Package ‘HDDesign’

June 9, 2016

Version 1.1

Date 2016-06-09

Title Sample Size Calculation for High Dimensional Classification Study

Author Meihua Wu <meihuawu@umich.edu>,
Brisa N. Sanchez <brisa@umich.edu>,
Peter X.K. Song <pxsong@umich.edu>,
Raymond Luu <raluu@umich.edu>,
Wen Wang <wangwen@umich.edu>

Maintainer Brisa N. Sanchez <brisa@umich.edu>

Description Determine the sample size requirement to achieve the target probability of correct classification (PCC) for studies employing high-dimensional features. The package implements functions to 1) determine the asymptotic feasibility of the classification problem; 2) compute the upper bounds of the PCC for any linear classifier; 3) estimate the PCC of three design methods given design assumptions; 4) determine the sample size requirement to achieve the target PCC for three design methods.

License GPL-2

NeedsCompilation yes

R topics documented:

HDDesign-package *Sample Size Calculation for High Dimensional Classification Study*

Description

This package facilitates the design of studies employing high dimensional features for binary classification. The package assumes that the study will build a linear classifier by first screening features and selecting those that appear important for classification, and then forming a linear predictor based on the selected features. The package implements functions to 1) determine the asymptotic feasibility of the classification problem; 2) compute the upper bounds of the PCC for any linear classifiers; 3) estimate the PCC of three design methods given design assumptions; 4) determine the sample size requirement to achieve the target PCC for various design methods.

Details

Package:	HDDesign
Type:	Package
Version:	1.1
Date:	2016-04-26
License:	GPL-2

Design of high-dimensional classification studies involves several aspects. Firstly, we need to consider the feasibility of the classification. For high dimensional classification, the important signals are sometimes so sparse and weak that the performance of any linear classifiers is no better than a random assignment classifier. We implement functions to determine the asymptotic feasibility of the classification problem based on the theory of the rare and weak model (Donoho and Jin 2009). If the classification is feasible, we need to determine the appropriate target PCC to calibrate the sample size calculation. The lower bound of the PCC corresponds to the worst case scenario where the classifier performs as poor as the random assignment classifier. Then the lower bound of the PCC is the prevalence of the dominating group. The upper bound of the PCC corresponds to the best case scenario where the classifier performs as good as the ideal classifier which uses full knowledge of the data generating mechanism by Dobbin and Simon (2007). We implement a function to compute the PCC of the ideal classifier. Then the target PCC can be chosen between the lower and upper bounds, depending on the budget and other constraints of the study. Furthermore, we need to incorporate the uncertainty in feature selection into our sample size calculations in the design stage to ensure the target PCC is achieved at the analysis stage. We have implemented the following feature selection methods: a procedure proposed by Dobbin and Simon (2007), thresholding by cross validation, and Higher Criticism Threshold (HCT) by Donoho and Jin (2009). We implement an efficient algorithm to calculate the sample size required when features are iid, for the case when HCT is used to build the classifier. The CV method is more computationally intensive. We also adapt the approaches for the cases when features are correlated, however the resulting sample sizes will often be conservative.

In this package we use the following assumptions and notations. We assume features have the same variance across groups, and for simplicity assume all variances equal 1. If the differences between the means of a feature between the groups is zero, then this feature does not differentially express between the groups and it is therefore not important for the purpose of classification. If, however, the difference is non zero, this feature is important and its effect size is half of the absolute value of the difference. We denote the effect size by mu0, the total number of features by p; the number of important features by m; and the total sample size for two groups by n; the prevalence of "group 1" in the population by p1 and the prevalence of "group -1" by 1-p1. Finally, pvalues for all pairwise associations are derived from the t-distribution instead of a normal distribution to take into account the fact that some studies may have a small sample size.

In the example below, we will illustrate how to use the functions implemented in this package to address the sample size calculation problem for studies employing high dimensional features for classification.

Author(s)

Meihua Wu <meihuawu@umich.edu> Brisa N. Sanchez <brisa@umich.edu> Peter X.K. Song <pxsong@umich.edu> Raymond Luu <raluu@umich.edu> Wen Wang <wangwen@umich.edu>

References

Donoho, D, and Jin, J. (2009). "Feature Selection by Higher Criticism Thresholding Achieves the Optimal Phase Diagram." Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences 367 (1906) : 4449-4470.

Dobbin, K.K., and Simon R.M. (2007). "Sample Size Planning for Developing Classifiers Using High-dimensional DNA Microarray Data." Biostatistics 8 (1): 101-117.

Sanchez, B.N., Wu, M., Song, P.X.K., and Wang W. (2016). "Study design in high-dimensional classification analysis." Biostatistics, in press.

Examples

```
# Consider the following design scenario:
# Prevalence of Group 1
p1=0.5
# Effect size
mu0=0.4
# Total number of features
p=500
# The number of important features
m=10


# Step 1: Feasibility of the classification for a study with about 100 individuals
which.region(mu0=mu0, p=p, m=m, n=100)
# return 4, indicating the classification belongs to the feasible region.

# Step 2: Upper bound of the PCC
ideal_pcc(mu0=mu0, m=m, p1=p1)
# return 0.8970484,
# So the target PCC can be chosen between 0.5 and 0.8970484.

# Step 3: Obtain the sample requirement for target PCC equal to 0.8
#Use method proposed by Dobbin and Simon (2007)
set.seed(1)
samplesize(target=0.8, nmin=20, nmax=100, ds_method, mu0=0.4, p=500, m=10)
#return sample size n=66

#Use cross validation(commented due to long waiting time)
#set.seed(1)
#samplesize(target=0.8, nmin=20, nmax=100, cv_method, mu0=0.4, p=500, m=10,
#alpha_list=10^((-10):(-2)), nrep=100)
#alpha_list should be a dense list of p-value cutoffs;
```

```
#here we only use a few values to ease computation of the example.
#return sample size n=78.

#Use HCT
set.seed(1)
samplesize(target=0.8, nmin=20, nmax=100, hct_method, mu0=0.4, p=500, m=10,
hct=hct_beta, alpha0=0.5, nrep=100)
#return sample size n=78.
```

cv_method *Formula-based PCC of a CV-based classifier*

Description

Determine the probability of correct classification (PCC) for a high dimensional classification study employing cross validation to determine an optimal p-value cutoff to select features that are included in a linear classifier.

Usage

```
cv_method(mu0, p, m, n, alpha_list, nrep, p1 = 0.5, ss = F, sampling.p=0.5)
```

Arguments

mu0	The effect size of the important features.
p	The number of the features in total.
m	The number of the important features.
n	The total sample size for the two groups.
alpha_list	The search grid for the p-value threshold.
nrep	The number of simulation replicates employed to compute the expected PCC and/or sensitivity and specificity.
p1	The prevalence of the group 1 in the population, default to 0.5.
ss	Boolean variable, default to FALSE. The TRUE value instruct the program to compute the sensitivity and the specificity of the classifier.
sampling.p	The assumed proportion of group 1 samples in the training data; default of 0.5 assumes groups are equally represented regardless of p1.

Details

Refer to Sanchez, Wu, Song, Wang 2016, Section 2.2 for the details of the algorithm. This function was used to produce Figure 2 of the paper.

Value

If ss=FALSE, the function returns the expected PCC. If ss=TRUE, the function returns a vector containing the expected PCC, sensitivity and specificity.

Author(s)

Meihua Wu <meihuawu@umich.edu> Brisa N. Sanchez <brisa@umich.edu> Peter X.K. Song <pxsong@umich.edu> Raymond Luu <raluu@umich.edu> Wen Wang <wangwen@umich.edu>

References

Sanchez, B.N., Wu, M., Song, P.X.K., and Wang W. (2016). "Study design in high-dimensional classification analysis." Biostatistics, in press.

Examples

```
set.seed(1)
cv_method(mu0=0.4, p=500, m=10, n=80, alpha_list=c(0.0000001, 0.0001, 0.01),
nrep=10, p1=0.6, ss=TRUE)
#return: 0.8012142 0.8210082 0.7714031
#alpha_list should be a dense list of p-value cutoffs;
#here we only use a few values to ease computation of the example.
```

`cv_method_corr` *Formula-based method to calculate the PCC of a CV-based classifier when features are correlated.*

Description

Determine the probability of correct classification (PCC) for a high dimensional classification study employing Cross validation classifier. This is similar to cv_method, but features generated are correlated.

Usage

```
cv_method_corr(mu0, p, m, n, alpha_list, nrep, p1 = 0.5, ss = F, pcorr,
chol.rho,sampling.p=0.5)
```

Arguments

`mu0`	The effect size of the important features.
`p`	The number of the features in total.
`m`	The number of the important features.
`n`	The total sample size for the two groups.
`alpha_list`	The search grid for the p-value threshold.
`nrep`	The number of simulation replicates employed to compute the expected PCC and/or sensitivity and specificity.
`p1`	The prevalence of the group 1 in the population, default to 0.5.
`ss`	Boolean variable, default to FALSE. The TRUE value instruct the program to compute the sensitivity and the specificity of the classifier.
`pcorr`	Number of correlated features.
`chol.rho`	Cholesky decomposition of the covariance of the pcorr features that are correlated. It is assumed that the m important features are part of the pcorr correlated features.
`sampling.p`	The assumed proportion of group 1 samples in the training data; default of 0.5 assumes groups are equally represented regardless of p1.

Details

Refer to Sanchez, Wu, Song, Wang 2015, Section 3 and Supplementary materials.

Value

If ss=FALSE, the function returns the expected PCC. If ss=TRUE, the function returns a vector containing the expected PCC, sensitivity and specificity.

Author(s)

Meihua Wu <meihuawu@umich.edu> Brisa N. Sanchez <brisa@umich.edu> Peter X.K. Song <pxsong@umich.edu> Raymond Luu <raluu@umich.edu> Wen Wang <wangwen@umich.edu>

References

Sanchez, B.N., Wu, M., Song, P.X.K., and Wang W. (2016). "Study design in high-dimensional classification analysis." Biostatistics, in press.

Examples

```
## Sigma_1 in the paper
#first block is pcorr x pcorr of compound symmetry
#other diagonal block is Identity; off diagonal blocks are 0
pcorr=10
p=500
rho.cs=.8
#create first block
rho=diag(c((1-rho.cs)*rep(1,pcorr),rep(1,p-pcorr)))+ matrix(c(rho.cs*
rep(1,pcorr),rep(0,p-pcorr)),ncol=1) %*% c(rep(1,pcorr),rep(0,p-pcorr))
chol.rho1.500=chol(rho[1:pcorr,1:pcorr])
lmax= max(eigen(rho)$values)
print(lmax)
set.seed(1)
cv_method_corr(mu0=0.4,p=500,m=10,n=80,alpha_list=c(0.0000001,0.0001,0.01),
nrep=10,p1=0.6,ss=TRUE,pcorr=pcorr,chol.rho=chol.rho1.500,sampling.p=0.5)
#return 0.6689385 0.6806896 0.6513119
#alpha_list should be a dense grid of pvalue cut-offs;
#three values are used here for simplicity of the example
```

cv_method_MC — *MC simulation-based method to calculate the PCC of a CV-based classifier; calculated using training and test datasets in MC simulations.*

Description

Determine the probability of correct classification (PCC) for a high dimensional classification study employing Cross validation classifier. In contrast to the cv_method this function also generates a test dataset so that the estimated PCC does not rely on the normal approximation for the PCC formula.

Usage

```
cv_method_MC(mu0, p, m, n, alpha_list, nrep, p1 = 0.5, ss = F, ntest,
sampling.p=0.5)
```

Arguments

`mu0` The effect size of the important features.

`p` The number of the features in total.

`m` The number of the important features.

`n` The total sample size for the two groups, that would be used to develop the classifier.

`alpha_list` The search grid for the p-value threshold. The examples below use only three values for the sake of giving examples that run quickly but this should ideally be a dense grid,

`nrep` The number of simulation replicates employed to compute the expected PCC and/or sensitivity and specificity.

`p1` The prevalence of the group 1 in the population, default to 0.5.

`ss` Boolean variable, default to FALSE. The TRUE value instruct the program to compute the sensitivity and the specificity of the classifier.

`ntest` Sample size for the test dataset.

`sampling.p` The assumed proportion of group 1 samples in the training data; default of 0.5 assumes groups are equally represented regardless of p1.

Details

Refer to Sanchez, Wu, Song, Wang 2016, Section 2.2. This function was used to verify that a given sample size achieves the target PCC in Table 1 of the manuscript.

Value

If ss=FALSE, the function returns the expected PCC. If ss=TRUE, the function returns a vector containing the expected PCC, sensitivity and specificity.

Author(s)

Meihua Wu <meihuawu@umich.edu> Brisa N. Sanchez <brisa@umich.edu> Peter X.K. Song <pxsong@umich.edu> Raymond Luu <raluu@umich.edu> Wen Wang <wangwen@umich.edu>

References

Sanchez, B.N., Wu, M., Song, P.X.K., and Wang W. (2016). "Study design in high-dimensional classification analysis." Biostatistics, in press.

Examples

```
set.seed(1)
cv_method_MC(mu0=0.4,p=500,m=10,n=80,alpha_list=c(0.0000001,0.0001,0.01),
nrep=10,p1=0.6,ss=TRUE,ntest=100)
#return: 0.818 0.882 0.754
#alpha_list should be a dense list of p-value cutoffs;
#here we only use a few values to ease computation of the example.
```

`cv_method_MC_corr` *MC simulation-based method to calculate the PCC of a CV-based classifier when features are correlated; uses training and testing datasets.*

Description

Determine the probability of correct classification (PCC) for a high dimensional classification study employing Cross validation classifier. This is similar to cv_method_MC, but instead features generated are correlated.

Usage

```
cv_method_MC_corr(mu0, p, m, n, alpha_list, nrep, p1 = 0.5, ss = F, ntest,
pcorr, chol.rho,sampling.p=0.5)
```

Arguments

`mu0` The effect size of the important features.

`p` The number of the features in total.

`m` The number of the important features.

`n` The total sample size for the two groups.

`alpha_list` The search grid for the p-value threshold.

`nrep` The number of simulation replicates employed to compute the expected PCC and/or sensitivity and specificity.

`p1` The prevalence of the group 1 in the population, default to 0.5.

`ss` Boolean variable, default to FALSE. The TRUE value instruct the program to compute the sensitivity and the specificity of the classifier.

`ntest` Sample size for the test dataset.

`pcorr` Number of correlated features.

`chol.rho` Cholesky decomposition of the covariance of the pcorr features that are correlated. It is assumed that the m important features are part of the pcorr correlated features.

`sampling.p` The assumed proportion of group 1 samples in the training data; default of 0.5 assumes groups are equally represented regardless of p1.

Details

Refer to Sanchez, Wu, Song, Wang 2016, supplementary materials. This function is used to verify if a study using the sample sizes in Table 1 of the manuscript attains the PCC target via MC simulations.

Value

If ss=FALSE, the function returns the expected PCC. If ss=TRUE, the function returns a vector containing the expected PCC, sensitivity and specificity.

Author(s)

Meihua Wu <meihuawu@umich.edu> Brisa N. Sanchez <brisa@umich.edu> Peter X.K. Song <px-song@umich.edu> Raymond Luu <raluu@umich.edu> Wen Wang <wangwen@umich.edu>

References

Sanchez, B.N., Wu, M., Song, P.X.K., and Wang W. (2016). "Study design in high-dimensional classification analysis." Biostatistics, in press.

Examples

```
## Sigma_1 in the paper
#first block is pcorr x pcorr of compound symmetry
#other diagonal block is Identity; off diagonal blocks are 0

pcorr=10
p=500
rho.cs=.8

#create first block
rho=matrix(rep(0,p^2),nrow=p)
rho[1:pcorr,1:pcorr]=rho.cs
diag(rho)=rep(1,p)

chol.rho1.500=chol(rho[1:pcorr,1:pcorr])

set.seed(1)
cv_method_MC_corr(mu0=0.4,p=500,m=10,n=80,alpha_list=c(0.0000001,0.0001,0.01),
nrep=10,p1=0.6,ss=TRUE,ntest=100,pcorr=10,chol.rho=chol.rho1.500)
#return: 0.623 0.670 0.576
#alpha_list should be a dense list of p-value cutoffs;
#here we only use a few values to ease computation of the example.
```

ds_method *Estimate PCC by DS Method*

Description

Determine the probability of correct classification (PCC) for studies employing high dimensional features for classification; uses the method proposed by (Dobbin and Simon 2007) to choose the p-value threshold for feature selection.

Usage

```
ds_method(mu0, p, m, n, p1=0.5, lmax=1, ss=F, sampling.p)
```

Arguments

`mu0` The effect size of the important features.

`p` The number of the features in total.

`m` The number of the important features.

`n` The total sample size for the two groups.

p1 The prevalence of the group 1 in the population, default to 0.5.

lmax The maximum eigenvalue of the variance-covariance matrix of the p features. Defaults to 1 which implies that the features are assumed i.i.d.

ss Boolean variable, default to FALSE. The TRUE value instruct the program to compute the sensitivity and the specificity of the classifier.

sampling.p The assumed proportion of group 1 samples in the training data; default of 0.5 assumes groups are equally represented regardless of p1.

Details

Refer to Dobbin and Simon (2007)

Value

If ss=FALSE, the function returns the expected PCC. If ss=TRUE, the function returns a vector containing the expected PCC, sensitivity and specificity.

Author(s)

Meihua Wu <meihuawu@umich.edu> Brisa N. Sanchez <brisa@umich.edu> Peter X.K. Song <pxsong@umich.edu> Raymond Luu <raluu@umich.edu> Wen Wang <wangwen@umich.edu>

References

Dobbin, K.K., and Simon R.M. (2007). "Sample Size Planning for Developing Classifiers Using High-dimensional DNA Microarray Data." Biostatistics 8 (1): 101-117.

Examples

```
ds_method(mu0=0.6, p=500, m=10, n=38, p1=0.5, lmax=1, ss=TRUE)
#[1] 0.9252471 0.9252471 0.9252471
```

---

hct_beta *Alternative HCT Procedure to Choose P-Value Threshold Based on Beta Distribution of P-Values.*

Description

This procedure chooses the p-value threshold for feature selection in a similar fashion to hct_empirical. However, it is based on the Beta distribution of the p-values. Only the features whose p-values are less than the thresold will be included in the classifier.

Usage

```
hct_beta(pvalue, p, n)
```

Arguments

pvalue A vector containing the p*alpha_0 smallest p-values; typically alpha_0=0.10

p The number of the features in total.

n The total sample size for the two groups.

Details

Refer to Sanchez, et al (2016), Section 3 and supplementary materials.

Value

The p-value threshold for feature selection. Only the features whose p-values are less than the threshold will be included in the classifier.

Author(s)

Meihua Wu <meihuawu@umich.edu> Brisa N. Sanchez <brisa@umich.edu> Peter X.K. Song <pxsong@umich.edu> Raymond Luu <raluu@umich.edu> Wen Wang <wangwen@umich.edu>

References

Sanchez, B.N., Wu, M., Song, P.X.K., and Wang W. (2016). "Study design in high-dimensional classification analysis." Biostatistics, in press.

Examples

```
hct_beta(pvalue=0.10,p=500,n=80)
# 0.1
```

hct_empirical *Original HCT Procedure to Choose P-Value Threshold for Feature Selection*

Description

This is the original Higher Criticism Threshold (HCT) procedure (Donoho and Jin 2009) to choose p-value threshold for feature selection. Only the features whose p-values are less than the thresold will be included in the classifier.

Usage

```
hct_empirical(pvalue, p, n)
```

Arguments

pvalue	A vector containing the p*alpha_0 smallest p-values.
p	The number of the features in total.
n	The total sample size for the two groups.

Details

Refer to (Donoho and Jin 2009)

Value

The p-value threshold for feature selection. Only the features whose p-values are less than the thresold will be included in the classifier.

Author(s)

Meihua Wu <meihuawu@umich.edu> Brisa N. Sanchez <brisa@umich.edu> Peter X.K. Song <pxsong@umich.edu> Raymond Luu <raluu@umich.edu> Wen Wang <wangwen@umich.edu>

References

Donoho, D. and Jin, J. 2009. "Feature Selection by Higher Criticism Thresholding Achieves the Optimal Phase Diagram." Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences 367 (1906): 4449-4470.

Examples

```
hct_empirical(pvalue=0.10,p=500,n=80)
# 0.1
```

`hct_method` *Estimate PCC of HCT Classifiers*

Description

Determine the probability of correct classification (PCC) for studies employing high dimensional features for classification. It is assumed that a Higher Criticism Threshold (HCT) is used to choose the p-value threshold for feature selection and that features meeting the threshold are regarded as important for classification. A linear combination of important features is assumed to form the classification rule, with all important features having equal weight. In addition to the original HCT procedure by Donoho and Jin (2009), two more procedures to choose p-value threshold have developed and implemented. This function generates a fraction (alpha0) of the smallest p-values, calculates the threshold, examines which p-values meet the p-value threshold, and uses the normal CDF to estimate the PCC of the classifier. Neither training nor testing data are used. (See Sanchez et al 2016.)

Usage

```
hct_method(mu0, p, m, n, hct, alpha0, nrep, p1 = 0.5, ss = F, sampling.p=0.5)
```

Arguments

`mu0` The effect size of the important features.

`p` The number of the features in total.

`m` The number of the important features.

`n` The total sample size for the two groups.

`hct` The HCT procedure employed to choose the p-value threshold for feature selection. There are two valid choices (case sensitive): 1) hct_empirical, the HCT procedure originally proposed by (Donoho and Jin 2009); 2) hct_beta, an alternative HCT procedure which makes use of the beta distribution of the p-values under the null;

`alpha0` The proportion of the smallest p-values we will consider in the HCT algorithm, typically 0.1.

`nrep` The number of simulation replicates employed to compute the expected PCC and/or sensitivity and specificity.

`p1` The prevalence of the group 1 in the population, default to 0.5.

`ss` Boolean variable, default to FALSE. The TRUE value instruct the program to compute the sensitivity and the specificity of the classifier.

`sampling.p` The assumed proportion of group 1 samples in the training data; default of 0.5 assumes groups are equally represented regardless of p1.

Value

If ss=FALSE, the function returns the expected PCC. If ss=TRUE, the function returns a vector containing the expected PCC, sensitivity and specificity.

Author(s)

Meihua Wu <meihuawu@umich.edu> Brisa N. Sanchez <brisa@umich.edu> Peter X.K. Song <pxsong@umich.edu> Raymond Luu <raluu@umich.edu> Wen Wang <wangwen@umich.edu>

References

Donoho, D, and Jin, J. (2009). "Feature Selection by Higher Criticism Thresholding Achieves the Optimal Phase Diagram." Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences 367 (1906): 4449-4470.

Sanchez, B.N., Wu, M., Song, P.X.K., and Wang W. (2016). "Study design in high-dimensional classification analysis." Biostatistics, in press.

Examples

```
set.seed(1)
hct_method(mu0=0.4, p=500, m=10, n=80, hct=hct_beta, alpha0=0.5, nrep=10,
p1 = 0.5, ss = TRUE)
#return: 0.807098 0.807098 0.807098
```

hct_method_corr *Estimate PCC of HCT Classifiers via implementation of Monte Carlo simulations with correlated features.*

Description

Determine the probability of correct classification (PCC) for studies employing high dimensional features for classification. Higher Criticisms Threshold (HCT) classifier is used to choose the p-value threshold for feature selection. In addition to the original HCT procedure by (Donoho and Jin 2009), two more procedures to choose p-value threshold have developed and implemented.

Usage

```
hct_method_corr(mu0, p, m, n, hct, alpha0, nrep, p1 = 0.5,
ss = F, pcorr, chol.rho, sampling.p=0.5)
```

Arguments

mu0	The effect size of the important features.
p	The number of the features in total.
m	The number of the important features.
n	The total sample size for the two groups.
hct	The HCT procedure employed to choose the p-value threshold for feature selection. There are two valid choices (case sensitive): 1) hct_empirical, the HCT procedure originally proposed by (Donoho and Jin 2009); 2) hct_beta, an alternative HCT procedure which makes use of the beta distribution of the p-values under the null
alpha0	The proportion of the smallest p-values we will consider in the HCT algorithm.
nrep	The number of simulation replicates employed to compute the expected PCC and/or sensitivity and specificity.
p1	The prevalence of the group 1 in the population, default to 0.5.
ss	Boolean variable, default to FALSE. The TRUE value instruct the program to compute the sensitivity and the specificity of the classifier.
pcorr	Number of correlated features.
chol.rho	Cholesky decomposition of the covariance of the pcorr features that are correlated. It is assumed that the m important features are part of the pcorr correlated features.
sampling.p	The assumed proportion of group 1 samples in the training data; default of 0.5 assumes groups are equally represented regardless of p1.

Value

If ss=FALSE, the function returns the expected PCC. If ss=TRUE, the function returns a vector containing the expected PCC, sensitivity and specificity.

Author(s)

Meihua Wu <meihuawu@umich.edu> Brisa N. Sanchez <brisa@umich.edu> Peter X.K. Song <pxsong@umich.edu> Raymond Luu <raluu@umich.edu> Wen Wang <wangwen@umich.edu>

References

Donoho, D., and Jin, J. (2009). "Feature Selection by Higher Criticism Thresholding Achieves the Optimal Phase Diagram." Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences 367 (1906) (November 13): 4449-4470.

Examples

```
## Sigma_1 in the paper
#first block is pcorr x pcorr of compound symmetry
#other diagonal block is Identity; off diagonal blocks are 0
pcorr=10
p=500
rho.cs=.8
#create first block
rho= diag(c((1-rho.cs)*rep(1,pcorr),rep(1,p-pcorr)))+ matrix(c(rho.cs*
rep(1,pcorr),rep(0,p-pcorr)), ncol=1) %*% c(rep(1,pcorr),rep(0,p-pcorr))
```

```
chol.rho1.500=chol(rho[1:pcorr,1:pcorr])
set.seed(1)
hct_method_corr(mu0=0.4,p=500,m=10,n=80,hct=hct_beta,alpha0=0.5,nrep=10,
p1=0.5,ss=TRUE,pcorr=pcorr,chol.rho=chol.rho1.500)
#return: 0.6672256 0.6672256 0.6672256
```

hct_method_MC *Estimate PCC of HCT Classifiers via implementation of Monte Carlo simulations, using training and testing datasets*

Description

Determine the probability of correct classification (PCC) for studies employing high dimensional features for classification. It is assumed that a Higher Criticism Threshold (HCT) is used to choose the p-value threshold for feature selection and that features meeting the threshold are important for classification. In addition to the original HCT procedure by Donoho and Jin (2009), two procedures to choose p-value threshold have been implemented (See hct_empirical and hct_beta). This function is similar to hct_method but this does not rely on the normal CDF to approximate the PCC. Instead training and testing datasets are generated at each iteration of the algorithm.

Usage

```
hct_method_MC(mu0, p, m, n, hct, alpha0, nrep, p1=0.5, ss=F, ntest,
sampling.p)
```

Arguments

mu0	The effect size of the important features.
p	The number of the features in total.
m	The number of the important features.
n	The total sample size for the two groups.
hct	The HCT procedure employed to choose the p-value threshold for feature selection. There are two valid choices (case sensitive): 1) hct_empirical, the HCT procedure originally proposed by Donoho and Jin (2009); 2) hct_beta, an alternative HCT procedure which makes use of the beta distribution of the p-values under the null;
alpha0	The proportion of the smallest p-values we will consider in the HCT algorithm.
nrep	The number of simulation replicates employed to compute the expected PCC and/or sensitivity and specificity.
p1	The prevalence of the group 1 in the population, default to 0.5.
ss	Boolean variable, default to FALSE. The TRUE value instruct the program to compute the sensitivity and the specificity of the classifier.
ntest	Sample size for the test dataset.
sampling.p	The assumed proportion of group 1 samples in the training data; default of 0.5 assumes groups are equally represented regardless of p1.

Value

If ss=FALSE, the function returns the expected PCC. If ss=TRUE, the function returns a vector containing the expected PCC, sensitivity and specificity.

Author(s)

Meihua Wu <meihuawu@umich.edu> Brisa N. Sanchez <brisa@umich.edu> Peter X.K. Song <pxsong@umich.edu> Raymond Luu <raluu@umich.edu> Wen Wang <wangwen@umich.edu>

References

Donoho, D., and Jin J. (2009). "Feature Selection by Higher Criticism Thresholding Achieves the Optimal Phase Diagram." Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences 367 (1906) (November 13): 4449-4470.

Sanchez, B.N., Wu, M., Song, P.X.K., and Wang W. (2016). "Study design in high-dimensional classification analysis." Biostatistics, in press.

Examples

```
set.seed(1)
hct_method_MC(mu0=0.4,p=500,m=10,n=80,hct=hct_beta,alpha0=0.5,nrep=10,p1=0.5,
ss=TRUE,ntest=100,sampling.p=0.5)
#return: 0.801 0.806 0.796
```

hct_method_MC_corr *Estimate PCC of HCT Classifiers constructed with correlated features using Monte Carlo simulations*

Description

Determine the probability of correct classification (PCC) for studies employing high dimensional features for classification. It is assumed that a Higher Criticism Threshold (HCT) is used to choose the p-value threshold for feature selection and that features meeting the threshold are important for classification. In addition to the original HCT procedure by (Donoho and Jin 2009), two procedures to choose p-value threshold have been implemented (See hct_empirical and hct_beta). This function is similar to hct_method_corr but this does not rely on the normal CDF to approximate the PCC. Instead training and testing datasets are generated at each iteration of the algorithm.

Usage

```
hct_method_MC_corr(mu0, p, m, n, hct, alpha0, nrep, p1=0.5,
ss=FALSE, ntest, pcorr, chol.rho,sampling.p)
```

Arguments

mu0	The effect size of the important features.
p	The number of the features in total.
m	The number of the important features.
n	The total sample size for the two groups.
hct	The HCT procedure employed to choose the p-value threshold for feature selection. There are two valid choices (case sensitive): 1) hct_empirical, the HCT procedure originally proposed by (Donoho and Jin 2009); 2) hct_beta, an alternative HCT procedure which makes use of the beta distribution of the p-values under the null;

`alpha0`	The proportion of the smallest p-values we will consider in the HCT algorithm.
`nrep`	The number of simulation replicates employed to compute the expected PCC and/or sensitivity and specificity.
`p1`	The prevalence of the group 1 in the population, default to 0.5.
`ss`	Boolean variable, default to FALSE. The TRUE value instruct the program to compute the sensitivity and the specificity of the classifier.
`ntest`	Sample size for the test dataset.
`pcorr`	Number of correlated features.
`chol.rho`	Cholesky decomposition of the covariance of the pcorr features that are correlated. It is assumed that the m important features are part of the pcorr correlated features.
`sampling.p`	The assumed proportion of group 1 samples in the training data; default of 0.5 assumes groups are equally represented regardless of p1.

Value

If ss=FALSE, the function returns the expected PCC. If ss=TRUE, the function returns a vector containing the expected PCC, sensitivity and specificity.

Author(s)

Meihua Wu <meihuawu@umich.edu> Brisa N. Sanchez <brisa@umich.edu> Peter X.K. Song <pxsong@umich.edu> Raymond Luu <raluu@umich.edu> Wen Wang <wangwen@umich.edu>

References

Donoho, D., and Jin J. (2009). "Feature Selection by Higher Criticism Thresholding Achieves the Optimal Phase Diagram." Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences 367 (1906) (November 13): 4449-4470.

Sanchez, B.N., Wu, M., Song, P.X.K., and Wang W. (2016). "Study design in high-dimensional classification analysis." Biostatistics, in press.

Examples

```
## Sigma_1 in the paper
#first block is pcorr x pcorr of compound symmetry
#other diagonal block is Identity; off diagonal blocks are 0
pcorr=10
p=500
rho.cs=.8
#create first block
rho=matrix(rep(0,p^2),nrow=p)
rho[1:pcorr,1:pcorr]=rho.cs
diag(rho)=rep(1,p)
chol.rho1.500=chol(rho[1:pcorr,1:pcorr])
set.seed(1)
hct_method_MC_corr(mu0=0.4, p=500, m=10, n=80, hct=hct_beta, alpha0=0.5, nrep=10,
p1 = 0.5, ss=TRUE, ntest=100, pcorr=10, chol.rho=chol.rho1.500,sampling.p=0.5)
#return: 0.673 0.686 0.660
```

`ideal_pcc` *Determine the Ideal PCC*

Description

Determine the probability of correct classification (PCC) for a study employing the ideal classifier. The ideal classifier is constructed assuming we know exactly the important features and their effect size. The ideal PCC is the uppper bound of the PCC of any linear classifiers.

Usage

```
ideal_pcc(mu0, m, p1 = 0.5)
```

Arguments

`mu0`	The effect size of the important features.
`m`	The number of the important features.
`p1`	The prevalence of the group 1 in the population, default to 0.5.

Value

The PCC of the ideal classifier.

Author(s)

Meihua Wu <meihuawu@umich.edu> Brisa N. Sanchez <brisa@umich.edu> Peter X.K. Song <pxsong@umich.edu> Raymond Luu <raluu@umich.edu> Wen Wang <wangwen@umich.edu>

References

Dobbin, Kevin K., and Richard M. Simon. 2007. "Sample Size Planning for Developing Classifiers Using High-dimensional DNA Microarray Data." Biostatistics 8 (1) (January 1): 101-117.

Examples

```
ideal_pcc(mu0=0.4, m=10, p1 = 0.6)
#return: 0.8999055
```

`samplesize` *Determine the Sample Size Requirement*

Description

Determine the sample size to achieve the target probability of correct classification (PCC) using various design methods. Sample sizes are chosen using a binary search algorithm between the range nmin to nmax.

Usage

```
samplesize(target, nmin, nmax, f, ...)
```

Arguments

`target`	Set the target probability of correct classifcation (PCC) for the study.
`nmin`	The mimimum sample size for both groups combined. Typically 0.05 smaller than the ideal PCC. It must be an even number.
`nmax`	The maximum sample size for both groups combined. So it must be an even number.
`f`	Specify the PCC estimation function: ds_method, cv_method, or hct_method
`...`	The design assumptions and other arguments for the PCC estimation function, f.

Value

The smallest sample size that achieves the target PCC.

Author(s)

Meihua Wu <meihuawu@umich.edu> Brisa N. Sanchez <brisa@umich.edu> Peter X.K. Song <pxsong@umich.edu> Raymond Luu <raluu@umich.edu> Wen Wang <wangwen@umich.edu>

Examples

```
set.seed(1)
samplesize(target=0.8, nmin=20, nmax=100, hct_method, mu0=0.4, p=500,
m=10, hct=hct_beta, alpha0=0.5, nrep=100)
#return: 78.0000000  0.8043205
```

`which.region` *Determine the Feasibility Region*

Description

Given the design assumption, determine which feasiblity region the design problem belongs to. The feasibility region is constructed from the asymptotic properties of the rare-and-weak model (Donoho and Jin 2009). The two groups are assumed to be equally proportioned, i.e. p_+1=p_-1=0.5. If the the problem is feasible, then the probability of correct classification (PCC) of the HCT classifer will approach 1 when the number of features goes to infinity. If the the problem is unfeasible, then the probability of correct classification (PCC) of any linear classifer will approach 0.5 when the number of features goes to infinity.

Usage

```
which.region(mu0, p, m, n)
```

Arguments

`mu0`	The effect size of the important features.
`p`	The number of the features in total.
`m`	The number of the important features.
`n`	The total sample size for the two groups.

Value

`0` The classification problem belongs to the unfeasible region.

`1` The classification problem belongs to the feasible region.

`2` The classification problem belongs to the feasible region.

`3` The classification problem belongs to the feasible region.

`4` The classification problem belongs to the feasible region.

Region 1-4 are all feasible regions. Their difference is discussed in more details in ().

Author(s)

Meihua Wu <meihuawu@umich.edu> Brisa N. Sanchez <brisa@umich.edu> Peter X.K. Song <pxsong@umich.edu> Raymond Luu <raluu@umich.edu> Wen Wang <wangwen@umich.edu>

References

Donoho, David, and Jiashun Jin. 2009. "Feature Selection by Higher Criticism Thresholding Achieves the Optimal Phase Diagram." Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences 367 (1906) (November 13): 4449-4470.

Examples

```
which.region(mu0=0.4, p=500, m=10, n=80)
#return: 4
```