Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

Estimation and Inference for Nonparametric Instrumental Variables Models: The R Package npiv

Jeffrey S. Racine
McMaster University

Timothy Christensen
New York University and
University College London

Xiaohong Chen
Yale University

Abstract

Nonparametric instrumental variables (NPIV) methods are widely used for causal inference in economics and have been adopted in other fields including machine learning, by way of illustration. This article discusses the R package **npiv**, which implements novel, fully data-driven NPIV estimation and inference methods proposed in Chen, Christensen, and Kankanala (2021a). 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) nonparametric uniform confidence bands based on the methods proposed in Chen and Christensen (2018). All methods subsume standard 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 relatively 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 non-economic areas including causal inference in biostatistics (Miao, Geng, and Tchetgen Tchetgen 2018) as well as reinforcement learning (Chen, Xu, Gulcehre, Paine, Gretton, de Freitas, and Doucet 2021b), and, more generally, causal machine learning (Kaddour, Lynch, Liu, Kusner, and Silva 2022). The basic premise underlying NPIV is that we are interested in estimating a nonparametric *structural function* 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. Unlike a standard nonparametric regression model, the function h_0 is not necessarily the conditional mean of Y given X 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. NPIV methods allow for the estimation of h_0 using a vector of *instrumental variables* W for which $\mathsf{E}[u|W] = 0$.[1] 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 for h_0 and derivatives of h_0 proposed in Chen *et al.* (2021a). The methods are developed in the context of sieve NPIV estimators, which are implemented in a very simple and straightforward manner using two-stage least-squares (TSLS) (Ai and Chen 2003; Newey and Powell 2003).

The main methods implemented in this package 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 and its derivatives. As shown in Chen *et al.* (2021a), the single data-driven choice of sieve dimension is adaptive to the optimal (i.e., minimax) sup-norm rate 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. All methods subsume standard 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.* (2021a) 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 of NPIV models and nonparametric regression models. Section 4 then illustrates the methods in the context of Engel curve estimation based on household consumption data.

2. Estimation and inference procedures

2.1. Sieve NPIV estimators

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 $\psi^J(x) = (\psi_{J1}(x), \ldots, \psi_{JJ}(x))^\top$ is a vector of basis functions and $c_J = (c_{J1}, \ldots, c_{JJ})^\top$ 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 $u = Y - h_0(X)$ satisfies $\mathsf{E}[u|W] = 0$. Provided bias_J 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 $K \geq J$ transformations $b^K(W) = (b_{K1}(W), \ldots, b_{KK}(W))^\top$ of W as instruments.

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

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

where $\mathbf{\Psi}_J = (\psi^J(X_1), \ldots, \psi^J(X_n))^\top$ and $\mathbf{B}_K = (b^K(W_1), \ldots, b^K(W_n))^\top$ 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 the projection matrix onto the instrument space,[2] and $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ is an $n \times 1$ vector. Sieve NPIV 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 $\partial^a \psi^J(x) = (\partial^a \psi_{J1}(x), \ldots, \partial^a \psi_{JJ}(x))^\top$ and

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

for $a = (a_1, \ldots, a_d) \in (\mathbb{N}_0)^d$ with $|a| = a_1 + \ldots + a_d$.

The estimator $\hat{h}_J$ collapses to a standard nonparametric series regression estimator

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

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

In practice, a researcher must choose bases $\psi_{J1}, \ldots, \psi_{JJ}$ and $b_{K1}, \ldots, b_{KK}$ and the tuning parameters J and K. Previously, Chen and Christensen (2018) showed that B-spline and certain 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 optimal rate in the special case of nonparametric regression (Belloni, Chernozhukov, Chetverikov, and Kato 2015; Chen and Christensen 2015), again under an appropriate choice of sieve dimension. We therefore use B-spline bases in what follows as these are easier to compute than wavelet bases. We generally construct bases for multivariate X or W by taking tensor products of univariate bases.[3]

The key tuning parameter to be chosen when implementing sieve NPIV estimators is the dimension J of the basis used to approximate h_0. Note that the bias of sieve NPIV estimators is decreasing in J because h_0 is approximated over a richer set of basis functions. On

[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.

the other hand, the variance (or sampling uncertainty) is increasing in J because there are more coefficients to estimate. Choices of J that appropriately balance bias against sampling uncertainty will lead to estimators that converge at optimal rates. Note, however, that the bias and sampling uncertainty depend on certain model regularities, such as the smoothness of h_0 and the strength of the instruments, which are unknown. It is therefore very important to have data-driven methods for choosing tuning parameters that can *adapt* to these unknown model regularities.

2.2. Data-driven choice of sieve dimension

Chen *et al.* (2021a) introduce a data-driven choice of J 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.

Before describing the procedure of Chen *et al.* (2021a) we first introduce some notation. Let $\mathcal{T}$ denote the set of feasible values of J. This set depends on the choice of basis. For instance, if X is d-dimensional and $\psi_{J1}, \ldots, \psi_{JJ}$ is the tensor product of r-th degree univariate B-spline bases, then $\mathcal{T} = \{(2^l + r)^d : l \in \mathbb{N}_0\}$. In particular, with a univariate cubic B-spline basis 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

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

The contrast $D_J(x) - D_{J_2}(x)$ compares the estimates of $\hat{h}_J(x) - h_0(x)$ and $\hat{h}_{J_2}(x) - h_0(x)$. Its variance is estimated by

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

where

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

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 $\widehat{\mathbf{U}}_{J,J_2}$ a $n \times n$ diagonal matrix whose ith diagonal entry is $\hat{u}_{i,J}\hat{u}_{i,J_2}$. 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.* (2021a) for choosing J in NPIV models is summarized in the following three steps:

1. Compute

$$\hat{\mathcal{J}} = \left\{ J \in \mathcal{T} : 0.1(\log \hat{J}_{\max})^2 \leq J \leq \hat{J}_{\max} \right\} \tag{3}$$

where

$$\hat{J}_{\max} = \min \left\{ J \in \mathcal{T} : J\sqrt{\log J}\hat{s}_J^{-1} \leq 10\sqrt{n} < J^+\sqrt{\log J^+}\hat{s}_{J^+}^{-1} \right\}. \tag{4}$$

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 $(\varpi_i)_{i=1}^n$.

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

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

The data-driven choice of sieve dimension is

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

Chen *et al.* (2021a) 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 optimal sup-norm rate in NPIV models. 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

$$\hat{J}_{\max} = \min\left\{ J \in \mathcal{T} : J\sqrt{\log J}\upsilon_n \leq 10\sqrt{n} < J^+\sqrt{\log J^+}\upsilon_n \right\} \tag{8}$$

with $\upsilon_n = \max\{1, (0.1 \log n)^4\}$, then compute $\hat{\mathcal{J}}$ as in (3) 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.* (2021a). 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

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

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

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

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

UCBs for derivatives $\partial^a h_0$ of the structural function in NPIV models are constructed analogously:

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

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

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

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

$$C^a_n(x) = \left[\partial^a\hat{h}_{\tilde{J}}(x) - \left(z^{a*}_{1-\alpha} + \hat{A}\theta^*_{1-\hat{\alpha}}\right)\hat{\sigma}^a_{\tilde{J}}(x),\ \partial^a\hat{h}_{\tilde{J}}(x) + \left(z^{a*}_{1-\alpha} + \hat{A}\theta^*_{1-\hat{\alpha}}\right)\hat{\sigma}^a_{\tilde{J}}(x)\right]. \tag{12}$$

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.* (2021a) establish uniform asymptotic coverage guarantees for these UCBs for both NPIV and nonparametric regression models. They also show that the UCBs contract at the optimal sup-norm rate (up to logarithmic factors).

2.4. Undersmoothed uniform confidence bands

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.

We now briefly summarize their 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

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

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

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

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

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

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

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the $N(0,1)$ distribution; Chen and Christensen (2015) and Belloni *et al.* (2015) provide 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 exampe, `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.

- `h.lower`: the lower UCB for the structural function.

- `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.
- `h.lower.deriv`: the lower UCB for the derivative of the structural function.
- `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 nonparameric 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 sup-norm. This choice may therefore differ from the choice of sieve dimension selected by other methods that target mean-square error (i.e., L^2 norm), such as cross validation, because the optimal choice for sup-norm and L^2 norm 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.* (2021a).

Estimators based on a deterministic choice of sieve dimension and undersmoothed UCBs can be computed by passing `J.x.segments`. Note that `K.w.segments` and `K.w.smooth` are both redundant for nonparametric regression and do not need to be specified. In this case, 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.

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`.
- `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 policy analysis and 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 fraction $h_0(x)$ of its total expenditure on food if log total expenditure was exogenously specified as x. As is well known (Banks, Blundell, and Lewbel 1997; Blundell, Browning, and Crawford 2003), total household expenditure is endogenous:

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

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

which follows from the fact that the household's consumption preferences determines its expenditure across categrories 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 household income (W_i) as an instrument for expenditure.

We use data from the 1995 British Family Expenditure Survey as in Blundell *et al.* (2007). We focus on the subset of 1655 married or cohabitating couples with the head of household employed and aged between 20 and 55 and with at most 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 household income. 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> Engel.kids <- Engel.kids[order(Engel.kids$logexp),]
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:

```
R> fm.food.dd <- npiv(food ~ logexp | logwages)
R> summary(fm.food.dd)

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

```
Nonparametric IV Model:

IV Regression Data: 1027 training points, in 1 endogenous variable(s)

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: 8.3 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, J.x.segments = 2,
+                           K.w.segments = 5)
```

The estimated Engel curves for food and plotted below, together with UCBs. Both the data-driven and undersmoothed estimates looks similar and are downward-sloping. This seems reasonable from an economic standpoint: food is a necessary good, and so it should decrease as a proportion of overall household expenditure as household expenditure increases. Note, however, that the data-driven estimate looks close to linear whereas the undersmoothed estimates and UCBs are quite wiggly. This is to be expected: as undersmoothed bands use a sub-optimally large choice of sieve dimension and will therefore be less "efficient" than data-driven bands.

These differences are also apparent when we compare the estimated derivatives of the Engel curves. The UCBs for derivatives are generally negative, but the zero function lies almost entirely inside both sets of UCBs. Therefore, while the estimates appear downwards-sloping, it is difficult to conclude that the Engel curves for food are significantly downwards-sloping.

We repeat the exercise to estimate an Engel curve for household budget share on fuel:

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

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

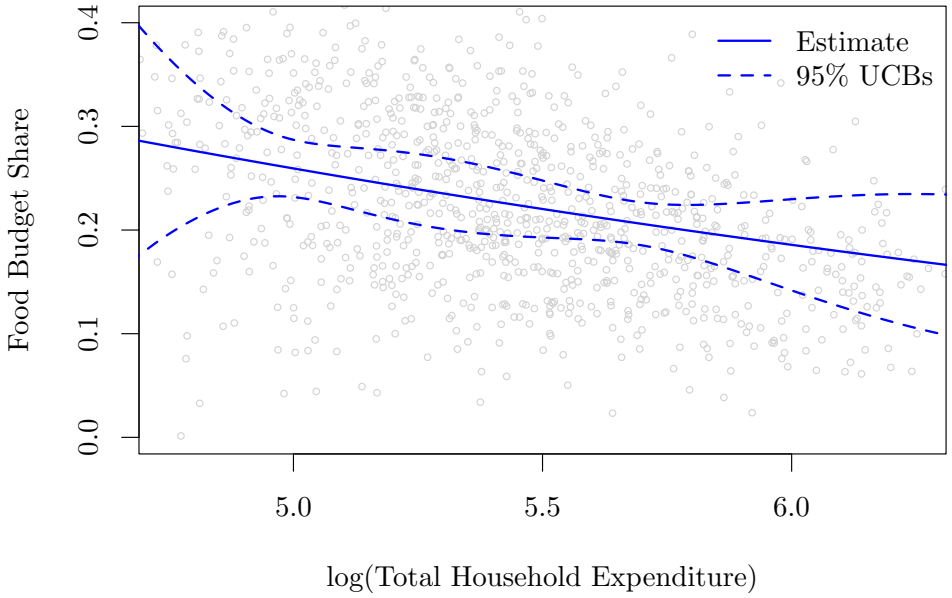

Figure 1: Engel Curve for Food: Data-driven Estimates and UCBs

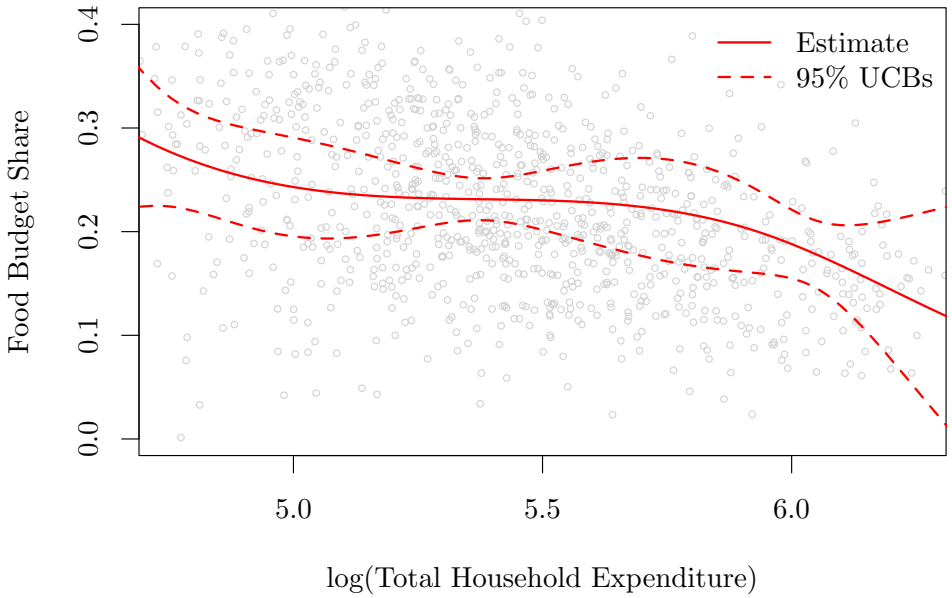

Figure 2: Engel Curve for Food: Estimates and UCBs based on Undersmoothing

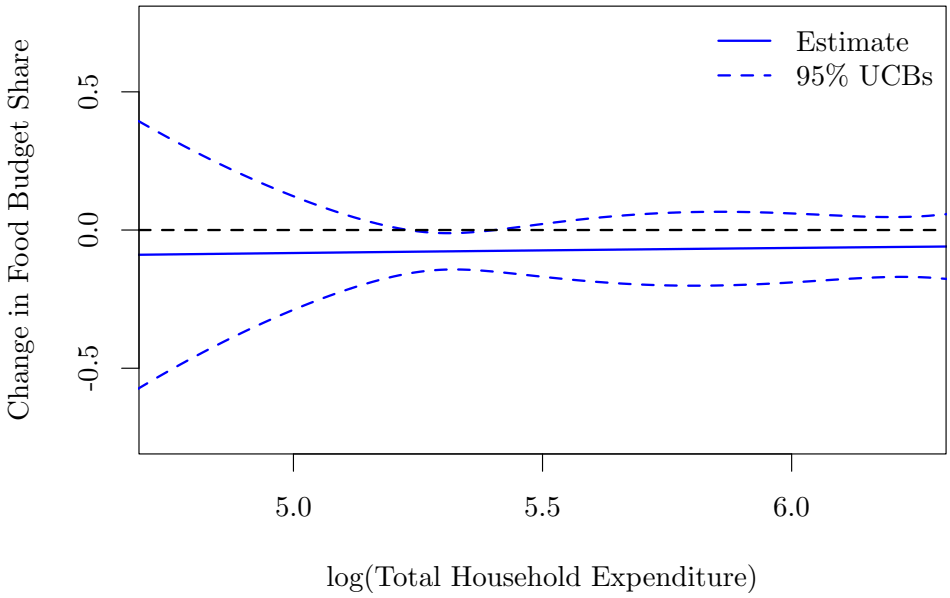

Figure 3: Engel Curve Derivative for Food: Data-driven Estimates and UCBs

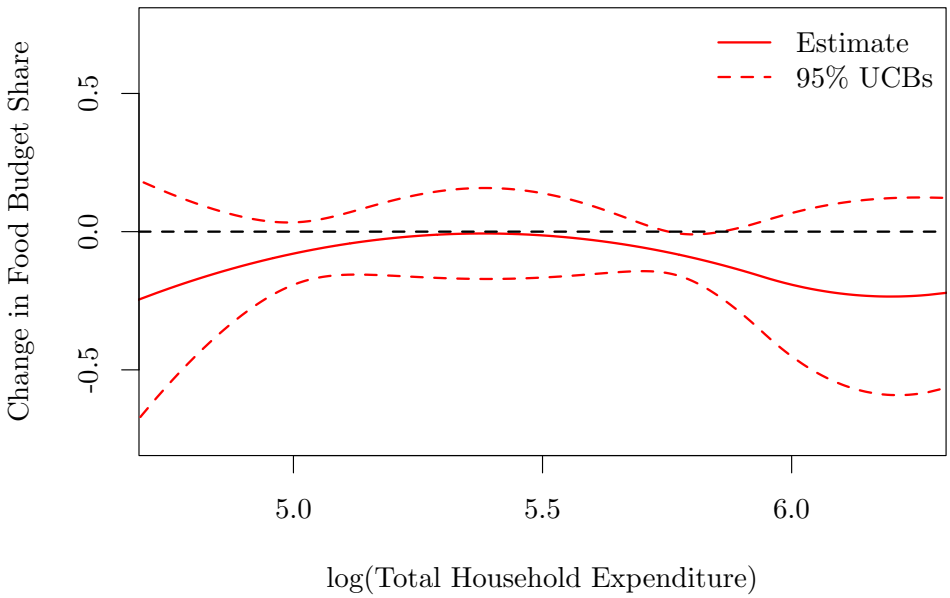

Figure 4: Engel Curve Derivative for Food: Estimates and UCBs based on Undersmoothing

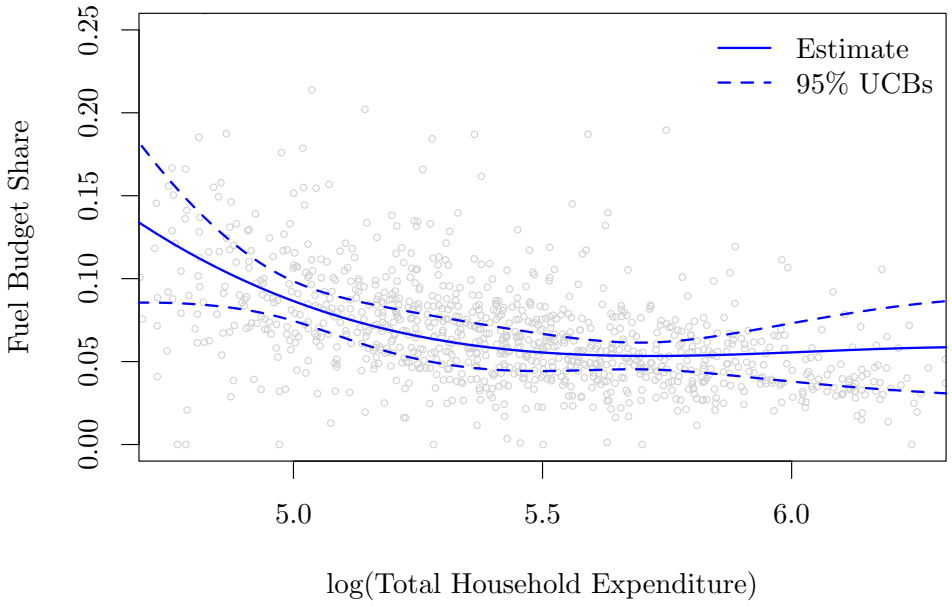

Figure 5: Engel Curve for Fuel: Data-driven Estimates and UCBs

```
Nonparametric IV Model:

IV Regression Data: 1027 training points, in 1 endogenous variable(s)

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: 3.6 seconds
```

We see from the estimated Engel curve and its derivative that budget shares on fuel decrease significantly at low income levels and then flatten out completely at moderate to high income 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.

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.

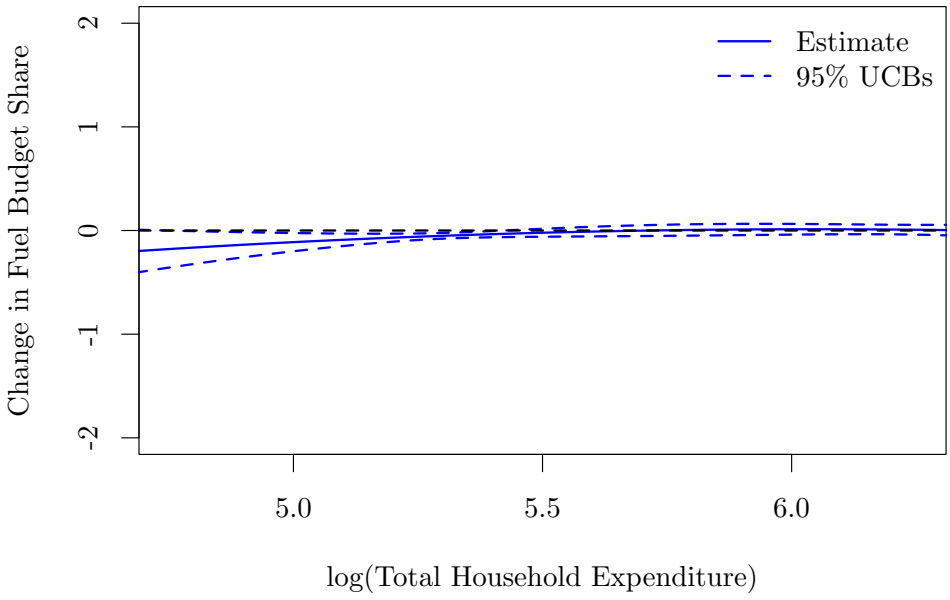

Figure 6: Engel Curve Derivative for Fuel: Data-driven Estimates and UCBs

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.

Blundell RW, Browning M, Crawford IA (2003). "Nonparametric Engel Curves and Revealed Preference." *Econometrica*, **71**(1), 205–240. URL http://www.jstor.org/stable/3082045.

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 (2021a). "Adaptive Estimation and Uniform Confidence Bands for Nonparametric Structural Functions and Elasticities." *arXiv preprint arXiv:2107.11869 [econ.EM]*. URL https://arxiv.org/abs/2107.11869.

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 Y, Xu L, Gulcehre C, Paine TL, Gretton A, de Freitas N, Doucet A (2021b). "On Instrumental Variable Regression for Deep Offline Policy Evaluation." *arxiv:2105.10148 [cs.lg]*.

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:

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

Timothy Christensen
New York University and
University College London
Department of Economics
E-mail: timothy.christensen@nyu.edu
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

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd