Package 'PDFEstimator'

August 24, 2023

Version 4.5

Date 2023-8-21

Title Multivariate Nonparametric Probability Density Estimator

Author Jenny Farmer <jfarmer@carolina.rr.com> and Donald Jacobs <djacobs1@uncc.ecu>

Maintainer Jenny Farmer <jfarmer@carolina.rr.com>

Description Farmer, J., D. Jacobs (2108) <DOI:10.1371/journal.pone.0196937>. A multivariate nonparametric density estimator based on the maximum-entropy method. Accurately predicts a probability density function (PDF) for random data using a novel iterative scoring function to determine the best fit without overfitting to the sample.

License GPL (>= 2)

Depends plot3D

Imports MultiRNG, methods

NeedsCompilation yes

Repository CRAN

Date/Publication 2023-08-24 07:30:24 UTC

R topics documented:

PDFEstimator-package *Nonparametric Probability Density Estimation and Analysis*

Description

This package provides tools for nonparametric density estimation according to the maximum entropy method described in Farmer and Jacobs (2018). PDFEstimator includes functionality for creating a robust data-driven estimate from a data sample requiring minimal user intervention, thus suitable for high-throughput applications.

Additionally, the package includes advanced plotting and visual diagnostics for confidence thresholding and identification of potentially poorly fitted regions of the estimate. These diagnostics are made available to other density estimation methods through a custom conversion utility, allowing for equitable comparison between estimates.

Details

Main function for estimating the density from a data sample:	`estimatePDF`
Customized plotting function for visual inspection and analysis:	`plot`
Plotting function for densities with 2 variables:	`plot2d`
Plotting function for densities with 3 variables:	`plot3d`
Conversion utility for estimates obtained by other methods:	`convertToPDFe`
Calculation of boundaries for user-defined confidence levels:	`getTarget`
Optional background shading outlining expected variance by position:	`plotBeta`
Utility for additional point approximation for an existing estimate:	`approximatePoints`

Author(s)

Jenny Farmer, University of North Carolina at Charlotte. <jfarmer@carolina.rr.com>.

Donald Jacobs, University of North Carolina at Charlotte. <djacobs1@uncc.edu>.

References

Farmer, J. and D. Jacobs (2018). "High throughput nonparametric probability density estimation." PLoS One 13(5): e0196937. doi: 10.1371/journal.pone.0196937.

`approximatePoints` *Approximate Data Points*

Description

Returns additional point estimates based on an existing estimate.

Usage

```
approximatePoints(estimate, estimationPoints)
```

Arguments

`estimate`	the pdfe object returned from estimatePDF or convertToPDFe
`estimationPoints`	a vector of additional points to estimate.

Details

This method approximates density estimates for the points specified by performing a linear interpolation on an existing probability density function. For a more precise point estimation, call estimatePDF with the estimationPoints argument.

Value

No return value, called for side effects

Author(s)

Jenny Farmer, Donald Jacobs

References

Farmer, J. and D. Jacobs (2018). "High throughput nonparametric probability density estimation." PLoS One 13(5): e0196937.

Examples

```
#Estimates a normal distribution with 1000 sample points using default
# parameters, then prints approximate probability density at points -3, 0, and 1

sampleSize = 1000
sample = rnorm(sampleSize, 0, 1)
dist = estimatePDF(sample)
approximatePoints(dist, c(-3, 0, 1))
```

`convertToPDFe` *Convert to pdfe*

Description

Converts an estimated probability density to a pdfe object type for plotting and analysis utilities within the PDFEstimator package.

Usage

```
convertToPDFe(sample, x, pdf)
```

Arguments

`sample`	original data sample estimated
`x`	estimated points
`pdf`	estimated probability density for each value in x

Details

The plotting functionality available in the PDFEstimator package requires a pdfe object type, generated by the estimatePDF() function. If an alternative estimation method is used, convertToPDFe() will convert it to a pdfe object type. The data sample and the x,y values of the alternative estimate must be provided.

Value

`pdfe`	a pdfe object type.

Author(s)

Jenny Farmer, Donald Jacobs

References

Farmer, J. and D. Jacobs (2018). "High throughput nonparametric probability density estimation." PLoS One 13(5): e0196937.

See Also

estimatePDF, plot.PDFe, lines.PDFe, summary.PDFe, print.PDFe

Examples

```
#Estimates a gamma distribution with 1000 sample points using the density() function
# and converts it to a pdfe object for advanced visual analysis.

sampleSize = 1000
sample = rgamma(sampleSize, shape = 1)
kde = density(sample)
kdeTOpdfe = convertToPDFe(sample, kde$x, kde$y)
plot(kdeTOpdfe, plotPDF = FALSE, plotSQR = TRUE, plotShading = TRUE, showOutlierPercent = 95)
```

`estimatePDF` *Nonparametric Density Estimation*

Description

Estimates the probability density function for a data sample.

Usage

```
estimatePDF(sample, pdfLength = NULL, estimationPoints = NULL,
lowerBound = NULL, upperBound = NULL, target = 70, lagrangeMin = 1,
lagrangeMax = 200, debug = 0, outlierCutoff = 7, smooth = TRUE)
```

Arguments

`sample`	the data sample from which to calculate the density estimate. If the sample has more than 1 column, the multivariate estimation function, estimatePDFmv(), is called instead.
`pdfLength`	the desired length of the estimate returned. Default value is calculated based on sample length. Overriding this calculation can increase or decrease the resolution of the estimate.
`estimationPoints`	a vector containing the points to estimate. If not specified, this is calculated automatically to span the entire sample data.
`lowerBound`	the lower bound of the PDF, if known. Default value is calculated based on the range of the data sample.
`upperBound`	the upper bound of the PDF, if known. Default value is calculated based on the range of the data sample.
`target`	a value from 1 to 100 representing the desired confidence percentage for the estimate score. The default of 70% represents the most likely score based on empirical simulations. A lower value may smooth estimates. A higher value tends to overfit to the sample and is not recommended.
`lagrangeMin`	minimum number of lagrange multipliers
`lagrangeMax`	maximum number of lagrange multipliers
`debug`	verbose output printed to console

outlierCutoff	outliers are automatically detected and removed according to the formula: < Q1 - outlierCutoff * IQR; or > Q3 + outlierCutoff * IQR, where Q1, Q3, and IQR represent the first quartile, third quartile, and inter-quartile range, respectively. Setting outlierCutoff = 0 turns off outlier detection.
smooth	minimizes noise in estimates, particularly in areas of low data density

Details

A nonparametric density estimator based on the maximum-entropy method. Accurately predicts a probability density function (PDF) for random data using a novel iterative scoring function to determine the best fit without overfitting to the sample.

Value

failedSolution	returns true if the pdf calculated is not considered an acceptable estimate of the data according to the scoring function.
threshold	represents the quality of the solution returned. Values of 40 to 70 indicate high confidence in the estimate. Values less than 5 are considered to be of poor quality. For more information on scoring see the referenced publication.
x	estimated range of density data
pdf	estimated probability density function
cdf	estimated cummulative density function
sqr	scaled quantile residual. Provides a sample-size invariant measure of the fluctuations in the estimate.
sqrSize	length of the returned scaled quantile residual. In most cases, this is the size of the input sample. Exceptions are if outliers are detected and/or if the failedSolution flag is true.
lagrange	values of lagrange multipliers. Can be used to reproduce the expansions for an analytical solution.
r	inverse of cdf for the sample.

Author(s)

Jenny Farmer, Donald Jacobs

References

Farmer, J. and D. Jacobs (2018). "High throughput nonparametric probability density estimation." PLoS One 13(5): e0196937.

Examples

```
#Estimates a normal distribution with 1000 sample points using default parameters

sampleSize = 1000
sample = rnorm(sampleSize, 0, 1)
dist = estimatePDF(sample)
```

estimatePDFmv *Multivariate Nonparametric Density Estimation*

Description

Estimates the multivariate probability density function for a data sample containing up to 3 variables.

Usage

```
estimatePDFmv(sample, debug = 0, resolution = NULL)
```

Arguments

sample	data sample from which to calculate the density estimate. Each column of data represents an independent variable.
debug	verbose output printed to console
resolution	grid length of data points for each independent variable.

Details

A multivariate nonparametric density estimator based on the maximum-entropy method. Accurately predicts a probability density function (PDF) for random data for 1, 2, or 3 variables.

Value

x	estimated range of density data
pdf	estimated probability density function

Author(s)

Jenny Farmer, Donald Jacobs

References

Farmer, J. and D. Jacobs (2018). "High throughput nonparametric probability density estimation." PLoS One 13(5): e0196937.

Examples

```
#Estimates a 2-variable normal distribution with 10000 sample points

library(MultiRNG)
nSamples = 5000
cmat = matrix(c(1.0, 0.0, 0.0, 1.0), nrow = 2, ncol = 2)
meanvec = c(0, 0)
sample = draw.d.variate.normal(no.row = nSamples, d = 2,
                               mean.vec = meanvec, cov.mat = cmat)
```

```
mvPDF = estimatePDFmv(sample)
```

`getTarget` *Define Target Outliers*

Description

calculates position-dependent threshold values about the mean according to a beta distribution with parameters k and (n + 1 - k), where k is the position and n is the total number of positions. These beta distributions represent probability per position for sort order statistics for a uniform distribution. This function returns a two-column matrix defining the upper and lower variances of the scaled quantile residual for the target threshold

Usage

```
getTarget(Ns, target)
```

Arguments

`Ns`	number of samples
`target`	target confidence threshold

Details

plotTarget is intended for use with plot.PDFe density estimation objects for plotting scaled quantile residuals, but can be called as a stand-alone user method as well.

Value

`bounds`	a two dimensional matrix defining the upper and lower variance boundaries for the requested target.

Author(s)

Jenny Farmer, Donald Jacobs

References

Farmer, J. and D. Jacobs (2018). "High throughput nonparametric probability density estimation." PLoS One 13(5): e0196937.

See Also

plot.PDFe

Examples

```
#returns boundaries of position-dependent variance calculated for 100 data samples
#  for a threshold of 40%
getTarget(100, 40)
```

`lines.PDFe` *Plot Lines Method for Nonparametric Density Estimation*

Description

The lines method for pdfEstimator objects.

Usage

```
  ## S3 method for class 'PDFe'
lines(x, showOutlierPercent = 0, outlierColor = "red3",
  lwd = 2, ...)
```

Arguments

`x`	an "estimatePDF" object
`showOutlierPercent`	specify confidence threshold for outliers
`outlierColor`	color for outliers positions outside of threshold defined in showOutlierPercent
`lwd`	line width for pdf. If plotPDF = FALSE and plotSQR = TRUE, then the sqr plot uses this line width
`...`	further plotting parameters

Value

No return value, called for side effects

Author(s)

Jenny Farmer, Donald Jacobs

References

Farmer, J. and D. Jacobs (2018). "High throughput nonparametric probability density estimation." PLoS One 13(5): e0196937.

Examples

```
plot(estimatePDF(rnorm(1000, 0, 1)))
lines(estimatePDF(rnorm(1000, 0, 1)), col = "gray")
```

plot.PDFe *Plot Method for Nonparametric Density Estimation*

Description

The plot method for pdfEstimator objects.

Usage

```
  ## S3 method for class 'PDFe'
plot(x, plotPDF = TRUE, plotSQR = FALSE,
  plotShading = FALSE, shadeResolution = 100,
  showOutlierPercent = 0, outlierColor = "red3", sqrPlotThreshold = 2,
  sqrColor = "steelblue4", type="l", lwd = 2, xlab = "x", ylab = "PDF",
                      legendcex = 0.9, ...)
```

Arguments

x	an "estimatePDF" object
plotPDF	plot the probability density function
plotSQR	plot the scaled quantile residual of the estimate
plotShading	plot a gray background shading representing the probability density of the scaled quantile residuals
shadeResolution	the number of sample points plotted in the background if plotShading = TRUE. Increasing resolution will provide sharper contours and take longer to plot.
showOutlierPercent	specify confidence threshold for outliers
outlierColor	color for outliers positions outside of threshold defined in showOutlierPercent
sqrPlotThreshold	magnitude of ylim above and below zero for SQR plot
sqrColor	color for sqr plot for positions within the threshold defined in showOutlierPercentage
type	plot type for pdf. If plotPDF = FALSE and plotSQR = TRUE, then the sqr plot uses this type
lwd	line width for pdf. If plotPDF = FALSE and plotSQR = TRUE, then the sqr plot uses this line width
xlab	x-axis label for pdf. If plotPDF = FALSE and plotSQR = TRUE, then the sqr plot uses this label
ylab	y-axis label for pdf. If plotPDF = FALSE and plotSQR = TRUE, then the sqr plot uses this label
legendcex	expansion factor for legend point size with sqr plot type, for plotPDF = FALSE and plotSQR = TRUE
...	further plotting parameters

Value

No return value, called for side effects

Author(s)

Jenny Farmer, Donald Jacobs

References

Farmer, J. and D. Jacobs (2018). "High throughput nonparametric probability density estimation." PLoS One 13(5): e0196937.

Examples

```
plot(estimatePDF(rnorm(1000, 0, 1)), plotSQR = TRUE, showOutlierPercent = 99)
```

---

plot2d *Plot two-dimensional probability density estimate*

Description

The plot method for two-dimensional pdfEstimator objects.

Usage

```
plot2d(x, xlab = "x", ylab = "y", zlab = "PDF")
```

Arguments

x	an "estimatePDFmv" object
xlab	x-axis label for pdf
ylab	y-axis label for pdf
zlab	z-axis label for pdf

Value

No return value, called for side effects

Author(s)

Jenny Farmer, Donald Jacobs

References

Farmer, J. and D. Jacobs (2018). "High throughput nonparametric probability density estimation." PLoS One 13(5): e0196937.

Examples

```
library(MultiRNG)
nSamples = 10000
cmat = matrix(c(1.0, 0.0, 0.0, 1.0), nrow = 2, ncol = 2)
meanvec = c(0, 0)
sample = draw.d.variate.normal(no.row = nSamples, d = 2,
                               mean.vec = meanvec, cov.mat = cmat)
mvPDF = estimatePDFmv(sample, resolution = 50)

plot2d(mvPDF)
```

`plot3d` *Plot three-dimensional probability density estimate*

Description

The plot method for three-dimensional pdfEstimator objects. Plots two-dimensional cross-sectional slices.

Usage

```
plot3d(x, xs = c(0), ys = c(0), zs = NULL, xlab = "X1", ylab = "X2", zlab = "X3")
```

Arguments

`x`	an "estimatePDFmv" object
`xlab`	x-axis label for pdf
`ylab`	y-axis label for pdf
`zlab`	z-axis label for pdf
`xs, ys, zs`	Vectors or matrices. Vectors specify the positions in x, y or z where the slices (planes) are to be drawn.

Value

No return value, called for side effects

Author(s)

Jenny Farmer, Donald Jacobs

References

Farmer, J. and D. Jacobs (2018). "High throughput nonparametric probability density estimation." PLoS One 13(5): e0196937.

plotBeta *Plot Diagnostic Shading*

Description

Plot background shading for density estimation based on the beta distribution for sort order statistics

Usage

```
plotBeta(samples, resolution = 100, xPlotRange, sqrPlotThreshold = 2)
```

Arguments

samples	a data sample for estimation
resolution	the number of sample points plotted in the contour
xPlotRange	the x-axis range for plotting
sqrPlotThreshold	magnitude of ylim above and below zero

Details

plotBeta is intended for use with the plot method in the PDFEstimator package for plotting pdfe density estimation objects.

Value

No return value, called for side effects

Author(s)

Jenny Farmer, Donald Jacobs

References

Farmer, J. and D. Jacobs (2018). "High throughput nonparametric probability density estimation." PLoS One 13(5): e0196937.

See Also

plot.PDFe

Index