# Package 'SPAtest'

January 10, 2020

**Type** Package

**Title** Score Test and Meta-Analysis Based on Saddlepoint Approximation

**Version** 3.0.2

**Date** 2020-01-09

**Author** Rounak Dey, Seunggeun Lee

**Maintainer** Rounak Dey <deyrnk@umich.edu>

**Description**
Performs score test using saddlepoint approximation to estimate the null distribution. Also prepares summary statistics for meta-analysis and performs meta-analysis to combine multiple association results. For the latest version, please check <https://github.com/leeshawn/SPAtest>.

**Depends** R (>= 3.0.0)

**License** GPL (>= 2)

**Repository** CRAN

**NeedsCompilation** no

**Date/Publication** 2020-01-10 13:50:03 UTC

## R topics documented:

Saddle_Prob *Calculate Saddlepoint p values (for external libraries)*

## Description

Function to calculate the SPA p value from score function. Intended to be used by external libraries.

## Usage

```
Saddle_Prob(q, mu, g, Cutoff=2,alpha,output="P",nodes.fixed,nodes.init)
Saddle_Prob_fast(q, g,mu,gNA,gNB,muNA,muNB,Cutoff=2,alpha,output,nodes.fixed,nodes.init)
```

## Arguments

| | |
|---|---|
| q | Numeric scalar, score statistic. |
| mu | Numeric vector, fitted probabilities from the null model. |
| g | Numeric vector, covariate adjusted genotypes. |
| gNA | Numeric vector, covariate adjusted genotypes where the observed genotype is zero. |
| gNB | Numeric vector, covariate adjusted genotypes where the observed genotype is non-zero. |
| muNA | Numeric vector, fitted probabilities from the null model where the observed genotype is zero. |
| muNB | Numeric vector, fitted probabilities from the null model where the observed genotype is non-zero. |
| Cutoff | An integer or the string "BE" denoting the standard deviation cutoff to be used. If Cutoff = "BE", the level-specific cutoff based on Berry-Esseen theorem is calculated. If the test statistic lies within the standard deviation cutoff of the mean, p-value based on traditional score test is returned. Default value is 2. |
| alpha | Significance level for the test(s), default value is $5 \times 10^{-8}$. Used only if Cutoff = "BE". |
| output | String specifying the output required. Possible values are "P" (default), "metaZ", "metaGC", and "metaspline". |
| nodes.fixed | Vector denoting the spline nodes for the spline based summary statistics, if you do not want to provide a fixed set of nodes instead of estimating the optimal set of nodes. Only applicable when the output is "metaspline". |
| nodes.init | Vector denoting the initial values of the spline nodes when you want to estimate the optimal set of spline nodes using the coordinate descent algorithm. Only applicable when the output is "metaspline". Ignored if nodes.fixed is provided. The node at 0 will be automatically added, no need to provide that. The number of finally selected nodes will be the same as in nodes.init. |

## Value

| | |
|---|---|
| p.value | p-value based on the saddlepoint approximation. If output = "P", it is unsigned. For all other choice of output, it is signed. |
| p.value.NA | p-value based on the normal approximation (traditional score test). If output = "P", it is unsigned. For all other choice of output, it is signed. |
| Is.converge | "TRUE" or "FALSE" denoting whether the root-finding algorithm for the saddlepoint equation has converged. |
| Score | Centered score statistic. |
| splfun | Outputs from fitting the CGF-Spline method, if output = "metaspline". |
| var | Variance of the score test statistic, if output = "metaspline". |

## See Also

ScoreTest_SPA

ScoreTest_SPA_wMeta

## Examples

```
## Not run:
Saddle_Prob(q, mu, g, Cutoff=2,alpha,output="P",nodes.fixed,nodes.init)
Saddle_Prob_fast(q, g,mu,gNA,gNB,muNA,muNB,Cutoff=2,alpha,output,nodes.fixed,nodes.init)

## End(Not run)
```

---

# ScoreTest_SPA *Score test based on saddlepoint approximation*

---

## Description

Performs score test using saddlepoint approximation to estimate the null distribution.

## Usage

```
ScoreTest_SPA(genos,pheno,cov,obj.null,method=c("fastSPA","SPA"),minmac=5,
Cutoff=2,alpha=5*10^-8,missing.id=NA,beta.out=FALSE,beta.Cutoff=5*10^-7)
```

## Arguments

| | |
|---|---|
| genos | A vector or matrix containing the genotypes or dosages. If matrix is provided then rows should correspond to SNPs and columns should correspond to subjects. Optional, but needed if obj.null is missing. |
| pheno | A vector containing the outcomes (phenotypes). Optional, but needed if obj.null is missing. |
| cov | A matrix or data frame containing the covariates. Optional, but needed if obj.null is missing. |

| | |
|---|---|
| obj.null | An object of class "SA_NULL". (Optional) |
| method | String specifying the p-value calculation method. Possible values are "fastSPA" (default) and "SPA". |
| minmac | Minimum minor allele count threshold to run SPA test, default value is $5$. |
| Cutoff | An integer or the string "BE" denoting the standard deviation cutoff to be used. If Cutoff = "BE", the level-specific cutoff based on Berry-Esseen theorem is calculated. If the test statistic lies within the standard deviation cutoff of the mean, p-value based on traditional score test is returned. Default value is 2. |
| alpha | Significance level for the test(s), default value is $5 \times 10^{-8}$. Used only if Cutoff = "BE". |
| missing.id | Missing value indicator. Numeric or NA, default value is NA. |
| beta.out | Logical indicating whether log odds ratios (beta parameters) are to be estimated, default value is FALSE. |
| beta.Cutoff | Maximum p-value threshold for beta parameters to be estimated, default value is $5 \times 10^{-7}$. |

## Details

genos can have discrete 0,1,2 values or continuous values between [0,2]. The genotype or dosage values can represent any of the major allele, minor allele, reference allele or alternate allele counts (or dosages), as long as it is consistent throughout the subjects.

genos can have missing values denoted by the missing.id argument. Such missing values will be imputed using mean imputation. pheno or cov cannot have missing values.

pheno and cov are ignored if obj.null is provided. If both obj.null and cov is missing, or obj.null is missing and cov=NULL, then the vector rep(1,n) is assigned to cov, where n is the number of subjects.

method = "SPA" is the basic saddlepoint approximation based test without the partially normal approximation improvement. method = "fastSPA" utilizes the partially normal approximation approach for improved efficiency, especially for rare variants.

Beta parameters are estimated using Firth's method, and thus computationally expensive. Therefore, it is recommended that beta parameters are only to be estimated when the p-value is very small (denoted by beta.Cutoff). The code for beta estimation is as implemented by Clement Ma in the EPACTS software.

## Value

| | |
|---|---|
| p.value | p-value based on the saddlepoint approximation. |
| p.value.NA | p-value based on the normal approximation (traditional score test). |
| Is.converge | "TRUE" or "FALSE" denoting whether the root-finding algorithm for the saddlepoint equation has converged. |
| beta | Genotype log-odds ratio estimate. |
| SEbeta | Standard error for the genotype log-odds ratio. |

## Author(s)

Rounak Dey, <deyrnk@umich.edu>

## References

Dey, R. et al., 2017. A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS. The American Journal of Human Genetics, Vol 101 (1), 37-49.

Ma, C. et al., 2013. Recommended Joint and Meta-Analysis Strategies for Case-Control Association Testing of Single Low-Count Variants. Genetic Epidemiology, Vol 37 (6), 539-550.

## See Also

ScoreTest_wSaddleApprox_NULL_Model

ScoreTest_SPA_wMeta

## Examples

```
## Not run:
ScoreTest_SPA(genos,pheno,cov,obj.null,method=c("fastSPA","SPA"),
minmac=5,Cutoff=2,alpha=5*10^-8,missing.id=NA,beta.out=FALSE,beta.Cutoff=5*10^-7)

## End(Not run)
```

---

# ScoreTest_SPA_wMeta *Prepare summary statistics for meta-analysis and perform SPA test.*

---

## Description

Performs score test using saddlepoint approximation and prepares summary statistics for meta-analysis.

## Usage

```
ScoreTest_SPA_wMeta(genos,pheno,cov,obj.null,method=c("fastSPA","SPA"),minmac=5,
Cutoff=2,alpha=5*10^-8,missing.id=NA,beta.out=FALSE,beta.Cutoff=5*10^-7,
output=c("P","metaZ","metaGC","metaspline"),nodes.fixed=NULL,
nodes.init=c(-100,-10,-1,1,10,100))
```

## Arguments

| | |
|---|---|
| genos | A vector or matrix containing the genotypes or dosages. If matrix is provided then rows should correspond to SNPs and columns should correspond to subjects. Optional, but needed if obj.null is missing. |
| pheno | A vector containing the outcomes (phenotypes). Optional, but needed if obj.null is missing. |
| cov | A matrix or data frame containing the covariates. Optional, but needed if obj.null is missing. |

| | |
|---|---|
| obj.null | An object of class "SA_NULL". (Optional) |
| method | String specifying the p-value calculation method. Possible values are "fastSPA" (default) and "SPA". |
| minmac | Minimum minor allele count threshold to run SPA test, default value is 5. |
| Cutoff | An integer or the string "BE" denoting the standard deviation cutoff to be used. If Cutoff = "BE", the level-specific cutoff based on Berry-Esseen theorem is calculated. If the test statistic lies within the standard deviation cutoff of the mean, p-value based on traditional score test is returned. Default value is 2. |
| alpha | Significance level for the test(s), default value is $5 \times 10^{-8}$. Used only if Cutoff = "BE". |
| missing.id | Missing value indicator. Numeric or NA, default value is NA. |
| beta.out | Logical indicating whether log odds ratios (beta parameters) are to be estimated, default value is FALSE. |
| beta.Cutoff | Maximum p-value threshold for beta parameters to be estimated, default value is $5 \times 10^{-7}$. |
| output | String specifying the output required. Possible values are "P" (default), "metaZ", "metaGC", and "metaspline". |
| nodes.fixed | Vector denoting the spline nodes for the spline based summary statistics, if you do not want to provide a fixed set of nodes instead of estimating the optimal set of nodes. Only applicable when the output is "metaspline". |
| nodes.init | Vector denoting the initial values of the spline nodes when you want to estimate the optimal set of spline nodes using the coordinate descent algorithm. Only applicable when the output is "metaspline". Ignored if nodes.fixed is provided. The node at 0 will be automatically added, no need to provide that. The number of finally selected nodes will be the same as in nodes.init. |

## Details

genos can have discrete 0,1,2 values or continuous values between [0,2]. The genotype or dosage values can represent any of the major allele, minor allele, reference allele or alternate allele counts (or dosages), as long as it is consistent throughout the subjects.

genos can have missing values denoted by the missing.id argument. Such missing values will be imputed using mean imputation. pheno or cov cannot have missing values.

pheno and cov are ignored if obj.null is provided. If both obj.null and cov is missing, or obj.null is missing and cov=NULL, then the vector rep(1,n) is assigned to cov, where n is the number of subjects.

method = "SPA" is the basic saddlepoint approximation based test without the partially normal approximation improvement. method = "fastSPA" utilizes the partially normal approximation approach for improved efficiency, especially for rare variants.

Beta parameters are estimated using Firth's method, and thus computationally expensive. Therefore, it is recommended that beta parameters are only to be estimated when the p-value is very small (denoted by beta.Cutoff). The code for beta estimation is as implemented by Clement Ma in the EPACTS software.

If output = "P", the output will be the same as running ScoreTest_SPA. No summary statistics for meta-analysis is provided.

If output = "metaZ", the output will be the same as with output = "P", except the p values in the output will be signed, and minor allele frequencies are also provided. This choice of output provide required summary statistics for Z score-based meta-analysis.

If output = "metaGC", the output will be the same as with output = "P", except the p values in the output will be signed, and genotype counts of homozygous minor and heterozygous genotypes are also provided. This choice of output provide required summary statistics for genotype count-based meta-analysis.

If output = "metaspline", the output will be the same as with output = "metaGC", additionally spline-based summary statistics are also provided. This choice of output provide required summary statistics for spline-based as well as genotype count-based meta-analysis.

## Value

| | |
|---|---|
| p.value | p-value based on the saddlepoint approximation. If output = "P", it is unsigned. For all other choice of output, it is signed. |
| p.value.NA | p-value based on the normal approximation (traditional score test). If output = "P", it is unsigned. For all other choice of output, it is signed. |
| Is.converge | "TRUE" or "FALSE" denoting whether the root-finding algorithm for the saddlepoint equation has converged. |
| beta | Genotype log-odds ratio estimate. |
| SEbeta | Standard error for the genotype log-odds ratio. |
| MAF | Minor allele frequencies. Only provided when output = "metaZ". |
| GCmat | Genotype counts for homozygous minor (column 1) and heterozygous genotypes (column 2). Only provided when output = "metaGC" or "metaspline". |
| spldata | Spline-based summary statistics on the CGF. Column 1 represents the raw score values, column 2 the variance of those scores, and the next columns represent nodes, first and second derivatives of the CGF in equal sizes. Only provided when output = "metaspline". |

## Author(s)

Rounak Dey, <deyrnk@umich.edu>

## References

Dey, R. et al., 2017. A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS. The American Journal of Human Genetics, Vol 101 (1), 37-49.

Ma, C. et al., 2013. Recommended Joint and Meta-Analysis Strategies for Case-Control Association Testing of Single Low-Count Variants. Genetic Epidemiology, Vol 37 (6), 539-550.

## See Also

ScoreTest_wSaddleApprox_NULL_Model

ScoreTest_SPA

SPAmeta

## Examples

```
## Not run:
ScoreTest_SPA_wMeta(genos,pheno,cov,obj.null,method=c("fastSPA","SPA"),
minmac=5,Cutoff=2,alpha=5*10^-8,missing.id=NA,beta.out=FALSE,beta.Cutoff=5*10^-7,
output=c("P","metaZ","metaGC","metaspline"),nodes.fixed=NULL,
nodes.init=c(-100,-10,-1,1,10,100))

## End(Not run)
```

---

`ScoreTest_wSaddleApprox_NULL_Model`

*Preparing the null model*

---

## Description

Prepares the null model object `SA_NULL` to be used in `ScoreTest_SPA` based on the outcome and the covariates.

## Usage

```
ScoreTest_wSaddleApprox_NULL_Model(formula, data=NULL)
```

## Arguments

| | |
|---|---|
| `formula` | An object of class "formula" (as used in the function "glm"). |
| `data` | An optional data frame, list or environment containing the variables in the model. |

## Value

`ScoreTest_wSaddleApprox_NULL_Model` returns an object of class `SA_NULL`.

## Author(s)

Rounak Dey, <deyrnk@umich.edu>

## See Also

`ScoreTest_SPA`

`ScoreTest_SPA_wMeta`

## Examples

```
## Not run:
ScoreTest_wSaddleApprox_NULL_Model(formula, data=NULL)

## End(Not run)
```

SPAmeta *Meta-analysis using SPA test summary statistics*

## Description

Performs meta-analysis using summary statistics obtained from SPA test within individual studies. It can be used to hybridize different meta-analysis techniques based on the available summary statistics.

## Usage

```
SPAmeta(pvalue.Z=NULL,MAF.Z=NULL,CCsize.Z=NULL,
pvalue.GC=NULL,GCmat=NULL,CCsize.GC=NULL,Cutoff.GC=2,
spldata=NULL,
Cutoff.meta=2)
```

## Arguments

| | |
|---|---|
| pvalue.Z | Vector denoting signed p values for which no other summary statistic is available and Z score conversion is suited. |
| MAF.Z | Vector denoting the minor allele frequencies in different studies corresponding to pvalue.Z. If missing, the different studies will be assumed to have no heterogeneity in the variants. Optional, if only Z score-based meta-analysis is being performed. Required, if being hybridized with other meta-analysis methods. |
| CCsize.Z | Matrix denoting the case-control sample sizes of the studies corresponding to pvalue.Z. Column 1 represents cases, column 2 controls. |
| pvalue.GC | Vector denoting signed p values for which genotype counts of the homozygous minor and heterozygous genotypes are available and the p values are calculated using the SPA test. These p values will be used in the meta-analysis using the genotype count-based meta-analysis strategy. |
| GCmat | Matrix denoting the genotype counts, column 1 represents the homozygous minor and column 2 the heterozygous genotypes. |
| CCsize.GC | Matrix denoting the case-control sample sizes of the studies corresponding to pvalue.GC. Column 1 represents cases, column 2 controls. |
| Cutoff.GC | An integer or vector specifying the Cutoffs used to calculate the SPA test p values within individual studies. |
| spldata | Vector, matrix or a list denoting the spline-based summary statistics as obtained from running ScoreTest_SPA_wMeta with output = "metaspline". Provide only for the studies with this information available. |
| Cutoff.meta | An integer denoting the fixed standard deviation cutoff to be used for the combined meta-analysis test. Default value is 2. |

## Details

If all studies with genotype count information used the same cutoff, then providing only a single number as `Cutoff.GC` is sufficient. Otherwise, a vector denoting the cutoff within each study is required. Currently, "BE" as a within-study cutoff is not accepted.

If provided as a matrix, the rows of `spldata` should correspond to the individual studies for which spline-based meta-analysis is to be applied. Column 1 should denote the raw within-study scores, column 2 the corresponding variances. Out of the next 3k columns, first k columns should denote the nodes, next k columns the functional values of the first derivative of the CGF, and the last k columns the functional values of the second derivative of the CGF. None of the elements of the matrix can be missing or NA. If different number of nodes were used in different studies, please provide them in list format. If spline-based summary statistic is available for only one study, then it can also be provided as a vector with 3k+2 elements where the elements denote similar statistics as denoted above in the case of a matrix. If provided as a list, the elements of the list correspond to the different studies. Each element is a vector which denote the score, variance, nodes, and first and second derivative values in this order for that particular study. The list type input is more suited when the number of nodes in different studies are different.

If only Z score-based meta-analysis is needed, provide only `pvalue.Z`, `MAF.Z`, and `CCsize.Z`. `MAF.Z` is used to properly weight the Z scores when there is between study heterogeneity present. It is optional to provide the `MAF.Z` information if only Z score-based meta-analysis is being performed. Then the studies will be assumed to be homogeneous if `MAF.Z` is missing. However, `MAF.Z` is required if the Z score-based meta-analysis is being hybridized with other meta-analysis methods. If only genotype count-based meta-analysis is needed, provide only `pvalue.GC`,`GCmat`,`CCsize.GC`, and `Cutoff.GC`. If only spline-based meta-analysis is needed, provide only `spldata`. If different studies provide different types of summary information, hybridize the methods accordingly.

For now, this package does not support having different minor alleles in different studies. If the minor alleles are different in different studies, some manual pre-processing is required, such as changing the signs of the p values, adjusting the MAF and genotype counts accordingly, and changing the signs of the scores, nodes and first derivatives in the spline-based summary statistics. In later implementations, we will try to provide a function for doing that pre-processing.

## Value

The signed meta-analysis p value is returned.

## Author(s)

Rounak Dey, <deyrnk@umich.edu>

## See Also

`ScoreTest_wSaddleApprox_NULL_Model`

`ScoreTest_SPA_wMeta`

## Examples

```
## Not run:
SPAmeta(pvalue.Z=NULL,MAF.Z=NULL,CCsize.Z=NULL,
pvalue.GC=NULL,GCmat=NULL,CCsize.GC=NULL,Cutoff.GC=2,
```

```
    spldata=NULL,
    Cutoff.meta=2)

    ## End(Not run)
```

# Index