Package 'RRmix'

June 3, 2016

Type Package

Title RRmix: Model-Based Classification with Simultaneous Adjustment for Unwanted Variation

Version 1.0

Date 2015-6-2

Author Muting Wan, Martin Wells, James Booth, Stephen Salerno

Maintainer Stephen Salerno <salerno1212@gmail.com>

Description NEED

License GPL (>=2)

LazyData TRUE

Depends R (>= 3.0.2)

Imports Rcpp (>= 0.12.4),
RcppArmadillo

LinkingTo Rcpp (>= 0.12.4),
RcppArmadillo

RoxygenNote 5.0.1

R topics documented:

`nfactors` *Number of factors to be estimated in the RRmix model*

Description

The number of factors to be used in the RRmix model such that `var.exp` of the variance in the data is explained.

Usage

```
nfactors(expr, var.exp = 0.8, plot = TRUE)
```

Arguments

`expr`	An `n` by `G` matrix of `G` gene expression variables, or compound abundances (standardized), on `n` samples.
`var.exp`	The minimum cumulative proportion of the variance required to be explained by the determined number of factors (`q.in`). Float on (0,1) (default = 0.8)
`plot`	boolean. Produce a scree plot of the variance explained by each factor (default = `TRUE`).

Details

This function performs Principal Component Analysis (PCA) on the covariance matrix for the raw data (`expr`). `var.exp` sets the criteria for the minimum variance explained by the determined number of factors (`q.in`) for the RRmix model. `plot = TRUE` produces a scree plot of the variance explained by each factor.

Value

The number of factors that cumulatively explain >= `var.exp`

References

Gao, C., Tignor, N. L., Salit, J., Strulovici-Barel, Y., Hackett, N. R., Crystal, R. G., and Mezey, J. G. (2014). HEFT: eQTL analysis of many thousands of expressed genes while simultaneously controlling for hidden factors. Bioinformatics, 30(3):369-376.

Examples

```
expr <- log(operators[, 1:12])  # Log-Transformed, First Two Operators
expr <- t(expr)                 # Transpose Matrix for n x G
q    <- nfactors(expr); q       # Number of Factors for 80% Variance
```

`operators` *Operator Batch Effect Data Frame*

Description

This dataset contains the relative abundances of 265 metabolites across a total of 24 samples. The samples were collected by four different operators, namely A, X, D, and Z. Each operator performed a metabolomics experiment on 6 samples: three untreated (control) replicates, and three treated replicates. Both operators A and X were given samples treated with the drug 2-deoxy-glucose (2DG). Metabolites are in rows (265 metabolites), and samples are in columns (24 samples).

Usage

```
operators
```

Format

A data frame with abundances of 265 filtered metabolites on 24 samples

Source

The Locasale Research Group - Department of Pharmacology & Cancer Biology, Duke University - Durham, North Carolina

References

NEED

Examples

```
dim(operators)      # [1] 265  24
summary(operators)  # Quantiles for each metabolite
```

RRmix *RRmix: Model-Based Classification with Simultaneous Adjustment for Unwanted Variation*

Description

NEED

Details

NEED

The number of hidden factors `q` is determined using the function `nfactors`, which is based on the method described in Gao et al. (2014). Principal Component Analysis (PCA) is performed on the correlation matrix of the raw data, and the number of factors is chosen based on how many principal components explain `var.exp` of the total variation, or through visual examination of the scree plot (`plot = TRUE`). Within a reasonable range of values, `RRmix` is robust to the value of `q` (Wan 2015).

Author(s)

Muting Wan, Martin Wells, James Booth, Stephen Salerno

References

Gao, C., Tignor, N. L., Salit, J., Strulovici-Barel, Y., Hackett, N. R., Crystal, R. G., and Mezey, J. G. (2014). HEFT: eQTL analysis of many thousands of expressed genes while simultaneously controlling for hidden factors. Bioinformatics, 30(3):369-376.

M. Wan, "Model-based classification with applications to high-dimensional data in bioinformatics," Ph.D. dissertation, Cornell University, 2015.

`runRRmix` *Emperical Bayes ECM Implementation of RRmix Algorithm*

Description

This function provides a variant on the Expectation-Maximization (EM) Algorithm for the estimation of the `RRmix` hierarchical mixture model parameters. Empirical Bayes inference is implemented via an EM algorithm with classification based on the posterior expectation of the latent indicator variables. This model-based classification method with simultaneous adjustment for unwanted variation is adapted to solve association detection problems for high-dimensional data in the presence of unwanted variation.

Usage

```
runRRmix(Y.in, SNP.in, Xc.in = matrix(nrow = 0, ncol = 0),
  betac.0 = matrix(nrow = 0, ncol = 0), sig20.0 = 1, sig21.0 = 0.1,
  p.0 = 0.05, er_tol.in = 10^(-3), q.in = 1)
```

Arguments

`Y.in` An `n` by `G` matrix of `G` gene expression variables, or compound abundances (standardized), on `n` samples.

`SNP.in` A `G` by `1` indicator vector for treatment condition or minor allele presence in a single SNP (standardized).

`Xc.in` `n` by `r` matrix of `r` known covariates (standardized, default = `matrix(nrow=0,ncol=0)`).

`betac.0` `n` by `r` matrix of initial values for EM estimated covariate term coefficients (default = `matrix(nrow=0,ncol=0)`).

`sig20.0` Initial value for EM estimated first variance component for $\beta_g|b_g$ (default = 1.0).

`sig21.0` Initial value for EM estimated second variance component for $\beta_g|b_g$ (default = 0.1).

`p.0` Initial value for EM estimated proportion of differentially abundant compounds (default = 0.05).

`er_tol.in` Convergence criterion for EM algorithm (default = 0.001).

`q.in` The number of factors to be estimated (default = 1).

Value

A list object containing the following named attributes:

`er.all` Vector of error/tolerances at each iteration of the EM algorithm until convergence.

`lc` Vector of likelihood values at each iteration of the EM algorithm until convergence.

`mu` Vector of EM estimated sample-level means over genes

`sig20` EM estimated first variance component for $\beta_g|b_g$.

`sig21` EM estimated second variance component for $\beta_g|b_g$.

`Lam` Estimated $n \times q$ loading matrix (Λ) for the factor analysis component of the model.

`sig2_g` Estimated compound-specific error variance.

`b_g` Vector of the posterior probabilities of differential abundance for each of the `G` compounds.

`p` EM estimated proportion of differentially abundant compounds.

`psi` EM estimated main effect of differential abundance.

`beta_g.1` Estimated $2 \times G$ coefficient matrix for the primary effect of treatment condition.

`beta_g.0` Initial $2 \times G$ coefficient matrix for the primary effect of treatment condition.

`F_g.1` Estimated $q \times G$ factor matrix for the factor analysis component of the model.

`F_g.0` Initial $q \times G$ factor matrix for the factor analysis component of the model.

`assoc.coefs` Vector of `G` associated coefficients for each of the `G` compounds.

`sig20.0` Initial value for EM estimated first variance component for $\beta_g | b_g$.

`sig21.0` Initial value for EM estimated second variance component for $\beta_g | b_g$.

`p.0` Initial value for EM estimated proportion of differentially abundant compounds.

`psi.0` Initial value for EM estimated main effect of differential abundance.

References

M. Wan, "Model-based classification with applications to high-dimensional data in bioinformatics," Ph.D. dissertation, Cornell University, 2015.

Examples

```
expr   <- log(operators[, 1:12])                 # Use log of First Two Operators' Abundances
expr    <- t(expr)                                     # Transpose Matrix for n x G
trmt    <- c(rep(0,3), rep(1,3), rep(0,3), rep(1,3))   # SNP Vector
results <- runRRmix(Y.in=expr, SNP.in=trmt, q.in=1)    # RRmix Results

names(results)                                         # Display Result Components
```

Index