Package ‘PCFAM’

January 20, 2025

Type Package

Title Computation of Ancestry Scores with Mixed Families and Unrelated Individuals

Version 1.0

Date 2017-03-20

Author Yi-Hui Zhou

Maintainer Yi-Hui Zhou `<yihui_zhou@ncsu.edu>`

Description We provide several algorithms to compute the genotype ancestry scores (such as eigenvector projections) in the case where highly correlated individuals are involved.

License GPL-2

LazyLoad yes

NeedsCompilation no

Repository CRAN

Date/Publication 2017-03-24 19:10:36 UTC

Contents

`PCFAM-package` *Computation of ancestry scores with mixed families and unrelated individuals*

Description

This package provides ancestry scores based on genotype data, and is robust to the presence of close-degree family members. Four main novel algorithms are represented: (i) Geometric rotation (within-family data orthogonalization); (ii) matrix substitution based on the decomposition of a target family-orthogonalized covariance matrix; (iii) covariance-preserving whitening, retaining covariances between unrelated pairs while orthogonalizing family members (Note: the function perfectwhiten generates a new dataset which keeps the same covariance structure as the original set); (iv) using family-averaged data to obtain loadings for projection of family members.

Details

Package:	PCFAM
Type:	Package
Version:	1.0
Date:	2016-10-11
License:	GPL 2
LazyLoad:	yes

Author(s)

Yi-Hui Zhou

Maintainer: Yi-Hui Zhou <yihui_zhou@ncsu.edu>

References

Computation of ancestry scores with mixed families and unrelated individuals. arXiv:1606.08416

Examples

```
X <- matrix(rbinom(1000*20,2,0.4),1000,20)
X[,1]=X[,2]*0.9
X=rowscale(X)
Xresid=residualize(X)
corXresid=cor(Xresid)
myfam=findfamilies(corXresid,0.1)
K=3
myms.pca=ms.pca(X,corXresid,0.1,K)
familyave.result=familyave(X,myfam,top=K)
```

`colcenter` *column centering of the data matrix*

Description

This function centerizes each column of the data matrix

Usage

```
colcenter(X)
```

Arguments

`X`	input data matrix

Value

return the data matrix with each column centered

Author(s)

Yi-Hui Zhou

References

Computation of ancestry scores with mixed families and unrelated individuals. Yi-Hui Zhou, J.S. Marron, Fred Wright, arXiv:1606.08416.

`cov.function` *Sample covariance calculator*

Description

Obtain a sample covariance matrix

Usage

```
cov.function(data.matrix)
```

Arguments

`data.matrix`	Input mxn data matrix

Value

return the nxn sample covariance matrix

Author(s)

Yi-Hui Zhou

References

Computation of ancestry scores with mixed families and unrelated individuals. arXiv:1606.08416.

Examples

```
X <- matrix(rbinom(1000*20,2,0.4),1000,20)
cov.X=cov.function(X)
```

`familyave` *Family average approach*

Description

This function implements the family-averaging algorithm, with loadings based on the combined data from singletons and family averages, then projected to all.

Usage

```
familyave(Xall,myfam, top = 5)
```

Arguments

`Xall`	The original input genotype dataset
`myfam`	The identified family IDs. Each singleton forms his/her own family.
`top`	The number ancestry scores desired.

Details

The function averages the genotype information in each family, re-inflates to have appropriate variability, andtreats as a 'singleton' for the purpose of loading calculation. Ancestry scores are obtained by projection to all.

Value

Output the top ancestry scores by combining family data with singletons

Author(s)

Yi-Hui Zhou

References

Computation of ancestry scores with mixed families and unrelated individuals. arXiv:1606.08416.

Examples

```
X <- matrix(rbinom(1000*20,2,0.4),1000,20)
X[,1]=X[,2]*0.9
X=rowscale(X)
Xresid=residualize(X)
corXresid=cor(Xresid)
myfam=findfamilies(corXresid,0.1)
K=3
familyave.result=familyave(X,myfam,top=K)
```

`fastcov` *Fast covariance function*

Description

This function can generate covariance matrix faster than the regular cov() function.

Usage

```
fastcov(X)
```

Arguments

`X`	input mxn data matrix

Value

Output nxn covariance matrix

Note

The input data matrix has to be column scaled in advance.

Author(s)

Yi-Hui Zhou,

References

Computation of ancestry scores with mixed families and unrelated individuals. arXiv:1606.08416.

`findfamilies` *Find families*

Description

This function searches for pairs of individuals with high kinship based on the genotype correlation matrix.

Usage

```
findfamilies(x, threshold = 0.4)
```

Arguments

`x`	The nxn correlation matrix of the input dataset.
`threshold`	This threshold is used to identify close-degree relatives. Recommended values are 0.4 to identify first-degree relatives, and 0.15 to identify first- and second-degree relatives.

Value

Output numerical family ID for each individual. Individuals with the same ID are judged to be family members.

Author(s)

Yi-Hui Zhou

References

Computation of ancestry scores with mixed families and unrelated individuals. arXiv:1606.08416.

`gr.pca` *The geometric rotation approach*

Description

This algorithm rotates scaled genotypes among family members so that they are mutually orthogonal.

Usage

```
gr.pca(data.input, index.family, myfam, weight, top, family.size, inflation)
```

Arguments

`data.input`	Input dataset, each row is for a genetic feature (SNP), each column is for individual. Data are typically number of minor alleles, possibly imputed.
`index.family`	Index vector to indicate the family id of each individual.
`myfam`	This value comes directly from the output of findfamilies().
`weight`	Weight is 0 by default. This is a deprecated weight value that can be used to control the amount of rotation performed. A weight of zero performs full orthogonalization, while a weight of 1 keeps the data unchanged.
`top`	The number of eigenvectors to be used.
`family.size`	The number of members in each family. Used to determine rotation angles.
`inflation`	The inflation of the data value is 0 under default. Deprecated.

Value

`data.new`	The new datamatrix after the geometric rotation
`topPCs`	The top eigenvectors
`topEigenvalue`	The top eigenvalues.

Author(s)

Yi-Hui Zhou

References

Computation of ancestry scores with mixed families and unrelated individuals. arXiv:1606.08416.

`ms.pca` *The matrix substitution approach*

Description

This function provides the matrix subsitution algorithm. The main idea is to replace the high covariance value entries in the covariance matrix which are produced by family members by a small value (e.g. median covariance).

Usage

```
ms.pca(X, corXresid, threshold, top)
```

Arguments

`X`	The input data matrix
`corXresid`	The correlation of the genotypes after residualization for any evidence of larger scale ancestry. Used to identify close-degree family members in a manner robust to large-scale ancestry.
`threshold`	Covariance values of identified family members are set to the threshold.
`top`	The number of ancestry scores to obtain.

Value

`eigenvector`	Eigenvectors after using the matrix substitution method
`myeigen`	The top eigenvalues and eigenvectors

Author(s)

Yi-Hui Zhou

References

Computation of ancestry scores with mixed families and unrelated individuals. arXiv:1606.08416.

`mysqrtm` *Matrix square root function*

Description

This function can find the matrix square root, without requiring a new package and often faster than other code.

Usage

```
mysqrtm(a, symmetric = F)
```

Arguments

`a`	The input matrix
`symmetric`	Default=FALSE. This argument indicates whether the input matrix is symmetric.

Details

Matrix B is said to be a square root of A if the matrix product BB is equal to A.

Value

returns the square root matrix B

`perfectwhiten` *The covariance preserving whitening function.*

Description

This algorithm generates a new scaled 'genotype' dataset which keeps the same covariance structure as the original data, except that family members have been made orthogonal to each other, and singletons are unchanged.

Usage

```
perfectwhiten(Xun, Xfam, delta = 3e-04, threshold = 0.35, eta = NULL, addfuzz = F)
```

Arguments

`Xun`	A matrix of (possibly scaled) genotypes, (number of SNPs)*(number of singletons)
`Xfam`	A matrix of (possibly scaled) genotypes, (number of SNPs)*(number of individuals belonging to families)
`delta`	A slight offset used to ensure that the target covariance matrix is of full rank
`threshold`	The correlation threshold used to determine pairs of relatives. The choice should be less than the degree desired. For example, 0.35 captures first degree relatives (expected correlation 0.5), 0.15 captures first and second degree relatives (expected correlation for second degree relatives is 0.25).
`eta`	This argument is the replacement value used for matrix substitution. The default is NULL, resulting in substitution by the median.
`addfuzz`	The default is FALSE. Deprecated.

Value

`Xplusscaled`	The row-scaled full genotype data, including both singletons and family members
`Y`	The (scaled) genotype matrix after whitening, and should have a covariance matrix very close to Mtarget. Column means are zero
`Ynotcolcentered`	The same as Y, but with column means matching those of Xplusscaled
`M`	The covariance matrix of the full data
`Mtilde`	The covariance matrix after matrix substitution of all family pairs identified with correlations exceedingeta
`whichbig`	The set of indexes of M that have correlation exceeding threshold
`covY`	The covariance matrix of Y, useful to compare to M or to Mtarget

Author(s)

Yi-Hui ZHou, Fred A. Wright

References

Computation of ancestry scores with mixed families and unrelated individuals. arXiv:1606.08416.

Examples

```
X <- matrix(rbinom(1000*20,2,0.4),1000,20)
X[,1]=X[,2]*0.9
X=rowscale(X)
Xresid=residualize(X)
library(PCFAM)
corXresid=cor(Xresid)
myfam=findfamilies(corXresid,0.1)
K=3
perfect.result=perfectwhiten(X[,which(myfam==0)],X[,which(myfam==1)])
```

`residualize` *Residualization and scale of the original genotype data*

Description

Thus function performs a simple residualization of a row-scaled genotype dataset, removing large-scale population stratification. Output is a residualized dataset appropriate for computing correlations such that family members can be easily identified. The function assumes X is row-scaled

Usage

```
residualize(X)
```

Arguments

`X`	The original input genotype dataset

Details

This function pre-treatment the data before applying the findfamily function.

Value

Outputs the new row-scaled genotype matrix after residualization

Author(s)

Yi-Hui Zhou

References

Computation of ancestry scores with mixed families and unrelated individuals. arXiv:1606.08416.

`rowcol` *Indicators for of the row and column of the original matrix*

Description

This function identifies the rows and columns of elements in a matrix, e.g. the family members identified based on the correlation matrix.

Usage

```
rowcol(I, J, elements)
```

Arguments

`I`	The number of rows of the matrix (scalar)
`J`	The number of columns of the matrix (scalar)
`elements`	A vector of matrix element indexes

Value

`whichrow`	The rows of elements in the matrix
`whichcol`	The columns of elements in the matrix

Author(s)

Yi-Hui ZHou, Fred A. Wright

References

Computation of ancestry scores with mixed families and unrelated individuals. arXiv:1606.08416.

Examples

```
X <- matrix(rbinom(1000*20,2,0.4),1000,20)
X[,1]=X[,2]*0.9
X=rowscale(X)
```

`rowscale` *Scale each row of a matrix*

Description

This function scales the input matrix so that each row mean is 0 and each row (sample) variance is 1.

Usage

```
rowscale(X)
```

Arguments

`X`	input data matrix

Value

Output the row-scaled matrix.

Author(s)

Yi-Hui ZHou, Fred A. Wright

References

Computation of ancestry scores with mixed families and unrelated individuals. arXiv:1606.08416.

Index