Package 'LPBkg'

October 5, 2019

Type Package

Title Detecting New Signals under Background Mismodelling

Version 1.2

Author Sara Algeri <salgeri@umn.edu>, Haoran Liu<liu00728@umn.edu>

Maintainer Sara Algeri <salgeri@umn.edu>

Description Given a postulated model and a set of data, the comparison density is estimated and the deviance test is implemented in order to assess if the data distribution deviates significantly from the postulated model. Finally, the results are summarized in a CD-plot as described in Algeri S. (2019) <arXiv:1906.06615>.

Depends R (>= 2.0.1), polynom

Imports orthopolynom,Hmisc,grDevices,graphics,stats

Encoding UTF-8

License GPL-3

LazyData true

NeedsCompilation no

Repository CRAN

Date/Publication 2019-10-04 23:40:02 UTC

R topics documented:

`BestM` *Chooses the size of the polynomial basis*

Description

Computes the deviance p-values considering different sizes of the polynomial basis and selects the one for which the deviance p-value is the smallest.

Usage

```
BestM(data, g, Mmax = 20, range = c(min(data),max(data)))
```

Arguments

`data`	A vector of data. See details.
`g`	The postulated model from which we want to assess if deviations occur.
`Mmax`	The maximum size of the polynomial basis from which a suitable value `M` is selected (the default is 20). See details.
`range`	Range of the data/ search region considered.

Details

The argument `data` collects the data for which we want to test if deviations occur from the postulated model specified in the argument `g`. As in Algeri, 2019, the sample specified under `data` corresponds to the source-free sample in the background calibration phase and to the physics sample in the signal search phase. The value `M` selected determines the smoothness of the estimated comparison density, with smaller values of M leading to smoother estimates. The deviance test is used to select the value `M` which leads to the most significant deviation from the postulated model. The default value for `Mmax` is set to 20. Notice that numerical issues may arise for larger values of `Mmax`.

Value

`pvals`	The deviance test p-value obtained for each values of `M` (from 1 to `Mmax`) considered.
`minp`	The minimum value of the deviance p-values observed.
`Msel`	The value of `M` at which the minimum deviance p-values is achieved.

Author(s)

Sara Algeri

References

S. Algeri, 2019. Detecting new signals under background mismodelling <arXiv:1906.06615>.

See Also

denoise, dhatL2.

Examples

```
#Generating data
x<-rnorm(1000,10,7)
data<-x[x>=10 & x<=20]

#Create suitable postulated quantile function of data
G<-pnorm(20,5,15)-pnorm(10,5,15)
g<-function(x){dnorm(x,5,15)/G}

Mmax=10
range=c(10,20)

BestM(data,g,Mmax,range)
```

---

c_alpha2 *Approximated quantiles*

Description

Approximates the quantiles of the supremum of the comparison density estimator using tube formulae and assuming that H_0 is true.

Usage

```
c_alpha2(M, IDs, alpha = 0.05, c_interval = c(1, 10))
```

Arguments

M	The size of the polynomial basis used to estimate the comparison density.
IDs	The IDs of the polynomial terms to be used out of the M considered.
alpha	Desired significance level.
c_interval	Lower and upper bounds for the quantile being computed.

Value

Approximated quantile of order 1-alpha of the supremum of the comparison density estimator.

Author(s)

Sara Algeri

References

S. Algeri, 2019. Detecting new signals under background mismodelling <arXiv:1906.06615>.

L.A. Wasserman, 2005. All of Nonparametric Statistics. Springer Texts in Statistics.

See Also

dhatL2.

Examples

```
c_alpha2(5, c(2,4), alpha = 0.05, c_interval = c(1, 10))
```

---

denoise *Coefficients of the denoised comparison density estimator*

Description

Selects the largest coefficients according to the AIC or BIC criterion.

Usage

```
denoise(LP, n, method)
```

Arguments

LP	Original vector of coefficients estimates. See details.
n	The dimension of the sample on which the estimates in LP have been obtained.
method	Either "AIC" or "BIC". See details.

Details

Give a vector of M coefficient estimates, the largest is selected according to the AIC or BIC criterion as described in Algeri, 2019 and Mukhopadhyay, 2017.

Value

Selected coefficient estimates.

Author(s)

Sara Algeri

References

S. Algeri, 2019. Detecting new signals under background mismodelling. <arXiv:1906.06615>.

S. Mukhopadhyay, 2017. Large-scale mode identification and data-driven sciences. Electronic Journal of Statistics 11 (2017), no. 1, 215–240.

See Also

Legj.

Examples

```
#generating data
x<-rnorm(1000,10,7)
xx<-x[x>=10 & x<=20]

#create suitable postulated quantile function
G<-pnorm(20,5,15)-pnorm(10,5,15)
g<-function(x){dnorm(x,5,15)/G}

#Vectorize quantile function
g<-Vectorize(g)
u<-g(xx)

Mmax=20
S<- as.matrix(Legj(u=u,m=Mmax))
n<-length(u)

LP <- apply(S,FUN="mean",2)

denoise(LP,n=n,method="AIC")
```

---

dhatL2 *CD-plot and adjusted deviance test*

Description

Construction of CD-plot and adjusted deviance test. The confidence bands are also adjusted for post-selection inference.

Usage

```
dhatL2(data, g, M = 6, Mmax = NULL, smooth = TRUE,
  criterion = "AIC", hist.u = TRUE, breaks = 20, ylim = c(0, 2.5),
  range = c(min(data),max(data)), sigma = 2)
```

Arguments

data	A vector of data. See details.
g	The postulated model from which we want to assess if deviations occur.
M	The desired size of the polynomial basis to be used.
Mmax	The maximum size of the polynomial basis from which M was selected (the default is 20). See details.

smooth	A logical argument indicating if a denoised solution should be implemented. The default is FALSE, meaning that the full solution should be implemented. See details.
criterion	If smooth=TRUE, the criterion with respect to which the denoising process should be implemented. The two possibilities are "AIC" or "BIC". See details.
hist.u	A logical argument indicating if the CD-plot should be displayed or not. The default is TRUE.
breaks	If hist.u=TRUE, the number of breaks of the CD-plot. The default is 20.
ylim	If hist.u=TRUE, the range of the y-axis of the CD-plot.
range	Range of the data/search region considered.
sigma	The significance level (in sigmas) with respect to which the confidence bands should be constructed. See details.

Details

The argument data collects the data for which we want to test if its distribution deviates from the one of the postulated model specified in the argument g. In Algeri, 2019, the sample specified under data corresponds to the source-free sample in the background calibration phase and to the physics sample in the signal search phase. The value M selected determines the smoothness of the estimated comparison density, with smaller values of M leading to smoother estimates. The deviance test is used to select the value M which leads to the most significant deviation from the postulated model. The default value for Mmax is set to 20. Notice that numerical issues may arise for larger values of Mmax. If smooth=TRUE the largest coefficient estimates are selected according to either the AIC or BIC criterion as described in Algeri, 2019 and Mukhopadhyay, 2017. If Mmax>1 and/or smooth=TRUE, post-selection Bonferroni's correction is automatically implemented to both the deviance test p-value and the confidence bands. The desired level of significance can be expressed as one minus the cdf of a standard normal evaluated at sigma (see Algeri, 2019).

Value

Deviance	Value of the deviance test statistic.
Dev_pvalue	Unadjusted p-value of the deviance test.
Dev_adj_pvalue	Post-selection Bonferroni adjusted p-value of the deviance test.
kstar	Number of coefficients selected by the denoising process. If smooth=FALSE, kstar=M.
dhat	Function corresponding to the estimated comparison density in the u domain.
dhat.x	Function corresponding to the estimated comparison density in the x domain.
SE	Function corresponding to the estimated standard errors of the comparison density in the u domain.
LBf1	Function corresponding to the lower bound of the confidence bands under in u domain.
UBf1	Function corresponding to the upper bound of the confidence bands in u domain.
f	Function corresponding to the estimated density of the data.

u	Vector of values corresponding to the cdf of the model specified in g evaluated at the vector data.
LP	Estimates of the coefficients.
G	Cumulative density function of the postulated model specified in the argument g.

Author(s)

Sara Algeri

References

S. Algeri, 2019. Detecting new signals under background mismodelling. <arXiv:1906.06615>.

S. Mukhopadhyay, 2017. Large-scale mode identification and data-driven sciences. Electronic Journal of Statistics 11 (2017), no. 1, 215–240.

See Also

Legj,BestM,denoise.

Examples

```
#generaing data
x<-rnorm(1000,10,7)
xx<-x[x>=10 & x<=20]

#create suitable postulated quantile function of data
G<-pnorm(20,5,15)-pnorm(10,5,15)
g<-function(x){dnorm(x,5,15)/G}

#Choose best M
Mmax=20
range=c(10,20)
m<-BestM(data=xx,g, Mmax,range)

# vectorize postulated quantile function
g<-Vectorize(g)
u<-g(xx)

#M has to be sufficient big, otherwise dhatL2 function will crush.
#So,here we set m eqaul 6 as an example
m<-6
comp.density<-dhatL2(data=xx,g, M=m, Mmax=Mmax,smooth=FALSE,criterion="AIC",hist.u=TRUE,breaks=20,
         ylim=c(0,2.5),range=range,sigma=2)
```

Legj *Evaluation of normalized shifted Legendre polynomials*

Description

Evaluates the a basis of normalized shifted Legendre polynomials over a specified data vector.

Usage

```
Legj(u, m)
```

Arguments

u	Data vector on which the polynomials are to be evaluated.
m	The size of the basis to be considered.

Value

Numerical values of the first m normalized shifted Legendre polynomials.

Examples

```
x<-rnorm(1000,10,7)
xx<-x[x>=10 & x<=20]
G<-pnorm(20,5,15)-pnorm(10,5,15)
g<-function(x){dnorm(x,5,15)/G}
g<-Vectorize(g)
u<-g(xx)
Mmax=20
s<-as.matrix(Legj(u,Mmax))
```

Index