Package: tightenBlock (via r-universe)

September 11, 2024

Type Package

Title Tightens an Observational Block Design by Balanced Subset Matching

Version 0.1.7

Author Paul Rosenbaum [aut, cre]

Maintainer Paul Rosenbaum <rosenbaum@wharton.upenn.edu>

Description Tightens an observational block design into a smaller design with either smaller or fewer blocks while controlling for covariates. The method uses fine balance, optimal subset matching (Rosenbaum, 2012 <doi:10.1198/jcgs.2011.09219>) and two-criteria matching (Zhang et al 2023 <doi:10.1080/01621459.2021.1981337>). The main function is tighten(). The suggested 'rrelaxiv' package for solving minimum cost flow problems: (i) derives from Bertsekas and Tseng (1988) <doi:10.1007/BF02288322>, (ii) is not available on CRAN due to its academic license, (iii) may be downloaded from GitHub at <https://github.com/josherrickson/rrelaxiv/>, (iv) is not essential to use the package.

License GPL-2

Encoding UTF-8

LazyData true

Imports stats, MASS, rcbalance

Suggests rrelaxiv

Additional_repositories https://errickson.net/rrelaxiv/

Depends R (>= 3.5.0)

NeedsCompilation no

Repository CRAN

Date/Publication 2023-12-15 08:00:02 UTC

Contents

`tightenBlock-package` *Tightens an Observational Block Design by Balanced Subset Matching*

Description

Tightens an observational block design into a smaller design with either smaller or fewer blocks while controlling for covariates. The method uses fine balance, optimal subset matching (Rosenbaum, 2012 <doi:10.1198/jcgs.2011.09219>) and two-criteria matching (Zhang et al 2023 <doi:10.1080/01621459.2021.198
The main function is tighten(). The suggested 'rrelaxiv' package for solving minimum cost flow problems: (i) derives from Bertsekas and Tseng (1988) <doi:10.1007/BF02288322>, (ii) is not available on CRAN due to its academic license, (iii) may be downloaded from GitHub at <https://github.com/josherrickson/rre
(iv) is not essential to use the package.

Details

The DESCRIPTION file:

Package:	tightenBlock
Type:	Package
Title:	Tightens an Observational Block Design by Balanced Subset Matching
Version:	0.1.7
Authors@R:	c(person("Paul", "Rosenbaum", role = c("aut", "cre"), email = "rosenbaum@wharton.upenn.edu"))
Author:	Paul Rosenbaum [aut, cre]
Maintainer:	Paul Rosenbaum <rosenbaum@wharton.upenn.edu>
Description:	Tightens an observational block design into a smaller design with either smaller or fewer blocks wh
License:	GPL-2
Encoding:	UTF-8
LazyData:	true
Imports:	stats, MASS, rcbalance
Suggests:	rrelaxiv
Additional_repositories:	https://errickson.net/rrelaxiv/
Depends:	R (>= 3.5.0)

Index of help topics:

aHDLt	Alcohol and HDL Cholesterol
addMahal	Rank-Based Mahalanobis Distance Matrix
addNearExact	Add a Near-exact Penalty to an Exisiting Distance Matrix.
makematch	Make a Match Using Two Criteria Matching with Optimal Subset Matching
makenetwork	Make the Network Used for Matching with Two Criteria
startcost	Initialize a Distance Matrix.
tighten	Tightening an Observational Block Design
tightenBlock-package	Tightens an Observational Block Design by Balanced Subset Matching

Author(s)

Paul Rosenbaum [aut, cre]

Maintainer: Paul Rosenbaum <rosenbaum@wharton.upenn.edu>

References

Bertsekas, D. P., Tseng, P. (1988) <doi:10.1007/BF02288322> The relax codes for linear minimum cost network flow problems. Annals of Operations Research, 13, 125-190.

Rosenbaum, P. R., Ross, R. N. and Silber, J. H. (2007) <10.1198/016214506000001059> Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. Journal of the American Statistical Association, 102(477), 75-83.

Rosenbaum, P. R. (2012) <doi:10.1198/jcgs.2011.09219> Optimal matching of an optimally chosen subset in observational studies. Journal of Computational and Graphical Statistics, 21(1), 57-71.

Rosenbaum, P. R. (2024) Tightening an observational block design to form an optimally balanced subdesign. Manuscript.

Zhang, B., D. S. Small, K. B. Lasater, M. McHugh, J. H. Silber, and P. R. Rosenbaum (2023) <doi:10.1080/01621459.2021.1981337> Matching one sample according to two criteria in observational studies. Journal of the American Statistical Association, 118, 1140-1151.

Examples

```
data(aHDLt)
result<-tighten(aHDLt,aHDLt$z,aHDLt$block,
       x=cbind(aHDLt$age,aHDLt$education),
       f=cbind(aHDLt$ibmi,(aHDLt$bmi>22.5)+(aHDLt$bmi>27.5)+(aHDLt$bmi>32.5)),
                ncontrols=2)
```

addMahal *Rank-Based Mahalanobis Distance Matrix*

Description

Adds a rank-based Mahalanobis distance to an exisiting distance matrix.

Usage

```
addMahal(costmatrix, z, X)
```

Arguments

costmatrix	An existing cost matrix with sum(z) rows and sum(1-z) columns. The function checks the compatability of costmatrix, z and X; so, it may stop with an error if these are not of appropriate dimensions. In particular, costmatrix may come from startcost().
z	A vector with z[i]=1 if individual i is treated or z[i]=0 if individual i is control. The rows of costmatrix refer to treated individuals and the columns refer to controls.
X	A matrix with length(z) rows containing covariates or a vector with length(z) containing a single covariate.

Details

The rank-based Mahalanobis distance is defined in section 9.3 of Rosenbaum (2020). Individual covariates are replaced by their ranks before computing the Mahalanobis distance. Even when ties are present, the untied variance of ranks is used. These adjustments improve the distance when: (i) a covariate contains an extreme outlier, causing its variance to increase, thereby causing the distance to ignore large differences in that covariate, and (ii) rare binary covariates that either match or mismatch by 1, and would have very small variances if the tied variance were used.

Value

A new distance matrix that is the sum of costmatrix and the rank-based Mahalanobis distances.

Author(s)

Paul R. Rosenbaum

References

Rosenbaum, P. R. (2020) <doi:10.1007/978-3-030-46405-9> Design of Observational Studies (2nd Edition). New York: Springer.

Rubin, D. B. (1980) <doi:10.2307/2529981> Bias reduction using Mahalanobis-metric matching. Biometrics, 36, 293-298.

Examples

```
data(aHDLt)

# names and corresponding rownames help when viewing output
rownames(aHDLt)<-aHDLt$SEQN
z<-aHDLt$z
names(z)<-aHDLt$SEQN

#
# First 12 people
aHDLt[1:12,]
#
# Start with a zero distance matrix.
dist<-startcost(z)
dist[1:3,1:9]
#
# Add Mahalanobis distances to the zero distance matrix.
dist<-addMahal(dist,z,cbind(aHDLt$age,aHDLt$education,aHDLt$female))
round(dist[1:3,1:9],2)

dim(dist)
sum(z)
sum(1-z)
```

`addNearExact` *Add a Near-exact Penalty to an Exisiting Distance Matrix.*

Description

Add a Near-exact Penalty to an Exisiting Distance Matrix.

Usage

```
addNearExact(costmatrix, z, exact, penalty = 1000)
```

Arguments

`costmatrix`	An existing cost matrix with sum(z) rows and sum(1-z) columns. The function checks the compatability of costmatrix, z and exact; so, it may stop with an error if these are not of appropriate dimensions. In particular, costmatrix may come from startcost().
`z`	A vector with z[i]=1 if individual i is treated or z[i]=0 if individual i is control. The rows of costmatrix refer to treated individuals and the columns refer to controls.
`exact`	A vector with the same length as z. Typically, exact represent a nominal covariate. Typically, exact is a vector whose coordinates take a small or moderate number of values.
`penalty`	One positive number.

Details

If the ith treated individual and the jth control have different values of exact, then the distance between them in costmatrix is increased by adding penalty.

Value

A penalized distance matrix.

Note

In two-criteria matching, addNearExact has a different effect when used on the left side of the network than on the right side. On the left side, it implements near-exact matching, but on the right side it implements fine or near-fine balance. Details follow.

On the left, a sufficiently large penalty will maximize the number of individuals exactly matched for exact. A smaller penalty will tend to increase the number of individuals matched exactly, without prioritizing one covariate over all others.

On the right, a sufficiently large penalty will strive for fine balance, and if that is infeasible, it will achieve near-fine balance. A smaller penalty will tend to increase balance, without prioritizing one covariate over all others.

In effect, Zubizarreta et al. (2011) seek near-fine, near-exact matching for the same covariate, essentially by placing it on both the left and the right, with a much smaller penalty on the left. Strive to balance the covariate, and if you can also pair for it, then so much the better.

If the left distance matrix is penalized, it will affect pairing and balance; however, if the right distance matrix is penalized it will affect balance only.

Adding several near-exact penalties for different covariates on the right distance matrix implements a Hamming distance on the joint distribution of those covariates, as discussed in Zhang et al. (2023). The tighten() function has the Hamming distance as an option and the example illustrates its use.

Near-exact matching for a nominal covariate is discussed and contrasted with exact matching in Sections 10.3 and 10.4 of Rosenbaum (2020). Near-exact matching is always feasible, because it implements a constraint using a penalty. Exact matching may be infeasible, but when feasible it may be used to speed up computations. For an alternative method of speeding computations, see Yu et al. (2020) who identify feasible constraints very quickly prior to matching with those constraints.

Author(s)

Paul R. Rosenbaum

References

Rosenbaum, P. R. (2020) <doi:10.1007/978-3-030-46405-9> Design of Observational Studies (2nd Edition). New York: Springer.

Yang, D., Small, D. S., Silber, J. H. and Rosenbaum, P. R. (2012) <doi:10.1111/j.1541-0420.2011.01691.x> Optimal matching with minimal deviation from fine balance in a study of obesity and surgical outcomes. Biometrics, 68, 628-636. (Extension of fine balance useful when fine balance is infeasible. Comes as close as possible to fine balance. Implemented in makematch() by placing a large near-exact penalty on a nominal/integer covariate x1 on the right distance matrix.)

Yu, R., Silber, J. H., Rosenbaum, P. R. (2020) <doi:10.1214/19-STS699> Matching methods for observational studies derived from large administrative databases. Statistical Science, 35, 338-355.

Zhang, B., D. S. Small, K. B. Lasater, M. McHugh, J. H. Silber, and P. R. Rosenbaum (2023) <doi:10.1080/01621459.2021.1981337> Matching one sample according to two criteria in observational studies. Journal of the American Statistical Association, 118, 1140-1151.

Zubizarreta, J. R., Reinke, C. E., Kelz, R. R., Silber, J. H. and Rosenbaum, P. R. (2011) <doi:10.1198/tas.2011.11072> Matching for several sparse nominal variables in a case control study of readmission following surgery. The American Statistician, 65(4), 229-238.

Examples

```
data(aHDLt)
rownames(aHDLt)<-aHDLt$SEQN
z<-aHDLt$z
names(z)<-aHDLt$SEQN
aHDLt[1:12,]
dist<-startcost(z)
dist<-addNearExact(dist,z,aHDLt$education)
dist[1:3,1:9]
```

---

aHDLt *Alcohol and HDL Cholesterol*

Description

A small observational study of light daily alcohol consumption and HDL cholesterol – so-called good cholesterol – derived from NHANES 2013-2014 and 2015-2016. There are 406 blocks of four individuals, making 1624 individuals in total. Blocks were matched for age, female and education in five ordered categories.

Usage

```
data("aHDLt")
```

Format

A data frame with 1624 observations on the following 14 variables.

`SEQN` NHANES ID number

`nh` NHANES 2013-2014 is 1314, and NHANES 2015-2016 is 1516

`age` Age in years

`female` 1=female, 0=male

`education` 1 is <9th grade, 3 is high school, 5 is a BA degree

`z` 1=light almost daily alcohol, 0=little or no alcohol last year.

`grp` Treated group and control groups. Daily=light almost daily alcohol, Never=fewer than 12 drinks during entire life, Rarely=more than 12 drinks in life, but fewer than 12 in the past year, and never had a period of daily binge drinking, PastBinge = a past history of binge drinking on most days, but currently drinks once a week or less. For details, see Rosenbaum (2023, Appendix).

`grpL` Short labels for plotting formed as the first letters of grp. `D < N < R < B`

`hdl` HDL cholesterol level mg/dL

`mmercury` Methylmercury level ug/L

`dentist1Y` 1 = been to the dentist in the past year, 0 = all other categories. Category 0 includes a very small number of people who did not know, refused to answer, or were otherwise missing.

`bmi` BMI or body mass index. A measure of obesity.

`ibmi` Four standard categories of BMI. 0 is BMI<25 (normal), 1 is BMI in [25,30) (overweight), 2 is BMI in [30,35) (obese), 3 is BMI >= 35 (morbidly obese).

`block` Block indicator, 1, 2, ..., 406. The 1624 observations are in 406 blocks, each of size 4.

Details

The data are from the US National Health and Nutrition Examination Survey. For extensive details, see the data appendix in Rosenbaum (2023).

There is a debate about whether light daily alcohol consumption – a single glass of red wine – shortens or lengthens life. LoConte et al. (2018) emphasize the important fact that alcohol is a carcinogen. Suh et al. (1992) claim reduced cardiovascular mortality brought about by an increase in high density high-density lipoprotein (HDL) cholesterol, the so-called good cholesterol. There is on-going debate about whether there are cardiovascular benefits, and if they exist, whether they are large enough to offset an increased risk of cancer. This example looks at a small corner of the larger debate, namely the effect on HDL cholesterol.

The example contains several attempts to detect unmeasured confounding bias, if present. There is a secondary outcome thought to be unaffected by alcohol consumption, namely methylmercury levels in the blood, likely an indicator of the consumption of fish, not of alcohol; see Pedersen et al. (1994) and WHO (2021). There are also three control groups, all with little present alcohol consumption, but with different uses of alcohol in the past; see the definition of variable grp above.

The data are used as an example in Rosenbaum (2022, 2023, 2024).

Source

US National Health and Nutrition Examination Survey (NHANES), 2013-2014 and 2015-2016.

References

LoConte, N. K., Brewster, A. M., Kaur, J. S., Merrill, J. K., and Alberg, A. J. (2018). Alcohol and cancer: a statement of the American Society of Clinical Oncology. Journal of Clinical Oncology 36, 83-93. <doi:10.1200/JCO.2017.76.1155>

Pedersen, G. A., Mortensen, G. K. and Larsen, E. H. (1994) Beverages as a source of toxic trace element intake. Food Additives and Contaminants, 11, 351–363. <doi:10.1080/02652039409374234>

Rosenbaum, P. R. (1987). <doi:10.1214/ss/1177013232> The role of a second control group in an observational study. Statistical Science, 2, 292-306.

Rosenbaum, P. R. (1989). <doi:10.2307/2531497> The role of known effects in observational studies. Biometrics, 45, 557-569.

Rosenbaum, P. R. (1989). <doi:10.1214/aos/1176347131> On permutation tests for hidden biases in observational studies. The Annals of Statistics, 17, 643-653.

Rosenbaum, P. R. (2014) Weighted M-statistics with superior design sensitivity in matched observational studies with multiple controls. Journal of the American Statistical Association, 109(507), 1145-1158 <doi:10.1080/01621459.2013.879261>

Rosenbaum, P. R. (2022) <doi:10.1080/00031305.2022.2063944> A New Transformation of Treated-Control Matched-Pair Differences for Graphical Display. American Statistician, 76(4), 346-352.

Rosenbaum, P. R. (2023) <doi:10.1111/biom.13558> Sensitivity analyses informed by tests for bias in observational studies. Biometrics, 79(1), 475-487.

Rosenbaum, P. R. (2024) Tightening an observational block design to form an optimally balanced subdesign. Manuscript.

Suh, I., Shaten, B. J., Cutler, J. A., and Kuller, L. H. (1992). Alcohol use and mortality from coronary heart disease: the role of high-density lipoprotein cholesterol. Annals of Internal Medicine 116, 881-887. <doi:10.7326/0003-4819-116-11-881>

World Health Organization (2021). Mercury and Health, <https://www.who.int/news-room/fact-sheets/detail/mercury-and-health>, (Accessed 30 August 2021).

Examples

```
data(aHDLt)
table(aHDLt$grp,aHDLt$grpL) # Short labels for plotting
boxplot(aHDLt$age~aHDLt$grp,xlab="Group",ylab="Age")
boxplot(aHDLt$education~aHDLt$grp,xlab="Group",ylab="Education")
table(aHDLt$female,aHDLt$grpL)
table(aHDLt$z,aHDLt$grpL)

# The sets were also matched for is.na(aHDLt$mmercury), for use
# in Rosenbaum (2023).  About half of the matched sets
# have values for mmercury.
table(is.na(aHDLt$mmercury),aHDLt$grp)

# See also the informedSen package for additional analysis


oldpar<-par(mfrow=c(1,2))

library(MASS)
huber2<-function(y){huber(y)$mu}

boxplot(aHDLt$hdl~aHDLt$grpL,ylab="HDL Cholesterol, mg/dL",cex.lab=1,
        cex.axis=.9,cex.main=.9,las=1,xlab="(i)")
axis(3,at=1:4,labels=round(tapply(aHDLt$hdl,aHDLt$grpL,huber2),1),
     cex.axis=.9)

boxplot(aHDLt$bmi~aHDLt$grpL,ylab="BMI",cex.lab=.9,
        cex.axis=1,cex.main=.9,las=1,xlab="(ii)")
axis(3,at=1:4,labels=round(tapply(aHDLt$bmi,aHDLt$grpL,huber2),1),
```

```
     cex.axis=.9)
par(oldpar)
```

`makematch` *Make a Match Using Two Criteria Matching with Optimal Subset Matching*

Description

The function makematch() is called by this package's main function, tighten(), as part of tightening an observational block design.

Usage

```
makematch(dat, costL, costR, ncontrols = 1, controlcosts = NULL,
          treatedcosts=NULL, large=100, solver="rrelaxiv")
```

Arguments

`dat`	A data frame. Typically, this is the entire data set. Part of it will be returned as a matched sample with some added variables.
`costL`	The distance matrix on the left side of the network, used for pairing. This matrix would most often be made by adding distances to a zero distance matrix created by startcost(), for instance, using addMahal(). In Figure 1 of Zhang et al. (2023), these are the costs on the left treated-control edges.
`costR`	The distance matrix on the right side of the network, used for balancing. This matrix would most often be made by adding distances to a zero distance matrix created by startcost(), for instance, using addNearExact(). If you do not need a right distance matrix, then initialize it to zero using startcost() and do not add additional distances to its intial form. In Figure 1 of Zhang et al. (2023), these are the costs on the right control-treated edges.
`ncontrols`	One positive integer, 1 for pair matching, 2 for matching two controls to each treated individual, etc.
`controlcosts`	An optional vector of costs used to penalize the control-control edges. For instance, one might penalize the use of controls with low propensity scores.
`treatedcosts`	An optional vector of costs that penalize the treated-treated edges used for subset matching in Rosenbaum (2012,2024). This option is available only if ncontrols=1. If treatedcosts = NULL, then the cost is set to the largest element of costL, costR and controlcosts multiplied-by-large to prevent subset matching. Otherwise, if treatedcosts is not NULL, then the ith coordinate of treatedcosts is the cost of not matching treated individual i.
`large`	A large positive number. Used only if treatedcosts = NULL. See treatedcosts.
`solver`	Determines the network optimization code that is used. Options are solver="rrelaxiv" and solver="rlemon". The Relax IV code of Bertsekas and Tseng (1988) is suggested, with solver="rrelaxiv", but is has an academic license, so various issues may arise. The makematch() function calls the callrelax() function in Pimentel's rcbalance package, whose documentation provides additional details, if needed.

Details

Implements the two-criteria matching method of Zhang et al (2022) with the possible addition of edges that permit some treated individuals to be removed rather than matched (Rosenbaum 2012). It is helpful to look at Figure 1 in Zhang et al. (2023) before using this function and Figure 4 in Rosenbaum (2024).

Value

Returns a matched data set. The matched rows of dat are returned with a new variable mset indicating the matched set. The returned file is sorted by mset and z.

Author(s)

Paul R. Rosenbaum

References

Bertsekas, D. P., Tseng, P. (1988) <doi:10.1007/BF02288322> The relax codes for linear minimum cost network flow problems. Annals of Operations Research, 13, 125-190.

Bertsekas, D. P. (1990) <doi:10.1287/inte.20.4.133> The auction algorithm for assignment and other network flow problems: A tutorial. Interfaces, 20(4), 133-149.

Bertsekas, D. P., Tseng, P. (1994) <http://web.mit.edu/dimitrib/www/Bertsekas_Tseng_RELAX4_!994.pdf> RELAX-IV: A Faster Version of the RELAX Code for Solving Minimum Cost Flow Problems.

Hansen, B. B. (2007) <https://www.r-project.org/conferences/useR-2007/program/presentations/hansen.pdf> Flexible, optimal matching for observational studies. R News, 7, 18-24. ('optmatch' package)

Pimentel, S. D. (2016) "Large, Sparse Optimal Matching with R Package rcbalance" <https://obsstudies.org/large-sparse-optimal-matching-with-r-package-rcbalance/> Observational Studies, 2, 4-23. (Discusses and illustrates the use of Pimentel's 'rcbalance' package.)

Rosenbaum, P. R. (1989) <doi:10.1080/01621459.1989.10478868> Optimal matching for observational studies. Journal of the American Statistical Association, 84(408), 1024-1032. (Discusses and illustrates fine balance using minimum cost flow in a network in section 3.2. This is implemented using makematch() by placing a large near-exact penalty on a nominal/integer covariate x1 on the right distance matrix.)

Rosenbaum, P. R., Ross, R. N. and Silber, J. H. (2007) <doi:10.1198/016214506000001059> Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. Journal of the American Statistical Association, 102, 75-83.

Rosenbaum, P. R. (2012) <doi:10.1198/jcgs.2011.09219> Optimal matching of an optimally chosen subset in observational studies. Journal of Computational and Graphical Statistics, 21(1), 57-71.

Rosenbaum, P. R. (2020) <doi:10.1007/978-3-030-46405-9> Design of Observational Studies (2nd Edition). New York: Springer.

Rosenbaum, P. R. (2024) Tightening an observational block design to form an optimally balanced subdesign. Manuscript.

Yang, D., Small, D. S., Silber, J. H. and Rosenbaum, P. R. (2012) <doi:10.1111/j.1541-0420.2011.01691.x> Optimal matching with minimal deviation from fine balance in a study of obesity and surgical outcomes. Biometrics, 68, 628-636. (Extension of fine balance useful when fine balance is infeasible.

Comes as close as possible to fine balance. Implemented in makematch() by placing a large near-exact penalty on a nominal/integer covariate x1 on the right distance matrix.)

Yu, R. (2023) <doi:10.1111/biom.13771> How well can fine balance work for covariate balancing? Biometrics, 79(3), 2346-2356.

Zhang, B., D. S. Small, K. B. Lasater, M. McHugh, J. H. Silber, and P. R. Rosenbaum (2023) <doi:10.1080/01621459.2021.1981337> Matching one sample according to two criteria in observational studies. Journal of the American Statistical Association, 118, 1140-1151. (This is the basic reference for two-criteria matching, which generalizes matching with fine balance.)

Zubizarreta, J. R., Reinke, C. E., Kelz, R. R., Silber, J. H. and Rosenbaum, P. R. (2011) <doi:10.1198/tas.2011.11072> Matching for several sparse nominal variables in a case control study of readmission following surgery. The American Statistician, 65(4), 229-238. (This paper combines near-exact matching and fine balance for the same covariate. It is implemented in makematch() by placing the same covariate on the left and the right.)

Examples

```
# This is a simple example to illustrate makematch().
# Please see the better match in the documentation for
# tighten() before continuing with this example.
#

###########
# Tighten match from 1-3 to 1-2 to adjust for BMI.
# BMI might be affected by alcohol consumption, so
# the primary comparison did not adjust for it.
###########

# The example below illustrates mechanics.
# See the same example in tighten() for a
# simpler construction of a better match.

data(aHDLt)
z<-aHDLt$z   #treatment indicator

# If you need to debug, it is helpful to have the same names for z
# and rownames for dat.
# These names are then used by functions that create distance matrices,
# including startcost(), addNearExact() and addMahal().
# If you use names, they will be checked for consistency and an error
# may result from incompatible names.

rownames(aHDLt)<-aHDLt$SEQN
names(z)<-aHDLt$SEQN

#  Create a zero cost (i.e., distance) matrix for the left and right
#  sides of the network.
left<-startcost(z)
right<-startcost(z)

left<-addNearExact(left,z,aHDLt$block) # Forces within-block matching.
#  Prefer to retain within blocks people who are closest for age, education.
```

```
left<-addMahal(left,z,cbind(aHDLt$education,aHDLt$age))

# Try to balance the categories of BMI which are out of balance.
right<-addNearExact(right,z,aHDLt$ibmi,penalty=20)

m<-makematch(aHDLt,left,right,ncontrols=2,large=10)

# Tightened blocks
table(aHDLt$z)
table(m$z)
table(table(aHDLt$block))
table(table(m$block))
table(table(m$mset))

boxplot(m$bmi[m$z==1],m$bmi[m$z==0],
        names=c("D","C"),ylab="BMI")

# Cost matrix for left side of network.
# Rows are treated, columns are control.
z[1:10]
round(left[1:5,1:4],1)

# Cost matrix for right side of network.
round(right[1:5,1:4],1)
```

`makenetwork` *Make the Network Used for Matching with Two Criteria*

Description

This function is of limited interest to most users, and is called by the makematch() function in the package. Makes the network used in the two-criteria matching method of Zhang et al (2023) with the possible addition of edges that permit some treated individuals to be removed rather than matched (Rosenbaum 2012,2024).

Usage

```
makenetwork(costL, costR, ncontrols = 1, controlcosts = NULL,
            treatedcosts=NULL, large=100)
```

Arguments

`costL`	The distance matrix on the left side of the network, used for pairing.
`costR`	The distance matrix on the right side of the network, used for balancing.
`ncontrols`	One positive integer, 1 for pair matching, 2 for matching two controls to each treated individual, etc.
`controlcosts`	An optional vector of nonnegative costs used to penalize the control-control edges.

treatedcosts	An optional vector of nonnegative costs used to penalize the treated-deletion edges. This option is available only if ncontrols=1. It is described in Rosenbaum (2024) and is closely related to the ideas in Rosenbaum (2012).
large	A large positive number. Used only if treatedcosts=NULL. See the parallel discussion in the documentation for makematch.

Details

This function creates the network depicted in Figure 1 of Zhang et al. (2023).

A minimum cost flow in this network is found by passing net to callrelax() in the package 'rcbalance'. If you use callrelax(), I strongly suggest you do this with solver set to 'rrelaxiv'. The 'rrelaxiv' package has an academic license. The 'rrelaxiv' package uses Fortran code from RELAX IV developed by Bertsekas and Tseng (1988, 1994) based on Bertsekas' (1990) auction algorithm.

Value

idtreated	Row identifications for treated individuals
idcontrol	Control identifications for control individuals
net	A network for use with callrelax in the 'rcbalance' package.

Author(s)

Paul R. Rosenbaum

References

Bertsekas, D. P., Tseng, P. (1988) <doi:10.1007/BF02288322> The relax codes for linear minimum cost network flow problems. Annals of Operations Research, 13, 125-190.

Bertsekas, D. P. (1990) <doi:10.1287/inte.20.4.133> The auction algorithm for assignment and other network flow problems: A tutorial. Interfaces, 20(4), 133-149.

Bertsekas, D. P., Tseng, P. (1994) <http://web.mit.edu/dimitrib/www/Bertsekas_Tseng_RELAX4_!994.pdf> RELAX-IV: A Faster Version of the RELAX Code for Solving Minimum Cost Flow Problems.

Rosenbaum, P. R. (2012) <doi:10.1198/jcgs.2011.09219> Optimal matching of an optimally chosen subset in observational studies. Journal of Computational and Graphical Statistics, 21(1), 57-71.

Rosenbaum, P. R. (2024) Tightening an observational block design to form an optimally balanced subdesign. Manuscript.

Zhang, B., D. S. Small, K. B. Lasater, M. McHugh, J. H. Silber, and P. R. Rosenbaum (2023) <doi:10.1080/01621459.2021.1981337> Matching one sample according to two criteria in observational studies. Journal of the American Statistical Association, 118, 1140-1151.

startcost *Initialize a Distance Matrix.*

Description

Creates an distance matrix of zeros of dimensions compatible with the treatment indicator vector z.

Usage

```
startcost(z)
```

Arguments

z	A vector with z[i]=1 if individual i is treated or z[i]=0 if individual i is control. The rows of costmatrix refer to treated individuals and the columns refer to controls. Although not strictly required, it is best that z has names that are the same as the names of the data frame dat that will be used in matching.

Value

A matrix of zeros with sum(z) rows and sum(1-z) columns. If z has names, then they become the row and column names of this matrix.

Author(s)

Paul R. Rosenbaum

Examples

```
# Although names are not required, they aid clarity.  The cost matrix
# has sum(z) rows and sum(1-z) columns, where z is a binary vector.
# If needed, it is easier to connect an entry in z to a row or column in
# cost if they have the same name.  If z has names, the data matrix should
# have the same names.
data(aHDLt)
rownames(aHDLt)<-aHDLt$SEQN
z<-aHDLt$z
names(z)<-aHDLt$SEQN
dist<-startcost(z)
dist[1:3,1:9]
```

`tighten` *Tightening an Observational Block Design*

Description

Implements a simple version of optimal balanced tightening of an observational block design. Finer control of the tightening may be obtained using the makematch() function; however, that requires more attention to detail.

Usage

```
tighten(dat, z, block, x = NULL, f = NULL,  ncontrols = 1,
    subset = NULL, pspace=10, solver="rrelaxiv")
```

Arguments

`dat`	A data frame for all data in the block design prior to tightening. Part of dat will be returned as a tightened block design with some added variables.
`z`	The vector z indicates treatment/control, where 1 indicates treated, and 0 indicates control. The number of rows of dat must equal length(z); otherwise, the tighten will stop with an error message.
`block`	A vector indicating the block. Every block must contain exactly one treated individual; that is, one individual whose value of z is 1. The minimum block size is two. The length of z must agree with the length of block. An error will result if these conditions are not met.
`x`	If not NULL, x is a matrix, dataframe or vector containing covariates to be used in a robust, rank based Mahalanobis distance (Rosenbaum 2020, Section 9.3). This is the within-block distance. If x is NULL, then the within-block distance is zero, and only between block distances affect tightening. In you want more information, see the documentation in this package for the function addMahal().
`f`	If not NULL, f determines the between block distances. In general, f describes fairly coarse nominal categories, either for one or several variables. If f is a vector or a factor, then it is used in fine or near-fine balance. If f is a matrix or a dataframe, then each column is a nominal variable, and they define a Hamming distance that counts the number of columns that differ. If f is a vector or a factor, then its length must equal the length of z. Otherwise, the number of rows of f must equal the length of z. If you want more information, see the documentation in this package for the function addNearExact().
`ncontrols`	If the blocks are initially of size J, with one treated and J-1 controls, then the tightened design will have blocks of size 1+ncontrols with one treated and ncontrols controls. If J-1 > 1, then ncontrols must be less than J-1. For instance, if J=4, then ncontrols might be either 1 or 2. If J=2 so J-1=1, then ncontrols must be 1 and subset must not be NULL.

`subset` If subset is NULL, then no blocks are discarded. Otherwise, subset should be a positive number. Roughly speaking, if subset=50, then tighten will prefer to discard a block than face a cost or distance of 50 by including the block. The distance here is the sum of within and between block distances. This option is available only if the blocks are of size two, i.e., pairs. For blocks larger than pairs, we prefer to reduce their size rather than eliminate them. An error will result if the subset is not NULL when the blocks are larger than pairs.

`pspace` This parameter is slightly technical, so it may be best to use the default until forced to do otherwise. Priorities in the matching are enforced by penalized costs. Most important, with the largest penalty, is to retain the original block structure. Second, if subset=NULL, then no blocks are to be deleted. Third, is the fine balance/Hamming distance constraint. The lowest priority with no penalty is the Mahalanobis distance within blocks. pspace is used to space the penalties, so priorities are respected. Increasing pspace emphasizes the relative importance of priorities, but may slow down the optimization. In general, the largest lower priority penalized distance is multiplied by pspace to produce the new penalty. The easist way to check that pspace is large enough is to increase it; if it is large enough, the match should be the same, but may take longer to compute.

`solver` Determines the network optimization code that is used. Options are solver="rrelaxiv" and solver="rlemon". The Relax IV code of Bertsekas and Tseng (1988) is suggested, with solver="rrelaxiv", but is has an academic license, so various issues may arise. The makematch() function calls the callrelax() function in Pimentel's rcbalance package, whose documentation provides additional details, if needed.

Details

The tighten function produces a simple version of an optimally tightened block design, combining a Mahalanobis distance within-blocks with some version of fine balancing between blocks. You can achieve finer control and subler effects using the makematch function, whose structure is closer to the mathematical structure of the tightening problem. The examples in the makematch function document these finer features.

Value

Returns a data set for a tightened block design. The matched rows of dat are returned with a new variable mset indicating the matched set. The returned file is sorted by mset and z.

Author(s)

Paul R. Rosenbaum

References

Bertsekas, D. P., Tseng, P. (1988) <doi:10.1007/BF02288322> The relax codes for linear minimum cost network flow problems. Annals of Operations Research, 13, 125-190.

Bertsekas, D. P. (1990) <doi:10.1287/inte.20.4.133> The auction algorithm for assignment and other network flow problems: A tutorial. Interfaces, 20(4), 133-149.

Bertsekas, D. P., Tseng, P. (1994) <http://web.mit.edu/dimitrib/www/Bertsekas_Tseng_RELAX4_!994.pdf> RELAX-IV: A Faster Version of the RELAX Code for Solving Minimum Cost Flow Problems.

Hansen, B. B. (2007) <https://www.r-project.org/conferences/useR-2007/program/presentations/hansen.pdf> Flexible, optimal matching for observational studies. R News, 7, 18-24. ('optmatch' package)

Pimentel, S. D. (2016) "Large, Sparse Optimal Matching with R Package rcbalance" <https://obsstudies.org/large-sparse-optimal-matching-with-r-package-rcbalance/> Observational Studies, 2, 4-23. (Discusses and illustrates the use of Pimentel's 'rcbalance' package.)

Rosenbaum, P. R. (1984) <doi:10.2307/2981697> The consequences of adjustment for a concomitant variable that has been affected by the treatment. Journal of the Royal Statistical Society Series A: Statistics in Society, 147(5), 656-666. (Related to the BMI example.)

Rosenbaum, P. R. (1989) <doi:10.1080/01621459.1989.10478868> Optimal matching for observational studies. Journal of the American Statistical Association, 84(408), 1024-1032. (Discusses and illustrates fine balance using minimum cost flow in a network in Section 3.2.)

Rosenbaum, P. R. (2006) <doi:10.1093/biomet/93.3.573> Differential effects and generic biases in observational studies. Biometrika, 93(3), 573-586. (Related to the dentist example.)

Rosenbaum, P. R. (2012) <doi:10.1198/jcgs.2011.09219> Optimal matching of an optimally chosen subset in observational studies. Journal of Computational and Graphical Statistics, 21(1), 57-71.

Rosenbaum, P. R. (2020) <doi:10.1007/978-3-030-46405-9> Design of Observational Studies (2nd Edition). New York: Springer. (Discusses fine balance and the rank-based Mahalanobis distance.)

Rosenbaum, P. R. (2024) Tightening an observational block design to form an optimally balanced subdesign. Manuscript.

Yang, D., Small, D. S., Silber, J. H. and Rosenbaum, P. R. (2012) <doi:10.1111/j.1541-0420.2011.01691.x> Optimal matching with minimal deviation from fine balance in a study of obesity and surgical outcomes. Biometrics, 68, 628-636. (Extension of fine balance useful when fine balance is infeasible. Comes as close as possible to fine balance.)

Yu, R. (2023) <doi:10.1111/biom.13771> How well can fine balance work for covariate balancing? Biometrics, 79(3), 2346-2356.

Zhang, B., D. S. Small, K. B. Lasater, M. McHugh, J. H. Silber, and P. R. Rosenbaum (2023) <doi:10.1080/01621459.2021.1981337> Matching one sample according to two criteria in observational studies. Journal of the American Statistical Association, 118, 1140-1151. (This method generalizes the concept of fine balance. Here, it is used with the Hamming distance when f is a matrix.)

Examples

```
# The two examples below are from Rosenbaum (2024).

###########
# Tighten match from 1-3 to 1-2 to adjust for BMI.
# BMI might be affected by alcohol consumption, so
# the primary comparison did not adjust for it.
# See Rosenbaum (1984).
###########

data(aHDLt)
result<-tighten(aHDLt,aHDLt$z,aHDLt$block,
```

```
        x=cbind(aHDLt$age,aHDLt$education),
        f=cbind(aHDLt$ibmi,(aHDLt$bmi>22.5)+(aHDLt$bmi>27.5)+(aHDLt$bmi>32.5)),
                 ncontrols=2)


omit<-aHDLt[!is.element(aHDLt$SEQN,result$SEQN),]
boxplot(result$bmi[result$z==1],result$bmi[result$z==0],omit$bmi,
    names=c("D","C","O"),ylab="BMI")
boxplot(result$hdl[result$z==1],result$hdl[result$z==0],omit$hdl,
    names=c("D","C","O"),ylab="HDL Cholesterol")

# Tightened blocks
table(aHDLt$z)
table(result$z)
table(table(aHDLt$block))
table(table(result$block))
table(table(result$mset))


##################
# Dentist in 1 Year Differential Effect
##################

#
#  Example of tightening in support of a differential
#  comparison to address a generic bias; see Rosenbaum (2006).
#
#
# First, create a data frame in which each block contains a
# differential comparison.
#
# Identify blocks in which the treated individual did not go
# to the dentist in the past year, but at lease one control did.
# Vector dife picks out the blocks that have this pattern.
data(aHDLt)
dif1<-tapply(((aHDLt$z==1)&(aHDLt$dentist1Y==0)),aHDLt$block,sum)==1
dif0<-tapply(((aHDLt$z==0)&(aHDLt$dentist1Y==1)),aHDLt$block,sum)>=1
dif<-dif1&dif0
dife<-as.vector(rbind(dif,dif,dif,dif))
elig<-((aHDLt$z==1)&(aHDLt$dentist1Y==0))|((aHDLt$z==0)&(aHDLt$dentist1Y==1))

# Now, form a data.frame for the people eligible for this comparison
# The tightened match will occur in this data.frame.
aHDLd<-cbind(aHDLt,elig)[dife,]
aHDLe<-aHDLd[((aHDLd$z==1)&(aHDLd$dentist1Y==0))|
            ((aHDLd$z==0)&(aHDLd$dentist1Y==1)),]
rm(dif,dif0,dif1,dife,elig,aHDLd)

#  With data frame aHDLe, several tightened block designs
#  are constructed.

x<-cbind(aHDLe$age,aHDLe$education)
```

```
# No pairs are deleted.  Education is not quite balanced.
m1<-tighten(aHDLe,aHDLe$z,aHDLe$block,x=x,f=aHDLe$education)
table(m1$z,m1$education)
tapply(m1$age,m1$z,summary)

# 6 pairs are deleted.  Education is alsmost balanced.
m2<-tighten(aHDLe,aHDLe$z,aHDLe$block,x=x,f=aHDLe$education,subset=150)
table(m2$z,m2$education)
tapply(m2$age,m2$z,summary)

# 13 pairs are deleted.  Perhaps too many. Education is balanced.
m3<-tighten(aHDLe,aHDLe$z,aHDLe$block,x=x,f=aHDLe$education,subset=50)
table(m3$z,m3$education)
tapply(m3$age,m3$z,summary)

oldpar<-par(mfrow=c(1,3))
barplot(t(table(m1$education,1-m1$z)),beside=TRUE,ylim=c(0,50),
        ylab="Count",xlab="Education Level",main="All 118 Pairs",
        col=gray.colors(2))
barplot(t(table(m2$education,1-m2$z)),beside=TRUE,ylim=c(0,50),
        ylab="Count",xlab="Education Level",main="Best 112 Pairs")
barplot(t(table(m3$education,1-m3$z)),beside=TRUE,ylim=c(0,50),
        ylab="Count",xlab="Education Level",main="Best 105 Pairs")
par(oldpar)
```

Index