CEM: Software for Coarsened Exact Matching*

Stefano M. Iacus† Gary King‡ Giuseppe Porro§

Version
December 27, 2009

Contents

*Forthcoming, *Journal of Statistical Software*

†Department of Economics, Business and Statistics, University of Milan, Via Conservatorio 7, I-20124 Milan, Italy; stefano.iacus@unimi.it

‡Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge MA 02138; `http://GKing.harvard.edu`, king@harvard.edu, (617) 495-2027.

§Department of Economics and Statistics, University of Trieste, P.le Europa 1, I-34127 Trieste, Italy; giuseppe.porro@econ.units.it.

1 Introduction

Overview This program is designed to improve causal inference via a method of matching that is widely applicable in observational data and easy to understand and use (if you understand how to draw a histogram, you will understand this method). The program implements the coarsened exact matching (CEM) algorithm, described below. CEM may be used alone or in combination with any existing matching method. This algorithm, and its statistical properties, are described in

> Stefano M. Iacus, Gary King, and Giuseppe Porro, "Matching for Causal Inference Without Balance Checking", `http://gking.harvard.edu/files/abs/cem-abs.shtml`.

Properties **?** show that CEM is a monotonoic imbalance bounding (MIB) matching method — which means both that the maximum imbalance between the treated and control groups may be chosen by the user ex ante, rather than discovered through the usual laborious process of ex post checking and repeatedly reestimating, and that adjusting the maximum imbalance on one variable has no effect on the maximum imbalance of any other.

This paper also shows that CEM bounds through ex ante user choice both the degree of model dependence and the average treatment effect estimation error, eliminates the need for a separate procedure to restrict data to common empirical support, meets the congruence principle, is robust to measurement error, works well with multiple imputation and

other methods for missing data, can be completely automated, and is fast computationally even with very large data sets. After preprocessing data with CEM, the analyst may then use a simple difference in means, or whatever matching method or statistical model they would have applied to the raw data. CEM also works for multicategory treatments, creating randomized blocks in experimental designs, and evaluating extreme counterfactuals.

Goal Matching is not a method of estimation; it is a way to preprocess a data set so that estimation of SATT based on the matched data set will be less "model-dependent" (i.e., less a function of apparently small and indefensible modeling decisions) than when based on the original full data set. Matching involves pruning observations that have no close matches on pre-treatment covariates in both the treated and control groups. The result is typically less model-dependence, bias, and (by removing heterogeneity) inefficiency (King and Zeng, 2006; Ho et al., 2007; **?**). If used for analyzing observational data, applications of CEM (and all other methods of causal inference) require an assumption of ignorability (a.k.a. "no omitted variable bias" or "no confounding").

The specific statistical goal is to estimate some version of a causal effect, such as the sample average treatment effect on the treated (the "SATT"). Thus, let Y_i be the dependent variable for unit i, T_i be a treatment variable, and X_i be a vector of pre-treatment control variables. Although CEM works as easily with multicategory treatment variables, we simplify this introductory description by assuming that T_i is dichotmous and takes on the value 1 for "treated" units and 0 for "control" units. Then define the treatment effect for treated units as the difference between two potential outcomes: $\text{TE}_i = Y_i(T_i = 1) - Y_i(T_i = 0)$, where $Y_i(T_i = 1) = Y_i$ is always obseved and $Y_i(T_i = 0)$, the value that Y_i would have taken on if it were the case that $T_i = 0$, is always unobserved. Then $Y_i(T_i = 0)$ is estimated with Y_j from matched controls (i.e., among units for which $X_i \approx X_j$), either directly, $\hat{Y}_i(T_i = 0) = Y_j(T_j = 0)$, or via a model, $\hat{Y}_i(T_i = 0) = \hat{g}(X_j)$. Then SATT can be computed as a simple average: $\text{SATT} = \frac{1}{n_T} \sum_{i \in \{T_i = 1\}} \text{TE}_i$.

Algorithm The CEM algorithm then involves three steps:

1. *Temporarily* coarsen each control variable in X as much as you are willing, for the purposes of matching. For example, years of education might be coarsened into grade school, middle school, high school, college, graduate school. Most researchers are intimately familiar with the concept and practice of coarsening, as it is widely used in applied data analyses in many fields, although unlike its present use coarsening for data analysis involves a permanent removal of information from the analysis and ultimate estimates.

2. Sort all units into strata, each of which has the same values of the coarsened X.

3. Prune from the data set the units in any stratum that do not include at least one treated and one control unit.

Following these three steps, the researcher can apply any method to the matched data that they might have to the raw data to estimate the causal effect, with the addition of a weight that equalizes the number of treated and control units within each stratum. Thus, any existing method of matching may be used within CEM strata to further prune the data in other ways, in which case the combined approach still inherits all of CEM's properties. Whether or not another method of matching is applied, one must compute the causal effect, either by a (weighted) difference in means in Y among the treated and control units, or with the application of a statistical model.

If the coarsened bins are set to zero width, then CEM returns the exact matching solution, in which case model dependence will be eliminated (other than the ignorability assumption), but too few observations may be left. If instead the coarsened bins are set too wide, then few observations will be discarded, but differences within the large strata must be spanned with a statistical model, in which case model dependence may be an issue.

What if the level of coarsening is set as large as the researcher finds reasonable, but the number of observations is still too small? This of course may happen with a single continuous covariate with little or no coarsening or due to higher order interactions among a large set of discrete covariates. We offer below a "progressive coarsening" procedure that may help you rethink some of your coarsening choices by indicating how many more observations you would recover by loosening up the coarsening level for each variable. But if the remaining sample is still too small, the only possibilites involve collecting more data; setting the coarsening level artificially large and relying on theory to rule out some of the model dependence; or living with the model dependence and increasing the uncertainty with which you draw your inferences and represent your conclusions. To be clear, in this situation, you are in a bind and no method of matching or analysis is likely to save you. Statistics of course is not magic and can only get you so far given limited information. When in this situation, it is best to recognize the limits of your data relative to the question you are asking, and decide whether to devote whatever resources at your disposal to collecting more data, developing better theory, dealing with the uncertainty, or choosing a different research project.

Measuring Balance Although CEM is MIB, the actual degree of imbalance achieved in the matched sample may be lower than the chosen maximum, and so we also introduce a simple and comprehensive *multivariate* imbalance measure (**?**). The measure is based on the $\mathcal{L}_1$ difference between the multidimensional histogram of all pretreatment covariates in the treated group and that in the control group. To do this, we first choose the number of bins for each continuous variable via standard automated univariate histogram methods and with categorical variables left as is. These bin sizes must be defined separately from and prior to the coarsening levels chosen for CEM. Although this initial choice poses all the usual issues and potential problems when choosing bins in drawing histograms, we use it only as a fixed reference to evaluate pre and post matching imbalance. (We also offer a function, which are logically similar to ROC curve for classification problems, that compute this measure for a very large number of coarsenings to see whether one matching solution dominates others; see `L1.profile`.)

Our functions compute these bin sizes automatically using automated histogram methods (and with smaller bins than would typically be chosen in running CEM), or they can optionally be set by the user, so long as this level is fixed through all subseqent matching. Then, we cross-tabulate the discretized variables as $X_1 \times \cdots \times X_k$ for the treated and control groups separately, and record the k-dimensional relative frequencies for the treated $f_{\ell_1 \cdots \ell_k}$ and control $g_{\ell_1 \cdots \ell_k}$ units. Finally, our measure of imbalance is the absolute difference over all the cell values:

$$\mathcal{L}_1(f, g) = \frac{1}{2} \sum_{\ell_1 \cdots \ell_k} |f_{\ell_1 \cdots \ell_k} - g_{\ell_1 \cdots \ell_k}| \tag{1}$$

and where the summation is over all cells of the multivariate histogram, but is feasible to compute because it contains at most n nonzero terms. The $\mathcal{L}_1$ measure[1] varies in $[0, 1]$. Perfect (up to discretization) global balance results in $\mathcal{L}_1 = 0$, and $\mathcal{L}_1 = 1$ indicates complete separation of the multimensional histograms. Any value in the interval $(0, 1)$ indicates the amount of difference between k-dimensional frequencies of the two groups.

CEM also offers several other measures of imbalance such as the global difference in means and the difference in means within the strata that are defined by every matching method.

2 Setup

2.1 Software Requirements

CEM works in conjunction with the R Project for Statistical Computing, and will run on any platform where R is installed (Windows, Linux, or Mac). R is available free for download at the Comprehensive R Archive Network (CRAN) at `http://cran.r-project.org/`. CEM has been tested on the most recent version of R.

CEM may be run by installing the program directly, as indicated below, or by using the alternative interface to CEM provided by MatchIt (`http://gking.harvard.edu/matchit`, (Ho et al., Forthcoming)). Using CEM directly is faster. The MatchIt interface is easier for some applications and works seemlessly with Zelig (`http://gking.harvard.edu/zelig`) for estimating causal effects after matching, but presently only offers a subset of features of the R version. A Stata verison of CEM is also available at the CEM web site, `http://gking.harvard.edu/cem`.

2.2 Installation

To install cem, type at the R command prompt,

```
> install.packages("cem")
```

and CEM will install itself onto your system automatically from CRAN. You may alternatively load the beta test version as

```
> install.packages("cem",repos="http://gking.harvard.edu")
```

[1] Prior to version 1.0.90 of the **cem** package the $\mathcal{L}_1$ measure did not include the factor $\frac{1}{2}$.

2.3 Loading CEM

You need to install CEM only once, but you must load it prior to each use. Do this at the R prompt:

```
> library(cem)
```

2.4 Updating CEM

We recommend that you periodically update CEM at the R prompt by typing:

```
> update.packages()
```

which will update all the libraries including CEM and load the new version of the package with

```
> library(cem)
```

3 A User's Guide

We show here how to use CEM through a simple running example: the National Supported Work (NSW) Demonstration data, also known as the Lalonde data set (Lalonde, 1986). This program provided training to selected individuals for 12-18 months and help finding a job in the hopes of increasing their' earnings. The treatment variable, `treated`, is 1 for participants (the treatment group) and 0 for nonparticipants (the control group). The key outcome variable is earnings in 1978 (`re78`).

Since participation in the program was not assigned strictly at random, we must control for a set of pretreatment variables by the CEM algorithm. These pre-treatment variables include age (`age`), years of education (`education`), marital status (`married`), lack of a high school diploma (`nodegree`), race (`black`, `hispanic`), indicator variables for unemployment in 1974 (`u74`) and 1975 (`u75`), and real earnings in 1974 (`re74`) and 1975 (`re75`). Some of these are dichotomous (`married`, `nodegree`, `black`, `hispanic`, `u74`, `u75`), some are categorical (`age` and `education`), and the earnings variables are continuous and highly skewed with point masses at zero. We modify these data by adding a (fictitious) variable to illustrate discrete responses called `q1`, the answer to a survey question asking before assignment for an opinion about this job training program, with possible responses *strongly agree*, *agree*, *neutral*, *strongly disagree*, *disagree*, and *no opinion*; note that the last category is not on the same ordered scale as the other responses. Ten percent of the observations have missing data (added randomly by us to illustrate how CEM deals with missingness). We call this new data set the `LeLonde` (intentionally misspelling Lalonde); the original, unmodified Lalonde (1986) data are contained in data set `LL`.

3.1 Basic Evaluation and Analysis of Unmatched Data

We begin with a naive estimate of SATT — the simple difference in means — which would be useful only if the in-sample distribution of pre-treatment covariates were the same in the treatment and control groups:

```
> require(cem)

How to use CEM? Type vignette("cem")

> data(LeLonde)
```

We remove missing data from the the data set before starting the analysis (we show better procedures for dealing with missing data in Section 3.6).

```
> Le <- data.frame(na.omit(LeLonde))
```

and then compute the size of the treated and control groups:

```
> tr <- which(Le$treated == 1)
> ct <- which(Le$treated == 0)
> ntr <- length(tr)
> nct <- length(ct)
```

Thus, the data include 258 treated units and 392 control units. The (unadjusted and therefore likely biased) difference in means is then:

```
> mean(Le$re78[tr]) - mean(Le$re78[ct])

[1] 759
```

Because the variable `treated` was not randomly assigned, the pre-treatment covariates differ between the treated and control groups. To see this, we focus on these pre-treatment covariates:

```
> vars <- c("age", "education", "black", "married", "nodegree",
+     "re74", "re75", "hispanic", "u74", "u75", "q1")
```

The overall imbalance is given by the $\mathcal{L}_1$ statistic:

We compute $\mathcal{L}_1$ statistic, as well as several unidimensional measures of imbalance via our `imbalance` function. In our running example:

```
> imbalance(group = Le$treated, data = Le[vars])
```

```
Multivariate Imbalance Measure: L1=0.902
Percentage of local common support: LCS=5.8%

Univariate Imbalance Measures:

           statistic   type        L1 min 25%   50%    75%     max
age        -0.252373 (diff) 5.102e-03   0   0   0.0   -1.0    -6.0
education   0.153635 (diff) 8.464e-02   1   0   1.0    1.0     1.0
black      -0.010323 (diff) 1.032e-02   0   0   0.0    0.0     0.0
married    -0.009551 (diff) 9.551e-03   0   0   0.0    0.0     0.0
nodegree   -0.081217 (diff) 8.122e-02   0  -1   0.0    0.0     0.0
re74      -18.160447 (diff) 5.551e-17   0   0 284.1  806.3 -2139.0
re75      101.501762 (diff) 5.551e-17   0   0 485.6 1238.4   490.4
hispanic   -0.010145 (diff) 1.014e-02   0   0   0.0    0.0     0.0
u74        -0.045582 (diff) 4.558e-02   0   0   0.0    0.0     0.0
u75        -0.065555 (diff) 6.556e-02   0   0   0.0    0.0     0.0
q1          7.494021 (Chi2) 1.067e-01  NA  NA    NA     NA      NA
```

Only the overall $\mathcal{L}_1$ statistic measure includes imbalance with respect to the joint distribution, including all interactions, of the covariates; in the case of our example, $\mathcal{L}_1$ =0.902. The unidimensional measures in the table are all computed for each variable separately.

The first column in the table of unidimensional measures, labeled `statistic`, reports the difference in means for numerical variables (indicated by the second column, `type`, reporting `(diff)`) or a chi-square difference for categorical variables (when the second column reports `(Chi2)`). The second column, labeled `L1`, reports the $\mathcal{L}_1^j$ measure, which is $\mathcal{L}_1$ computed for the j-th variable separately (which of course does not include interactions). The remaining columns in the table report the difference in the empirical quantile of the distributions of the two groups for the 0th (min), 25th, 50th, 75th, and 100th (max) percentiles for each variable. When the variable type is `Chi2`, the only variable-by-variable measure that is defined in this table is $\mathcal{L}_1^j$; others are reported missing.

This particular table shows that variables `re74` and `re75` are imbalanced in the raw data in many ways and variable `age` is balanced in means but not in the quantiles of the two distributions. This table also illustrates the point that balancing only the means between the treated and control groups does not necessarily guarantee balance in the rest of the distribution. Most important, of course, is the overall $\mathcal{L}_1$ measure, since even if the marginal distribution of every variable is perfectly balanced, the joint distribution can still be highly imbalanced.

As an aside, we note that for convenience that the function `imbalance` allows you to drop some variables before computation:

```
todrop <- c("treated", "re78")
imbalance(group=Le$treated, data=Le, drop=todrop)
```

3.2 Coarsened Exact Matching

We now apply the coarsened exact matching algorithm by calling the function `cem`. The CEM algorithm performs exact matching on coarsened data to determine matches and then passes on the uncoarsened data from observations that were matched to estimate the causal effect. Exact matching works by first sorting all the observations into strata, each of which has identical values for all the coarsened pre-treatment covariates, and then discarding all observations within any stratum that does not have at least one observation for each unique value of the treatment variable.

To run this algorithm, we must choose a type of coarsening for each covariate. We show how this is done this via a fully automated procedures in Section 3.2.1. Then we show how to use explicit prior knowledge to choose the coarsening in Section 3.2.2, which is normally preferable when feasible.

In CEM, the treatment variable may be *dichotomous* or *mutichotomous*. Alternatively, `cem` may be used for *randomized block experiments* without specifying a treatment variable; in this case no strata are deleted and the treatment variable is (randomly) assigned to units within each strata to ensure that each has at least one observation assigned each value of the treated variable.

3.2.1 Automated Coarsening

In our running example we have a dichotomous treatment variable. In the following code, we match on all variables but `re78`, which is the outcome variable and so should never be included. Hence we proceed specifying `"re78"` in argument `drop`:

```
> mat <- cem(treatment = "treated", data = Le, drop = "re78")
```

The output object `mat` contains useful information about the match, including a (small) table about the number of observations in total, matched, and unmatched, as well as the results of a call to the `imbalance` function for information about the quality of the matched data (unless `eval.imbalance` is set to `FALSE`). Since `cem` bounds the imbalance ex ante, the most important information in `mat` is the number of observations matched. But the results also give the imbalance in the matched data using the same measures as that in the original data described in Section 3.1. Thus,

```
> mat

           G0  G1
All       392 258
Matched    95  84
Unmatched 297 174


Multivariate Imbalance Measure: L1=0.605
```

```
Percentage of local common support: LCS=25.8%

Univariate Imbalance Measures:

           statistic   type        L1 min 25% 50%   75%    max
age          0.09405 (diff) 0.000e+00   0   0   1   0.0    0.0
education   -0.02222 (diff) 2.222e-02   0   0   0   0.0    0.0
black        0.00000 (diff) 0.000e+00   0   0   0   0.0    0.0
married      0.00000 (diff) 0.000e+00   0   0   0   0.0    0.0
nodegree     0.00000 (diff) 0.000e+00   0   0   0   0.0    0.0
re74       149.64180 (diff) 0.000e+00   0   0   0 463.3  889.5
re75       158.75212 (diff) 0.000e+00   0   0   0 843.7 -640.9
hispanic     0.00000 (diff) 0.000e+00   0   0   0   0.0    0.0
u74          0.00000 (diff) 0.000e+00   0   0   0   0.0    0.0
u75          0.00000 (diff) 0.000e+00   0   0   0   0.0    0.0
q1           2.08335 (Chi2) 1.388e-17  NA  NA  NA    NA     NA
```

We can see from these results the number of observations matched and thus retained, as well as those which were pruned because they were not comparable. By comparing the imbalance results to the original imbalance table given in the previous section, we can see that a good match can produce a substantial reduction in imbalance, not only in the means, but also in the marginal and joint distributions of the data.

The function `cem` also generates weights for use in the evaluation of imbalance measures and estimates of the causal effect (stored in `mat$w`).

3.2.2 Coarsening by Explicit User Choice

The power and simplicity of CEM comes from choosing the coarsening yourself rather than using the automated algorithm as in the previous section. Choosing the coarsening enables you to set the maximum level of imbalance ex ante, which is a direct function of the coarsening you choose. By controlling the coarsening, you also put an explicit bound on the degree of model dependence and the SATT estimation error.

Fortunately, the coarsening is a fundamentally substantive act, almost synonymous with the measurement of the original variables. In other words, if you know something about the data you are analyzing, you almost surely have enough information to choose the coarsening. (And if you don't know something about the data, you might ask why you are analyzing it in the first place!)

In general, we want to set the coarsening for each variable so that substantively indistinguishable values are grouped and assigned the same numerical value. Groups may be of different sizes if appropriate. Recall that any coarsening during CEM is used only for matching; the original values of the variables are passed on to the analysis stage for all matched observations.

The function `cem` treats categorical and numerical variables differently.

For *categorical* variables, we use the `grouping` option. For example, variable `q1` has the following levels

```
> levels(Le$q1)

[1] "agree"             "disagree"          "neutral"
[4] "no opinion"        "strongly agree"    "strongly disagree"
```

Notice that the levels are not ordered in the original data set. One can possibly tranform the variable `q1` from `factor` to ordered factor using the command `ordered` in R or may want to group them in three groups as follows:

```
> q1.grp <- list(c("strongly agree", "agree"), c("neutral", "no opinion"),
+     c("strongly disagree", "disagree"))
```

For *numerical* variables, we use the `cutpoints` option. Thus, for example, in the US educational system, the following discretization of years of education corresponds to different levels of school

Grade school	0–6
Middle school	7–8
High school	9–12
College	13–16
Graduate school	>16

Using these natural breaks in the data to create the coarsening is generally a good approach and usually better than caliper matching in this context, as it would disregard these meaningful breaks. (The venerable technique of caliper matching of course may be useful for certain other types of data.) Because in our data, no respondents fall in the last category,

```
> table(Le$education)

  3   4   5   6   7   8   9  10  11  12  13  14  15
  1   5   4   6  12  55 106 146 173 113  19   9   1
```

we define the cutpoints as:

```
> educut <- c(0, 6.5, 8.5, 12.5, 17)
```

and run `cem` adding only the `grouping` and `cutpoints` options, leaving the rest unchanged:

```
> mat1 <- cem(treatment = "treated", data = Le, drop = "re78",
+     cutpoints = list(education = educut), grouping = list(q1.grp))
> mat1
```

```
           G0  G1
All       392 258
Matched   137  91
Unmatched 255 167


Multivariate Imbalance Measure: L1=0.713
Percentage of local common support: LCS=19.8%

Univariate Imbalance Measures:

           statistic   type        L1 min 25% 50%   75%    max
age         1.809e-01 (diff) 0.000e+00   0   0   1   0.0    0.0
education   5.716e-02 (diff) 5.754e-02   0   1   0   0.0   -2.0
black      -1.110e-16 (diff) 6.245e-17   0   0   0   0.0    0.0
married    -6.939e-18 (diff) 1.145e-16   0   0   0   0.0    0.0
nodegree   -1.110e-16 (diff) 6.939e-17   0   0   0   0.0    0.0
re74        1.022e+02 (diff) 0.000e+00   0   0   0 463.3  889.5
re75        1.161e+02 (diff) 0.000e+00   0   0   0 843.7 -640.9
hispanic   -6.939e-18 (diff) 5.898e-17   0   0   0   0.0    0.0
u74         0.000e+00 (diff) 2.776e-17   0   0   0   0.0    0.0
u75        -1.110e-16 (diff) 1.110e-16   0   0   0   0.0    0.0
q1          2.064e+00 (Chi2) 7.633e-17  NA  NA  NA    NA     NA
```

As we can see, this matching solution differs from that resulting from our automated approach in the previous section. For comparison, the automatic cutpoints produced by `cem` are stored in the output object in slot `breaks`. So, for example, our automated coarsening produced:

```
> mat$breaks$education

 [1]  3.0  4.2  5.4  6.6  7.8  9.0 10.2 11.4 12.6 13.8 15.0
```

whereas we can recover our personal choice of cutpoints as

```
> mat1$breaks$education

[1]  0.0  6.5  8.5 12.5 17.0
```

3.3 Progressive coarsening

Although the maximum imbalance is fixed ex ante by the user's coarsening choices, the number of observations matched is determined as a consequence of the matching procedure. If you are dissatisfied with the number of observations available after matching, and you feel that it is substantively appropriate to coarsen further, then just increase the coarsening (by

using fewer cutpoints). The result will be additional matches and of course a concommitant increase in the maximum possible imbalance between the treated and control groups. This is easy with CEM because CEM is a monotonic imbalance bounding (MIB) method, which means that increasing the imbalance on one variable (by widening the coarsened bin sizes) will not change the maximum imbalance on any other variable. MIB thus enables you to tinker with the solution one variable at a time to quickly produce a satisfactory result, if one is feasible.

If, however, you feel that additional coarsening is not appropriate, than too few obser-ations may indicate that your data contains insufficient information to estimate the causal effects of interest without model dependence; in that situation, you either give up or will have to attempt adjusting for the pre-treatment covariates via modeling assumptions.

Suppose, instead, that you are unsure whether to coarsen further or how much to coarsen, and are willing to entertain alternative matching solutions. We offer here an automated way to compute these solutions. The idea is to relax the initial `cem` solution selectively and automatically, to prune equivalent solutions, and to present them in a convenient manner so that users can ascertain where the difficulties in matching in these data can be found and what choices would produce which outcomes in terms of the numbers of observations matched.

For categorical variables, the algorithm considers the numerical values associated to each level of the variable. In R the numerical values associated to the levels go from 1 to the number of levels, say k. The coarsening occurs by partitioning the interval $[1, k]$ into intervals of growing size starting from, say, $\epsilon = 1$. So in all cases, coarsening occurs by grouping adjacent levels from "left" to "right". This is of course not completely appropriate for pure unordered categorical variables, but to treat them in a proper way additional combinatorics would be necessary. The progressive coarsening is instead intended as an instrument which gives a feeling on which variable is more likely to prevent matching for a given data set.

We start by illustrating what happens when we relax a CEM solution "by hand". The following three runs show the effect on the matching solution (in terms of the number of observations and imbalance) when the coarsening for one variable (`age`) is relaxed from 10 to 6 to 3 bins. As can be seen, fewer cutpoints (which means larger bins) produces more matched units and high maximum (and in this case actual) imbalance:

```
> cem("treated", Le, cutpoints = list(age = 10), drop = "re78",
+     grouping = q1.grp)

           G0  G1
All       392 258
Matched    87  77
Unmatched 305 181


Multivariate Imbalance Measure: L1=0.644
Percentage of local common support: LCS=22.5%
```

```
Univariate Imbalance Measures:

           statistic   type        L1 min 25%   50%   75%    max
age        1.405e-01 (diff) 0.000e+00   0   0   1.0  -1.0    0.0
education -4.762e-03 (diff) 4.762e-03   0   0   0.0   0.0    0.0
black      0.000e+00 (diff) 0.000e+00   0   0   0.0   0.0    0.0
married    0.000e+00 (diff) 5.551e-17   0   0   0.0   0.0    0.0
nodegree   0.000e+00 (diff) 0.000e+00   0   0   0.0   0.0    0.0
re74       1.453e+02 (diff) 0.000e+00   0   0 378.5 385.3  889.5
re75       7.227e+01 (diff) 0.000e+00   0   0 275.9 695.1 -640.9
hispanic   0.000e+00 (diff) 5.551e-17   0   0   0.0   0.0    0.0
u74       -5.551e-17 (diff) 2.776e-17   0   0   0.0   0.0    0.0
u75        0.000e+00 (diff) 0.000e+00   0   0   0.0   0.0    0.0
q1         3.611e+00 (Chi2) 0.000e+00  NA  NA    NA    NA     NA

> cem("treated", Le, cutpoints = list(age = 6), drop = "re78",
+     grouping = q1.grp)

           G0  G1
All       392 258
Matched   131 101
Unmatched 261 157


Multivariate Imbalance Measure: L1=0.729
Percentage of local common support: LCS=18.8%

Univariate Imbalance Measures:

          statistic   type        L1 min 25%   50%   75%    max
age         0.20165 (diff) 0.000e+00   0   0   0.0   0.0    0.0
education  -0.02871 (diff) 2.871e-02   0   0   0.0   0.0    0.0
black       0.00000 (diff) 0.000e+00   0   0   0.0   0.0    0.0
married     0.00000 (diff) 0.000e+00   0   0   0.0   0.0    0.0
nodegree    0.00000 (diff) 0.000e+00   0   0   0.0   0.0    0.0
re74      149.25677 (diff) 0.000e+00   0   0 369.3 481.7  889.5
re75      106.52881 (diff) 0.000e+00   0   0 202.6 556.8 -640.9
hispanic    0.00000 (diff) 0.000e+00   0   0   0.0   0.0    0.0
u74         0.00000 (diff) 0.000e+00   0   0   0.0   0.0    0.0
u75         0.00000 (diff) 0.000e+00   0   0   0.0   0.0    0.0
q1          1.25381 (Chi2) 1.388e-17  NA  NA    NA    NA     NA
```

```
> cem("treated", Le, cutpoints = list(age = 3), drop = "re78",
+     grouping = q1.grp)

            G0  G1
All        392 258
Matched    173 123
Unmatched  219 135


Multivariate Imbalance Measure: L1=0.783
Percentage of local common support: LCS=14.0%

Univariate Imbalance Measures:

            statistic   type        L1 min 25%    50%    75%   max
age       -2.923e-01 (diff) 5.551e-17   0  -1    0.0   -1.0  -7.0
education -7.724e-03 (diff) 7.724e-03   0   0    0.0    0.0   0.0
black      1.110e-16 (diff) 6.245e-17   0   0    0.0    0.0   0.0
married    6.939e-18 (diff) 5.898e-17   0   0    0.0    0.0   0.0
nodegree   1.110e-16 (diff) 6.939e-17   0   0    0.0    0.0   0.0
re74      -6.242e+01 (diff) 5.551e-17   0   0    0.0 -570.8 957.1
re75      -9.322e+01 (diff) 5.551e-17   0   0 -228.7 -400.8 640.9
hispanic   3.469e-18 (diff) 5.725e-17   0   0    0.0    0.0   0.0
u74        0.000e+00 (diff) 2.776e-17   0   0    0.0    0.0   0.0
u75        5.551e-17 (diff) 2.776e-17   0   0    0.0    0.0   0.0
q1         1.951e+00 (Chi2) 7.633e-17  NA  NA     NA     NA    NA
```

We automate this *progressive coarsening* procedure here in the `relax.cem` function. This function starts with the output of `cem` and relaxes variables one (`depth=1`), two (`depth=2`), or three (`depth=3`) at a time, while optionally keeping unchanged a chosen subset of the variables which we know well or have important effects on the outcome (`fixed`). The function also allows one to specify the minimal number of breaks of each variable (the default limit being 1). We begin with this example (the argument `perc=0.3` is passed to the plot function and implies that only the solutions with at least 30% of the units are matched)

```
> tab <- relax.cem(mat, Le, depth = 1, perc = 0.3)

Executing 47 different relaxations

  |
  |=======                                                          |  10%
  |
  |==============                                                   |  20%
```

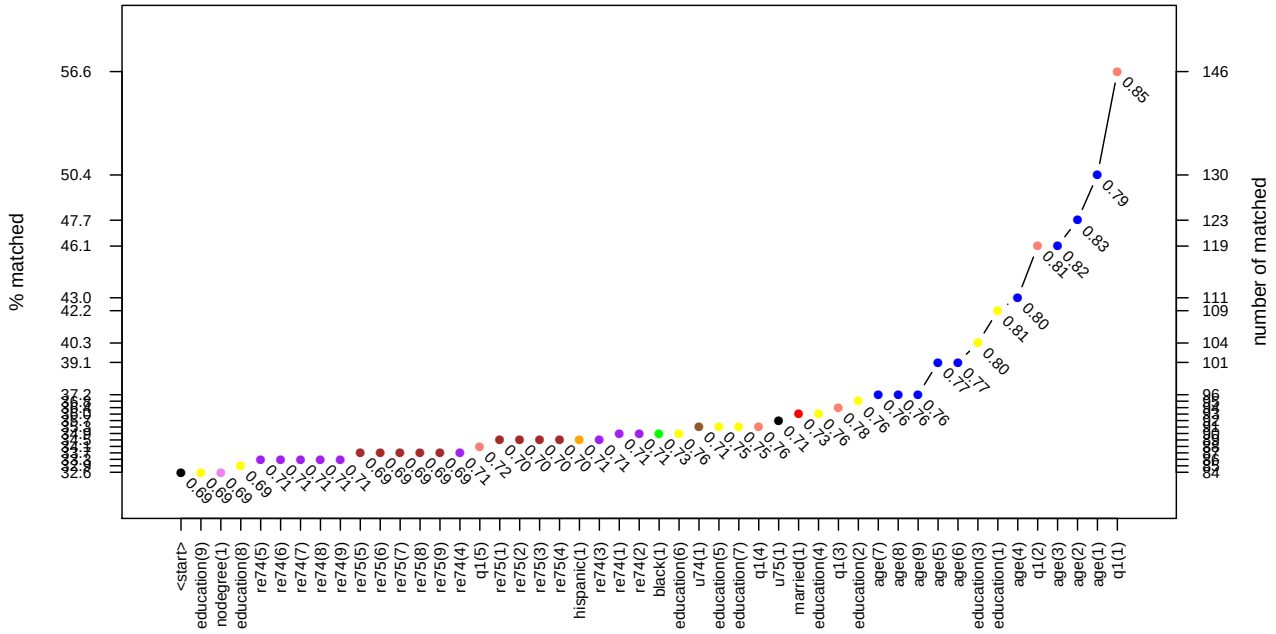

Figure 1: Example of the graphical output of `relax.cem`.

```
|
|====================                                                   |  30%
|
|============================                                           |  40%
|
|==================================                                     |  50%
|
|========================================                               |  60%
|
|================================================                       |  70%
|
|==========================================================             |  80%
|
|====================================================================   |  90%
|
|======================================================================| 100%
```

After all possible coarsening relaxations are attempted, the function returns a list of tables, one per group (i.e. treated and control). Each row of the tables contain information about the number of treated and control units matched, the value of the $\mathcal{L}_1$ measure, and the type of relaxation made. Each table is the sorted according to the number of treated (or

control) units matched.

The user may want to see the output of `tab$G1` or `tab$G0` but these tables may be very long, and so we provide a method `plot` to view these tables more conveniently. The output of `plot(tab)` is plotted in Figure 1 from which it is seen that the most difficult variables to match are `age` and `education`. On the x-axis of the plot the variable and the number of equally sized bins used for the coarsening are used (color-coded by variable). On the y-axis on the right is the absolute number of treated units matched, while the left side y-axis reports the same number in percentages. The numbers below the dots in the graphs represent the $\mathcal{L}_1$ measure corresponding to that matching solution. This graph also gives a feeling of the MIB behaviour of `cem`. When the tables produced by `relax.cem` are too large, the `plot` function, allows for some reduction like printing only the best matching solutions (in the terms of number of treated units matched), removing duplicates (i.e. different coarsenings may lead to the same matching solution), or printing only solution where at least some percentage of treated units, have been matched, or a combination of these. For more information refer to the reference manual for the function `relax.plot` which can be called directly instead of plot.

Here is one example of use of `plot` in which we specify that only solutions with at least 60% of the treated units are matched and duplicated solutions are removed. The output can be seen in Figure 2

```
> plot(tab, group = "1", perc = 0.35, unique = TRUE)
```

3.4 Restricting the Matching Solution to a k-to-k Match

By default, CEM uses maximal information, resulting in strata that may include different numbers of treated and control units. To compensate for the differential strata sizes, `cem` also returns weights to be used in subsequent analyses. Although this is generally the best option, a user with enough data may opt for a k-to-k solution to avoid the slight inconvenience of needing to use weights.

The function `k2k` accomplishes this by pruning observations from a `cem` solution within each stratum until the solution contains the same number of treated and control units in all strata. Pruning occurs within a stratum (for which observations are indistuinguishable to cem proper) by using nearest neighbor selection using a distance function specified by the user (including `euclidean`, `maximum`, `manhattan`, `canberra`, `binary`, or `minkowski`). By default `method` is set to `NULL`, which means random matching inside `cem` strata, an option that may reduce the chance for bias. (For the Minkowski distance the power can be specified via the argument `mpower`. For more information on `method != NULL`, refer to `dist` help page.)

Here is an example of this approach. First, by running `cem`:

```
> mat <- cem(treatment = "treated", data = Le, drop = "re78")
> mat
```

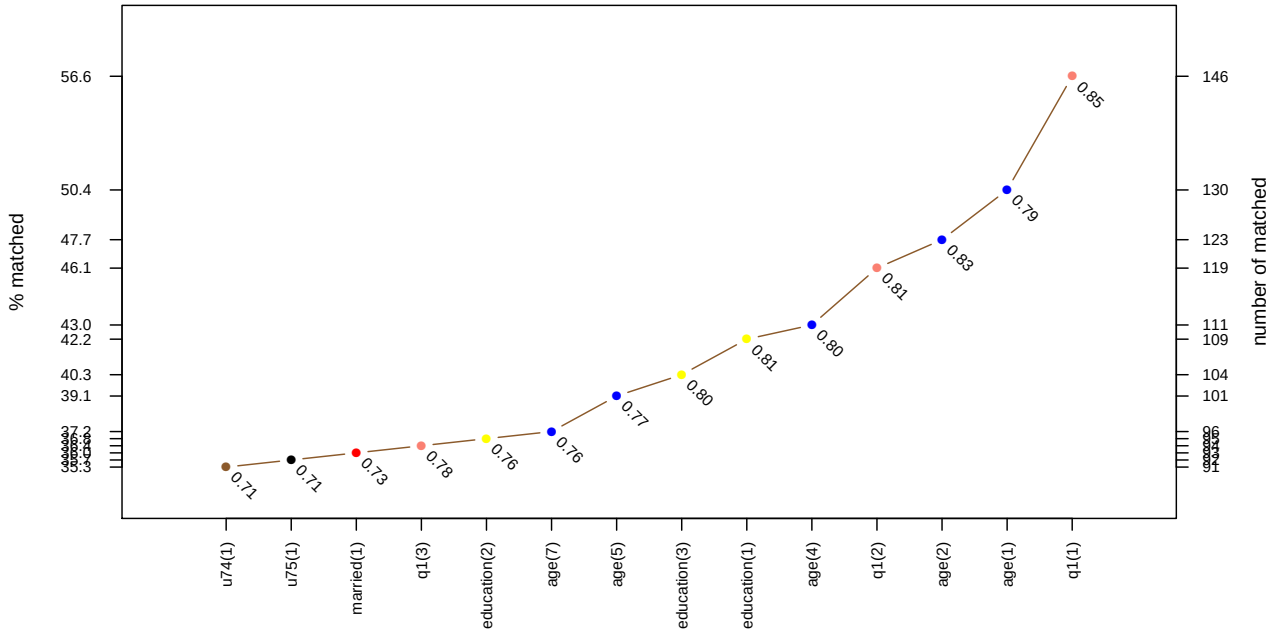

Figure 2: Example of reduced graphical output of `relax.cem`.

```
           G0  G1
All       392 258
Matched    95  84
Unmatched 297 174


Multivariate Imbalance Measure: L1=0.605
Percentage of local common support: LCS=25.8%

Univariate Imbalance Measures:

          statistic   type        L1 min 25% 50%   75%    max
age         0.09405 (diff) 0.000e+00   0   0   1   0.0    0.0
education  -0.02222 (diff) 2.222e-02   0   0   0   0.0    0.0
black       0.00000 (diff) 0.000e+00   0   0   0   0.0    0.0
married     0.00000 (diff) 0.000e+00   0   0   0   0.0    0.0
nodegree    0.00000 (diff) 0.000e+00   0   0   0   0.0    0.0
re74      149.64180 (diff) 0.000e+00   0   0   0 463.3  889.5
re75      158.75212 (diff) 0.000e+00   0   0   0 843.7 -640.9
hispanic    0.00000 (diff) 0.000e+00   0   0   0   0.0    0.0
u74         0.00000 (diff) 0.000e+00   0   0   0   0.0    0.0
```

```
u75          0.00000 (diff) 0.000e+00   0   0   0   0.0    0.0
q1           2.08335 (Chi2) 1.388e-17  NA  NA  NA    NA     NA
```

```
> mat$k2k
```

```
[1] FALSE
```

and now pruning to a k-to-k solution, using the euclidean distance within CEM strata:

```
> mat2 <- k2k(mat, Le, "euclidean", 1)
> mat2
```

```
           G0  G1
All       392 258
Matched    70  70
Unmatched 322 188


Multivariate Imbalance Measure: L1=0.605
Percentage of local common support: LCS=25.8%

Univariate Imbalance Measures:

           statistic   type        L1 min 25% 50%   75%    max
age          0.09405 (diff) 0.000e+00   0   0   1   0.0    0.0
education   -0.02222 (diff) 2.222e-02   0   0   0   0.0    0.0
black        0.00000 (diff) 0.000e+00   0   0   0   0.0    0.0
married      0.00000 (diff) 0.000e+00   0   0   0   0.0    0.0
nodegree     0.00000 (diff) 0.000e+00   0   0   0   0.0    0.0
re74       149.64180 (diff) 0.000e+00   0   0   0 463.3  889.5
re75       158.75212 (diff) 0.000e+00   0   0   0 843.7 -640.9
hispanic     0.00000 (diff) 0.000e+00   0   0   0   0.0    0.0
u74          0.00000 (diff) 0.000e+00   0   0   0   0.0    0.0
u75          0.00000 (diff) 0.000e+00   0   0   0   0.0    0.0
q1           2.08335 (Chi2) 1.388e-17  NA  NA  NA    NA     NA
```

```
> mat2$k2k
```

```
[1] TRUE
```

Alternatively, we can produce the same result in one step by adding the `k2k=TRUE` option to the original `cem` call.

3.5 Estimating the Causal Effect from `cem` output

Using the output from `cem`, we can estimate SATT via the `att` function. The simplest approach requires a weighted difference in means (unless `k2k` was used, in which case no weights are required). For convenience, we compute this as a regression of the outcome variable on a constant and the treatment variable,

```
> data(LL)
> mat <- cem(treatment = "treated", data = LL, drop = "re78")
> est <- att(mat, re78 ~ treated, data = LL)
> est

           G0  G1
All       425 297
Matched   222 163
Unmatched 203 134

Linear regression model on CEM matched data:

SATT point estimate: 550.962564 (p.value=0.368242)
95% conf. interval: [-647.777701, 1749.702830]
```

where the SATT estimate is the coefficient on the `treated` variable, in our case 550.96. The function `att` allows for R's standard `formula` interface and, by default, uses a linear model to estimate the att using the weights produced by `cem`.

If exact matching (i.e., without coarsening) was chosen this procedure is appropriate as is. In other situations, with some coarsening, some imbalance remains in the matched data. The remaining imbalance is strictly bounded by the level of coarsening, which can be seen by any remaining variation within the coarsened bins. Thus, a reasonable approach in this common situation is to attempt to adjust for the remaining imbalance via a statistical model. (Modeling assumptions for models applied to the matched data are much less consequential than they would otherwise be because CEM is known to strictly bound the level of model dependence.) To apply a statistical model to control for the remaining imbalance, we use the `formula` interface in `att`. For example:

```
> est2 <- att(mat, re78 ~ treated + re74, data = LL)
> est2

           G0  G1
All       425 297
Matched   222 163
Unmatched 203 134

Linear regression model on CEM matched data:
```

```
SATT point estimate: 553.113736 (p.value=0.362760)
95% conf. interval: [-636.606542, 1742.834014]
```

The user can also specify the option `model` which accepts one of the following string arguments

- `linear` or `lm` (the default) for linear model, when the treatment effect is supposed to be homogeneous

```
> att(mat, re78 ~ treated + re74, data = LL, model = "linear")

           G0  G1
All       425 297
Matched   222 163
Unmatched 203 134

Linear regression model on CEM matched data:

SATT point estimate: 553.113736 (p.value=0.362760)
95% conf. interval: [-636.606542, 1742.834014]
```

- `linear-RE` or `lme` for linear model with random effects in cem strata, for non-homogeneous treatment effect

```
> att(mat, re78 ~ treated + re74, data = LL, model = "linear-RE")

           G0  G1
All       425 297
Matched   222 163
Unmatched 203 134

Linear random effect model on CEM matched data:

SATT point estimate: 552.448961 (p.value=0.000000)
95% conf. interval: [364.067068, 740.830853]
```

- `logistic` or `logit` for dichotomous response variable[2], for homogeneous treatment effect.

- `forest` or `rf` for random forest model, also for non-homogeneous treatment effect. It accepts continuous, dichotomous or counting outcomes.

```
> att(mat, re78 ~ treated + re74, data = LL, model = "forest")
```

[2]We do not provide an example here, model the syntax is the same for the other models.

```
            G0  G1
All        425 297
Matched    222 163
Unmatched  203 134

Random forest model on CEM matched data:

SATT point estimate: 539.779654 (p.value=0.547602)
95% conf. interval: [-1219.489250, 2299.048558]
```

All the above models run on the CEM matched subsamples, so the quantity of interest may change in case of non-homogeneous treatment effect. The option `extrapolate`, if set `TRUE`, extrapolates each of the above models also to the set of treated units not matched. In this case the quantity of interest is kept fixed but the estimation is more model dependent.

```
> att(mat, re78 ~ treated + re74, data = LL, model = "linear",
+     extra = TRUE)

           G0  G1
All       425 297
Matched   222 163
Unmatched 203 134

Linear regression model with extrapolation:

SATT point estimate: 674.337762 (p.value=0.347286)
95% conf. interval: [-236.091964, 1584.767489]

> att(mat, re78 ~ treated + re74, data = LL, model = "linear-RE",
+     extra = TRUE)

           G0  G1
All       425 297
Matched   222 163
Unmatched 203 134

Linear random effect model with extrapolation:

SATT point estimate: 902.087484 (p.value=0.000000)
95% conf. interval: [816.567245, 987.607724]

> att(mat, re78 ~ treated + re74, data = LL, model = "rf", extra = TRUE)
```

```
           G0  G1
All       425 297
Matched   222 163
Unmatched 203 134

Random forest model with extrapolation:

SATT point estimate: 25.972220 (p.value=0.968503)
95% conf. interval: [-1263.224645, 1315.169084]
```

As Figure 3 shows, it is also possible to plot the results of the SATT estimation as follows

```
> est3 <- att(mat, re78 ~ treated + re74, data = LL)
> est3

           G0  G1
All       425 297
Matched   222 163
Unmatched 203 134

Linear regression model on CEM matched data:

SATT point estimate: 553.113736 (p.value=0.362760)
95% conf. interval: [-636.606542, 1742.834014]

> plot(est3, mat, LL, vars = c("education", "age", "re74", "re75"))
```

For more information, see the reference manual entry for `att`.

3.6 Matching and Missing Data

Almost all previous methods of matching assume the absence of any missing values. In contrast, CEM offers two valid approaches to dealing with missing values (item nonresponse). In the first, where we treat missing values as one of the values of the variables, is appropriate when "`NA`" is a valid value that is not really missing (such as when "no opinion" really means no opinion); see Section 3.6.1. The other is a special procedure to allow for multiply imputed data in CEM, as described in Section 3.6.2.

3.6.1 Matching on Missingness

In the next example, we use our original `LeLonde` data with missing values and we compare the result with `Le` from which we dropped the `NA` values. For comparability, we use the same cutpoints we used in Section 3.2 on the `Le` data. The cutpoints are contained in `mat$breaks`

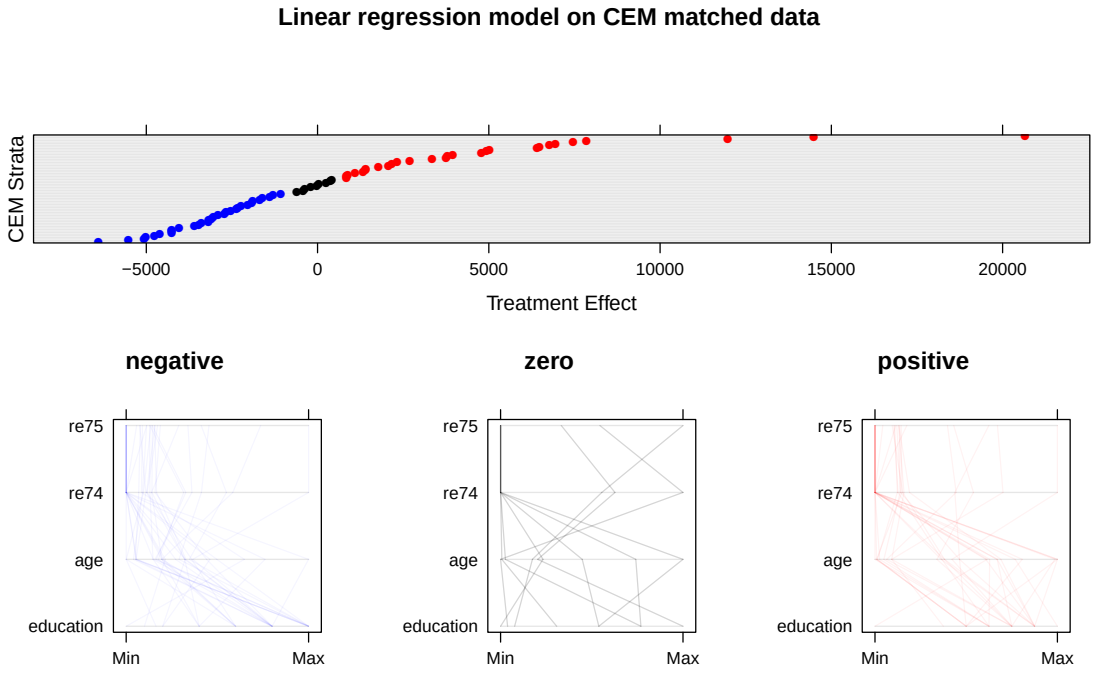

Figure 3: Example of a plot of the output of `att`. The top panel gives observation-level causal effect estimates sorted in numerical order and colored in ranges – negative (in blue), not significantly different from zero (black), or positive (red). For each range of effects, the bottom panel gives parallel plots; each line in a parallel plot represents the (covariate) characteristics of a single observation.

```
> mat3 <- cem("treated", LeLonde, drop = "re78", cutpoints = mat$breaks,
+     grouping = list(q1 = q1.grp))

The data contain missing values. CEM will match on them; see the manual for other option

> mat3

           G0  G1
All       425 297
Matched   134 101
Unmatched 291 196


Multivariate Imbalance Measure: L1=0.723
Percentage of local common support: LCS=19.9%
```

```
Univariate Imbalance Measures:

           statistic   type        L1 min 25% 50% 75%    max
age        1.893e-01 (diff) 0.000e+00   0   0   0   0    0.0
education  0.000e+00 (diff) 0.000e+00   0   0   0   0    0.0
black     -1.110e-16 (diff) 5.551e-17   0   0   0   0    0.0
married    0.000e+00 (diff) 0.000e+00   0   0   0   0    0.0
nodegree   0.000e+00 (diff) 0.000e+00   0   0   0   0    0.0
re74       4.563e+01 (diff) 0.000e+00   0   0   0 543  889.5
re75       8.120e+01 (diff) 0.000e+00   0   0   0 816 -640.9
hispanic   0.000e+00 (diff) 0.000e+00   0   0   0   0    0.0
u74        0.000e+00 (diff) 2.776e-17   0   0   0   0    0.0
u75        0.000e+00 (diff) 0.000e+00   0   0   0   0    0.0
q1         3.822e+00 (Chi2) 7.084e-02  NA  NA  NA  NA     NA
```

and we compare the above with the solution obtained by dropping the observations with missing data

```
> mat4 <- cem("treated", Le, drop = "re78", cutpoints = mat$breaks,
+     grouping = list(q1 = q1.grp))
> mat4

           G0  G1
All       392 258
Matched   132 100
Unmatched 260 158


Multivariate Imbalance Measure: L1=0.734
Percentage of local common support: LCS=19.3%

Univariate Imbalance Measures:

           statistic   type        L1 min 25% 50% 75%    max
age        1.919e-01 (diff) 0.000e+00   0   0   0   0    0.0
education 0.000e+00 (diff) 0.000e+00   0   0   0   0    0.0
black     0.000e+00 (diff) 3.469e-18   0   0   0   0    0.0
married   0.000e+00 (diff) 0.000e+00   0   0   0   0    0.0
nodegree  0.000e+00 (diff) 0.000e+00   0   0   0   0    0.0
re74      4.583e+01 (diff) 0.000e+00   0   0   0 543  889.5
re75      8.287e+01 (diff) 0.000e+00   0   0   0 816 -640.9
hispanic  3.469e-18 (diff) 1.735e-18   0   0   0   0    0.0
u74       0.000e+00 (diff) 0.000e+00   0   0   0   0    0.0
```

```
u75        0.000e+00 (diff) 2.776e-17   0   0   0   0    0.0
q1         3.698e+00 (Chi2) 7.279e-02  NA  NA  NA  NA     NA
```

and, as expected, the two solutions differ somewhat. The gain (in terms of number of matched units) decreases as the number of covariates increases.

3.6.2 Matching Multiply Imputed Data

Consider a data set to be matched, some of which is missing. One approach to analyzing data with missing values is *multiple imputation*, which involves creating m (usually about $m = 5$) data sets, each of which is the same as the original except that the missing values have been imputed in each. Uncertainty in the values of the missing cells is represented by variation in the imputations across the different imputed data sets (King et al., 2001).

As an example we take the original `LeLonde` data with missing values

```
> summary(LeLonde)
```

```
    treated           age          education        black        married
 Min.   :0.000   Min.   :17.0   Min.   : 3.0   Min.   :0.0   Min.   :0.000
 1st Qu.:0.000   1st Qu.:19.0   1st Qu.: 9.0   1st Qu.:1.0   1st Qu.:0.000
 Median :0.000   Median :23.0   Median :10.0   Median :1.0   Median :0.000
 Mean   :0.411   Mean   :24.5   Mean   :10.3   Mean   :0.8   Mean   :0.163
 3rd Qu.:1.000   3rd Qu.:27.0   3rd Qu.:11.0   3rd Qu.:1.0   3rd Qu.:0.000
 Max.   :1.000   Max.   :55.0   Max.   :16.0   Max.   :1.0   Max.   :1.000
                 NA's   : 8.0   NA's   : 8.0   NA's   :1.0   NA's   :9.000
    nodegree          re74            re75            re78
 Min.   :0.000   Min.   :    0   Min.   :    0   Min.   :    0
 1st Qu.:1.000   1st Qu.:    0   1st Qu.:    0   1st Qu.:    0
 Median :1.000   Median :  824   Median :  941   Median : 4033
 Mean   :0.778   Mean   : 3662   Mean   : 3051   Mean   : 5486
 3rd Qu.:1.000   3rd Qu.: 5237   3rd Qu.: 3993   3rd Qu.: 8813
 Max.   :1.000   Max.   :39571   Max.   :37432   Max.   :60308
 NA's   :5.000   NA's   :    8   NA's   :    4   NA's   :    7
    hispanic           u74             u75                          q1
 Min.   : 0.000   Min.   :0.000   Min.   :0.000   agree            :111
 1st Qu.: 0.000   1st Qu.:0.000   1st Qu.:0.000   disagree         :121
 Median : 0.000   Median :0.000   Median :0.000   neutral          :129
 Mean   : 0.104   Mean   :0.453   Mean   :0.402   no opinion       :117
 3rd Qu.: 0.000   3rd Qu.:1.000   3rd Qu.:1.000   strongly agree   :121
 Max.   : 1.000   Max.   :1.000   Max.   :1.000   strongly disagree:118
 NA's   :11.000   NA's   :3.000   NA's   :3.000   NA's             :  5
```

Now we use `Amelia` package (Honaker, King and Blackwell, 2006) to create multiply imputed data sets:

```
> require(Amelia)

##
## Amelia II: Multiple Imputation
## (Version 1.2-13, built: 2009-08-12)
## Copyright (C) 2005-2009 James Honaker, Gary King and Matthew Blackwell
## Refer to http://gking.harvard.edu/amelia/ for more information
##

> set.seed(123)
> imputed <- amelia(LeLonde, noms = c("black", "hispanic", "treated",
+     "married", "nodegree", "u74", "u75", "q1"))

-- Imputation 1 --

 1  2  3

-- Imputation 2 --

 1  2  3  4

-- Imputation 3 --

 1  2  3

-- Imputation 4 --

 1  2  3  4

-- Imputation 5 --

 1  2  3

> imputed <- imputed$imputations[1:5]
```

Now `imputed` contains a list of 5 multiply imputed versions of `LeLonde`. We pass this list to the `cem` function in the argument `datalist` and `cem` produces a set of multiply imputed solutions, as usual with the original uncoarsened values of the variables, but now assigning each multiply imputed observation to the strata where it falls most frequently. The output of `cem` is a list of `cem.match` solutions (named `match1`, `match2`,..., `match5`). (Be sure to also name the original data frame in option `data` or `cem` will merely run the basic algorithm five separate times on each of the input data sets, a procedure that can be useful for batch processing of data to be matched, but is not recommended for multiply imputed data sets since the strata will not be the same across the data sets.) For example:

```
> mat2 <- cem("treated", datalist = imputed, drop = "re78", data = LeLonde,
+     grouping = list(q1 = q1.grp))
> mat2
```

```
           G0  G1
All       425 297
Matched   132  98
Unmatched 293 199


Multivariate Imbalance Measure: L1=0.703
Percentage of local common support: LCS=22.3%

Univariate Imbalance Measures:

          statistic   type       L1    min 25% 50%    75%    max
age       -0.141910 (diff) 0.000000    0.0  -1   0    0.0    0.0
education -0.003269 (diff) 0.017007    0.0   0   0    0.0    0.0
black      0.000000 (diff) 0.000000    0.0   0   0    0.0    0.0
married   -0.003061 (diff) 0.005102    0.0   0   0    0.0    0.0
nodegree  -0.002041 (diff) 0.000000    0.0   0   0    0.0    0.0
re74      99.066521 (diff) 0.000000 -250.6   0   0  252.3 -197.1
re75      44.619415 (diff) 0.000000    0.0   0   0 -128.5  640.9
hispanic   0.001531 (diff) 0.007653    0.0   0   0    0.0    0.0
u74        0.004082 (diff) 0.000000    0.0   0   0    0.0    0.0
u75        0.000000 (diff) 0.000000    0.0   0   0    0.0    0.0
q1         2.141127 (Chi2) 0.061905     NA  NA  NA     NA     NA
```

Now we estimate SATT via the usual multiple imputation combining formulas (averaging the point estimates and within and between variances, as usual; see King et al. 2001). The function `att` implements these procedures:

```
> out <- att(mat2, re78 ~ treated, data = imputed)
> out
```

```
Linear regression model on CEM matched data:

SATT point estimate: 1177.182780 (p.value=0.123993)
95% conf. interval: [-322.747427, 2677.112987]
```

3.7 Creating paired samples

In same cases, it is useful to apply CEM so some data to create paired matched samples. Given an output of `cem`, the function `pair` produces two sets of indexes corresponding to pair matched units.

```
> data(LL)
> mat <- cem(data = LL, drop = "re78")
> psample <- pair(mat, data = LL)

Total number of units paired in CEM strata: 352
Total number of units matched: 722
```

each pair of observation has a different strata number

```
> table(psample$paired)

  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
 21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
 41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
 61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
 81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
```

not all observations can be matched in cem strata

```
> psample$paired[1:100]

15993 15994 15995 15996 15997 15998 15999 16000 16001 16002 16003 16004 16005
   NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
16006 16007 16008 16009 16010 16011 16012 16013 16014 16015 16016 16017 16018
   NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
16019 16020 16021 16022 16023 16024 16025 16026 16027 16028 16029 16030 16031
   NA    NA    NA    NA    NA     1    NA    NA    NA     1    NA    NA    NA
16032 16033 16034 16035 16036 16037 16038 16039 16040 16041 16042 16043 16044
   NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
16045 16046 16047 16048 16049 16050 16051 16052 16053 16054 16055 16056 16057
    3    NA    NA     4    NA    NA    NA    NA    NA    NA    NA    NA    NA
```

```
16058 16059 16060 16061 16062 16063 16064 16065 16066 16067 16068 16069 16070
   NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA     3     4
16071 16072 16073 16074 16075 16076 16077 16078 16079 16080 16081 16082 16083
   NA     6    NA    NA     7    NA    NA    NA    NA    NA    NA    NA    NA
16084 16085 16086 16087 16088 16089 16090 16091 16092
    8    NA    NA     9    NA     9    NA    NA    NA
```

the remaining observations are then matched and the final list of all paired units is contained in the filed `full.paired`

```
> table(psample$full.paired)

  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
 21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
 41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
 61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
 81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
```

```
321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
361
  2
```

```
> psample$full.paired[1:10]

15993 15994 15995 15996 15997 15998 15999 16000 16001 16002
  265   268   240   189   196   195   267   177   227   232
```

paired; if they cannot be paired, the function indicates which units are left without a mate.

```
> mat1 <- cem(data = LL[-1, ], drop = "re78")
> psample <- pair(mat1, data = LL[-1, ])

Total number of units paired in CEM strata: 352
Total number of units matched: 720
Unit corresponding to row `15994', not paired
```

```
> table(psample$full.paired)

  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
 21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
 41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
 61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
 81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220
```

2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
221	222	223	224	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239	240
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
241	242	243	244	245	246	247	248	249	250	251	252	253	254	255	256	257	258	259	260
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
261	262	263	264	265	266	267	268	269	270	271	272	273	274	275	276	277	278	279	280
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
281	282	283	284	285	286	287	288	289	290	291	292	293	294	295	296	297	298	299	300
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
301	302	303	304	305	306	307	308	309	310	311	312	313	314	315	316	317	318	319	320
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
321	322	323	324	325	326	327	328	329	330	331	332	333	334	335	336	337	338	339	340
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
341	342	343	344	345	346	347	348	349	350	351	352	353	354	355	356	357	358	359	360
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2

4 Reference to CEM's Functions

4.1 cem: Coarsened Exact Matching

Description

Implementation of Coarsened Exact Matching

Usage

```
cem(treatment=NULL, data = NULL, datalist=NULL, cutpoints = NULL,
    grouping = NULL, drop=NULL, eval.imbalance = TRUE, k2k=FALSE,
method=NULL, mpower=2, L1.breaks = NULL, verbose = 0)
```

Arguments

`treatment`	character, name of the treatment variable
`data`	a data.frame
`datalist`	a list of optional multiply imputed data.frame's
`cutpoints`	named list each describing the cutpoints for numerical variables (the names are variable names). Each list element is either a vector of cutpoints, a number of cutpoints, or a method for automatic bin contruction. See Details.
`grouping`	named list, each element of which is a list of groupings for a single categorical variable. See Details.
`drop`	a vector of variable names in the data frame to ignore during matching
`eval.imbalance`	Boolean. See Details.
`k2k`	boolean, restrict to k-to-k matching? Default = `FALSE`
`method`	distance method to use in `k2k` matching. See Details.
`mpower`	power of the Minkowski distance. See Details.
`L1.breaks`	list of cutpoints for the calculation of the L1 measure.
`verbose`	controls level of verbosity. Default=0.

Details

When specifying cutpoints, several automatic methods may be chosen, including "`sturges`" (Sturges' rule, the default), "`fd`" (Freedman-Diaconis' rule), "`scott`" (Scott's rule) and "ss" (Shimazaki-Shinomoto's rule). See references for a description of each rule.

The `grouping` option is a list where each element is itself a list. For example, suppose for variable `quest1` you have the following possible levels `"no answer", NA, "negative", "neutral", "positive"` and you want to collect `("no answer", NA, "neutral")` into a single group, then the `grouping` argument should contain `list(quest1=list(c("no`

answer", NA, "neutral"))). Or if you have a discrete variable elements with values 1:10 and you want to collect it into groups "1:3,NA", "4", "5:9", "10" you specify in grouping the following list list(elements=list(c(1:3,NA), 5:9)). Values not defined in the grouping are left as they are. If cutpoints and groupings are defined for the same variable, the groupings take precedence and the corresponding cutpoints are set to NULL.

verbose: a number greater or equal to 0. The higher, the more info are provided during the execution of the algorithm.

If eval.imbalance = TRUE (the default), cem$imbalance contains the imbalance measure by absolute difference in means for numerical variables and chi-square distance for categorical variables. If FALSE then cem$imbalance is set to NULL. If data contains missing data, the imbalance measures are not calculated.

If L1.breaks is missing, the default rule to calculate cutpoints is the Scott's rule.

If k2k is set to TRUE, the algorithm return strata with the same number of treated and control units per stratum, otherwise all the matched units are returned (default). When k2k = TRUE, the user can choose a method (between 'euclidean', 'maximum', 'manhattan', 'canberra', 'binary' and 'minkowski') for nearest neighbor matching inside each cem strata. By default method is set to 'NULL', which means random matching inside cem strata. For the Minkowski distance the power can be specified via the argument mpower'. For more information on method != NULL, refer to dist help page.

By default, cem treats missing values as distinct categories and matches observations with missing values in the same variable in the same stratum provided that all the remaining (corasened) covariates match.

If argument data is non-NULL and datalist is NULL, CEM is applied to the single data set in data.

Argument datalist is a list of (multiply imputed) data frames (i.e., with missing cell values imputed). If data is NULL, the function cem is applied independently to each element of the list, resulting in separately matched data sets with different numbers of treated and control units.

When data and datalist are both non-NULL, each multiply imputed observation is assigned to the stratum in which it has been matched most frequently. In this case, the algorithm outputs the same matching solution for each multiply imputed data set (i.e., an observation, and the number of treated and control units matched, in one data set has the same meaning in all, and is the same for all)

Value

Returns an object of class cem.match if only data is not NULL or an object of class cem.match.list, which is a list of objects of class cem.match plus a field called unique which is true only if data and datalist are not both NULL. A cem.match object is a list with the following slots:

call	the call
strata	vector of stratum number in which each observation belongs, NA if the observation has not been matched
n.strata	number of strata generated
vars	report variables names used for the match
drop	variables removed from the match
breaks	named list of cutpoints, eventually NULL
treatment	name of the treatment variable
groups	factor, each observation belong to one group generated by the treatment variable
n.groups	number of groups identified by the treatment variable
group.idx	named list, index of observations belonging to each group
group.len	sizes of groups
tab	summary table of matched by group
imbalance	NULL or a vector of imbalances. See Details.

Author(s)

Stefano Iacus, Gary King, and Giuseppe Porro

References

Stefano Iacus, Gary King, Giuseppe Porro, "Matching for Casual Inference Without Balance Checking: Coarsened Exact Matching," http://gking.harvard.edu/files/abs/cem-abs.shtml

Examples

```
data(LL)


todrop <- c("treated","re78")

imbalance(LL$treated, LL, drop=todrop)

# cem match: automatic bin choice
mat <- cem(treatment="treated", data=LL, drop="re78")
mat

# cem match: user choiced coarsening
```

```
re74cut <- hist(LL$re74, br=seq(0,max(LL$re74)+1000, by=1000),plot=FALSE)$breaks
re75cut <- hist(LL$re75, br=seq(0,max(LL$re75)+1000, by=1000),plot=FALSE)$breaks
agecut <- hist(LL$age, br=seq(15,55, length=14),plot=FALSE)$breaks
mycp <- list(re75=re75cut, re74=re74cut, age=agecut)
mat <- cem(treatment="treated",data=LL, drop="re78",cutpoints=mycp)
mat


# cem match: user choiced coarsening, k-to-k matching
mat <- cem(treatment="treated",data=LL, drop="re78",cutpoints=mycp,k2k=TRUE)
mat

# mahalnobis matching: we use MatchIt
if(require(MatchIt)){
mah <- matchit(treated~age+education+re74+re75+black+hispanic+nodegree+married+u74+u75,
   distance="mahalanobis", data=LL)
mah
#imbalance
imbalance(LL$treated, LL, drop=todrop, weights=mah$weights)
}

# Multiply Imputed data
# making use of Amelia for multiple imputation
if(require(Amelia)){
 data(LL)
 n <- dim(LL)[1]
 k <- dim(LL)[2]

 set.seed(123)

 LL1 <- LL
 idx <- sample(1:n, .3*n)
 invisible(sapply(idx, function(x) LL1[x,sample(2:k,1)] <<- NA))

 imputed <- amelia(LL1,noms=c("black","hispanic","treated","married",
                              "nodegree","u74","u75"))
 imputed <- imputed$imputations[1:5]
# without information on which observation has missing values
 mat1 <- cem("treated", datalist=imputed, drop="re78")
 mat1

# ATT estimation
 out <- att(mat1, re78 ~ treated, data=imputed)
```

```
# with information about missingness
 mat2 <- cem("treated", datalist=imputed, drop="re78", data=LL1)
 mat2

# ATT estimation
 out <- att(mat2, re78 ~ treated, data=imputed)
}
```

4.2 att: Example of ATT estimation from CEM output

Description

An example of ATT estimation from CEM output

Usage

```
att(obj, formula, data, model="linear", extrapolate=FALSE, ntree=2000)
## S3 method for class 'cem.att':
plot(x, obj, data, vars=NULL, plot=TRUE, ecolors, ...)
## S3 method for class 'cem.att':
summary(object, ...)
```

Arguments

obj	a cem.match or cem.match.list object
formula	a model formula. See Details.
data	a single data.frame or a list of data.frame's in case of cem.match.list
model	one model. See Details.
extrapolate	extrapolate the CEM restriced estimate to the whole data. Default = FALSE.
ntree	number of trees to generate in random forest model. Default = 2000.
x	the output from the att function
vars	a vector of variable names to be used in the parallel plots. By default all variables involved in data matching are used.
object	an object of class cem.att function
plot	if TRUE the plot is produced, otherwise only calculations are made.
ecolors	a vector of three colors respectively for positive, zero and negative treatment effect. Default c("blue","black","red").
...	passed to the plot function or to printCoefmat for the method summary

Details

Argument model can be lm, linear for linear regression model; logit for the the logistic model; lme, linear-RE for the linear model with random effects. Also rf, forest for the randomforest algorithm.

If the outcome is y and the treatment variable is T, then a formula like y ~ T will produce the simplest estimate the ATT: with lm, it is just the coefficient on T, which is the same as the difference in means, weighted by CEM stratum size. Users can add covariates to

span any remaining imbalance after the match, such as `y ~ T + age + sex`, to adjust for variables `age` and `sex`.

In the case of multiply imputed datasets, the model is applied to each single matched data and the ATT and is the standard error estimated using the standard formulas for combining results of multiply imputed data.

When `extrapolate` = `TRUE`, the estimate model is extrapolated to the whole set of data.

There is a `print` method for the output of `att`. Specifying the option `TRUE` in a `print` command gives complete output from the estimated model when availalble.

Value

A matrix of estimates with their standard error, or a list in the case of `cem.match.list`. For `plot.att` a list of strata estimated treatment effect and group ("positive", "negative", "zero") and individual treatment and effect and group. The individual treatment effect and group is given by the treatment effect of the strata. Similarly for the group ("positive", "negative", "zero"). Also, colors associated to estimated treatment effects are returned for easy subsequent plotting.

Author(s)

Stefano Iacus, Gary King, and Giuseppe Porro

References

Stefano Iacus, Gary King, Giuseppe Porro, "Matching for Casual Inference Without Balance Checking: Coarsened Exact Matching," http://gking.harvard.edu/files/abs/cem-abs.shtml

Examples

```
data(LL)

# cem match: automatic bin choice
mat <- cem(treatment="treated",data=LL, drop="re78")
mat
mat$k2k

# ATT estimate
homo1 <- att(mat, re78~treated,  data=LL)
rand1 <- att(mat, re78~treated,  data=LL, model="linear-RE")
rf1 <- att(mat, re78~treated,  data=LL, model="rf")

homo2 <- att(mat, re78~treated,  data=LL, extra=TRUE)
```

```
rand2 <- att(mat, re78~treated,  data=LL, model="linear-RE", extra=TRUE)
rf2 <- att(mat, re78~treated,  data=LL, model="rf", extra=TRUE)

homo1
summary(homo1)

rand1
rf1

homo2
rand2
rf2

plot( homo1, mat, LL, vars=c("age","education","re74","re75"))
plot( rand1, mat, LL, vars=c("age","education","re74","re75"))
plot( rf1, mat, LL, vars=c("age","education","re74","re75"))

plot( homo2, mat, LL, vars=c("age","education","re74","re75"))
plot( rand2, mat, LL, vars=c("age","education","re74","re75"))
plot( rf2, mat, LL, vars=c("age","education","re74","re75"))


# reduce the match into k2k using euclidean distance within cem strata
mat2 <- k2k(mat, LL, "euclidean", 1)
mat2
mat2$k2k

# ATT estimate after k2k
att(mat2, re78~treated, data=LL)

# example with missing data
# using multiply imputated data
# we use Amelia for multiple imputation
if(require(Amelia)){
 data(LL)
 n <- dim(LL)[1]
 k <- dim(LL)[2]

# we generate missing values in 30% of the rows of LL data
# randomly in one colum per row
 LL1 <- LL
 idx <- sample(1:n, .3*n)
 invisible(sapply(idx, function(x) LL1[x,sample(2:k,1)] <<- NA))
```

```
 imputed <- amelia(LL1)
 imputed <- imputed$imputations[1:5]

 mat <- cem("treated", datalist=imputed, data=LL1, drop="re78")

 print(mat)

 att(mat, re78 ~ treated, data=imputed)
}
```

4.3 `cemspace`: Exploration tool for CEM

Description

Exploration tool for CEM

Usage

```
cemspace(treatment=NULL, data = NULL, R=100, grouping = NULL, drop=NULL,
L1.breaks = NULL, plot = TRUE, fixed = NULL, minimal = 1, maximal = 5,
M=250, raw.profile=NULL)
```

Arguments

`treatment`	character, name of the treatment variable.
`data`	a data.frame.
`R`	number of possible random coarsening for the CEM.
`grouping`	named list, each element of which is a list of groupings for a single categorical variable. For more details see `cem`.
`drop`	a vector of variable names in the data frame to ignore during matching
`L1.breaks`	list of cutpoints for the calculation of the L1 measure.
`plot`	plot the space of solutions?
`fixed`	vector of variable names which will not be relaxed.
`minimal`	the minimal number of intervals acceptable after relaxation. Should be a nameed list of positive integers or if a number, this is applied to all variables.
`maximal`	the maximal number of intervals acceptable after relaxation. Should be a nameed list of positive integers or if a number, this is applied to all variables.
`M`	number of possible random coarsening for the L1 measure
`raw.profile`	and object of class `L1profile`. If passed, the `L1.breaks` are ignored.

Details

This is a tool to help the user to explore different cem solutions by choosing random coarsenings. The algorithm tries `R` random choiches of coarsenings into intervals between `minimal` and `maximal` for numerical, integer or ordered factors. It drops or include dichotomous or boolean variables.

Calling directly `plot` on the output of `cemspace` has the same effect of calling directly `imbspace.plot`.

If you want to relax a given cem solution, use the function `imbspace` instead.

Value

`val` an invisible object of class `imbalance.space`.

Author(s)

Stefano Iacus, Gary King, and Giuseppe Porro

References

Stefano Iacus, Gary King, Giuseppe Porro, "Matching for Casual Inference Without Balance Checking: Coarsened Exact Matching," http://gking.harvard.edu/files/abs/cem-abs.shtml

See Also

`imbspace.plot`, `cemspace`

Examples

```
## Not run:
data(LL)
tmp <- cemspace("treated", LL, drop="re78", M=50)

## End(Not run)
```

4.4 DW: Dehejia-Wahba dataset

Description

A subset of the Lalonde dataset (see cited reference).

Usage

```
data(DW)
```

Format

A data frame with 445 observations on the following 10 variables.

`treated` treated variable indicator

`age` age

`education` years of education

`black` race indicator variable

`married` marital status indicator variable

`nodegree` indicator variable of not possessing a degree

`re74` real earnings in 1974

`re75` real earnings in 1975

`re78` real earnings in 1978 (post treatment outcome)

`hispanic` ethnic indicator variable

`u74` unemployment in 1974 indicator variable

`u75` unemployment in 1975 indicator variable

Source

see references

References

Dehejia, R., Wahba, S. (1999) "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94, 1053-1062.

4.5 imbalance: Calculates several imbalance measures

Description

Calculates several imbalance measures for the original and matched data sets

Usage

```
imbalance(group, data, drop=NULL, breaks = NULL, weights)
```

Arguments

group	the group variable
data	the data
drop	a vector of variable names in the data frame to ignore
breaks	a list of vectors of cutpoints used to calculate the L1 measure. See Details.
weights	weights

Details

This function calculates several imbalance measures. For numeric variables, the difference in means (under the column statistic), the difference in quantiles and the L1 measure is calculated. For categorical variables the L1 measure and the Chi-squared distance (under column statistic) is calculated. Column type reports either (diff) or (Chi2) to indicate the type of statistic being calculated.

If breaks is not specified, the Scott automated bin calculation is used (which coarsens less than Sturges, which used in cem). Please refer to cem help page. In this case, breaks are used to calculate the L1 measure.

This function also calculate the global L1 imbalance measure. If breaks is missing, the default rule to calculate cutpoints is the Scott's rule. See L1.meas help page for details.

Value

An object of class imbalance which is a list with the following two elements

tab	Table of imbalance measures
L1	The global L1 measure of imbalance

Author(s)

Stefano Iacus, Gary King, and Giuseppe Porro

References

Stefano Iacus, Gary King, Giuseppe Porro, "Matching for Casual Inference Without Balance Checking: Coarsened Exact Matching," http://gking.harvard.edu/files/abs/cem-abs.shtml

Examples

```
data(LL)

todrop <- c("treated","re78")

imbalance(LL$treated, LL, drop=todrop)

# cem match: automatic bin choice
mat <- cem(treatment="treated", data=LL, drop="re78")
```

4.6 imbspace: Diagnostic tool for CEM

Description

Diagnostic tools for CEM

Usage

```
imbspace(obj, data, depth = 1, L1.breaks = NULL,
plot = TRUE, fixed = NULL, minimal = 1, maximal = 6,
M=250, raw.profile=NULL)
```

Arguments

obj	an object of class cem.match
data	the original data.
depth	if 1, relaxes up to dropping one var, if 2 relaxes (up to dropping) two vars, etc.
L1.breaks	list of cutpoints for the calculation of the L1 measure.
plot	plot the space of solutions?
fixed	vector of variable names which will not be relaxed.
minimal	the minimal number of intervals acceptable after relaxation. Should be a nameed list of positive integers or if a number, this is applied to all variables.
maximal	the maximal number of intervals acceptable after relaxation. Should be a nameed list of positive integers or if a number, this is applied to all variables.
M	number of possible random coarsening for the L1 measure
raw.profile	and object of class L1profile. If passed, the L1.breaks are ignored.

Details

This is a diagnostic tool to help the user in the search of different choices of coarsenings. The algorithm tries all possible combination of coarsenings into intervals between minimal and maximal one variable at time, for pairs, triplets, etc depending on the value of depth.

Calling directly plot on the output of imbspace has the same effect of calling directly imbspace.plot.

Value

val	an invisible object of class imbalance.space.

Author(s)

Stefano Iacus, Gary King, and Giuseppe Porro

References

Stefano Iacus, Gary King, Giuseppe Porro, "Matching for Casual Inference Without Balance Checking: Coarsened Exact Matching," http://gking.harvard.edu/files/abs/cem-abs.shtml

See Also

`imbspace.plot`

4.7 k2k: Reduction to k2k Matching

Description

Reduces a CEM output to a k2k matching

Usage

```
k2k(obj, data, method=NULL, mpower=2, verbose=0)
```

Arguments

obj	an object as output from cem
data	the original data.frame used by cem
method	distance method to use in k2k matching. See Details.
mpower	power of the Minkowski distance. See Details.
verbose	controls level of verbosity. Default=0.

Details

This function transforms a typical cem matching solution to a k-to-k match, with k variable along strata: i.e., in each stratum generated by cem, the match is reduce to have the same number of treated and control units. (This option will delete some data that matched well, and thus likely increase the variance, but it means that subsequent analyses do not require weights.)

The user can choose a method (between 'euclidean', 'maximum', 'manhattan', 'canberra', 'binary' and 'minkowski') for nearest neighbor matching inside each cem strata. By default method is set to 'NULL', which means random matching inside cem strata. For the Minkowski distance the power can be specified via the argument mpower'. For more information on method != NULL, refer to dist help page.

After k2k the weights of each matched observation are set to unity.

Value

obj	an object of class cem.match

Author(s)

Stefano Iacus, Gary King, and Giuseppe Porro

References

Stefano Iacus, Gary King, Giuseppe Porro, "Matching for Casual Inference Without Balance Checking: Coarsened Exact Matching," http://gking.harvard.edu/files/abs/cem-abs.shtml

Examples

```
data(LL)

# cem match: automatic bin choice
mat <- cem(treatment="treated", data=LL, drop="re78")
mat
mat$k2k

# ATT estimate
att(mat, re78 ~ treated, data=LL)


# transform the match into k2k
mat2 <- k2k(mat, LL, "euclidean", 1)
mat2
mat2$k2k

# ATT estimate after k2k
att(mat2, re78 ~ treated, data=LL)
```

4.8 L1.meas: Evaluates L1 distance between multidimensional histograms

Description

Evaluates L1 distance between multidimensional histograms

Usage

```
L1.meas(group, data, drop=NULL, breaks = NULL, weights)
```

Arguments

group	the group variable
data	the data
drop	a vector of variable names in the data frame to ignore
breaks	a list of vectors of cutpoints; if not specified, automatic choice will be made
weights	weights

Details

This function calculates the L1 distance on the k-dimensional histogram in order to measure the level of imbalance in a matching solution.

If `breaks` is not specified, the Scott automated bin calculation is used (which coarsens less than Sturges, which used in `cem`). Please refer to `cem` help page. In this case, breaks are used to calculate the L1 measure.

When choosing `breaks` for L1, a very fine coarsening (many cut points) produces values of L1 close to 1. A very mild coarsening (very fex cutpoints), is not able to discriminate, i.e. L1 close to 0 (particularly true when the number of observations is small with respect to the number of continuous variables). The `L1.profile` function shows how to compare matching solutions for any level of (i.e., without regard to) coarsening.

This code also calculate the Local Common Support (LCS) measure, which is the proportion of non empty k-dimensional cells of the histogram which contain at least one observation per group.

Value

An object of class `L1.meas` which is a list with the following fields

L1	The numerical value of the L1 measure
breaks	A list of cutpoints used to calculate the L1 measure
LCS	The numerical value of the Local Common Support proportion

Author(s)

Stefano Iacus, Gary King, and Giuseppe Porro

References

Stefano Iacus, Gary King, Giuseppe Porro, "Matching for Casual Inference Without Balance Checking: Coarsened Exact Matching," http://gking.harvard.edu/files/abs/cem-abs.shtml

Examples

```
data(LL)
L1.meas(LL$treated,LL, drop=c("treated","re78"))
```

4.9 `L1.profile`: Calculates L1 distance for different coarsenings

Description

Calculates L1 distance for different coarsenings

Usage

```
L1.profile(group, data, drop = NULL, min.cut = 2, max.cut = 12,
weights, plot = TRUE, add = FALSE, col = "red",
lty = 1, M=100, useCP=NULL)
```

Arguments

`group`	the group variable
`data`	the data
`drop`	a vector of variable names in the data frame to ignore
`min.cut`	minimum number of cut points per variable
`max.cut`	maximum number of cut points per variable
`weights`	weights
`useCP`	a list which elements is a list of cutpoints, usually passed from a previous instance of `L1.profile`. If not `NULL` these coarsenings are used instead of generating them randomly.
`M`	number of random coarsenings
`plot`	plot a graph?
`add`	add graph to an existing plot? Makes sense only if `plot` is `TRUE`
`col`	draw in specified color
`lty`	draw using specified lty

Details

The L1 measure depends on the coarsening chosen to calculate it, and as such the comparison of different matching solutions may differ depending on this somewhat arbitrary choice. This function computes L1 for a random range of possible coarsenings. The point of this function is that if one matching solution has a lower L1 than another, then it dominates without regard to the choice of coarsening. A graphic display conveys the results succinctly. (The logic is similar to that for ROC curves used for classification algorithms.) (This degree of coarsening should remain fixed for different CEM runs.)

For each variables the function generates a random number of cutpoints between `min.cut` and `max.cut` in which to cut the support of each variable. This procedure is repeated `M` times. The out is sorted in increasing values of L1 just for graphical representation.

Non numeric variables are not grouped, i.e. no coarsening occurs.

A `plot` method exists for the returned object.

Value

An invisible object of class `L1profile` which contains a named list of coarsenings and values of the L1 measure for each coarsening.

Author(s)

Stefano Iacus, Gary King, and Giuseppe Porro

References

Stefano Iacus, Gary King, Giuseppe Porro, "Matching for Casual Inference Without Balance Checking: Coarsened Exact Matching," http://gking.harvard.edu/files/abs/cem-abs.shtml

Examples

```
## Not run:
data(LL)
for(i in c(4:6,10:12))
 LL[[i]] <- factor(LL[[i]])

imb0 <- L1.profile(LL$treated,LL, drop=c("treated","re78"))

if(require(MatchIt)){
 m2 <- matchit(treated ~ black + hispanic + married + nodegree + u74 + u75 + education +
  age + re74 + re75, data=LL, distance="logit")

 m3 <- matchit(treated ~ black + hispanic + married + nodegree + u74 + u75 + education +
  age + re74 + re75, data=LL, distance="mahalanobis")

 L1.profile(LL$treated,LL, drop=c("treated","re78"),
  weights=m2$w, add=TRUE, col="green", lty=2, useCP=imb0$CP)

 L1.profile(LL$treated,LL, drop=c("treated","re78"),
  weights=m3$w, add=TRUE, col="orange", lty=3, useCP=imb0$CP)
}

m1 <- cem("treated", LL, drop="re78")

L1.profile(LL$treated,LL, drop=c("treated","re78"),
 weights=m1$w>0, add=TRUE, col="blue", lty=4, useCP=imb0$CP)
```

```
legend(5, 0.9, legend=c("raw data", "pscore", "mahalanobis", "cem"), lty=1:4, col=c("red", "

## End(Not run)
```

4.10 LL: Lalonde dataset

Description

Lalonde experimental dataset (see cited reference).

Usage

```
data(LL)
```

Format

A data frame with 722 observations on the following 10 variables.

`treated` treatment variable indicator

`age` age

`education` years of education

`black` race indicator variable

`married` marital status indicator variable

`nodegree` indicator variable for not possessing a degree

`re74` real earnings in 1974

`re75` real earnings in 1975

`re78` real earnings in 1978 (post-treatment outcome)

`hispanic` ethnic indicator variable

`u74` unemployment in 1974 indicator variable

`u75` unemployment in 1975 indicator variable

Source

see references

References

Lalonde, R. (1986) "Evaluating the Econometric Evaluations of Training Programs," *American Economic Review*, 76, 604-620.

4.11 `LeLonde`: Modified Lalonde dataset

Description

This is a modified version of the Lalonde experimental dataset used for explanatory reasons only.

Usage

```
data(LL)
```

Format

A data frame with 722 observations on the following 11 variables.

`treated` treatment variable indicator

`age` age

`education` years of education

`black` race indicator variable

`married` marital status indicator variable

`nodegree` indicator variable for not possessing a degree

`re74` real earnings in 1974

`re75` real earnings in 1975

`re78` real earnings in 1978 (post-treatment outcome)

`hispanic` ethnic indicator variable

`u74` unemployment in 1974 indicator variable

`u75` unemployment in 1975 indicator variable

`q1` answer to survey question n1

Details

This data is a copy of the original Lalonde (1986) data set (see `LL`) with 10% of missing data and an additional variable `q1` which is the fictituous answer to the questionarie on "Agreement on this job training program".

Source

see references

References

Lalonde, R. (1986) "Evaluating the Econometric Evaluations of Training Programs," *American Economic Review*, 76, 604-620.

4.12 pair: Produces a paired sample out of a CEM match solution

Description

Produces a paired sample out of a CEM match solution

Usage

```
pair(obj, data, method=NULL, mpower=2, verbose=0)
```

Arguments

obj	an object as output from cem
data	the original data.frame used by cem
method	distance method to use in k2k matching. See Details.
mpower	power of the Minkowski distance. See Details.
verbose	controls level of verbosity. Default=0.

Details

This function returns a vector of paired matched units index.

The user can choose a method (between 'euclidean', 'maximum', 'manhattan', 'canberra', 'binary' and 'minkowski') for nearest neighbor matching inside each cem strata. By default method is set to 'NULL', which means random matching inside cem strata. For the Minkowski distance the power can be specified via the argument mpower'. For more information on method != NULL, refer to dist help page.

Value

obj	a list with the fields paired, full.paired, reservoir and reservoir2. The latter contain the indexes of the unmatched units.

Author(s)

Stefano Iacus, Gary King, and Giuseppe Porro

References

Stefano Iacus, Gary King, Giuseppe Porro, "Matching for Casual Inference Without Balance Checking: Coarsened Exact Matching," http://gking.harvard.edu/files/abs/cem-abs.shtml

Examples

```
data(LL)

# cem match: automatic bin choice
mat <- cem(data=LL, drop="re78")

# we want a set of paired units
psample <- pair(mat, data=LL)
table(psample$paired)
psample$paired[1:100]

table(psample$full.paired)
psample$full.paired[1:10]



# cem match: automatic bin choice, we drop one row from the data set
mat1 <- cem(data=LL[-1,], drop="re78")

# we want a set of paired units but we have an odd number of units in the data
psample <- pair(mat1, data=LL[-1,])
table(psample$full.paired)
```

4.13 relax.cem: Diagnostic tool for CEM

Description

Diagnostic tools for CEM

Usage

```
relax.cem(obj, data, depth=1, verbose = 1, L1.breaks=NULL, plot=TRUE, fixed=NULL,
 shifts=NULL, minimal=NULL, use.coarsened=TRUE, eval.imbalance=TRUE, ...)
relax.plot(tab, group="1", max.terms=50, perc=.5, unique=FALSE, colors=TRUE)
```

Arguments

obj	an object of class cem.
data	the original data.
verbose	controls the level of verbosity.
L1.breaks	list of cutpoints for the calculation of the L1 measure.
plot	plot the solutions?
tab	the output table from relax.cem.
fixed	vector of variable names which will not be relaxed.
max.terms	plot only the last best results of relax.cem.
shifts	a vector of proportions of shifts.
minimal	the minimal number of intervals acceptable after relaxation. Should be a nameed list of positive integers.
group	character string denoting group id. Defaults to "1".
perc	only plot if percentage of matched units is greater than perc.
unique	only plot different solutions (in terms of matched units).
depth	if 1, relaxes up to dropping one var, if 2 relaxes (up to dropping) two vars, etc.
use.coarsened	used coarsened values for continuous variables.
colors	If TRUE each variable is plotted in a different colour.
eval.imbalance	If TRUE L1 measure is evaluated at each iteration.
...	passed to the relax.plot function.

Details

relax.cem starts from a cem solution (as given by cem) and tries several relaxed coarsenings on the variables. Coarsenings corresponds to dividing the support of each variable into a decreasing number of intervals of the same length (even if in the starting solution intervals are of different lengths). Because CEM is MIB, the number of matched units increases as the number of intervals decrease. All variables are coarsened into k intervals along a sequence which starts from the original number of intervals and decreases to 10 intervals by 2, then continues from 10 down to 1 intervals by 1. If minimal is specified, variables are coarsened down to that minimal value.

To observe MIB property of CEM use.coarsened (default) should be set to TRUE; otherwise the coarsening of the continuous variable will be recalculated at each iteration and there is no guarantee of monotonicity.

relax.cem outputs a list of tables. Each table is named Ggroup where group is the id of the group. Each Ggroup table is ordered in increasing order of matched units of group group. Columns PercGgroup and Ggroup report percentage and absolute number of matched units for each group. Column Relaxed indicates which relaxation has been done, with something like "V1(4), V3(5)", which means "variable V1 has been split in 4 intervals of the same length and variable V3 into five intervals". Thus, the number of intervals is reported in parenthases and if equal to 1 means that the corresponding variable is excluded from affecting the match (i.e. all observations are assigned to the same interval).

If shifts is not null, each coarsening is shifted accordingly (see shift.cem for additional details). In case of shifting "S:" appears in the labels.

The relax.plot, plot all the different relaxation in increasing order of number of treated units matched. For each coarsening it also reports the value of the L1 measure. The table generated by relax.cem may contain many entries. By default, only a portion of best coarsenings are plotted (option max.terms). In addition, the user can specify to plot the corasening for which at least a certain percentage of treated units have been matched (option perc, by default 50 In addition, of several different coarsenings which lead to the same number of treated units matched, the user can specify to plot only one of them using the option unique = TRUE (default).

If L1.breaks are NULL they are taken from the cem object if available or calculated atumatically as in cem.

Calling directly plot on the output of cem.relax has the same effect of calling directly relax.plot.

Value

tab	an invisible object containing the tabs and the L1breaks used

Author(s)

Stefano Iacus, Gary King, and Giuseppe Porro

References

Stefano Iacus, Gary King, Giuseppe Porro, "Matching for Casual Inference Without Balance Checking: Coarsened Exact Matching," http://gking.harvard.edu/files/abs/cem-abs.shtml

See Also

cem

Examples

```
## Not run:
data(LL)

mat <- cem(treatment="treated",data=LL, drop="re78")
mat
tab <- relax.cem(mat, LL, depth=1, plot=FALSE)

relax.plot(tab, group="1")
plot(tab, group="1")
relax.plot(tab, group="1", unique=TRUE)
relax.plot(tab, group="1", perc=0.6)
relax.plot(tab, group="1", perc=0.6,unique=TRUE)


tab1 <- relax.cem(mat, LL, depth=1, minimal=list(re74=6, age=3, education=3, re75=5))
tab2 <- relax.cem(mat, LL, depth=1, minimal=list(re74=6, age=3,
                  education=3, re75=5), shifts=0.01)
tab3 <- relax.cem(mat, LL, depth=1, minimal=list(age=3, education=3),
                  fixed=c("re74","re75"))
tab4 <- relax.cem(mat, LL, depth=2, minimal=list(age=4,
                   education=3,re75=6),plot=FALSE, fixed="re74")
relax.plot(tab4)
relax.plot(tab4, unique=TRUE)
relax.plot(tab4, perc=0.7)

## End(Not run)
```

4.14 `shift.cem`: Diagnostic tool for CEM

Description

Diagnostic tools for CEM. Applies leftward and rightward shifts of the cutpoints.

Usage

```
shift.cem(obj, data, shifts=NULL, verbose=0, plot=TRUE)
```

Arguments

`obj`	and object of class cem
`data`	the original data
`shifts`	a vector of proportions of shifts
`verbose`	controls the level of verbosity
`plot`	whether to plot a graphic representation of the search

Details

For each variable, shift all the cutpoints left and right by `shifts` times the smallest epsilon of the coarsening. Shifting to the right produces a new cell on the left; shift to the left, adds a new cell to the coarsening on the right. Only positive proportions should be used; the algorithm will produce shifting on the left or on the right. The best shifting of the original cem match is produced as output, where best is defined in terms of the maximal total number of matched units `mT+mC` (see below).

By default, the function returns minimal information about the execution of the algorithm. By setting a value greater than 0 in option `verbose` more feedback on the process is returned.

Option `plot` = `TRUE` plots the number of treated units matched `mT`, the number of control units matched `mC`, and the sum `mT+mC`, as a function of the shifts.

Value

`tab`	an invisible object containing a new cem object

Author(s)

Stefano Iacus, Gary King, and Giuseppe Porro

References

Stefano Iacus, Gary King, Giuseppe Porro, "Matching for Casual Inference Without Balance Checking: Coarsened Exact Matching," http://gking.harvard.edu/files/abs/cem-abs.shtml

See Also

cem

Examples

```
## Not run:
data(LL)

m74 <- max(LL$re74, na.rm=TRUE)
s74 <- seq(0,m74,by=sd(LL$re74))
l74 <- length(s74)
if(max(s74) < m74) s74 <- c(s74, m74)

m75 <- max(LL$re75, na.rm=TRUE)
s75 <- seq(0,m75,by=sd(LL$re75))
l75 <- length(s75)
if(max(s75) < m75) s75 <- c(s75, m75)

mybr = list(re74=s74,
 re75 = s75,
 age = hist(LL$age,plot=FALSE)$breaks,
 education = hist(LL$education,plot=FALSE)$breaks)

mat <- cem(treatment="treated",data=LL, drop="re78",cut=mybr)
mat

shift.cem(mat, data=LL, shifts=seq(0.01, 0.5, length=10), verb=1)

## End(Not run)
```

References

Ho, Daniel E., Kosuke Imai, Gary King and Elizabeth A. Stuart. Forthcoming. "MatchIt: Nonparametric Preprocessing for Parametric Causal Inference." *Journal of Statistical Software* . http://gking.harvard.edu/matchit.

Ho, Daniel, Kosuke Imai, Gary King and Elizabeth Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15:199–236. http://gking.harvard.edu/files/abs/matchp-abs.shtml.

Honaker, James, Gary King and Matthew Blackwell. 2006. "Amelia II: A Program for Missing Data.". http://gking.harvard.edu/amelia.

King, Gary, James Honaker, Anne Joseph and Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." *American Political Science Review* 95(1):49–69. http://gking.harvard.edu/files/abs/evil-abs.shtml.

King, Gary and Langche Zeng. 2006. "The Dangers of Extreme Counterfactuals." *Political Analysis* 14(2):131–159. http://gking.harvard.edu/files/abs/counterft-abs.shtml.

Lalonde, Robert. 1986. "Evaluating the Econometric Evaluations of Training Programs." *American Economic Review* 76:604–620.