Package 'aberrant'

August 16, 2011

Type Package

Title A robust clustering algorithm for outlier identification

Version 1.0

Depends R (>= 2.13.0), MCMCpack, mvtnorm, ellipse

Date 2011-08-11

Author Celine Bellenguez and Chris CA Spencer

Maintainer Chris CA Spencer <chris.spencer@well.ox.ac.uk>

Description High-throughput genotyping arrays provide an efficient way to survey single nucleotide polymorphisms (SNPs) across the genome in large numbers of individuals. Downstream analysis of the data, for example in genome-wide association studies (GWAS), often involves statistical models of genotype frequencies across individuals. The complexities of the sample collection process and the potential for errors in the experimental assay can lead to biases and artefacts in an individual's inferred genotypes. Rather than attempting to model these complications, it has become standard practice to remove individuals whose genome-wide data differs from the sample at large. This package contains a simple, but robust,clustering algorithm to identify samples with atypical summaries of genome-wide variation.

License GPL (>= 2)

LazyLoad yes

R topics documented:

aberrant.corr 2
aberrant.ind 3
aberrant.plot 4

Index **6**

`aberrant.corr` *Outlier identification for correlated summary statistics*

Description

Outlier identification based on two summary statistics. A correlation between the two statistics is modeled.

Usage

```
aberrant.corr(x, lambda, niter=10000, prior_df, prior_scale,
      alpha, beta, standardize = TRUE, verbose = TRUE)
```

Arguments

`x`	Vector, matrix or numeric data frame, with two summary statistics in column.
`lambda`	Ratio of the standard deviations of outlying and normal individuals.
`niter`	Number of samples to be generated by the Gibbs sampling. Default to 10000.
`prior_df`	Degree of freedom of the Inverse-Wishart distribution that describes prior information on the covariance matrix of the means of the distributions describing variability of the summary statistics.
`prior_scale`	Scale matrix of the Inverse-Wishart distribution that describes prior information on the covariance matrix of the means of the distributions describing variability of the summary statistics.
`alpha`	First shape parameter of the Beta distribution describing prior information on the probability that an individual is an outlier.
`beta`	Second shape parameter of the Beta distribution describing prior information on the probability that an individual is an outlier.
`standardize`	A logical indicating whether the summary statistics should be standardized.
`verbose`	logical. If TRUE, verbose output is generated during the Gibbs sampling.

Value

`group`	Vector indicating whether an individual is an outlier (=1) or not (=0). Individuals are in the same order as in the initial data matrix x.
`posterior`	Vector indicating the posterior probability for an individual to be an outlier.
`lambda`	Ratio of the standard deviations of outlying and normal individuals used.
`post_mean`	Posterior mean of the summary statistics for the normal individuals.
`post_var`	Posterior covariance matrix of the summary statistics for the normal individuals.
`standardize`	Logical indicating if summary statistics were standardized.
`inlier`	Indices of normal individuals.
`outlier`	Indices of outlying individuals.

Author(s)

Celine Bellenguez and Chris CA Spencer

References

Celine Bellenguez, Amy Strange, Colin Freeman, Wellcome Trust Case Control Consortium 2, Chris CA Spencer. A robust clustering algorithm for identifying problematic samples in genome-wide association studies. Bioinformatics.

Examples

```
x<-rmvt(1000, sigma=matrix ( c ( 1, 0.5 , 0.5 , 1 ) , 2 , 2 ), df=3)
aberrant.corr(x, lambda=10 , prior_df=10 , prior_scale=diag(1,2),
     alpha = 1 , beta = 20 )
```

---

aberrant.ind *Outlier identification for independent summary statistics*

Description

Outlier identification based on one summary statistic or two summary statistics that are considered independent.

Usage

```
aberrant.ind(x, lambda, niter=10000, prior_df, prior_scale,
     alpha, beta, standardize = TRUE, verbose = TRUE)
```

Arguments

x	Vector, matrix or numeric data frame, with one or two summary statistics in column.
lambda	Ratio of the standard deviations of outlying and normal individuals.
niter	Number of samples to be generated by the Gibbs sampling. Default to 10000.
prior_df	Vector of degrees of freedom for the Scaled Inverse Chi-Square distribution that describes prior information on the variances of the means of the distributions describing variability of the summary statistics.
prior_scale	Vector of scale parameters for the Scaled Inverse Chi-Square distribution that describes prior information on the variances of the means of the distributions describing variability of the summary statistics.
alpha	First shape parameter of the Beta distribution describing prior information on the probability that an individual is an outlier.
beta	Second shape parameter of the Beta distribution describing prior information on the probability that an individual is an outlier.
standardize	A logical indicating whether the summary statistics should be standardized.
verbose	logical. If TRUE, verbose output is generated during the Gibbs sampling.

Value

group	Vector indicating whether an individual is an outlier (=1) or not (=0). Individuals are in the same order as in the initial data matrix x.
posterior	Vector indicating the posterior probability for an individual to be an outlier.
lambda	Ratio of the standard deviations of outlying and normal individuals used.
post_mean	Posterior mean of the summary statistics for the normal individuals.
post_var	Posterior covariance matrix of the summary statistics for the normal individuals.
standardize	Logical indicating if summary statistics were standardized.
inlier	Indices of normal individuals.
outlier	Indices of outlying individuals.

Author(s)

Celine Bellenguez and Chris CA Spencer

References

Celine Bellenguez, Amy Strange, Colin Freeman, Wellcome Trust Case Control Consortium 2, Chris CA Spencer. A robust clustering algorithm for identifying problematic samples in genome-wide association studies. Bioinformatics.

Examples

```
x <- rmvt ( 1000 , sigma = diag ( 2 ) , df = 3 )
aberrant.ind ( x, lambda=10, prior_df=c(10,10), prior_scale=c(0.5,0.5),
     alpha=1, beta=20 )
```

---

aberrant.plot *Scatter plot of summary statistics*

Description

Draws a scatter plot of the summary statistics showing outliers.

Usage

```
aberrant.plot(x)
```

Arguments

x	output from function aberrant.ind or aberrant.corr

Details

"Normal" individuals are coloured with a gradation from black to grey, with darker colours denoting higher density of individuals. Outliers are coloured with a gradation from orange to red, with darker colours denoting higher posterior probability of being an outlier. 99% confidence ellipse of the inferred inlier distribution is shown as a dashed grey line.

Author(s)

Celine Bellenguez and Chris CA Spencer

References

Celine Bellenguez, Amy Strange, Colin Freeman, Wellcome Trust Case Control Consortium 2, Chris CA Spencer. A robust clustering algorithm for identifying problematic samples in genome-wide association studies. Bioinformatics.

Examples

```
x<-rmvt(1000, sigma=matrix(c(1,0.5,0.5,1),2,2), df=3)
res<-aberrant.corr(x, lambda=10, prior_df=10, prior_scale=diag(1,2),
     alpha=1, beta=20)
aberrant.plot(res)
```

Index

aberrant.corr, 2
aberrant.ind, 3
aberrant.plot, 4