Macro TDRDAS_SEPCLASS User's Guide
30 December 2019

TDRDA Set Methods

True discovery rate degree of association (TDRDA) set analysis is a method to identify sets of predictors for which the true degree of association with clinical outcome or state is more than a specified value while controlling the false discovery rate (Benjamini and Hochberg, 1995). The methodology is described in detail by Crager (2010).

The method starts with a set of predictors, each having (1) an estimated degree of association with clinical outcome or state and (2) a standard error of the estimate. These might be, for example, a standardized log hazard ratio and its standard error from a Cox proportional hazards regression analysis, or a standardized log odds ratio estimate and its standard error from a logistic regression analysis. Given the estimates and standard errors, and a specified true discovery rate $1-\gamma$ (for example, $1-\gamma = 0.80$ or 0.90) the method identifies set of predictors among which $100(1-\gamma)\%$ can be expected to have true absolute degree of association higher than a given value θ. Then, letting θ vary over a grid of values, the method calculates for each predictor a maximum lower bound (MLB) absolute degree of association for which the predictor belongs to a TDRDA set with true discovery rate $1-\gamma$.

We are apt to focus further research on the predictors with the strongest associations. Because of regression to the mean, the absolute degrees of association that these predictors show in subsequent research is likely to be less than the "naïve" estimate of the degree of association in the current study. To get a realistic estimate of the degree of association we could expect for each predictor in subsequent research, the TDRDA set method computes an estimated degree of association corrected for regression to the mean (RM).

The MLB absolute degree of association and the RM-corrected estimate can be displayed together in a concise bar-chart summary of the analysis results. The graph includes all predictors for which the false discovery rate for Wald tests of non-zero degree of

association is γ or less, since all TDRDA sets with true discovery rate $1-\gamma$ are refinements of this set.

Separate Class Analysis

Analyses that control the false discovery rate lose power to identify even predictors with strong association when the proportion of predictors with little or no association increases. To avoid this loss of power, we might consider dividing a very large predictor set up into smaller predictor sets and analyzing each set separately, ignoring the others. However, this approach has the disadvantage that for small predictor sets, the asymptotics and efficiencies of false discovery rate analyses that derive from having a large number of predictors may be lost.

Efron (2008) developed an alternative strategy called "separate class" analysis. Efron divides the predictor set into mutually exclusive classes and calculates a false discovery rate within each class, but he utilizes information from all of the predictors in this calculation. This preserves the asymptotics of the FDR calculation while allowing the data analyst to concentrate on specific, smaller classes of interest.

Separate Class TDRDA Set Analysis

Separate class analysis methods can be applied to TDRDA set analysis. The false discovery rate calculations used to establish the TDRDA sets and maximum lower bounds are made using the separate class methodology. Correction for regression to the mean is made within each predictor class.

Since the inference is by predictor class, it makes sense to present the TDRDA set bar chart sorted by class. Efron (2008) shows that if we identify predictors in each class using separate class methods and then combine the identified genes into a single list, the false discovery rate in the combined list is asymptotically equal to the common false discovery rate used in each class. Therefore, it is also reasonable to present the TDRDA set bar chart without sorting by gene class.

The details of the separate class TDRDA set analysis calculation are described in Crager and Ahmed (2014).

Macro TDRDAS_SEPCLASS

The macro TDRDAS carries out the calculations for the separate class TDRDA set analysis. The macro takes as input a data set containing a gene identifier, a gene class identifier, the estimated degree of association and its standard error, with one record per predictor. The macro produces an output data set containing the MLB absolute degree of association and RM-corrected estimate of the degree of association for each gene. At the user's discretion, the output data set may contain transformed values of the MLB and RM-corrected estimates. For example, if the input estimates are log hazard ratios, the user may wish to use the exponential transform to express the MLB and RM-corrected estimates as hazard ratios. The macro also produces a TDRDA set bar chart showing the (transformed) MLB and RM-corrected estimates. At the user's discretion, the bar chart and output data set may be sorted (1) by class, then by MLB and RM-corrected estimate, or (2) simply by MLB and RM-corrected estimate.

The macro is called as follows:

```
%macro tdrdas_sepclass(
  /* Input Specification*/   indsn=, predictorname=, class=,
                             estimate=, stderr=, estzero=,
  /* Analysis Parameters */  accuracy=,oneminusq=,lambda=,
  /* Output Specification */ outdsn=, transformx=,
                             MLBvar=, RMCEstvar=,
   /* Graph Options*/        refinterval=, measure=, graphlabel=,
                             goutpath=, graphname=, maxgenesppg=,
                             graphcolor1=, graphcolor2=,sortbyclass=,
                             class_value_graphname=
              );
```

The macro parameters are described in Table 1.

Table 1. Macro TDRDAS_SEPCLASS Parameters

Parameter	Type	Required?	Default Value	Description
indsn	$	Yes	—	Input data set name
precitorname	$	Yes	—	Name of predictor
class	$/#	Yes	—	Predictor class identifier. This variable may be character or numeric.
estimate	#	Yes	—	Input data set variable that contains the estimated degrees of association
stderr	#	Yes	—	Input data set variable that contains the standard errors of the estimated degrees of association
estzero	#	Yes	—	Value of the association estimate that corresponds to no association. If the association is measured by log hazard ratio or log odds ratio, then estzero = 0.
accuracy	#	No	0.001	Accuracy to which MLB degree of association is calculated.
oneminusq	#	No	0.9	True discovery rate (equal to 1 minus the acceptable false discovery rate).
lambda	#	No	0.5	The value of the tuning parameter λ in Storey's method for estimating the proportion of true null hypotheses.
outdsn	$	Yes	—	Name of the output data set.
transformx	$	No	exp(x)	Optional transformation to be applied to the MLB absolute degrees of association and RM-corrected estimates. The parameter must include an "(x)" with no spaces between x and each parenthesis. Every time x appears in the function, it must be encased in parentheses.
MLBvar	#	No	MLB	Name of the output data set variable that will contain the (possibly transformed) MLB absolute degree of association for each gene.

Table 1. Macro TDRDAS_SEPCLASS Parameters

Parameter	Type	Required?	Default Value	Description
RMCEstvar	#	No	RMCest	Name of the output data set variable that will contain the (possibly transformed) RM-corrected estimates of the degree of association for each gene.
refinterval	#	No	0.05	Spacing of reference lines on the graph. Used in the axis statement as in "axis1 order=(1 to 1.75 by &refinterval.)"
measure	$	No	Degree of Association	The measure of the degree of association. This is used in the legend for the graph and in variable labels. If you change this, you will also need to change the value for graphlabel. The value selected for this parameter is case sensitive.
graphlabel	$	No	Standardized Degree of Association for Predictor with Outcome	This parameter allows the user to modify the label on the graph.
goutpath	$	No	–	The goutpath parameter can be used to redirect the graph to a folder other than the one in which the calling SAS program is running.
graphname	$	No	–	Name of the file that will contain the TDRDAS bar chart. **Note:** Do not add &timestamp or the file extension; the macro will add them.
maxgenesppg	#	No	150	The maximum number of genes to be shown on each page of the graph. The value selected should be between 1 and 150.
graphcolor1	$	No	black	The color of the bars on the graph representing the genes with positive association.
graphcolor2	$	No	gray	The color of the bars on the graph representing the genes with negative association.

Table 1. Macro TDRDAS_SEPCLASS Parameters

Parameter	Type	Required?	Default Value	Description
sortbyclass	$	No	yes	If this parameter is set to yes, the TDRDA set bar chart chart and output data set are sorted first by class, then by MLB and RM-corrected estimate. If no, chart and output data set are sorted only by MLB and RM-corrected estimate.
class_value_graphname	$	No	–	Optional filename of graph of logistic regression estimates of class probabilities given z-score values. These plots can be useful to make sure the cubic spline fit is reasonable, and to show which classes are enriched for truly associated genes. Also, the slope of the fitted lines should be close to zero around z=0. A substantial departure from this indicates that the null distribution of the test statistic is not the same across classes, which suggests a possible study design or conduct issue. If this parameter is not specified, no graph is produced.

References

Benjamini Y, Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**:289–300.

Crager MR, Ahmed M (2014). Separate Class True Discovery Rate Degree of Association Sets for Biomarker Identification. *Journal of Biopharmaceutical Statistics* **24**:1022-1034.

Crager MR. (2010). Gene identification using true discovery rate of association degree of association sets and estimates corrected for regression to the mean. *Statistics in Medicine* **29**:33–45. DOI 10.1002/sim.3789.

Efron B (2008). Simultaneous inference: When should hypothesis testing problems be combined? *Annals of Applied Statistics* **2**:197–223.