positive definite weighting matrix. Then the conditional moment restrictions model (2.1) is equivalent to

$$\left\| [\Sigma(\cdot)]^{-1/2} m(\cdot, \alpha) \right\|_{L^p(X)} = 0 \quad \text{when } \alpha = \alpha_0$$

for some $p \in [1, \infty]$. Newey and Powell (2003) and Ai and Chen (2003) independently proposed the quadratic minimum distance (MD) criterion $Q(\alpha) = E\left[m(X, \alpha)'\{\Sigma(X)\}^{-1} m(X, \alpha)\right]$ (i.e.,

[5] See, e.g., Imbens (2002), Kitamura (2007), Hansen (2014), Parente and Smith (2014) for recent reviews on various estimation and testing methods for $E[g(Y, X; \theta_0)] = 0$.

$L^2(X)-$norm), and the sieve MD estimation:

$$\min_{\alpha\in\Theta\times\mathcal{H}_{k(n)}} \widehat{Q}_n(\alpha), \quad \widehat{Q}_n(\alpha) = \frac{1}{n}\sum_{t=1}^{n} \widehat{m}(X_t,\alpha)'\{\widehat{\Sigma}(X_t)\}^{-1}\widehat{m}(X_t,\alpha), \tag{2.12}$$

where $\widehat{m}(x,\alpha)$ and $\widehat{\Sigma}(x)$ are any consistent estimators of $m(x,\alpha)$ and $\Sigma(x)$, respectively. When $\widehat{\Sigma}(x) = \widehat{\Sigma}_0(x)$ is a consistent estimator of the optimal weighting $\Sigma_0(x) = Var(\rho(Y,X;\alpha_0)|X=x)$, $\widehat{Q}_n^0(\alpha) = \frac{1}{n}\sum_{t=1}^{n}\widehat{m}(X_t,\alpha)'\{\widehat{\Sigma}_0(X_t)\}^{-1}\widehat{m}(X_t,\alpha)$ is called the optimally weighted MD criterion, which leads to semiparametric efficient estimation of θ_0 for the model (2.1) (see Ai and Chen (2003), Chen and Pouzo (2009)). For a general sieve MD criterion, Chen and Pouzo (2015) considered any consistent nonparametric estimator $\widehat{m}(x,\alpha)$ that is linear in $\rho(Z,\alpha)$:

$$\widehat{m}(x,\alpha) \equiv \sum_{i=1}^{n}\rho(Z_i,\alpha)A_n(X_i,x) \tag{2.13}$$

where $A_n(X_i,x)$ is a known measurable function of $\{X_j\}_{j=1}^n$ for all x, whose expression varies according to different nonparametric procedures such as series, kernel, local linear regression, and nearest neighbors. Series and kernel LS estimators are the most widely used in economics:

- If $A_n(X_i,x) = A_{LS}(X_i,x) = p^{J_n}(X_i)'(P'P)^{-}p^{J_n}(x)$ then $\widehat{m}(x,\alpha)$ is the series least squares (LS) estimator (2.14):

$$\widehat{m}_{LS}(x,\alpha) = \left(\sum_{i=1}^{n}\rho(Z_i,\alpha)p^{J_n}(X_i)'\right)(P'P)^{-}p^{J_n}(x), \tag{2.14}$$

  where $\{p_j\}_{j=1}^{\infty}$ is a sequence of known basis functions that can approximate any square integrable functions of X well, $p^{J_n}(X) = (p_1(X),...,p_{J_n}(X))'$, $P' = (p^{J_n}(X_1),...,p^{J_n}(X_n))$, and $(P'P)^{-}$ is the generalized inverse of the $J_n\times J_n-$matrix $P'P$. See Newey and Powell (2003), Ai and Chen (2003), Chen and Pouzo (2009).

- If $A_n(X_i,x) = A_K(X_i,x) = 1\{x\in\mathcal{X}_n\}K\left(\frac{X_i-x}{a_n}\right)/\sum_{j=1}^{n}K\left(\frac{X_j-x}{a_n}\right)$ then $\widehat{m}(x,\alpha)$ is the kernel conditional mean estimator (2.15):

$$\widehat{m}_K(x,\alpha)=1\{x\in\mathcal{X}_n\}\frac{\sum_{i=1}^n\rho(Z_i,\alpha)K\left(\frac{X_i-x}{a_n}\right)}{\sum_{j=1}^n K\left(\frac{X_j-x}{a_n}\right)}, \tag{2.15}$$

where $K:\mathbb{R}^{d_x}\to\mathbb{R}$ is a known symmetric function, a_n a bandwidth satisfying $a_n\to 0$ as $n\to\infty$, and the indicator function $1\{x\in\mathcal{X}_n\}$ is to trim the boundaries of the support of $\{X_j\}_{j=1}^n$. See Ai and Chen (1999) for details.

For better finite sample performance of the SMD (2.12), it is better to use delete-t observation version in computing $\widehat{m}(X_t,\alpha)$ in (2.13) (likewise in (2.14) and (2.15)):

$$\widehat{m}(X_t,\alpha)=\sum_{i=1,i\neq t}^n\rho(Z_i,\alpha)A_n(X_i,X_t). \tag{2.16}$$

Another criterion for model (2.1) is the following sieve conditional EL:

$$\min_{\alpha\in\Theta\times\mathcal{H}_{k(n)}}\widehat{Q}_n(\alpha),\quad \widehat{Q}_n(\alpha)=\frac{1}{n}\sum_{t=1}^n\sup_{\lambda\in\widehat{\Lambda}_n(X_t,\alpha)}\sum_{j=1}^n A_K(X_j,X_t)\log\left(1+\lambda'\rho(Z_j,\alpha)\right)$$

where $\widehat{\Lambda}_n(X_t,\alpha)=\left\{\lambda\in\mathbb{R}^{d_\rho}:1+\lambda'\rho(Z_j,\alpha)>0,\ j=1,...,n\right\}$. See Zhang and Gijbels (2003) and Otsu (2011) for details. Of course one could also study sieve conditional GEL and other related criteria for model (2.1).

Some examples, such as Robinson's (1988) partly linear regression and Ichimura's (1993) single index regression, of the model (2.1) satisfy $m(X,\alpha)-m(X,\alpha_0)=\rho(Y,X,\alpha)-\rho(Y,X,\alpha_0)$ for any α. Then, instead of applying the SMD estimation or the sieve conditional EL, one could simply perform the following sieve generalized least squares (GLS) regression:

$$\min_{\alpha\in\Theta\times\mathcal{H}_{k(n)}}\frac{1}{n}\sum_{i=1}^n\rho(Z_i,\alpha)'[\widehat{\Sigma}(X_i)]^{-1}\rho(Z_i,\alpha). \tag{2.17}$$

See Ai and Chen (2007) for combination of SMD and sieve GLS for estimation of the more general model (2.8) with different information sets.

2.2.2 Criteria based on unconditional moments of increasing dimension

Note that $E[\rho(Z,\alpha_0)|X]=0$ if and only if the following increasing number of unconditional moment restrictions hold:

$$E[\rho(Z,\alpha_0)p_j(X)]=0,\ j=1,2,...,J_n, \tag{2.18}$$

where $\{p_j(X), j=1,2,...,J_n\}$ is a sequence of known basis functions that can approximate any real-valued square integrable function of X well as $J_n\to\infty$. It is now obvious that the semi-nonparametric conditional moment restrictions (2.1) can be estimated using any criteria $Q()$ (and $\widehat{Q}_n()$) for the set of unconditional moment restrictions of increasing dimension (2.18).

A typical quadratic MD criterion for model (2.18) is the following sieve GMM:

$$\min_{\alpha\in\Theta\times\mathcal{H}_{k(n)}}\widehat{Q}_n(\alpha),\quad \widehat{Q}_n(\alpha)=\widehat{g}_n(\alpha)'\widehat{W}\widehat{g}_n(\alpha) \tag{2.19}$$

with $\widehat{g}_n(\alpha)=\frac{1}{n}\sum_{t=1}^n\rho(Z_t,\alpha)\otimes p^{J_n}(X_t)$, and $\widehat{W}$ is a possibly random $d_\rho J_n\times d_\rho J_n-$weighting matrix of increasing dimension (that is introduced for potential efficiency gains). Sieve GMM (2.19) was suggested in Ai and Chen (2003) and Chen (2007), and studied in Sueishi (2014), Tao (2015) and Chernozhukov, Newey and Santos (2015) subsequently.

Another criterion for model (2.18) is the following sieve (unconditional) EL:

$$\min_{\alpha\in\Theta\times\mathcal{H}_{k(n)}}\widehat{Q}_n(\alpha),\quad \widehat{Q}_n(\alpha)=\sup_{\lambda\in\widehat{\Lambda}_n(\alpha)}\frac{1}{n}\sum_{t=1}^n\log\left(1+\lambda'[\rho(Z_t,\alpha)\otimes p^{J_n}(X_t)]\right)$$

where $\widehat{\Lambda}_n(\alpha)=\left\{\lambda\in\mathbb{R}^{d_\rho J_n}:1+\lambda'[\rho(Z_t,\alpha)\otimes p^{J_n}(X_t)]>0,\ t=1,...,n\right\}$; see, e.g., Chang, Chen and Chen (2014). Of course one could also study sieve (unconditional) GEL:

$$\min_{\alpha\in\Theta\times\mathcal{H}_{k(n)}}\widehat{Q}_n(\alpha),\quad \widehat{Q}_n(\alpha)=\sup_{\lambda\in\widehat{\Lambda}_n(\alpha)}\frac{1}{n}\sum_{t=1}^n\psi\left(\lambda'[\rho(Z_t,\alpha)\otimes p^{J_n}(X_t)]\right)$$

where $\psi() : \Psi \to [0, \infty]$ is a concave function and

$\widehat{\Lambda}_n(\alpha) = \left\{\lambda \in \mathbb{R}^{d_\rho J_n} : \lambda'[\rho(Z_t, \alpha) \otimes p^{J_n}(X_t)] \in \Psi,\ t = 1, ..., n\right\}$. See Smith (1997), Donald, Imbens and Newey (2003) and Parente and Smith (2014) for various choices of $\psi()$ and other related criteria for unconditional moment restrictions (2.18) with increasing dimension.

2.2.3 Computation and Heuristic choices of regularization parameters

Although many different criteria could be used in PSE (2.9) to estimate $\alpha_0 = (\theta_0', h_0)$ for the model (2.1), some criterion functions are much easier to compute than others in the presence of unknown functions (h) with endogeneity. Without unkown h, theoretical statistics and econometrics papers recommend EL and GEL over MD and GMM for better asymptotic second-order properties in efficient estimation of θ, although MD and GMM are easier to compute with commonly used sample size in empirical work in economics. For model (2.1) with nonparametric endogeneity, (penalized) sieve MD and sieve GMM are much easier to compute. In particular, for nonparametric quantile IV, quantile partially additive IV, quantile varying coefficient IV, quantile single-index IV, quantile transformation IV regression examples of model (2.1), the optimal weighting in SMD criterion is known, and hence we could always use the optimally weighted MD criterion $\widehat{Q}_n^0(\alpha)$ with $\widehat{\Sigma}_0(x) = \Sigma_0(x) = \gamma(1 - \gamma) \times I_{d_\rho}$, which leads to computationally simple yet semiparametrically efficient estimator for θ_0 (see Chen and Pouzo, 2009).

There is no formal theoretical result on data-driven choices of smoothing parameters for the various PSE estimators for the general model (2.1) yet. Based on the sieve GMM interpretation (2.19) of the original sieve MD estimator with series LS estimator (2.14) of $m(X, \alpha)$, Ai and Chen (2003) suggested $d_\rho \times J_n \geq d_\theta + k(n)$, $J_n/n \to 0$ and $k(n) \to \infty$ slowly. Blundell, Chen and Kristensen (2007) and Chen and Pouzo (2009, 2012, 2015) present detailed Monte Carlo studies and Engel curve real data applications in terms of choices of smoothing parameters that are consistent with their theoretical conditions for the optimal rates of convergence of the penalized SMD estimators. They recommend using the penalized SMD estimators with finite dimensional linear sieves (typically splines) with small penalty: $\lambda_n \to 0$ fast (i.e., very close to zero), $k(n) \to \infty$ slowly, $d_\rho \times J_n = c \times k(n)$

for c slightly bigger than 1 and $J_n/n \to 0$; see these papers for details. There are a few very recent papers on data-driven choices of smoothing parameters for various estimators of the NPIV model (2.2) $E[Y_1 - h_0(Y_2)|X] = 0$; see Section 4 for details.

Many members of the general model (2.1) have a scalar-valued regression or quantile regression residual function $\rho(Z, \alpha)$. A computationally attractive and stable procedure is the (penalized) SMD estimation (2.12) using the identity weighting $\widehat{\Sigma}(X) = 1$, the series LS estimator (2.14) as $\widehat{m}(X, \alpha)$ for the conditional mean function $m(X, \alpha) = E[\rho(Z, \alpha)|X]$, and a linear sieve $\pi' q^{k(n)}$ for $h \in \mathcal{H}$. The procedure could be expressed as

$$\min_{\theta \in \Theta, \pi} \left\{ [\mathbf{R}(\theta, \pi)]' P(P'P)^{-} P' [\mathbf{R}(\theta, \pi)] + \lambda_n \widehat{P}_n(\pi' q^{k(n)}) \right\} \tag{2.20}$$

where $\mathbf{R}(\theta, \pi) = (\rho(Z_1, \theta, \pi' q^{k(n)}), ..., \rho(Z_n, \theta, \pi' q^{k(n)}))'$ and $P' = (p^{J_n}(X_1), ..., p^{J_n}(X_n))$. Obviously $J_n \geq d_\theta + k(n)$, $J_n/n \to 0$, $k(n) \to \infty$ slowly and $\lambda_n \to 0$ fast (i.e., very close to zero or could be set to zero). As already mentioned, this simple criterion (2.20) (even with $\lambda_n = 0$) automatically leads to semiparametric efficient estimation of θ_0 for various quantile IV examples of (2.1).

Example 1: Partially additive IV regression The model is

$$Y_1 = Y_2' \theta_0 + h_{01}(Y_3) + h_{02}(Y_0) + u, \quad E[u|X] = 0, \tag{2.21}$$

where $Y_1 \in \mathbb{R}, Y_2 \in \mathbb{R}^{d_\theta}$, $Y_0, Y_3 \in [0,1]$ are endogenous, and $X = (X_1, X_2, X_3) \in \mathcal{X} \subset \mathbb{R}^{d_\theta + 2}$ are conditioning variables. The parameters of interest are $\alpha = (\theta', h_1, h_2) \in \Theta \times \mathcal{H}_1 \times \mathcal{H}_2 \equiv \mathcal{A}$. For simplicity we assume that $Var(Y_2) > 0$, $\mathcal{H}_1 = \{h_1 \in C^2([0,1]) : \int [\nabla^2 h_1(y_3)]^2 dy_3 < \infty\}$ and $\mathcal{H}_2 = \{h_2 \in C^2([0,1]) : \int [\nabla^2 h_2(y_0)]^2 dy_0 < \infty, h_2(0.5) = c\}$ for a known finite constant c. Under mild regularity conditions the true parameters $\alpha_0 \in \mathcal{A}$ are identified, and can be consistently estimated via a (penalized) SMD procedure. We can take $\mathcal{A}_n = \Theta \times \mathcal{H}_{1n} \times \mathcal{H}_{2n}$ as a sieve space with $\mathcal{H}_{1n} = \{h_1(y_3) = q^{k_{1n}}(y_3)' \pi_1 : \int [\nabla^2 h_1(y_3)]^2 dy_3 \leq c_1 \log n\}$ and $\mathcal{H}_{2n} = \{h_2(y_0) = q^{k_{2n}}(y_0)' \pi_2 : \int [\nabla^2 h_2(y_0)]^2 dy_0 \leq c_2 \log n,\ h_2(0.5) = c\}$, where $q^{k_{1n}}()$, $q^{k_{2n}}()$ are either a polynomial spline basis with equally spaced (according to empirical quantile the support) knots or a 3rd order cardinal B-spline basis.

Example (2.21) reduces to the Monte Carlo example in Chen (2007, p. 5580) when $Y_2 = X_1, Y_0 = X_2$ become exogenous. It also becomes the Monte Carlo experiment 1 in Section 6 when $Y_0 = X_2$ becomes exogenous. Nevertheless, all these examples could be estimated using the same (penalized) sieve MD procedure (2.20) with the residual function $\rho(Z,\alpha) = Y_1 - (Y_2'\theta + h_1(Y_3) + h_2(Y_0))$, which becomes a penalized 2SLS (without constraints):

$$\min_{\theta,\pi} (\mathbf{Y}_1 - \mathbf{Y}_2\theta - \mathbf{Q}\pi)' P(P'P)^- P' (\mathbf{Y}_1 - \mathbf{Y}_2\theta - \mathbf{Q}\pi) + \sum_{\ell=1}^{2} \lambda_\ell \pi_\ell' C_\ell \pi_\ell \tag{2.22}$$

where $\mathbf{R}(\theta,\pi) = \mathbf{Y}_1 - \mathbf{Y}_2\theta - \mathbf{Q}\pi$, with $\mathbf{Y}_1 = (Y_{1,1}, ..., Y_{1,n})'$, $\mathbf{Y}_2 = (Y_{2,1}, ..., Y_{2,n})'$, $\pi = (\pi_1', \pi_2')'$, $\mathbf{Q}_1 = (q^{k_{1,n}}(Y_{3,1}), ..., q^{k_{1,n}}(Y_{3,n}))'$, $\mathbf{Q}_2 = (q^{k_{2,n}}(Y_{0,1}), ..., q^{k_{2,n}}(Y_{0,n}))'$ and $\mathbf{Q} = (\mathbf{Q}_1', \mathbf{Q}_2')'$. And the penalty $\lambda_n \widehat{P}_n(\pi' q^{k(n)}) = \sum_{\ell=1}^{2} \lambda_\ell \pi_\ell' C_\ell \pi_\ell$, with $C_\ell = \int [\nabla^2 q^{k_{\ell,n}}(y)][\nabla^2 q^{k_{\ell,n}}(y)]' dy$, $\pi_\ell' C_\ell \pi_\ell = \int [\nabla^2 h_\ell(y)]^2 dy$ for $\ell = 1, 2$. with small penalty terms $\lambda_1, \lambda_2 \geq 0$. The problem (2.22) has a simple closed form solution as presented in Chen (2007, p. 5580-83).

Example 2: Single-index IV regression The model is:

$$Y_1 = h_0(Y_3 + Y_2'\theta_0) + u, \quad E[u|X] = 0. \tag{2.23}$$

with $\dim(X) \geq 2$ and $\alpha_0 = (\theta_0', h_0) \in \Theta \times \mathcal{H}$. See Chen et al (2014) for sufficient conditions for identification of α_0. We can estimate α_0 using the sieve MD procedure (2.20) with a residual function $\rho(Z,\alpha) = Y_1 - h(Y_3 + Y_2'\theta)$, and a sieve space $\mathcal{A}_n \equiv \Theta \times \mathcal{H}_n$, $\mathcal{H}_n = \left\{h \in \mathcal{H} : h(.) = \pi' q^{k(n)}\right\}$, that is, $\widehat{\alpha} = (\widehat{\theta}', \widehat{h}) \in \mathcal{A}_n$ solves

$$\min_{\theta,\pi} [\mathbf{Y}_1 - \mathbf{Q}(\theta)\pi]' P(P'P)^- P' [\mathbf{Y}_1 - \mathbf{Q}(\theta)\pi]$$

where $\mathbf{R}(\theta,\pi) = \mathbf{Y}_1 - \mathbf{Q}(\theta)\pi$, with $\mathbf{Q}(\theta) = (q^{k(n)}(Y_{3,1} + Y_{2,1}'\theta), ..., q^{k(n)}(Y_{3,n} + Y_{2,n}'\theta))'$. For a computationally simpler estimation procedure, we can follow the profile SMD procedure suggested by Blundell et al (2007). First, for each fixed $\theta \in \Theta$ we estimate $\widetilde{h}(y_3, y_2; \theta) = q^{k(n)}(y_3 + y_2'\theta)'\widetilde{\pi}(\theta)$ via 2SLS

$$\begin{aligned}\widetilde{\pi}(\theta) &= \arg\min_{\pi}[\mathbf{Y}_1 - \mathbf{Q}(\theta)\pi]'P(P'P)^{-}P'[\mathbf{Y}_1 - \mathbf{Q}(\theta)\pi] \\ &= \left(\mathbf{Q}(\theta)'P(P'P)^{-}P'\mathbf{Q}(\theta)\right)^{-}\mathbf{Q}(\theta)'P(P'P)^{-}P'\mathbf{Y}_1.\end{aligned}$$

Second, obtain $\widehat{\theta}_n$ as the solution to

$$\widehat{\theta}_n = \arg\min_{\theta\in\Theta}[\mathbf{Y}_1 - \mathbf{Q}(\theta)\widetilde{\pi}(\theta)]'P(P'P)^{-}P'[\mathbf{Y}_1 - \mathbf{Q}(\theta)\widetilde{\pi}(\theta)].$$

Lastly, estimate $h_0(y_3 + y_2'\theta_0)$ by $\widehat{h}_n(y_3 + y_2'\widehat{\theta}_n) = q^{k(n)}(y_3 + y_2'\widehat{\theta}_n)'\widetilde{\pi}(\widehat{\theta}_n)$.

2.2.4 Consistency and Convergence rates of nonparametric part with endogeneity

Suppose that $\alpha_0 \in (\mathcal{A}, ||.||_s)$ is (point) identified by the model (2.1) (see Newey and Powell (2003), Chen et al. (2014) and references therein for sufficient conditions for identification). Newey and Powell (2003) derived the consistency of SMD estimators assuming compact parameter space and smooth residuals $\rho(Z,\alpha)$ (in α_0). Chen and Pouzo (2012) present a general consistency theorem for approximate PSE estimators, allowing for ill-posed inverse problems, non-compact parameter spaces, flexible penalty functions and non-smooth residuals $\rho()$ (in α_0). In particular, they allow for fast growing sieve space (for h) with L^1 penalty on function h or its derivatives, which is similar to LASSO.

For NPIV model (2.2) $E[Y_1 - h_0(Y_2)|X] = 0$, Hall and Horowitz (2005) and Chen and Reiss (2011) establish the minimax lower bound in $||.||_{L^2(Y_2)}-$loss for estimation of $h_0()$, Chen and Christensen (2015) derived the minimax lower bound in $||.||_{L^2(Y_2)}-$loss for estimation of derivatives of $h_0()$, and in $||.||_\infty-$loss for estimation of $h_0()$ and its derivatives. The sieve NPIV estimator of Blundell, Chen and Kristensen (2007) and the modified series NPIV estimator of Horowitz (2011) are shown to achieve the optimal convergence rate in $||.||_{L^2(Y_2)}-$norm. Recently Chen and Christensen (2015) obtained the optimal sup-norm rate of the sieve NPIV estimator for estimating h_0 and its derivatives. Interestingly, the optimal sup-norm rate coincides with the optimal L^2-norm rate for severely ill-posed case, and is up to a factor of $(\log(n))^{\varepsilon}$ with $0 < \varepsilon < 1/2$ for mildly ill-posed case. The sup-norm rate result is very useful for inference on nonlinear welfare functionals of $h_0()$.

For general model (2.1) that include NPIV and nonparametric quantile IV as special cases, Chen and Pouzo (2012) first establish the Hilbert-norm rate of convergence for the PSMD estimators under high level regularity conditions that allow for any nonparametric consistent estimators $\widehat{m}(X, \alpha)$ of the conditional mean functions $m(X, \alpha) = E[\rho(Z, \alpha)|X]$. They then provide low level sufficient conditions in terms of the series LS estimator (2.14). In particular, they show that the PSMD estimators for general model (2.1) can achieve the same optimal rate in $||.||_{L^2(Y_2)}$ as that for the NPIV. Unfortunately, besides the NPIV model, there is no sup-norm rate results for estimating h_0 of a general model (2.1) with nonparametric endogeneity.

2.2.5 Shape Restrictions and Shape-preserving Sieves

Economic theory often provides shape restrictions, such as additivity, non-negativity, monotonicity, convexity, concavity, homogeneity of the unknown function $h(.)$ (e.g., Matzkin, 1994). Imposing shape restrictions often help with nonparametric identification. See, for example, Chen and Pouzo (2012, theorem A.1) for using strictly convex penalty to regain identification for a class of partially identified nonparametric IV models, and Freyberger and Horowitz (2013) for imposing monotonicity to obtain tighter identified set in a partially identified NPIV model with discrete endogenous regressor. Functions that are known to satisfy shape restrictions can be well approximated by various kinds of shape-preserving sieves, including shape-preserving B-splines, Bernstein polynomials, and certain wavelet sieves. See, e.g., DeVore (1977a, 1977b), Anastassiou and Yu (1992a, 1992b), Dechevsky and Penev (1997), Chui (1992 Chapter 4, 6), Chen (2007), and Wang and Ghosh (2012). Nonparametric estimation and testing with shape restrictions have been studied in both statistics and econometrics literature. See Groeneboom and Jongbloed (2014), and Han and Wellner (2016) for detailed treatments and up-to-date references about nonparametric estimation and inference under shape constraints for models without endogeneity. For a smooth nonparametric function in models with or without endogeneity, it is known that imposing shape restriction improves finite sample performance but does not affect the optimal nonparametric convergence rate, while testing against some shape restricted alternative could be more powerful than that against alternatives without shape restrictions. See, e.g., Meyer (2008) for inference using shape-restricted regression splines without endogeneity, Blundell, Horowitz and Parey (2012, 2015) for improving finite sample behavior of their demand curve estimation by imposing Slutsky inequality on demand function, Grasmair, Scherzer and Vanhems (2013) for the asymptotic properties of a NPIV model (2.2) with a general set of constraints, Chetverikov and Wilhelm (2015) for estimation and testing in a nonparametric regression with endogeneity under a monotone IV assumption and a monotonicity restriction of $h(.)$.

3 Sieve Inferences on Functionals of Nonparametric Endogeneity

In many applications of the semi-nonparametric conditional moment restrictions model (2.1), we are interested in inference on a vector-valued linear and/or nonlinear functional $\phi : \mathcal{A} \to \mathbb{R}^{d_\phi}$. For example, consider a model $E[\rho(Y_1, \theta_0, h_0(Y_2))|X] = 0$, linear functionals of $\alpha = (\theta', h) \in \mathcal{A}$ could be an Euclidean functional $\phi(\alpha) = \theta$, a point evaluation functional $\phi(\alpha) = h(\overline{y}_2)$ (for $\overline{y}_2 \in \text{supp}(Y_2)$), a weighted derivative functional $\phi(h) = \int w(y_2)\nabla h(y_2)dy_2$ and others; nonlinear functionals include a quadratic functional $\int w(y_2)\left|h(y_2)\right|^2 dy_2$, a quadratic derivative functional $\int w(y_2)\left|\nabla h(y_2)\right|^2 dy_2$, exact consumer surplus and deadweight loss functionals of an endogenous demand function h (see Vanhems (2010), Blundell, Horowitz and Parey (2012), Chen and Christensen (2015)).

Let $\widehat{\alpha}_n = (\widehat{\theta}'_n, \widehat{h}_n)$ be a consistent estimator of $\alpha_0 = (\theta'_0, h_0)$ that is identified by the semi-nonparametric model (2.1). Then $\phi(\widehat{\alpha}_n)$ is a simple plug-in estimator of the functional of interest $\phi(\alpha_0)$. And $\phi(\widehat{\alpha}_n) - \phi(\alpha_0)$ typically converges to zero at either a root-n rate or a slower than root-n rate. In the literature, $\phi : \mathcal{A} \to \mathbb{R}^{d_\phi}$ is sometimes called a regular (or smooth or bounded) functional if $\phi(\alpha_0)$ can be estimated at a $\sqrt{n}-$rate. And $\phi()$ is called a irregular (or non-smooth or unbounded) functional if $\phi(\alpha_0)$ can be best estimated at a slower than $\sqrt{n}$−rate.

3.1 The simpler case when $\phi(\alpha) = \theta$ is regular

For the conditional moment restrictions (2.1) with i.i.d. data, Chamberlain (1992) and Ai and Chen (1999, 2003) derive the semiparametric efficiency bound for θ_0 in (2.1). Let $\widehat{\alpha}_n = (\widehat{\theta}', \widehat{h})$ be the sieve MD estimator (2.12). Under a set of regularity conditions, Ai and Chen (1999, 2003) establish $\sqrt{n}(\widehat{\theta} - \theta_0) \overset{d}{\to} \mathcal{N}(0, V_\theta)$ for a finite positive definite matrix V_θ, and provide simple consistent variance estimators for V_θ. They also show that the optimally weighted sieve MD estimator achieves the semiparametric efficiency bound of θ_0 in (2.1). Their results are subsequently extended by Otsu (2011) and Sueishi (2014) to sieve EL and sieve GMM estimation of (2.1) respectively. All these papers assume that the entire parameter space $\mathcal{A} \equiv \Theta \times \mathcal{H}$ is compact under a strong metric $||.||_s = ||.||_e + ||.||_H$ and that the residual function $\rho(Z, \alpha)$ is pointwise differentiable in $\alpha_0 = (\theta_0', h_0)$. Chen and Pouzo (2009) relax these assumptions. They show that, for the general model (2.1) with nonparametric endogeneity, the penalized SMD estimator $\widehat{\alpha}_n = (\widehat{\theta}', \widehat{h})$ can simultaneously achieve root-n asymptotic normality of $\widehat{\theta}$ and the optimal nonparametric convergence rate of $\widehat{h}$ (in a strong norm $|| \cdot ||_H$), allowing for possibly nonsmooth residuals and/or a noncompact (in $|| \cdot ||_H$) function space ($\mathcal{H}$) or noncompact sieve spaces ($\mathcal{H}_{k(n)}$). This result is very useful to applied researchers since the same regularization parameters chosen to achieve the optimal rate for estimating h_0 are valid for $\sqrt{n}(\widehat{\theta} - \theta_0) \overset{d}{\to} \mathcal{N}(0, V_\theta)$. In addition, Chen and Pouzo (2009) show that a simple weighted bootstrap procedure can consistently estimate the limiting distribution of the (penalized) SMD $\widehat{\theta}$, which is very useful when the residual function $\rho(Z; \theta, h(\cdot))$ is non-smooth in $\alpha_0 = (\theta_0', h_0)$, such as in a partially linear quantile IV regression example $E[1\{Y_3 \leq Y_1'\theta_0 + h_0(Y_2)\}|X] = \gamma \in (0, 1)$ (see proposition 5.1 in Chen and Pouzo 2009).

All these papers could only conduct inference on $\phi(\alpha_0) = \theta_0$ of the model (2.1) when θ_0 is assumed to be regular (i.e., root-n estimable), however.

3.2 Possibly irregular functional $\phi(\alpha)$ of model (2.1)

For the semi-nonparametric conditional moment restrictions (2.1) with nonparametric endogeneity, it is in general difficult to check whether a functional $\phi(\alpha)$ is regular or irregular. Let $\widehat{\phi}_n \equiv \phi(\widehat{\alpha}_n)$ be the *plug-in (penalized) SMD estimator* of $\phi(\alpha_0)$. Recently, Chen and Pouzo (2015) established the asymptotic normality of $\widehat{\phi}_n$ of $\phi(\alpha_0)$ that could be slower than root-n estimable. They also establish asymptotic distributions of sieve Wald, sieve quasi-likelihood ratio (QLR), and sieve score statistics for the hypothesis of $\phi(\alpha_0) = \phi_0$, regardless of whether $\phi(\alpha_0)$ is root-n estimable or not. Some of their inference results are summarized in this subsection.

3.2.1 Sieve t (or Wald) statistic

Under some regularity conditions and regardless of whether $\phi(\alpha_0)$ is $\sqrt{n}$ estimable, Chen and Pouzo (2015) show that

$$\sqrt{n} V_{\phi,n}^{-1/2} \left(\phi(\widehat{\alpha}_n) - \phi(\alpha_0)\right) = -\sqrt{n}\mathbb{Z}_n + o_p(1) \xrightarrow{d} \mathcal{N}(0, I_{d_\phi})$$

where $\mathbb{Z}_n \equiv \frac{1}{n}\sum_{i=1}^n V_{\phi,n}^{-1/2} S_{n,i}^*$ and

$$S_{n,i}^* \equiv \left(\frac{dm(X_i,\alpha_0)}{d\alpha}[v_n^*]\right)' \Sigma(X_i)^{-1}\rho(Z_i,\alpha_0) \tag{3.1}$$

is the *sieve score*, $V_{\phi,n} = Var\left(S_{n,i}^*\right)$ is the *sieve variance.*

For notational simplicity we focus on a real-valued functional $\phi : \mathcal{A} \to \mathbb{R}$ and sieve t statistic in this subsection. See Appendix A in Chen and Pouzo (2015) for vector-valued and increasing dimensional functionals and the corresponding sieve Wald statistic. Then intuitively, the functional $\phi(\alpha_0)$ is $\sqrt{n}$-estimable if $\limsup_n V_{\phi,n} < \infty$, and is slower-than-$\sqrt{n}$-estimable if $\limsup_n V_{\phi,n} = \infty$.

The sieve variance $V_{\phi,n}$ has a *closed form* expression resembling the "delta-method" variance for a parametric MD problem:

$$V_{\phi,n} = Var\left(S_{n,i}^*\right) = F_n' D_n^- \mho_n D_n^- F_n, \tag{3.2}$$

where $\overline{q}^{k(n)}(\cdot) \equiv \left(\mathbf{1}'_{d_\theta}, q^{k(n)}(\cdot)'\right)'$ is a $(d_\theta + k(n)) \times 1$ vector with $\mathbf{1}_{d_\theta}$ a $d_\theta \times 1$ vector of 1's,

$$F_n \equiv \frac{d\phi(\alpha_0)}{d\alpha}[\overline{q}^{k(n)}(\cdot)] \equiv \frac{\partial \phi(\theta_0 + \theta, h_0 + \beta' q^{k(n)}(\cdot))}{\partial \gamma'}|_{\gamma=0}$$

and $\gamma \equiv (\theta', \beta')'$ are $(d_\theta + k(n)) \times 1$ vectors, $\frac{d\phi(\alpha_0)}{dh}[q^{k(n)}(\cdot)'] \equiv \frac{\partial \phi(\theta_0, h_0 + \beta' q^{k(n)}(\cdot))}{\partial \beta}|_{\beta=0}$, and

$$D_n = E\left[\left(\frac{dm(X,\alpha_0)}{d\alpha}[\overline{q}^{k(n)}(\cdot)']\right)' \Sigma(X)^{-1} \left(\frac{dm(X,\alpha_0)}{d\alpha}[\overline{q}^{k(n)}(\cdot)']\right)\right],$$

$$\mho_n = E\left[\left(\frac{dm(X,\alpha_0)}{d\alpha}[\overline{q}^{k(n)}(\cdot)']\right)' \Sigma(X)^{-1}\Sigma_0(X)\Sigma(X)^{-1} \left(\frac{dm(X,\alpha_0)}{d\alpha}[\overline{q}^{k(n)}(\cdot)']\right)\right],$$

where $\frac{dm(X,\alpha_0)}{d\alpha}[\overline{q}^{k(n)}(\cdot)'] \equiv \frac{\partial E[\rho(Z,\theta_0+\theta,h_0+\beta' q^{k(n)}(\cdot))|X]}{\partial \gamma}|_{\gamma=0}$ is a $d_\rho \times (d_\theta + k(n))$ matrix.

The closed form expression of $V_{\phi,n}$ immediately leads to simple consistent plug-in sieve variance estimators; one of which is

$$\widehat{V}_{\phi,n} = \widehat{F}_n' \widehat{D}_n^- \widehat{\mho}_n \widehat{D}_n^- \widehat{F}_n, \tag{3.3}$$

where $\widehat{F}_n \equiv \frac{d\phi(\widehat{\alpha}_n)}{d\alpha}[\overline{q}^{k(n)}(\cdot)] \equiv \frac{\partial \phi(\widehat{\theta}+\theta, (\widehat{\beta}+\beta)' q^{k(n)}(\cdot))}{\partial \gamma'}|_{\gamma=0}$, $\widehat{\Sigma}_i = \widehat{\Sigma}(X_i)$ and $\widehat{\rho}_i = \rho(Z_i, \widehat{\alpha}_n)$,

$$\widehat{D}_n = \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{d\widehat{m}(X_i,\widehat{\alpha})}{d\alpha}[\overline{q}^{k(n)}(\cdot)']\right)' \widehat{\Sigma}_i^{-1} \left(\frac{d\widehat{m}(X_i,\widehat{\alpha})}{d\alpha}[\overline{q}^{k(n)}(\cdot)']\right)\right],$$

$$\widehat{\mho}_n = \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{d\widehat{m}(X_i,\widehat{\alpha})}{d\alpha}[\overline{q}^{k(n)}(\cdot)']\right)' \widehat{\Sigma}_i^{-1}\widehat{\rho}_i\widehat{\rho}_i'\widehat{\Sigma}_i^{-1} \left(\frac{d\widehat{m}(X_i,\widehat{\alpha})}{d\alpha}[\overline{q}^{k(n)}(\cdot)']\right)\right].$$

Theorem 4.2 in Chen and Pouzo (2015) then presents the asymptotic normality of the sieve (Student's) t statistic:

$$\sqrt{n}\widehat{V}_{\phi,n}^{-1/2}\left(\phi(\widehat{\alpha}_n) - \phi(\alpha_0)\right) \overset{d}{\to} \mathcal{N}(0,1). \tag{3.4}$$

Example 1: Partially Additive IV Regression (2.21) (Continued) For this example, $\alpha = (\theta, h_1, h_2) \in \Theta \times \mathcal{H}_1 \times \mathcal{H}_2$, $\rho(Z,\alpha) = Y_1 - (Y_2'\theta + h_1(Y_3) + h_2(Y_0))$, $u = \rho(Z,\alpha_0)$, $m(X,\alpha) = E[Y_1 - (Y_2'\theta + h_1(Y_3) + h_2(Y_0))\,|X]$ and $\widehat{\Sigma}(X) = \Sigma(X) = 1$. To apply sieve t statistic (3.4) for inference on a functional $\phi(\alpha_0)$, we just need to compute a plug-in estimator $\widehat{V}_{\phi,n}$ for the sieve variance $V_{\phi,n} = F_n' D_n^- \mho_n D_n^- F_n$. Note that

$$F_n = \left(\frac{\partial\phi(\alpha_0)}{\partial\theta'}, \frac{d\phi(\alpha_0)}{dh_1}[q^{k_{1n}}(.)'], \frac{d\phi(\alpha_0)}{dh_2}[q^{k_{2n}}(.)'] \right)',$$

$\gamma = (\theta', \beta_1', \beta_2')'$, $\frac{d\phi(\alpha_0)}{dh_1}[q^{k_{1n}}(.)'] = \frac{\partial\phi(\theta_0, h_{01}+\beta_1' q^{k_{1n}}(.), h_{02})}{\partial\beta_1}\Big|_{\beta_1=0}$, and

$$D_n = E\left[\left(E[\left(Y_2', q^{k_{1n}}(Y_3)', q^{k_{2n}}(Y_0)'\right)|X] \right)' \left(E[\left(Y_2', q^{k_{1n}}(Y_3)', q^{k_{2n}}(Y_0)'\right)|X] \right) \right],$$

$$\mho_n = E\left[\left(E[\left(Y_2', q^{k_{1n}}(Y_3)', q^{k_{2n}}(Y_0)'\right)|X] \right)' u^2 \left(E[\left(Y_2', q^{k_{1n}}(Y_3)', q^{k_{2n}}(Y_0)'\right)|X] \right) \right].$$

Then a consistent sieve variance estimator is $\widehat{V}_{\phi,n} = \widehat{F}_n' \widehat{D}_n^- \widehat{\mho}_n \widehat{D}_n^- \widehat{F}_n$ with

$$\widehat{D}_n = \widehat{S}'\widehat{G}^{-1}\widehat{S}, \quad \widehat{\mho}_n = \widehat{S}'\widehat{G}^{-1}\widehat{\Omega}\widehat{G}^{-1}\widehat{S} \tag{3.5}$$

where $\widehat{S} = P'(\mathbf{Y}_2, \mathbf{Q}_1, \mathbf{Q}_2)/n$, $\widehat{G} = P'P/n$ and $\widehat{\Omega} = n^{-1}\sum_{i=1}^n \widehat{u}_i^2 p^{J_n}(X_i) p^{J_n}(X_i)'$ with $\widehat{u}_i = Y_{1,i} - Y_{2,i}'\widehat{\theta}_n - \widehat{h}_{1,n}(Y_{3,i}) - \widehat{h}_{2,n}(Y_{0,i})$. We note that this is a standard 2SLS variance estimator with $\left(Y_2', q^{k_{1n}}(Y_3)', q^{k_{2n}}(Y_0)'\right)$ endogenous variables and $p^{J_n}(X)'$ instruments.

3.2.2 Sieve QLR statistic

When the generalized residual function $\rho(Y, X, \alpha)$ is not pointwise smooth at α_0, instead of sieve t (or sieve Wald) statistic, we could use sieve quasi likelihood ratio for constructing confidence set of $\phi(\alpha_0)$ and for hypothesis testing of $H_0 : \phi(\alpha_0) = \phi_0 \in \mathbb{R}^{d_\phi}$ against $H_1 : \phi(\alpha_0) \neq \phi_0$. Denote

$$\widehat{QLR}_n(\phi_0) \equiv n \left(\inf_{\alpha \in \mathcal{A}_{k(n)} : \phi(\alpha)=\phi_0} \widehat{Q}_n(\alpha) - \widehat{Q}_n(\widehat{\alpha}_n) \right) \tag{3.6}$$

as the *sieve quasi likelihood ratio* (SQLR) statistic. It becomes an *optimally weighted SQLR* statistic, $\widehat{QLR}^0_n(\phi_0)$, when $\widehat{Q}_n(\alpha)$ is the optimally weighted MD criterion $\widehat{Q}^0_n(\alpha)$. Regardless of whether $\phi(\alpha_0)$ is $\sqrt{n}$ estimable or not, Chen and Pouzo (2015) show that $\widehat{QLR}^0_n(\phi_0)$ is asymptotically chi-square distributed $\chi^2_{d_\phi}$ under the null H_0, and diverges to infinity under the fixed alternatives H_1, and is asymptotically noncentral chi-square distributed under local alternatives. One could compute $100(1-\tau)\%$ confidence set for $\phi(\alpha_0)$ as

$$\left\{ r \in \mathbb{R}^{d_\phi} : \ \widehat{QLR}_n^0(r) \leq c_{\chi^2_{d_\phi}}(1-\tau) \right\},$$

where $c_{\chi^2_{d_\phi}}(1-\tau)$ is the $(1-\tau)$-th quantile of the $\chi^2_{d_\phi}$ distribution. For nonparametric quantile IV, quantile partially additive IV, quantile varying coefficient IV, quantile single-index IV, quantile transformation IV regression examples of model (2.1), the residual function $\rho(Y, X, \alpha)$ is not pointwise smooth at α_0, but we could always use the optimally weighted SQLR statistic $\widehat{QLR}_n^0(\phi_0)$ to construct confidence set for $\phi(\alpha_0)$; see Chen and Pouzo (2009, 2015).

Bootstrap sieve QLR statistic. Chen and Pouzo (2015) propose a bootstrap version of the SQLR statistic. Let $\widehat{QLR}_n^B$ denote a bootstrap SQLR statistic:

$$\widehat{QLR}_n^B(\widehat{\phi}_n) \equiv n \left(\inf_{\alpha \in \mathcal{A}_{k(n)} : \phi(\alpha) = \widehat{\phi}_n} \widehat{Q}_n^B(\alpha) - \inf_{\alpha \in \mathcal{A}_{k(n)}} \widehat{Q}_n^B(\alpha) \right), \tag{3.7}$$

where $\widehat{\phi}_n \equiv \phi(\widehat{\alpha}_n)$, and $\widehat{Q}_n^B(\alpha)$ is a bootstrap version of $\widehat{Q}_n(\alpha)$:

$$\widehat{Q}_n^B(\alpha) \equiv \frac{1}{n} \sum_{i=1}^n \widehat{m}^B(X_i, \alpha)' \widehat{\Sigma}(X_i)^{-1} \widehat{m}^B(X_i, \alpha), \tag{3.8}$$

where $\widehat{m}^B(x, \alpha)$ is a bootstrap version of $\widehat{m}(x, \alpha)$, which is computed in the same way as that of $\widehat{m}(x, \alpha)$ except that we use $\omega_{i,n}\rho(Z_i, \alpha)$ instead of $\rho(Z_i, \alpha)$. Here $\{\omega_{i,n} \geq 0\}_{i=1}^n$ is a sequence of bootstrap weights that has mean 1 and is independent of the original data $\{Z_i\}_{i=1}^n$. Typical weights include an i.i.d. weight $\{\omega_i \geq 0\}_{i=1}^n$ with $E[\omega_i] = 1$, $E[|\omega_i - 1|^2] = 1$ and $E[|\omega_i - 1|^{2+\epsilon}] < \infty$ for some $\epsilon > 0$, or a multinomial weight (i.e., $(\omega_{1,n}, ..., \omega_{n,n}) \sim Multinomial(n; n^{-1}, ..., n^{-1})$). For example, if $\widehat{m}(x, \alpha)$ is a series LS estimator (2.14) of $m(x, \alpha)$, then $\widehat{m}^B(x, \alpha)$ is a bootstrap series LS estimator of $m(x, \alpha)$, defined as:

$$\widehat{m}^B(x, \alpha) \equiv \left(\sum_{i=1}^n \omega_{i,n} \rho(Z_i, \alpha) p^{J_n}(X_i)' \right) (P'P)^- p^{J_n}(x). \tag{3.9}$$

They establish that under the null H_0, the fixed alternatives H_1 or the local alternatives, the conditional distribution of $\widehat{QLR}_n^B(\widehat{\phi}_n)$ (given the data) always converges to the asymptotic null distribution of $\widehat{QLR}_n(\phi_0)$. Let $\widehat{c}_n(a)$ be the $a-th$ quantile of the distribution of $\widehat{QLR}_n^B(\widehat{\phi}_n)$ (conditional on the data $\{Z_i\}_{i=1}^n$). Then for any $\tau \in (0,1)$, we have $\lim_{n\to\infty} \Pr\{\widehat{QLR}_n(\phi_0) > \widehat{c}_n(1-\tau)\} = \tau$ under the null H_0, $\lim_{n\to\infty} \Pr\{\widehat{QLR}_n(\phi_0) > \widehat{c}_n(1-\tau)\} = 1$ under the fixed alternatives H_1, and $\lim_{n\to\infty} \Pr\{\widehat{QLR}_n(\phi_0) > \widehat{c}_n(1-\tau)\} > \tau$ under the local alternatives. We could thus construct a $100(1-\tau)\%$ confidence set using the bootstrap critical values:

$$\left\{ r \in \mathbb{R}^{d_\phi} : \ \widehat{QLR}_n(r) \leq \widehat{c}_n(1-\tau) \right\}. \tag{3.10}$$

The bootstrap consistency holds for possibly non-optimally weighted SQLR statistic and possibly irregular functionals, without the need to compute standard errors.

For model (2.1) with smooth residuals, Chen and Pouzo (2015) established the validity of bootstrap sieve Wald and bootstrap sieve score statistics in their appendices.

3.2.3 Closely related inference results

Since sieve MD, sieve GMM, sieve EL and sieve GEL criteria are all asymptotically first-order equivalent, it is easy to see that all the inference results of Chen and Pouzo (2015) for sieve MD based criterion carry through to those based on sieve GMM, sieve EL and sieve GEL. Indeed, under conditions similar to Ai and Chen (2003) and Chen and Pouzo (2009, 2015), Tao (2015) establishes the same results for sieve Wald, sieve GMM and sieve score test statistics using sieve GMM criterion for model (2.1) when the residual function $\rho(Z,\alpha)$ is smooth in α_0. Pouzo (2015) establishes the validity of bootstrap QLR statistic based on unconditional GEL for functionals of increasing dimension. Also see Ai and Chen (2007, 2012) for sieve MD estimation and inference results for more general models (2.4-2.5) and (2.8).

4 Sieve NPIV: rate-adaptivity and uniform inference

4.1 Sup-norm rate-adaptive sieve NPIV estimation

There is no formal theoretical result on data-driven choices of regularization parameters for any PSE for the general models (2.1) yet. For the NPIV model (2.2) $E[Y_1 - h_0(Y_2)|X] = 0$, there is some very recent work on choice of regularization parameters.

The sieve NPIV (or series 2SLS) estimator of $h_0(.)$ in (2.2) can be written as

$$\widehat{h}_K(y_2) \equiv q^K(y_2)\widehat{\pi}_K = q^K(y_2)' \left(\mathbf{Q}_K' P_J (P_J' P_J)^- P_J' \mathbf{Q}_K\right)^- \mathbf{Q}_K' P_J (P_J' P_J)^- P_J' \mathbf{Y}_1 \tag{4.1}$$

where the subscript K of $\widehat{h}$ indicates the sieve dimension approximating the unknown h_0, and

$$\begin{aligned}
\mathbf{Q}_K &= \left(q^K(Y_{2,1}), \ldots, q^K(Y_{2,n})\right)' && \text{with } q^K(y_2) = (q_1(y_2), \ldots, q_K(y_2))', \\
P_J &= \left(p^J(X_1), \ldots, p^J(X_n)\right)' && \text{with } p^J(x) = (p_1(x), \ldots, p_J(x))', \\
\mathbf{Y}_1 &= (Y_{1,1}, \ldots, Y_{1,n})', && J \geq K.
\end{aligned}$$

Here K is the key regularization parameter, whereas J is a smoothing parameter.[6]

When $K = J$, $q^K(.) = p^J(.)$ are orthonormal bases in $L^2([0,1])$ such as the Legendre Polynomial, the estimator (4.1) becomes identical to Horowitz's (2011) modified series NPIV estimator. Horowitz (2014) proposed an adaptive procedure to choose the regularization parameter $K = J$ for his estimator by choosing K to minimize the sample analog of an approximate asymptotic integrated mean-square error. He also showed that his adaptive procedure leads to near L^2-norm rate adaptivity for estimation of h_0 by a factor of $\sqrt{\log(n)}$. Breunig and Johannes (2015) applied Lepski's (1990) method to choose K and derived a near L^2-norm rate adaptivity for estimation of linear functionals of h_0.

Recently Chen and Christensen (2015, CC) obtained the optimal sup-norm rate of the sieve

[6] We use the notation in Chen and Pouzo (2009, 2012 and 2015) in this review. Chen and Christensen (2015) used J as the regularization parameter and K as the smoothing paramter.

NPIV estimator (4.1) for estimating h_0 and its derivatives. Both the sup-norm and the L^2-norm convergence rates of the sieve NPIV estimator depend on the sieve measure of ill-posendess introduced by Blundell et al (2007), which is defined as:

$$\tau_K = \sup_{h \in \mathcal{H}_K : h \neq 0} \frac{||h||_{L^2(Y_2)}}{||Th||_{L^2(X)}} \tag{4.2}$$

Here, $T : L^2(Y_2) \to L^2(X)$ is a conditional expectation operator given by $Th(x) = E[h(Y_2)|X = x]$, and $\mathcal{H}_K$ is the sieve space for h. Precisely, while the bias part of the sieve NPIV estimator $\widehat{h}_K$ depends only on the smoothness of h_0 (which in general decreases as K increases), the variance part of $\widehat{h}_K$, as measured in sup-norm or L^2-norm, increases with τ_K (which increases with K). Based on this bias-variance trade-off in K, CC adopted Lepski's balance principle to propose a data-driven choice of K that could achieve the optimal sup-norm convergence rate for the NPIV model.

To illustrate, fixing the smoothing parameter J as a known function called $J(.) : \mathbb{N} \to \mathbb{N}$, of the regularization parameter K such that $J(K) \geq K$. CC show that τ_K can be estimated by

$$\widehat{\tau}_K = \frac{1}{s_{min}\left((P_J' P_J/n)^{-1/2}(P_J' \mathbf{Q}_K/n)(\mathbf{Q}_K' \mathbf{Q}_K/n)^{-1/2}\right)} \tag{4.3}$$

where $s_{min}(A)$ is the minimum singular value of the matrix A, and $A^{-1/2}$ denotes the inverse of the square root of a positive definite matrix A.

Let $K_{min} = \lfloor \log(\log(n)) \rfloor$ and

$$\widehat{K}_{\max} = \min\left\{K > K_{\min} : \widehat{\tau}_K \, |\zeta(K)|^2 \sqrt{\log(\log(K))(\log(n))/n} \geq 1\right\}$$

where $|\zeta(K)|^2 = K$ if $\mathbf{Q}_K$ and P_J are spanned by a spline, wavelet, or cosine basis, and $|\zeta(K)|^2 = K^2$ if $\mathbf{Q}_K$ and P_J are spanned by orthogonal polynomial basis. Define $\widehat{I}_K = \left\{k \in \mathcal{K} : K_{min} \leq k \leq \widehat{K}_{max}\right\}$, where $\mathcal{K}$ denotes the sequence of regularizing sieve dimensions. The data-driven index set is defined as

$$\widehat{\mathcal{K}} = \left\{k \in \widehat{I}_K : \left|\left|\widehat{h}_k - \widehat{h}_l\right|\right|_\infty \leq \sqrt{2}\overline{\sigma}\left(\widehat{V}_{\sup}(k) + \widehat{V}_{\sup}(l)\right) \quad \text{for all } l \in \widehat{I}_K \text{ with } l \geq k\right\},$$

with

$$\widehat{V}_{\sup}(k) = \widehat{\tau}_k \xi_k \sqrt{(\log(n))/(n\widehat{e}_k)}$$

where $\xi_k = \sup_{y_2} ||q^k(y_2)||_{\ell^1}$, $\widehat{e}_k = \lambda_{min}(\mathbf{Q}_k'\mathbf{Q}_k/n)$, and a finite constant $\overline{\sigma}^2 \geq \sup_x E[(Y_1 - h_0(Y_2))^2|X = x]$.

The data-driven choice of optimal K is

$$\widehat{K} = \arg\min_{k \in \widehat{\mathcal{K}}} k.$$

CC show that such a choice is sup-norm rate adaptive for sieve NPIV estimation of h_0 and its derivatives.

4.2 Bootstrap uniform confidence band for nonlinear functional processes

Previously Horowitz and Lee (2012) provided a uniform confidence band for $\{h_0(y_2) : y_2 \in [0,1]\}$ based on Horowitz's (2011) modified series NPIV estimator. Recently CC present a bootstrap uniform confidence band for a possibly nonlinear functional process of h_0, which could be used for inference on consumer surplus functional process of an endogenous demand function.

Let $Z^n = \{(Y_{1i}, Y_{2i}, X_i)\}_{i=1}^n$ denote the original data, and $\{w_i\}_{i=1}^n$ be a bootstrap sample of iid random variables drawn independently of the data satisfying $E[w_i|Z^n] = 0, E[w_i^2|Z^n] = 1$ and $E[|w_i|^{2+\varepsilon}|Z^n] < \infty$ for some $\varepsilon \geq 1$. For example, w_i can be $\mathcal{N}(0,1)$ or Mammen's (1993) two-point distribution. CC proposed a sieve score bootstrap procedure to obtain the uniform confidence bands for general nonlinear functional processes $\{\phi_t(h_0) : t \in \mathcal{T}\}$ of h_0 in a NPIV model, where $\mathcal{T}$ is an index set and h_0 is estimated using the sieve NPIV estimator (4.1) with $K \in (\widehat{K}+1, \widehat{K}_{\max}]$, a possibly measurable function of data Z^n that ensures that sup-norm bias is of a smaller order of the sup-norm standard derivation.

Define the sieve score bootstrap process $\{\mathbb{Z}_n^*(t) : t \in \mathcal{T}\}$

$$\mathbb{Z}_n^*(t) = \frac{\frac{\partial \phi_t(\widehat{h})}{\partial h}[q^K]' \widehat{D}^- \widehat{S}' \widehat{G}^{-1}}{\sqrt{\widehat{V}_{\phi,t}}} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n p^J(X_i) \widehat{u}_i w_i \right) \tag{4.4}$$

where $\widehat{D} = \widehat{S}'\widehat{G}^{-1}\widehat{S}$ with $\widehat{S} = P'\mathbf{Q}/n$, $\widehat{G} = P'P/n$, and $\widehat{V}_{\phi,t} = \frac{\partial \phi_t(\widehat{h})}{\partial h}[q^K]'\widehat{D}^- \widehat{\mho} \widehat{D}^- \frac{\partial \phi_t(\widehat{h})}{\partial h}[q^K]$ is a consistent sieve variance estimator, with $\widehat{\mho} = \widehat{S}'\widehat{G}^{-1}\widehat{\Omega}\widehat{G}^{-1}\widehat{S}$ and $\widehat{\Omega} = n^{-1}\sum_{i=1}^n \widehat{u}_i^2 p^J(X_i)p^J(X_i)'$ and $\widehat{u}_i = Y_{1,i} - \widehat{h}(Y_{2,i})$. Note that for fixed t, this sieve variance estimator $\widehat{V}_{\phi,t}$ is the same as that in (3.5).

Let $\mathbb{P}^* \{\}$ denote the probability measure of the bootstrap innovations $\{w_i\}_{i=1}^n$ conditional on the original data Z^n. Under some regularity conditions allowing for both mildly- or severely- illposed NPIV model, CC showed that

$$\sup_{s \in \mathbb{R}} \left| \mathbb{P}\left\{ \sup_{t \in \mathcal{T}} \left| \frac{\sqrt{n}\left(\phi_t(\widehat{h}) - \phi_t(h_0)\right)}{\sqrt{\widehat{V}_{\phi,t}}} \right| \leq s \right\} - \mathbb{P}^* \left\{ \sup_{t \in \mathcal{T}} |\mathbb{Z}_n^*(t)| \leq s \right\} \right| = o_p(1).$$

This could be used to construct uniform confidence bands for nonlinear functional process of a NPIV model.

5 Semiparametric Two-step GMM

The methods of sieves and penalizations are versatile, and can be used in multi-step estimation of complicated models that are ubiquitous in diverse empirical economics. See, e.g., Engle and Gonzalez-Rivera (1991), Gallant, Hansen and Tauchen (1990), Conley and Dupor (2003), Engle and Rangel (2008), Kawai (2011), Chen, Favilukis and Ludvigson (2013), Arcidiacono and Miller (2011), Nevo (2011) to name only a few. See Chen (2007, 2013), Ackerberg, et al (2012, 2014) for additional references. This section reviews results for sieve semiparametric two-step GMM estimation and inference with iid data.[7]

[7] See Chen and Liao (2015) for sieve semiparametric two-step GMM with weakly dependent data

The Model. Let $\{Z_i\}_{i=1}^n = \{(Y_i', X_i')'\}_{i=1}^n$ be a random sample from the probability distribution of $Z = (Y', X')'$. Let $g(\cdot) : \mathbb{R}^{d_Z} \times B \times \mathcal{A} \to \mathbb{R}^{d_g}$ be a vector of measurable functions with $\infty > d_g \geq d_\beta \geq 1$, B is a compact subset in $\mathbb{R}^{d_\beta}$ with a non-empty interior, and $\mathcal{A}$ is an infinite dimensional (nuisance) function space. Let $Q(\cdot) : \mathcal{A} \to \mathbb{R}$ be a non-random criterion function. A semiparametric structural model specifies that

$$\mathbb{E}[g(Z, \beta, \alpha_0(\cdot, \beta))] = 0 \text{ at } \beta = \beta_0 \in \text{int}(B), \tag{5.1}$$

and for any fixed $\beta \in B$, $\alpha_0(\cdot, \beta) \in \mathcal{A}$ solves

$$Q(\alpha_0) = \inf_{\alpha \in \mathcal{A}} Q(\alpha). \tag{5.2}$$

If the nuisance parameter $\alpha_0(.)$ were known, the finite dimensional parameter of interest β_0 is over-identified by d_g moment conditions in (5.1). But $\alpha_0(.)$ is unknown, except that it is identified by (5.2). As in Newey (1994) and Chen Linton and van Keilegom (CLvK 2003), we allow for $\alpha_0(\cdot, \beta) \in \mathcal{A}$ to depend on β and data. We use a simplified notation, $(\beta_0, \alpha_0) \equiv (\beta_0, \alpha_0(, \beta_0))$, throughout this section. The GMM moment function $g(Z_i, \beta, \alpha(\cdot))$ is allowed to depend on the entire nuisance functions $\alpha(\cdot)$ and not just their values at observed data points. Finally, the parameter $\alpha(\cdot)$ could consist of both finite dimensional parameter θ and infinite dimensional functions $h(\cdot)$; that is, $\alpha(\cdot, \beta) = (\theta', h(\cdot, \beta)) \in \Theta \times \mathcal{H} = \mathcal{A}$ as in Section 2, except that $\dim(\Theta)$ could be zero in many semiparametric two-step GMM applications which corresponds to $\mathcal{A} = \mathcal{H}$ and $\alpha(\cdot, \beta) = h(\cdot, \beta)$.

Sieve semiparametric two-step GMM estimation. In the first-step the unknown nuisance functions $\alpha_0(\cdot)$ is estimated via an approximate sieve extremum estimation, i.e.,

$$\widehat{Q}_n(\widehat{\alpha}_n) \leq \inf_{\alpha \in \mathcal{A}_{k(n)}} \widehat{Q}_n(\alpha) + o_p(n^{-1}), \tag{5.3}$$

where $\widehat{Q}_n(.)$ is a random criterion function such that $\sup_{\alpha \in \mathcal{A}_{k(n)}} \left| \widehat{Q}_n(\alpha) - Q(\alpha) \right| = o_p(1)$, and $\mathcal{A}_{k(n)} = \Theta \times \mathcal{H}_{k(n)}$ is a sieve space for $\mathcal{A} = \Theta \times \mathcal{H}$ as reviewed in Section 2. In the second-step, the first-step sieve extremum estimator $\widehat{\alpha}_n$ is plugged into some unconditional moment conditions and the unknown β_0 is estimated by GMM

$$\widehat{\beta}_n = \arg\min_{\beta\in B} \left[\frac{1}{n}\sum_{i=1}^{n} g(Z_i,\beta,\widehat{\alpha}_n)\right]' W_n \left[\frac{1}{n}\sum_{i=1}^{n} g(Z_i,\beta,\widehat{\alpha}_n)\right] \tag{5.4}$$

where W_n is a $d_g \times d_g$ positive definite (possibly random) matrix, with $plim_n W_n = W$.

The definition of sieve semiparametric two-step GMM estimation consists of equations (5.3) and (5.4). As demonstrated in Chen (2007, 2013), sieve extremum estimation in the first-step (5.3) is very flexible and can estimate unknown functions in most semi-nonparametric models by different choices of criterion $\widehat{Q}_n()$ and sieves $\mathcal{A}_{k(n)}$. For example, if $\alpha_0(.)$ is identified as a solution to $\inf_{\alpha\in\mathcal{A}} -E[\varphi(Z,\alpha)]$ for a measurable function $\varphi(Z,\alpha) : \mathbb{R}^{d_z} \times \mathcal{A} \to \mathbb{R}$, then one can use sieve M-estimation (e.g., Least Square, Quantile, quasi MLE) with $Q(\alpha) = -E[\varphi(Z,\alpha)]$ and $\widehat{Q}_n(\alpha) = -\frac{1}{n}\sum_{i=1}^{n}\varphi(Z_i,\alpha)$. If $\alpha_0(.)$ is identified through conditional moment restrictions model (2.1) $E[\rho(Z,\alpha_0)|X] = 0$, then one can use the SMD estimation with $Q(\alpha) = E[m(X,\alpha)'m(X,\alpha)]/2$ and $\widehat{Q}_n(\alpha) = \frac{1}{2n}\sum_{i=1}^{n}\widehat{m}(X_i,\alpha)'\widehat{m}(X_i,\alpha)$, where $\widehat{m}(X,\alpha)$ is any consistent estimator (say 2.13) of the conditional moment function $m(X,\alpha) = E[\rho(Z,\alpha)|X]$ in Section 2. Of course sieve GMM or sieve EL as described in Section 2 for model (2.1) could also be used in the first-step (5.3). Likewise, instead of the second-step GMM (5.4), one could apply EL or GEL to estimate β_0 identified by model (5.1). All these different variations give the same asymptotic variance for estimating β_0 in terms of first-order asymptotic theory. Therefore, we only review the simplest GMM estimator (5.4) in the second step.

Root-n asymptotic normality of the second-step GMM estimator

Let $G(\beta,\alpha) \equiv E[g(Z,\beta,\alpha)]$. For any $(\beta,\alpha) \in B \times \mathcal{A}$, we denote the ordinary derivative of $G(\beta,\alpha)$ with respect to β as $\Gamma_1(\beta,\alpha)$. Let $\Gamma_1 = \Gamma_1(\beta_0,\alpha_0)$ and $\Gamma_1' W \Gamma_1$ be non-singular. Under some regularity conditions (see e.g., Newey (1994), CLvK, Chen (2007), Chen and Liao (2015)), the second-step GMM estimator $\widehat{\beta}_n$ is $\sqrt{n}$-estimable and satisfies

$$\sqrt{n}(\widehat{\beta}_n - \beta_0) = -(\Gamma_1' W \Gamma_1)^{-1} \Gamma_1' W \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g\left(Z_i, \beta_0, \widehat{\alpha}_n\right) + o_p(1) \xrightarrow{d} \mathcal{N}\left(0, V_\beta\right)$$

where $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g\left(Z_i, \beta_0, \widehat{\alpha}_n\right) \xrightarrow{d} \mathcal{N}\left(0, V_1\right)$ for a finite positive definite V_1, and

$$V_\beta = \left(\Gamma_1' W \Gamma_1\right)^{-1} \left(\Gamma_1' W V_1 W \Gamma_1\right) \left(\Gamma_1' W \Gamma_1\right)^{-1}. \tag{5.5}$$

We call the estimator $\widehat{\beta}_n$ with $W_n = V_1^{-1} + o_p(1)$ "semiparametric two-step optimally weighted GMM" since its asymptotic variance becomes $V_\beta = (\Gamma_1' V_1 \Gamma_1)^{-1}$, the smallest among the class of semiparametric two-step GMM estimators. See Ai and Chen (2012), Ackerberg, et al (2014) and Chen and Santos (2015) for conditions under which it achieves the full semiparametric efficiency bound for β_0 of the models (5.1)-(5.2).

Note that V_1 captures the effect (or influence) of the first-step nonparametric estimation of α_0 on the second-step GMM estimation of β_0. For any $(\beta, \alpha) \in B \times \mathcal{A}$, let $\Gamma_2(\beta, \alpha)[v] = \left(\Gamma_{2,1}(\beta, \alpha)[v], \ldots, \Gamma_{2,d_g}(\beta, \alpha)[v]\right)'$ $\left.\frac{\partial G(\beta, \alpha + \tau v)}{d\tau}\right|_{\tau=0}$ denote the pathwise derivative of $G(\beta, \alpha)$ with respect to $\alpha \in \mathcal{A}$ in the direction v with $\{\alpha + \tau v : \tau \in [0, 1]\} \subset \mathcal{A}$. Under mild regularity conditions

$$V_1 = \lim_{n \to \infty} Var\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(Z_i, \beta_0, \alpha_0) + \sqrt{n} \Gamma_2(\beta_0, \alpha_0)[\widehat{\alpha}_n - \alpha_0]\right).$$

When $\sqrt{n}\Gamma_2(\beta_0, \alpha_0)[\widehat{\alpha}_n - \alpha_0] = o_p(1)$, which is essentially the asymptotic orthogonality condition $\sqrt{n}G(\beta_0, \widehat{\alpha}_n) = o_p(1)$ of Andrews (1994, equation (2.12)), we have $V_1 = Var\left(g\left(Z, \beta_0, \alpha_0\right)\right)$, and hence the first-step nonparametric estimation of α_0 does not affect the second-step GMM estimation of β_0. For typical semiparametric two step estimation we have $\sqrt{n}\Gamma_2(\beta_0, \alpha_0)[\widehat{\alpha}_n - \alpha_0] = O_p(1)$ and the first-step nonparametric estimation of α_0 will affect the second-step estimation of β_0.

Sieve approximations to V_1 and V_β. When $\sqrt{n}\Gamma_2(\beta_0, \alpha_0)[\widehat{\alpha}_n - \alpha_0] = O_p(1)$ it is generally difficult to calculate V_1 (and hence V_β) in closed forms when the first-step semi-nonparametric model (5.2) is complicated, say when the first-step model contains several unknown functions.

Built upon insight from Newey (1994), Ai and Chen (2007), Ackerberg et al (2012) and under some regularity conditions, Chen and Liao (2015) established that the unknown asymptotic variance V_1 could be approximated by a sieve variance $V_{1,n}^* = Var\left(S_{n,i}^*\right)$, where

$$S_{n,i}^* = g(Z_i, \beta_0, \alpha_0) + \Big(\Delta(Z_i, \alpha_0)[v_{1,n}^*], \ldots, \Delta(Z_i, \alpha_0)[v_{d_g,n}^*]\Big)', \tag{5.6}$$

and for each $j = 1, ..., d_g$, the sieve adjustment term $\Delta(Z_i, \alpha_0)[v_{j,n}^*]$ satisfies

$$E\left(\Delta(Z_i, \alpha_0)[v_{j,n}^*]\right) = -\left[\frac{\partial}{\partial\tau} Q(\alpha_0 + \tau v_{j,n}^*)\right]\Bigg|_{\tau=0} \quad \text{and } v_{j,n}^*(.) = \overline{q}^{k(n)}(.)' D_n^- F_{j,n}$$

where $\overline{q}^{k(n)}(.)$ is a vector of linear basis used to approximate $\mathcal{A}_{k(n)}$, $F_{j,n} = \Gamma_{2,j}(\beta_0, \alpha_0)[\overline{q}^{k(n)}]$, and D_n is a $k(n) \times k(n)$ positive definite matrix such that

$$\gamma' D_n \gamma \equiv \left[\frac{\partial^2}{\partial\tau^2} Q\left(\alpha_0(.) + \tau \overline{q}^{k(n)}(.)'\gamma\right)\right]\Bigg|_{\tau=0} \quad \text{for all} \quad \gamma \in \mathbb{R}^{k(n)}.$$

Note that $S_{n,i}^*$ given in (5.6) could be viewed as a sieve score term for estimating β_0 via the sieve semiparametric two-step GMM procedure. Let

$$V_{1,n}^* = Var\left(S_{n,i}^*\right) \quad \text{and} \quad V_{\beta,n} = \left(\Gamma_1' W \Gamma_1\right)^{-1}\left(\Gamma_1' W V_{1,n}^* W \Gamma_1\right)\left(\Gamma_1' W \Gamma_1\right)^{-1}, \tag{5.7}$$

both could be computed in closed forms once the linear sieve for $\mathcal{A}_{k(n)}$ is chosen. Chen and Liao (2015) establishes that $V_1 = \lim_{n\to\infty} V_{1,n}^*$, $V_\beta = \lim_{n\to\infty} V_{\beta,n}$ and

$$\sqrt{n}\,(V_{\beta,n})^{-1/2}\,(\widehat{\beta}_n - \beta_0) \xrightarrow{d} \mathcal{N}(0, I_{d_\beta}).$$

Sieve Wald and Overidentification Hansen's J tests

The closed form sieve variance expression (5.7) immediately implies that one can estimate $V_{\beta,n}$ by component-wise empirical analog just as in Subsection 3.2.1. Given a consistent estimator of the semiparametric two-step GMM variance $\widehat{V}_{\beta,n}$, we have

$$\sqrt{n}\widehat{V}_{\beta,n}^{-1/2}\left(\widehat{\beta}_n - \beta_0\right) \xrightarrow{d} \mathcal{N}(0, I_{d_\beta}).$$

Furthermore, we can conduct inference about β_0 through the standard Wald test of $\beta = \beta_0$ from

$$\mathcal{W}_n = n(\widehat{\beta}_n - \beta_0)'\widehat{V}_{\beta,n}^{-1}(\widehat{\beta}_n - \beta_0) \xrightarrow{d} \chi^2_{d_\beta}.$$

Here χ^2_d stands for Chi-squared distribution with d degrees of freedom.

Similarly, we can construct Hansen style overidentification J test of $E[g(Z, \beta_0, \alpha_0)] = 0$. Let $\widehat{W_n}$ be a positive definite weighting matrix such that $\widehat{W_n} = (V^*_{1,n})^{-1} + o_p(1)$. The overidentification test statistic is

$$\mathcal{J}_n = \left[n^{-1/2}\sum_{i=1}^{n} g(Z_i, \widehat{\beta}_n, \widehat{\alpha}_n)\right]' \widehat{W_n} \left[n^{-1/2}\sum_{i=1}^{n} g(Z_i, \widehat{\beta}_n, \widehat{\alpha}_n)\right],$$

and $\mathcal{J}_n \xrightarrow{d} \chi^2_{d_g - d_\beta}$ under the null. See Chen and Liao (2015) for details.

For weakly dependent time series data, Chen and Liao (2015) first propose a consistent sieve estimator of the $Avar(\widehat{\beta}_n)$ for a sieve semiparametric two-step GMM estimator when the first step unknown function is estimated via sieve M or MD estimation. They then show that this consistent estimate of the *semiparametric* $Avar(\widehat{\beta}_n)$ is *numerically identical* to the estimate of the *parametric* asymptotic variance using the standard parametric two-step framework for time series data. These results greatly simplify the computation of semiparametric standard errors of semiparametric two-step GMM estimators for time series models.

Control Function (CF) Approach

As mentioned in the introduction, estimation and inference for semiparametric models with endogeneity via CF approach are typically conducted in semiparametric two-step or multi-step, where a nonparametric conditional mean or quantile regression without endogeneity, or a conditional choice probability or other reduced form unknown functions of exogenous variables are estimated in the first step. Semiparametric CF approach is widely used in empirical studies in labor economics and industrial organization. See, e.g., Heckman (1979), Heckman and Robb (1985), Olley and Pakes (1996), Wooldridge (2009), Arcidiacono and Miller (2011), Ackerberg, Caves and Frazer (2015), De Loecker, Goldberg, Khandelwal and Pavcnik (2015), Gandhi, Navarro and Rivers (2015). Existing theoretical work on sieve semiparametric and nonparametric CF approach include, but are not limited to, Newey, Powell and Vella (1999), Ai and Chen (2007, 2012), Imbens and Newey (2009), Blundell, Kristensen and Matzkin (2013), Ackerberg, Chen and Hahn (2012), Ackerberg, Chen, Hahn and Liao (2014). The theory for the sieve semiparametric two-step GMM remains valid for sieve semiparametric CF problems. In particular, when the reduced form unknown functions in a semiparametric or a nonparametric CF problem are approximated and estimated via finite dimensional linear sieves, the asymptotic variance of the $\widehat{\beta}_n$ in the final stage could be consistently estimated by the variance estimator for the corresponding parametric CF problem (as that in Wooldridge 2002).

6 Simulation

This section evaluates the performance of the SMD estimation procedures for the NPIV model using Monte Carlo simulation.[8] See Blundell, Chen and Kristensen (2007), Chen and Pouzo (2009, 2012, 2015), Chen and Christensen (2015), Horowitz (2011, 2014) for additional Monte Carlo studies and empirical applications of sieve NPIV and NPQIV estimators.

6.1 Experiment 1: Partially Linear Additive IV Regression

We first consider a partially linear additive IV regression similar to Example 1 and the Monte Carlo example in Chen (2007, p. 5580). The true model is $Y_1 = Y_2\theta_0 + h_{01}(Y_3) + h_{02}(X_2) + \varepsilon$ with $\theta_0 = 1$, $h_{01}(Y_3) = 1/(1+\exp(-Y_3))$ and $h_{02}(X_2) = \log(1+X_2)$ with location constraint $h_{02}(0.5) = \log(3/2)$. Y_2 and Y_3 are endogenous, having $Y_2 = X_1 + 0.5\varepsilon_2 + e$ and $Y_3 = \Phi(U_3 + 0.5\varepsilon_3)$; $X_2 \sim \text{uniform}[0,1]$,

[8] We conducted the Monte Carlo experiments using R. Sample codes using STATA and R are available upon request.

$X_1 = \Phi(U_2)$, $X_3 = \Phi(U_3)$, $\varepsilon = (\varepsilon_1+\varepsilon_2+\varepsilon_3)/3$, $\varepsilon_1, \varepsilon_2, \varepsilon_3, U_2, U_3$ are independent normally distributed with mean 0 and variance 1 and e is normally distributed with mean 1 and variance 0.1. $\Phi(.)$ is the standard normal CDF. We use the SMD estimation procedure described in Example 1 to obtain the estimates of $\alpha_0 = (\theta_0, h_{01}, h_{02})$. Table 1 reports the performance of the SMD estimator of $\widehat{\theta}$ and functions $\widehat{h}_1, \widehat{h}_2$ evaluated at a point as well as their sieve variance estimators with respect to the Monte Carlo standard variance. It is apparent that the simple sieve variance formulation of $\widehat{\theta}$ performs well even when the regressor is endogenous.

Table 1: Model in Experiment 1

	θ	$\widehat{\text{SE}}(\theta)$	$\lvert\text{bias}(h_1(\overline{y_3}))\rvert$	$\widehat{\text{SE}}(h_1(\overline{y_3}))$	$\lvert\text{bias}(h_2(\overline{x_2}))\rvert$	$\widehat{\text{SE}}(h_2(\overline{x_2}))$
Model 1	1.0146	0.0658 (0.0670)	0.0547	0.0648 (0.0668)	0.0010	0.0031 (0.0036)
Model 2	1.0139	0.0693 (0.0682)	0.0363	0.0987 (0.0941)	0.0010	0.0033 (0.0041)

We generate a 1000 observation i.i.d. sample with 1000 Monte Carlo replications. The column θ refers to the Monte Carlo average of estimator $\widehat{\theta}$; the column $\lvert\text{bias}(h_1(\overline{y_3}))\rvert$ refers to the Monte Carlo average of the bias in absolute value of the estimated function h_1 evaluated at a point $\overline{y_3}$ (median of y_3); the column $\widehat{\text{SE}}(h_1(\overline{y_3}))$ refers to the Monte Carlo median of the estimated standard error (square-root of the sieve variance) of $h_1(\overline{y_3})$ with Monte Carlo standard error displayed beneath it in parenthesis. Similarly for $\lvert\text{bias}(h_2(\overline{x_2})\rvert$ and $\widehat{\text{SE}}(h_2(\overline{x_2}))$. Model 1 uses both Legendre Polynomial to approximate h_i and Model 2 uses cubic spline to approximate h_1 and Legendre Polynomial to approximate h_2. The sieve dimensions are chosen using simple AIC procedure. Other choices of sieve are considered and they yield similar results. Due to lack of space, they are not reported here.

6.2 Experiment 2: NPIV Adaptive Procedure

Next, we use the Monte Carlo design in Chen and Christensen (2015, CC) to illustrate performance of adaptive procedures. The true model is $Y_1 = h_0(Y_2)+\varepsilon$, where $h_0(Y_2) = \log(|6Y_2-3|+1)\text{sign}(Y_2-0.5)$. We generate (ε, V^*, X^*) from

$$\begin{bmatrix} \varepsilon \\ V^* \\ X^* \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\right),$$

and set $Y_2 = \Phi((X^* + V^*)/\sqrt{2})$, $X = \Phi(X^*/\sqrt{2})$. We first implement Horowitz's (2014) data-driven procedure using a Legendre Polynomial basis orthonormalized with respect to the $L^2([0,1])$ inner product. Then we use CC's adaptive procedure using both a Cubic B-spline with interior knots placed evenly and a Legendre Polynomial. Table 2 presents the MC average sup-norm, L^2-norm loss, the sup-norm relative error ratio $||\widehat{h}_{\widehat{K}} - h_0||_\infty / ||\widehat{h}_{K_\infty} - h_0||_\infty$ where $K_\infty = \arg\min_{k \in I_K} ||\widehat{h}_K - h_0||_\infty$ and the L^2-norm relative error ratio $||\widehat{h}_{\widehat{K}} - h_0||_{L^2(Y_2)} / ||\widehat{h}_{K_{L^2}} - h_0||_{L^2(Y_2)}$ where $K_{L^2} = \arg\min_{k \in I_k} ||\widehat{h}_K - h_0||_{L^2(Y_2)}$. I_k represents the set of the possible tuning parameter pairs for each adaptive procedure.[9] The table can be understood as the follow, the sup-norm loss of CC's data-driven estimator $\widehat{h}$ using cubic Basis-Spline to approximate the unknown function $h(.)$ is at most 6.04% larger than that of the infeasible estimator.

Table 2: Model in Experiment 2

	q^K	p^J	sup ratio	L^2 ratio	sup	L^2
H	Leg	Leg	1.8622	1.3386	0.4283	0.1614
CC	4	5	1.0554	1.0179	0.3879	0.1488
CC	4	Leg	1.0604	1.0195	0.3641	0.1430
CC	Leg	Leg	1.1378	1.1910	0.2476	0.1310

H and CC stand for Horowitz and Chen and Christensen (2015) procedure, respectively. q^K and p^J stand for the bases used to approximate the unknown function $h_0(.)$ and the conditional mean function $m(X, \alpha)$. Leg, 4 and 5 stand for the Legendre Polynomial, 4^{th} order (Cubic) Basis-Spline, and 5^{th} order (Quartic) Basis-Spline, respectively. J is chosen to be $2K$ for the CC procedure.

Next, we consider conducting inference over regular functional $\phi(h) = \int w(y_2)\nabla h(y_2)dy_2$, and irregular functionals $\phi(h) = \int h(y_2)dy_2$, and $\phi(h) = \int w(y_2)(h(y_2))^2 dy_2$ where $w(y) = 6(y - y^2)$ is a known weighting function over $[0,1]$. In addition, we estimate the weighted average derivative $\mathbb{E}[w(Y_2)\nabla h(Y_2)]$ using the plug-in estimator $\frac{1}{n}\sum_{i=1}^{n} w(Y_{2i})\nabla\widehat{h}(Y_{2i})$. For each functional, we calculate the sieve variance as in Chen and Pouzo (2015) and Chen and Liao (2015) then we compute the sieve student-t statistic. Table 3 reports the MC rejection frequencies for t test $\sqrt{n}\frac{\phi(\widehat{h}) - \phi(h_0)}{\sqrt{\widehat{V}_{\phi,n}}}$; table 4 reports the MC standard deviation and bias of $\phi(\widehat{h})$.[10] From the Monte Carlo experiments we

[9] We use the notation in Chen and Pouzo (2009, 2012 and 2015) consistently throughout this section. CC used J as the regularization parameter and K as the smoothing parameter.

[10] Tables 2, 3 and 4 are generated with 3000 Monte Carlo replications and sample size $n = 1000$.

observe that the choice of both basis and sieve dimension $\widehat{K}$ do not matter for regular functionals. In contrast, it is recommended to choose $\widehat{K}_{max}$ to inflate variance and reduce bias for irregular functionals. This case is well-illustrated under the Legendre Polynomial basis example where $\widehat{K}_{max}$ corresponds to a higher order polynomial. We refer readers to Blundell, et al. (2007) and Chen and Pouzo (2012) for more Monte Carlo results.

Table 3: Model in Experiment 2

regular functional			$\int w(y_2)\nabla h(y_2)dy_2$		$\mathbb{E}[w(Y_2)\nabla h(Y_2)]$	
	q^K	p^J	5%	10%	5%	10%
$\widehat{K}$	4	5	0.0560	0.1030	0.0477	0.0943
$\widehat{K}_{max}$	4	5	0.0517	0.0983	0.0423	0.0877
$\widehat{K}$	4	Leg	0.0600	0.1020	0.0473	0.0967
$\widehat{K}_{max}$	4	Leg	0.0527	0.1000	0.0459	0.0887
$\widehat{K}$	Leg	Leg	0.0613	0.1203	0.0667	0.1200
$\widehat{K}_{max}$	Leg	Leg	0.0507	0.0950	0.0430	0.0883
irregular functional			$\int w(y_2)\left(h(y_2)\right)^2 dy_2$		$\int_{[0.05,0.95]} \nabla h(y_2)dy_2$	
	q^K	p^J	5%	10%	5%	10%
$\widehat{K}$	4	5	0.0633	0.1123	0.0463	0.0960
$\widehat{K}_{max}$	4	5	0.0330	0.0630	0.0433	0.0897
$\widehat{K}$	4	Leg	0.0647	0.1150	0.0503	0.1000
$\widehat{K}_{max}$	4	Leg	0.0313	0.0637	0.0437	0.0937
$\widehat{K}$	Leg	Leg	0.5807	0.6953	0.6913	0.7833
$\widehat{K}_{max}$	Leg	Leg	0.0350	0.0623	0.0417	0.0823

Using CC procedure with B-Spline, Figure 1 displays the estimated function, the pointwise confidence bands and the 95% uniform confidence bands for a representative sample. Further, we construct the uniform confidence bands for $h_0(.)$ over the support $[0.05, 0.95]$ using the Horowitz and Lee (2012) and CC procedures with 1000 bootstrap replications. The sieve dimensions for their procedures are chosen adaptively using Horowitz (2014) and CC respectively. Table 5 reports the MC coverage probabilities. Both uniform bands perform well.

Table 4: Model in Experiment 2

regular functional			$\int w(y_2)\nabla h(y_2)dy_2$		$\mathbb{E}[w(Y_2)\nabla h(Y_2)]$	
	q^K	p^J	sd	$\vert\text{bias}\vert$	sd	$\vert\text{bias}\vert$
$\widehat{K}$	4	5	0.1751	0.0025	0.1728	0.0118
$\widehat{K}_{max}$	4	5	0.1797	0.0025	0.1906	0.0116
$\widehat{K}$	4	Leg	0.1741	0.0003	0.1718	0.0086
$\widehat{K}_{max}$	4	Leg	0.1787	0.0005	0.1902	0.0132
$\widehat{K}$	Leg	Leg	0.1599	0.0443	0.1547	0.0472
$\widehat{K}_{max}$	Leg	Leg	0.1798	0.0036	0.1912	0.0175
irregular functional			$\int w(y_2)\left(h(y_2)\right)^2 dy_2$		$\int_{[0.05,0.95]} \nabla h(y_2)dy_2$	
	q^K	p^J	sd	$\vert\text{bias}\vert$	sd	$\vert\text{bias}\vert$
$\widehat{K}$	4	5	0.1484	0.0041	0.3144	0.0113
$\widehat{K}_{max}$	4	5	0.1988	0.0502	0.3177	0.0076
$\widehat{K}$	4	Leg	0.1420	0.0015	0.3016	0.0245
$\widehat{K}_{max}$	4	Leg	0.1680	0.0430	0.3056	0.0234
$\widehat{K}$	Leg	Leg	0.0503	0.1060	0.1440	0.2208
$\widehat{K}_{max}$	Leg	Leg	0.2124	0.0354	0.3219	0.0046

Table 5: Model in Experiment 2

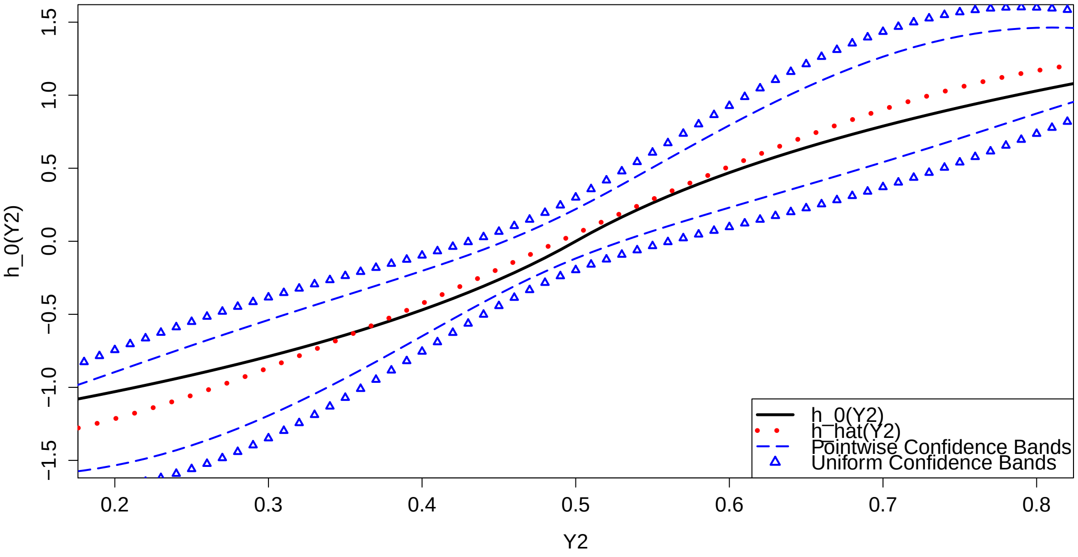
Coverage Probabilities	90% CI	95% CI	99% CI
Horowitz and Lee	0.892	0.946	0.982
Chen and Christensen	0.898	0.946	0.984

This table is constructed with 500 Monte Carlo replications with sample size of 1000. We use the orthonormalized Legendre Polynomial for Horowitz's procedure, and we use a Cubic B-Spline for q^K and Quartic B-Spline for p^J with $J = 2K$ for CC's procedure with K chosen adaptively. Horowitz and Lee's bootstrap procedure requires estimation of the unknown function for each bootstrap sample, and hence is computationally more expensive than CC's.

6.3 Experiment 3: Bootstrap Uniform Confidence Bands for Functional of NPIV

The score bootstrap uniform confidence bands established in CC apply to general functionals of NPIV. To illustrate, we generate data from $Y_1 = h_0(Y_2, Y_3) + \varepsilon$, $h_0(Y_2, Y_3) = \sin(\pi Y_2)\exp(Y_3)$ and $\mathbb{E}[\varepsilon|X_1, X_2] = 0$. Y_2 and Y_3 are endogenous, having $Y_2 = \Phi(X_1 - X_2 + 0.5\varepsilon + e_1)$, $Y_3 = \Phi(X_1 X_2 + 0.3\varepsilon + e_2)$, X_1 and X_2 are standard uniformly distributed, e_1, e_2 are standard normally distributed and ε is normally distributed with mean 0 and variance 0.1. h_0 is estimated using PSMD with a Cubic Spline tensor product basis, and a penalty function taken as the squared L^2 norm of

Figure 1: Model in Experimen 2
95% Pointwise Bands and Uniform Bands of h_0(Y2)



the second partial derivative of h with respect to y_2. The sieve dimension is chosen using a simple AIC procedure. Figure 2 displays the uniform confidence bands of functional $\frac{\partial h_0(y_2,0.5)}{\partial y_2}$, the partial derivative of h_0 with respect to y_2 evaluated at $y_3 = 0.5$.

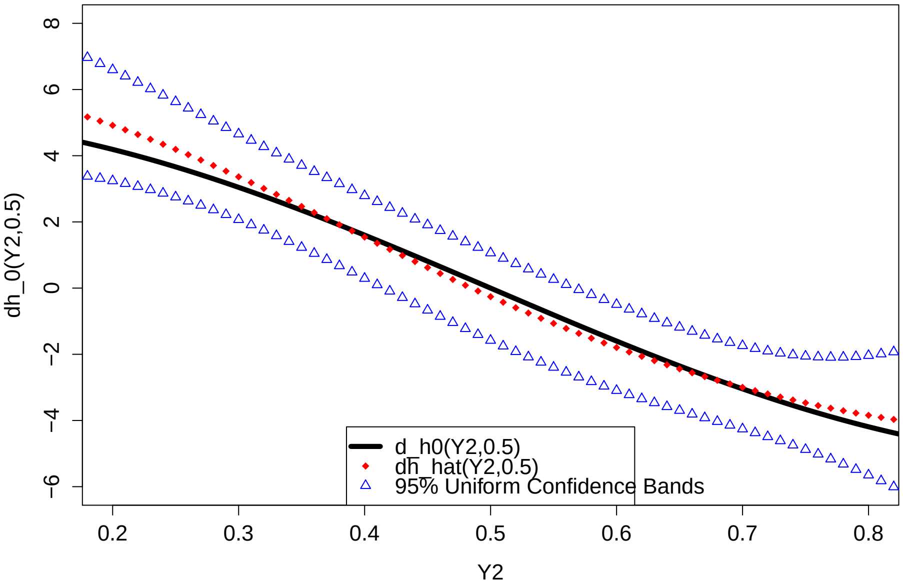
7 Concluding Remarks

In this brief review, we have described (penalized) sieve estimation and inference for general semiparametric and nonparametric models with endogeneity. Currently available large sample theories are mostly developed for sieve MD and sieve two-step or multi-step GMM procedures for models with endogeneity under i.i.d. data and using (penalized) finite-dimensional linear sieves. The easy implementation of linear sieve MD procedures is illustrated via examples and Monte Carlo studies.

We conclude by briefly mentioning additional existing works and unsolved problems in the literature on inference for models with nonparametric endogeneity.

Additional estimation and inference results: (1) There are many papers on Kernel-based estimation of NPIV and NPQIV under i.i.d. data. See Hall and Horowitz (2005), Horowitz and Lee (2007), Carrasco, Florens and Renault (2007), Darolles, Fan, Florens and Renault (2011), Gagliar-

Figure 2: Model in Experiment 3
95% Uniform Confidence Bands of dh_0(Y2,0.5)



dini and Scaillet (2012) and the references therein. There are also a few papers on Bayesian method for NPIV with independent data. See Liao and Jiang (2011), Florens and Simoni (2012), Kato (2013) and the references therein. (2) There are some published papers on kernel- or sieve- based specification tests for NPIV and NPQIV under i.i.d. data. See Horowitz (2006, 2011, 2012, 2014), Breunig (2015a, 2015b) and the references therein. (3) There are also some very recent works on sieve inference on partially identified nonparametric conditional moment restrictions with endogeneity under i.i.d. data. See, for example, Santos (2012) for NPIV and Hong (2012) for NPQIV via sieved version of Bieren's type test, Tao (2015) and Chernozhukov, Newey and Santos (2015) for partially identified conditional moment restrictions via sieve GMM.

Some open questions: (1) Besides a few papers on NPIV with i.i.d. data, there is nothing about data-driven choices of smoothing parameters for linear sieve estimation and inference on general nonparametric conditional moment restrictions with endogeneity. (2) There is no published work on any second-order asymptotic theories for linear sieve estimation and inference on nonparametric conditional moment restrictions with endogeneity. It is easy to conjecture that, given the same finite dimensional linear sieve approximation, the sieve GEL would have second-order refinement over the optimally weighted sieve MD and sieve GMM. However there is no formal theoretical proofs yet. (3) Sieve MD and sieve GMM methods have been used to estimate conditional moment restrictions containing unknown functions of endogeneous variables in empirical studies with temporal or/and spatial dependent data sets. However, the general inferential theory allowing for dependent data has not been fully developed yet.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

Chen is grateful to Steve Berry and Richard Blundell for their encouragements to write such a review. We thank Richard Blundell and the referees for useful comments, and Cowles Foundation for research support. The usual disclaimer applies.

References

[1] Ackerberg, D., K. Caves and G. Frazer (2015) “Identification Properties of Recent Production Function Estimators ”, *Econometrica*, 83, 2411-2451.

[2] Ackerberg, D., X. Chen, and J. Hahn (2012) “A Practical Asymptotic Variance Estimator for Two-Step Semiparametric Estimators”, *Review of Economics and Statistics*, 94, 482-498.

[3] Ackerberg, D., X. Chen, J. Hahn and Z. Liao (2014) “Asymptotic Efficiency of Semiparametric Two-step GMM”, *Review of Economic Studies*, 81, 919-943.

[4] Ai, C. and X. Chen (1999): “A Kernel Estimation of Semiparametric Conditional Moment Restriction Models with Different Conditioning Variables for Different Equations,” *Working Paper.*

[5] Ai, C. and X. Chen (2003): “Efficient Estimation of Conditional Moment Restrictions Models Containing Unknown Functions,” *Econometrica,* 71, 1795-1843.

[6] Ai, C. and X. Chen (2007): “Estimation of Possibly Misspecified Semiparametric Conditional Moment Restriction Models with Different Conditioning Variables,” *Journal of Econometrics* 141, 5-43.

[7] Ai, C. and X. Chen (2012): "The Semiparametric Efficiency Bound for Models of Sequential Moment Restrictions Containing Unknown Functions," *Journal of Econometrics* 170, 442-457.

[8] Anastassiou, G., Yu, X. (1992a): "Monotone and Probabilistic Wavelet Approximation," *Stochastic Analysis and Applications* 10, 251-264.

[9] Anastassiou, G., Yu, X. (1992b): "Convex and Convex-Probabilistic Wavelet Approximation," *Stochastic Analysis and Applications* 10, 507-521.

[10] Andrews, D. (1994): "Asymptotics for Semi-parametric Econometric Models via Stochastic Equicontinuity," *Econometrica* 62, 43-72.

[11] Arcidiacono, P. and R. Miller (2011): "Conditional Choice Probability Estimation of Dynamic Discrete Choice Models with Unobserved Heterogeneity," *Econometrica* 79, 1823-1868.

[12] Bajari, P., H. Hong, and D. Nekipelov (2011) "Game Theory and Econometrics. A Survey of Some Recent Research,"Working Paper, Stanford and UC Berkeley.

[13] Berry, S., and P. Haile (2014) "Identification in Differentiated Products Markets Using Market Level Data", *Econometrica*, 82, 1749-1797.

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

[15] Blundell, R., J. Horowitz, and M. Parey (2012) "Measuring the price responsiveness of Gasoline Demand: Economic Shape Restrictions and Nonparametric Demand Estimation ", *Quantitative Economics*, 3(1), 29-51.

[16] Blundell, R., J. Horowitz, and M. Parey (2015) "Nonparametric Estimation of a Heterogeneous Demand Function under Slutsky Inequality Restriction ", *Working Paper* CWP54/13, commap.

[17] Blundell, R., D. Kristensen, and R. Matzkin (2013) "Control Functions and Simultaneous Equation Methods", *American Economic Review*, 103, 563-69.

[18] Blundell, R. and R. Matzkin (2014) "Conditions for the Existence of Control Functions in Nonseparable Simultaneous Equations Models ", *Quantitative Economics*, 5, 271-295.

[19] Blundell, R. and J.L. Powell (2003) "Endogeneity in Nonparametric and Semiparametric Regression Models", in M. Dewatripont, L.P. Hansen and S.J. Turnovsky (eds.), *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, Vol. 2. Cambridge, UK: Cambridge University Press.

[20] Bontemps, C., and D. Martimort (2013) "Identification and Estimation of Incentive Contracts Under Asymmetric Information: An Application to the French Water Sector,"Working Paper, Toulouse School of Economics.

[21] Breunig, C. (2015a) "Goodness-of-Fit Tests Based on Series Estimators in Nonparametric Instrumental Regression "*Journal of Econometrics,* 184(2) 328-346.

[22] Breunig, C. (2015b) "Specification Testing in Nonparametric Instrumental Quantile Regression " *Working Paper.*

[23] Breunig, C. and J. Johannes (2015) "Adaptive Estimation of Functionals in Nonparametric Instrumental Regression, "*Econometric Theory* 1-43.

[24] Carrasco, M., J.-P. Florens and E. Renault (2007) "Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization", in J.J. Heckman and E.E. Leamer (eds.), *The Handbook of Econometrics*, vol. 6. Amsterdam: North-Holland.

[25] Chamberlain, G. (1992a) “Efficiency Bounds for Semiparametric Regression”, *Econometrica*, 60, 567-596.

[26] Chamberlain, G. (1992b) “Comment: sequential moment restrictions in panel data,” *Journal of Business and Economic Statistics* 10, 20-26.

[27] Chang, J., X.S. Chen, and X. Chen (2015) “High Dimensional Generalized Empirical Likelihood or Moment Restrictions with Dependent Data, ” *Journal of Econometrics*, 185, 283-304.

[28] Chen, X. (2007) “Large Sample Sieve Estimation of Semi-Nonparametric Models”, in J.J. Heckman and E.E. Leamer (eds.), *The Handbook of Econometrics*, vol. 6B. Amsterdam: North-Holland.

[29] Chen, X. (2013) “Penalized Sieve Estimation and Inference of Semi-Nonparametric Dynamic Models: A Selective Review ” *2010 World Congress of the Econometric Society Book Volumes*, Cambridge University Press.

[30] Chen, X. and T. Christensen (2015) “Optimal Sup-norm Rates, Adaptivity And Inference In Nonparametric Instrumental Variable Estimation ”Cowles Foundation Discussion Paper No. 1923R

[31] Chen, X., V. Chernozhukov, S. Lee and W. Newey (2014) “Local Identification of Nonparametric and Semiparametric Models”, *Econometrica*, 82, 785-809.

[32] Chen, X., J. Favilukis and S. Ludvigson (2003) “An Estimation of Economic Models with Recursive Preferences ,” *Quantitative Economics*, 4(1), 39-83.

[33] Chen, X., O. Linton and I. van Keilegom (2003) “Estimation of Semiparametric Models when the Criterion Function is not Smooth”, *Econometrica*, 71, 1591-1608.

[34] Chen, X. and Z. Liao (2015) “Sieve Semiparametric Two-step GMM Under Weak Dependence, ” *Journal of Econometrics* 189 163-186.

[35] Chen, X., Z. Liao and Y. Sun (2014) “Sieve Inference on Semi-nonparametric Time Series Models, ” *Journal of Econometrics* 178(3) 639-658.

[36] Chen, X. and S. Ludvigson (2009) “Land of Addicts? An Empirical Investigation of Habit-Based Asset Pricing Models”, *Journal of Applied Econometrics,* 24, 1057-1093.

[37] Chen, X. and D. Pouzo (2009) “Efficient Estimation of Semiparametric Conditional Moment Models with Possibly Nonsmooth Residuals”, *Journal of Econometrics*, 152, 46–60.

[38] Chen, X., and D. Pouzo (2012) “Estimation of Nonparametric Conditional Moment Models with Possible Nonsmooth Generalized Residuals ”, *Econometrica* 80 277-321

[39] Chen, X., and D. Pouzo (2015) “Sieve Quasi Likelihood Ratio Inference on Semi/nonparametric Conditional Moment Models ”, *Econometrica* 83 1013-1079

[40] Chen, X. and M. Reiß (2011) “On Rate Optimality for Ill-Posed Inverse Problems in Econometrics”, *Econometric Theory*, 27, 497-521.

[41] Chen, X. and A. Santos (2015) “Overidentification in Regular Models”, *Cowles Foundation Discussion Paper*, d1999.

[42] Chen, X. and X. Shen (1998) “Sieve Extremum Estimates for Weakly Dependent Data”, *Econometrica*, 66, 289-314.

[43] Chesher, A. (2003) “Identification in Nonseparable Models”, *Econometrica*, 77, 1405-41.

[44] Chernozhukov, V., and C. Hansen (2005) “An IV Model of Quantile Treatment Effects”, *Econometrica,* 73, 245-261.

[45] Chernozhukov, V., and C. Hansen (2013) "Quantile Models With Endogeneity", *Annual Reviews of Economics,* 5:57-81

[46] Chernozhukov, V., G.W. Imbens, and W.K. Newey (2007) "Instrumental Variable Estimation of Nonseparable Models", *Journal of Econometrics,* 139, 4-14.

[47] Chernozhukov, V., W. Newey and A. Santos (2015) "Constrained Conditional Moment Restriction Models ", Working Paper.

[48] Chetverikov, D. and W. Wilhelm (2015) "Nonparametric Instrumental Variable Estimation Under Monotonicity ", Working Paper.

[49] Chui, C (1992) An Introduction to Wavelets. Academic Press, Inc., San Diego

[50] Conley, T. and W. Dupor (2003) "A Spatial Analysis of Sectoral Complementarity," *Journal of Political Economy* 111, 311-352.

[51] Darolles, S, Y. Fan, J.-P. Florens, and E. Renault (2011) "Nonparametric Instrumental Regression", *Econometrica,* 79, 1541-1566.

[52] De Loecker, P. Goldberg, A. Khandelwal and N. Pavcnik (2015) "Prices, Markups and Trade Reform", *Econometrica,* Forthcoming.

[53] Dechevsky, L., and S. Penev (1997) "On Shape-preserving Probabilistic Wavelet Approximators ", *S-tochastic Analysis and Applications,*15(2), 187-215.

[54] DeVore, R.A. (1977a) "Monotone Approximation by Splines", *SIAM Journal on Mathematical Analysis,* 8, 891-905.

[55] DeVore, R.A. (1977b) "Monotone Approximation by Polynomials", *SIAM Journal on Mathematical Analysis,* 8, 906-921.

[56] DeVore, R.A. and G. G. Lorentz (1993) *Constructive Approximation.* Springer-Verlag, Berlin.

[57] Donald, S., G.W. Imbens, and W.K. Newey (2003) "Empirical Likelihood Estimation and Consistent Tests with Conditional moment Restrictions", *Journal of Econometrics,* 117, 55-93.

[58] Engle, R. and G. Gonzalez-Rivera (1991) "Semiparametric ARCH Models", *Journal of Business and Economic Statistics*, 9, 345-359.

[59] Engle, R. and J. G. Rangel (2008) "The Spline-GARCH Model for Low-Frequency Volatility and Its Global Macroeconomic Gauses", *The Review of Financial Studies*, 21 1187-1222.

[60] Fan, J. and I. Gijbels (1996) *Local Polynomial Modelling and Its Applications.* London: Chapman and Hall.

[61] Florens, J.P. and A. Simoni. (2012) "Nonparametric Estimation of an Instrumental Regression: A Quasi-Bayesian Approach Based On Regularized Posterior, *Journal of Econometrics* 170 458-475.

[62] Florens, J.P., J. Heckman, C. Meghir and E. Vytlacil. (2008) "Identification of Treatment Effects Using Control Functions in Models with Continuous Endogenous Treatment and Heterogeneous Effects", *Econometrica*, 76, 1191-1206.

[63] Freyberger, J., and J. Horowitz. (2013) "Identification and Shape Restrictions in Nonparametric Instrumental Variables Estimation, *Working Paper* CWP31/13, commas.

[64] Gagliardini, P. and O. Scailet (2012) "Semiparametric Estimation of Conditional Constrained Heterogenous Processes: Asset Pricing Applications", *Econometrica*, 80, 1533-1562.

[65] Gallant, A.R., Hansen, L.P., Tauchen, G., "Using Conditional Moments of Asset Payoffs to Infer the Volatility of Intertemporal Marginal Rates of Substitution, "Journal of Econometrics 45, 141-179.

[66] Gallant, A.R. and G. Tauchen (1989) "Semiparametric Estimation of Conditional Constrained Heterogenous Processes: Asset Pricing Applications", *Econometrica*, 57, 1091-1120.

[67] Gandhi, A., S. Navarro and D. Rivers (2015) "Which Moments to Match? " *Journal of Political Economy* forthcoming.

[68] Grasmair, M., O. Scherzer and A. Vanhems (2013) "Nonparametric Instrumental Regression with non-convex Constraints"*Inverse Problems* 29.

[69] Grenander, U. (1981) *Abstract Inference*, New York: Wiley Series.

[70] Groeneboom, P. and G. Jongbloed (2014) *Nonparametric Estimation under Shape Constraints*, Cambridge: Cambridge University Press. Series.

[71] Hahn, J. and G. Ridder (2013) "Asymptotic Variance of Semiparametric Estimators with Generated Regressors",*Econometrica*, 81, 315-340.

[72] Hall, P. and J. Horowitz (2005): "Nonparametric Methods for Inference in the Presence of Instrumental Variables", *Annals of Statistics*, 33, 2904-2929.

[73] Han, Q. and J. Wellner (2016): "Multivariate Convex Regression: Global Risk Bounds and Adaptation", arXiv preprint arXiv:1601.06844

[74] Hansen, L.P. (1982) "Large Sample Properties of Generalized Method of Moments Estimators",*Econometrica*, 50, 1029-1054.

[75] Hansen, L.P. (2014) "Nobel Lecture: Uncertainty Outside and Inside Economic Models",*Journal of Political Economy*, vol. 122, issue 5, 945-987.

[76] Hausman, J (1987) "Specification and Estimation of Simultaneous Equation Models", in Zvi Griliches and Michael D. Intriligator (eds.), *The Handbook of Econometrics*, vol. 1. Amsterdam: North-Holland.

[77] Heckman, J.J. (1979) "Sample Selection Bias as a Specification Error ",*Econometrica*, vol. 47, issue 1, 153-161.

[78] Heckman, J.J. and R. Robb (1985): "Alternative Methods for Evaluating the Impact of Interventions An Overview ", *Journal of Econometrics*, 30, 239-267.

[79] Heckman, J.J. and B, Singer (1984): "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data ", *Econometrica*, 68, 839-874.

[80] Hong, S. (2012) "Inference in Semiparametric Conditional Moment Models with Partial Identification, "Working Paper.

[81] Horowitz, J. (2006) "Testing a Parametric Model Against a Nonparametric Alterative with Identification through Instrumental Variables", *Econometrica*, 74 521-538.

[82] Horowitz, J. (2011) "Applied Nonparametric Instrumental Variables Estimation", *Econometrica*, 79, 347–394.

[83] Horowitz, J. (2013) "Ill-Posed Inverse Problems in Economics,"*Annual Review of Economics*, 6, 21-51

[84] Horowitz, J. (2014) "Adaptive Nonparametric Instrumental Variables Estimation: Empirical Choice of the Regularization Parameter,"*Journal of Econometrics*, 180, 158-173.

[85] Horowitz, J. and S. Lee (2007) "Nonparametric Instrumental Variables Estimation of a Quantile Regression Model", *Econometrica*, 75, 1191–1208.

[86] Horowitz, J. and S. Lee (2012) “Uniform Confidence Bands for Functions Estimated Nonparametrically with Instrumental Variables,” *Journal of Econometrics*, 168, 175-188.

[87] Ichimura, H. (1993) “Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single Index Models”, *Journal of Econometrics,* 58, 71-120.

[88] Imbens, G. (2002) “Generalized Method of Moments and Empirical Likelihood,” *Journal of Business and Economic Statistics,* 20(4), 493-506.

[89] Imbens, G. and W. Newey (2009) “Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity ”, *Econometrica*, 77, 1481-1512.

[90] Kato, K. (2013) “Quasi-Bayesian Analysis of Nonparametric Instrumental Variables Model, ”*Annals of Statistics*, 41, 2359-2390.

[91] Kawai, K. (2011) “Auction Design and the Incentives to Invest: Evidence from Procurement Auctions, ”Working Paper, NYU Stern.

[92] Kitamura, Y. (2007) “Empirical Lieklihood Methods in Econometrics: Theory and Practice,”*in Advances in Economics and Econometrics: Ninth World Congress of the Econometric Society* Cambridge University Press

[93] Lepskii, O. V. (1990) “On a Problem of Adaptive Estimation in Gaussian White Noise, ”*Theory of Probability and its Applications* 35(3), 454-466.

[94] Liao, Y. and W. Jiang (2011). “Posterior Consistency of Nonparametric Conditional Moment Restricted Models, ”*Annals of Statistics* 39 3003-3031.

[95] Mammen, E. (1993) “Bootstrap and Wild Bootstrap for High Dimensional Linear Models, ”*Annals of Statistics* 21(1), 255-285.

[96] Mammen, E., C. Rothe and M. Schienle (2012) “Nonparametric Regression with Nonparametrically Generated Covariates, ”*Annals of Statistics* 40, 1132-1170.

[97] Mammen, E., C. Rothe and M. Schienle (2016) “Semiparametric Estimation with Generated Covariates, ”*Econometric Theory* forthcoming.

[98] Matzkin, R.L. (1994) “Restrictions of Economic Theory in Nonparametric Methods”, in R.F. Engle III and D.F. McFadden (eds.), *The Handbook of Econometrics*, vol. 4. Amsterdam: North-Holland.

[99] Matzkin, R.L. (2007) “Nonparametric Identification”, Chapter 73 in J.J. Heckman and E.E. Leamer (eds.), *The Handbook of Econometrics*, vol. 6B. Amsterdam: North-Holland.

[100] Matzkin, R.L. (2013) “Nonparametric Identification in Structural Economic Models”, *Annual Review of Economics,* Vol 5.

[101] Merlo, A. and A. De Paula (2015) “Identification and Estimation of Voter Preferences”, *CeMMAP Working Paper 50/15.*

[102] Meyer, M. (2008) “Inference Using Shape-restricted Regression Splines”, *The Annals of Applied Statistics,* 2, 1013-1033.

[103] Newey, W.K. (1994) “The Asymptotic Variance of Semiparametric Estimators”, *Econometrica,* 62, 1349-1382.

[104] Newey, W.K. (1997) “Convergence Rates and Asymptotic Normality for Series Estimators”, *Journal of Econometrics,* 79, 147-168.

[105] Newey, W.K. and J.L Powell (2003) “Instrumental Variable Estimation of Nonparametric Models”, *Econometrica,* 71, 1565-1578. Working paper version, 1989.

[106] Newey, W.K., J.L. Powell and F. Vella (1999) “Nonparametric Estimation of Triangular Simultaneous Equations Models”, *Econometrica,* 67, 565-603.

[107] Nevo, A. (2011) “Empirical Models of Consumer Behavior, ” *Annual Review of Economics* 3:51-75.

[108] Olley, S. and A. Pakes (1996) “The Dynamics of Productivity in the Telecommunications Equipment Industry, ” *Econometrica,* 65, 1263-1297.

[109] Otsu, T. (2011) “Empirical Likelihood Estimation of Conditional Moment Restriction Models with Unknown Functions”, *Econometric Theory*, 27, 8-46.

[110] Pakes, A. and S. Olley (1995) “A Limit Theorem for A Smooth Class of Semiparametric Estimators,” *Journal of Econometrics*, 65, 295-332.

[111] Parente, P., and R. Smith (2014) “Recent Developments in Empirical Likelihood and Related Methods,” *Annual Review of Economics*, vol. 6(1), 77-102, 08.

[112] Pinkse, J., Slade, M. E. and Bret, C. (2002), “Spatial Price Competition: A Semiparametric Approach. ”, *Econometrica*, 70, 1111-1153.

[113] Pouzo, D. (2015) “Bootstrap Consistency for Quadratic Forms of Sample Averages with Increasing Dimension ”, *Electronic Journal of Statistics,* Forthcoming.

[114] Robinson, P. (1988) “Root-N-Consistent Semiparametric Regression”, *Econometrica,* 56, 931-954.

[115] Santos, A. (2012) “Inference in Nonparametric Instrumental Variables with Partial Identification”, *Econometrica*, 80 213-275.

[116] Smith, R. (1997) “Alternative Semi-Parametric Likelihood Approaches to Generalized Method of Moments Estimation ”, *Economic Journal,* 107, 509-519.

[117] Sueishi, N. (2014) “Efficient Estimation via Conditional Moment Restrictions Containing Unknown Functions ”, Working Paper, Kyoto University.

[118] Tao, J. (2015) “Inference for Point and Partially Identified Semi-Nonparametric Conditional Moment Models ”, Working Paper.

[119] Vanhems, A. (2010) “Non-parametric estimation of exact consumer surplus with endogeneity in price ”, *Econometrics Journal,* 13, S80-S98.

[120] Wang, J., S.K. Ghosh (2012) “Shape Restricted Nonparametric Regression With Berstein Polynomials ”, *Computational Statistics and Data Analysis,* 56, 2729-2741.

[121] Wooldridge, J. (2002) *Econometric Analysis of Cross Section and Panel Data*, MIT Press.

[122] Wooldridge, J. (2009) “On Estimating Firm-Level Production Functions Using Proxy Variables to Control for Unobservables ”, *Economics Letters,* 104, 112-114.

[123] Zhang, J. and I. Gijbels (2003) “Sieve Empirical Likelihood and Extensions of the Generalized Least Squares”, *Scandinavian Journal of Statistics*, 30, 1-24.