Package 'refreg'

October 14, 2022

Type Package

Title Conditional Multivariate Reference Regions

Version 0.1.1

Maintainer Oscar Lado-Baleato <oscarlado.baleato@usc.es>

Description An R package for estimating conditional multivariate reference regions.
The reference region is non parametrically estimated using a kernel density estimator.
Covariates effects on the multivariate response means vector and variance-covariance
matrix, thus on the region shape, are estimated by flexible additive predictors.
Continuous covariates non linear effects might be estimated using penalized splines smoothers.
Confidence intervals for the covariates estimated effects might be derived from
bootstrap resampling. Kernel density bandwidth can be estimated with different methods, including
a method that optimize the region coverage. Numerical, and graphical, summaries
can be obtained by the user in order to evaluate reference region performance with real data.
Full mathematical details can be found in <doi:10.1002/sim.9163> and <doi:10.1007/s00477-020-01901-1>.

Imports
ks,KernSmooth,mgcv,gridExtra,ggplot2,RColorBrewer,sp,stringr,foreach,doParallel,pracma,misc3d,rgl,mbend,matrixcalc

License GPL-2

Encoding UTF-8

LazyData true

RoxygenNote 7.1.2

Depends R (>= 2.10)

NeedsCompilation no

Author Oscar Lado-Baleato [cre, aut],
Javier Roca-Pardinas [aut, ctb],
Carmen Cadarso-Suarez [aut, ths],
Gude Francisco [aut, ths]

Repository CRAN

Date/Publication 2022-04-18 14:20:02 UTC

R topics documented:

ACE   *Alternating Conditional Expectation (ACE) algorithm*

Description

This function implements Alternating Conditional Expectation algorithm (ACE, Hastie and Tibshirani, 1990). This is a bivRegr's inner function for estimating variance, and correlation models.

Usage

```
ACE(
  y = "y",
  predictor = "~s(x)",
  restriction = "positive",
  eps = 0.01,
  itmax = 10,
  data = data,
  ...
)
```

Arguments

y	A character defining the response variable.
predictor	The regression predictor, following the mgcv package structure (see example below).
restriction	Type of restriction to be imposed to the response variable, "positive" for variance, and "correlation" for the correlation.
eps	A number defining the allowed estimation error, default = 0.01.
itmax	Maximum number of iterations of the algorithm, default = 10.
data	A data frame containing the response, and predictor variables.
...	Additional mgcv::gam() parameters to be modified by the user.

Value

This function returns a mgcv::gam() fit for a transformed response.

References

Hastie, T. & Tibshirani, J. (1990) Generalized additive models. CRC press. London.

Examples

```
n <- 100
x <- runif(n, -1, 1)
y <- x^2 + rnorm(n, sd = 0.1)
df <- data.frame(y, x)
plot(df$x, df$y)
m1 <- ACE(
  y = "y", predictor = "~s(x)", restriction = "positive",
  eps = 0.01, itmax = 10, data = df
)$fit
nw <- data.frame(x = seq(-1, 1, 0.1))
abline(h = 0, col = 2)
lines(exp(predict(m1, newdata = nw)) ~ nw$x, col = 3, lwd = 2)
legend("top", legend = c("ACE fit", "Zero"), lty = 1, lwd = 2, col = c(3, 2), bty = "n")
```

---

aegis	*A Estrada Glycation and Inflammation Study (AEGIS) Data*

Description

The aegis dataset was obtained from the Estrada Glycation and Inflammation (AEGIS) project. A cross sectional, population based study that was performed in the municipality of A Estrada (Galicia, NW Spain). The study objective was to investigate the association between glycation, inflammation status, lifestyle and common diseases, and reasons for glycemic markers discordances (Gude et al., 2017).

Usage

```
aegis
```

Format

A data frame containing 1516 observations, and 7 variables:

id Anonymized patient identificator.

gender A factor variable describing patient gender, (female, and male).

age Patients' age.

dm Diabetes mellitus indicator (no, and yes).

fpg Fasting plasma glucose levels (mg/dL).

hba1c Glycated hemoglobin percentage.

fru Fructosamine concentration (mg/dL).

References

Gude F, Diaz–Vidal P, Rua–Perez C, et al. Glycemic variability and its association with demographics and lifestyles in ageneral adult population. J Diabetes Sci Technol. 2017; 11(4): 780–790

---

bivRegion   *Bivariate reference region estimation*

Description

This functions estimate a probabilistic/reference region for bivariate data. It is based on a kernel density estimation. It may be applied to a set of bivariate data points, or to a bivRegr object. In the former case, the function will estimate a bivariate reference region for the model standarized residuals.

Usage

```
bivRegion(
  Y = fit,
  H_choice = "Hcov",
  tau = 0.95,
  k = 20,
  display_plot = TRUE,
  shape = NULL,
  ...
)
```

Arguments

Y	A set of bivariate data points, or a bivRegr object.
H_choice	Kernel bandwidth selection method: "plug.in" for plug.in method, "LSCV" for least squate cross valiation, "SCV" for smooth cross validation, and "Hcov" for a bandwidth selection method which optimize the region coverage.
tau	A number or vector defining the desired coverage(s) of the bivariate reference region.
k	In case of using "Hcov" the number of k fold cross validations replicates to be performed.
display_plot	A logical indicating if plot must be displayed during "Hcov" bandwidht estimation procedure. The plot depicts region's coverage, evaluated with k fold cross validation, depending on kernel bandwidth value.
shape	Shape parameter modulating the final shape of the bivariate probabilistic/reference region by hand.
...	Additional parameters to be modified in KernSmooth::bkde2D() function by the user (e.g. gridsize).

Value

This function return a region or a set of regions containing a given percentage of bivariate data points.

References

Duong, T. (2019) ks: Kernel Smoothing. R package version 1.11.6. https://CRAN.R–project.org/package=ks.

Matt Wand (2020). KernSmooth: Functions for Kernel Smoothing Supporting Wand & Jones (1995). R package version 2.23–18. https://CRAN.R–project.org/package=KernSmooth

Examples

```
Y <- cbind(rnorm(100), rnorm(100))
Y <- as.data.frame(Y)
names(Y) <- c("y1", "y2")
region <- bivRegion(Y, tau = 0.95, shape = 2)
plot(region)
```

---

bivRegr	*Bivariate regression model*

Description

This function estimates the covariates effects on the means vector, and variance covariance matrix from a bivariate variable. Non linear effects might be estimated for continuous covariates using penalized spline smoothers.

Usage

```
bivRegr(f = f, data = data, ace.eps = 0.01, ace.itmax = 25)
```

Arguments

f	A list of five formulas defining the covariates effects in both responses means, variances, and in their correlation. The formulas follow the same structure as mgcv::gam() (see example below).
data	A data frame containing the reponses, and predictor variables values.
ace.eps	A number defining the error rate in the ACE algorithm.
ace.itmax	A number defining the maximum number of ACE algorithm iterations.

Value

This function returns the covariates effect on the means, variances, and correlation of a bivariate response variable.

Examples

```
# Bivariate reference region for fasting plasma glucose (fpg)
# and glycated hemoglobin (hba1c) levels depending on age

dm_no <- subset(aegis, aegis$dm == "no") # select healthy patients
# 1.1) Define predictors
mu1 <- fpg ~ s(age)
mu2 <- hba1c ~ s(age)
var1 <- ~ s(age)
var2 <- ~ s(age)
rho <- ~ s(age)
f <- list(mu1, mu2, var1, var2, rho)

fit <- bivRegr(f, data = dm_no)

# 1.2) Depict the estimated covariates effects
plot(fit, eq = 1)
plot(fit, eq = 2)
plot(fit, eq = 3)
plot(fit, eq = 4)
plot(fit, eq = 5)
# 1.2.1) Depict the estimated covariates effects with CI (Not Run)

s0 <- summary_boot(fit, B = 100) # no parallelization
# s1 = summary_boot(fit,B=100,parallel=TRUE) #parallelization
plot(s0, eq = 1)


# 1.3) Obtain the reference region in the standarized residuals
region <- bivRegion(fit, tau = 0.95, shape = 2)
plot(region)
```

```
# 1.4) Identify those patients located outside the reference region
summary(region)

# 1.5) Depict the conditional reference region for two ages
plot(region,
  cond = TRUE, newdata = data.frame(age = c(20, 50)), col = "grey", pch = "*",
  reg.lwd = 2, reg.lty = 2
)
```

Hcov	*Kernel bandwidth selection method based on bivariate density contours coverage*

Description

This function implements a method for estimating bivariate kernel bandwidth based on data covarage. The method starts with the plug in estimate (which usually overfits the data), and then increase this bandwidth value until the desired coverage is obtained. Region coverage is evaluated in an out sample design, using a k fold cross validation scheme.

Usage

```
Hcov(Y, shape = seq(1, 10, 0.5), k = 20, tau = 0.9, display_plot = TRUE)
```

Arguments

Y	A matrix containing bivariate data values.
shape	A sequence of values which controls plug in estimator increasing.
k	A number indicating k fold cross validations to be performed.
tau	The desired region coverage
display_plot	A logical indicating if a plot must be displaying, during the function estimation process, summarizing the results.

Value

This function return a diagonal kernel bandwidth matrix.

---

plot.bivRegion *Plot a bivRegion object*

Description

This function allow to depict the estimated bivariate reference/probabilistic region, in the estandarized residuals scale (cond=FALSE), or for any covariate value (cond=TRUE).

Usage

```
## S3 method for class 'bivRegion'
plot(
  x,
  tau = 0.95,
  newdata = NULL,
  reg.col = NULL,
  reg.lwd = 1,
  reg.lty = NULL,
  axes = TRUE,
  axes.col = "black",
  axes.lwd = 2L,
  cond = FALSE,
  add = FALSE,
  legend = TRUE,
  ...
)
```

Arguments

x	A bivRegion object.
tau	A number, or vector, defining the desired coverage(s) of the bivariate reference region.
newdata	If cond=FALSE, a data.frame with new values to be depicted in the standarized residuals scale. If cond=TRUE, a data frame containing covariate values for which the reference/probabilistic region will be depicted.
reg.col	Region line colour, in case of more than one tau it can be a vector.
reg.lwd	Region line width, in case of more than one tau it can be a vector.
reg.lty	Region line type, in case of more than one tau it can be a vector.
axes	Logical; if TRUE (and cond=FALSE), vertical and horizontal lines are added indicating four quadrants in the model residuals scale.
axes.col	Axes colour.
axes.lwd	Axes line width.
cond	A logical argument, if TRUE a conditional reference region is depicted.

add	A logical argument, if TRUE the conditional reference region is depicted over a pre existing plot.
legend	A logical argument, if TRUE a legend is given along with the reference region.
...	Further plot parameters.

Value

This function return a graphical representation for a bivRegion object.

Examples

```
Y <- cbind(rnorm(100), rnorm(100))
Y <- as.data.frame(Y)
names(Y) <- c("y1", "y2")
reg <- bivRegion(Y, tau = 0.95, shape = 2)
plot(reg)
```

---

plot.bivRegr   *Plot method for bivRegr fit*

Description

This function takes an bivRegr object and plots the estimated effects for the conditional response means, variances or correlation. summary_boot function must be applied by the user in order to get the estimated effects confidence intervals.

Usage

```
## S3 method for class 'bivRegr'
plot(x, eq = 1, newdata = NULL, ...)
```

Arguments

x	A bivRegr fit.
eq	A number indicating the model effects to be depicted; 1 = first response mean, 2 = second response mean, 3 = first response variance, 4 = second response variance, and 5 = correlation model.
newdata	An optional data frame containing covariates values.
...	Additional plot arguments.

Value

This function returns a plot for bivRegr mean, variance and correlation models.

---

plot.refcurve	*Plot univariate conditional quantile models curves (i.e. reference curves)*

Description

This function depict the univariate conditional quantile model based on the non parametric location scale model fitted with the refcurv function.

Usage

```
## S3 method for class 'refcurve'
plot(
  x,
  newdata = data.frame(x = seq(0, 1, 0.01)),
  tau = seq(0.1, 0.9, 0.2),
  ...
)
```

Arguments

x	A refcurv object.
newdata	A data frame defining a sequence of the predictor variables values.
tau	A number or vector defining desired quantile.
...	Additional plot options.

Value

This function returns a plot of the refcurve model.

---

plot.summary_boot	*Default summary_boot plotting*

Description

This function takes the bivRegr bootstrap replicates obtained with summary_boot function, and plots the parametric, and smooth effects for each model.

Usage

```
## S3 method for class 'summary_boot'
plot(x, eq = 1, select = NULL, ...)
```

Arguments

x	A summary_boot object.
eq	A number indicating the model effects to be depicted; 1 = first response mean, 2 = second response mean, 3 = first response variance, 4 = second response variance, and 5 = correlation model.
select	An optional parameter to represent an specific effect for each equation.
...	Additional plot parameters, not yet implemented.

Value

This function returns a ggplot2 plot for the estimated effects along with bootstrap 95% confidence intervals.

---

plot.trivRegion	*Default trivRegion plotting*

Description

This function allow to depict in an interactive rgl plot the estimated trivariate reference/probabilistic region. If cond=FALSE it showes trivariate standarized residuals, while with cond=TRUE it represents the region shape for any covariate(s) value.

Usage

```
## S3 method for class 'trivRegion'
plot(
  x,
  cond = FALSE,
  planes = FALSE,
  newdata = NULL,
  add = FALSE,
  reg.col = NULL,
  incol = "grey",
  legend = FALSE,
  ...
)
```

Arguments

x	A trivRegion object.
cond	A logical argument, if TRUE a conditional reference region is depicted.
planes	Logical; if TRUE, planes are added indicating (x=0,y=0,z=0).
newdata	If cond==TRUE, a data frame containing covariate values for which the reference/probabilistic region will be depicted.

add	A logical argument, if TRUE the conditional reference region is depicted over a pre existing plot.
reg.col	Region contour colour.
incol	Colours for the points included inside the reference region.
legend	A logical argument, if TRUE a legend is given along with the reference region.
...	Further rgl plot parameters.

Value

This function return an interactive rgl plot of a bivRegion object.

---

pollution	*SO2 and Nox concentrations*

Description

The pollution dataset contains the SO2 and Nox concentrations in the As Pontes thermal power plant surrondings, during a year. The data was analyzed from a bivariate perspective in Roca–Pardiñas et al. (2021).

Usage

```
pollution
```

Format

A data frame with 6179 observations and 13 variables:

Date The observation day, hour, and minute.

So2 So2 air concentration at the present time.

Nox Nox air concentration at the present time.

So2_0 So2 air concentration registered 30 minutes before than the So2 one.

Nox_0 Nox air concentration registered 30 minutes before than the Nox one.

So2_1 So2 air concentration registered 45 minutes before than the So2 one.

Nox_1 Nox air concentration registered 45 minutes before than the Nox one.

So2_2 So2 air concentration registered 60 minutes before than the So2 one.

Nox_2 Nox air concentration registered 60 minutes before than the Nox one.

So2_3 So2 air concentration registered 75 minutes before than the So2 one.

Nox_3 Nox air concentration registered 75 minutes before than the Nox one.

So2_4 So2 air concentration registered 90 minutes before than the So2 one.

Nox_4 Nox air concentration registered 90 minutes before than the Nox one.

References

Roca–Pardinas, J., Ordonez, C., & Lado Baleato, O. (2021). Nonparametric location scale model for the joint forecasting of SO2 and NOx pollution episodes. Stochastic Environmental Research and Risk Assessment, 35(2), 231–244.

---

pollution_episode   *SO2 and Nox concentrations during a pollution episode*

Description

SO2 and Nox concentrations in the As Pontes thermal power plant surrondings, during a pollution episode.

Usage

```
pollution_episode
```

Format

A data frame with 288 observations and 13 variables:

Date The observation day, hour and minute.

So2 So2 air concentration at the present time.

Nox Nox air concentration at the present time.

So2_0 So2 air concentration registered 30 minutes before than the So2 one.

Nox_0 Nox air concentration registered 30 minutes before than the Nox one.

So2_1 So2 air concentration registered 45 minutes before than the So2 one.

Nox_1 Nox air concentration registered 45 minutes before than the Nox one.

So2_2 So2 air concentration registered 60 minutes before than the So2 one.

Nox_2 Nox air concentration registered 60 minutes before than the Nox one.

So2_3 So2 air concentration registered 75 minutes before than the So2 one.

Nox_3 Nox air concentration registered 75 minutes before than the Nox one.

So2_4 So2 air concentration registered 90 minutes before than the So2 one.

Nox_4 Nox air concentration registered 90 minutes before than the Nox one.

---

predict.bivRegion   *Prediction for a bivRegion object*

Description

This function takes a bivRegion object and allow to obtain region limits for a given covariate value if cond=TRUE. If not, it can be applied to a new dataset to evaluate which points will be included inside the standarized region (for instance, if we estimate a reference region with healthy patients results, we can know where non healthy patients will be located).

Usage

```
## S3 method for class 'bivRegion'
predict(object, tau = 0.95, newdata = NULL, cond = FALSE, ...)
```

Arguments

object	A bivRegion object.
tau	A number or vector defining the desired coverage(s) of the bivariate reference/probabilistic region.
newdata	An optional data frame containing new covariate values, or new observations.
cond	A logical argument, if TRUE, a matrix of values defining the conditional reference region limits is given.
...	Additional prediction parameters.

Value

This function returns reference region limits values for a given covariate value (if cond=TRUE), or which observations falls outside the estimated region (if cond=FALSE).

---

predict.bivRegr   *Predict method for bivRegr*

Description

Obtains predictions for a bivRegr object, that is the bivRegr means, variances, and correlation for given covariate values.

Usage

```
## S3 method for class 'bivRegr'
predict(object, newdata = NULL, ...)
```

Arguments

`object`	A bivRegr fit.
`newdata`	A data frame defining the covariate values for prediction. Default is NULL and the prediction will be done in the original data.
`...`	Additional predict options.

Value

This function returns prediction of bivRegr mean, variance and correlation models.

---

refcurv	*Univariate reference curve model*

Description

This function obtain univariate conditional quantiles as described in Martinez-Silva et. al (2016).

Usage

```
refcurv(mu = "y~s(x)", sigma = "~s(x)", data = data)
```

Arguments

`mu`	A formula object for the response mean model following the mgcv package structure (see example below).
`sigma`	a formula object for fitting a model to the response variance (see example below).
`data`	A data frame containing both the response, and predictor variables.

Details

In the Martinez Silva et. al (2016) the non linear effects of the continuous covariates are estimating through polynomial kernel smoother, in this package we implement the same methodology but using penalized splines in order to reduce computational cost.

Value

This function returns univariate conditional quantiles estimated using a non parametric location scale model.

References

Martinez–Silva, I., Roca–Pardinas, J., & Ordonez, C. (2016). Forecasting SO2 pollution incidents by means of quantile curves based on additive models. Environmetrics, 27(3), 147–157.

Examples

```
#--- Glycation hemoglobin reference curve depending on age
dm_no <- subset(aegis, aegis$dm == "no")
fit1 <- refcurv(mu = "hba1c~s(age)", sigma = "~s(age)", data = dm_no)
plot(fit1, newdata = data.frame(age = 18:90), tau = c(0.025, 0.05, 0.10, 0.90, 0.95, 0.975))
```

---

summary.bivRegion   *bivRegion summary method*

Description

This function takes an bivRegion object and indicates which observations are located outside the estimated reference region and why.

Usage

```
## S3 method for class 'bivRegion'
summary(object, tau = 0.95, ...)
```

Arguments

object	A bivRegion object.
tau	The data coverage proportion previously obtained by the bivRegion function.
...	Additional summary options.

Value

This function indicates the region apparent coverage, and which patients are located outside the estimated reference region.

---

summary_boot   *bivRegr summary function*

Description

This function perform a bootstrap procedure in order to obtain confidence intervals for the bivRegr estimated effects. The function allow to paralelize the bootstrap resampling scheme using doParalell, and foreach libraries.

Usage

```
summary_boot(object, B = 100, parallel = FALSE, cores = NULL)
```

Arguments

object	A bivRegr fit.
B	A number indicating the bootstrap iterations.
parallel	A logical indicating if bootstrap parallelization must be applied.
cores	If parallel = TRUE, a number indicating computer cores to be used during parallelization. If NULL the function use all cores available but one.

Value

This function returns the bootstrap replicates of the bivRegr sub models. Results might be checked applying plot.summary_boot().

---

trivRegion	*Trivariate reference region estimation*

Description

This functions estimate a probabilistic/reference region for trivariate data. It is based on a non parametric kernel density estimation. It can only be applied to a trivRegr object, and for one single tau.

Usage

```
trivRegion(fit, tau = 0.9)
```

Arguments

fit	A trivRegr object.
tau	A number defining the desired coverage of the trivariate reference region.

Value

This function return a region containing a given percentage of trivariate data points.

References

Duong, T. (2019) ks: Kernel Smoothing. R package version 1.11.6. https://CRAN.R–project.org/package=ks.

---

trivRegr   *Trivariate regression model*

Description

This function estimates the covariates effects on the means vector, and variance covariance matrix of a trivariate variable. Non linear effects might be estimated for continuous covariates using penalized splines.

Usage

```
trivRegr(f = f, data = data)
```

Arguments

f   A list of 9 formulas defining the covariates effects in three responses means, in
                  their variances, and in their correlations. The formulas follow the mgcv::gam()
                  structure.

data   A data frame containing the reponses, and predictor variables values.

Value

This function returns the covariates effect on the means, variances, and correlation for a trivariate response variable.

Examples

```
dm_no <- subset(aegis, aegis$dm == "no")

# Model formulas
mu1 <- fpg ~ s(age)
mu2 <- hba1c ~ s(age)
mu3 <- fru~s(age)
var1 <- ~ s(age)
var2 <- ~ s(age)
var3 <- ~ s(age)
theta12 <- ~ s(age)
theta13 <- ~ s(age)
theta23 <- ~ s(age)
f <- list(mu1, mu2, mu3, var1, var2, var3, theta12, theta13, theta23)

# Model fit
fit <- trivRegr(f, data = dm_no)
# Trivariate region estimation
region <- trivRegion(fit, tau = 0.95)
plot(region, col = 2, planes = TRUE)
plot(region,
  cond = TRUE, newdata = data.frame(age = c(20, 80)),
  xlab = "FPG, mg/dl", ylab = "HbA1c, %", zlab = "Fru, mg/dL"
```

)

Index