Package 'lookout'

July 21, 2026

Type Package

Title Leave One Out Kernel Density Estimates for Outlier Detection

Version 2.0.2

Maintainer Sevvandi Kandanaarachchi <sevvandik@gmail.com>

Description Outlier detection using leave-one-out kernel density estimates and extreme value theory. The bandwidth for kernel density estimates is computed using persistent homology, a technique in topological data analysis. Using peak-over-threshold method, a generalized Pareto distribution is fitted to the log of leave-one-out kde values to identify outliers.

License GPL-3

Encoding UTF-8

BugReports https://github.com/sevvandi/lookout/issues

Imports evd, ggplot2, mlpack, RANN, robustbase, stats, tidyr

Suggests knitr, rmarkdown, testthat (>= 3.0.0)

URL https://sevvandi.github.io/lookout/,
https://github.com/sevvandi/lookout

Config/testthat/edition 3

Config/roxygen2/version 8.0.0

NeedsCompilation no

Author Sevvandi Kandanaarachchi [aut, cre] (ORCID: <https://orcid.org/0000-0002-0337-0395>), Rob Hyndman [aut] (ORCID: <https://orcid.org/0000-0002-2140-5352>), Chris Fraley [ctb]

Repository CRAN

Date/Publication 2026-07-21 08:30:19 UTC

Contents

autoplot.lookoutliers 2
autoplot.persistingoutliers 3
find_tda_bw 4
lookout 5
lookout_ts 6
mvscale 7
persisting_outliers 9

Index **11**

`autoplot.lookoutliers` *Plots outliers identified by lookout algorithm.*

Description

Scatterplot of two columns from the data set with outliers highlighted.

Usage

```
## S3 method for class 'lookoutliers'
autoplot(object, columns = 1:2, ...)
```

Arguments

`object`	The output of the function `lookout`.
`columns`	Which columns of the original data to plot (specified as either numbers or strings)
`...`	Other arguments currently ignored.

Value

A ggplot object.

Examples

```
X <- rbind(
  data.frame(
    x = rnorm(500),
    y = rnorm(500)
  ),
  data.frame(
    x = rnorm(5, mean = 10, sd = 0.2),
    y = rnorm(5, mean = 10, sd = 0.2)
  )
)
lo <- lookout(X)
autoplot(lo)
```

`autoplot.persistingoutliers` *Plots outlier persistence for a range of significance levels.*

Description

This function plots outlier persistence for a range of significance levels using the algorithm lookout, an outlier detection method that uses leave-one-out kernel density estimates and generalized Pareto distributions to find outliers.

Usage

```
## S3 method for class 'persistingoutliers'
autoplot(object, alpha = object$alpha, ...)
```

Arguments

`object`	The output of the function `persisting_outliers`.
`alpha`	The significance levels to plot.
`...`	Other arguments currently ignored.

Value

A ggplot object.

Examples

```
X <- rbind(
  data.frame(
    x = rnorm(500),
    y = rnorm(500)
  ),
  data.frame(
    x = rnorm(5, mean = 10, sd = 0.2),
    y = rnorm(5, mean = 10, sd = 0.2)
  )
)
plot(X, pch = 19)
outliers <- persisting_outliers(X, scale = FALSE)
autoplot(outliers)
```

find_tda_bw *Identifies bandwidth for outlier detection.*

Description

This function identifies the bandwidth that is used in the kernel density estimate computation. The function uses topological data analysis (TDA) to find the badnwidth.

Usage

```
find_tda_bw(X, fast = TRUE, gamma = 0.98, use_differences = FALSE)
```

Arguments

X	The numerical input data in a data.frame, matrix or tibble format.
fast	If TRUE (default), makes the computation faster by sub-setting the data for the bandwidth calculation.
gamma	Parameter for bandwidth calculation giving the quantile of the Rips death radii to use for the bandwidth. Default is 0.98. Ignored under the old version; where the lower limit of the maximum Rips death radii difference is used. Also ignored if bw is provided.
use_differences	If TRUE, the bandwidth is set to the lower point of the maximum Rips death radii differences. If FALSE, the gamma quantile of the Rips death radii is used. Default is FALSE.

Value

The bandwidth

Examples

```
X <- rbind(
  data.frame(
    x = rnorm(500),
    y = rnorm(500)
  ),
  data.frame(
    x = rnorm(5, mean = 10, sd = 0.2),
    y = rnorm(5, mean = 10, sd = 0.2)
  )
)
find_tda_bw(X, fast = TRUE)
```

lookout *Identifies outliers using the algorithm lookout.*

Description

This function identifies outliers using the algorithm lookout, an outlier detection method that uses leave-one-out kernel density estimates and generalized Pareto distributions to find outliers.

Usage

```
lookout(
  X,
  alpha = 0.01,
  beta = 0.9,
  gamma = 0.98,
  bw = NULL,
  gpd = NULL,
  scale = TRUE,
  fast = NROW(X) > 1e+05,
  old_version = FALSE
)
```

Arguments

X	The numerical input data in a data.frame, matrix or tibble format.
alpha	The level of significance. Default is 0.01. So there is a 1/100 chance of any point being falsely classified as an outlier.
beta	The quantile threshold used in the GPD estimation. Default is 0.90. To ensure there is enough data available, values greater than 0.90 are set to 0.90.
gamma	Parameter for bandwidth calculation giving the quantile of the Rips death radii to use for the bandwidth. Default is 0.98. Ignored under the old version; where the lower limit of the maximum Rips death radii difference is used. Also ignored if bw is provided.
bw	Bandwidth parameter. If NULL (default), the bandwidth is found using Persistent Homology.
gpd	Generalized Pareto distribution parameters. If NULL (the default), these are estimated from the data.
scale	If TRUE, the data is standardized. Using the old version, unit scaling is applied so that each column is in the range [0,1]. Under the new version, robust rotation and scaling is used so that the columns are approximately uncorrelated with unit variance. Default is TRUE.
fast	If TRUE (default), makes the computation faster by sub-setting the data for the bandwidth calculation.
old_version	Logical indicator of which version of the algorithm to use. Default is FALSE, meaning the newer version is used.

Value

A list with the following components:

`outliers`	The set of outliers.
`outlier_probability`	The GPD probability of the data.
`outlier_scores`	The outlier scores of the data.
`bandwidth`	The bandwdith selected using persistent homology.
`kde`	The kernel density estimate values.
`lookde`	The leave-one-out kde values.
`gpd`	The fitted GPD parameters.

References

Kandanaarachchi, S, and Hyndman, RJ (2022) Leave-one-out kernel density estimates for outlier detection, *J Computational & Graphical Statistics*, **31**(2), 586-599. https://robjhyndman.com/publications/lookout/.

Hyndman, RJ, Kandanaarachchi, S, and Turner, K (2026) When lookout meets crackle: Anomaly detection using kernel density estimation, in preparation. https://robjhyndman.com/publications/lookout2.html

Examples

```
X <- rbind(
  data.frame(
    x = rnorm(500),
    y = rnorm(500)
  ),
  data.frame(
    x = rnorm(5, mean = 10, sd = 0.2),
    y = rnorm(5, mean = 10, sd = 0.2)
  )
)
lo <- lookout(X)
lo
autoplot(lo)
```

`lookout_ts` *Identifies outliers in univariate time series using the algorithm lookout.*

Description

This is the time series implementation of lookout which identifies outliers in the double differenced time series.

Usage

```
lookout_ts(x, scale = FALSE, ...)
```

Arguments

x	The input univariate time series.
scale	If TRUE, the data is standardized. Using the old version, unit scaling is applied so that each column is in the range [0,1]. Under the new version, robust rotation and scaling is used so that the columns are approximately uncorrelated with unit variance. Default is TRUE.
...	Other arguments are passed to lookout.

Value

A lookout object.

See Also

lookout

Examples

```
set.seed(1)
x <- arima.sim(list(order = c(1, 1, 0), ar = 0.8), n = 200)
x[50] <- x[50] + 10
plot(x)
lo <- lookout_ts(x)
lo
```

---

mvscale *Compute robust multivariate scaled data*

Description

A multivariate version of base::scale(), that takes account of the covariance matrix of the data, and uses robust estimates of center, scale and covariance by default. The centers are removed using medians, the scale function is the IQR, and the covariance matrix is estimated using a robust OGK estimate. The data are scaled using the Cholesky decomposition of the inverse covariance. Then the scaled data are returned.

Usage

```
mvscale(
  object,
  center = stats::median,
  scale = robustbase::s_Qn,
  cov = robustbase::covOGK,
  warning = TRUE
)
```

Arguments

object	A vector, matrix, or data frame containing some numerical data.
center	A function to compute the center of each numerical variable. Set to NULL if no centering is required.
scale	A function to scale each numerical variable. When cov = robustbase::covOGK(), it is passed as the sigmamu argument.
cov	A function to compute the covariance matrix. Set to NULL if no rotation required.
warning	Should a warning be issued if non-numeric columns are ignored?

Details

Optionally, the centering and scaling can be done for each variable separately, so there is no rotation of the data, by setting cov = NULL. Also optionally, non-robust methods can be used by specifying center = mean, scale = stats::sd(), and cov = stats::cov(). Any non-numeric columns are retained with a warning.

Value

A vector, matrix or data frame of the same size and class as object, but with numerical variables replaced by scaled versions.

Author(s)

Rob J Hyndman

See Also

base::scale(), stats::sd(), stats::cov(), robustbase::covOGK(), robustbase::s_Qn()

Examples

```
# Univariate z-scores (no rotation)
z <- mvscale(faithful, center = mean, scale = sd, cov = NULL, warning = FALSE)
# Non-robust scaling with rotation
z <- mvscale(faithful, center = mean, cov = stats::cov, warning = FALSE)
# Robust scaling and rotation
z <- mvscale(faithful, warning = FALSE)
```

persisting_outliers *Computes outlier persistence for a range of significance values.*

Description

This function computes outlier persistence for a range of significance values, using the algorithm lookout, an outlier detection method that uses leave-one-out kernel density estimates and generalized Pareto distributions to find outliers.

Usage

```
persisting_outliers(
  X,
  alpha = seq(0.01, 0.1, by = 0.01),
  st_qq = 0.9,
  scale = TRUE,
  num_steps = 20,
  old_version = FALSE
)
```

Arguments

X	The input data in a matrix, data.frame, or tibble format. All columns should be numeric.
alpha	Grid of significance levels.
st_qq	The starting quantile for death radii sequence. This will be used to compute the starting bandwidth value.
scale	If TRUE, the data is scaled. Default is TRUE. Which scaling method is used depends on the old_version parameter. See lookout for details.
num_steps	The length of the bandwidth sequence.
old_version	Logical indicator of which version of the algorithm to use.

Value

A list with the following components:

out	A 3D array of N x num_steps x num_alpha where N denotes the number of observations, num_steps denote the length of the bandwidth sequence, and num_alpha denotes the number of significance levels. This is a binary array and the entries are set to 1 if that observation is an outlier for that particular bandwidth and significance level.
bw	The set of bandwidth values.
gpdparas	The GPD parameters used.
lookoutbw	The bandwidth chosen by the algorithm lookout using persistent homology.

Examples

```
X <- rbind(
  data.frame(
    x = rnorm(500),
    y = rnorm(500)
  ),
  data.frame(
    x = rnorm(5, mean = 10, sd = 0.2),
    y = rnorm(5, mean = 10, sd = 0.2)
  )
)
plot(X, pch = 19)
outliers <- persisting_outliers(X, scale = FALSE)
outliers
autoplot(outliers)
```

Index

autoplot.lookoutliers, 2
autoplot.persistingoutliers, 3

base::scale(), *7*, *8*

find_tda_bw, 4

lookout, 5, *7*, *9*
lookout_ts, 6

mvscale, 7

persisting_outliers, 9

robustbase::covOGK(), *8*
robustbase::s_Qn(), *8*

stats::cov(), *8*
stats::sd(), *8*