Package 'lpdensity'

July 13, 2019

Title Local Polynomial Density Estimation and Inference

Version 1.0

Author Matias D. Cattaneo, Michael Jansson, Xinwei Ma

Maintainer Xinwei Ma <x1ma@ucsd.edu>

Description Without imposing stringent distributional assumptions or shape restrictions, nonparametric density estimation has been popular in economics and other social sciences for counterfactual analysis, program evaluation, and policy recommendations. This package implements a novel density estimator based on local polynomial regressions, documented in Cattaneo, Jansson and Ma (2019a, b) <arXiv:1811.11512, arXiv:1906.06529>: lpdensity() to construct local polynomial based density (and derivatives) estimator; lpbwdensity() to perform data-driven bandwidth selection; and lpdensity.plot() for density plot with robust confidence interval.

Imports ggplot2

Depends R (>= 3.1)

License GPL-2

Encoding UTF-8

LazyData true

RoxygenNote 6.1.1

NeedsCompilation no

Repository CRAN

Date/Publication 2019-07-12 22:30:21 UTC

R topics documented:

lpdensity-package 2
lpbwdensity 2
lpdensity 4
lpdensity.plot 6

Index **9**

lpdensity-package *lpdensity: Local Polynomial Density Estimation and Inference*

Description

Without imposing stringent distributional assumptions or shape restrictions, nonparametric density estimation has been popular in economics and other social sciences for counterfactual analysis, program evaluation, and policy recommendations. This package implements a novel density estimator based on local polynomial regression, documented in Cattaneo, Jansson and Ma (2019a): `lpdensity` to construct local polynomial based density estimator, `lpbwdensity` to perform data-driven bandwidth selection, and `lpdensity.plot` for density plot with robust confidence interval.

The companion software article, Cattaneo, Jansson and Ma (2019b), provides further implementation details and illustrations with simulated data. For related `Stata` and `R` packages useful for nonparametric data analysis and statistical inference, visit https://sites.google.com/site/nppackages/.

Author(s)

Matias D. Cattaneo, Princeton University. <cattaneo@princeton.edu>.

Michael Jansson, University of California Berkeley. <mjansson@econ.berkeley.edu>.

Xinwei Ma (maintainer), University of California San Diego. <x1ma@ucsd.edu>.

References

M.D. Cattaneo, M. Jansson and X. Ma. (2019a). Simple Local Polynomial Density Estimators. *Journal of the American Statistical Association*, forthcoming.

M.D. Cattaneo, M. Jansson and X. Ma. (2019b). `lpdensity`: Local Polynomial Density Estimation and Inference. Working paper.

lpbwdensity *Data-driven Bandwidth Selection for Local Polynomial Density Estimators*

Description

`lpbwdensity` implements the bandwidth selector for local polynomial based density (and derivatives) estimation, proposed in Cattaneo, Jansson and Ma (2019a). See Cattaneo, Jansson and Ma (2019b) for more implementation details and illustrations.

Companion command: `lpdensity` for local polynomial density estimation.

For more details, and related Stata and R packages useful for empirical analysis, visit https://sites.google.com/site/nppackages/.

Usage

```
lpbwdensity(data, grid = NULL, bwselect = c("mse-dpi", "imse-dpi",
  "mse-rot", "imse-rot"), p = NULL, v = NULL,
  kernel = c("triangular", "uniform", "epanechnikov"), Cweights = NULL,
  Pweights = NULL, regularize = TRUE)
```

Arguments

data	Numeric vector or one dimensional matrix / data frame, the raw data.
grid	Numeric vector or one dimensional matrix / data frame, the grid on which density is estimated. When set to default, grid points will be chosen as 0.05-0.95 percentiles of the data, with 0.05 step size.
bwselect	String, the method for data-driven bandwidth selection. This option will be ignored if bw is provided. Can be (1) "mse-dpi" (default, mean squared error-optimal bandwidth selected for each grid point); or (2) "imse-dpi" (integrated MSE-optimal bandwidth, common for all grid points); (3) "mse-rot" (rule-of-thumb bandwidth with Gaussian reference model); and (4) "imse-rot" (integrated rule-of-thumb bandwidth with Gaussian reference model).
p	Integer, nonnegative, the order of the local-polynomial used to construct point estimates. (Default is 2.)
v	Integer, nonnegative, the derivative of distribution function to be estimated. 0 for the distribution function, 1 (default) for the density funtion, etc.
kernel	String, the kernel function, should be one of "triangular", "uniform" or "epanechnikov".
Cweights	Numeric vector or one dimensional matrix / data frame, the weights used for counterfactual distribution construction. Should have the same length as sample size. This option will be ignored if bwselect is "ROT" or "IROT".
Pweights	Numeric vector or one dimensional matrix / data frame, the weights used in sampling. Should have the same length as sample size, and nonnegative. This option will be ignored if bwselect is "ROT" or "IROT".
regularize	TRUE (default) or FALSE, whether the bandwidth should be regularized. When set to TRUE, the bandwidth is chosen such that at least 20 + p + 1 are available locally.

Value

BW	A matrix containing (1) grid (grid points), (2) bw (bandwidths), and (3) nh (effective/local sample sizes).
opt	A list containing options passed to the function.

Author(s)

Matias D. Cattaneo, Princeton University. <cattaneo@princeton.edu>.

Michael Jansson, University of California Berkeley. <mjansson@econ.berkeley.edu>.

Xinwei Ma (maintainer), University of California San Diego. <x1ma@ucsd.edu>.

References

M.D. Cattaneo, M. Jansson and X. Ma. (2019a). Simple Local Polynomial Density Estimators. *Journal of the American Statistical Association*, forthcoming.

M.D. Cattaneo, M. Jansson and X. Ma. (2019b). lpdensity: Local Polynomial Density Estimation and Inference. Working paper.

See Also

lpdensity.

Examples

```
# Generate a random sample
set.seed(42); X <- rnorm(2000)

# Construct bandwidth
summary(lpbwdensity(X))
```

---

lpdensity *Local Polynomial Density Estimation and Inference*

Description

lpdensity implements the local polynomial regression based density (and derivatives) estimator proposed in Cattaneo, Jansson and Ma (2019a). This command can also be used to obtain smoothed estimates of cumulative distribution functions. See Cattaneo, Jansson and Ma (2019b) for more implementation details and illustrations.

Companion command: lpbwdensity for data-driven bandwidth selection, and lpdensity.plot for density plot with robust confidence interval.

For more details, and related Stata and R packages useful for empirical analysis, visit https://sites.google.com/site/nppackages/.

Usage

```
lpdensity(data, grid = NULL, bw = NULL, p = NULL, q = NULL,
  v = NULL, bwselect = c("mse-dpi", "imse-dpi", "mse-rot", "imse-rot"),
  kernel = c("triangular", "uniform", "epanechnikov"), Cweights = NULL,
  Pweights = NULL, scale = NULL)
```

Arguments

data	Numeric vector or one dimensional matrix / data frame, the raw data.
grid	Numeric vector or one dimensional matrix / data frame, the grid on which density is estimated. When set to default, grid points will be chosen as 0.05-0.95 percentiles of the data, with 0.05 step size.

bw	Numeric vector or one dimensional matrix / data frame, the bandwidth used for estimation. Can be (1) a positive scalar (common bandwidth for all grid points); or (2) a positive numeric vector specifying bandwidths for each grid point (should be the same length as grid).
p	Integer, nonnegative, the order of the local-polynomial used to construct point estimates. (Default is 2.)
q	Integer, nonnegative, the order of the local-polynomial used to construct pointwise confidence interval (a.k.a. the bias correction order). Default is p+1. When specified the same as p, no bias correction will be performed. Otherwise it should be strictly larger than p.
v	Integer, nonnegative, the derivative of distribution function to be estimated. 0 for the distribution function, 1 (default) for the density funtion, etc.
bwselect	String, the method for data-driven bandwidth selection. This option will be ignored if bw is provided. Can be (1) "mse-dpi" (default, mean squared error-optimal bandwidth selected for each grid point); or (2) "imse-dpi" (integrated MSE-optimal bandwidth, common for all grid points); (3) "mse-rot" (rule-of-thumb bandwidth with Gaussian reference model); and (4) "imse-rot" (integrated rule-of-thumb bandwidth with Gaussian reference model).
kernel	String, the kernel function, should be one of "triangular", "uniform" or "epanechnikov".
Cweights,	Numeric vector or one dimensional matrix / data frame, the weights used for counterfactual distribution construction. Should have the same length as sample size.
Pweights	Numeric vector or one dimensional matrix / data frame, the weights used in sampling. Should have the same length as sample size and nonnegative.
scale	Numeric, scaling factor for the final estimate. This parameter controls how estimates are scaled. For example, setting this parameter to 0.5 will scale down both the point estimates and standard errors by half. By default it is 1. This parameter is used if only part of the sample is used for estimation, and should not be confused with Cweights or Pweights.

Value

Estimate	A matrix containing (1) grid (grid points), (2) bw (bandwidths), (3) nh (effective/local sample sizes), (4) f_p (point estimates with p-th order local polynomial), (5) f_q (point estimates with q-th order local polynomial, only if option q is nonzero), (6) se_p (standard error corresponding to f_p), and (7) se_q (standard error corresponding to f_q).
opt	A list containing options passed to the function.

Author(s)

Matias D. Cattaneo, Princeton University. <cattaneo@princeton.edu>.

Michael Jansson, University of California Berkeley. <mjansson@econ.berkeley.edu>.

Xinwei Ma (maintainer), University of California San Diego. <x1ma@ucsd.edu>.

References

M.D. Cattaneo, M. Jansson and X. Ma. (2019a). Simple Local Polynomial Density Estimators. *Journal of the American Statistical Association*, forthcoming.

M.D. Cattaneo, M. Jansson and X. Ma. (2019b). lpdensity: Local Polynomial Density Estimation and Inference. Working paper.

See Also

lpbwdensity and lpdensity.plot.

Examples

```
# Generate a random sample
set.seed(42); X <- rnorm(2000)

# Estimate density and report results
est1 <- lpdensity(data = X, bwselect = "imse-dpi")
summary(est1)
```

---

lpdensity.plot *Local Polynomial Density Plot with Robust Confidence Intervals*

Description

lpdensity.plot plots estimated density/derivatives. This command can also be used to plot smoothed distribution function. See Cattaneo, Jansson and Ma (2019b) for more implementation details and illustrations.

Companion command: lpdensity for local polynomial based density and derivatives estimation.

For more details, and related Stata and R packages useful for empirical analysis, visit https://sites.google.com/site/nppackages/.

Usage

```
lpdensity.plot(..., alpha = NULL, type = NULL, CItype = NULL,
  title = "", xlabel = "", ylabel = "", lty = NULL, lwd = NULL,
  lcol = NULL, pty = NULL, pwd = NULL, pcol = NULL,
  CIshade = NULL, CIcol = NULL, legendTitle = NULL,
  legendGroups = NULL)
```

Arguments

...	Objects returned by lpdensity.
alpha	Numeric scalar between 0 and 1, the significance level for plotting confidence regions. If more than one is provided, they will be applied to data series accordingly.

type	String, one of "line" (default), "points" or "both", how the point estimates are plotted. If more than one is provided, they will be applied to data series accordingly.
CItype	String, one of "region" (shaded region, default), "line" (dashed lines), "ebar" (error bars), "all" (all of the previous) or "none" (no confidence region), how the confidence region should be plotted. If more than one is provided, they will be applied to data series accordingly.
title, xlabel, ylabel	Strings, title of the plot and labels for x- and y-axis.
lty	Line type for point estimates, only effective if type is "line" or "both". 1 for solid line, 2 for dashed line, 3 for dotted line. For other options, see the instructions for ggplot2 or par. If more than one is provided, they will be applied to data series accordingly.
lwd	Line width for point estimates, only effective if type is "line" or "both". Should be strictly positive. For other options, see the instructions for ggplot2 or par. If more than one is provided, they will be applied to data series accordingly.
lcol	Line color for point estimates, only effective if type is "line" or "both". 1 for black, 2 for red, 3 for green, 4 for blue. For other options, see the instructions for ggplot2 or par. If more than one is provided, they will be applied to data series accordingly.
pty	Scatter plot type for point estimates, only effective if type is "points" or "both". For options, see the instructions for ggplot2 or par. If more than one is provided, they will be applied to data series accordingly.
pwd	Scatter plot size for point estimates, only effective if type is "points" or "both". Should be strictly positive. If more than one is provided, they will be applied to data series accordingly.
pcol	Scatter plot color for point estimates, only effective if type is "points" or "both". 1 for black, 2 for red, 3 for green, 4 for blue. For other options, see the instructions for ggplot2 or par. If more than one is provided, they will be applied to data series accordingly.
CIshade	Numeric, opaqueness of the confidence region, should be between 0 (transparent) and 1. Default is 0.2. If more than one is provided, they will be applied to data series accordingly.
CIcol	color for confidence region. 1 for black, 2 for red, 3 for green, 4 for blue. For other options, see the instructions for ggplot2 or par. If more than one is provided, they will be applied to data series accordingly.
legendTitle	String, title of legend.
legendGroups	String Vector, group names used in legend.

Value

A stadnard ggplot object is returned, hence can be used for further customization.

Author(s)

Matias D. Cattaneo, Princeton University. <cattaneo@princeton.edu>.

Michael Jansson, University of California Berkeley. <mjansson@econ.berkeley.edu>.

Xinwei Ma (maintainer), University of California San Diego. <x1ma@ucsd.edu>.

References

M.D. Cattaneo, M. Jansson and X. Ma. (2019a). Simple Local Polynomial Density Estimators. *Journal of the American Statistical Association*, forthcoming.

M.D. Cattaneo, M. Jansson and X. Ma. (2019b). lpdensity: Local Polynomial Density Estimation and Inference. Working paper.

See Also

lpdensity

Examples

```
# Generate a random sample
set.seed(42); X <- rnorm(2000)

# Generate a density discontinuity at 0
X <- X - 0.5; X[X>0] <- X[X>0] * 2

# Density estimation, left of 0 (scaled by the relative sample size)
est1 <- lpdensity(data = X[X<=0], grid = seq(-2.5, 0, 0.05), bwselect = "imse-dpi",
  scale = sum(X<=0)/length(X))
# Density estimation, right of 0 (scaled by the relative sample size)
est2 <- lpdensity(data = X[X>0],  grid = seq(0, 2, 0.05), bwselect = "imse-dpi",
  scale = sum(X>0)/length(X))

# Plot
lpdensity.plot(est1, est2, legendTitle="My Plot", legendGroups=c("Left", "Right"))
```

Index

_PACKAGE *(*lpdensity-package*)*, 2

ggplot2, *7*

lpbwdensity, *2*, 2, *4*, *6*
lpdensity, *2*, *4*, 4, *6*, *8*
lpdensity-package, 2
lpdensity.plot, *2*, *4*, *6*, 6

par, *7*