Package 'LPRelevance'

April 23, 2021

Type Package

Title Relevance-Integrated Statistical Inference Engine

Version 3.2

Date 2021-04-08

Author Subhadeep Mukhopadhyay, Kaijun Wang

Maintainer Kaijun Wang <kaijunwang.19@gmail.com>

Description Provide methods to perform customized inference at individual level by taking contextual covariates into account. Three main functions are provided in this package: (i) LASER(): it generates specially-designed artificial relevant samples for a given case; (ii) g2l.proc(): computes customized fdr(z|x); and (iii) rEB.proc(): performs empirical Bayes inference based on LASERs. The details can be found in Mukhopadhyay, S., and Wang, K (2021, <arXiv:2004.09588>).

Imports leaps,locfdr,Bolstad2,reshape2,ggplot2,polynom,glmnet,caret

Depends R (>= 3.5.0), stats, BayesGOF, MASS

License GPL-2

NeedsCompilation no

Repository CRAN

Date/Publication 2021-04-22 22:10:02 UTC

R topics documented:

LPRelevance-package *Relevance-Integrated Statistical Inference Engine*

Description

How to individualize a global inference method? The goal of this package is to provide a systematic recipe for converting classical global inference algorithms into customized ones. It provides methods that perform individual level inferences by taking contextual covariates into account. At the heart of our solution is the concept of "artificially-designed relevant samples", called LASERs– which pave the way to construct an inference mechanism that is simultaneously efficiently estimable and contextually relevant, thus works at both macroscopic (overall simultaneous) and microscopic (individual-level) scale.

Author(s)

Subhadeep Mukhopadhyay, Kaijun Wang

Maintainer: Kaijun Wang <kaijunwang.19@gmail.com>

References

Mukhopadhyay, S., and Wang, K (2021) "On The Problem of Relevance in Statistical Inference". <arXiv:2004.09588>

data.dti *DTI data.*

Description

A diffusion tensor imaging study comparing brain activity of six dyslexic children versus six normal controls. Two-sample tests produced z-values at $N = 15443$ voxels (3-dimensional brain locations), with each $z_i \sim N(0, 1)$ under the null hypothesis of no difference between the dyslexic and normal children.

Usage

```
data(data.dti)
```

Format

A data frame with 15443 observations on the following 4 variables.

`coordx` A list of x coordinates

`coordy` A list of y coordinates

`coordz` A list of z coordinates

`z` The z-values.

Source

http://statweb.stanford.edu/~ckirby/brad/LSI/datasets-and-programs/datasets.html

References

Efron, B. (2012). "Large-scale inference: empirical Bayes methods for estimation, testing, and prediction". Cambridge University Press.

---

funnel *A stylized simulated example.*

Description

A large-scale heterogeneous dataset used in our paper.

Usage

```
data("funnel")
```

Format

A data frame with 3565 observations on the following 3 variables.

`x` A list of covariate values.

`z` A list of z-values.

`tags` Binary vector of labels, 1 indicates a data point is a signal.

References

Mukhopadhyay, S., and Wang, K (2021) "On The Problem of Relevance in Statistical Inference". <arXiv:2004.09588>

---

g2l.proc *Procedures for global and local inference.*

Description

This function performs customized fdr analyses tailored to each individual cases.

Usage

```
g2l.proc(X, z, X.target = NULL, z.target = NULL, m = c(4, 6), alpha = 0.1,
nbag = NULL, nsample = length(z), lp.reg.method = "lm",
null.scale = "QQ", approx.method = "direct", ngrid = 2000,
centering = TRUE, coef.smooth = "BIC", fdr.method = "locfdr",
plot = TRUE, rel.null = "custom", locfdr.df = 10,
fdr.th.fixed = NULL, parallel = FALSE, ...)
```

Arguments

X	A n-by-d matrix of covariate values
z	A length n vector containing observations of z values.
X.target	A k-by-d matrix providing k sets of covariates for target cases to investigate. Set to NULL to investigate all cases and provide global inference results.
z.target	A vector of length k, providing the target z values to investigate
m	An ordered pair. First number indicates how many LP-nonparametric basis to construct for each X, second number indicates how many to construct for z. Default: m=c(4,6).
alpha	Confidence level for determining signals.
nbag	Number of bags of parametric bootstrapped samples to use for each target case, each time a new set of relevance samples will be generated for analysis, and the resulting fdr curves are aggregated together by taking the mean values. Set to NULL to disable.
nsample	Number of relevance samples generated for each case. The default is the size of the input z-statistic.
lp.reg.method	Method for estimating the relevance function and its conditional LP-Fourier coefficients. We currently support three options: lm (inbuilt with subset selection), glmnet, and knn.
null.scale	Method of estimating null standard deviation from the laser samples. Available options: "IQR", "QQ" and "locfdr"
approx.method	Method used to approximate customized fdr curve, default is "direct".When set to "indirect", the customized fdr is computed by modifying pooled fdr using relevant density function.
ngrid	Number of gridpoints to use for computing customized fdr curve.
centering	Whether to perform regression-adjustment to center the data, default is TRUE.
coef.smooth	Specifies the method to use for LP coefficient smoothing (AIC or BIC). Uses BIC by default.
fdr.method	Method for controlling false discoveries (either "locfdr" or "BH"), default choice is "locfdr".
plot	Whether to include plots in the results, default is TRUE.
rel.null	How the relevant null changes with x: "custom" denotes we allow it to vary with x, and "th" denotes fixed.
locfdr.df	Degrees of freedom to use for locfdr()
fdr.th.fixed	Use fixed fdr threshold for finding signals. Default set to NULL, which finds different thresholds for different cases.
parallel	Use parallel computing for obtaining the relevance samples, mainly used for very huge nsample, default is FALSE.
...	Extra parameters to pass to other functions. Currently only supports the arguments for knn().

Value

A list containing the following items:

`macro`	Available when `X.target` set to `NULL`, contains the following items:
`result`	A list of global inference results:
`X`	Matrix of covariates, same as input `X`.
`z`	Vector of observations, same as input `z`.
`probnull`	A vector of length n, indicating how likely the observed z belongs to local null.
`signal`	A binary vector of length n, discoveries are indicated by 1.
`plots`	A list of plots for global inference:
`signal_x`	A plot of signals discovered, marked in red
`dps_xz`	A scatterplot of z on x, colored based on the discovery propensity scores, only available when `fdr.method = "locfdr"`.
`dps_x`	A scatterplot of discovery propensity scores on x, only available when `fdr.method = "locfdr"`.
`micro`	Available when `X.target` are provided with values, contains the following items:
`result`	Customized estimates for null probabilities for target X and z
`result$signal`	A binary vector of length k, discoveries in the target cases are indicated by 1
`global`	Pooled global estimates for null probabilities for target X and z
`plots`	Customized fdr plots for the target cases.
`m.lp`	Same as input `m`

Author(s)

Subhadeep Mukhopadhyay, Kaijun Wang

Maintainer: Kaijun Wang <kaijunwang.19@gmail.com>

References

Mukhopadhyay, S., and Wang, K (2021) "On The Problem of Relevance in Statistical Inference". <arXiv:2004.09588>

Examples

```
data(funnel)
X<-funnel$x
z<-funnel$z
##macro-inference using locfdr and LASER:
g2l_macro<-g2l.proc(X,z)
g2l_macro$macro$plots
```

```
#Microinference for the DTI data: case A with x=(18,55) and z=3.95
data(data.dti)
X<- cbind(data.dti$coordx,data.dti$coordy)
z<-data.dti$z
g2l_x<-g2l.proc(X,z,X.target=c(18,55),z.target=3.95,nsample =3000)
g2l_x$micro$plots$fdr.1+ggplot2::coord_cartesian(xlim=c(0,4))
g2l_x$micro$result[4]
```

---

kidney *Kidney data.*

Description

This data set records age and kidney function of $N = 157$ volunteers. Higher scores indicates better function.

Usage

```
data(kidney)
```

Format

A data frame with 157 observations on the following 2 variables.

`x` A list of patients' age.

`z` A list of kidney scores.

Source

http://statweb.stanford.edu/~ckirby/brad/LSI/datasets-and-programs/datasets.html

References

Efron, B. (2012). "Large-scale inference: empirical Bayes methods for estimation, testing, and prediction". Cambridge University Press.

Lemley, K. V., Lafayette, R. A., Derby, G., Blouch, K. L., Anderson, L., Efron, B., & Myers, B. D. (2007). "Prediction of early progression in recently diagnosed IgA nephropathy." Nephrology Dialysis Transplantation, 23(1), 213-222.

LASER *Generates Artificial RELevance Samples.*

Description

This function generates the artificial relevance samples (LASER).These are "sharpened" z-samples manufactured by the relevance-function $d_x(z)$.

Usage

```
LASER( X,z, X.target, m=c(4,6), nsample=length(z), lp.reg.method='lm',
       coef.smooth='BIC', centering=TRUE,parallel=FALSE,...)
```

Arguments

X	A n-by-d matrix of covariate values
z	A length n vector containing observations of z values.
X.target	A k-by-d matrix providing k sets of target points for which the LASERs are required.
m	An ordered pair. First number indicates how many LP-nonparametric basis to construct for each X, second number indicates how many to construct for z. Default: m=c(4,6)
nsample	Number of relevance samples to generate for each case.
lp.reg.method	Method for estimating the relevance function and its conditional LP-Fourier coefficients. We currently support thee options: lm (inbuilt with subset selection), glmnet, and knn.
centering	Whether to perform regression-adjustment to center the data, default is TRUE.
coef.smooth	Specifies the method to use for LP coefficient smoothing (AIC or BIC). Uses BIC by default.
parallel	Use parallel computing for obtaining the relevance samples, mainly used for very huge nsample, default is FALSE.
...	Extra parameters to pass to other functions. Currently only supports the arguments for knn().

Value

A list containing the following items:

data	The relevant samples at X.target.
LPcoef	Parameters of the relevance function $d_x(x)$.

Author(s)

Subhadeep Mukhopadhyay, Kaijun Wang

Maintainer: Kaijun Wang <kaijunwang.19@gmail.com>

References

Mukhopadhyay, S., and Wang, K (2021) "On The Problem of Relevance in Statistical Inference". <arXiv:2004.09588>

Examples

```
data(funnel)
X<-funnel$x
z<-funnel$z
z.laser.x30<-LASER(X,z,X.target=30,m=c(4,8))$data
hist(z.laser.x30,50)
```

`rEB.Finite.Bayes` *Relevance-Integrated Finite Bayes.*

Description

Performs custom-tailored Finite Bayes inference via LASERs.

Usage

```
rEB.Finite.Bayes(X,z,X.target,z.target,m=c(4,6),m.EB=8, B=10, centering=TRUE,
            nsample=min(1000,length(z)), g.method='DL',LP.type='L2',  sd0=NULL,
            theta.set.prior=seq(-2.5*sd(z),2.5*sd(z),length.out=500),
        theta.set.post=seq(z.target-2.5*sd(z),z.target+2.5*sd(z),length.out=500),
            post.alpha=0.8,  plot=TRUE, ...)
```

Arguments

`X`	A n-by-d matrix of covariate values
`z`	A length n vector containing observations of target random variable.
`X.target`	A length d vector providing the set of covariates for the target case.
`z.target`	the target z to investigate
`m`	An ordered pair. First number indicates how many LP-nonparametric basis to construct for each X, second number indicates how many to construct for z.
`m.EB`	The truncation point reflecting the concentration of true nonparametric prior density π around known prior distribution g
`B`	Number of bags of bootstrap samples for Finite Bayes.
`centering`	Whether to perform regression-adjustment to center the data, default is TRUE.
`nsample`	Number of relevance samples generated for the target case.
`g.method`	Suggested method for finding parameter estimates $\hat{\mu}$ and $\hat{\tau}^2$ for normal prior: "DL" uses Dersimonian and Lard technique; "SJ" uses Sidik-Jonkman; 'REML' uses restricted maximum likelihood; and "MoM" uses a method of moments technique.

LP.type	User selects either "L2" for LP-orthogonal series representation of relevance density function d or "MaxEnt" for the maximum entropy representation. Default is L2.
sd0	Fixed standard deviation for $z\|\theta$. Default is NULL, the standard error will be calculated from data.
theta.set.prior	This indicates the set of grid points to compute prior density.
theta.set.post	This indicates the set of grid points to compute posterior density.
post.alpha	The alpha level for posterior HPD interval.
plot	Whether to display plots for prior and posterior of Relevance Finite Bayes.
...	Extra parameters to pass to LASER function.

Value

A list containing the following items:

prior	Relevant Finite Bayes prior results.
$prior.fit	Prior density curve estimation.
posterior	Relevant empirical Bayes posterior results.
$post.fit	Posterior density curve estimation.
$post.mode	Posterior mode for $\pi(\theta\|z, \boldsymbol{x})$.
$post.mean	Posterior mean for $\pi(\theta\|z, \boldsymbol{x})$.
$post.mean.sd	Standard error for the posterior mean.
$HPD.interval	The HPD interval for posterior $\pi(\theta\|z, \boldsymbol{x})$.
g.par	Parameters for $g = N(\mu, \tau^2)$.
LP.coef	Reports the LP-coefficients of the relevance function $d_x(x)$.
sd0	Initial estimate for null standard errors.
plots	The plots for prior and posterior density.

Author(s)

Subhadeep Mukhopadhyay, Kaijun Wang

Maintainer: Kaijun Wang <kaijunwang.19@gmail.com>

References

Mukhopadhyay, S., and Wang, K (2021) "On The Problem of Relevance in Statistical Inference". <arXiv:2004.09588>

Examples

```
data(funnel)
X<-funnel$x
z<-funnel$z
X.target=30
z.target=4.49
rFB.out=rEB.Finite.Bayes(X,z,X.target,z.target,B=5,nsample=1000,m=c(4,8),m.EB=8,
                         theta.set.prior=seq(-4,4,length.out=500),
                    theta.set.post=seq(0,5,length.out=500),cred.interval=0.8,parallel=FALSE)
rFB.out$plots$prior
rFB.out$plots$post
```

`rEB.proc` *Relevance-Integrated Empirical Bayes Inference*

Description

Performs custom-tailored empirical Bayes inference via LASERs.

Usage

```
rEB.proc(X, z, X.target, z.target, m = c(4, 6), nbag = NULL, centering = TRUE,
lp.reg.method = "lm", coef.smooth = "BIC", nsample = min(length(z),2000),
theta.set.prior = NULL, theta.set.post = NULL, LP.type = "L2",
g.method = "DL", sd0 = NULL, m.EB = 8, parallel = FALSE,
avg.method = "mean", post.curve = "HPD", post.alpha = 0.8,
color = "red", ...)
```

Arguments

`X`	A n-by-d matrix of covariate values
`z`	A length n vector containing observations of target random variable.
`X.target`	A length d vector providing the set of covariates for the target case.
`z.target`	the target z to investigate
`m`	An ordered pair. First number indicates how many LP-nonparametric basis to construct for each X, second number indicates how many to construct for z.
`nbag`	Number of bags of parametric bootstrapped samples to use, set to `NULL` to disable.
`centering`	Whether to perform regression-adjustment to center the data, default is TRUE.
`lp.reg.method`	Method for estimating the relevance function and its conditional LP-Fourier coefficients. We currently support thee options: lm (inbuilt with subset selection), glmnet, and knn.

coef.smooth	Specifies the method to use for LP coefficient smoothing (AIC or BIC). Uses BIC by default.
nsample	Number of relevance samples generated for the target case.
theta.set.prior	This indicates the set of grid points to compute prior density.
theta.set.post	This indicates the set of grid points to compute posterior density.
LP.type	User selects either "L2" for LP-orthogonal series representation of relevance density function d or "MaxEnt" for the maximum entropy representation. Default is L2.
g.method	Suggested method for finding parameter estimates $\hat{\mu}$ and $\hat{\tau}^2$ for normal prior: "DL" uses Dersimonian and Lard technique; "SJ" uses Sidik-Jonkman; 'REML' uses restricted maximum likelihood; and "MoM" uses a method of moments technique.
sd0	Fixed standard deviation for $z|\theta$. Default is NULL, the standard error will be calculated from data.
m.EB	The truncation point reflecting the concentration of true nonparametric prior density π around known prior distribution g
parallel	Use parallel computing for obtaining the relevance samples, mainly used for very huge nsample, default if FALSE.
avg.method	For parametric bootstrapping, this specifies how the results from different bags are aggregated. ("mean" or "median".)
post.curve	For plotting, this specifies what to show on posterior curve. "HPD" provides HPD interval, "band" gives confidence band.
post.alpha	Confidence level to use when plotting posterior confidence band, or the alpha level for HPD interval.
color	The color of the plots.
...	Extra parameters to pass to other functions. Currently only supports the arguments for knn().

Value

A list containing the following items:

result	Contains relevant empirical Bayes prior and posterior results.
sd0	Initial estimate for null standard errors.
prior	Relevant empirical Bayes prior results.
$g.par	Parameters for $g = N(\mu, \tau^2)$.
$g.method	Method used for finding the parameter estimates $\hat{\mu}$ and $\hat{\tau}^2$ for g.
$LP.coef	Reports the LP-coefficients of the relevance function $d_x(x)$.
posterior	Relevant empirical Bayes posterior results.
$post.mode	Posterior mode for $\pi(\theta|z, \boldsymbol{x})$.

`$post.mean`	Posterior mean for $\pi(\theta\|z,\boldsymbol{x})$.
`$post.mean.sd`	Standard error for the posterior mean, when using parametric bootstrap.
`$HPD.interval`	The HPD interval for posterior $\pi(\theta\|z,\boldsymbol{x})$.
`$post.alpha`	same as input `post.alpha`.
`plots`	The plots for prior and posterior density.

Author(s)

Subhadeep Mukhopadhyay, Kaijun Wang

Maintainer: Kaijun Wang <kaijunwang.19@gmail.com>

References

Mukhopadhyay, S., and Wang, K (2021) "On The Problem of Relevance in Statistical Inference". <arXiv:2004.09588>

Examples

```
data(funnel)
X<-funnel$x
z<-funnel$z
X.target=60
z.target=4.49
rEB.out<-rEB.proc(X,z,X.target,z.target,m=c(4,8),
theta.set.prior=seq(-2,2,length.out=200),
theta.set.post=seq(-2,5,length.out=200),
centering=TRUE,m.EB=6,nsample=1000)
rEB.out$plots$rEB.post
rEB.out$plots$rEB.prior
```

Index