Package 'svapls'

February 20, 2015

Type Package

Title Surrogate variable analysis using partial least squares in a gene expression study.

Version 1.4

Date 2013-09-19

Author Sutirtha Chakraborty, Somnath Datta and Susmita Datta

Maintainer Sutirtha Chakraborty <statistuta@gmail.com>

Depends R (>= 2.0), class, stats, pls

Description Accurate identification of genes that are truly differentially expressed over two sample varieties, after adjusting for hidden subject-specific effects of residual heterogeneity.

License GPL-3

Collate fitModel.R svpls.R hfp.R

NeedsCompilation no

Repository CRAN

Date/Publication 2013-09-20 08:13:19

R topics documented:

svapls-package *Surrogate variable analysis using Partial Least Squares in a gene expression data*

Description

The package svapls contains functions that are intended for the identification, correction and visualization of the hidden variability owing to a variety of unknown subject/sample specific effects of residual heterogeneity in a gene expression data.

Details

Package:	svapls
Type:	Package
Version:	1.4
Date:	2013-09-19
License:	GPL-3

The package can be used to find the genes that are truly differentially expressed between two types of samples (tissue types, biological conditions like Cancer/Non- Cancer samples, etc.), after adjusting for the hidden factors of residual heterogeneity in the data. The function svpls detects the truly positive genes after correcting for the hidden variation and also provides a modified gene expression matrix which is free from the spurious effects of the residual expression heterogeneity. Another important function hfp produces a heat- map representing the intensity of latent variability due to the unknown sample- specific factors, for any specified set of genes and subjects.

fitModel, svpls and hfp

Author(s)

Sutirtha Chakraborty, Somnath Datta and Susmita Datta.

Maintainer: Sutirtha Chakraborty <statistuta@gmail.com>

References

Sutirtha Chakraborty, Somnath Datta and Susmita Datta. (2012) Surrogate Variable Analysis Using Partial Least Squares in Gene Expression Studies. Bioinformatics.

Examples

```
data(hidden_fac.dat)
fit <- svpls(10,10,hidden_fac.dat,pmax = 5)
fit$genes
Y.corrected <- fit$Y.corr

data(hidden_fac.dat)
```

```
gen <- paste("g",c(1:15,50:65),sep="")
sub <- paste("S",c(1:5,11:17),sep="")

hfp(fit,gen,sub,hidden_fac.dat)
```

`fitModel` *Function to fit an ANCOVA model to the log transformed gene expression data, with a certain specified number of surrogate variables.*

Description

This function begins its operation by fitting a standard ANOVA model to the gene expression data, with the gene, variety main effects and their mutual interaction. The residuals from the fit of this model and the original gene expression values are then respectively organized into two matrices E and Y, where each column corresponds to a certain gene. Now E is regressed on Y by Partial Least Squares (PLS) and a specified number of scores are extracted as the estimates of the latent components from their respective column spaces. The scores in the Y-space are used as surrogate variables along with the gene and variety interaction effects with the first score and the usual effects from the standard ANOVA model, in order to fit an ANCOVA model to the data. The function returns the results from this fit.

Usage

```
fitModel(k1, k2, Y, n.surr)
```

Arguments

`k1`	Number of subjects/samples under variety 1.
`k2`	Number of subjects/samples under variety 2.
`Y`	The log transformed gene expression data, with genes along the rows and subjects/samples along the columns.
`n.surr`	The specified number of surrogate variables.

Value

`mu.hat`	Intercept (general mean effect).
`G.hat`	Main effects of the genes.
`V.hat`	Main effects of the varieties.
`GV.hat`	Gene-Variety interaction effects.
`sc`	Values of the Surrogate variables (computed only when `n.surr`>0).
`beta.hat`	Coefficients of the surrogate variables (computed only when `n.surr`>0).
`GZ1.hat`	Interaction effects of the genes with the first surrogate variable (computed only when `n.surr`>0).
`VZ1.hat`	Interaction effects of the varieties with the fist surrogate variable (computed only when `n.surr`>0).

vhat.gvh	Variances of the estimators for the gene-variety interaction effects.
MSE	Mean Squarred Error for the fitted model.
AIC	Value of the Akaike's Information Criterion (AIC) for the fitted model.

Author(s)

Sutirtha Chakraborty, Somnath Datta and Susmita Datta.

References

Sutirtha Chakraborty, Somnath Datta and Susmita Datta. (2012) Surrogate Variable Analysis Using Partial Least Squares in Gene Expression Studies. Bioinformatics. Martens, H., Naes, T. (1989) Multivariate Calibration. Chicestor:Wiley.

See Also

svpls, hfp

Examples

```
data(hidden_fac.dat)

## Fitting an ANCOVA model with 5 surrogate variables
fit <- fitModel(10,10,hidden_fac.dat,n.surr = 5)
print(fit)
```

---

hfp *Function to construct a heatmap of the hidden variation in the gene expression data.*

Description

The function hfp produces a plot of the PLS imputed estimate of the hidden variability in the data, derived from the optimal model, corresponding to an user-specified set of genes and subjects/samples.

Usage

```
hfp(obj, gen, ind, Y)
```

Arguments

obj	An svpls object.
gen	An user-specified set of genes.
ind	An user-specified set of subjects.
Y	A log transformed gene expression matrix with genes along the rows and subjects/samples along the columns.

Value

A heatmap of the hidden variability corresponding to the specified set of genes and subjects, attributable to the unknown subject-specific factors in the gene expression data.

Author(s)

Sutirtha Chakraborty, Somnath Datta and Susmita Datta.

References

Sutirtha Chakraborty, Somnath Datta and Susmita Datta. (2012) Surrogate Variable Analysis Using Partial Least Squares in Gene Expression Studies. Bioinformatics.

See Also

heatmap, fitModel, svpls

Examples

```
## Fitting the optimal ANCOVA model to the data gives:
data(hidden_fac.dat)
fit <- svpls(10,10,hidden_fac.dat,pmax = 5)

## Specifying the sets of genes and subjects
gen <- paste("g",c(1:15,50:65),sep="")
sub <- paste("S",c(1:5,11:17),sep="")

hfp(fit,gen,sub,hidden_fac.dat)
```

---

hidden_fac.dat *A gene expression data affected by a hidden variable.*

Description

The dataset contains the log transformed expression levels of 500 genes over 20 subjects distributed equally between two varieties 1 and 2. The data is affected by the unknown effects from a hidden confounder whose effect changes over the two sample varieties.

Usage

```
data(hidden_fac.dat)
```

Format

A data frame with 500 observations on the following 20 variables.

`S1` a numeric vector

`S2` a numeric vector

`S3` a numeric vector

`S4` a numeric vector

`S5` a numeric vector

`S6` a numeric vector

`S7` a numeric vector

`S8` a numeric vector

`S9` a numeric vector

`S10` a numeric vector

`S11` a numeric vector

`S12` a numeric vector

`S13` a numeric vector

`S14` a numeric vector

`S15` a numeric vector

`S16` a numeric vector

`S17` a numeric vector

`S18` a numeric vector

`S19` a numeric vector

`S20` a numeric vector

Examples

```
data(hidden_fac.dat)
## maybe str(hidden_fac.dat) ; plot(hidden_fac.dat) ...
```

`svpls` *Function for identfying the optimal ANCOVA model and detecting the genes that are truly differentially expressed between the two types of samples.*

Description

This function calls `fitModel` repeatedly to fit a series of ANCOVA models along with the standard ANOVA model, to the log transformed gene expression data. The model with the minimum AIC is selected as the optimal one and its corresponding estimated effects are then used to perform a multiple testing of differential expression, over all the genes, using the Benjamini-Hochberg correction.

Usage

```
svpls(k1, k2, Y, pmax = 3, fdr = 0.05)
```

Arguments

k1	Number of subjects/samples under variety 1.
k2	Number of subjects/samples under variety 2.
Y	The log transformed gene expression data, with genes along the rows and subjects/samples along the columns.
pmax	Maximum number of surrogate variables to be incorporated in the ANCOVA model (means pmax ANCOVA models are fitted to the data). By default, it is taken as 3.
fdr	The specified False Discovery Rate (FDR) for multiple testing of differential expression, using the Benjamini-Hochberg correction. By Default it is taken as 0.05.

Value

opt.model	The optimal model. 1 denotes the standard ANOVA model.
PLS.imp	PlS imputed estimate of the hidden expression heterogeneity, evaluated from the optimal model (applicable only when opt.model>1).
Y.corr	Corrected gene expression matrix after adjusting for the hidden effects (applicable only when opt.model>1).
pvalues	p-values from the tests with the effects estimated from the standard ANOVA model (returned only when opt.model=1.
pvalues.adj	Adjusted p-values after correcting for the hidden effects (applicable only when opt.model>1).
genes	Genes that are deemed to be differentially expressed from the multiple hypotheses testing with effects estimated from the optimal model.
AIC.opt	AIC value for the optimal model.

Author(s)

Sutirtha Chakraborty, Somnath Datta and Susmita Datta.

References

Hirotsugu, A. (1980) Likelihood and the Bayes Procedure. The Institute of Statistical Mathematics, Tokyo., Benjamini, Y and Hochberg, Y (1995) Controlling the false discovery rate : a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society.

See Also

fitModel, hfp

Examples

```
## Loading the first dataset
data(hidden_fac.dat)

## Fitting the optimal ANCOVA model to the data gives:
fit <- svpls(10,10,hidden_fac.dat,pmax = 5)

## The optimal ANCOVA model, its AIC value and the positive genes detected from it are given by:
fit$opt.model

fit$AIC.opt

fit$genes

## The corrected gene expression matrix obtained after removing the effects of
## the hidden variability is given by:

Y.corrected <- fit$Y.corr
```

Index