Journal of Statistical Software

November 2012, Volume 51, Code Snippet 1. *http://www.jstatsoft.org/*

%GI: A SAS Macro for Measuring and Testing Global Imbalance of Covariates within Subgroups

Furio Camillo
University of Bologna

Ida D'Attoma
University of Bologna

Abstract

The global imbalance (GI) measure is a way for checking balance of baseline covariates that confound efforts to draw valid conclusions about treatment effects on outcomes of interest. In addition, GI is tested by means of a multivariate test. The GI measure and its test overcome some limitations of the common way for assessing the presence of imbalance in observed covariates that were discussed in D'Attoma and Camillo (2011). A user written SAS macro called %GI, to simultaneously measure and test global imbalance of baseline covariates is described. Furthermore, %GI also assesses global imbalance by subgroups obtained through several matching or classification methods (e.g., cluster analysis, propensity score subclassification, Rosenbaum and Rubin 1984), no matter how many groups are examined. %GI works with mixed categorical, ordinal and continuous covariates. Continuous baseline covariates need to be split into categories. It also works in the multi-treatment case. The use of the %GI macro will be illustrated using two artificial examples.

Keywords: global imbalance measure, global imbalance test, subgroups, multi-treatment, SAS.

1. Introduction

Assessing balance of non-equivalent groups is fundamental before estimating effects of treatments on outcomes of interest, especially in the presence of observational data where the rule that governs treatment assignment is generally unknown, and either units are self-selected into treatments or they are non randomly selected to receive a treatment. Various methods are used to balance groups with unequal distribution of covariates – i.e., matching, cluster analysis, propensity score (PS) adjustments. The most widely used and applied in various fields is the PS adjustment (Rosenbaum and Rubin 1983). PS is the conditional probability that a unit will be assigned to the treatment condition based on a set of observed covariates. Then, propensity score adjustments (e.g., PS subclassification) are used to balance groups

with unequal distributions of covariates. Another method is cluster analysis that balances unequal group distributions by stratifying on clusters based on several covariates (Camillo and D'Attoma 2010; D'Attoma and Camillo 2011; Peck 2005). Unlike PS, cluster analysis does not create a single aggregate score, permitting each covariate to maintain its functional form. In conjunction with the increasing use of methods to balance non-equivalent groups different criteria for checking balance have been proposed (Rosenbaum and Rubin 1984; Rubin 2001; Baser 2006), but most of them assess separately balance variable-by-variable. In this paper a macro that allows to simultaneously measure and test imbalance of a set of baseline covariates is provided. The macro computes and tests the global imbalance (GI) measure introduced in D'Attoma and Camillo (2011), mainly based on the concept of between-groups inertia of a factorial predictor space. According to it, perfect balance occurs when the between-groups inertia equals zero; whereas, perfect imbalance occurs when the between-groups inertia equals the total inertia (I_T), which indicates that the observed total variability of the $\mathbf{X}$-space is completely due to the selection mechanism. Thus, the proposed GI measure in data varies in $[0, I_T]$. A SAS macro `%GI` is described that measures and tests global imbalance on subgroups. The macro mainly uses the SAS/IML language. The final summary about balance is saved in a SAS dataset in the directory specified by the user. This paper after a review of several tools that address the balance checking problem, briefly introduces the GI measure and its related test, then describes the macro -its arguments, implementation and the output dataset it produces- and finally presents two examples demonstrating the use of the macro for assessing the global imbalance of a set of baseline covariates by subgroups: One for the binary treatment case and another for the multi-treatment case.

2. Measuring balance: A review

The success of various methods in reducing bias of the estimated effects mainly depends on the balance criterion adopted. According to Rubin (2001), balance concerns similarity in covariate distributions across treatment groups. As reported in Ho, Imai, King, and Stuart (2007) it holds when the treatment ($\mathbf{T}$) and the covariates ($\mathbf{X}$) are unrelated such that $\tilde{p}(\mathbf{X} \mid \mathbf{T} = 1) = \tilde{p}(\mathbf{X} \mid \mathbf{T} = 0)$; where $\tilde{p}$ denotes the observed empirical density of data. Balance is commonly evaluated by conducting hypothesis testing. The standard practice involves the use of t-test for the difference in means for each continuous covariate or the χ^2 test for each categorical covariate. However, this practice starts to be criticized (to cite few: (Imai, King, and Stuart 2008; Iacus, King, and Porro 2011). The main critique is that researchers used to ignore the multivariate balance. In recent works such multivariate aspect starts to be taken into account (Hansen and Bowers 2008; Li, Maasoumi, and Racine 2009; Camillo and D'Attoma 2010; D'Attoma and Camillo 2011; Iacus *et al.* 2011). Hansen and Bowers (2008) propose a simultaneous balance test on multiple $\mathbf{X}$. The hypothesis of no association between a treatment variable and the $\mathbf{X}$ covariates is assessed by comparing the differences of means (or regression coefficients), without standardization, to their distribution under hypothetical shuffles of the treatment variable, a permutation or randomization distribution (Bowers, Fredrickson, and Hansen 2011). This test balances not only on each X separately, but also on all linear combinations of them. Its law is a χ^2-approximation. The test is implemented within the **RItools** package and the `xBalance` function (Bowers *et al.* 2011). It should work also when the treatment variable is not binary, but it doesn't seem clear with which kind of covariates (categorical, continuous, ordinal) it works. Li *et al.* (2009) propose

a nonparametric test for equality of two multivariate densities with mixed categorical and continuous data. The test statistic, I_n (Li *et al.* 2009), is constructed based on the integrated squared density difference given by $I = \int [f(x)dF(x) + g(x)dG(x) - f(x)dG(x) - g(x)dF(x)]$ where $F(\cdot)$ and $G(\cdot)$ are the cumulative distribution function for X and Y. X and Y are the multivariate vectors of dimension $q+r$ where q denotes the number of continuous variables and r the number of discrete/categorical variables. Li *et al.* (2009) demonstrate that under the null of equality of distributions the I_n statistics can be approximate by a normal standard. The test is implemented within the R software using the **np** package (Hayfield and Racine 2008) and the `npdeneqtest` function (Li *et al.* 2009; Racine 2012). The test has the advantage of working with mixed categorical and continuous data. Iacus *et al.* (2011) propose a multivariate imbalance measure based on the L_1 difference between the multidimensional histogram of all pre-treatment covariates in the treatment group and that in the control group. To obtain the measure, they cross-tabulate the discretized variables and the categorical variables as $X_1 \times X_2 \ldots \times X_k$ for the treated and control groups separately, and record the k-dimensional relative frequencies for the treated $f_{l_1 \ldots l_k}$ and control $g_{l_1 \ldots l_k}$ units. Finally, they take the absolute difference over all the cell values:

$$L_1(f, g) = \frac{1}{2} \sum_{l_1 \ldots l_k} |f_{l_1 \ldots l_k} - g_{l_1 \ldots l_k}| \tag{1}$$

The L_1 measure is implemented within R using the **cem** package and the `imbalance` function (Iacus, King, and Porro 2009). It is also implemented in Stata using the **cem** package and the `imb` function (Blackwell, Iacus, and King 2009). An undoubted advantage of such a measure is its simplicity and intuitive interpretation. Furthermore, it should work with multicategory treatments and with any kind of variables.

3. Description of the GI measure and its related test

The present section provides a brief description of the GI measure and its related test. For a more comprehensive treatment of the theoretical framework within the GI measure and its related test are developed see Camillo and D'Attoma (2010) and D'Attoma and Camillo (2011), and for an application see Peck, Camillo, and D'Attoma (2010). The `%GI` macro uses the SAS/IML language to compute the GI measure expressed as

$$GI = \frac{1}{Q} \sum_{t=1}^{T} \sum_{j=1}^{J_Q} \frac{b_{tj}^2}{k_{.t} k_{.j}} - 1 \tag{2}$$

where Q denotes the number of baseline covariates, T denotes the number of treatment levels, J_Q denotes the set of all categories of the Q baseline variables, b_{tj} is the number of units with category $j \in J_Q$ in the treatment group $t \in T$, $k_{.t}$ is the group size $t \in T$, and $k_{.j}$ is the number of units with category $j \in J_Q$. The GI measure is the result of using the conditional multiple correspondence analysis (MCA) framework (Escofier 1988) to quantify the between groups inertia[1]. In fact, when the dependence among categorical baseline covariates ($\mathbf{X}$) and the treatment assignment ($\mathbf{T}$) is outside the control of researchers, displaying the relationship

[1]The term inertia is used by analogy with the definition in applied mathematics of moment of inertia which stands for the integral of mass times the squared distance to the centroid (Greenacre 1984).

among them on a factorial space represents a first step for discovering the hidden relationship. In the presence of dependence, any descriptive factorial analysis may exhibit this link. Commonly, the problem of the factorial decomposition of the variance related to the juxtaposition of the **X** matrix and **T** is faced within the MCA framework (Lebart, Morineau, and Warwick 1984). With reference to MCA, the structure of the data matrix eigenvectors and eigenvalues decomposition process, could be strongly influenced by the presence of an external conditioning variable (i.e., the treatment assignment **T**). Hence, a conditional analysis is used in order to isolate the part of the variability of the **X**-space due to **T**. With reference to the Huygens' inertia decomposition of total inertia (I_T) as within-groups (I_W) and between-groups (I_B), conditional MCA (Escofier 1988) consists in the factorial decomposition of the within-group inertia. In this sense, it could be also considered as an intra analysis that detects and describes differences among units within each group by not considering the effect due to the partition's structure induced by the non random selection process. The space generated by the conditional MCA is continuous, and thus, in the computation of distances between groups and between units, becomes possible to use a standard metric based on the criterion of the variance minimization. The key result of using conditional MCA is represented by the quantified between groups inertia that represents the measure of global imbalance in data (D'Attoma and Camillo 2011).

Then, to determine the significance of the detected imbalance, `%GI` macro performs a multivariate imbalance test. The null hypothesis of no dependence among **X** and **T** is specified as

$$H_0 : I_W = I_T \tag{3}$$

On the basis of the asymptotic distribution function of I_B (Estadella, Aluja, and Thi-Henestrosa 2005) expressed as

$$I_B \sim \frac{\chi^2_{(T-1)(J-1),\alpha}}{nQ} \tag{4}$$

The interval of plausible values for GI is defined as

$$GI \in \left(0, \frac{\chi^2_{(T-1)(J-1),\alpha}}{nQ}\right) \tag{5}$$

With n as the sample size, Q as the number of baseline covariates and $\chi^2_{(T-1)(J-1)}$ as the χ^2 value with $(T-1)(J-1)$ degrees of freedom. If the measured GI is outside the interval, then the null hypothesis of no dependence among **X** and **T** is rejected and data are deemed unbalanced. The main advantage of the GI measure is its simplicity of interpretation. The proposed measure varies in $[0, I_t]$. Perfect balance occurs when $I_B = 0$; whereas, perfect imbalance occurs when $I_W = 0$ and $I_B = I_T$ which indicates that the observed total variability of the X-space is completely due to the influence of conditioning (T). An index that ranges between 0 and 1 is represented by the Multivariate Imbalance Coefficient (MIC) which is defined as one minus the ratio between the within-groups inertia relative to the total inertia:

$$MIC = 1 - \frac{I_W}{I_T} \tag{6}$$

$MIC = 0$ denotes perfect balance; whereas, $MIC = 1$ indicates perfect imbalance. The GI measure works with categorical nominal or ordinal variables. Continuous variables need to be previously discretized. Furthermore, it also works in a multitreatment environment.

It is very similar for its simplicity to the measure proposed by Iacus *et al.* (2011). At the same time, it is more exhaustive than the L_1 measure since it considers the variability of a global space, its decomposition in between and within variability and also the asymptotic distribution function of I_B, that allows to define an interval of plausible values of the Global Imbalance measure. In addition, the use of the Bart table and the Burt band (for more details see D'Attoma and Camillo 2011) in the computation of the between groups inertia is more exhaustive than cross-tabulating the discretized and categorical variables for the treated and control groups separately. The Burt table is the symmetric matrix of all two-way crosstabulations between the categorical, nominal or ordinal variables, and has an analogy to the covariance matrix of continuous variables. It simultaneously displays information on the occurrence of category combinations (frequencies) for all variables. The Burt Band crosses the categories of the **X** variables with **T** levels. Furthermore, in the computation of the between groups inertia a more appropriate distance measure is used that is the χ^2 metric. The χ^2 metric includes a coefficient that re-evaluates elements with low frequency and resizes those with high frequency by weighting each element by the inverse of its importance on the total frequencies. Such a metric avoids to pay attention in the data pre-processing to the equilibrium between categories. It will be no more necessary to avoid categories with low frequency or variables with a lot of categories. The `%GI` macro produces as output a SAS dataset that reports for each group (e.g., a PS bin, a cluster, a stratum): The group size (`n`), the number of units in the treatment group 1 (`n_t1`), the number of units in the treatment group 2 (`n_t2`), the number of units in the treatment group n (`n_tn`), the group identifier (`id_clu`), the Global Imbalance measure (`GI`), the upper limit of the interval of the plausible values (`CHI`), the significance level used in the balancing test (`alpha`), the number of treatment levels in the entire dataset considered (`multitreat`), the MIC coefficient (`MIC`), the number of treatment levels in the specific subgroup (`LEVELT`) and the balance summary (`Balance`). `Balance` equals `yes` if the group is balanced, equals `no` if the group is unbalanced and equals `no common support` if units are observed only in a particular state without units in the other state.

4. List of parameters in the macro

Based on the GI measure and its related test presented in Section 3, a SAS/IML (SAS Institute Inc. 2008) macro program to measure and test Global Imbalance is written. A complete list of the parameters in `%GI` is as follows:

```
%GI(library=, dsn=, out=, firstclu=, lastclu=, id=, group_var=,
  balance_var=, Q=, treat=, alpha=, multitreat=);
```

where

- `library`: Name of the directory in which information is located.
- `dsn`: Name of the SAS data set to be read. It must contain Q categorical covariates, the treatment indicator variable, the ID variable and the group membership variable. A group could be the result of any classification analysis conducted separately before running `%GI`.
- `out`: Name of the SAS output data set.

- `firstclu`: Number of the first group to analyze. It is a numeric value.
- `lastclu`: Number of the last group to analyze. It is a numeric value.
- `id`: ID variable.
- `group_var`: Name of the variable that denotes the group membership.
- `balance_var`: Includes the name(s) of the baseline categorical variable(s) to be balance checked. The name(s) may be listed in any order and separated by blanks. The variable(s) must be numeric. No missing values are allowed.
- `Q`: Number of categorical variables on which simultaneously check imbalance.
- `treat`: Name of the treatment indicator variable. It must be a numerical value.
- `alpha`: Significance level to be used in testing GI.
- `multitreat`: Denotes the number of treatment levels.

The macro computes for each group the GI measure using the SAS/IML language. At this end, first it counts treatment and control units for each group, then creates a disjunctive table for each group. In particular, to compute the GI measure the following matrices will be created and used within the IML procedure: **Z** that includes the Q baseline categorical covariates in disjunctive form, **L** that includes the t treatment levels indicators, **B** that is the Burt table, the result of the inner product of the indicator matrix **Z**, the **Band** matrix that is a contingency table which crosses the categories of the baseline categorical covariates with each treatment level. Before quitting, the `%GI` macro deletes temporary datasets created during the implementation to avoid cluttering and errors in case the macro is invoked again.

5. Examples

This section will work through the use of the `%GI` macro with two artificial examples: One considering a binary treatment case and another a multi-treatment environment. Raw data as well as code for performing both example analyses are provided with this article. In the binary treatment case results are compared to the L_1 distance measure of Iacus *et al.* (2011), the I_n test statistic of Li *et al.* (2009) and to the Hansen and Bowers test. Whereas, in the multitreatment case, for the sake of brevity, only the GI index is considered.

5.1. Binary treatment case

The dataset

The present example measures the effect of a binary treatment **T** on an outcome of interest by subgroups. Assume that no random assignment to treatment conditions is feasible. The aim is to show how the macro works to check balance. Assume to have a dataset containing 1775 instances and five baseline categorical variables with different levels: X_1 with two levels, X_2 with two levels, X_3 with three levels, X_4 with two levels and X_5 with two levels. All possible combinations of covariates ($2 \times 2 \times 3 \times 2 \times 2 = 48$ cells) are considered. Units within each of

Combination	$Y(1)$	$Y(0)$	ATE
$X_1 = 1; X_3 = 1$	$= 0.88X_1 + 2X_2 + 3.33X_3 + 0.31X_4 + 8X_5$	$= Y(1) - 19.18$	19.18
$X_1 = 1; X_3 = 2$	$= 6X_1 + 3.3X_2 + 4.1X_3 + 0.31X_4 + 8X_5$	$= Y(1) + 29.24$	-29.24
$X_1 = 1; X_3 = 3$	$= -2.3X_1 + 0.99X_2 - 3X_3 + 0.31X_4 + 8X_5$	$= Y(1) - 2.41$	2.41
Others	$= 7X_1 + 0.5X_2 - 0.31X_3 + 0.31X_4 + 8X_5$	$= Y(1) - 25.6$	25.60

Table 1: Simulated outcomes.

those 48 cells are allocated on the basis of different proportions (π) to a different treatment level in order to create dependence among $\mathbf{X}$ and $\mathbf{T}$ (see Table 9 in Appendix A).

Assume that all covariates involved in the assignment to treatment process are perfectly known and no confounding variables exist. Assume this dataset is available before seeing any outcome. Suppose that the 1775 instances are classified employing a cluster analysis on multiple correspondence analysis coordinates and that 29 groups are chosen on the basis of the visual inspection of the resulting dendrogram. Before estimating any treatment effect the GI is measured and tested in each subgroup. Since in real applications may be less appropriate to expect that treatment effects are the same on all units than considering treatment effect heterogeneity within subgroups, heterogeneous treatment effects are simulated. At this end, different potential outcomes $Y(0)$ and $Y(1)$ are generated for observations with a different set of covariates combinations (Table 1).

Then, the observed outcomes ($Y_{i,obs}$) are obtained as in the following equation:

$$Y_{i,obs} = T_i Y_i(1) + (1 - T_i) Y_i(0) \tag{7}$$

By design, different average treatment effects (ATE) exist: 19.18; -29.24; 2.41; 25.6. The effect of treatment is estimated as the comparison of means between treatment and comparison cases. The dataset called `example_binary` contains the following information: The ID, the treatment indicator variable (`t`), the baseline covariates (`X1, X2, X3, X4, X5`), the group membership indicator (`Cluster`) and the observed continuous outcomes (`Y_obs`)[2]. Assume that the data are in SAS format and are stored in the directory specified by the user. Finally, invoke the `%GI` macro.

The implementation

Specify the `%INCLUDE` statement to indicate the location of the macro file, input values for various arguments as shown in the previous section and in the code below and invoke the macro. The macro will create a new SAS file, save it as `balance_binary` in the specified directory. This would be accomplished with the following macro call in SAS:

```
%GI(library = work, dsn = example_binary, out = balance_binary,
  firstclu = 1, lastclu = 29, id = id, group_var = cluster,
  balance_var = X1 X2 X3 X4 X5, Q=5, treat = t, alpha = 0.05,
  multitreat = 2)
```

[2]The observed outcome does not enter in the assessment of balance. It is not used when the macro is invoked. But, it will be used after balance checking in order to understand if to a situation of balance corresponds a reduction of bias in the treatment effect estimation. In this sense, the correspondence between balance and bias reduction is considered as a measure of the success of the method used to check balance.

Work.Balance_binary

	n_t1	n_t2	id_clu	GI	CHI	MIC	Balance	alpha	multitreat	LEVELT	n
1	5	45	1	0	0.038	0	Yes	0.05	2	2	50
2	15	15	2	0	0.0738	0	Yes	0.05	2	2	30
3	20	20	3	0	0.0474	0	Yes	0.05	2	2	40
4	24	16	4	0.025	0.0554	0.125	Yes	0.05	2	2	40
5	10	40	5	0	0.038	0	Yes	0.05	2	2	50
6	8	42	6	0	0.038	0	Yes	0.05	2	2	50
7	25	25	7	0	0.038	0	Yes	0.05	2	2	50
8	25	25	8	0	0.038	0	Yes	0.05	2	2	50
9	35	15	9	0	0.038	0	Yes	0.05	2	2	50
10	30	20	10	0	0.038	0	Yes	0.05	2	2	50
11	48	2	11	0	0.038	0	Yes	0.05	2	2	50
12	25	25	12	0	0.038	0	Yes	0.05	2	2	50
13	5	45	13	0	0.038	0	Yes	0.05	2	2	50
14	5	45	14	0	0.038	0	Yes	0.05	2	2	50
15	25	25	15	0	0.038	0	Yes	0.05	2	2	50
16	99	1	16	0	0.019	0	Yes	0.05	2	2	100
17	45	5	17	0	0.038	0	Yes	0.05	2	2	50
18	30	30	18	0	0.0369	0	Yes	0.05	2	2	60
19	5	75	19	0	0.0237	0	Yes	0.05	2	2	80
20	11	56	20	0.0623	0.0376	0.1558	No	0.05	2	2	67
21	25	23	21	0.005	0.0586	0.0083	Yes	0.05	2	2	48
22	33	33	22	0	0.0382	0	Yes	0.05	2	2	66
23	60	20	23	0	0.0237	0	Yes	0.05	2	2	80
24	5	75	24	0	0.0237	0	Yes	0.05	2	2	80
25	8	72	25	0	0.0237	0	Yes	0.05	2	2	80
26	5	75	26	0	0.0237	0	Yes	0.05	2	2	80
27	8	66	27	0.0183	0.038	0.0304	Yes	0.05	2	2	74
28	95	5	28	0	0.019	0	Yes	0.05	2	2	100
29	17	83	29	0.0123	0.0252	0.0308	Yes	0.05	2	2	100

Figure 1: The output dataset in the binary treatment example.

The output

The output includes a dataset called `balance_binary` (Figure 1) that contains information about balance for each subgroup.

Specifically, it displays the number of units within treatment group (`n_t1`), the number of units in the control group (`n_t2`), the group membership indicator (`Id_clu`), the GI measure (`GI`), the upper limit of the confidence interval (`CHI`), the MIC coefficient (`MIC`), the significance level (`alpha`), the number of treatment levels present in the entire dataset (`multitreat`) and the balance result (`Balance`). `Balance` equals `yes` if the GI measure is lower than the upper limit of the interval; otherwise, it equals `no`. Finally, `Balance` equals `no common support` in case lacks the common support. As reported in Figure 1 only one group over 29 is deemed unbalanced. In the remaining 28 balanced clusters the treatment effect of interest is computed. Table 2 reports the estimated effects with standard errors and shows that in 25 over 28 clusters the simulated heterogeneous effects are exactly reproduced. Results do support the conclusion that the estimated effects are unbiased where baseline characteristics are by the GI measure computation exogenous to the treatment and this is confirmed by the percent bias reduction (Rubin 1973) that reaches its maximum in almost all balanced subgroups.

Comparison of results

Results in terms of L_1 distance are obtained running the `imbalance` function of the **cem**

Cluster	Effect	Std. error	Bias reduction (%)
1	2.41	0	100
2	2.41	0.0552	100
3	25.6	0	100
4	22.99	0.876	69
5	25.6	0	100
6	25.6	0	100
7	19.18	0	100
8	19.18	0	100
9	−29.24	0	100
10	−29.24	0	100
11	19.18	0	100
12	19.18	0	100
13	25.6	0	100
14	25.6	0	100
15	2.41	0	100
16	19.18	0	100
17	19.18	0	100
18	25.6	0.9904	100
19	25.6	0	100
20	–	–	–
21	−16.54	5.0859	73
22	25.6	0.8505	100
23	−29.24	0	100
24	2.41	0	100
25	25.6	0	100
26	25.6	0	100
27	14.52	1.53	−141
28	19.18	0	100
29	25.62	0.0408	100

Table 2: Estimated effects by clusters.

Sample size	1775	
Treated	751	Balance
Untreated	1024	
GI	0.079	No
Interval	$(0; 0.002)$	
L_1	0.584	No
I_n	121.462	No
p value	<2.22e-16	
H&B		
χ^2	552	No
p value	4.62e-117	

Table 3: Binary treatment case: Comparison of results on the entire dataset.

Cluster	GI	Balance	L_1	Balance	I_n	p value	Balance	H&B p value	Balance
1	0	Yes	0	Yes	0	<2.22e-16	No	0.019	No
2	0	Yes	0	Yes	−0.233	0.909	Yes	1	Yes
3	0	Yes	0	Yes	0	<2.22e-16	No	2.57e-78	No
4	0.025	Yes	0.312	Yes	1.221	0.015	No	0.027	No
5	0	Yes	0	Yes	0	<2.22e-16	No	8.69e-21	No
6	0	Yes	0	Yes	0	<2.22e-16	No	2.11e-27	No
7	0	Yes	0	Yes	0	<2.22e-16	No	6.56e-98	No
8	0	Yes	0	Yes	0	<2.22e-16	No	6.56e-98	No
9	0	Yes	0	Yes	0	<2.22e-16	No	9.43e-88	No
10	0	Yes	0	Yes	0	<2.22e-16	No	5.07e-89	No
11	0	Yes	0	Yes	0	<2.22e-16	No	4.29e-30	No
12	0	Yes	0	Yes	0	<2.22e-16	No	6.56e-98	No
13	0	Yes	0	Yes	0	<2.22e-16	No	0	No
14	0	Yes	0	Yes	0	<2.22e-16	No	0	No
15	0	Yes	0	Yes	0	<2.22e-16	No	6.56e-98	No
16	0	Yes	0	Yes	No common support		–	0	No
17	0	Yes	0	Yes	0	<2.22e-16	No	2.76e-14	No
18	0	Yes	0	Yes	−0.348	0.882	Yes	1	Yes
19	0	Yes	0	Yes	0	<2.22e-16	No	0	No
20	0.062	No	0.566	No	6.849	0.002	No	9.48e-05	No
21	0.005	Yes	0.132	Yes	−0.675	0.827	Yes	0.755	Yes
22	0	Yes	0	Yes	−0.310	0.889	Yes	1	Yes
23	0	Yes	0	Yes	0	<2.22e-16	No	3.44e-14	No
24	0	Yes	0	Yes	0	<2.22e-16	No	0	No
25	0	Yes	0	Yes	0	<2.22e-16	No	0	No
26	0	Yes	0	Yes	0	<2.22e-16	No	0	No
27	0.018	Yes	0.477	Yes	3.711	0.002	No	0.107	Yes
28	0	Yes	0	Yes	0	<2.22e-16	No	0	No
29	0.012	Yes	0.255	Yes	2.069	0.007	No	0.053	Yes

Table 4: Comparison of results by subgroups in the binary treatment case.

package in R (Iacus *et al.* 2009)[3]. Results from the I_n test statistic of Li *et al.* (2009) are obtained within R software using the **np** package and the `npdeneqtest` function (Racine 2012). Finally, results about the Hansen and Bowers simultaneous balance test are obtained using the **RItools** package and the `xbalance` function (Bowers *et al.* 2011). First, balance on the overall dataset is assessed using our GI measure and its related test and compared to the L_1 distance, the I_n statistic and the Hansen and Bowers (H&B) simultaneous test. As emerges from Table 3 all compared measures let us conclude that balance does not hold in the entire dataset. By considering subgroups, our GI measure and L_1 distance give the same results (Table 4). As confirmed by the treatment effect estimation and the percent bias reduction reported in Table 2 both measures correctly assess balance. Whereas, only in 5 clusters over 29 the I_n statistic confirms GI and L_1 results. An important difference between the I_n statistic and the other measures concerns the definition of common support. According to all measures, with the exception of the I_n statistic, the common support set holds if at least one observation is present in all treatment options. This allows to measure balance in any case and let practitioners to define how much restrictive the definition of common support

[3]The multitreatment version of the L_1 distance could be obtained using the most recent R version of the **cem** package.

must be. We suppose the I_n statistic fails in checking balance because data are not a mix of continuous and discrete/categorical variables. Also the Hansen and Bowers test fails in detecting balance. It gives results different from those of GI and L_1 in 22 clusters over 29. We can conclude that only the GI measure and the L_1 distance correctly detect balance, and this conclusion is supported by the estimated effects that are unbiased in almost all cases, as showed by the percent bias reduction reported in Table 2.

5.2. Multitreatment case

The dataset

In the present example, a treatment **T** with 3 levels is considered. Assume to have a dataset containing 15645 instances and five baseline categorical variables with different levels: X_1 with two levels, X_2 with two levels, X_3 with three levels, X_4 with two levels and X_5 with two levels. As in the binary case, all possible combinations of covariates ($2\times2\times3\times2\times2 = 48$ cells) are considered. Then, units within each of those 48 cells are allocated on the basis of different proportions (π) to a different treatment level in order to create dependence among **X** and **T** (Table 9, Appendix A). As in the binary treatment example, this dataset is available before seeing any outcome. After balance checking, in order to verify if treatment effects are unbiased in balanced clusters, the presence of heterogeneous treatment effects is simulated. Assume to have 3 multiple treatments ($T = \{1, 2, 3\}$). Therefore, each subject has 3 potential outcomes, $Y(1)$, $Y(2)$ and $Y(3)$. At this end, $Y(1), Y(2), Y(3)$ are generated for each observation who receives a treatment t. For each unit, potential outcomes, $Y_i(t)$, are generated with the following model and assuming a zero error term and a zero intercept:

$$Y_i(t) = \alpha(t) + \beta_1(t)X_{1i} + \beta_2(t)X_{2i} + \beta_3(t)X_{3i} + \beta_4(t)X_{4i} + \beta_5(t)X_{5i} \tag{8}$$

In particular, four different set of parameters are generated in order to create heterogeneous treatment effects (Table 5). Despite three potential outcomes exist, only one outcome under the assigned treatment can be observed. Following Feng, Zhou, Zou, Fan, and Li (2012) the observed outcome for each subject i, $Y_{i,obs}$, is computed as:

$$Y_{i,obs} = \sum_{t=1}^{m} Y_i(t) I(T_i = t) \tag{9}$$

where $I(T_i = t)$ is the indicator of receiving treatment t:

$$\begin{aligned} I(T) &= 1, \text{ if } T = t \\ &= 0, \text{ otherwise} \end{aligned} \tag{10}$$

Parameters			Set 1	Set 2	Set 3	Set 4
$\beta_1(1)$	$\beta_1(2)$	$\beta_1(3)$	$[1, -7.5, 4]$	$[1, -7.5, 44]$	$[-1, -7.5, -44]$	$[0.5, -7.5, -4]$
$\beta_2(1)$	$\beta_2(2)$	$\beta_2(3)$	$[3, 2, 11]$	$[0.3, -3, 11]$	$[-0.3, -2, -11]$	$[3, -2, -1]$
$\beta_3(1)$	$\beta_3(2)$	$\beta_3(3)$	$[2, 4, 8.3]$	$[2.2, -4, 83]$	$[-2.2, -4, -83]$	$[22, -4, -8.3]$
$\beta_4(1)$	$\beta_4(2)$	$\beta_4(3)$	$[6, 3.3, 10]$	$[0.6, -3.3, 10]$	$[-0.6, -3.3, -10]$	$[0.6, -3.3, 10]$
$\beta_5(1)$	$\beta_5(2)$	$\beta_5(3)$	$[7, 0.31, -15]$	$[7, -0.31, 15]$	$[-7, -0.31, -15]$	$[17, -0.31, -1.5]$

Table 5: Parameter sets.

Parameters	ATE_{12}	ATE_{13}	ATE_{23}
Set 1	30.05	5.22	−24.82
Set 2	106.96	118.77	11.80
Set 3	8.58	327.69	319.10
Set 4	40.95	−246.22	−287.18

Table 6: Simulated true treatment effects in multi-treatment case.

n_{t1}	n_{t2}	n_{t3}	n	GI	CHI	MIC	Balance
5198	5377	5070	15645	0.0007	0.0004	0.00057	No

Table 7: Balance in the overall data in multi-treatment case.

and $Y_i(t)$ denotes the potential outcome of subject i if the subject has been assigned treatment t. Finally, the true ATEs are estimated. If all the potential outcomes are observed, the ATE of treatment j versus treatment k with $j \neq k$ are estimated by

$$ATE_{jk} = \frac{1}{n}\sum_{i=1}^{n} Y_i(j) - \frac{1}{n}\sum_{i=1}^{n} Y_i(k) \tag{11}$$

As displayed in Table 6, by design, 3 true treatment effects exist for each simulated parameter set and they are estimated as the comparison of means of potential outcomes.[4] Once the simulated data are ready, balance is checked on the overall simulated data. Being the overall data unbalanced (Table 7) a subgroup analysis is performed using a cluster analysis on multiple correspondence analysis coordinates. On the basis of the examination of the dendrogram, 29 groups are retained. Before estimating the treatment effect of interest the GI is measured and tested in each subgroup. At this end, assume that the data are in SAS format and are stored in the directory specified by the user. Finally, invoke the the `%GI` macro. The dataset called `example_multi` contains the following information: The ID, the multi-treatment indicator (t), the baseline covariates $(x_1, x_2, x_3, x_4, x_5)$, the group membership indicator (cluster) and the observed continuous outcome (Y_{obs}).

The implementation

Specify the `%INCLUDE` statement to indicate the location of the macro file, input values for various arguments as shown in the previous section and in the code below and invoke the macro. The macro will create a new SAS file, save it as `balance_multi` in the specified directory. This would be accomplished with the following macro call in SAS:

```
%GI(library = work, dsn = example_multi, out = balance_multi, id = id,
  firstclu = 1, lastclu = 29,group_var = cluster,
  balance_var = X1 X2 X3 X4 X5, Q = 5, treat = t, alpha = 0.05,
  multitreat = 3)
```

The output

The output will include a dataset called `balance_multi` (Figure 2).

[4] In fact, given 3 treatment groups, all possible comparisons between all couples of means are $3(3-1)/2$.

Work.Balance_multi

	n_t1	n_t2	n_t3	id_clu	GI	CHI	MIC	Balance	alpha	multitreat	LEVELT	n
1	214	213	184	1	0.0066	0.006	0.0331	No	0.05	3	3	611
2	211	168	193	2	0.0063	0.0064	0.0316	Yes	0.05	3	3	572
3	227	243	166	3	0.0033	0.0058	0.0164	Yes	0.05	3	3	636
4	198	252	227	4	0.0068	0.0054	0.0341	No	0.05	3	3	677
5	245	198	227	5	0.0043	0.0055	0.0217	Yes	0.05	3	3	670
6	255	243	209	6	0.0025	0.0052	0.0125	Yes	0.05	3	3	707
7	247	199	250	7	0.0013	0.0053	0.0067	Yes	0.05	3	3	696
8	114	136	112	8	0	0.0086	0	Yes	0.05	3	3	362
9	135	136	135	9	0	0.0076	0	Yes	0.05	3	3	406
10	155	257	250	10	0.0003	0.0055	0.0013	Yes	0.05	3	3	662
11	236	195	162	11	0.0003	0.0062	0.0015	Yes	0.05	3	3	593
12	215	222	213	12	0.0057	0.0056	0.0287	No	0.05	3	3	650
13	209	231	184	13	0.0047	0.0059	0.0236	Yes	0.05	3	3	624
14	101	134	125	14	0	0.0086	0	Yes	0.05	3	3	360
15	133	132	106	15	0	0.0084	0	Yes	0.05	3	3	371
16	157	219	229	16	0.0012	0.0061	0.0058	Yes	0.05	3	3	605
17	161	204	229	17	0.0043	0.0062	0.0215	Yes	0.05	3	3	594
18	140	136	124	18	0	0.0078	0	Yes	0.05	3	3	400
19	273	177	225	19	0.0005	0.0054	0.0024	Yes	0.05	3	3	675
20	183	210	183	20	0.0041	0.0064	0.0207	Yes	0.05	3	3	576
21	105	129	130	21	0	0.0085	0	Yes	0.05	3	3	364
22	135	135	109	22	0	0.0082	0	Yes	0.05	3	3	379
23	221	169	212	23	0.0057	0.0061	0.0284	Yes	0.05	3	3	602
24	188	273	228	24	0.0004	0.0053	0.0022	Yes	0.05	3	3	689
25	135	126	82	25	0	0.009	0	Yes	0.05	3	3	343
26	93	122	139	26	0	0.0088	0	Yes	0.05	3	3	354
27	208	202	171	27	0.0038	0.0063	0.0189	Yes	0.05	3	3	581
28	109	132	98	28	0	0.0091	0	Yes	0.05	3	3	339
29	195	184	168	29	0.0043	0.0067	0.0214	Yes	0.05	3	3	547

Figure 2: The output dataset in the multi-treatment example.

It displays the number of units within treatment group 1 (`n_t1`), the number of units in the treatment group 2 (`n_t2`), the number of units in the treatment group 3 (`n_t3`) the group size (`n`), the group membership indicator (`Id_clu`), the GI measure (`GI`), the upper limit of the confidence interval (`CHI`), the significance level (`alpha`), the MIC coefficient (`MIC`), the number of treatment levels present in the entire dataset (`multitreat`), the number of treatment levels present in the specific subgroup (`LEVELT`) and the balance result (`Balance`). `Balance` equals `yes` if the GI measure is lower than the upper limit of the interval; otherwise, it equals `no`. Finally, `Balance` equals `no common support` in case lacks the common support[5]. As reported in Figure 2, only 3 groups over 29 are deemed unbalanced. In the remaining clusters the treatment effects of interest are computed. Table 8 reports the estimated effects with simultaneous confidence limits in brackets and shows that, on average, the bias is reduced around 60%.

We acknowledge that it is a result not so excellent as that obtained in the binary case (Table 2). Such a result might be due to the increased number of combinations of treatment levels and covariates. At the same time, we consider the result satisfactory if compared to bias reduction obtained by adopting a PS Subclassification analysis[6](Figure 3), where, on average, bias is reduced around 30% in case the propensity score is forced to be split in 29 bins. From Figure 3

[5]For the multiple treatment case the common support set is in general determined by the minimum of the maximum and the maxima of the minimum participation probabilities for the various treatment options (Frölich, Hesmati, and Lechner 2004)

[6]The propensity score is estimated using a generalized multinomial logit and the SAS `catmod proc` and using as independent variables all the five simulated variables X_1, X_2, X_3, X_4, X_5

Cluster	Estimated Effects			Bias reduction (%)		
	T1 vs. T2	T1 vs. T3	T2 vs. T3	T1 vs. T2	T1 vs. T3	T2 vs. T3
2	32.67 [31.85; 33.50]	18.52 [17.72; 19.32]	-14.15 [−15.00; −13.30]	93.54	81.85	67.37
3	35.10 [34.42; 35.78]	15.71 [14.96; 16.46]	-19.39 [−20.13; −18.65]	87.55	85.687	83.39
5	87.20 [86.52; 87.90]	97.24 [96.57; 97.90]	10.03 [9.33; 10.73]	45.62	46.52	54.71
6	70.24 [69.63; 70.85]	76.74 [76.10; 77.37]	6.50 [5.85; 7.14]	−1.04	−4.39	−35.11
7	82.71 [81.55; 83.87]	86.46 [85.37; 87.55]	3.75 [2.59; 4.90]	33.27	19.75	−105.08
8	94.4 [94.4; 94.4]	105.5 [105.5; 105.5]	11.1 [11.1; 11.1]	65.43	67.03	81.93
9	77.11 [77.11; 77.11]	87.00 [87.00; 87.00]	9.89 [9.89; 9.89]	17.86	21.09	51.14
10	10.56 [9.36; 11.76]	337.48 [336.22; 338.63]	326.86 [325.81; 327.91]	96.82	96.07	97.50
11	126.74 [126.47; 127.00]	145.64 [145.36; 145.92]	18.90 [18.61; 19.19]	45.57	33.26	−80.41
13	10.37 [9.67; 11.07]	318.75 [318.00; 319.49]	308.38 [307.65; 309.10]	97.13	96.41	96.55
14	120.21 [120.21; 120.21]	133.00 [133.00; 133.00]	12.79 [12.79; 12.79]	63.54	64.65	75.06
15	125.21 [125.21; 125.21]	137.00 [137.00; 137.00]	11.79 [11.79; 11.79]	49.78	54.72	99.49
16	74.60 [49.54; 99.64]	227.00 [202.17; 251.81]	152.39 [129.75; 175.04]	−6.42	59.60	46.44
17	3.81 [2.99; 4.64]	326.29 [325.49; 327.09]	322.47 [321.72; 323.23]	92.29	99.44	98.92
18	137.5 [137.5; 137.5]	151.5 [151.5; 151.5]	14.0 [14.0; 14.0]	15.96	18.70	44.27
19	113.95 [113.69; 114.22]	123.24 [122.99; 123.48]	9.28 [9.00; 9.56]	80.76	88.89	35.62
20	47.09 [46.32; 47.86]	−255.22 [−256.02; −254.43]	−302.31 [−303.08; −301.55]	79.33	97.23	94.87
21	120.42 [135.8; 135.8]	135.80 [135.80; 135.80]	15.38 [15.38; 15.38]	62.96	57.70	9.16
22	115.42 [115.42; 115.42]	131.80 [131.80; 131.80]	16.38 [16.38; 16.38]	76.72	67.63	9.16
23	40.90 [39.72; 42.08]	−245.18 [−246.29; −244.06]	−286.08 [−287.28; −284.88]	99.80	99.68	99.63
24	38.10 [37.04; 39.17]	−236.00 [−237.11; −234.90]	−274.11 [−275.12; −273.10]	90.36	95.85	95.57
25	12.3 [12.3; 12.3]	323.00 [323.00; 323.00]	310.60 [310.60; 310.60]	94.02	98.12	97.27
26	38.31 [38.31; 38.31]	−242.10 [−242.10; −242.10]	−280.41 [−280.41; −280.41]	91.06	98.73	97.70
27	95.81 [95.13; 96.50]	108.67 [107.96; 109.38]	12.85 [12.14; 13.57]	69.32	74.91	73.54
28	99.21 [99.21; 99.21]	106.70 [106.70; 106.70]	7.49 [7.49; 7.49]	78.67	70.02	−9.92
29	79.33 [57.40; 101.26]	−69.65 [−92.11; −47.19]	−148.99 [−171.76; −126.22]	−29.36	45.62	53.16

Table 8: Effects by subgroups in multi-treatment case.

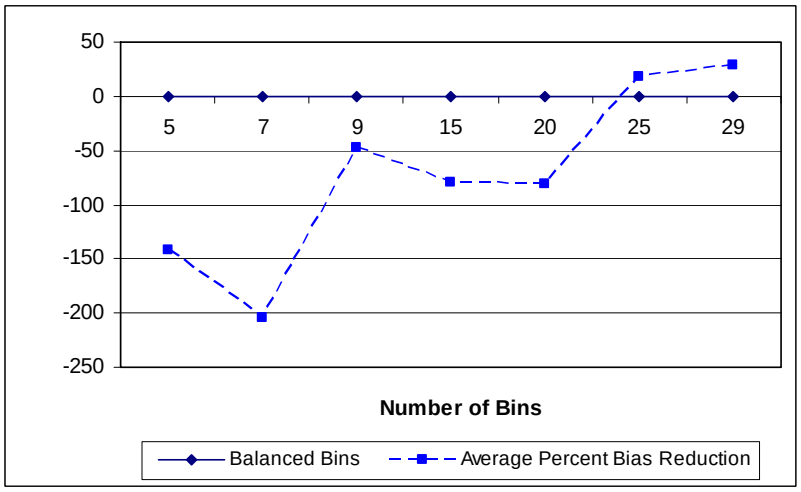

Figure 3: Balance and percent bias reduction in PS subclassification.

emerges that in most of splitting cases bias is increased rather than reduced and such a result might be due to the fact that propensity score subclassification in the present multi-treatment example is not able to balance pre-treatment covariates.

6. Concluding remarks

The `%GI` macro enables one to measuring and testing balance of categorical, ordinal and continuous covariates according to the GI measure and its related test introduced in D'Attoma and Camillo (2011). The macro is illustrated using two artificial examples showing that it works in both binary and multi-treatment environments. The macro described in this paper encourages analysts to globally checking balance rather than performing variable-by-variable tests, which do not consider interactions among baseline covariates. Compared to other measures, in the binary case it correctly detects balance as the L_1 distance, and such correcteness is supported by the percent bias reduction that reaches its maximum in most of examined clusters. The I_n test statistic and the Hansen and Bowers test fail in correctly detect balance. The I_n test statistic and the Hansen and Bowers test probably fail for two main reasons. First, the nature of covariates used in the present examples might be not appropriate for the mentioned two tests. Second, the two tests might be influenced by the sample size of the groups being compared. Furthermore, the I_n test statistic is designed to work with mixed categorical and continuous data. We think it might not work when data are not mixed, but exclusively categorical or continuous as in the examples here presented. For what concerns the multi-treatment case, only the GI measure is considered. In terms of percent bias reduction its performance is worse than in the binary case and such a result might be due to the increased number of combinations of treatment levels and combinations. It was not our intent here to provide proofs of the theoretical superiority of GI measure over other examined measures; instead, we provide a brief introduction to the concept of GI and a simple illustration of its computation using the proposed `%GI` SAS macro. Our main goal has been to show how the macro works in both binary and multi-treatment case. Comparing

the GI measures and %GI macro to the other measures and their related tools, we learn that not all examined tools work with all kind of covariates and not all tools produce the same results. The limit of our proposed measure is that it does not work with continuous covariates that must be previously discretized with some discretization method that we do not suggest. In sum, the macro makes easy to check balance by subgroups on which estimate binary or multiple treatment effects of interest under non-experimental conditions, where a subgroup could be the result of any classification analysis or a bin of a PS subclassification (Dehejia and Wahba 2002). In doing that, the multivariate structure of data is taken into account. The main strength of the %GI macro is that it allows to solve complex problems, because especially in a data mining perspective, it does not suffer from the number of variables and observations. This paves the way for applications business-oriented (e.g., marketing or redemption campaigns) that might need a continuous monitoring. In fact the use of the %GI macro for its simplicity makes easy to monitor the effect of any kind of private or public policy by subgroups and even in a continuous way and, as such, might reverse the concept of evaluation, that might be considered not only as a one-time action, but as a process.

References

Baser O (2006). "Too Much Ado about Propensity Score Models? Comparing Methods of Propensity Score Matching." *Value in Health*, **9**, 377–385.

Blackwell M, Iacus SM, King G (2009). "**cem**: Coarsened Exact Matching in Stata." *The Stata Journal*, **9**(4), 524–546.

Bowers J, Fredrickson M, Hansen B (2011). ***RItools**: Randomization Inference Tools*. R package version 0.1-11, URL http://CRAN.R-project.org/package=RItools.

Camillo F, D'Attoma I (2010). "A New Data Mining Approach to Estimate Causal Effects of Policy Interventions." *Expert System with Applications*, **37**(1), 171–181.

D'Attoma I, Camillo F (2011). "A Multivariate Strategy to Measure and Test Global Imbalance in Observational Studies." *Expert System with Applications*, **38**(4), 3451–3460.

Dehejia RH, Wahba S (2002). "Propensity Score-Matching Methods for Nonexperimental Causal Studies." *The Review of Economics and Statistics*, **84**(1), 151–161.

Escofier B (1988). "Analyse des Correspondances Multiples Conditionelle." In E Diday (ed.), *Data Analysis and Informatics*, pp. 333–342. Elsevier Science, North Holland, Amsterdam.

Estadella JD, Aluja T, Thi-Henestrosa S (2005). "Distribution of the Inter and Intra Inertia in Conditional MCA." *Computational Statistics*, **20**(3), 449–463.

Feng P, Zhou X, Zou Q, Fan M, Li X (2012). "Generalized Propensity Score for Estimating the Average Treatment Effect of Multiple Treatments." *Statistics in Medicine*, **31**, 681–697.

Frölich M, Hesmati A, Lechner M (2004). "A Microeconometric Evaluation of Rehabilitation of Long-Term Sickness in Sweden." *Journal of Applied Econometrics*, **19**(3), 375–396.

Greenacre MJ (1984). *Theory and Applications of Correspondence Analysis*. Academic Press.

Hansen BB, Bowers J (2008). "Covariate Balance in Simple, Stratified and Clustered Comparative Studies." *Statistical Science*, **23**(2), 219–236.

Hayfield T, Racine JS (2008). "Nonparametric Econometrics: The **np** Package." *Journal of Statistical Software*, **27**(5), 1–32. URL http://www.jstatsoft.org/v27/i05/.

Ho DE, Imai K, King G, Stuart EA (2007). "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis*, **15**, 199–236.

Iacus SM, King G, Porro G (2009). "**cem**: Software for Coarsened Exact Matching." *Journal of Statistical Software*, **30**(9), 1–27. URL http://www.jstatsoft.org/v30/i09/.

Iacus SM, King G, Porro G (2011). "Multivariate Matching Methods That Are Monotonic Imbalance Bounding." *Journal of the American Statistical Association*, **106**(493), 345–361.

Imai K, King G, Stuart EA (2008). "Misunderstanding between Experimentalists and Observationalists about Causal Inference." *Journal of the Royal Statistical Society A*, **171**(2), 481–502.

Lebart L, Morineau A, Warwick KM (1984). *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices*. John Wiley & Sons.

Li Q, Maasoumi E, Racine JS (2009). "A Nonparametric Test for Equality of Distributions with Mixed Categorical and Continuous Data." *Journal of Econometrics*, **148**, 186–200.

Peck LR (2005). "Using Cluster Analysis in Program Evaluation." *Evaluation Review*, **29**(2), 178–196.

Peck LR, Camillo F, D'Attoma I (2010). "A Promising New Approach to Eliminating Selection Bias." *Canadian Journal of Program Evaluation*, **24**(2), 31–56.

Racine JS (2012). "Entropy-Based Inference Using R and the **np** Package: A Primer." R package vignette, version 0.40-13, URL http://CRAN.R-project.org/package=np.

Rosenbaum PR, Rubin DB (1983). "The Central Role of Propensity Score in Observational Studies for Causal Effects." *Biometrika*, **70**(1), 41–55.

Rosenbaum PR, Rubin DB (1984). "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association*, **79**(387), 516–524.

Rubin DB (1973). "Matching to Remove Bias in Observational Studies." *Biometrics*, **29**, 159–183.

Rubin DB (2001). "Using Propensity Scores to Help Design Observational Studies: Applications to the Tobacco Litigation." *Health Services and Outcome Research Methodology*, **2**, 169–188.

SAS Institute Inc (2008). *SAS/IML 9.2 User's Guide*. SAS Institute Inc., Cary, NC. URL http://www.sas.com/.

A. Design details

						Binary treatment			Multi-treatment			
Combinations	X_1	X_2	X_3	X_4	X_5	$\pi_{t=1}$	$\pi_{t=0}$	N_{tot}	$\pi_{t=1}$	$\pi_{t=2}$	$\pi_{t=3}$	N_{tot}
1	1	1	1	1	1	50%	50%	50	34.15%	42.15%	23.69%	325
2	2	1	1	1	1	50%	50%	40	33.15%	32.32%	34.53%	362
3	1	2	1	1	1	50%	50%	50	36.01%	26.57%	37.41%	286
4	2	2	1	1	1	50%	50%	30	31.62%	30.48%	37.89%	351
5	1	1	2	1	1	63%	38%	8	26.42%	38.35%	35.23%	352
6	2	1	2	1	1	6%	94%	50	35.10%	40.40%	24.50%	302
7	1	2	2	1	1	70%	30%	50	31.23%	25.58%	43.19%	301
8	2	2	2	1	1	6%	94%	80	32.15%	38.94%	28.91%	339
9	1	1	3	1	1	20%	80%	5	32.80%	29.20%	38.00%	250
10	2	1	3	1	1	20%	80%	50	28.06%	37.22%	34.72%	360
11	1	2	3	1	1	6%	94%	80	40.18%	32.74%	27.08%	336
12	2	2	3	1	1	16%	84%	50	35.85%	35.58%	28.57%	371
13	1	1	1	2	1	99%	1%	100	22.06%	37.06%	40.88%	340
14	2	1	1	2	1	50%	50%	40	39.13%	36.52%	24.35%	345
15	1	2	1	2	1	90%	10%	50	36.50%	37.39%	26.11%	337
16	2	2	1	2	1	50%	50%	20	33.25%	33.50%	33.25%	406
17	1	1	2	2	1	75%	25%	80	26.27%	34.46%	39.27%	354
18	2	1	2	2	1	10%	90%	50	36.56%	28.67%	34.77%	279
19	1	2	2	2	1	60%	40%	50	28.17%	36.51%	35.32%	252
20	2	2	2	2	1	10%	90%	80	42.03%	31.19%	26.78%	295
21	1	1	3	2	1	4%	96%	50	39.36%	36.73%	23.91%	343
22	2	1	3	2	1	30%	70%	10	39.72%	31.36%	28.92%	287
23	1	2	3	2	1	25%	75%	4	23.44%	37.19%	39.38%	320
24	2	2	3	2	1	20%	80%	5	39.87%	34.31%	25.82%	306
25	1	1	1	1	2	95%	5%	100	28.69%	37.06%	40.88%	251
26	2	1	1	1	2	50%	50%	20	37.31%	35.17%	27.52%	327
27	1	2	1	1	2	96%	4%	50	43.30%	23.05%	33.64%	321
28	2	2	1	1	2	50%	50%	6	39.42%	26.67%	33.91%	345
29	1	1	2	1	2	50%	50%	4	28.19%	40.95%	30.86%	337
30	2	1	2	1	2	20%	80%	5	39.82%	28.32%	31.86%	339
31	1	2	2	1	2	60%	40%	10	42.00%	31.00%	27.00%	301
32	2	2	2	1	2	6%	94%	80	41.07%	24.11%	34.82%	336
33	1	1	3	1	2	10%	90%	50	23.73%	40.82%	35.44%	316
34	2	1	3	1	2	30%	70%	10	35.00%	34.00%	31.00%	400
35	1	2	3	1	2	50%	50%	20	26.29%	39.46%	34.35%	294
36	2	2	3	1	2	17%	83%	30	25.72%	33.12%	41.16%	311
37	1	1	1	2	2	90%	10%	10	29.75%	41.10%	29.14%	326
38	2	1	1	2	2	50%	50%	30	40.00%	52.00%	8.00%	25
39	1	2	1	2	2	50%	50%	50	41.94%	35.16%	22.90%	310
40	2	2	1	2	2	50%	50%	10	31.49%	37.57%	30.94%	362
41	1	1	2	2	2	50%	50%	20	25.69%	42.01%	32.29%	288
42	2	1	2	2	2	10%	90%	50	35.62%	35.62%	28.76%	379
43	1	2	2	2	2	60%	40%	10	30.98%	42.02%	26.99%	326
44	2	2	2	2	2	38%	63%	8	28.85%	35.44%	35.71%	364
45	1	1	3	2	2	50%	50%	50	30.94%	26.98%	42.09%	278
46	2	1	3	2	2	40%	60%	10	27.11%	41.57%	31.33%	332
47	1	2	3	2	2	50%	50%	10	23.39%	40.35%	36.26%	342
48	2	2	3	2	2	10%	90%	50	39.31%	26.42%	34.28%	318

Table 9: Simulation designs.

Affiliation:

Furio Camillo, Ida D'Attoma
Department of Statistical Sciences
University of Bologna
via Belle Arti 41
40126 Bologna, Italy
E-mail: furio.camillo@unibo.it, Ida.dattoma2@unibo.it

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 51, Code Snippet 1 *Submitted:* 2011-11-15
November 2012 *Accepted:* 2012-08-20