Package 'MAMSE'

January 7, 2017

Type Package

Title Calculation of Minimum Averaged Mean Squared Error (MAMSE) Weights

Version 0.2-1

Date 2017-07-07

Author Jean-Francois Plante <jfplante@hec.ca>

Maintainer Jean-Francois Plante <jfplante@hec.ca>

Description Calculates the nonparametric adaptive MAMSE weights for univariate, right-censored or multivariate data. The MAMSE weights can be used in a weighted likelihood or to define a mixture of empirical distribution functions. The package includes functions for the MAMSE weighted Kaplan-Meier estimate and for MAMSE weighted ROC curves.

Depends R (>= 2.4.0)

License GPL-2

NeedsCompilation yes

Repository CRAN

Date/Publication 2017-01-07 16:16:32

R topics documented:

MAMSE-package *Minimum Averaged Mean Squared Error (MAMSE) Weights.*

Description

This package provides algorithms to calculate the nonparametric adaptive MAMSE weights. The MAMSE weights can be used for the weighted likelihood (see references below), or as mixing probabilities to define mixtures of empirical distributions. They provide a framework to borrow strenght with minimal assumptions.

Details

Package:	MAMSE
Type:	Package
Version:	0.2
Date:	2016-01-20
License:	GPL-2

Function MAMSE calculates the MAMSE weights for univariate data, right-censored data, or for the copula underlying the distribution of multivariate data. The function WKME is used to compute the MAMSE-weighted Kaplan-Meier estimate with (optional) bootstrap confidence intervals. The function roc calculates MAMSE-weighted ROC curves.

Author(s)

Jean-Francois Plante, <jfplante(at)hec.ca>.

Maintainer: Jean-Francois Plante, <jfplante(at)hec.ca>.

References

J.-B. Débordès & J.-F. Plante (2009). Combining ROC curves using MAMSE weighted distributions. *Cahier du GERAD G-2015-69*.

F. Hu and J. V. Zidek (2002). The weighted likelihood, *The Canadian Journal of Statistics*, **30**, 347–371.

J.-F. Plante (2007). *Adaptive Likelihood Weights and Mixtures of Empirical Distributions*. Unpublished doctoral dissertation, University of British Columbia.

J.-F. Plante (2008). Nonparametric adaptive likelihood weights. *The Canadian Journal of Statistics*, **36**, 443-461.

J.-F. Plante (2009). Asymptotic properties of the MAMSE adaptive likelihood weights. *Journal of Statistical Planning and Inference*, **139**, 2147-2161.

J.-F. Plante (2009). About an adaptively weighted Kaplan-Meier estimate. *Lifetime Data Analysis*, **15**, 295-315.

X. Wang (2001). *Maximum weighted likelihood estimation*, unpublished doctoral dissertation, Department of Statistics, The University of British Columbia.

See Also

MAMSE, WKME, roc.

Examples

```
set.seed(2009)

# MAMSE weights for univariate data
x=list(rnorm(25),rnorm(250,.1),rnorm(100,-.1))
wx=MAMSE(x)

# Weighted Likelihood estimate for the mean (Normal model)
sum(wx*sapply(x,mean))

#MAMSE weights for copulas
rho=c(.25,.3,.15,.2)
r=2*sin(rho*pi/600)
y=list(0,0,0,0)
for(i in 1:4){
  sig=matrix(c(1,r,r,1),2,2)
  y[[i]]=matrix(rnorm(150),nc=2)
}
wy=MAMSE(y)

# Weighted coefficient of correlation
sum(wy*sapply(y,cor,method="spearman")[2,])

#MAMSE weights for right-censored data

z=list(0,0,0)
for(i in 1:3){
  zo=rexp(100)
  zc=pmin(rexp(100),rexp(100),rexp(100))
  z[[i]]=cbind(pmin(zo,zc),zo<=zc)
}

MAMSE(z,.5,surv=TRUE)

allz=pmin(.5,c(z[[1]][z[[1]][,2]==1,1],z[[2]][z[[2]][,2]==1,1],
   z[[3]][z[[3]][,2]==1,1]))
K=WKME(z,.5,time=sort(unique(c(0,.5,allz,allz-.0001))))
plot(K$time,K$wkme,type='l',col="blue",xlab="x",ylab="P(X<=x)",
   ylim=c(0,.5))
lines(K$time,K$kme[,1],col="red")
legend(0,.5,c("Weighted Kaplan-Meier","Kaplan-Meier"),
   col=c("blue","red"),lty=c(1,1))

# MAMSE-weighted ROC curve

set.seed(2016)
nh=c(50,25,70,100)
nd=c(40,20,50,80)
```

```
muh=c(1.5,1,1.7,1.2)
mud=c(0,.2,.5,.4)

# Target curve
FPR=seq(0,1,.01)
TPR=pnorm(qnorm(FPR,mean=muh[1]),mean=mud[1])

simh=list()
simd=list()

for(i in (1:length(nh))){
  simh[[i]]=rnorm(nh[i],mean=muh[i])
  simd[[i]]=rnorm(nd[i],mean=mud[i])
}

par(mfrow=c(1,2))
plot(roc(simh,simd),col="red")
lines(roc(simh[[1]],simd[[1]]),col="blue")
lines(FPR,TPR,col="gray")
title("Empirical ROC curves")

plot(roc(simh,simd,method="normal"),col="red")
lines(roc(simh[[1]],simd[[1]],method="normal"),col="blue")
lines(FPR,TPR,col="gray")
title("Parametric ROC curves")
```

KM *Functions for use by WKME.*

Description

Functions used by WKME to compute the MAMSE-weighted Kaplan-Meier estimate.

MAMSE *Minimum Averaged Mean Squared Error Weights*

Description

Computes the MAMSE weights (see references below for their definition).

Usage

```
MAMSE(x,surv=FALSE,ub=NULL,lb=0,MCint=FALSE,nMC=10000)
```

Arguments

x	A list of m samples. Elements of the list must be vectors of matrices. If they are vectors, the univariate MAMSE weights are computed. Matrices should have n lines with one p-dimensional datum per line. The data are automatically tranformed into rescaled ranks by the function ranked. The MAMSE weights for copulas are then calculated. For survival MAMSE weights, use the argument surv=TRUE and provide an n by 2 matrix where the second column is an indicator (delta) of whether the time in column 1 is observed (delta=1) or censored (delta=0).
surv	Controls the calculation of the survival MAMSE weights rather that the multivariate version for copulas.
ub	if surv=TRUE, the upper bound for the integral of the MAMSE criterion.
lb	If surv=TRUE, the lower bound for the integral of the MAMSE criterion.
MCint	When MAMSE weights are calculated for copulas, MCint=TRUE allows to proceed with Monte Carlo integration. The laternative MCint=TRUE will estimate the integral on the grid [1/n1,2/n1,...,1]^p which does not scale well with the number of dimensions p.
nMC	When MCint=TRUE, nMC controls the number of samples used to approximate the integral.

Details

Provided a list of samples, this function returns the Minimum Averaged Mean Squared Error weights. The MAMSE weights can be used in a weighted likelihood, or to define mixtures of empirical distributions. In both cases, the methodology is used to infer on Population 1 while borrowing strength from the other samples provided. Refer to the articles below for the exact definition of the MAMSE weights, their asymptotic properties and simulations results, as well as additional information about the weighted likelihood.

Value

A vector of p elements containing the MAMSE weights for each of the populations.

References

F. Hu and J. V. Zidek (2002). The weighted likelihood, *The Canadian Journal of Statistics*, **30**, 347–371.

J.-F. Plante (2007). *Adaptive Likelihood Weights and Mixtures of Empirical Distributions*. Unpublished doctoral dissertation, University of British Columbia.

J.-F. Plante (2008). Nonparametric adaptive likelihood weights. *The Canadian Journal of Statistics*, **36**, 443-461.

J.-F. Plante (2009). Asymptotic properties of the MAMSE adaptive likelihood weights. *Journal of Statistical Planning and Inference*, **139**, 2147-2161.

J.-F. Plante (2009). About an adaptively weighted Kaplan-Meier estimate. *Lifetime Data Analysis*, **15**, 295-315.

X. Wang (2001). *Maximum weighted likelihood estimation*, unpublished doctoral dissertation, Department of Statistics, The University of British Columbia.

See Also

MAMSE-package, WKME.

Examples

```
set.seed(2009)

# MAMSE weights for univariate data
x=list(rnorm(25),rnorm(25,.1),rnorm(25,.2))
MAMSE(x)

#MAMSE weights for copulas
y=list(matrix(rnorm(150),nc=2),matrix(rnorm(150),nc=2),
   matrix(rnorm(150),nc=2))
MAMSE(y)
MAMSE(y,MCint=TRUE)

#MAMSE weights for right-censored data
z=list(cbind(rexp(50),rbinom(50,1,.5)),cbind(rexp(50,1.1),
   rbinom(50,1,.5)),cbind(rexp(50,.9),rbinom(50,1,.5)))
MAMSE(z,3,surv=TRUE)

#For more examples, see help on "MAMSE-package"
```

`MAMSEpo` *Functions for use by MAMSE.*

Description

Functions used by MAMSE to compute the MAMSE weights.

`Progesterone` *Data sets: Progesterone level for detecting ectopic pregnancies and natural abortions.*

Description

Débordès & Plante (2015) extract data from figures in published articles. All papers show the level of progesterone as a diagnotic variable for ectopic pregnancies or natural abortions from other causes. The object `Progesterone` is a list of lists containing samples from the different papers from which data are obtained.

Usage

```
data(Progesterone)
```

Format

List of lists of vectors.

Source

Data were extracted from ublished figures in Dumps et al. (2002), Florio et al. (2007), Gelder et al. (1991), Grosskinsky et al. (1993), Hanita et al. (2012), Ledger et al. (1994), O'leary et al. (1996), Peterson et al. (1992), Riss et al. (1989), Stewart et al.(1995), and Witt et al. (1990). All measurements are in nmol/l. Note that the data were *extracted* from the *figures* of the paper and as such, contain error due to their conversion back into numbers.

References

J.-B. Débordès & J.-F. Plante (2009). Combining ROC curves using MAMSE weighted distributions. *Cahier du GERAD G-2015-69*.

Dumps, P., Meisser, A., Pons, D., Morales, M. A., Anguenot, J.-L., Campana, A., and Bischof, P. (2002). Accuracy of single measurements of pregnancy-associated plasma protein-a, human chorionic gonadotropin and progesterone in the diagnosis of early pregnancy failure. European Journal of Obstetrics & Gynecology and Reproductive Biology 100, 174–180.

Florio, P., Severi, F. M., Bocchi, C., Luisi, S., Mazzini, M., Danero, S., Torricelli, M., and Petraglia, F. (2007). Single serum activin a testing to predict ectopic pregnancy. The Journal of Clinical Endocrinology & Metabolism 92, 1748–1753.

Gelder, M., Boots, L., and Younger, J. (1991). Use of a single random serum progesterone value as a diagnostic aid for ectopic pregnancy. Fertility and sterility 55, 497–500.

Grosskinsky, C., Hage, M., Tyrey, L., Christakos, A., and Hughes, C. (1993). hcg, progesterone, alpha-fetoprotein, and estradiol in the identification of ectopic pregnancy. Obstetrics and gynecology 81, 705–709.

Hanita, O., Hanisah, A., et al. (2012). Potential use of single measurement of serum progesterone in detecting early pregnancy failure. Malaysian J Pathol 34, 41–46.

Ledger, W., Sweeting, V., and Chatterjee, S. (1994). Rapid diagnosis of early ectopic pregnancy in an emergency gynaecology serviceare measurements of progesterone, intact and free β human chorionic gonadotrophin helpful? Human reproduction 9, 157–160.

O'Leary, P., Nichols, C., Feddema, P., Lam, T., and Aitken, M. (1996). Serum progesterone and human chorionic gonadotrophin measurements in the evaluation of ectopic pregnancy. Australian and New Zealand journal of obstetrics and gynaecology 36, 319–323.

Peterson, C., Kreger, D., Delgado, P., and Hung, T. (1992). Laboratory and clinical comparison of a rapid versus a classic progesterone radioimmunoassay for use in determining abnormal and ectopic pregnancies. American journal of obstetrics and gynecology 166, 562–566.

Riss, P. A., Radivojevic, K., and Bieglmayer, C. (1989). Serum progesterone and human, Biometrics, December 2015 chorionic gonadotropin in very early pregnancy: implications for clinical management. European Journal of Obstetrics & Gynecology and Reproductive Biology 32, 71–77.

Stewart, B. K., Nazar-Stewart, V., and Toivola, B. (1995). Biochemical discrimination of pathologic pregnancy from early, normal intrauterine gestation in symptomatic patients. American journal of clinical pathology 103, 386–390.

Witt, B., Wolf, G., Wainwright, C., Johnston, P., and Thorneycroft, I. (1990). Relaxin, ca-125, progesterone, estradiol, schwangerschaft protein, and human chorionic gonadotropin as predictors of outcome in threatened and nonthreatened pregnancies. Fertility and sterility 53, 1029–1036.

See Also

roc, linkhealthy.

`ranked` *Function used by MAMSE to transform the data into rescaled ranks.*

Description

Functions used by MAMSE to convert the data into rescaled ranks. By default, `ranked=function(x){ rank(x)/(length(x` The user can redefine the function according to other rescalings.

`roc` *Receiver Operating Characteristic (ROC) Curves*

Description

Computes the ROC curve (nonparametric or parametric based on likelihood) for single populations or a weighted ROC curve for lists of populations. The MAMSE weights are used by default for the multiple populations case.

Usage

```
roc(healthy, diseased, wh=NULL, wd=NULL, FPR=NULL, method="np",
   smalldiseased=TRUE, AUC=FALSE, nFPR=201)
```

Arguments

`healthy`	A single numeric vector with the values of the diagnostic variable for the healthy group, or a list of `m` samples (each a numeric vector) from healthy subjects from different populations. When relevant (when MAMSE weights are used), the first sample (`healthy[[1]]`) is deemed to come from the population of interest and the `m-1` other samples are used to borrow strength.

`diseased` A single numeric vector with the values of the diagnostic variable for the diseased group, or a list of `m` samples (each a numeric vector) from diseased subjects from different populations. The number of populations in `healthy` and `diseased` must match, and it is assumed that they are presented in the same order (i.e. the `j^{th}` element of both lists are from the same population. When relevant (when MAMSE weights are used), the first sample (`healthy[[1]]`) is deemed to come from the population of interest and the `m-1` other samples are used to borrow strength.

`wh` Weights for the healthy population. If healthy is a vector, `wh` is a numeric vector of the same length that sums to one and if `wh` is `NULL`, equal weights are given to each datum. If healthy is a list, `wh` is a numeric vector of length `m` that sums to one and if `wh` is `NULL`, MAMSE weights are calculated.

`wd` Weights for the diseased population. If healthy is a vector, `wd` is a numeric vector of the same length that sums to one and if `wd` is `NULL`, equal weights are given to each datum. If healthy is a list, `wh` is a numeric vector of length `m` that sums to one and if `wh` is `NULL`, MAMSE weights are calculated.

`FPR` Numeric vector giving the values of `FPR` (the x-axis) where the ROC curve should be computed. If `FPR` is `NULL`, the default is to keep every step in the nonparametric settings, or to split the `[0,1]` interval in `nFPR` steps (keeping both 0 and 1).

`method` Allowed values are `"np"` for nonparametric ROC curves, `"lognormal"` for a parametric curve based on the log-normal distribution or `"normal"` for a parametric curve based on the normal distribution. In the parametric cases, plug-in estimates are used with the (possibly weighted) likelihood.

`smalldiseased` By default, it is assumed that diseased subjects tend to have smaller values than healthy ones, but `smalldiseased=FALSE` can be used when diseased present large values of the diagnostic variable.

`AUC` If `AUC=TRUE`, the Area Under the Curve will be calculated and returned. Note that AUC will not be calculated if a manually provided `FPR` does not start at 0 and end at 1.

`nFPR` If `FPR` is not provided in the parametric setting, it will be generated with equal steps between 0 and 1 (including those bounds).

Details

This function returns the ROC curve based on the provided data sets. The method can be either parametric (normal or log-normal) or nonparametric. Multiple samples can be used and weighted. MAMSE weights are used by default. The first sample appearing in the lists of data is then deemed to come from the population of interest. The function returns a list of point (FPR,TPR) that can be plotted to see the ROC curve. The points where the function is evaluated can be controlled by specifying FPR manually. By default, it is assumed that small values of the diagnostic variable indicate a disease, but the option `smalldiseased` can be used if small values are for healthy subjects.

Value

S3 object of type `roc` which is a list with the values TPR (vector with true positive rates for different thresholds), FPR (false positive rate for the corresponding threshold) and AUC (Area under the ROC curve). A method for plot has been defined for easier display (see exemples below).

References

J.-B. Débordès & J.-F. Plante (2009). Combining ROC curves using MAMSE weighted distributions. *Cahier du GERAD G-2015-69*.

See Also

MAMSE-package, MAMSE.

Examples

```
data(Progesterone)
healthy=lapply(Progesterone,function(x){x$viable})
diseased=lapply(Progesterone,function(x){sort(c(x$ecto,x$abort))})

par(mfrow=c(2,2))

plot(roc(healthy[[1]],diseased[[1]],AUC=TRUE))
title("Empirical ROC curve based on Ledger (1994)")
plot(roc(healthy[[1]],diseased[[1]],AUC=TRUE,method="lognormal"))
title("Parametric ROC curve based on Ledger (1994)")

plot(roc(healthy,diseased,AUC=TRUE))
title("MAMSE-weighted empirical ROC curve")
plot(roc(healthy,diseased,AUC=TRUE,method="lognormal"))
title("MAMSE-weighted parametric ROC curve")
```

WKME *Kaplan-Meier Estimate*

Description

Computes the weighted Kaplan-Meier estimate over some time points with optional confidence intervals.

Usage

```
WKME(x,ub,lb=0,time=NULL,boot=NULL,REP=1000)
```

Arguments

x	A list of m samples. Each element is an n by 2 matrix whose second column is an indicator of whether the time in column 1 is observed (1) or censored (0).
lb,ub	Lower and upper bounds of the integral of the MAMSE criterion.
time	A vector of times at which to compute the Kaplan-Meier estimate.
boot	When NULL, bootstrap confidence intervals are not generated. Otherwise must be a number in (0,1) corresponding to the coverage probability of the bootstrap intervals to be built.
REP	When bootstrap is used, controls the number of pseudo-sample to generate.

Details

This function calculates the weighted Kaplan-Meier estimate and can provide pointwise bootstrap confidence intervals.

Value

List of elements:

`x`	Sorted list of the times (observed and censored) from each samples
`weight`	The size of the jump that the Kaplan-Meier estimate allocates to each time in `x`.
`time`	Vector of time points where the function is evaluated.
`kme`	The Kaplan-Meier estimate for Population 1 evaluated at `time`.
`kmeCI`	Pointwise bootstrap confidence interval for `kme`.
`wkme`	The weighted Kaplan-Meier estimate evaluated at `time`.
`wkmeCI`	Pointwise bootstrap confidence interval for `wkme`.

References

J.-F. Plante (2007). *Adaptive Likelihood Weights and Mixtures of Empirical Distributions*. Unpublished doctoral dissertation, University of British Columbia.

J.-F. Plante (2009). About an adaptively weighted Kaplan-Meier estimate. *Lifetime Data Analysis*, 15, 295-315.

See Also

MAMSE-package, WKME.

Examples

```
set.seed(2009)
x=list(
  cbind(rexp(20),sample(c(0,1),20,replace=TRUE)),
  cbind(rexp(50),sample(c(0,1),50,replace=TRUE)),
  cbind(rexp(100),sample(c(0,1),100,replace=TRUE))
)

allx=pmin(1,c(x[[1]][x[[1]][,2]==1,1],x[[2]][x[[2]][,2]==1,1],
   x[[3]][x[[3]][,2]==1,1]))
K=WKME(x,1,time=sort(unique(c(0,1,allx,allx-.0001))),boot=.9,REP=100)
# Only 100 bootstrap repetitions were used to get a fast enough
# calculation on a CRAN check.

plot(K$time,K$wkme,type='l',col="blue",xlab="x",
   ylab="P(X<=x)",ylim=c(0,1))
lines(K$time,K$kme[,1],col="red")

lines(K$time,K$wkmeCI[1,],lty=2,col="blue")
lines(K$time,K$wkmeCI[2,],lty=2,col="blue")
```

```
lines(K$time,K$kmeCI[1,],lty=2,col="red")
lines(K$time,K$kmeCI[2,],lty=2,col="red")
legend(.1,.9,c("Weighted Kaplan-Meier","Kaplan-Meier"),
   col=c("blue","red"),lty=c(1,1))
```

Index