Package 'dotgen'

October 27, 2024

Title Gene-Set Analysis via Decorrelation by Orthogonal Transformation

Version 0.1.1

Date 2024-10-27

Description

Decorrelates a set of summary statistics (i.e., Z-scores or P-values per SNP) via Decorrelation by Orthogonal Transformation (DOT) approach and performs gene-set analyses by combining transformed statistic values; operations are performed with algorithms that rely only on the association summary results and the linkage disequilibrium (LD). For more details on DOT and its power, see Olga (2020) <doi:10.1371/journal.pcbi.1007819>.

Depends R (>= 3.5.0), mvtnorm

License GPL (>= 2)

URL https://github.com/xiaoran831213/dotgen

Encoding UTF-8

Author Olga Vsevolozhskaya [aut] (<https://orcid.org/0000-0001-9376-3645>),
Dmitri Zaykin [aut] (<https://orcid.org/0000-0002-9444-2859>),
Xiaoran Tong [aut, cre] (<https://orcid.org/0000-0002-4648-3330>)

Maintainer Xiaoran Tong <xt@uky.edu>

RoxygenNote 7.2.3

NeedsCompilation no

Repository CRAN

Date/Publication 2024-10-27 12:10:02 UTC

Contents

cst *Correlation among association test statistics*

Description

Calculates the correlation among genetic association test statistics.

Usage

```
cst(g, x = NULL)
```

Arguments

g	matrix of genotype, one row per sample, one column per variant, missing values allowed.
x	matrix of covariates, one row per sample, no missing values allowed.

Details

When no covariates are present in per-variant association analyses, that is, `x==NULL`, correlation among test statistics is the same as the correlation among variants, `cor(g)`.

With covariates, correlation among test statistics is not the same as `cor(g)`. In this case, `cst()` takes the generalized inverse of the entire correlation matrix, `corr(cbind(g, x))`, and then inverts back only the submtarix containing genotype variables, `g`.

If Z-scores were calculated based on genotypes with some missing values, the correlation among test statistics will be reduced by the amount that can be theoretically derived. It can be shown that this reduced correlation can be calculated by imputing the missing values with the averages of non-missing values. Therefore, by default, `cst()` fills missing values in each variant with the average of non-missing values in that same variant (i.e., imputation by average, `imp_avg()`). Other imputation methods are also available (see topic imp for other techniques that may improve power), but note that techniques other than the imputation by average requires one to re-run the association analyses with imputed variants to ensure the correlation among new statistics (i.e., Z-scores) and the correlation among imputed variants are identical. Otherwise, Type I error may be inflated for decorrelation-based methods.

Value

Correlation matrix among association test statistics.

See Also

imp, `imp_avg()`

Examples

```
## get genotype and covariate matrices
gno <- readRDS(system.file("extdata", 'rs208294_gno.rds', package="dotgen"))
cvr <- readRDS(system.file("extdata", 'rs208294_cvr.rds', package="dotgen"))

## correlation among association statistics, covariates involved
res <- cst(gno, cvr)
print(res[1:4, 1:4])

## genotype matrix with 2% randomly missing data
g02 <- readRDS(system.file("extdata", 'rs208294_g02.rds', package="dotgen"))
cvr <- readRDS(system.file("extdata", 'rs208294_cvr.rds', package="dotgen"))
res <- cst(g02, cvr)
print(res[1:4, 1:4])
```

---

dot *Decorrelation by Orthogonal Transformation (DOT)*

Description

`dot()` decorrelates genetic association test statistics by special symmetric orthogonal transformation.

Usage

```
dot(Z, C, tol.cor = NULL, tol.egv = NULL, ...)
```

Arguments

`Z`	vector of association test statistics (i.e., Z-scores).
`C`	correlation matrix among the association test statistics, as obtained by `cst()`.
`tol.cor`	tolerance threshold for the largest correlation absolute value.
`tol.egv`	tolerance threshold for the smallest eigenvalue.
`...`	additional parameters.

Details

Genetic association studies typically provide per-variant test statistics that can be converted to asymptotically normal, signed Z-scores. Once those Z-scores are transformed to independent random variables, various methods can be applied to combine them and obtain SNP-set overall association.

`dot()` uses per-variant genetic association test statistics and the correlation among them to decorrelate Z-scores.

To estimate the correlation among genetic association test statistics, use `cst()`. If P-values and estimated effects (i.e, beta coefficients) are given instead of test statistics, `zsc()` can be used to recover the test statistics (i.e., Z-scores).

`tol.cor`: variants with correlation too close to 1 in absolute value are considered to be collinear and only one of them will be retained to ensure that the LD matrix is full-rank. The maximum value for tolerable correlation is 1 - `tol.cor`. The default value for `tol.cor` is `sqrt(.Machine$double.eps)`.

`tol.egv`: negative and close to zero eigenvalues are truncated from matrix `D` in `H = EDE'`. The corresponding columns of `E` are also deleted. Note the the dimention of the square matrix `H` does not change after this truncation. See DOT publication in the reference below for more details on definitions of `E` and `D` matrices. The default eigenvalue tolerance value is `sqrt(.Machine$double.eps)`.

A number of methods are available for combining de-correlated P-values, see dot_sst for details.

Value

a list with return values.

`X`: association test statistics, de-correlated.

`H`: orthogonal transformation, such that `X = H \%*\% Z`.

`M`: effective number of variants after de-correlation.

`L`: effective number of eigenvalues after truncation.

References

Vsevolozhskaya, O. A., Shi, M., Hu, F., & Zaykin, D. V. (2020). DOT: Gene-set analysis by combining decorrelated association statistics. PLOS Computational Biology, 16(4), e1007819.

See Also

`cst()`, `zsc()`, dot_sst

Examples

```
## get genotype and covariate matrices
gno <- readRDS(system.file("extdata", 'rs208294_gno.rds', package="dotgen"))
cvr <- readRDS(system.file("extdata", 'rs208294_cvr.rds', package="dotgen"))

## estimate the correlation among association test statistics
sgm <- cst(gno, cvr)

## get the result of genetic association analysis (P-values and effects)
res <- readRDS(system.file("extdata", 'rs208294_res.rds', package="dotgen"))

## recover Z-score statistics
stt <- with(res, zsc(P, BETA))

## decorrelate Z-scores by DOT
result <- dot(stt, sgm)
print(result$X)          # decorrelated statistics
print(result$H)          # orthogonal transformation

## sum of squares of decorrelated statistics is a chi-square
ssq <- sum(result$X^2)
pvl <- 1 - pchisq(ssq, df=result$L)
```

```
print(ssq)            # sum of squares = 35.76306
print(pvl)            # chisq P-value =  0.001132132
```

---

dot_sst *Methods for combining decorrelated summary statistics*

Description

Decorrelates and combines per-variant genetic association test statistics, Z, given the correlation matrix among them, C.

Usage

```
dot_chisq(Z, C, ...)

dot_fisher(Z, C, ...)

dot_art(Z, C, k = NULL, ...)

dot_arta(Z, C, k = NULL, w = NULL, ...)

dot_rtp(Z, C, k = NULL, ...)

dot_tpm(Z, C, tau = 0.05, ...)
```

Arguments

Z	vector of association test statistics (i.e., Z-scores).
C	matrix of correlation among the test statistics, as obtained by cst().
...	additional parameters
k	combine k smallest (decorrelated) P-values.
w	weight assigned to partial sums in ARTA implementation; default is 1.
tau	combine (decorrelated) P-values no large than tau; default is 0.05.

Details

These functions first call dot() to decorrelate the genetic association test statistics and then provide various options to combine independent statistics or corresponding P-values into the overall statistic and P-value.

The two rank truncated tests (i.e., dot_art(), dot_rtp()) require an additional parameter k that specifes the number of smallest (decorrelated) P-values to combine. By default, k equals half of the number of variants. The adaptive rank truncation method, dot_arta(), determines the optimal truncation value between 1 and k.

The truncated product method, dot_tpm(), combines P-values at least as small as tau (0.05 by default). If tau is equal to 1, then dot_tpm() provides the same result as dot_fisher() (i.e.,

Fisher's method for combining P-values). Similarly, if k is equal to the total number of tests, the results of `dot_art()` and `dot_rtp()` will be the same as that of `dot_fisher()`.

Reference (**a**) below details how to combine decorrelated test statistics or P-values via `dot_art()`, `dot_rtp()` and `dot_arta()`; reference (**b**) details `dot_tpm()` method.

Value

a list of

`X`: decorrelated association statistics.

`H`: orthogonal transformation, such that `X = H \%*\% Z`.

`Y`: the overall combined statistic.

`P`: the P-value corresponding to `Y`.

for Augmented Rank Truncated Adaptive (ARTA) test,

k: the number of decorrelated P-values that were adaptively picked.

for Truncated Product Method (TPM),

k: the number of decorrelated P-values $\leq$ `tau`.

Functions

- `dot_chisq()`: decorrelation followed by a Chi-square test.
- `dot_fisher()`: decorrelated Fisher's combined P-value test.
- `dot_art()`: decorrelated Augmented Rank Truncated (ART) test.
- `dot_arta()`: decorrelated Augmented Rank Truncated Adaptive (ARTA) test.
- `dot_rtp()`: decorrelated Rank Truncated Product (RTP) test.
- `dot_tpm()`: decorrelated Truncated Product Method (TPM) test.

References

(a) Vsevolozhskaya, O. A., Hu, F., & Zaykin, D. V. (2019). *Detecting weak signals by combining small P-values in genetic association studies.* Frontiers in genetics, 10, 1051.

(b) Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H., & Weir, B. S. (2002). *Truncated product method for combining P-values.* Genetic Epidemiology, 22(2), 170-185.

See Also

`dot()`

Examples

```
## get the test statistics and pre-calculated LD matrix
stt <- readRDS(system.file("extdata", 'art_zsc.rds', package="dotgen"))
sgm <- readRDS(system.file("extdata", 'art_ldm.rds', package="dotgen"))


## decorrelated chi-square test
result <- dot_chisq(stt, sgm)
print(result$Y)  # 37.2854
print(result$P)  # 0.0003736988

## decorrelated Fisher's combined P-value chi-square test
result <- dot_fisher(stt, sgm)
print(result$Y)  # 58.44147
print(result$P)  # 0.0002706851

## decorrelated augmented rank truncated (ART) test.
result <- dot_art(stt, sgm, k=6)
print(result$Y)  # 22.50976
print(result$P)  # 0.0006704994

## decorrelated Augmented Rank Truncated Adaptive (ARTA) test
result <- dot_arta(stt, sgm, k=6)
print(result$Y)  # -1.738662
print(result$k)  #  5 smallest P-values are retained
print(result$P)  #  0.003165 (varies)

## decorrelated Rank Truncated Product (RTP)
result <- dot_rtp(stt, sgm, k=6)
print(result$Y)  # 22.6757
print(result$P)  # 0.0007275518

## decorrelated Truncated Product Method (TPM)
result <- dot_tpm(stt, sgm, tau=0.05)
print(result$Y)  #  1.510581e-08
print(result$k)  #  6 P-values <= tau
print(result$P)  #  0.0007954961
```

`imp` *Impute missing genotypes*

Description

Impute missing genotype calls with values inferred from non-missing ones.

Usage

```
imp_avg(g, ...)

imp_cnd(g, ...)
```

Arguments

g	genotype matrix, one row per sample, and one column per variant.
...	additional parameters.

Details

A seemingly naive way to impute a missing value is to use the average of all non-missing values per variant, `imp_avg()`. Besides simplicity, this imputation by average has the advantage of approximating the correlation among test statistics (i.e., Z-scores) when the original association analyses were performed with missing values unfilled, which is a common practice. This naive approach is the defualt for the correlation calculator `cst()`.

An advanced imputation approach is based on the conditional expectation method, `imp_cnd()`, that explores the relationship between variants and borrows information from variants other than the target one when making guesses. The sample correlation among variants imputed this way is closer to the true LD, and may improve power. However, after this imupation one must re-run the association analyses with imputed variants to avoid inflation in Type I error rates.

Value

imputed genotype matrix without any missing values.

Functions

- `imp_avg()`: imputation by average.
- `imp_cnd()`: imputation by conditional expectation

---

zsc *Calculate Z-scores from P-values and estimated effects*

Description

`zsc()` recovers Z-scores from P-values and corresponding effect directions (or beta coefficients) reported by a genetic association analysis.

Usage

```
zsc(P, BETA)
```

Arguments

P	vector of P-values.
BETA	vector of effect directions or beta coefficients.

Details

For any genetic variant, its two-sided P-value (p) and the sign of estimated effect (β) is used to recover the Z-score (z), that is, $z = sign(\beta)\Phi^{-1}(p/2)$.

Value

A vector of Z-scores.

See Also

dot()

Examples

```
## result of per-variant analysis (P-values and estimated effects)
res <- readRDS(system.file("extdata", 'rs208294_res.rds', package="dotgen"))

## recover Z-score statistics
stt <- with(res, zsc(P, BETA))

## checking
stopifnot(all.equal(pnorm(abs(stt), lower.tail = FALSE) * 2, res$P))
```

Index