Package ‘freqMAP’

February 19, 2015

Type Package

Title Frequency Moving Average Plots (MAP) of Multinomial Data by a Continuous Covariate

Version 0.2

Date 2009-12-30

Author Colin McCulloch

Maintainer Colin McCulloch <colin.mcculloch@themccullochgroup.com>

Description A frequency moving average plot (MAP) is estimated from a multinomial data and a continuous covariate. The frequency MAP is a moving average estimate of category frequencies, where frequency means and posterior bounds are estimated. Comparisons of two frequency MAPs as well as odds ratios can be plotted.

License GPL (>= 2)

LazyLoad yes

Repository CRAN

Date/Publication 2012-10-29 08:58:47

NeedsCompilation no

R topics documented:

freqMAP-package ........................................... 2
freqMAP ....................................................... 3
genotypebyage ............................................. 6
plot.freqMAP .................................................. 7
posterior.comparison.freqMAP .......................... 10
print.freqMAP ................................................ 12
summary.freqMAP ......................................... 13

Index 14
Description

This package can be used to estimate a frequency moving average plots (MAP) from multinomial data and a continuous covariate. The frequency MAP is a moving average estimate of category frequencies, where frequency means and posterior bounds are estimated. Comparisons of two frequency MAPs as well as odds ratios can be plotted.

Details

Package: freqMAP
Type: Package
Version: 0.1
Date: 2008-09-24
License: GPL (>= 2)

This package allows one to visualize frequency data (e.g. genotype) as a function of a continuous covariate (e.g. age). Two populations can be visualized on the same plot and inference can be made about significant differences in the frequencies and odds ratios comparing two unmatched populations. This can be useful for exploring possible confounding effects in a case control study. In this application, one frequency MAP would be used for cases and another MAP for controls. Then one would plot the frequency MAP of, say, genotype frequencies as a function of age in both groups. Interactions between age and genotype would show up as non-parallel frequency MAPs in the two groups. This information can then be used for formulating interactions in the subsequent logistic regression modeling.

At each of a pre-specified set of values of the continuous covariate, the frequency data is modeled as Multinomial with an unknown true category frequency vector. The prior on the true category frequency vector is assumed to be uniform (Dirichlet(1,...,1)). Then samples are generated from the Dirichlet posterior distribution on the true category frequency vector. The central 95% posterior interval on all true category frequencies are estimated from the posterior samples.

The function `freqMAP` creates a `freqMAP` object and samples from the posterior on true category frequencies. The object can then be plotted using `plot.freqMAP`. The posterior distributions from two unmatched populations contained in two `freqMAP` objects can be compared using `posterior.comparison.freqMAP`. `plot.freqMAP` can also plot two populations at once as well as odds ratios comparing the two populations.

The example below analyzes a typical dataset one would find in a genetic association study, where genotypes are collected for a group of probands and controls of varying ages. See `genotypebyage` for more details on the dataset. The example code shows how to generate MAP plots by genotype and by allele.
freqMAP

Author(s)
Colin McCulloch <colin.mcculloch@themccullochgroup.com>

References

Examples

data("genotypebyage")

####An analysis by genotype

#Create a freqMAP object for probands and controls, respectively.
fm.cont <- freqMAP(genotypebyage[,c("age","genotype")],
                   x=seq(25,95,by=5),x.label="age",hw=5)
fm.prob <- freqMAP(genotypebyage[,c("age","genotype")],
                   x=seq(25,95,by=5),x.label="age",hw=5)

#Plot the two MAPs overlaid.
plot(fm.cont,fm.prob,legend=c("controls","probands"),show.p.value.legenh=TRUE)

####An analysis by allele

#Reformat the dataset to stack the two alleles into one column. The
#resulting dataset has twice as many rows as genotypebyage
allelebyage <- rbind(cbind(genotypebyage[,c("age","subject.type")],
                      allele=genotypebyage$allele1.stringsAsFactors=FALSE),
                     cbind(genotypebyage[,c("age","subject.type")],
                      allele=genotypebyage$allele2.stringsAsFactors=FALSE))

#Create a freqMAP object for probands and controls, respectively.
fm.cont <- freqMAP(allelebyage[,c("age","allele")],
                   x=seq(25,95,by=5),x.label="age",hw=5)
fm.prob <- freqMAP(allelebyage[,c("age","allele")],
                   x=seq(25,95,by=5),x.label="age",hw=5)

plot(fm.cont,fm.prob,legend=c("controls","probands"),show.p.value.legenh=TRUE)
Description

This function creates a frequency MAP object from categorical data and a continuous covariate. The frequency MAP is a moving average estimate of category frequencies, where frequency means and posterior bounds are estimated.

Usage

freqMAP(dat, x, x.label, hw, cat.names = NULL, cat.short = NULL, num.samples = 1e+05)

Arguments

dat A dataframe with two columns. The first column should have the numeric value over which we are calculating the moving average (e.g. age), the 2nd column is a string vector giving the category (e.g. genotype).
x A numeric vector at which to calculate the moving average (relative to the first column of dat).
x.label A string which will be used to name the covariate column (first column) of the output moving average dataframe, cat.ma. The name of the covariate column in dat is not used for this purpose.
hw The half-width of the moving average. See Details below.
cat.names Optional. The categories to analyze. The default behavior is to use all unique values in dat[,2]. See Details below.
cat.short Optional. A string vector of the same length as cat.names which gives short forms of the category names. These shortforms will be used in subsequent plots.
um.samples The number of samples to generate from the posterior on the true category frequencies.

Details

The following calculations are performed independently for each element of x:

First, a frequency moving average is generated by binning the category data in dat[,2] into buckets x[i]+/-hw by the value in dat[,1]. Then the observed category frequency is tabulated for all categories. (Note that with certain choices of x and hw, data in dat[,2] can be counted multiple times in multiple buckets.) Next, the frequency data is modeled as Multinomial with an unknown true category frequency vector. The prior on the true category frequency vector is assumed to be uniform ( Dirichlet(1,...,1) ). Samples are generated from the Dirichlet posterior distribution on the true category frequency vector. The central 95% posterior interval (CPI) on all true category frequencies is estimated from the posterior samples. See Value below for details on how results are tabulated.

If it is specified, cat.names must contain at least all of the unique values of dat[,2]. It can be useful to define extra elements of cat.names if you know that there are other possible categories that were not observed in dat[,2] simply due to finite sampling of low probability categories.

Along with the example given below, see freqMAP-package for an example based on genotype data.
Value

A list with the following elements:

- **cat.ma**: A dataframe with the following columns. \( x \) gives the \( x \) argument passed in. \( n \) gives the number of category observations falling in the bucket centered at \( x \). Columns titled by the elements of \( \text{cat.names} \) are the observed frequencies for each category. Columns with names ending in \( .lpi \) and \( .upi \) give the lower and upper bounds of the 95% CPI for each category.

- **post.samples**: Three dimensional array containing the posterior samples. The first dimension indexes the sample, the second dimension indexes the category, the third dimension indexes the bucket centered at \( x \).

- **cat.names**: The category names analyzed.

- **cat.short**: Shortforms of the category names. If not supplied, then this will equal \( \text{cat.names} \).

- **hw**: The bucket halfwidth passed in.

- **x.label**: The name of the continuous covariate. All functions using the object created by this function will search for element \( \text{obj}\$\text{cat.ma}[\text{"obj}\$\text{x.label"}] \) as the continuous covariate.

The returned object has class \( \text{c("freqMAP","list")} \).

Author(s)

Colin McCulloch <colin.mcculloch@themccullochgroup.com>

References


See Also

- `plot.freqMAP`
- `summary.freqMAP`
- `posterior.comparison.freqMAP`

Examples

```r
# Make two sets of 2-category frequency data, y1 & y2, which both vary as
# a function of a continuous variable x
x <- runif(2000, min=-2, max=2)
y1 <- c("a","b")[(1+rbinom(n=length(x), size=rep(1,length(x)), prob=dnorm(x/2))]
y2 <- c("a","b")[(1+rbinom(n=length(x), size=rep(1,length(x)), prob=dnorm(x/5))]

# Create the frequency MAP objects for y1 and y2
fp1 <- freqMAP(data.frame(x=x, y=y1, stringsAsFactors=FALSE),
               x=seq(-2,2,by=.2), x.label="x", hw=.2)
fp2 <- freqMAP(data.frame(x=x, y=y2, stringsAsFactors=FALSE),
               x=seq(-2,2,by=.2), x.label="x", hw=.2)
```
# Examine the frequency MAP objects
summary(fp1)
print(fp2)

# Compare the posterior distributions on the two frequency MAPs
pc <- posterior.comparison.freqMAP(group1=fp1,group2=fp2)

# Three example plots
plot(fp1,ylim=matrix(c(0,1),nrow=length(fp1$cat.names),ncol=2,byrow=TRUE))
plot(fp1,fp2,type="freq",legend=c("y1","y2"),show.p.value.legend=TRUE)
plot(fp1,fp2,type="or")

---

Simulated Genotype Data by Age for 4000 Probands and 4000 Controls

Description

A simple simulated dataset comprising 4000 probands and 4000 controls, where subject age is tabulated against their genotype.

Usage

data(genotypebyage)

Format

A data frame with 8000 observations on the following 5 variables.

age  Subject age (a numeric vector)
subject.type  Subject type: proband or control (a character vector)
al allele1  Subject’s first allele: "a" or "b" (a character vector)
allele2  Subject’s second allele: "a" or "b" (a character vector)
genotype  Subject’s genotype, coded "allele1.allele2" (a character vector)

Details

This dataset is used for the examples in the main help page for the freqMAP: freqMAP-package. See there for usage details. In MAP plots as a function of age, the genotype frequency curves in this dataset demonstrate non-homogeneous frequency patterns. For instance, at ages above 70, the "b.b" genotype becomes more frequent in probands than in younger probands. To compensate, the "a.a" genotype is less frequent in older probands than younger ones. The "a.b" genotype in probands is more homogeneous over all ages. Control genotype frequencies are also relatively homogeneous over all ages.
References


Examples

data(genotypebyage)

plot.freqMAP

Plotting Frequency Moving Average Plot (MAP) Objects

Description

This function can plot a single frequency MAP object, can overlay two frequency MAP object plots, and can plot odds ratios (OR) comparing two frequency MAP objects.

Usage

## S3 method for class 'freqMAP'
plot(x, y = NULL, xlim = NULL, ylim = NULL,
     legend = NULL,
     show.p.value.legend = FALSE,
     type = "freq",
     p.value.bar.alpha = c(0.05, 0.01),
     p.value.bar.color = c("gray90", "darkgray"),
     cex = 1, cex.axis = 1, cex.lab = 1, cex.main = 1,
     pch.x = 2, lty.x = 1, lwd.x = 1, col.x = "red",
     pch.y = 1, lty.y = 2, lwd.y = 1, col.y = "blue",
     cex.legend = 1,
     layout.matrix = NULL,...)

Arguments

x  
A freqMAP object, usually the result of a call to \texttt{freqMAP}

y  
An optional freqMAP object, usually the result of a call to \texttt{freqMAP}

xlim  
Limits on the x-axis. Must be a vector of length 2. The same limits will be used for all category or OR plots. If not specified, then the function will calculate reasonable limits.

ylim  
Limits on the y-axis. If type="freq", then ylim must be a matrix with 2 columns and as many rows as the length of \texttt{x$cat.names}. In this case, row i will be used as the y limits on the frequency plot of category i. If type="or", then ylim must be a vector of length 2 specifying the common y limits for all odds ratio plots. If ylim is not specified, then the function will calculate reasonable limits which will be common on all plots.
plot.freqMAP

Legend

A 2-vector of strings specifying legend text for the x and y freqMAP objects when type="freq". This argument is ignored when type="or" or if y is not specified.

show.p.value.legend

When set to TRUE, a legend is generated for the two levels of Bayesian p-values plotted. This argument is ignored when y is not specified.

type

The type of plot to generate. Must be either "freq" or "or". See Details below.

p.value.bar.alpha

A 2-vector of p-value cutoffs for the p-value bar at the bottom of plots. This argument is ignored if y is not specified. See Details below.

p.value.bar.color

A 2-vector of strings specifying the shading colors for the p-value bar at the bottom of plots. This argument is ignored if y is not specified. See Details below.

cex

The size of plotting characters.

cex.axis

The value passed to cex.axis in all plots. See ?par for details.

cex.lab

The value passed to cex.lab in all plots. See ?par for details.

cex.main

The value passed to cex.main in all plots. See ?par for details.

pch.x

The plotting character used for the freqMAP object in x.

lty.x

The line type lty for the posterior bounds in the freqMAP object in x.

lwd.x

The line width lwd for the posterior bounds in the freqMAP object in x.

col.x

The plotting color for the freqMAP object in x.

pch.y

The plotting character used for the freqMAP object in y.

lty.y

The line type lty for the posterior bounds in the freqMAP object in y.

lwd.y

The line width lwd for the posterior bounds in the freqMAP object in y.

col.y

The plotting color for the freqMAP object in y.

cex.legend

The cex parameter passed to all legends. See ?legend for details.

layout.matrix

The matrix passed to layout() prior to plotting. When this argument is not used, the default behavior is to show all plots on the same figure stacked vertically.

Details

Three types of plots are possible. First, if y is not specified, then a frequency MAP will be plotted for each category in the x object. The mean is plotted surrounded by 95% central posterior intervals (CPI). Second, if y is specified and type="freq", then a frequency MAP will be plotted for each category in the x and y objects, where the two frequency MAPs are overlaid on each category plot. Third, if y is specified and type="or", then a plot is made for all odds ratios comparing each pair of categories. The posterior mean OR and 95% CPI on the OR are plotted. The title of each OR plot gives the direction of the OR. For example, if there are two categories "a" and "b", and the plot title says "b" vs. "a", this indicates the OR plotted is \( \frac{\hat{t}_{xb} \hat{t}_{ya}}{\hat{t}_{xa} \hat{t}_{yb}} \), where \( \hat{t}_{xj} \) and \( \hat{t}_{yj} \)
indicate the true frequency of category j in the x and y MAPs, respectively. If type="or", plotting characters, line types, line widths, and colors are controlled by pch.x, lty.x, lwd.x, and col.x.

If y is specified, then x and y must have the same values in their elements [["cat.names"]], [["cat.short"]], and [["hw"]]. The [["x.label"]]] element can be different. (This element indicates the name of the column in the [["cat.ma"]]] element which contains the x values when plotting.)

When y is specified, a gray bar is added to the bottom of the plot when the two frequency plots are "significantly" different. When type="freq", significance is defined as the posterior Pr(category frequency in MAP x > category frequency in MAP y) > alpha/2 or < (1-alpha/2). When type="or", significance is defined as the posterior Pr( OR(category 1 vs. category 2) > 1 ) > alpha/2 or < (1-alpha/2). The value of alpha is defined at two levels by the p.value.bar .alpha argument and plotted in light gray (higher alpha) and dark gray (lower alpha). The shading colors of the two alpha levels can be overridden by the p.value.bar .color argument. This can be thought of as a Bayesian analog to a two-sided p-value. See posterior.comparison.freqMAP for further details.

Specifying a legend or asking for a p-value legend, while not specifying xlim, will cause the x axis limits to be expanded 20% on the left side to give room for the legend.

Along with the example given below, see freqMAP-package for an example based on genotype data.

Author(s)

Colin McCulloch <colin.mcculloch@themccullochgroup.com>

References


See Also

freqMAP, summary.freqMAP, layout

Examples

# Make two sets of 2-category frequency data, y1 & y2, which both vary as
# a function of a continuous variable x
set.seed(5)
x <- runif(2000, min=-2, max=2)
y1 <- c("a","b")%*%rbinom(n=length(x),size=rep(1,length(x)),prob=pnorm(x/2))
y2 <- c("a","b")%*%rbinom(n=length(x),size=rep(1,length(x)),prob=pnorm(x/5))

# Create the frequency MAP objects for y1 and y2
fp1 <- freqMAP(data.frame(x=x,y=y1,stringsAsFactors=FALSE),
x=seq(-2,2,by=.2),x.label="x",hw=.2)
fp2 <- freqMAP(data.frame(x=x,y=y2,stringsAsFactors=FALSE),
x=seq(-2,2,by=.2),x.label="x",hw=.2)
# Examine the frequency MAP objects
summary(fp1)
print(fp2)

# Compare the posterior distributions on the two frequency MAPs
pc <- posterior.comparison.freqMAP(group1=fp1, group2=fp2)

# Three example plots
plot(fp1, ylim=matrix(c(0, 1), nrow=2, ncol=2, byrow=TRUE))
plot(fp1, fp2, type="freq", legend=c("y1", "y2"), show.p.value.legend=TRUE)
plot(fp1, fp2, type="freq", legend=c("y1", "y2"), show.p.value.legend=TRUE,
     p.value.bar.color=c("green", "yellow"))
plot(fp1, fp2, type="or")

---

posterior.comparison.freqMAP  

*Compare Two Frequency Moving Average Plot (MAP) Objects*

**Description**

This function compares the posterior distributions on two frequency MAPs from two unmatched populations. Posterior summaries are calculated of the population frequency differences and log odds ratios.

**Usage**

```r
posterior.comparison.freqMAP(group1, group2)
```

**Arguments**

- `group1`: A freqMAP object
- `group2`: A freqMAP object

**Details**

Posterior summaries are calculated by simply comparing the posterior samples inside the two freqMAP objects. The probability that group1 frequency is greater than group2 frequency is calculated, the log odds ratio mean and 95% central posterior interval (CPI) is also calculated for each pair of categories.

For any given category, the posterior distributions on the frequencies in the two populations are assumed independent. Namely the category data in the two populations is assumed to be unmatched.

The two frequency MAPs `group1` and `group2` must have the same values in their elements `[["cat.names"],["hw"],` and `[["x.label"]]]`. Further, the values of the continuous covariate in `...$cat.ma[..]$x.label]` must be the same in both groups.
Value

Dataframe with the following columns:

- **x**: The values of the continuous covariate.
- **n1**: The number of observations in group 1 falling into the bucket centered at \( x[i] \).
- **n2**: The number of observations in group 2 falling into the bucket centered at \( x[i] \).
- **.gr1.gt.gr2**: (One column for each category) The posterior probability that group 1 frequency for this category is greater than the frequency in group 2.
- **.*.lor.mean**: (One column for each possible pair of categories) The posterior mean on the log odds ratio for each pair of categories.
- **.*.lor.lpi & .* lor.cpi**: (One column for each possible pair of categories) The posterior lower (lpi) and upper bounds (upi) on the CPI on the log odds ratio for each pair of categories.

Author(s)

Colin McCulloch <colin.mcculloch@themccullochgroup.com>

See Also

`freqMAP`, `plot.freqMAP`

Examples

```r
# Make two sets of 2-category frequency data, y1 & y2, which both vary as a function of a continuous variable x
x <- runif(2000, min=-2, max=2)
y1 <- c("a","b")[(1+rinom(n=length(x),size=rep(1,length(x))),prob=pnorm(x/2))]
y2 <- c("a","b")[(1+rinom(n=length(x),size=rep(1,length(x)),prob=pnorm(x/5)))]

# Create the frequency MAP objects for y1 and y2
fp1 <- freqMAP(data.frame(x=x,y=y1,stringsAsFactors=FALSE), x=seq(-2,2,by=.2),x.label="x",hw=.2)
fp2 <- freqMAP(data.frame(x=x,y=y2,stringsAsFactors=FALSE), x=seq(-2,2,by=.2),x.label="x",hw=.2)

# Examine the frequency MAP objects
summary(fp1)
print(fp2)

# Compare the posterior distributions on the two frequency MAPs
pc <- posterior.comparison.freqMAP(group1=fp1,group2=fp2)

# Three example plots
plot(fp1,ylim=matrix(c(0,1),nrow=length(fp1$cat.names),ncol=2,byrow=TRUE))
plot(fp1,fp2,type="freq",legend=c("y1","y2"),show.p.value.legend=TRUE)
plot(fp1,fp2,type="or")
```
print.freqMAP  

Printing Frequency MAP Objects

Description

Print method for freqMAP objects.

Usage

```r
## S3 method for class 'freqMAP'
print(x, ...)
```

Arguments

- `x` An object of class freqMAP, usually a result of a call to `freqMAP`
- `...` Not used at this time.

Details

`print` prints the result of `summary(x)`

Examples

```r
# Make two sets of 2-category frequency data, y1 & y2, which both vary as
# a function of a continuous variable x
x <- runif(2000, min=-2, max=2)
y1 <- c("a","b")[1+rbinom(n=length(x), size=rep(1,length(x)), prob=pnorm(x/2))]
y2 <- c("a","b")[1+rbinom(n=length(x), size=rep(1,length(x)), prob=pnorm(x/5))]

# Create the frequency MAP objects for y1 and y2
fp1 <- freqMAP(data.frame(x=x,y=y1,stringsAsFactors=FALSE),
               x=seq(-2,2,by=.2), x.label="x", hw=.2)
fp2 <- freqMAP(data.frame(x=x,y=y2,stringsAsFactors=FALSE),
               x=seq(-2,2,by=.2), x.label="x", hw=.2)

# Examine the frequency MAP objects
summary(fp1)
print(fp2)

# Three example plots
plot(fp1, ylim=matrix(c(0,1), nrow=2, ncol=2, byrow=TRUE))
plot(fp1,fp2,type="freq", legend=c("y1","y2"), show.p.value.legend=TRUE)
plot(fp1,fp2,type="or")
```
**summary.freqMAP**

**Summarizing Frequency MAP Objects**

**Description**

Summary method for freqMAP objects.

**Usage**

```r
## S3 method for class 'freqMAP'
summary(object,...)
```

**Arguments**

- `object` An object of class `freqMAP`, usually a result of a call to `freqMAP`.
- `...` Not used at this time.

**Value**

`summary` returns the category moving average table of the object, element `[["cat.ma"]]

**Examples**

```r
# Make two sets of 2-category frequency data, y1 & y2, which both vary as a function of a continuous variable x
x <- runif(2000, min=-2, max=2)
y1 <- c("a","b")[1+rbinom(n=length(x),size=rep(1,length(x)),prob=pnorm(x/2))]
y2 <- c("a","b")[1+rbinom(n=length(x),size=rep(1,length(x)),prob=pnorm(x/5))]

# Create the frequency MAP objects for y1 and y2
fp1 <- freqMAP(data.frame(x=x,y=y1,stringsAsFactors=FALSE),
                x=seq(-2,2,by=.2),x.label="x",hw=.2)
fp2 <- freqMAP(data.frame(x=x,y=y2,stringsAsFactors=FALSE),
                x=seq(-2,2,by=.2),x.label="x",hw=.2)

# Examine the frequency MAP objects
summary(fp1)
print(fp2)

# Three example plots
plot(fp1,ylim=matrix(c(0,1),nrow=2,ncol=2,byrow=TRUE))
plot(fp1,fp2,type="freq",legend=c("y1","y2"),show.p.value.legend=TRUE)
plot(fp1,fp2,type="or")
```
Index

*Topic attribute
  print.freqMAP, 12
  summary.freqMAP, 13
*Topic datasets
  genotypebyage, 6
*Topic hplot
  plot.freqMAP, 7
*Topic htest
  freqMAP, 3
  freqMAP-package, 2
  posterior.comparison.freqMAP, 10
*Topic models
  freqMAP, 3
  freqMAP-package, 2
  posterior.comparison.freqMAP, 10
*Topic multivariate
  freqMAP, 3
  freqMAP-package, 2
*Topic package
  freqMAP-package, 2
*Topic smooth
  freqMAP, 3
  freqMAP-package, 2
  freqMAP, 2, 3, 7, 9, 11–13
  freqMAP-package, 2
  genotypebyage, 2, 6
  layout, 9
  plot.freqMAP, 2, 5, 7, 11
  posterior.comparison.freqMAP, 2, 5, 9, 10
  print.freqMAP, 12
  summary.freqMAP, 5, 9, 13