npiv: Estimation and Confidence Bands for Nonparametric Instrumental Variables Models Using R

Timothy Christensen 🆔
University College London

Xiaohong Chen 🆔
Yale University

Jeffrey S. Racine 🆔
McMaster University

Abstract

Nonparametric instrumental variables (NPIV) methods are widely used in economics, causal inference, and machine learning. This article discusses the R package **npiv**, which implements novel, fully data-driven NPIV estimation and inference methods proposed in Chen, Christensen, and Kankanala (2024). The main methods implemented in this package are (i) data-driven, rate-optimal choices of sieve dimension for sieve NPIV estimators and (ii) data-driven uniform confidence bands for the structural function and its derivatives. Additional functionality allows for constructing (non-data driven) uniform confidence bands based on the methods proposed in Chen and Christensen (2018). All methods subsume nonparametric regression as a special case.

Keywords: nonparametric instrumental variables, nonparametric regression, minimax rate-adaptive estimation, adaptive uniform confidence bands, R.

1. Introduction

Nonparametric instrumental variables (NPIV) methods are well-established, widely used, and have been adopted by practitioners in a range of empirical fields. While there is a long tradition of their use in demand estimation in economics (Blundell, Chen, and Kristensen 2007; Blundell, Horowitz, and Parey 2017; Berry and Haile 2014), NPIV methods have recently been adopted in causal inference in biostatistics (Miao, Geng, and Tchetgen Tchetgen 2018), reinforcement learning (Chen, Xu, Gulcehre, Paine, Gretton, De Freitas, and Doucet 2022), and causal machine learning (Kaddour, Lynch, Liu, Kusner, and Silva 2022). In this paper, we introduce the R package **npiv**. The package makes a novel set of estimation and inference methods for NPIV models, including with data-driven choice of tuning parameters, newly accessible to R users.

We are interested in estimating a nonparametric *structural function*, denoted h_0, linking a scalar outcome variable Y and a vector of regressors X through the relation

$$Y = h_0(X) + u, \tag{1}$$

where u is a disturbance term. The function h_0 is not necessarily the conditional mean of Y given X (as in nonparametric regression) but rather has a particular *structural* or *causal* interpretation. Equivalently, X is *endogenous* in the sense that $\mathsf{E}[u|X]$ is not necessarily zero.[1] NPIV methods allow for the estimation of h_0 using a vector of *instrumental variables* W for which $\mathsf{E}[u|W] = 0$. For instance, in a demand estimation context the variable Y denotes the quantity demanded, X denotes the price, and W is a "cost-shifter" that causes variation in the marginal cost of production. Derivatives of h_0 may also be of interest as they have important "structural" interpretations as elasticities or other marginal effects.

This article discusses the R package **npiv**, which implements novel, fully data-driven NPIV estimation and inference methods proposed in Chen *et al.* (2024). The methods are developed in the context of sieve NPIV estimators, which are implemented in a very simple manner using two-stage least-squares (TSLS) (Ai and Chen 2003; Newey and Powell 2003). The main methods implemented in **npiv** are (i) a data-driven choice of sieve dimension for sieve NPIV estimators and (ii) data-driven uniform confidence bands (UCBs) for the structural function h_0 and its derivatives. As shown in Chen *et al.* (2024), the single data-driven choice of sieve dimension is minimax sup-norm rate adaptive for estimating both h_0 and derivatives of h_0. Moreover, the data-driven *uniform* confidence bands contain the entire true structural function and its derivatives with desired asymptotic coverage probability and contract at the optimal rate (up to log terms). Additional functionality allows for constructing (non-data driven) UCBs based on the methods proposed in Chen and Christensen (2018) as well as *pointwise* confidence intervals for $h_0(x)$ or its derivatives at a point x based on the method of Chen and Pouzo (2015). All methods subsume nonparametric regression as a special case.

The remainder of this article is organized as follows. Section 2 reviews sieve NPIV estimators and describes the data-driven methods from Chen *et al.* (2024) and the UCB constructions from Chen and Christensen (2018). Section 3 gives a description of the main functionality of the package **npiv** and describes its implementation. Section 4 then illustrates the methods in the context of Engel curve estimation based on household consumption data.

2. Method

2.1. Sieve NPIV estimators

The exposition in this section closely follows Chen *et al.* (2024). We being by briefly describing sieve NPIV estimators. Consider approximating an unknown structural function h_0 using a linear combination of J basis functions, denoted $\psi_{J1}, \ldots, \psi_{JJ}$:

$$h_0(x) \approx \sum_{j=1}^{J} c_{Jj} \psi_{Jj}(x) = (\psi^J(x))' c_J, \tag{2}$$

---

[1]To simplify exposition, we shall omit the "almost surely" qualifier when specifying conditional moment restrictions.

where $\psi^J(x) = (\psi_{J1}(x), \dots, \psi_{JJ}(x))^\top$ is a vector of basis functions and $c_J = (c_{J1}, \dots, c_{JJ})^\top$ is a vector of coefficients. Combining (1) and (2), we obtain

$$Y = (\psi^J(X))^\top c_J + \text{bias}_J + u\,,$$

where $\text{bias}_J = h_0(X) - (\psi^J(X))^\top c_J$ and $u = Y - h_0(X)$. Provided the bias term is "small" relative to u, we have an approximate linear IV model where $\psi^J(X)$ is a $J \times 1$ vector of "endogenous variables" and c_J is a vector of unknown "parameters". One can then estimate c_J using TSLS using a vector of $K \geq J$ transformations $b^K(W) = (b_{K1}(W), \dots, b_{KK}(W))^\top$ of W as instruments.

Given data $(X_i, Y_i, W_i)_{i=1}^n$, the TSLS estimator of c_J is simply

$$\hat{c}_J = \left(\boldsymbol{\Psi}_J^\top \mathbf{P}_K \boldsymbol{\Psi}_J\right)^- \boldsymbol{\Psi}_J^\top \mathbf{P}_K \mathbf{Y}\,,$$

where $\boldsymbol{\Psi}_J = (\psi^J(X_1), \dots, \psi^J(X_n))^\top$ and $\mathbf{B}_K = (b^K(W_1), \dots, b^K(W_n))^\top$ are $n \times J$ and $n \times K$ matrices, $\mathbf{P}_K = \mathbf{B}_K (\mathbf{B}_K^\top \mathbf{B}_K)^- \mathbf{B}_K^\top$ is a $K \times K$ matrix,[2] and $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ is an $n \times 1$ vector. Let

$$\partial^a h(x) = \frac{\partial^{|a|} h(x)}{\partial^{a_1} x_1 \dots \partial^{a_d} x_d}\,,$$

for $a = (a_1, \dots, a_d) \in (\mathbb{N}_0)^d$ with $|a| = a_1 + \dots + a_d$. Estimators of h_0 and its derivative $\partial^a h_0$ are given by

$$\hat{h}_J(x) = (\psi^J(x))^\top \hat{c}_J\,, \qquad \partial^a \hat{h}_J(x) = (\partial^a \psi^J(x))^\top \hat{c}_J\,,$$

where $\partial^a \psi^J(x) = (\partial^a \psi_{J1}(x), \dots, \partial^a \psi_{JJ}(x))^\top$. The estimator $\hat{h}_J$ collapses to a nonparametric series regression estimator

$$\hat{h}_J(x) = \psi^J(x)^\top (\boldsymbol{\Psi}_J^\top \boldsymbol{\Psi}_J)^- \boldsymbol{\Psi}_J^\top \mathbf{Y}$$

in the special case in which $W = X$, $K = J$, and $(b_{K1}, \dots, b_{KK}) = (\psi_{J1}, \dots, \psi_{JJ})$. The data-driven methods for choosing J and for constructing UCBs for h_0 and its derivatives therefore carry over to nonparametric regression as a special case.

In practice, a researcher must choose bases $\psi_{J1}, \dots, \psi_{JJ}$ and $b_{K1}, \dots, b_{KK}$ and the tuning parameters J and K. Chen and Christensen (2018) showed that B-spline and Cohen–Daubiches–Vial wavelet bases lead to estimators of h_0 and its derivatives that converge at the minimax sup-norm rate under an appropriate choice of J and K. These bases also lead to estimators that converge at the minimax rate in the special case of nonparametric regression (Belloni, Chernozhukov, Chetverikov, and Kato 2015; Chen and Christensen 2015). We therefore use B-spline bases in what follows as these are easier to compute than wavelet bases. Bases for multivariate X or W are generally constructed by taking tensor products of univariate bases.[3]

2.2. Data-driven choice of sieve dimension

Chen *et al.* (2024) introduce a data-driven choice of sieve dimension that leads to estimators of both h_0 and its derivatives that converge at the optimal sup-norm rate. The choice of K is pinned down by J in their procedure, so we write $K(J)$, $b^{K(J)}(W)$, $\mathbf{B}_{K(J)}$ and $\mathbf{P}_{K(J)}$ in what follows.

---

[2] We let A^- denote the generalized (or Moore–Penrose) inverse of a matrix A.
[3] The exception is when X is multivariate and we wish to impose an additively separable structure on h_0.

We first introduce some notation. Let $\mathcal{T} \subset \mathbb{N}_0$ denote the set of candidate values of J. This set depends on the choice of basis. If X is d-dimensional, then $\mathcal{T} = \{(2^l + r)^d : l \in \mathbb{N}_0\}$ for a tensor-product basis of B-splines of order r. For example, with univariate cubic B-splines ($r = 4$) we have $\mathcal{T} = \{4, 5, 7, 11, ...\}$. Let $J^+ = \min\{j \in \mathcal{T} : j > J\}$ be the smallest sieve dimension in $\mathcal{T}$ exceeding J. Let $\mathbf{M}_J = (\mathbf{\Psi}_J^\top \mathbf{P}_{K(J)} \mathbf{\Psi}_J)^- \mathbf{\Psi}_J^\top \mathbf{P}_{K(J)}$. For $J, J_2 \in \mathcal{T}$ with $J_2 > J$, define the contrast

$$D_J(x) - D_{J_2}(x) = (\psi^J(x))^\top \mathbf{M}_J \hat{\mathbf{u}}_J - (\psi^{J_2}(x))^\top \mathbf{M}_{J_2} \hat{\mathbf{u}}_{J_2}.$$

Its variance is estimated by

$$\hat{\sigma}_{J,J_2}^2(x) := \hat{\sigma}_J^2(x) + \hat{\sigma}_{J_2}^2(x) - 2\tilde{\sigma}_{J,J_2}(x),$$

where

$$\hat{\sigma}_J^2(x) = (\psi^J(x))^\top \mathbf{M}_J \widehat{\mathbf{U}}_{J,J} \mathbf{M}_J^\top \psi^J(x)$$

and

$$\tilde{\sigma}_{J,J_2}(x) = (\psi^J(x))^\top \mathbf{M}_J \widehat{\mathbf{U}}_{J,J_2} \mathbf{M}_{J_2}^\top \psi^{J_2}(x),$$

with $\widehat{\mathbf{U}}_{J,J_2}$ an $n \times n$ diagonal matrix whose ith diagonal entry is $\hat{u}_{i,J}\hat{u}_{i,J_2}$ for $\hat{u}_{i,J} = Y_i - \hat{h}_J(X_i)$. A bootstrap version of the contrast is given by

$$D_J^*(x) - D_{J_2}^*(x) = (\psi^J(x))^\top \mathbf{M}_J \hat{\mathbf{u}}_J^* - (\psi^{J_2}(x))^\top \mathbf{M}_{J_2} \hat{\mathbf{u}}_{J_2}^*,$$

with $\hat{\mathbf{u}}_J^* = (\hat{u}_{1,J}\varpi_1, \ldots, \hat{u}_{n,J}\varpi_n)^\top$ denoting a multiplier bootstrap version of $\hat{\mathbf{u}}_J$, where $(\varpi_i)_{i=1}^n$ are IID $N(0,1)$ draws independent of the data. Finally, let $\hat{s}_J$ be the smallest singular value of $(\mathbf{B}_{K(J)}^\top \mathbf{B}_{K(J)})^{-1/2} (\mathbf{B}_{K(J)}^\top \mathbf{\Psi}_J)(\mathbf{\Psi}_J^\top \mathbf{\Psi}_J)^{-1/2}$.

The data-driven procedure from Chen *et al.* (2024) for choosing J in NPIV models is summarized in the following three steps:

1. Compute

$$\hat{J}_{\max} = \min\left\{J \in \mathcal{T} : J\sqrt{\log J}\,\hat{s}_J^{-1} \le 10\sqrt{n} < J^+\sqrt{\log J^+}\,\hat{s}_{J^+}^{-1}\right\}, \tag{3}$$

$$\hat{\mathcal{J}} = \left\{J \in \mathcal{T} : 0.1(\log \hat{J}_{\max})^2 \le J \le \hat{J}_{\max}\right\}. \tag{4}$$

2. Let $\hat{\alpha} = \min\{0.5, (\log(\hat{J}_{\max})/\hat{J}_{\max})^{1/2}\}$. For each independent draw of $(\varpi_i)_{i=1}^n$, compute

$$\sup_{\{(x,J,J_2)\in\mathcal{X}\times\hat{\mathcal{J}}\times\hat{\mathcal{J}}:J_2>J\}} \left|\frac{D_J^*(x) - D_{J_2}^*(x)}{\hat{\sigma}_{J,J_2}(x)}\right|. \tag{5}$$

Let $\theta_{1-\hat{\alpha}}^*$ denote the $(1 - \hat{\alpha})$ quantile of the sup statistic (5) across a large number of independent draws of $(\varpi_i)_{i=1}^n$ (conditional on the data).

3. Let $\hat{J}_n = \max\{J \in \hat{\mathcal{J}} : J < \hat{J}_{\max}\}$ and

$$\hat{J} = \min\left\{J \in \hat{\mathcal{J}} : \sup_{(x,J_2)\in\mathcal{X}\times\hat{\mathcal{J}}:J_2>J} \left|\frac{D_J(x) - D_{J_2}(x)}{\hat{\sigma}_{J,J_2}(x)}\right| \le 1.1\theta_{1-\hat{\alpha}}^*\right\}. \tag{6}$$

The data-driven choice of sieve dimension is

$$\tilde{J} = \min\{\hat{J}, \hat{J}_n\}. \tag{7}$$

Chen *et al.* (2024) show that the single data-driven choice $\tilde{J}$ leads to estimators $\hat{h}_{\tilde{J}}$ and $\partial^a \hat{h}_{\tilde{J}}$ of h_0 and $\partial^a h_0$ that converge at the minimax sup-norm rate for NPIV estimation. They also show that these adaptivity guarantees carry over to nonparametric regression models with $W = X$, $K = J$, and $(b_{K1}, \ldots, b_{KK}) = (\psi_{J1}, \ldots, \psi_{JJ})$. In that case, the procedure is implemented largely as above, but with the following slight modifications:

1.$'$ Compute

$$\hat{J}_{\max} = \min \left\{ J \in \mathcal{T} : J\sqrt{\log J}\, v_n \leq 10\sqrt{n} < J^+ \sqrt{\log J^+}\, v_n \right\} \qquad (8)$$

with $v_n = \max\{1, (0.1 \log n)^4\}$, then compute $\hat{\mathcal{J}}$ as in (4) with this choice of $\hat{J}_{\max}$.

2.$'$ As above.

3.$'$ Take $\tilde{J} = \hat{J}$ for $\hat{J}$ defined in (6).

2.3. Data-driven uniform confidence bands

We now describe the method for constructing UCBs for h_0 and its derivatives proposed in Chen *et al.* (2024). The approach builds on the data-driven choice of J introduced in the previous subsection. Recall $\hat{\mathcal{J}}$ and $\tilde{J}$ from above. Let $\hat{A} = \log\log \tilde{J}$ and

$$\hat{\mathcal{J}}_- = \begin{cases} \{J \in \hat{\mathcal{J}} : J < \hat{J}_n\} & \text{if } \tilde{J} = \hat{J}, \\ \hat{\mathcal{J}} & \text{if } \tilde{J} = \hat{J}_n. \end{cases}$$

UCBs for the structural function h_0 in NPIV models are constructed in the following steps:

4. For each independent draw of $(\varpi_i)_{i=1}^n$, compute

$$\sup_{(x,J) \in \mathcal{X} \times \hat{\mathcal{J}}_-} \left| \frac{D_J^*(x)}{\hat{\sigma}_J(x)} \right|. \qquad (9)$$

   Let $z_{1-\alpha}^*$ denote the $(1 - \alpha)$ quantile of the sup statistic (9) across a large number of independent draws of $(\varpi_i)_{i=1}^n$ (conditional on the data).

5. Construct the $100(1 - \alpha)\%$ UCB

$$C_n(x) = \left[\hat{h}_{\tilde{J}}(x) - \left(z_{1-\alpha}^* + \hat{A}\theta_{1-\hat{\alpha}}^* \right) \hat{\sigma}_{\tilde{J}}(x), \ \hat{h}_{\tilde{J}}(x) + \left(z_{1-\alpha}^* + \hat{A}\theta_{1-\hat{\alpha}}^* \right) \hat{\sigma}_{\tilde{J}}(x) \right]. \qquad (10)$$

Let

$$D_J^{a*}(x) = (\partial^a \psi^J(x))' \mathbf{M}_J \hat{\mathbf{u}}_J^*,$$

and

$$\hat{\sigma}_J^{a2}(x) = (\partial^a \psi^J(x))' \mathbf{M}_J \hat{\mathbf{U}}_{J,J} \mathbf{M}_J'(\partial^a \psi^J(x)).$$

UCBs for $\partial^a h_0$ in NPIV models are constructed analogously:

4.′ For each independent draw of $(\varpi_i)_{i=1}^n$, compute

$$\sup_{(x,J)\in\mathcal{X}\times\hat{\mathcal{J}}_-}\left|\frac{D_J^{a*}(x)}{\hat{\sigma}_J^a(x)}\right|. \tag{11}$$

Let $z_{1-\alpha}^{a*}$ denote the $(1-\alpha)$ quantile of sup statistic (11) across a large number of independent draws of $(\varpi_i)_{i=1}^n$ (conditional on the data).

5.′ Construct the $100(1-\alpha)\%$ UCB

$$C_n^a(x) = \left[\partial^a\hat{h}_{\bar{J}}(x) - \left(z_{1-\alpha}^{a*} + \hat{A}\theta_{1-\hat{\alpha}}^*\right)\hat{\sigma}_{\bar{J}}^a(x), \ \partial^a\hat{h}_{\bar{J}}(x) + \left(z_{1-\alpha}^{a*} + \hat{A}\theta_{1-\hat{\alpha}}^*\right)\hat{\sigma}_{\bar{J}}^a(x)\right]. \tag{12}$$

The above UCB constructions apply equally to nonparametric regression models in the special case in which $W = X$, $K = J$, and $(b_{K1},\ldots,b_{KK}) = (\psi_{J1},\ldots,\psi_{JJ})$.

Chen *et al.* (2024) establish uniform asymptotic coverage guarantees for these UCBs for both NPIV and nonparametric regression. They also show that the UCBs contract at the optimal sup-norm rate (up to logarithmic factors).

2.4. Uniform confidence bands based on undersmoothing

An alternative UCB construction based on a deterministic (i.e., non-data-driven) choice of sieve dimension is proposed by Chen and Christensen (2018) for NPIV models. The construction also applies to nonparametric regression in the special case in which $W = X$, $K = J$, and $(b_{K1},\ldots,b_{KK}) = (\psi_{J1},\ldots,\psi_{JJ})$. The theoretical justification for this approach is based on *undersmoothing*. That is, the sieve dimension J should be larger than what would be optimal for estimation, so that bias is of smaller order than sampling uncertainty. Note that this approach requires a user-supplied choice of sieve dimension J.

We now briefly summarize the Chen and Christensen (2018) procedure using the notation introduced above. UCBs for h_0 and $\partial^a h_0$ are constructed as follows:

1. For each independent draw of $(\varpi_i)_{i=1}^n$, compute

$$\sup_{x\in\mathcal{X}}\left|\frac{D_J^*(x)}{\hat{\sigma}_J(x)}\right| \ \text{ for } h_0, \quad \text{ or } \quad \sup_{x\in\mathcal{X}}\left|\frac{D_J^{a*}(x)}{\hat{\sigma}_J^a(x)}\right| \ \text{ for } \partial^a h_0. \tag{13}$$

Let $z_{1-\alpha,J}^*$ and $z_{1-\alpha,J}^{a*}$ denote the $(1-\alpha)$ quantile of these sup statistics across a large number of independent draws of $(\varpi_i)_{i=1}^n$ (conditional on the data).

2. Construct the $100(1-\alpha)\%$ UCBs

$$C_{n,J}(x) = \left[\hat{h}_J(x) - z_{1-\alpha,J}^*\hat{\sigma}_J(x), \ \hat{h}_J(x) + z_{1-\alpha,J}^*\hat{\sigma}_J(x)\right] \text{ for } h_0,$$

or

$$C_{n,J}^a(x) = \left[\partial^a\hat{h}_J(x) - z_{1-\alpha,J}^{a*}\hat{\sigma}_J^a(x), \ \partial^a\hat{h}_J(x) + z_{1-\alpha,J}^{a*}\hat{\sigma}_J^a(x)\right] \text{ for } \partial^a h_0.$$

2.5. Pointwise confidence intervals based on undersmoothing

Finally, we note that it may sometimes be of interest to construct confidence intervals (CIs) for $h_0(x)$ or $\partial^a h_0(x)$ at a certain point x. To this end, one may use the following constructions for $100(1 - \alpha)\%$ CIs:

$$\left[\hat{h}_J(x) - z_{1-\alpha/2}\hat{\sigma}_J(x), \ \hat{h}_J(x) + z_{1-\alpha/2}\hat{\sigma}_J(x) \right] \text{ for } h_0(x),$$

or

$$\left[\partial^a \hat{h}_J(x) - z_{1-\alpha/2}\hat{\sigma}_J^a(x), \ \partial^a \hat{h}_J(x) + z_{1-\alpha/2}\hat{\sigma}_J^a(x) \right] \text{ for } \partial^a h_0(x),$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the $N(0,1)$ distribution; see Chen and Pouzo (2015) and Appendix D of Chen and Christensen (2018) for a formal justification for these constructions. These CIs can be (and sometimes are) plotted alongside estimates of h_0 or $\partial^a h_0$, but it should be understood that they are only valid *pointwise*. That is, they will generally be too narrow to contain the entire function h_0 or $\partial^a h_0$ with desired asymptotic coverage probability. By contrast, the above UCB constructions will contain the true structural function or its derivatives with desired asymptotic coverage probability.

3. Package description

The contributed R package **npiv** is used to estimate and construct UCBs for a nonparametric structural function h_0 and its derivatives $\partial^a h_0$ in both NPIV and nonparametric regression models. In this section, we describe the main functionality of this package. The current version of the package can be downloaded from GitHub:

```
library(devtools); install_github('JeffreyRacine/npiv')
```

3.1. Use in NPIV models

The main function is `npiv()`, which can be called with the following syntax:

```
npiv(formula, data, newdata=NULL, basis=c("tensor","additive","glp"),
    J.x.degree=3, J.x.segments=NULL,
    K.w.degree=4, K.w.segments=NULL, K.w.smooth=2, ...)
```

The required arguments include

- `formula`: a symbolic formula specifying the model to be estimated. The syntax is the same as the package **ivreg** (Fox, Kleiber, Zeileis, and Kuschnig 2021). For example, `y ~ x1 + x2 | x1 + z1 + z2` where `y` is the dependent variable, `x1` is an exogenous regressor, `x2` is an endogenous regressor, and `z1` and `z2` are instrumental variables.

- `data`: an optional data frame containing the variables in the model.

- `newdata`: optional data frame collecting the values of `x` variables at which to evaluate the estimator (e.g. for plotting). Will evaluate at sample data points if `NULL`.

- `basis`: type of basis to use if `x` variables are multivariate. Default is `tensor` for a tensor-product basis. Can also use `additive` for an additively-separable structural function or `glp` for generalized B-spline polynomial bases (for further details, see documentation for the package **crs** (Racine, Nie, and Ripley 2022)).

- `J.x.degree`: degree of the B-spline used to approximate the structural function. Default is 3 (cubic spline).

- `J.x.segments`: number of segments (interior knots plus 1) for the B-spline used to approximate the structural function. Note that integers for both `J.x.segments` and `K.w.segments` must be passed to estimate the model with a deterministic J (and K) and return undersmoothed UCBs. Te default is `NULL`, in which case both `J.x.segments` and `K.w.segements` are data-determined using the procedure described above and data-driven UCBs are returned.

- `K.w.degree`: degree of the B-spline used to approximate the nonparametric first stage. Default is 4 (quartic spline).

- `K.w.segments`: number of segments (interior knots plus 1) for the B-spline used to approximate the structural function. Note that integers for both `J.x.segments` and `K.w.segments` must be passed to estimate the model with a deterministic J (and K) and return undersmoothed UCBs. The default is `NULL`, in which case both `J.x.segments` and `K.w.segements` are data-determined using the procedure described above and data-driven UCBs are returned.

- `K.w.smooth`: a non-negative integer. The basis for the nonparametric first-stage uses `2^{K.w.smooth}` more B-spline segments for each instrument than the basis approximating the structural function. Default is 2. Setting `K.w.smooth=0` uses the same number of segments for `x` and `w`.

The output of the function `npiv()` is a 'npiv' object. The `summary()` method prints a brief summary of the estimated model. The `npiv` object includes the following components:

- `h`: the estimated structural function evaluated at the sample data (or evaluation data, if provided).

- `h.upper`: the upper UCB for the structural function evaluated at the sample data (or evaluation data, if provided).

- `h.lower`: the lower UCB for the structural function evaluated at the sample data (or evaluation data, if provided).

- `deriv`: estimated derivative of the structural function evaluated at the sample data (or evaluation data, if provided).

- `h.upper.deriv`: the upper UCB for the derivative of the structural function evaluated at the sample data (or evaluation data, if provided).

- `h.lower.deriv`: the lower UCB for the derivative of the structural function evaluated at the sample data (or evaluation data, if provided).

- `asy.se`: pre-asymptotic standard errors for the estimator of the structural function evaluated at the sample data (or evaluation data, if provided). These may be used to construct (pointwise) confidence intervals for $h_0(x)$ based on undersmoothing when a deterministic sieve dimension is used.

- `deriv.asy.se`: pre-asymptotic standard errors for the estimator of the derivative of the structural function evaluated at the sample data (or evaluation data, if provided). These may be used to construct (pointwise) confidence intervals for $\partial^a h_0(x)$ based on undersmoothing when a deterministic sieve dimension is used.

- `K.w.segments`: value of `K.w.segments` used (will be data-determined if not provided).

- `J.x.segments`: value of `J.x.segments` used (will be data-determined if not provided).

- `beta`: vector of estimated spline coefficients $\hat{c}_J$.

3.2. Use in nonparametric regression models

The command `npiv()` may also be used for data-driven choice of sieve dimension and UCB construction in nonparametric regression models. In this case, `formula` should pass the same explanatory variables as instrumental variables (and in the same order). Calling

```
npiv(y ~ x1 + x2 | x1 + x2)
```

will perform nonparametric regression of `y` on `x1` and `x2`, with a data-driven choice of sieve dimension. Note that this choice of sieve dimension will be rate-optimal for estimation of h_0 and its derivatives under sup-norm loss. This choice may differ from the choice of sieve dimension selected by other methods that target mean-square error (i.e., L^2-norm loss), such as cross validation, because the optimal choices of sieve dimension for estimation under sup-norm and L^2-norm losses diverge at different rates. As before, data-driven UCBs for the conditional mean function and its derivative will also be computed by default using the procedure of Chen *et al.* (2024).

Estimators based on a deterministic choice of sieve dimension and undersmoothed UCBs can be computed by passing `J.x.segments`. Both `K.w.segments` and `K.w.smooth` are redundant and do not need to be specified. Undersmoothed UCBs are constructed using the procedure of Chen and Christensen (2018); see also Belloni *et al.* (2015). Standard errors `asy.se` and `deriv.asy.se` may also be used to construct (pointwise) confidence intervals for $h_0(x)$ and $\partial^a h_0(x)$ based on undersmoothing, using the method described in Section 2.5.

3.3. Additional options

By default, `npiv` will perform estimation of the structural function h_0 and construct 95% UCBs for h_0 and $\frac{dh_0}{dx}$ if x is scalar or

$$\frac{\partial h_0(x_1, \ldots, x_d)}{\partial x_1}$$

if x is multivariate. The following optional arguments can be used to modify the UCB constructions:

- `alpha`: nominal size of the uniform confidence bands. Default is 0.05 for 95% uniform confidence bands.

- `deriv.index`: integer indicating for which regressor to compute the derivative. Default is 1 for differentiation with respect to x_1.

- `deriv.order`: integer indicating the order of derivative to be computed. Default is 1.

- `ucb.h`: whether to compute a uniform confidence band for the structural function. Default is `TRUE`.

- `ucb.deriv`: whether to compute a uniform confidence band for the derivative of the structural function. Default is `TRUE`.

Run time can be improved by passing `ucb.h = FALSE` or `ucb.deriv = FALSE` when a UCB for h_0 or its derivative are not required.

4. Engel curve estimation

We now illustrate the usage of `npiv` with an empirical application to nonparametric Engel curve estimation (Blundell *et al.* 2007). Engel curves explain how consumers' expenditure shares on different categories of goods and services vary as a function of total household expenditure. Understanding consumers' consumption and substitution patterns across expenditure categories is important for conducting policy analysis and for modeling demand. Engel curve estimation is therefore the subject of a large amount of empirical work in economics. There is also a long tradition of estimating Engel curves via "flexible" (i.e., nonparametric or semi-parametric) methods which recognize correctly that consumers' expenditure shares vary nonlinearly with total expenditure.[4]

We wish to estimate the Engel curve h_0 for a household's expenditure share on a particular category, say food. The model is

$$Y_i = h_0(X_i) + u_i$$

where Y_i is expenditure share for household i on food and X_i is log total household expenditure. The "structural" interpretation of the model is that the household would spend the

---

[4]Interestingly, the original empirical exercise performed by Engel over 150 years ago used precursors of nonparametric statistical methods (Chai and Moneta 2010).

fraction $h_0(x)$ of its total expenditure on food if log total expenditure was exogenously specified as x. As is well known (see, e.g., Banks, Blundell, and Lewbel (1997)), total household expenditure is endogenous:

$$\mathsf{E}[Y_i|X_i] \neq h_0(X_i),$$

which follows from the fact that the household's consumption preferences determines its expenditure across categories and therefore its total expenditure. To deal with this endogeneity, we shall follow Banks *et al.* (1997), Blundell *et al.* (2007), and several other works and use log gross earnings of the household head (W_i, hereafter log wages) as an instrument for log total expenditure.

We use data from the 1995 British Family Expenditure Survey as in Blundell *et al.* (2007). The data consist of 1655 married or cohabitating couples with the head of household employed and aged between 20 and 55 and with zero, one, or two children. The data are available as part of **npiv** in the data frame `Engel95`.

```
R> data("Engel95", package = "npiv")
R> head(Engel95)
```

```
         food   catering   alcohol       fuel       motor       fares    leisure
1 0.1494936 0.10459980 0.00000000 0.07554642 0.19400656 0.03446044 0.16236387
2 0.3646063 0.03262203 0.01532990 0.09762080 0.19944200 0.03679176 0.13520971
3 0.2104212 0.06524783 0.07918546 0.03367232 0.25255781 0.01002152 0.17813644
4 0.0786906 0.11526845 0.07013521 0.05841701 0.00000000 0.24973512 0.02726338
5 0.2824394 0.05508218 0.07504724 0.15707207 0.21778214 0.00000000 0.03753635
6 0.1663630 0.15928116 0.17427441 0.02050462 0.09070033 0.00000000 0.14901404
     logexp logwages nkids
1 4.877561 5.533113     0
2 5.094241 5.371289     0
3 5.781675 6.001043     0
4 5.756296 5.860758     0
5 5.279863 6.569341     0
6 5.821521 5.879247     0
```

The first seven columns of `Engel95` list household budget shares across different expenditure categories. The variables `logexp` and `logwages` are log total household expenditure and log wages The final variable `nkids` indicates whether the household has zero children (if `nkids = 0`; 628 households in total), or 1-2 two children (if `nkids = 1`; 1027 households in total).

We first estimate the Engel curve for food. To try to ensure a relatively homogeneous set of households, we focus on the subset of households with 1-2 children.

```
R> Engel.kids <- subset(Engel95, nkids == 1)
R> attach(Engel.kids)
```

To use a data-driven choice of sieve dimension and construct data-driven UCBs for h_0 and its derivative, we first call `npiv()` without specifying the sieve dimensions. We also pass a grid of points spanning $[4.75, 6.25]$ for plotting the estimator and confidence bands:

```
R> newdata <- data.frame(logexp = seq(4.75, 6.25, length.out = 1000))
R> fm.food.dd <- npiv(food ~ logexp | logwages, newdata = newdata)
R> summary(fm.food.dd)


Call:
npiv.formula(formula = food ~ logexp | logwages, newdata = newdata)

Nonparametric IV Model:

IV Regression Data: 1027 training points, and 1000 evaluation points, in 1 endogenous vari

B-spline degree for instruments:            4
B-spline segments for instruments:          4

B-spline degree for endogenous predictors:   3
B-spline segments for endogenous predictors: 1

Estimation time: 2.2 seconds
```

As X and W are both scalar and we use the default cubic B-spline basis for X and quartic B-spline basis for W, the data-driven choice of sieve dimension are therefore $J = 4$ and $K = 8$ (the sum of segments and degree for each basis). The run time represents the total time to perform data-driven choice of sieve dimension and to construct the data-driven UCBs for h_0 and its derivative.

To compare data-driven and undersmoothed UCBs, we estimate h_0 and its derivative again, this time using a deterministic choice of J and K by passing both `J.x.segments` and `K.w.segments`. As the justification for the UCBs is undersmoothing, we must choose a J and K larger than the optimal value so that bias is of smaller order than sampling uncertainty. We choose to estimate with `J.x.segments = 2` and `K.w.segments = 5`, which corresponds to $J = 5$ and $K = 9$.

```
R> fm.food.det <- npiv(food ~ logexp | logwages, newdata = newdata,
+                      J.x.segments = 2, K.w.segments = 5)
```

The estimated Engel curves for food are plotted below, together with their UCBs (using `h.upper` and `h.lower`). Both the data-driven and undersmoothed estimates are downward-sloping. This seems reasonable from an economic standpoint: food is a necessary good, and so it should decrease as a proportion of total household expenditure as total household expenditure increases. Note, however, that the data-driven estimate looks close to linear whereas the undersmoothed estimate is more wiggly. The undersmoothed UCBs are also wider over much of the support of total expenditure because they use a sub-optimally large choice of sieve dimension and are therefore less "efficient" than the data-driven bands. We also plot 95% pointwise confidence intervals (PCIs) based on undersmoothing (using `asy.se` and a $N(0, 1)$ critical value). As expected, these are narrower than the UCBs based on undersmoothing as they are only required to contain the true function at each point, rather than over the entirety of its support. So while PCIs can be used for drawing inference about
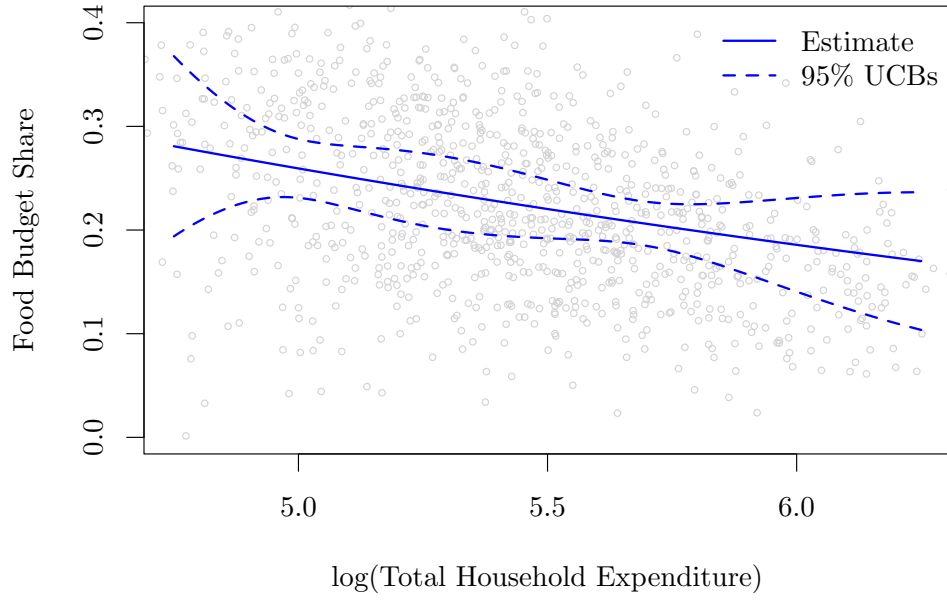
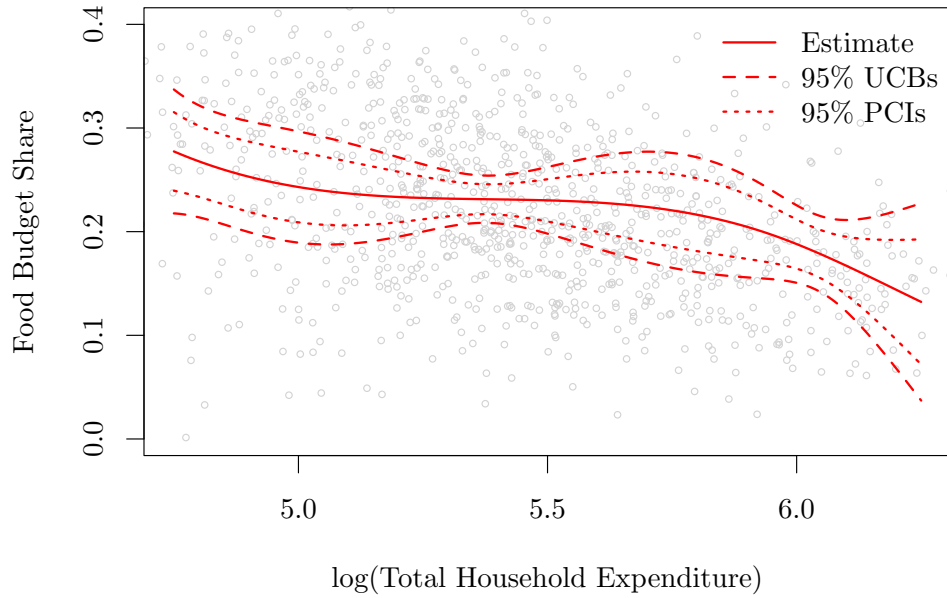Figure 1: Engel Curve for Food: Data-driven Estimates and UCBs

Figure 2: Engel Curve for Food: Estimates, UCBs, and PCIs based on Undersmoothing
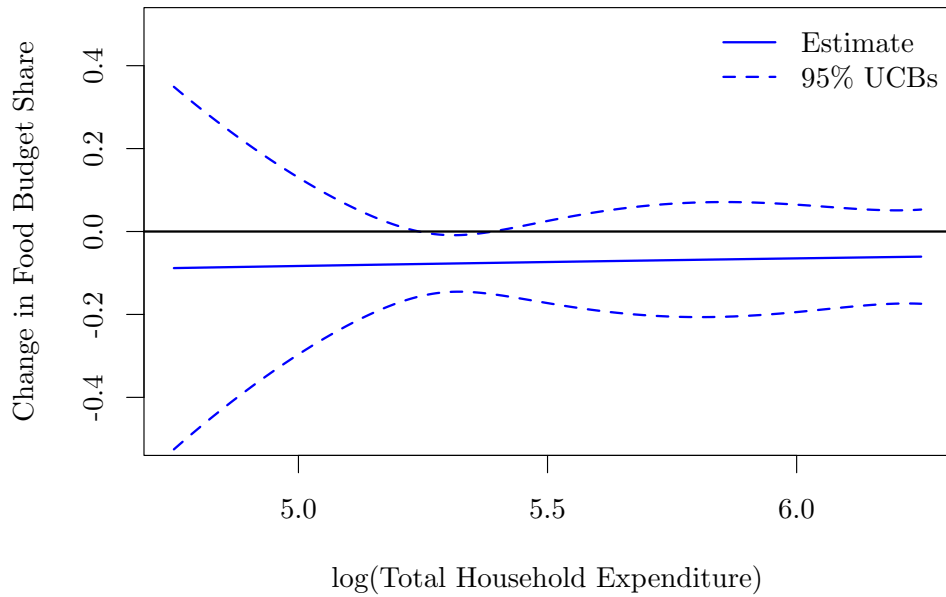
Figure 3: Engel Curve Derivative for Food: Data-driven Estimates and UCBs

the value of the function at a given point, they cannot be used for inferring the true shape of the function.

These differences between the data-driven and undersmoothed methods are also apparent when we compare the estimated derivatives of the Engel curves and their corresponding UCBs (using `h.lower.deriv` and `h.upper.deriv`). The zero function lies significantly above the data-driven UCBs over a sub-interval of total expenditure, indicating that the true Engel curve is significantly downwards-sloping over this region. Therefore, one may conclude from the data-driven UCBs that the Engel curve for food differs significantly from a constant.

Conversely, the UCBs based on undersmoothing are wider, and appear to contain the zero function over the full range of total expenditure. Hence, one cannot reject a flat Engel curve using the UCBs based on undersmoothing. We also report PCIs for the derivative based on undersmoothing (using on `deriv.asy.se` and $N(0,1)$ critical values). These are useful for testing hypotheses about the derivative at a *given* point, but that is a different problem from testing whether the derivative is significantly different from zero at *any* point over its support.

We repeat the above exercise, this time estimating an Engel curve for household budget share on fuel:

```
R> fm.fuel.dd <- npiv(fuel ~ logexp | logwages, newdata = newdata)
R> summary(fm.fuel.dd)


Call:
npiv.formula(formula = fuel ~ logexp | logwages, newdata = newdata)

Nonparametric IV Model:

IV Regression Data: 1027 training points, and 1000 evaluation points, in 1 endogenous vari
```
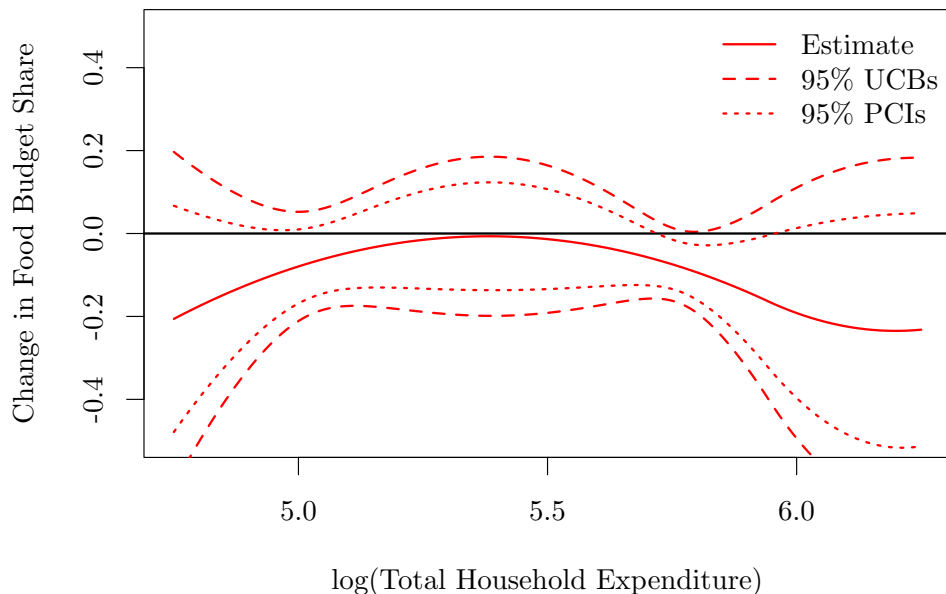
Figure 4: Engel Curve Derivative for Food: Estimates, UCBs, and PCIs based on Under-smoothing

```
B-spline degree for instruments:            4
B-spline segments for instruments:          4

B-spline degree for endogenous predictors:   3
B-spline segments for endogenous predictors: 1

Estimation time: 1.8 seconds
```

We see from the estimated Engel curve and its derivative that budget shares on fuel decrease significantly at low total expenditure levels and then flatten out completely at moderate to high total expenditure levels. As there is much less variation in households' fuel expenditure than their food expenditure, these curves are estimated with greater precision and the UCBs are correspondingly much narrower.

5. Conclusion

The package **npiv** makes available a set of novel, fully-data driven estimation and inference methods for NPIV models estimated using B-spline bases. The methods include (i) a data-driven choice of sieve dimension that leads to estimators of the structural function and its derivative that converge at the minimax sup-norm rate, and (ii) data-driven uniform confidence bands for the structural function and its derivative that contract at the minimax rate (up to log terms). Additional functionality allows for the construction of uniform confidence bands and pointwise confidence intervals based on undersmoothing. The design is user-friendly with syntax similar to the widely used package **ivreg** (Fox *et al.* 2021) for linear IV models.
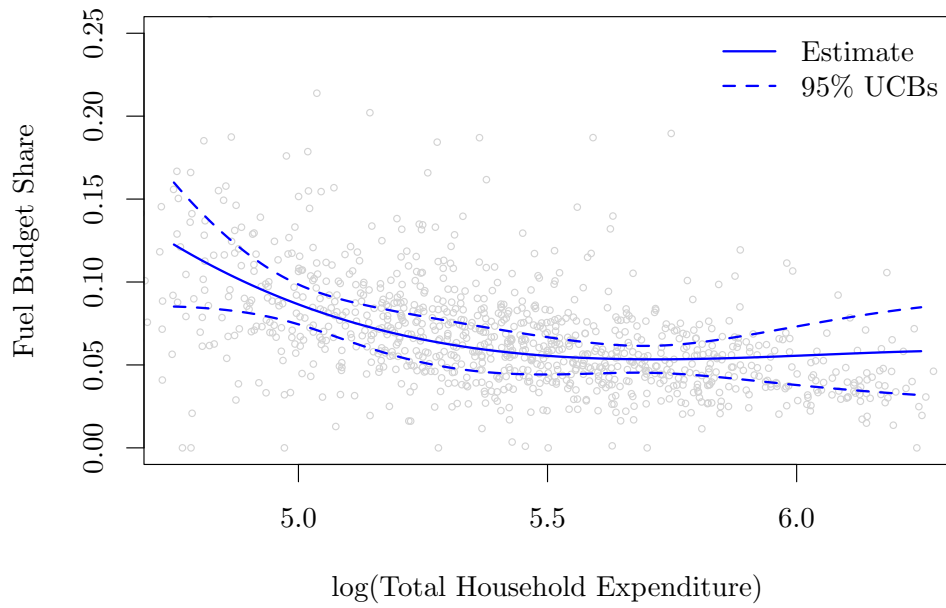
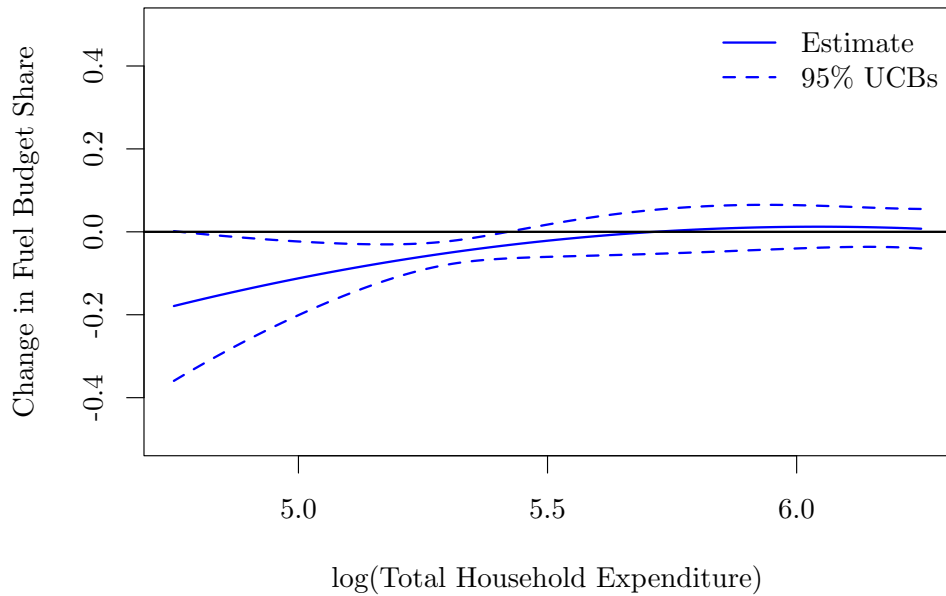Figure 5: Engel Curve for Fuel: Data-driven Estimates and UCBs

Figure 6: Engel Curve Derivative for Fuel: Data-driven Estimates and UCBs

Overall, the methods should be useful to researchers engaged in demand estimation, causal inference, and reinforcement learning to name a few. We welcome suggestions from the community for additional features to be implemented.

References

Ai C, Chen X (2003). "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions." *Econometrica*, **71**(6), 1795–1843. `doi:https://doi.org/10.1111/1468-0262.00470`. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0262.00470`.

Banks J, Blundell R, Lewbel A (1997). "Quadratic Engel Curves and Consumer Demand." *The Review of Economics and Statistics*, **79**(4), 527–539. `doi:10.1162/003465397557015`.

Belloni A, Chernozhukov V, Chetverikov D, Kato K (2015). "Some new asymptotic theory for least squares series: Pointwise and uniform results." *Journal of Econometrics*, **186**(2), 345–366. `doi:https://doi.org/10.1016/j.jeconom.2015.02.014`.

Berry ST, Haile PA (2014). "Identification in Differentiated Products Markets Using Market Level Data." *Econometrica*, **82**(5), 1749–1797. `doi:https://doi.org/10.3982/ECTA9027`. URL `https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA9027`.

Blundell R, Chen X, Kristensen D (2007). "Semi-nonparametric IV estimation of shape-invariant Engel curves." *Econometrica*, **75**(6), 1613–1669. `doi:https://doi.org/10.1111/j.1468-0262.2007.00808.x`.

Blundell R, Horowitz J, Parey M (2017). "Nonparametric Estimation of a Nonseparable Demand Function under the Slutsky Inequality Restriction." *The Review of Economics and Statistics*, **99**(2), 291–304.

Chai A, Moneta A (2010). "Retrospectives: Engel Curves." *Journal of Economic Perspectives*, **24**(1), 225–40. `doi:10.1257/jep.24.1.225`. URL `https://www.aeaweb.org/articles?id=10.1257/jep.24.1.225`.

Chen X, Christensen T, Kankanala S (2024). "Adaptive Estimation and Uniform Confidence Bands for Nonparametric Structural Functions and Elasticities." *Review of Economic Studies*, **forthcoming**. URL `https://doi.org/10.1093/restud/rdae025`.

Chen X, Christensen TM (2015). "Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions." *Journal of Econometrics*, **188**(2), 447–465. `doi:https://doi.org/10.1016/j.jeconom.2015.03.010`.

Chen X, Christensen TM (2018). "Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression." *Quantitative Economics*, **9**(1), 39–84. `doi:https://doi.org/10.3982/QE722`. `https://onlinelibrary.wiley.com/doi/pdf/10.3982/QE722`.

Chen X, Pouzo D (2015). "Sieve Wald and QLR inferences on semi/nonparametric conditional moment models." *Econometrica*, **83**(3), 1013–1079.

Chen Y, Xu L, Gulcehre C, Paine TL, Gretton A, De Freitas N, Doucet A (2022). "On instrumental variable regression for deep offline policy evaluation." *Journal of Machine Learning Research*, **23**(1). ISSN 1532-4435.

Fox J, Kleiber C, Zeileis A, Kuschnig N (2021). **ivreg**: *Instrumental-Variables Regression by '2SLS', '2SM', or '2SMM', with Diagnostics*. R package version 0.6-1, URL https://CRAN.R-project.org/package=ivreg.

Kaddour J, Lynch A, Liu Q, Kusner MJ, Silva R (2022). "Causal Machine Learning: A Survey and Open Problems." *arxiv:2206.15475 [cs.lg]*.

Miao W, Geng Z, Tchetgen Tchetgen EJ (2018). "Identifying causal effects with proxy variables of an unmeasured confounder." *Biometrika*, **105**(4), 987–993. ISSN 0006-3444. doi:10.1093/biomet/asy038. https://academic.oup.com/biomet/article-pdf/105/4/987/27121264/asy038.pdf, URL https://doi.org/10.1093/biomet/asy038.

Newey WK, Powell JL (2003). "Instrumental Variable Estimation of Nonparametric Models." *Econometrica*, **71**(5), 1565–1578. doi:https://doi.org/10.1111/1468-0262.00459. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0262.00459.

Racine JS, Nie Z, Ripley BD (2022). **crs**: *Categorical Regression Splines*. R package version 0.15-36, URL https://CRAN.R-project.org/package=crs.

Affiliation:

Timothy Christensen
University College London
Department of Economics
E-mail: t.christensen@ucl.ac.uk
URL: https://tmchristensen.com

Xiaohong Chen
Yale University
Cowles Foundation for Research in Economics and
Department of Economics
E-mail: xiaohong.chen@yale.edu
URL: https://economics.yale.edu/people/faculty/xiaohong-chen

Jeffrey S. Racine
McMaster University
Department of Economics and
Graduate Program in Statistics
E-mail: racinej@mcmaster.ca
URL: https://socialsciences.mcmaster.ca/people/racinej