Package: isva (via r-universe)

Type Package

Title Independent Surrogate Variable Analysis

Version 1.9

Date 2017-1-14

Author Andrew E Teschendorff

Maintainer Andrew Teschendorff <a.teschendorff@ucl.ac.uk>

Depends qvalue, fastICA, JADE

Description Independent Surrogate Variable Analysis is an algorithm
for feature selection in the presence of potential confounding
factors (see Teschendorff AE et al 2011, <doi:10.1093/bioinformatics/btr171>).

License GPL-2

LazyLoad yes

NeedsCompilation no

Date/Publication 2017-01-14 01:35:23

Repository https://aet21.r-universe.dev

RemoteUrl https://github.com/cran/isva

RemoteRef HEAD

RemoteSha 5487ae5bfc201bccca2d44035f739a1d1bae22f7

Contents

DoISVA   *Feature selection using independent surrogate variables*

Description

Given a data matrix and a phenotype of interest, this function performs feature selection to identify features associated with the phenotype of interest in the presence of potential confounding factors. The algorithm first finds the variation in the data matrix not associated with the phenotype of interest (using a linear model), and subsequently performs Independent Component Analysis (ICA) on this residual variation matrix. The number of independent components to be inferred can be prespecified or estimated using Random Matrix Theory (RMT). Independent Surrogate Variables (ISVs) are constructed from the independent components and provide estimates of the effect of confounders on the data. If potential confounders are unknown (default NULL option) there will be as many ISVs as there are independent components in the residual variation space. If potential confounders are known (either exactly or subject to error/uncertainty) the algorithm will select only those independent components that correlate with the confounders. If potential confounders are specified it can happen that ISVA will not select any ISVs because none of the independent components correlates with the confounders. In this scenario ISVA should be rerun with the default (NULL) option. The constructed ISVs are finally included as covariates in a multivariate regression model to identify features that correlate with the phenotype of interest independently of the potential confounders. There are two implementations of ICA offered: JADE and fastICA. We note that the former will result in the same solution (therefore deterministic), whereas fastICA may result in convergence to different local minima for different runs. In the latter case, a consensus solution is advised if results vary between runs.

Usage

```
DoISVA(data.m, pheno.v, cf.m = NULL, factor.log, pvthCF = 0.01,
 th = 0.05, ncomp = NULL,icamethod=c("JADE","fastICA"))
```

Arguments

data.m	Data matrix: rows label features, columns label samples. It is assumed that number of features is much larger than number of samples.
pheno.v	Numeric vector of length equal to number of columns of data matrix. At present only numeric (ordinal) phenotypes are supported, so categorical phenotypes are excluded.
cf.m	Matrix of potential confounding factors. Rows label samples, Columns label confounding factors, which may be numeric or categorical. The default option (NULL) is for the case where potential confounding factors are not known or irrelevant.
factor.log	A logical vector of same length as columns of cf.m. FALSE indicates factor is to be treated as a numeric, TRUE as categorical.
pvthCF	P-value threshold to call a significant association between an independent surrogate variable and a confounding factor. By default this is 0.01.

th	False discovery rate threshold for feature selection. By default this is 0.05.
ncomp	Number of independent surrogate variables to look for. By default this is NULL, and estimation is performed using Random MatrixTheory.
icamethod	Method implementing ICA to be used. Must be either JADE or fastICA.

Value

A list with following entries:

spv	Sorted P-values.
rk	Ranked index of features.
qv	Estimated sorted q-values (False Discovery Rate).
ndeg	Number of differentially altered features.
deg	Indices of differentially altered features.
lm	Matrix of significant feature regression statistics and P-values.
isv	Matrix of selected independent surrogate variables (ISVs).
nsv	Number of selected ISVs.
pvCF	P-value matrix of associations between factors (phenotype of interest plus confounding factors) and inferred ISVs. Note that this may be a larger set than the selected ISVs.
selisv	Column indices of selected ISVs.

Author(s)

Andrew E Teschendorff

References

Independent Surrogate Variable Analysis to deconvolve confounding factors in large-scale microarray profiling studies. Teschendorff AE, Zhuang JJ, Widschwendter M. Bioinformatics. 2011 Jun 1;27(11):1496-505.

Examples

```
### Example

### load in simulated data
data(simdataISVA);
data.m <- simdataISVA$data;
pheno.v <- simdataISVA$pheno;

## factors matrix (two potential confounding factors, e.g chip and cohort)
factors.m <- cbind(simdataISVA$factors[[1]],simdataISVA$factors[[2]]);
colnames(factors.m) <- c("CF1","CF2");

### Estimate number of significant components of variation
rmt.o <- EstDimRMT(data.m);
```

```
print(paste("Number of significant components=",rmt.o$dim,sep=""));
### this makes sense since 1 component is associated with the
### the phenotype of interest, while the other two are associated
### with the confounders
ncp <- rmt.o$dim-1 ;

### Do ISVA
### run with the confounders as given
isva.o <- DoISVA(data.m,pheno.v,factors.m,factor.log=rep(FALSE,2),
pvthCF=0.01,th=0.05,ncomp=ncp,icamethod="fastICA");

### Evaluation (ISVs should correlate with confounders)
### modeling of CFs
print(cor(isva.o$isv,factors.m));
### this shows that CFs are reconstructed fairly well

### sensitivity (fraction of detected true positives)
print(length(intersect(isva.o$deg,simdataISVA$deg))/length(simdataISVA$deg));

### PPV (1-false discovery rate)
print(length(intersect(isva.o$deg,simdataISVA$deg))/length(isva.o$deg));

### run not knowing what confounders there are and with ncp=3 say.
isva2.o <- DoISVA(data.m,pheno.v,cf.m=NULL,factor.log=rep(FALSE,2),
pvthCF=0.01,th=0.05,ncomp=3,icamethod="fastICA");

### sensitivity (fraction of detected true positives)
print(length(intersect(isva2.o$deg,simdataISVA$deg))/length(simdataISVA$deg));

### PPV (1-false discovery rate)
print(length(intersect(isva2.o$deg,simdataISVA$deg))/length(isva2.o$deg));
```

EstDimRMT   *Estimates dimensionality of a data set using Random Matrix Theory*

Description

Given the data matrix, it estimates the number of significant components of variation by comparing the observed distribution of spectral eigenvalues to the theoretical one under a Gaussian Orthogonal Ensemble (GOE). Specifically, a spectral decomposition of the data covariance matrix is performed and the number of eigenvalues larger than the theoretical maximum predicted by the GOE is taken as an estimate of the number of significant components.

Usage

```
EstDimRMT(data.m,plot=TRUE)
```

Arguments

`data.m`	Data matrix. Rows label features, Columns samples.
`plot`	Logical. Plots Eigenvalue densities if true.

Value

A list with following objects

`cor`	Data covariance matrix.
`dim`	Estimated intrinsic dimensionality of data.
`estdens`	Empirical density of eigenvalues.
`thdens`	Theoretical density of eigenvalues.

Author(s)

Andrew E Teschendorff

References

Random matrix approach to cross correlations in financial data. Plerou et al. Physical Review E (2002), Vol.65.

Examples

```
## see example for DoISVA
```

---

isva *Independent Surrogate Variable Analysis*

Description

Independent Surrogate Variable Analysis is an algorithm for feature selection in the presence of potential confounding factors, specially designed for the analysis of large-scale high-dimensional quantitative genomic data (e.g microarrays). It uses Independent Component Analysis (ICA) to model the confounding factors as independent surrogate variables (ISVs). These ISVs are included as covariates in a multivariate regression model to subsequently identify features that correlate with a phenotype of interest independently of these confounders. Two ICA implementations are offered: JADE from the JADE R-package and fastICA from the fastICA R-package.

Details

Package:	isva
Type:	Package
Version:	1.9
Date:	2017-01-13
License:	GPL-2
LazyLoad:	yes

There are two internal functions. One function (EstDimRMT) performs the dimensionality estimation using a Random Matrix Theory approximation. The other function (isvaFn) is the main engine function and performs the modelling of confounding factors using Independent Component Analysis (ICA). Briefly, ICA is applied on the residual variation orthogonal to that of the phenotype of interest. DoISVA is the main user function, performing feature selection using the constructed independent surrogate variables as covariates.

Author(s)

Andrew E Teschendorff Maintainer:<a.teschendorff@ucl.ac.uk>

References

Independent Surrogate Variable Analysis to deconvolve confounding factors in large-scale microarray profiling studies. Teschendorff AE, Zhuang JJ, Widschwendter M. Bioinformatics. 2011 Jun 1;27(11):1496-505.

---

isvaFn	*Main engine function for inference of independent surrogate variables (ISVs)*

Description

This is the main engine function which infers the statistically independent surrogate variables (ISVs) by performing Independent Component Analysis (ICA) on the residual variation matrix. It uses either the ICA implementation of JADE or the one from the fastICA R-package. The residual variation matrix reflects the variation orthogonal to that of a phenotype of interest and is inferred using a linear model.

Usage

```
isvaFn(data.m, pheno.v, ncomp = NULL,icamethod)
```

Arguments

`data.m`	Data matrix. Rows label features. Columns label samples.
`pheno.v`	Numeric vector encoding phenotype of interest.
`ncomp`	Optionally specify number of ISVs to look for. By default will use Approximate Random Matrix Theory to infer this number.
`icamethod`	The ICA method to be used. Input value is taken from DoISVA.

Value

A list with following entries:

`n.isv`	Number of inferred ISVs.
`isv`	Matrix of inferred ISVs.

Author(s)

Andrew E Teschendorff

References

Independent Surrogate Variable Analysis to deconvolve confounding factors in large-scale microarray profiling studies. Teschendorff AE, Zhuang JJ, Widschwendter M. Bioinformatics. 2011 Jun 1;27(11):1496-505.

Examples

```
## see example for DoISVA
```

---

simdataISVA   *Simulated data for ISVA*

Description

A synthetic data set of 750 features and 50 samples with a binary phenotype and two confounding factors. Relative effect size of confounding factors (CFs) to that of phenotype of interest is 2. For further details please see reference.

Usage

```
simdataISVA
```

Format

This synthetic data set is a list object containing the following elements: (i) data is the data matrix (750 features, 50 samples), (ii) pheno is a binary phenotype vector, (iii) factors is a list of length two containing the two binary confounding factors, (iv) deg is the index vector of those truly differentially "expressed" features, (v) degL is a list of index vectors for features truly differentially altered (first element,degL[[1]]=deg) and those features affected by CFs (2nd and 3rd elements).

References

Independent Surrogate Variable Analysis to deconvolve confounding factors in large-scale microarray profiling studies. Teschendorff AE, Zhuang JJ, Widschwendter M. Bioinformatics. 2011 Jun 1;27(11):1496-505.

Index