Package 'trioGxE'

February 20, 2015

Type Package

Title A data smoothing approach to explore and test gene-environment interaction in case-parent trio data

Version 0.1-1

Date 2013-04-02

Author Ji-Hyung Shin <shin@sfu.ca>, Brad McNeney <mcneney@sfu.ca>, Jinko Graham <jgraham@sfu.ca>

Maintainer Ji-Hyung Shin <shin@sfu.ca>

Depends msm, mgcv, gtools

LazyData true

Description The package contains functions that 1) estimates gene-environment interaction between a SNP and a continuous non-genetic attribute by fitting a generalized additive model to case-parent trio data, 2) produces graphical displays of estimated interaction, 3) performs permutation test of gene-environment interaction; 4) simulates informative case-parent trios.

License GPL-2

NeedsCompilation no

Repository CRAN

Date/Publication 2013-04-03 01:39:24

R topics documented:

hypoTrioDat *Simulated data for a hypothetical example*

Description

A simulated data set with 1000 informative case-parent trios to illustrate the `trioGxE` package.

Usage

```
data(hypoTrioDat)
```

Format

data frame with columns:

	column name	type	details
[,1]	parent 1	numeric	copy number of the putative risk allele carried by the first parent (0, 1, 2)
[,2]	parent 2	numeric	copy number of the putative risk allele carried by the other parent (0, 1, 2)
[,3]	child	numeric	copy number of the putative allele carried by the child (0, 1, 2)
[,4]	subpop	numeric	membership for two subpopulations 0, 1
[,5]	attr	numeric	non-genetic attribute value of the child

Note

The data contains 1000 informative case-parent trios, each of which has at least one heterozygous parent. The trios were generated, using `trioSim` under recessive quadratic gene-environment interaction, from a stratified population composed of two equal-sized subpopulations. The two subpopulations have different distributions for the SNP and the non-genetic attribute. For the SNP, the risk allele frequencies for the first (`subpop=0`) and second (`subpop=1`) subpopulations are 0.1 and 0.9, respectively. For the non-genetic attribute, the distributions in subpopulations 0 and 1 are Normal with subpopulation-specific means -0.8 and 0.8, respectively, and common variance 0.36.

See Also

`trioSim`

plot.trioGxE *Graphical display of gene-environment interaction between a SNP and a continuous non-genetic attribute in case-parent trio data*

Description

The function `plot.trioGxE` uses the calculations made in `trioGxE` and plots the point- and interval-estimates of gene-environment interaction between a single nucleotide polymorphism (SNP) and a continuously varying environmental or non-genetic covariate in case-parent trio data.

Usage

```
## S3 method for class 'trioGxE'
plot(x, se = TRUE, seWithGxE.only = TRUE, ylim = NULL, yscale = TRUE,
     xlab = NULL, ylab = NULL, rugplot = TRUE, ...)
```

Arguments

x	A returned object produced by trioGxE function.
se	A logical or a positive number. When TRUE (default), upper and lower lines are added to the plots at 2 standard errors above and below the fitted values of the interaction functions. When it is a positive number, lines are added at se standard errors above and below the fitted interaction values. When FALSE, no standard error lines are plotted.
seWithGxE.only	If TRUE, the associated standard errors reflect the uncertainty in the estimates of the gene-environment interaction functions only. If FALSE, the standard errors include the uncertainty in the genetic main effect estimates.
ylim	Either a list holding two length-2 numeric vectors that give different y-coordinate ranges for the two plots, or a single length-2 vector that gives equal y-coordinate ranges for both plots.
yscale	If TRUE (default), the same y-axis scale is chosen for each plot. Ignored if ylim is supplied.
xlab	An optional string setting the title for the x-axis.
ylab	An optional string setting the title for the y-axis.
rugplot	Logical indicating whether to add rug representation of the data to the plots. Default (TRUE) adds rugs.
...	Further graphical parameters passed to plot, such as col, lwd, etc.

Details

The function produces two plots in a 2 x 1 layout. The first plot in the left panel displays the estimated gene-environment interaction (GxE) curve related to GRR_1, the genotype relative risks (GRRs) among the individuals with one copy of the putative risk allele compared to those with zero copies. The right panel displays the estimated GxE curve related to GRR_2, the GRRs among the individuals with two copies of the risk allele compared to those with one copy.

When object$penmod="codominant" (with se=TRUE), confidence intervals are plotted for both interaction curves that are related to GRR_1 and GRR_2. When object$penmod="dominant", the confidence intervals are plotted only in the left panel, but not in the right panel because GRR_2 is not estimated but set to be 1 under this penetrance mode. Similarly, when object$penmod="recessive", the confidence intervals are plotted only in the right panel, but not in the right panel because GRR_1 is not estimated but set to be 1 under this penetrance mode. When object$penmod="additive", equivalent confidence intervals are plotted in both panels, which display equivalent fitted curves. This is because GRR_1 and GRR_2 are set to be equivalent under the log-additive or multiplicative penetrance mode.

When se is TRUE or a positive number, standard error lines are plotted based on the calculations of the Bayesian posterior variance estimates of the generalized additive model parameters for GRRs (Wood, 2006).

Author(s)

Ji-Hyung Shin <shin@sfu.ca>, Brad McNeney <mcneney@sfu.ca>, Jinko Graham <jgraham@sfu.ca>

References

Shin, J.-H. (2012): *Inferring gene-environment interaction from case-parent trio data: evaluation of and adjustment for spurious $G \times E$ and development of a data-smoothing method to uncover true $G \times E$*, Ph.D. thesis, Statistics and Actuarial Science, Simon Fraser University: URL https://theses.lib.sfu.ca/sites/all/files/public_copies/etd7214-j-shin-etd7214jshin.pdf.

Wood, S. (2006): *Generalized Additive Models: An Introduction with R*, Boca Raton, FL: Chapman & Hall/CRC.

See Also

trioGxE, test.trioGxE, trioSim

Examples

```
data(hypoTrioDat)

## fitting a co-dominant model to the hypothetical data
simfit <- trioGxE(data=hypoTrioDat,pgenos=c("parent1","parent2"),cgeno="child",cenv="attr",
                  k=c(5,5),knots=NULL,sp=NULL)

## produce the graphical display of the point- and interval-estimates of GxE curve
plot.trioGxE(simfit) # or just plot(simfit)
```

`test.trioGxE` *Test of gene-environment interaction between a SNP and a continuous non-genetic covariate from case-parent trio data.*

Description

Performs permutation test of gene-environment interaction based on the associated penalized maximum likelihood estimates obtained by fitting a generalized additive model to case-parent trio data.

Usage

```
test.trioGxE(object, data = NULL, nreps, level = 0.05, early.stop = FALSE,
             fix.sp = FALSE, output = NULL, return.data = FALSE,
             return.object = FALSE, ...)
```

Arguments

`object`	A returned object from `trioGxE` function. When `NULL`, a data set of case-parent trios must be provided (through 'data' argument).
`data`	Trio data to be passed into `trioGxE` when 'object' is not provided.
`nreps`	Desired number of permutation replicates.
`fix.sp`	When `TRUE`, the approximated null distribution of the test statistic is obtained by computing the test statistic by fitting each simulated data set under fixed values of the smoothing parameters. When `FALSE` (Default), the null distribution is obtained by fitting each simulated data set while simultaneously estimating the smoothing parameter values.
`level`	Desired significance level for the test.
`early.stop`	When `TRUE`, sampling is terminated early when the number of test statistics that are more extreme than or as extreme as the observed test statistic equals `nreps*level`. `nreps` $\times$ `level` values larger than the observed test statistic are obtained.
`output`	A character string specifying the name of the output file that writes the values of the test statistics calculated for the actual and simulated data set. When `NULL` (Default), no written output file is produced.
`return.data`	When `TRUE`, the original data set is returned.
`return.object`	When `TRUE`, the fitting object for the original data set is returned.
`...`	Arguments passed to `trioGxE`: when `data` is provided, instead of `trioGxE` class object, parameters of `trioGxE` must be provided through ....

Details

Suppose k_1 and k_2 are the numbers of knots used to represent the interaction functions f_1 and f_2, respectively, via cubic regression spline functions. Let $\boldsymbol{c}_1 = (c_{11}, ..., c_{1K_1-1})'$ and $\boldsymbol{c}_2 = (c_{21}, ..., c_{2K_2-1})'$ are the spline coefficient vectors for f_1 and f_2 that satisfy model identifiability constraints.

The function `test.trioGxE` calculates test statistic T,

$$T = t(\hat{\boldsymbol{c}})\mathrm{V}^{-1}(\boldsymbol{c})\hat{\boldsymbol{c}},$$

where $\boldsymbol{c} = (\boldsymbol{c}_1', \boldsymbol{c}_2')'$ and V_c is a square matrix of size (k_1+k_2-2), formed by extracting the rows and columns, corresponding to the spline coefficients from the Bayesian posterior variance-covariance matrix calculated in `trioGxE`.

If the actual data were fitted under the co-dominant penetrance mode (i.e., `object$penmod="codominant"`), the test statistic T represents an *overall* test of GxE, where

$$\mathrm{H}_0 : \boldsymbol{c} = 0.$$

Depending on the context, an investigator may also want to perform individual tests: $\mathrm{H}_{01} : \boldsymbol{c}_1 = \mathbf{0}$ and $\mathrm{H}_{02} : \boldsymbol{c}_2 = \mathbf{0}$. For example, when the null hypothesis is rejected, the user may want to know which of the two interaction function is not zero (i.e., which curve is not flat). For the *individual*

tests, `test.trioGxE` calculates the permutation p-values based on the Monte-Carlo distributions of the individual test statistics T_1 and T_2, where

$$T_h = t(\hat{\boldsymbol{c}}_h)\mathrm{V}^{-1}(\boldsymbol{c}_h)\hat{\boldsymbol{c}}_h, h = 1, 2.$$

Under the dominant, log-additive (multiplicative) or recessive penetrance model, T can be viewed as an individual test since $\boldsymbol{c}_2 = \mathbf{0}$, $\boldsymbol{c}_1 = \boldsymbol{c}_2$ and $\boldsymbol{c}_1 = \mathbf{0}$, respectively, under the dominant, log-additive and recessive models. For example, under the dominant penetrance model, $T \equiv T_1$ because $\boldsymbol{c}_2 = \mathbf{0}$, and $T_2 = 0$.

As the analysis is conditional on parental genotypes, the distribution of the test statistic under H_0 is calculated by shuffling the column that holds the values of the non-genetic covariate within mating types. This can be justifiable based on the fact that under no interaction, the SNP and the non-genetic covariate are independent of each other within a random affected trio when they are independent within a trio from the general population (Umbach and Weinberg, 2000).

The distribution of the test statistics can be obtained in two ways: either under fixed smoothing parameters (`fixed.sp=TRUE`) or under varying smoothing parameters (`fixed.sp=FALSE`). Under the fixed smoothing parameters, the penalized iteratively re-weighted least squares procedure is performed for each simulated data set under the same smoothing parameter values. Under varying smoothing parameters, smoothing parameters are estimated for each simulated data set. Therefore, the test under `fixed.sp=FALSE` accounts for the extra uncertainty introduced by the smoothing parameter estimation.

To save computation time, the user can use 'early-termination' option (Besag and Clifford, 1991). Under this option, sampling is terminated when the number of the simulated data sets reaches `nreps*{level}` $<$ `nreps` when the evidence is not strong enough to reject the null hypothesis at the given significance level (`level`). For example, if the user specifies `nreps=1000` and `level=0.05`, the test terminates when the number of data sets that have test statistic values that are more extreme than or as extreme as the observed test statistic value reaches 50.

Value

`GxE.test` Either a 3- or 1-column matrix. When the actual data was fitted under a co-dominant penetrance mode (i.e., `object$penmod = "codominant"`), a 3-column matrix is returned, where the first column holds the values for T for the original and the generated data sets, and the second and third columns hold the values of T_1 and T_2, respectively for the same data sets. When the actual data was fitted under a non-co-dominant penetrance mode (e.g., dominant), `GxE.test` is retuned as a matrix with a sigle column holding T.

`p.value` If `object$penmod = "codominant"`, it is returned as a vector holding three values, where the first value indicates the overall p-value obtained from the distribution of T, and the other two values indicate the individual p-values obtained from the distributions of T_1 and T_2. Under `object$penmod` is `dominant`, `additive` or `recessive`, it is returned as a single p-value calculated based on T.

Author(s)

Ji-Hyung Shin <shin@sfu.ca>, Brad McNeney <mcneney@sfu.ca>, Jinko Graham <jgraham@sfu.ca>

References

Umbach, D. and Weinberg, C. (2000). The use of case-parent triads to study joint effects of genotype and exposure. *Am J Hum Gen*, **66**:251-61.

Besag, J. and P. Clifford (1991). Sequential Monte Carlo p-values. *Biometrika*, **78**:301-304.

See Also

trioGxE, plot.trioGxE, trioSim

Examples

```
data(hypoTrioDat)
example.fit <- trioGxE(hypoTrioDat, pgenos = c("parent1","parent2"), cgeno = "child",
                       cenv = "attr",penmod="codominant", k=c(5,5))
# A toy example with 'few' permutation replicates
example.test <- test.trioGxE(example.fit, nreps=10, early.stop = FALSE,
                             output=NULL)


## Not run:

## More proper examples of permutation tests with 5000 replicates

## Example1: does not generate an output file containing test statistic values
example.test1 <- test.trioGxE(example.fit, nreps=5000, early.stop = TRUE,
                             output=NULL)
## Example 2: generates an output file 'myoutput.out' containing test statistic values
example.test2 <- test.trioGxE(example.fit, nreps=5000, early.stop = TRUE,
                             output="myoutput.out")

## End(Not run)
```

---

trioGxE — *Generalized additive model estimation of gene-environment interaction using data from case-parent trios*

Description

trioGxE estimates statistical interaction (GxE) between a single nucleotide polymorphism (SNP) and a continuous environmental or non-genetic attributes in case-parent trio data by fitting a generalized additive model (GAM) using a penalized iteratively re-weighted least squares algorithm.

Usage

```
trioGxE(data, pgenos, cgeno, cenv,
        penmod = c("codominant","dominant","additive","recessive"),
        k = NULL, knots = NULL, sp = NULL, lsp0 = NULL, lsp.grid = NULL,
```

```
          control = list(maxit = 100, tol = 1e-07, trace = FALSE),
          testGxE = FALSE, return.data = TRUE, ...)
```

Arguments

data	a data frame with columns for parental genotypes, child genotypes and child environmental/non-genetic attribute. See 'Details' section for the required format.
pgenos	a length-2 vector of character strings specifying the names of the two columns in data that hold parental genotypes.
cgeno	a character string specifying the name of the column in data that holds the child genotypes.
cenv	a character string specifying the name of the column in data that holds the non-genetic attribute being examined for interaction with genotype.
penmod	the penetrance mode of the genetic and interaction effects: "codominant" (default), "dominant", "additive", or "recessive".
k	an optional vector or single value specifying the desired number(s) of knots to be used for the cubic spline basis construction of smooth function(s) representing GxE. When penmod="codominant", a length-2 vector with positive integers must be provided to specify the numbers of knots (or basis dimensions) for the two interaction functions. Otherwise, a single positive integer must be provided. The minimum value for each integer is 3. The default basis dimension is either k=c(5,5) or k=5. See 'Details' section for more information.
knots	knot positions for the cubic spline basis construction. When penmod="codominant", a list of two vectors must be provided. For the other penetrance modes, a single vector must be provided. When NULL (default), knots will be placed at equally-spaced quantiles of the distribution of E within trios from appropriate parental mating types. If both knots and k are provided, the argument k is ignored. See 'Details' section for more information.
sp	smoothing parameters for the interaction functions. When penmod="codominant", a vector with two non-negative numbers must be provided. Otherwise, a single non-negative number must be provided. When NULL (default), a double (under the co-dominant mode) or a single (under a non-co-dominant mode) 1-dimensional grid search finds the optimal smoothing parameter values.
lsp0	an optional length-2 numeric vector or a single numeric value used for choosing trial values of log smoothing parameters in the grid search for the optimal smoothing parameters. When NULL (default), trioGxE takes the log of smoothing parameter estimates obtained by applying a likelihood approach that makes inference of GxE conditional on the parental genotypes, non-genetic attribute and partial information on child genotypes.
lsp.grid	trial values of log smoothing parameters used in the grid search for smoothing parameters. When penmod= "codominant", a list of two vectors of lengths ≥ 2 must be provided. As the vector is longer, the grid becomes more refined. When the penetrance mode is not co-dominant, a single vector must be provided. When lsp.grid=NULL (default), the function take the vectors of length 6 obtained by using the truncated normal distributions constructed based on lsp0.

`control` a list of convergence parameters for the penalized iteratively re-weighted least squares (PIRLS) procedure:

- `maxit`: positive integer giving the maximal number of PIRLS iterations
- `tol`: positive convergence tolerance in terms the relative difference in penalized deviances (`pdev`) between iterations: $|\texttt{pdev} - \texttt{pdev}_{old}|/(|\texttt{pdev}| + 0.1) < \texttt{tol}$
- `trace`: logical indicating if output should be produced for each PIRLS iteration.

`testGxE` a logical specifying whether the fitting is for testing interaction. Default is `FALSE`. User should not modify this argument.

`return.data` a logical specifying whether the original data should be returned. If `TRUE` (default), the data is returned.

`...` sets the arguments for `control`, which includes `maxit`, `tol` or `trace`.

Details

`trioGxE` fits data from case-parent trios to a GAM with smooth functions representing gene-environment interaction ($G \times E$).

The `data` input must be a data frame containing the following 4 columns (of any order):

- [,1] number (0, 1 or 2) of copies of a putative risk allele carried by the mother
- [,2] number of copies of a putative risk allele carried by the father
- [,3] number of copies of a putative risk allele carried by the affected child (G)
- [,4] value of a continuous environmental/non-genetic attribute measured on the child (E)

The function `trioGxE` does basic error checking to ensure that only the trios that are consistent with Mendelian segregation law with complete genotype, environment and parental genotype information. The function determines which trios are from *informative* parental mating types. An informative parental pair has at least one heterozygote; such parental pair can have offspring that are genetically different. Under the assumption that the parents transmit the alleles to their child under Mendel's law, with no mutation, there are three types of informative mating types $G_p = 1, 2, 3$:

- $G_p = 1$: if one parent is heterozygous, and the other parent is homozygous for the non-risk allele
- $G_p = 2$: if one parent is heterozygous, and the other parent is homozygous for the risk allele
- $G_p = 3$: if both parents are heterozygous

Since GxE occurs when genotype relative risks vary with non-genetic attribute values $E = e$, GxE inference is based on the attribute-specific genotype relative risks, $\text{GRR}_h(e)$, expressed as

$$\text{GRR}_h(e) = \frac{P(D = 1|G = h, E = e)}{P(D = 1|G = h - 1, E = e)} = \exp\left(\gamma_h + f_h(e)\right), \tilde{}\tilde{} h = 1, 2,$$

where $D = 1$ indicates the affected status, γ_1 and γ_2 represent genetic main effect, and $f_1(e)$ and $f_2(e)$ are unknown smooth functions. The functions $f_h(e)$ represent GxE since GRRs vary with $E = e$ only when $f_1(e) \neq 0$ or $f_2(e) \neq 0$ vary with E. The expressions are followed by assuming a log-linear model of disease risk in Shin et al. (2010).

Under the co-dominant penetrance mode (i.e., `penetrance="codominant"`), $\text{GRR}_1(e)$ and $\text{GRR}_2(e)$ are estimated using the information in the trios from the informative mating types $G_p = 1$, 3 and

those from $G_p = 2$, 3, respectively. Under a non-co-dominant penetrance mode, only one GRR function, $\text{GRR}(e) = \gamma + f(e)$, is estimated from an appropriate set of informative trios. Under the dominant penetrance mode (i.e., `penetrance="dominant"`), because $\text{GRR}_2(e)$ is 1 (i.e., $\gamma_2 = 0$ and $f_2(e) = 0$), $\text{GRR}(e) \equiv \text{GRR}_1(e)$ is estimated based on the trios from $G_p = 1$ and 3. Under the recessive penetrance mode (i.e., `penetrance="recessive"`), because $\text{GRR}_1(e)$ is 1 (i.e., $\gamma_1 = 0$ and $f_1(e) = 0$), $\text{GRR}(e) \equiv \text{GRR}_2(e)$ is estimated based on the trios from $G_p = 2$ and 3. Under the multiplicative or log-additive penetrance model (`penetrance="additive"`), since $\text{GRR}_1(e) = \text{GRR}_2(e)$ (i.e., $\gamma_1 = \gamma_2$ and $f_1 = f_2$), $\text{GRR}(e) \equiv \text{GRR}_1(e) \equiv \text{GRR}_2(e)$ is estimated based on all informative trios.

The interaction functions are approximated by cubic regression spline functions defined by the knots specified through the arguments `k` and `knots`. For each interaction function, at least three knots are chosen within the range of the observed non-genetic attribute values. Under the co-dominant mode, `k[1]` and `k[2]` knots, respectively, located at `knots[[1]]` and `knots[[2]]` are used to construct the basis for $f_1(e)$ and $f_2(e)$, respectively. By default, a total of 5 knots are placed for each interaction function: three interior knots at the 25th, 50th and 75th quantiles and two boundary knots at the endpoints of the data in trios from $G_p = 1$ or 3, for $f_1(e)$, and in trios from $G_p = 2$ or 3, for $f_2(e)$. Similarly, under a non-co-dominant penetrance mode, when `knots=NULL`, `k` knots are chosen based on the data in trios from $G_p = 1$ and 3 (dominant mode); in trios from all informative mating types (log-additive mode); and in trios from $G_p = 2$ or 3 (recessive mode). A standard model identifiability constraint is imposed on each interaction function, which involves the sum of the interaction function over all observed attribute values of cases in the appropriate set of informative trios.

For smoothing parameter estimation, `trioGxE` finds the optimal values using either a double (if co-dominant) or a single 1-dimensional grid search constructed based on the arguments `lsp0` and `lsp.grid`. When `lsp0 = NULL`, `trioGxE` takes the log smoothing parameter estimates obtained from a likelihood approach that makes inference of GxE conditional on parental mating types, non-genetic attribute and partial information on child genotypes (Duke, 2007). When `lsp.grid = NULL`, `trioGxE` takes the following 6 numbers to be the trial values of the log-smoothing parameter for each interaction function: -20 and 20, `lsp0` and the quartiles of the truncated normal distributions constructed based on `lsp0`. At each trial value in `lsp.grid`, the prediction error criterion function, UBRE (un-biased risk estimator, is minimized to find the optimal smoothing parameter. For more details on how to estimate the smoothing parameters, see Appendix B.3 in Shin (2012).

For variance estimation, `trioGxE` uses a Bayesian approach (Wood, 2006); the resulting Bayesian credible intervals have been shown to have good frequentist coverage probabilities as long as the bias is a relatively small fraction of the mean squared error (Marra and Wood, 2012)

Value

`coefficients`	a vector holding the spline coefficient estimates for $\hat{f}_1$ and/or $\hat{f}_2$. The length of the vector is equal to the total number of knots used for constructing the bases of the interaction curves. For example, under the default basis dimension with co-dominant penetrance mode, the vector has size 10 (i.e., 5 for f_1 and the other 5 for f_2).
`control`	a list of convergence parameters used for the penalized iteratively re-weighted least squares (PIRLS) procedure
`data`	original data passed to `trioGxE()` as an argument: returned when `return.data=TRUE`

`edf` a vector of effective degrees of freedom for the model parameters (see page 171 in Wood, 2006 for details).

`Gp` a vector containing the values of parental mating types G_p (see 'Details' for the definition)

`lsp0` log smoothing parameter(s) used in the grid search. Not returned (i.e., `NULL`) when smoothing parameters were not estimated but provided by the user.

`lsp.grid` trial values of the log smoothing parameter(s) used in the grid search. Not returned when smoothing parameters were not estimated but provided by the user.

`penmod` the penetrance mode under which the data were fitted.

`qrc` a list containing the QR-decomposition results used for imposing the identifiability constraints. See `qr` for the list of values.

`smooth` a list with components:

`model.mat`: The design matrix of the problem. The number of rows is equal to $n_1 + n_2 + 2n_3$, where n_m is the number of mth informative mating types. The number of columns is equal to the size of `coefficient`.

`pen.mat`: penalty matrix with size equal to the size of `coefficient`.

`bs.dim`: number of knots used for basis construction.

`knots`: knot positions used for basis construction.

`sp` optimal smoothing parameter values calculated from UBRE optimization or smoothing parameter values provided by the user.

`sp.user` logical whether or not the smoothing parmeter values were provided by the user. If `FALSE`, `sp` contains the smoothing parameter values estimated by the UBRE optimization.

`terms` list of character strings of column names in `data` corresponding to the child genotypes, parental genotypes and child's non-genetic attributes.

`triodata` Formatted data passed into the internal fitting functions. To be used in `test.trioGxE` function.

`ubre` the minimum value of the un-biased risk estimator (UBRE), measure of predictability for the interaction function estimators $\hat{f}_1$ or $\hat{f}_2$. Not returned when smoothing parameters were not estimated but provided by the user.

`ubre.val` a list or a vector of ubre values corresponding to the trial values of smoothing parameter(s) in `lsp.grid`.

`Vp` Bayesian posterior variance-covariance matrix for the coefficients. The size the matrix is the same as that of `coefficient`.

Author(s)

Ji-Hyung Shin <shin@sfu.ca>, Brad McNeney <mcneney@sfu.ca>, Jinko Graham <jgraham@sfu.ca>

References

Duke, L. (2007): *A graphical tool for exploring SNP-by-environment interaction in case-parent trios*, M.Sc. Thesis, Statistics and Actuarial Science, Simon Fraser University:

URL http://www.stat.sfu.ca/content/dam/sfu/stat/alumnitheses/Duke-2007.pdf.

Marra, G., Wood, S.N. (2012). Coverage properties of confidence intervals for generalized additive model components. *Scand J Stat*, **39**: 53-74.

Shin, J.-H., McNeney, B. and Graham, J. (2010). On the use of allelic transmission rates for assessing gene-by-environment interaction in case-parent trios. *Ann Hum Gen*, **74**: 439-51.

Shin, J.-H. (2012): *Inferring gene-environment interaction from case-parent trio data: evaluation of and adjustment for spurious $G \times E$ and development of a data-smoothing method to uncover true $G \times E$*, Ph.D. Thesis, Statistics and Actuarial Science, Simon Fraser University: URL https://theses.lib.sfu.ca/sites/all/files/public_copies/etd7214-j-shin-etd7214jshin.pdf.

Wood, S. (2006): *Generalized Additive Models: An Introduction with R*, Boca Raton, FL: Chapman & Hall/CRC.

See Also

`trioSim`, `plot.trioGxE`, `test.trioGxE`

Examples

```
## fitting a co-dominant model
data(hypoTrioDat)
simfit <- trioGxE(data=hypoTrioDat,pgenos=c("parent1","parent2"),cgeno="child",cenv="attr",
                  k=c(5,5),knots=NULL,sp=NULL)

## fitting a dominant model to the hypothetical data
simfit.dom <- trioGxE(data=hypoTrioDat,pgenos=c("parent1","parent2"),cgeno="child",cenv="attr",
                      penmod="dom",k=5,knots=NULL,sp=NULL)
```

---

trioSim *Simulate informative case-parent trios*

Description

`trioSim()` simulates parental genotypes, child genotypes, environmental attribute and sub-population membership on affected trios with *informative* mating types from a stratified population. All genotypes are at a test locus that is linked to a causal locus.

Usage

```
trioSim(n, popfs, hapfs, edists, recomb = 0, riskmod, batchsize = 1000)
```

Arguments

n	A desired number of informative case-parent trios to simulate.
popfs	A vector of sub-population frequencies whose length is equal to the number of sub-populations.
hapfs	A list comprised of vectors of haplotype frequencies. One haplotype frequency vector for each sub-population. See Details for the assumed order of haplotypes.
edists	A list comprised of functions to simulate the environment attribute. One simulation function for each sub-population.
recomb	Recombination frequency between the test and causal locus. Currently not implemented. The function will stop execution if a non-zero value is specified.
riskmod	A function to evaluate the risk (probability) of disease. The function should take two arguments. The first is the child's genotype, and the second is the environmental attribute.
batchsize	Size of the batches of trios to simulate. See Details for more information.

Details

The function simulates trios from a stratified population. Population stratification is controlled by the user's choice of sub-population sizes, haplotype frequencies in each sub-population and the distribution of the environmental attribute in each sub-population. Given sub-population sizes, the degree of population stratification increases with greater differences in the distributions of the haplotype frequency and the environmental attribute among sub-populations.

The function first simulates sub-population membership for each trio using the sub-population frequencies supplied by the user in the argument popfs. Conditional on sub-population, parental haplotypes Hp are simulated assuming Hardy-Weinberg proportions using the subpopulation-specific haplotype frequencies in the argument hapfs. Haplotype frequencies should be in the order N0, N1, R0, R1, where N and R denote the non-risk and risk alleles at the causal locus, and 0 and 1 denote the non-index and index alleles at the test locus.

To save computation time, we only considered informative parental mating types by simulating one parent from the conditional distribution given that the parent is heterozygous and simulating the other parent without any restrictions. Conditional on parental haplotypes, child haplotypes are sampled according to Mendel's laws. From the sampled haplotypes of the parents and children, their genotypes for the causal and test loci are extracted.

Assuming conditional independence between the gene and the environmental attribute given sub-population, the environmental attribute for each trio is simulated conditional on sub-population using the subpopulation-specific simulation functions in the argument edists. Finally, disease status is simulated according to the risk model in the argument riskmod; only those trios with affected children are retained.

To speed up computation, the rejection sampling of trios is done in batches of size 'batchsize' until a desired number of affected trios is obtained. In simulation studies we have performed, choosing batchsize on the order of 1/3 the desired number of trios appeared to be the fastest.

Value

A data frame with columns

parent1	Test locus genotypes for one parent (heterozygous) coded as 0, 1, 2 representing the number of copies of the index allele.
parent2	Test locus genotype for the other parent.
child	Test locus genotypes for the child.
subpop	Sub-population membership for the trio. Sub-populations are numbered $0, 1, ..., k-1$, where k is the number of sub-populations.
attr	The environmental attribute.

Author(s)

Ji-Hyung Shin <shin@sfu.ca>, Brad McNeney <mcneney@sfu.ca>, Jinko Graham <jgraham@sfu.ca>

See Also

trioGxE, plot.trioGxE, test.trioGxE

Examples

```
# Generate case-parent trio from a population composed of
# two equal sized subpopulations.

# Set up list of functions to sample from each E distribution

e1<-function(n) {
return(rnorm(n,mean=(-0.8),sd=sqrt(1-.8^2)))
}
e2<-function(n) {
return(rnorm(n,mean=(0.8),sd=sqrt(1-.8^2)))
}

# Set up haplotype frequency distributions in the two subpopulations:
# The first subpopulation has the risk allele frequency of 0.1, where as
# the second subpopulation's frequency is 0.9.

# Set up risk model function.
## Simulate informative case-parent trios under additive linear GxE with a negative slope

riskmod<-function(G,E) {
n<-length(G)
# Baseline risk. Affects disease prevalence.
# The higher the prevalence, the less time wasted
# rejecting unaffected trios.
k<-(-2)

betaG<-log(3)/2

# Interaction
betaGE<-(-0.1)

# quadratic GxE
rr<-exp(k+betaG*G + betaGE*G*E)
```

```
rr[rr>1]<-1 # It is up to the user to make sure there are
# no probabilities greater than one.

D<-rbinom(n=n,size=1,prob=rr)
return(D)
}


# Simulate trio data under haplotype-environment dependence
# when marker locus is causal locus.
# allele frequency in subpop 0 is 0.1, allele frequency in subpop 1 is 0.9.
hapf1=c(0.9, 0, 0, 0.1)
hapf2=c(0.1, 0, 0, 0.9)
simdat.HEdep<-trioSim(n=3000,popfs=c(0.5,0.5),riskmod=riskmod,
                      edists=list(e1,e2),hapfs=list(hapf1,hapf2),
                      recomb=0,batchsize=1000)

# Simulate trio data under haplotype-environment independence
# when marker locus is causal locus.
# allele frequency in subpop 0 and subpop 1 is 0.1.
hapf1=hapf2=c(0.9, 0, 0, 0.1)
simdat.HEindep<-trioSim(n=3000,popfs=c(0.5,0.5),riskmod=riskmod,
                        edists=list(e1,e2),hapfs=list(hapf1,hapf2),
                        recomb=0,batchsize=1000)
```

Index