Package: npiv (via r-universe)

January 8, 2025

Type Package

Title Nonparametric Instrumental Variables Estimation and Inference

Version 0.1.3

Date 2024-12-10

Imports progress, MASS, Formula, withr

Maintainer Timothy Christensen <timothy.christensen@yale.edu>

Description Implements methods introduced in Chen, Christensen, and Kankanala (2024) <doi:10.1093/restud/rdae025> for estimating and constructing uniform confidence bands for nonparametric structural functions using instrumental variables, including data-driven choice of tuning parameters. All methods in this package apply to nonparametric regression as a special case.

NeedsCompilation yes

Author Jeffrey S. Racine [aut], Timothy Christensen [aut, cre], Patrick Alken [ctb], Rhys Ulerich [ctb], Simon N. Wood [ctb]

Repository CRAN

Date/Publication 2025-01-08 15:10:02 UTC

Contents

`npiv-package` *Nonparametric Instrumental Variables Estimation and Inference*

Description

This package implements the nonparametric instrumental variables estimation and inference methods described in Chen, Christensen, and Kankanala (2024) and Chen and Christensen (2018). The function `npiv` estimates the nonparametric structural function `h0` using B-splines and constructs uniform confidence bands for `h0`. The function `npiv_choose_J` performs data-driven choice of sieve dimension. All methods in this package apply to estimation and inference for nonparametric regression as a special case.

Details

This package provides a function `npiv(...)` with a simple interface for performing nonparametric instrumental variable estimation and inference.

Given a dependent variable vector `Y`, matrix of endogenous regressors `X`, and matrix of instruments `W`, `npiv` nonparametrically estimates the structural function `h0` and its derivative using B-splines. `npiv` can also be used for estimting the conditional mean `h0` of `Y` given `X`, as as well as the derivative of the conditional mean function, by nonparametric regression.

The function `npiv` also constructs uniform confidence bands for `h0` and its derivative.

Sieve dimensions are determined in a data-dependent way if not provided by the user via the function `npiv_choose_J`, which implements the methods described in Chen, Christensen, and Kankanala (2024). This data-driven choice of sieve dimension ensures estimators of `h0` and its derivatives converge at the optimal sup-norm rate. The resulting uniform confidence bands for `h0` and its derivative contract within a logarithmic factor of the optimal rate. In this way, `npiv` facilitates fully data-driven estimation and uniform inference on `h0` and its derivative.

If sieve dimensions are provided by the user, `npiv` implements the bootstrap-based procedure of Chen and Christensen (2018) to construct uniform confidence bands for `h0` and its derivative.

Author(s)

Jeffrey S. Racine <racinej@mcmaster.ca>, Timothy Christensen <timothy.christensen@yale.edu>

Maintainer: Timothy Christensen <timothy.christensen@yale.edu>

References

Chen, X. and T. Christensen (2018). “Optimal Sup-norm Rates and Uniform Inference on Nonlinear Functionals of Nonparametric IV Regression.” *Quantitative Economics*, **9**(1), 39-85. doi:10.3982/QE722

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

`Engel95` *1995 British Family Expenditure Survey*

Description

This dataset is based on a sample taken from the British Family Expenditure Survey for 1995. It includes households consisting of married or cohabiting couples with an employed head of household, aged between 25 and 55 years, and with at most two children. There are 1655 household-level observations in total.

Usage

```
data("Engel95")
```

Format

A data frame with 10 columns, and 1655 rows.

food expenditure share on food, of type `numeric`

catering expenditure share on catering, of type `numeric`

alcohol expenditure share on alcohol, of type `numeric`

fuel expenditure share on fuel, of type `numeric`

motor expenditure share on motor, of type `numeric`

fares expenditure share on fares, of type `numeric`

leisure expenditure share on leisure, of type `numeric`

logexp logarithm of total expenditure, of type `numeric`

logwages logarithm of total earnings, of type `numeric`

nkids '0' indicates no children, '1' indicates 1-2 children, of type `numeric`

Source

Richard Blundell and Dennis Kristensen

References

Blundell, R., X. Chen and D. Kristensen (2007). "Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves." *Econometrica*, **75**(6), 1613-1669. doi:10.1111/j.14680262.2007.00808.x

Chen, X. and T. Christensen (2018). "Optimal Sup-norm Rates and Uniform Inference on Nonlinear Functionals of Nonparametric IV Regression." *Quantitative Economics*, **9**(1), 39-85. doi:10.3982/QE722

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

Examples

```
## Load data
data("Engel95", package = "npiv")

## Sort on logexp (the regressor) for plotting purposes
Engel95 <- Engel95[order(Engel95$logexp),]
attach(Engel95)
logexp.eval <- seq(4.5,6.5,length=100)

## Estimate the Engel curve for food using logwages as an instrument
food_engel <- npiv(food, logexp, logwages, X.eval = logexp.eval)

## Plot the estimated function and uniform confidence bands
plot(food_engel, showdata = TRUE)
```

---

gsl.bs *GSL (GNU Scientific Library) B-spline/B-spline Derivatives*

Description

`gsl.bs` generates the B-spline basis matrix for a polynomial spline and (optionally) the B-spline basis matrix derivative of a specified order with respect to each predictor

Usage

```
gsl.bs(...)
## Default S3 method:
gsl.bs(x,
       degree = 3,
       nbreak = 2,
       deriv = 0,
       x.min = NULL,
       x.max = NULL,
       intercept = FALSE,
       knots = NULL,
       ...)
```

Arguments

`x`	the predictor variable. Missing values are not allowed
`degree`	degree of the piecewise polynomial - default is ‘3’ (cubic spline)
`nbreak`	number of breaks in each interval - default is ‘2’
`deriv`	the order of the derivative to be computed-default if `0`
`x.min`	the lower bound on which to construct the spline - defaults to `min(x)`
`x.max`	the upper bound on which to construct the spline - defaults to `max(x)`
`intercept`	if ‘TRUE’, an intercept is included in the basis; default is ‘FALSE’

`knots`	a vector (default `knots="NULL"`) specifying knots for the spline basis (default enables uniform knots, otherwise those provided are used)
`...`	optional arguments

Details

Typical usages are (see below for a list of options and also the examples at the end of this help file)

```
B <- gsl.bs(x,degree=10)
B.predict <- predict(gsl.bs(x,degree=10),newx=xeval)
```

Value

`gsl.bs` returns a `gsl.bs` object. A matrix of dimension 'c(length(x), degree+nbreak-1)'. The generic function `predict` extracts (or generates) predictions from the returned object.

A primary use is in modelling formulas to directly specify a piecewise polynomial term in a model. See `https://www.gnu.org/software/gsl/` for further details.

Author(s)

Jeffrey S. Racine <racinej@mcmaster.ca>

References

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

See Also

`bs`

Examples

```
## Plot the spline bases and their first order derivatives
x <- seq(0,1,length=100)
matplot(x,gsl.bs(x,degree=5),type="l")
matplot(x,gsl.bs(x,degree=5,deriv=1),type="l")

## Regression example
n <- 1000
x <- sort(runif(n))
y <- cos(2*pi*x) + rnorm(n,sd=.25)
B <- gsl.bs(x,degree=5,intercept=FALSE)
plot(x,y,cex=.5,col="grey")
lines(x,fitted(lm(y~B)))
```

npiv *Nonparametric Instrumental Variable Estimation and Inference*

Description

npiv performs nonparametric a structural function h0 and its derivatives using a B-spline sieve. It also constructs uniform confidence bands for h0 and its derivative.

Sieve dimensions are determined in a data-dependent way if not provided by the user, via the methods described in Chen, Christensen, and Kankanala (2024). This data-driven choice of sieve dimension ensures estimators of h0 and its derivatives converge at the optimal sup-norm rate. The resulting uniform confidence bands for h0 and its derivatives also converge at the minimax rate up to log factors; see Chen, Christensen, and Kankanala (2024).

If sieve dimensions are provided by the user, npiv implements the bootstrap-based procedure of Chen and Christensen (2018) to construct uniform confidence bands based on undersmoothing for h0 and its derivatives.

The methods in npiv apply to estimation and inference on a nonparametric regression function as a special case.

Usage

```
npiv(...)

## S3 method for class 'formula'
npiv(formula,
     data=NULL,
     newdata=NULL,
     subset=NULL,
     na.action="na.omit",
     call,
     ...)

## Default S3 method:
npiv(Y,
     X,
     W,
     X.eval=NULL,
     X.grid=NULL,
     alpha=0.05,
     basis=c("tensor","additive","glp"),
     boot.num=99,
     check.is.fullrank=FALSE,
     deriv.index=1,
     deriv.order=1,
     grid.num=50,
     J.x.degree=3,
     J.x.segments=NULL,
```

```
K.w.degree=4,
K.w.segments=NULL,
K.w.smooth=2,
knots=c("uniform","quantiles"),
progress=TRUE,
ucb.h=TRUE,
ucb.deriv=TRUE,
W.max=NULL,
W.min=NULL,
X.min=NULL,
X.max=NULL,
...)
```

Arguments

formula	a symbolic description of the model to be fit.
data	an optional data frame containing the variables in the model.
newdata	an optional data frame in which to look for variables with which to predict (i.e., predictors in X passed in X.eval which must contain identically named variables).
subset	an optional vector specifying a subset of observations to be used in the fitting process (see additional details about how this argument interacts with data-dependent bases in the 'Details' section of the model.frame documentation).
na.action	a function which indicates what should happen when the data contain NAs. The default is set by the na.action setting of options, and is na.fail if that is unset. The 'factory-fresh' default is na.omit. Another possible value is NULL, no action. Value na.exclude can be useful.
call	the original function call (this is passed internally by npiv). It is not recommended that the user set this.
Y	dependent variable vector.
X	matrix of endogenous regressors.
W	matrix of instrumental variables. Set W=X for nonparametric regression.
X.eval	optional matrix of evaluation data for the endogenous regressors.
X.grid	optional vector of grid points for X when determining model complexity. Default (X.grid=NULL) uses 50 equally spaced points (can be changed in grid.num) over the support of each X variable.
alpha	nominal size of the uniform confidence bands. Default is 0.05 for 95% uniform confidence bands.
basis	basis type (if X or W are multivariate), a character string. Options are: tensor tensor product basis. Default option. additive additive basis for additively separable models. glp generalized B-spline polynomial basis.
boot.num	number of bootstrap replications.

check.is.fullrank
: check that X and W have full rank. Default is FALSE.

deriv.index
: integer indicating the column of X for which to compute the derivative.

deriv.order
: integer indicating the order of derivative to be computed.

grid.num
: number of grid points for each X variable if X.grid is not provided.

J.x.degree
: B-spline degree (integer or vector of integers of length ncol(X)) for approximating the structural function. Default is degree=3 (cubic B-spline).

J.x.segments
: B-spline number of segments (integer or vector of integers of length ncol(X)) for approximating the structural function. Default is NULL. If either J.x.segments=NULL or K.w.segments=NULL, these are both chosen automatically using npiv_choose_J.

K.w.degree
: B-spline degree (integer or vector of integers of lenth ncol(W)) for estimating the nonparametric first-stage. Default is degree=4 (quartic B-spline).

K.w.segments
: B-spline number of segments (integer or vector of integers of length ncol(W)) estimating the nonparametric first stage. Defulat is NULL. If either J.x.segments=NULL or K.w.segments=NULL, these are both chosen automatically using npiv_choose_J.

K.w.smooth
: non-negative integer. 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.

knots
: knots type, a character string. Options are:

  quantiles interior knots are placed at equally spaced quantiles (equal number of observations lie in each segment).

  uniform interior knots are placed at equally spoaced intervals over the support of the variable. Default option.

progress
: whether to display progress bar or not. Default is TRUE.

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.

W.min
: lower bound on the support of each W variable. Default is min(W).

W.max
: upper bound on the support of each W variable. Default is max(W).

X.min
: lower bound on the support of each X variable. Default is min(X).

X.max
: upper bound on the support of each X variable. Default is max(X).

...
: optional arguments

Details

npiv estimates and constructs uniform confidence bands for a nonparametric structural function h_0 and its derivatives in the model $Y = h_0(X) + U, \quad E[U|W] = 0 \quad \text{(almost surely)}$. Estimation is performed using nonparametric two-stage least-squares with a B-spline sieve. The key tuning parameter is the dimension J of the sieve used to approximate h_0. The dimension is tuned via modifying the number and placement of interior knots in the B-spline basis (equivalently, the number of segments of the basis). Sieve dimensions can be user-provided or data-determined using the procedure of Chen, Christensen, and Kankanala (2024).

Typical usages mirror `ivreg` (see above and below for a list of options and the example at the bottom of this document)

```
foo <- npiv(y~x|w)
foo <- npiv(y~x1+x2|w1+w2)
foo <- npiv(Y=y,X=x,W=w)
```

`npiv` can be used in two ways:

1. Data-driven sieve dimension is invoked if either `K.w.segments` or `J.x.segments` are unspecified or `NULL` (the default). Sieve dimensions are chosen automatically using `npiv_choose_J`. Uniform confidence bands for h_0 and its derivatives are constructed using the data-driven method of Chen, Christensen, and Kankanala (2024).

2. The user may specify the sieve dimensions of both bases by specifying values for `K.w.segments` and `J.x.segments`. Uniform confidence bands for h_0 and its derivatives are constructed using the method of Chen and Christensen (2018).

`npiv` can also be used for estimation and inference on a nonparametric regression function by setting `W=X`.

Value

`npiv` returns a `npiv` object. The generic function `fitted` extracts the estimated values for the sample (or evaluation data, if provided), while the generic function `residuals` extracts the sample residuals. The generic function `summary` provides a simple model summary. The generic function `plot` also plots the estimated function and derivative, together with uniform confidence bands.

The function `npiv` returns a list with the following components:

`h`	estimated structural function evaluated at the sample data (or evaluation data, if provided).
`residuals`	residuals for the sample data.
`deriv`	estimated 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)
`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).
`deriv.index`	index for the estimated derivative.
`deriv.order`	order of the estimated derivative.
`K.w.degree`	value of `K.w.degree` used.
`K.w.segments`	value of `K.w.segments` used (will be data-determined if not provided).
`J.x.degree`	value of `J.x.degree` used.
`J.x.segments`	value of `J.x.segments` used (will be data-determined if not provided).
`beta`	vector of estimated spline coefficients.

Author(s)

Jeffrey S. Racine <racinej@mcmaster.ca>, Timothy Christensen <timothy.christensen@yale.edu>

References

Chen, X. and T. Christensen (2018). "Optimal Sup-norm Rates and Uniform Inference on Nonlinear Functionals of Nonparametric IV Regression." *Quantitative Economics*, **9**(1), 39-85. doi:10.3982/QE722

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

See Also

npiv_choose_J

Examples

```
## load data
data("Engel95", package = "npiv")

## sort on logexp (the regressor) for plotting purposes
Engel95 <- Engel95[order(Engel95$logexp),]
attach(Engel95)

## Estimate the Engel curve for food using logwages as an instrument
fm1 <- npiv(food ~ logexp | logwages)

## Plot the estimated Engel curve and data-driven uniform confidence bands
plot(logexp,food,
     ylab="Food Budget Share",
     xlab="log(Total Household Expenditure)",
     xlim=c(4.75, 6.25),
     ylim=c(0, 0.4),
     main="",
     type="p",
     cex=.5,
     col="lightgrey")
lines(logexp,fm1$h,col="blue",lwd=2,lty=1)
lines(logexp,fm1$h.upper,col="blue",lwd=2,lty=2)
lines(logexp,fm1$h.lower,col="blue",lwd=2,lty=2)

## Estimate the Engel curve using pre-specified sieve dimension
## (dimension 5 for logexp, dimension 9 for logwages)
fm2 <- npiv(food ~ logexp | logwages,
            J.x.segments = 2,
            K.w.segments = 5)

## Plot uniform confidence bands based on undersmoothing
lines(logexp,fm2$h.upper,col="red",lwd=2,lty=2)
lines(logexp,fm2$h.lower,col="red",lwd=2,lty=2)
```

```
## Plot pointwise confidence bands based on pre-asymptotic standard errors
lines(logexp,fm2$h+1.96*fm2$asy.se,col="red",lwd=2,lty=3)
lines(logexp,fm2$h-1.96*fm2$asy.se,col="red",lwd=2,lty=3)

legend("topright",
       legend=c("Data-driven Estimate",
                "Data-driven UCBs",
                "Undersmoothed UCBs",
                "Pointwise CBs"),
       col=c("blue","blue","red","red"),
       lty=c(1,2,2,3),
       lwd=c(2,2,2,2))

## Plot the data-driven estimate of the derivative of the Engel curve
plot(logexp,fm1$deriv,col="blue",lwd=2,lty=1,type="l",
     ylab="Derivative of Food Budget Share",
     xlab="log(Total Household Expenditure)",
     xlim=c(4.75, 6.25),
     ylim=c(-1,1))

## Plot data-driven uniform confidence bands for the derivative
lines(logexp,fm1$h.upper.deriv,col="blue",lwd=2,lty=2)
lines(logexp,fm1$h.lower.deriv,col="blue",lwd=2,lty=2)

## Plot uniform confidence bands based on undersmoothing
lines(logexp,fm2$h.upper.deriv,col="red",lwd=2,lty=2)
lines(logexp,fm2$h.lower.deriv,col="red",lwd=2,lty=2)

## Plot pointwise confidence bands based on pre-asymptotic standard errors
lines(logexp,fm2$deriv+1.96*fm2$deriv.asy.se,col="red",lwd=2,lty=3)
lines(logexp,fm2$deriv-1.96*fm2$deriv.asy.se,col="red",lwd=2,lty=3)

legend("topright",
       legend=c("Data-driven Estimate",
                "Data-driven UCBs",
                "Undersmoothed UCBs",
                "Pointwise CBs"),
       col=c("blue","blue","red","red"),
       lty=c(1,2,2,3),
       lwd=c(2,2,2,2))
```

`npiv_choose_J` *Data-driven Choice of Sieve Dimension for Nonparametric Instrumental Variables Estimation and Inference*

Description

`npiv_choose_J` implements the data-driven choice of sieve dimension developed in Chen, Christensen, and Kankanala (2024) for nonparametric instrumental variables estimation using a B-spline sieve. It applies to nonparametric regression as a special case.

Usage

```
npiv_choose_J(Y,
              X,
              W,
              X.grid = NULL,
              J.x.degree = 3,
              K.w.degree = 4,
              K.w.smooth = 2,
              knots = c("uniform", "quantiles"),
              basis = c("tensor", "additive", "glp"),
              X.min = NULL,
              X.max = NULL,
              W.min = NULL,
              W.max = NULL,
              grid.num = 50,
              boot.num = 99,
              check.is.fullrank = FALSE,
              progress = TRUE)
```

Arguments

`Y` dependent variable vector.

`X` matrix of endogenous regressors.

`W` matrix of instrumental variables. Set `W=X` for nonparametric regression.

`X.grid` vector of grid point(s). Default uses 50 equally spaced points over the support of each `X` variable.

`J.x.degree` B-spline degree (integer or vector of integers of length `ncol(X)`) for approximating the structural function. Default is `degree=3` (cubic B-spline).

`K.w.degree` B-spline degree (integer or vector of integers of lenth `ncol(W)`) for estimating the nonparametric first-stage. Default is `degree=4` (quartic B-spline).

`K.w.smooth` non-negative integer. 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`.

`knots` knots type, a character string. Options are:

`quantiles` interior knots are placed at equally spaced quantiles (equal number of observations lie in each segment).

`uniform` interior knots are placed at equally spoaced intervals over the support of the variable. Default option.

`basis` basis type (if `X` or `W` are multivariate), a character string. Options are:

`tensor` tensor product basis. Default option.

`additive` additive basis for additively separable models.

`glp` generalized B-spline polynomial basis.

`X.min` lower bound on the support of each `X` variable. Default is `min(X)`.

X.max	upper bound on the support of each X variable. Default is max(X).
W.min	lower bound on the support of each W variable. Default is min(W).
W.max	upper bound on the support of each W variable. Default is max(W).
grid.num	number of grid points for each X variable if X.grid is not provided.
boot.num	number of bootstrap replications.
check.is.fullrank	check that X and W have full rank. Default is FALSE.
progress	whether to display progress bar or not. Default is TRUE.

Value

J.hat.max	largest element of candidate set of sieve dimensions searched over.
J.hat.n	second largest element of candidate set of sieve dimensions searched over.
J.hat	bootstrap-based Lepski choice of sieve dimension.
J.tilde	data-driven choice of sieve dimension using the method of Chen, Christensen, and Kankanala (2024). Minimum of J.hat and J.hat.n.
J.x.seg	data-driven number of segments for X using the method of Chen, Christensen, and Kankanala (2024).
K.w.seg	data-driven number of segments for W using the method of Chen, Christensen, and Kankanala (2024).
theta.star	Lepski critical value used in determination of J.hat.

Author(s)

Jeffrey S. Racine <racinej@mcmaster.ca>, Timothy Christensen <timothy.christensen@yale.edu>

References

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

See Also

npiv

Examples

```
library(MASS)

## Simulate the data
n <- 10000
cov.ux <- 0.5
var.u <- 0.1
mu <- c(1,1,0)
Sigma <- matrix(c(1.0,0.85,cov.ux,
                  0.85,1.0,0.0,
```

```
                  cov.ux,0.0,1.0),
                3,3,
                byrow=TRUE)
foo <- mvrnorm(n = n,
               mu,
               Sigma)
X <- 2*pnorm(foo[,1],mean=mu[1],sd=sqrt(Sigma[1,1])) -1
W <- 2*pnorm(foo[,2],mean=mu[2],sd=sqrt(Sigma[2,2])) -1
U <- foo[,3]
## Cosine structural function
h0 <- sin(pi*X)
Y <- h0 + sqrt(var.u)*U

npiv_choose_J(Y,X,W)
```

Index