Package: PCObw (via r-universe)

September 8, 2024

Type Package

Title Bandwidth Selector with Penalized Comparison to Overfitting Criterion

Version 0.0.1

Date 2023-03-16

Author S. Varet

Maintainer S. Varet <suzanne.varet@universite-paris-saclay.fr>

Description Bandwidth selector according to the Penalised Comparison to Overfitting (P.C.O.) criterion as described in Varet, S., Lacour, C., Massart, P., Rivoirard, V., (2019) <https://hal.archives-ouvertes.fr/hal-02002275>. It can be used with univariate and multivariate data.

Depends R (>= 3.5.0)

License GPL (>= 2)

Imports Rcpp (>= 0.12.17), stats(>= 3.5.2)

LinkingTo Rcpp, RcppEigen

RoxygenNote 7.2.2

Suggests knitr, rmarkdown, mvtnorm

VignetteBuilder knitr

Encoding UTF-8

NeedsCompilation yes

Repository CRAN

Date/Publication 2023-03-23 22:32:13 UTC

Contents

PCObw-package *Bandwidth Selector with Penalized Comparison to Overfitting Criterion*

Description

Bandwidth selector according to the Penalised Comparison to Overfitting (P.C.O.) criterion as described in Varet, S., Lacour, C., Massart, P., Rivoirard, V., (2019). It can be used with univariate and multivariate data.

Details

bw.L2PCO(x_i, ...)

bw.L2PCO.diag(x_i, ...)

select the optimal bandwith according to PCO criterion where x_i are the data (a numeric matrix or a numeric vector).

Author(s)

S. Varet.

Maintainer: S. Varet <suzanne.varet@universite-paris-saclay.fr>

References

Varet, S., Lacour, C., Massart, P., Rivoirard, V., (2019). *Numerical performance of Penalized Comparison to Overfitting for multivariate kernel density estimation*. hal-02002275. https://hal.archives-ouvertes.fr/hal-02002275

Examples

```
# load univariate data
data("gauss_1D_sample")

# computes the optimal bandwith for the sample x_i with all parameters set to their default value
bw.L2PCO(gauss_1D_sample)
```

bw.L2PCO *Compute the full PCO bandwith*

Description

bw.L2PCO tries to minimise the PCO criterion (described and studied in Varet, S., Lacour, C., Massart, P., Rivoirard, V., (2019)) with a gold section search. For multivariate data, it searches for a full matrix.

Usage

```
bw.L2PCO(
  x_i,
  nh = 40,
  K_name = "gaussian",
  binning = FALSE,
  nb = 32,
  tol = 1e-06,
  adapt_nb_bin = FALSE,
  nb_bin_vect = NULL
)
```

Arguments

x_i	the observations. Must be a matrix with d column and n lines (d the dimension and n the sample size)
nh	the maximum number of PCO criterion evaluations during the golden section search. The default value is 40. The golden section search stop once this value is reached or if the tolerance is achieved, and return the middle of the interval.
K_name	name of the kernel. Can be 'gaussian', 'epanechnikov', or 'biweight'. The default value is 'gaussian'.
binning	default value is FALSE, that is the function computes the exact PCO criterion. If set to TRUE allows to use binning.
nb	is the number of bins to use when binning is TRUE. For multivariate x_i, nb corresponds to the number of bins per dimension. The default value is 32.
tol	is the maximum authorized length of the interval which contains the optimal h for univariate data. For multivariate data, it corresponds to the length of each hypercube axe. The golden section search stop once this value is achieved or when nh is reached and return the middle of the interval. Its default value is 10^(-6).
adapt_nb_bin	is a boolean used for univariate x_i. If set to TRUE, authorises the function to increase the number of bins if, with nb bins, the middle of the initial interval is not an admissible solution, that is if the criterion at the middle is greater than the mean of the criterion at the bounds of the initial interval of search.
nb_bin_vect	can be set to have a different number of bins on each dimension

Value

a scalar for univariate data or a matrix for multivariate data corresponding to the optimal bandwidth according to the PCO criterion

References

Varet, S., Lacour, C., Massart, P., Rivoirard, V., (2019). *Numerical performance of Penalized Comparison to Overfitting for multivariate kernel density estimation*. hal-02002275. https://hal.archives-ouvertes.fr/hal-02002275

See Also

[stats::nrd0()], [stats::nrd()], [stats::ucv()], [stats::bcv()] and [stats::SJ()] for other univariate bandwidth selection and [stats::density()] to compute the associated density estimation.

[ks::Hlscv()], [ks::Hbcv()], [ks::ns()] for other multivariate bandwidth selection.

Examples

```
# an example with simulated univariate data

# load univariate data
data("gauss_1D_sample")

# computes the optimal bandwith for the sample x_i with all parameters set to their default value
bw.L2PCO(gauss_1D_sample)


# an example with simulated multivariate data

# load multivariate data
data("gauss_mD_sample")

# computes the optimal bandwith for the sample x_i with all parameters set to their default value
# generates a warning since the tolerance value is not reached
bw.L2PCO(gauss_mD_sample)

# To avoid this warning, it is possible to increase the parameter nh
bw.L2PCO(gauss_mD_sample, nh = 80)
```

---

bw.L2PCO.diag *Compute the diagonal PCO bandwith*

Description

`bw.L2PCO.diag` tries to minimise the PCO criterion (described and studied in Varet, S., Lacour, C., Massart, P., Rivoirard, V., (2019)) with a gold section search. For multivariate data, it searches for a diagonal matrix.

Usage

```
bw.L2PCO.diag(
  x_i,
  nh = 40,
  K_name = "gaussian",
  binning = FALSE,
  nb = 32,
  tol = 1e-06,
  adapt_nb_bin = FALSE,
  nb_bin_vect = NULL
)
```

Arguments

x_i	the observations. Must be a matrix with d column and n lines (d the dimension and n the sample size)
nh	the maximum of possible bandwiths tested. The default value is 40.
K_name	name of the kernel. Can be 'gaussian', 'epanechnikov', or 'biweight'. The default value is 'gaussian'.
binning	can be set to TRUE or FALSE. The value TRUE allows to use binning. The default value FALSE computes the exact PCO criterion.
nb	is the number of bins to use when binning is TRUE. For multivariate x_i, nb corresponds to the number of bins per dimension.
tol	is the maximum authorized length of the interval which contains the optimal h for univariate data. For multivariate data, it corresponds to the length of each hypercube axe. The golden section search stop once this value is achieved or when nh is reached and return the middle of the interval. Its default value is 10^(-6).
adapt_nb_bin	is a boolean used for univariate x_i. If set to TRUE, authorises the function to increase the number of bins if, with nb bins, the middle of the initial interval is not an admissible solution, that is if the criterion at the middle is greater than the mean of the criterion at the bounds of the initial interval of search.
nb_bin_vect	can be set to have a different number of bins on each dimension

Value

a scalar for univariate data or a vector (the diagonal of the matrix) for multivariate data corresponding to the optimal bandwidth according to the PCO criterion

References

Varet, S., Lacour, C., Massart, P., Rivoirard, V., (2019). *Numerical performance of Penalized Comparison to Overfitting for multivariate kernel density estimation*. hal-02002275. https://hal.archives-ouvertes.fr/hal-02002275

See Also

[stats::nrd0()], [stats::nrd()], [stats::ucv()], [stats::bcv()] and [stats::SJ()] for other univariate bandwidth selection and [stats::density()] to compute the associated density estimation.

[ks::Hlscv.diag()], [ks::Hbcv.diag()], [ks::ns.diag()] for other multivariate bandwidth selection.

Examples

```
# an example with simulated univariate data

# load univariate data
data("gauss_1D_sample")

# computes the optimal bandwith for the sample x_i with all parameters set to their default value
bw.L2PCO.diag(gauss_1D_sample)


# an example with simulated multivariate data

# load multivariate data
data("gauss_mD_sample")

# computes the optimal bandwith for the sample x_i with all parameters set to their default value
# generates a warning since the tolerance value is not reached
bw.L2PCO.diag(gauss_mD_sample)

# To avoid this warning, it is possible to increase the parameter nh
bw.L2PCO.diag(gauss_mD_sample, nh = 80)
```

---

gauss_1D_sample *Univariate sample*

Description

A univariate sample of 100 realisations of a gaussian law with mean 0 and standard deviation 1

Usage

```
data("gauss_1D_sample")
```

Format

A vector with 100 rows

gauss_mD_sample *Multivariate sample*

Description

A 2D sample of 100 realisations of a gaussian law with mean (0, 0) and covariance matrix | 1 0.9 | | 0.9 1 |

Usage

```
data("gauss_mD_sample")
```

Format

A matrix with 100 rows and 2 columns

Index