

# Package ‘sda’

January 22, 2012

**Version** 1.2.1

**Date** 2012-01-22

**Title** Shrinkage Discriminant Analysis and CAT Score Variable Selection

**Author** Miika Ahdesmaki, Verena Zuber, and Korbinian Strimmer

**Maintainer** Korbinian Strimmer <strimmer@uni-leipzig.de>

**Depends** R (>= 2.10.0), lattice, entropy (>= 1.1.7), corpcor (>= 1.6.2), fdrtool (>= 1.2.8)

**Suggests**

**Description** This package provides an efficient framework for high-dimensional linear and diagonal discriminant analysis with variable selection. The classifier is trained using James-Stein-type shrinkage estimators and predictor variables are ranked using CAT scores (correlation-adjusted t-scores). Variable selection error is controlled using false non-discovery rates or higher criticism scores.

**License** GPL (>= 3)

**URL** <http://strimmerlab.org/software/sda/>

**Repository** CRAN

**Date/Publication** 2012-01-22 08:10:01

## R topics documented:

sda-package . . . . .	2
catscore . . . . .	3
centroids . . . . .	4
khan2001 . . . . .	6
predict.sda . . . . .	7
sda . . . . .	8
sda.ranking . . . . .	10
singh2002 . . . . .	13

<b>Index</b>	<b>14</b>
--------------	-----------

---

sda-package

*The sda package*

---

## Description

This package performs linear discriminant analysis (LDA) and diagonal discriminant analysis (DDA) with variable selection using correlation-adjusted t (CAT) scores.

The classifier is trained using James-Stein-type shrinkage estimators. Variable selection is based on ranking predictors by CAT scores (LDA) or t-scores (DDA). A cutoff is chosen by false non-discovery rate (FNDR) or higher criticism (HC) thresholding.

This approach is particularly suited for high-dimensional classification with correlation among predictors. For details see Zuber and Strimmer (2009) and Ahdesm\`aki and Strimmer (2010).

Typically the functions in this package are applied in three steps:

- Step 1: feature selection with `sda.ranking`,
- Step 2: training the classifier with `sda`, and
- Step 3: classification using `predict.sda`.

The accompanying web site (see below) provides example R scripts to illustrate the functionality of this package.

## Author(s)

Miika Ahdesm\`aki, Verena Zuber and Korbinian Strimmer (<http://strimmerlab.org/>)

## References

Ahdesm\`aki, A., and K. Strimmer. 2010. Feature selection in omics prediction problems using cat scores and false non-discovery rate control. *Ann. Appl. Stat.* 4: 503-519. Preprint available from <http://arxiv.org/abs/0903.2003>.

Zuber, V., and K. Strimmer. 2009. Gene ranking and biomarker discovery under correlation. *Bioinformatics* 25: 2700-2707. Preprint available from <http://arxiv.org/abs/0902.0751>.

See website: <http://strimmerlab.org/software/sda/>

## See Also

`catscore`, `sda.ranking`, `sda`, `predict.sda`.

---

catscore                      *Estimate CAT scores and t-scores*

---

**Description**

catscore computes CAT scores (correlation-adjusted t-scores) between the group centroids and the pooled mean.

**Usage**

```
catscore(Xtrain, L, diagonal=FALSE, shrink=FALSE, verbose=TRUE)
```

**Arguments**

Xtrain	A matrix containing the training data set. Note that the rows correspond to observations and the columns to variables.
L	A factor with the class labels of the training samples.
diagonal	for diagonal=FALSE (the default) CAT scores are computed; otherwise with diagonal=TRUE t-scores.
shrink	Use empirical estimates or a shrinkage estimator for the CAT score.
verbose	Print out some info while computing.

**Details**

CAT scores generalize conventional t-scores to account for correlation among predictors (Zuber and Strimmer 2009). If there is no correlation then CAR scores reduce to t-scores. The squared CAR scores provide a decomposition of Hotelling's  $T^2$  statistic.

CAT scores for two classes are described in Zuber and Strimmer (2009), for the multi-class case see Ahdesmaki and Strimmer (2010).

**Value**

catscore returns a matrix containing the cat score (or t-score) between each group centroid and the pooled mean for each feature.

**Author(s)**

Verena Zuber, Miika Ahdesmaki and Korbinian Strimmer (<http://strimmerlab.org>).

**References**

Ahdesmaki, A., and K. Strimmer. 2010. Feature selection in omics prediction problems using cat scores and false non-discovery rate control. *Ann. Appl. Stat.* 4: 503-519. Preprint available from <http://arxiv.org/abs/0903.2003>.

Zuber, V., and K. Strimmer. 2009. Gene ranking and biomarker discovery under correlation. *Bioinformatics* 25: 2700-2707. Preprint available from <http://arxiv.org/abs/0902.0751>.

**See Also**

[sda.ranking](#), [carscore](#),.

**Examples**

```
# load sda library
library("sda")

#####
# training data #
#####

# prostate cancer set
data(singh2002)

# training data
Xtrain = singh2002$x
Ytrain = singh2002$y
dim(Xtrain)

#####
# shrinkage t-score (DDA setting - no correlation) #
#####

tstat = catscore(Xtrain, Ytrain, diagonal=TRUE, shrink=TRUE)
dim(tstat)
tstat[1:10,]

#####
# shrinkage CAT score (LDA setting - with correlation) #
#####

cat = catscore(Xtrain, Ytrain, diagonal=FALSE, shrink=TRUE)
dim(cat)
cat[1:10,]
```

---

centroids

*Group Centroids and (Pooled) Variances*


---

**Description**

centroids computes group centroids, the pooled mean and pooled variance, and optionally the group specific variances.

**Usage**

```
centroids(x, L, var.groups=FALSE, centered.data=FALSE, shrink=FALSE, verbose=TRUE)
```

**Arguments**

x	A matrix containing the data set. Note that the rows are sample observations and the columns are variables.
L	A factor with the group labels.
var.groups	Estimate group-specific variances.
centered.data	Return column-centered data matrix.
shrink	Use empirical estimates or a shrinkage estimator for the variances.
verbose	Provide some messages while computing.

**Details**

If option `shrink=TRUE` then the shrinkage estimators `var.shrink` from Opgen-Rhein and Strimmer (2007) and `cor.shrink` from Sch\"afer and Strimmer (2005) are used.

Details on the algorithm for efficiently computing the power of the shrinkage correlation matrix are given in Zuber and Strimmer (2009).

**Value**

`centroids` returns a list with the following components:

samples	a vector containing the samples sizes in each group,
means	the group means and the pooled mean,
variances	the group-specific and the pooled variances, and
centered.data	a matrix containing the centered data.

**Author(s)**

Korbinian Strimmer (<http://strimmerlab.org>).

**See Also**

[var.shrink](#), [powcor.shrink](#).

**Examples**

```
# load sda library
library("sda")

## prepare data set
data(iris) # good old iris data
X = as.matrix(iris[,1:4])
Y = iris[,5]

## estimate centroids and empirical pooled variances
centroids(X, Y)

## also compute group-specific variances
centroids(X, Y, var.groups=TRUE)
```

```
## use shrinkage estimator for the variances
centroids(X, Y, var.groups=TRUE, shrink=TRUE)

## return centered data
xc = centroids(X, Y, centered.data=TRUE)$centered.data
apply(xc, 2, mean)

## compute pooled inverse correlation matrix
powcor.shrink(xc, alpha=-1)
```

---

khan2001

*Childhood Cancer Study of Khan et al. (2001)*

---

### Description

Gene expression data (2308 genes for 88 samples) from the microarray study of Khan et al. (2001).

### Usage

```
data(khan2001)
```

### Format

khan2001\$x is a 88 x 2308 matrix containing the expression levels. Note that rows correspond to samples, and columns to genes. The row names are the original image IDs, and the column names the original probe labels.

khan2001\$y is a factor containing the diagnosis for each sample ("BL", "EWS", "NB", "non-SRBCT", "RMS").

khan2001\$descr provides some annotation for each gene.

### Details

This data set contains measurements of the gene expression of 2308 genes for 88 observations: 29 cases of Ewing sarcoma (EWS), 11 cases of Burkitt lymphoma (BL), 18 cases of neuroblastoma (NB), 25 cases of rhabdomyosarcoma (RMS), and 5 other (non-SRBCT) samples.

### Source

The data are described in Khan et al. (2001) and can be obtained from [http://cbbp.thep.lu.se/pub/Preprints/01/lu\\_tp\\_01\\_06\\_supp.html](http://cbbp.thep.lu.se/pub/Preprints/01/lu_tp_01_06_supp.html). Note that the values in khan.data\$x are additionally logarithmized (using natural `log`) for normalization.

### References

Khan et al. 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine* 7:673–679.

**Examples**

```
# load sda library
library("sda")

# load full Khan et al (2001) data set
data(khan2001)
dim(khan2001$x) # 88 2308
hist(khan2001$x)
khan2001$y # 5 levels

# data set containing the SRBCT samples
get.srbct = function()
{
  data(khan2001)
  idx = which( khan2001$y == "non-SRBCT" )
  x = khan2001$x[-idx,]
  y = factor(khan2001$y[-idx])
  descr = khan2001$descr[-idx]

  list(x=x, y=y, descr=descr)
}

srbct = get.srbct()
dim(srbct$x) # 83 2308
hist(srbct$x)
srbct$y # 4 levels
```

---

predict.sda

*Shrinkage Discriminant Analysis 3: Prediction Step*


---

**Description**

predict.sda performs class prediction.

**Usage**

```
## S3 method for class 'sda'
predict(object, Xtest, feature.idx, verbose=TRUE, ...)
```

**Arguments**

object	An sda fit object obtained from the function sda.
Xtest	A matrix containing the test data set. Note that the rows correspond to observations and the columns to variables.
feature.idx	A vector indicating which features to employ for prediction (if unspecified all features will be used).
verbose	Report shrinkage intensities (sda) and number of used features (predict.sda).
...	Additional arguments for generic predict.

**Value**

`predict.sda` predicts class probabilities for each test sample and returns a list with two components:

<code>class</code>	a factor with the most probable class assignment for each test sample, and
<code>posterior</code>	a matrix containing the respective class posterior probabilities.

**Author(s)**

Miika Ahdeml"aki and Korbinian Strimmer (<http://strimmerlab.org>).

**See Also**

[sda](#), [sda.ranking](#).

**Examples**

```
# see the examples at the "sda" help page
```

---

sda

*Shrinkage Discriminant Analysis 2: Training Step*

---

**Description**

`sda` trains a LDA or DDA classifier using James-Stein-type shrinkage estimation.

**Usage**

```
sda(Xtrain, L, diagonal=FALSE, verbose=TRUE)
```

**Arguments**

<code>Xtrain</code>	A matrix containing the training data set. Note that the rows correspond to observations and the columns to variables.
<code>L</code>	A factor with the class labels of the training samples.
<code>diagonal</code>	Chooses between LDA (default, <code>diagonal=FALSE</code> ) and DDA ( <code>diagonal=TRUE</code> ).
<code>verbose</code>	Print out some info while computing.

**Details**

In order to train the LDA or DDA classifier, three separate shrinkage estimators are employed:

- class frequencies: the estimator [freqs.shrink](#) from Hausser and Strimmer (2008),
- variances: the estimator [var.shrink](#) from Opgen-Rhein and Strimmer (2007),
- correlations: the estimator [cor.shrink](#) from Sch" afer and Strimmer (2005).

Note that the three corresponding regularization parameters are obtained analytically without resorting to computer intensive resampling.

**Value**

sda trains the classifier and returns an sda object with the following components needed for the subsequent prediction:

regularization a vector containing the three estimated shrinkage intensities,  
 prior the estimated class frequencies,  
 predcoef matrix containing the coefficients used for prediction

**Author(s)**

Miika Ahdesmaki and Korbinian Strimmer (<http://strimmerlab.org>).

**References**

Ahdesmaki, A., and K. Strimmer. 2010. Feature selection in omics prediction problems using cat scores and false non-discovery rate control. Ann. Appl. Stat. 4: 503-519. Preprint available from <http://arxiv.org/abs/0903.2003>.

**See Also**

[predict.sda](#), [sda.ranking](#), [freqs.shrink](#), [var.shrink](#), [invcor.shrink](#).

**Examples**

```
# load sda library
library("sda")

#####
# training and test data #
#####

# data set containing the SRBCT samples
get.srbct = function()
{
  data(khan2001)
  idx = which( khan2001$y == "non-SRBCT" )
  x = khan2001$x[-idx,]
  y = factor(khan2001$y[-idx])
  descr = khan2001$descr[-idx]

  list(x=x, y=y, descr=descr)
}
srbct = get.srbct()

# training data
Xtrain = srbct$x[1:63,]
Ytrain = srbct$y[1:63]
Xtest = srbct$x[64:83,]
Ytest = srbct$y[64:83]
```

```
#####
# classification with correlation (shrinkage LDA) #
#####

sda.fit = sda(Xtrain, Ytrain)
ynew = predict(sda.fit, Xtest)$class # using all 2308 features
sum(ynew != Ytest)

#####
# classification with diagonal covariance (shrinkage DDA) #
#####

sda.fit = sda(Xtrain, Ytrain, diagonal=TRUE)
ynew = predict(sda.fit, Xtest)$class # using all 2308 features
sum(ynew != Ytest)

#####
# for complete example scripts illustrating classification with #
# feature selection visit http://strimmerlab.org/software/sda/ #
#####
```

---

sda.ranking

*Shrinkage Discriminant Analysis 1: Predictor Ranking*


---

## Description

sda.ranking determines a ranking of predictors by computing CAT scores (correlation-adjusted t-scores) between the group centroids and the pooled mean.

plot.sda.ranking provides a graphical visualization of the top ranking features..

## Usage

```
sda.ranking(Xtrain, L, diagonal=FALSE, fdr=TRUE, plot.fdr=FALSE, verbose=TRUE)
## S3 method for class 'sda.ranking'
plot(x, top=40, ...)
```

## Arguments

Xtrain	A matrix containing the training data set. Note that the rows correspond to observations and the columns to variables.
L	A factor with the class labels of the training samples.
diagonal	Chooses between LDA (default, diagonal=FALSE) and DDA (diagonal=TRUE).
fdr	compute FDR values and HC scores for each feature.
plot.fdr	Show plot with estimated FDR values.
verbose	Print out some info while computing.
x	An "sda.ranking" object – this is produced by the sda.ranking() function.
top	The number of top-ranking features shown in the plot (default: 40).
...	Additional arguments for generic plot.

**Details**

For each predictor variable and centroid a shrinkage CAT scores of the mean versus the pooled mean is computed. The overall ranking of a feature is determined by the sum of the squared cat scores across all centroids. For the diagonal case (LDA) the (shrinkage) CAT score reduces to the (shrinkage) t-score. Thus in the two-class diagonal case the feature are simply ranked according to the (shrinkage) t-scores.

Calling `sda.ranking` is step 1 in a classification analysis with the `sda` package. Steps 2 and 3 are `sda` and `predict.sda`

See Ahdesm\`aki and Strimmer (2010) for details on multi-class CAT scores, Zuber and Strimmer (2009) for CAT scores in general. For shrinkage t scores see Opgen-Rhein and Strimmer (2007).

**Value**

`sda.ranking` returns a matrix with the following columns:

<code>idx</code>	original feature number
<code>score</code>	sum of the squared CAT scores across groups - this determines the overall ranking of a feature
<code>cat</code>	for each group and feature the cat score of the centroid versus the pooled mean

If `fdr=TRUE` then additionally local false discovery rate (FDR) values as well as higher criticism (HC) scores are computed for each feature (using `fdrtool`).

**Author(s)**

Miiika Ahdesm\`aki, Verena Zuber and Korbinian Strimmer (<http://strimmerlab.org>).

**References**

Ahdesm\`aki, A., and K. Strimmer. 2010. Feature selection in omics prediction problems using cat scores and false non-discovery rate control. *Ann. Appl. Stat.* 4: 503-519. Preprint available from <http://arxiv.org/abs/0903.2003>.

Opgen-Rhein, R., and K. Strimmer. 2007. Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Statist. Appl. Genet. Mol. Biol.* 6:9.

Zuber, V., and K. Strimmer. 2009. Gene ranking and biomarker discovery under correlation. *Bioinformatics* 25: 2700-2707. Preprint available from <http://arxiv.org/abs/0902.0751>.

**See Also**

[catscore](#), [sda](#), [predict.sda](#).

**Examples**

```
# load sda library
library("sda")

#####
# training data #
```

```
#####  
  
# prostate cancer set  
data(singh2002)  
  
# training data  
Xtrain = singh2002$x  
Ytrain = singh2002$y  
  
#####  
# feature ranking (diagonal covariance) #  
#####  
  
# ranking using t-scores (DDA)  
ranking.DDA = sda.ranking(Xtrain, Ytrain, diagonal=TRUE)  
ranking.DDA[1:10,]  
  
# plot t-scores for the top 40 genes  
plot(ranking.DDA, top=40)  
  
# number of features with local FDR < 0.8  
# (i.e. features useful for prediction)  
sum(ranking.DDA[, "lfd"] < 0.8)  
  
# number of features with local FDR < 0.2  
# (i.e. significant non-null features)  
sum(ranking.DDA[, "lfd"] < 0.2)  
  
# optimal feature set according to HC score  
plot(ranking.DDA[, "HC"], type="l")  
which.max( ranking.DDA[1:1000, "HC"] )  
  
#####  
# feature ranking (full covariance) #  
#####  
  
# ranking using CAT-scores (LDA)  
ranking.LDA = sda.ranking(Xtrain, Ytrain, diagonal=FALSE)  
ranking.LDA[1:10,]  
  
# plot t-scores for the top 40 genes  
plot(ranking.LDA, top=40)  
  
# number of features with local FDR < 0.8  
# (i.e. features useful for prediction)  
sum(ranking.LDA[, "lfd"] < 0.8)  
  
# number of features with local FDR < 0.2  
# (i.e. significant non-null features)  
sum(ranking.LDA[, "lfd"] < 0.2)  
  
# optimal feature set according to HC score
```

```
plot(ranking.LDA[, "HC"], type="l")
which.max( ranking.LDA[1:1000, "HC"] )
```

---

singh2002

*Prostate Cancer Study of Singh et al. (2002)*

---

### Description

Gene expression data (6033 genes for 102 samples) from the microarray study of Singh et al. (2002).

### Usage

```
data(singh2002)
```

### Format

singh2002\$x is a 102 x 6033 matrix containing the expression levels. The rows contain the samples and the columns the genes.

singh2002\$y is a factor containing the diagnosis for each sample ("cancer" or "healthy").

### Details

This data set contains measurements of the gene expression of 6033 genes for 102 observations: 52 prostate cancer patients and 50 healthy men.

### Source

The data are described in Singh et al. (2001) and are provided in exactly the form as used by Efron (2008) - see <http://www-stat.stanford.edu/~ckirby/brad/papers/Ebaydata.R>.

### References

D. Singh et al. 2002. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 1:203–209.

Efron, B. 2008. Empirical Bayes estimates for large-scale prediction problems. Technical Report, Stanford University.

### Examples

```
# load sda library
library("sda")

# load Singh et al (2001) data set
data(singh2002)
dim(singh2002$x) # 102 6033
hist(singh2002$x)
singh2002$y # 2 levels
```

# Index

## \*Topic **datasets**

khan2001, [6](#)  
singh2002, [13](#)

## \*Topic **multivariate**

catscore, [3](#)  
centroids, [4](#)  
predict.sda, [7](#)  
sda, [8](#)  
sda-package, [2](#)  
sda.ranking, [10](#)

carscore, [4](#)  
catscore, [2](#), [3](#), [11](#)  
centroids, [4](#)  
cor.shrink, [5](#), [8](#)

fdrtool, [11](#)  
freqs.shrink, [8](#), [9](#)

invcor.shrink, [9](#)

khan2001, [6](#)

log, [6](#)

plot.sda.ranking (sda.ranking), [10](#)  
powcor.shrink, [5](#)  
predict.sda, [2](#), [7](#), [9](#), [11](#)

sda, [2](#), [8](#), [8](#), [11](#)  
sda-package, [2](#)  
sda.ranking, [2](#), [4](#), [8](#), [9](#), [10](#)  
singh2002, [13](#)

var.shrink, [5](#), [8](#), [9](#)