

Multivariate Matching Methods That are Monotonic Imbalance Bounding*

Stefano M. Iacus[†] Gary King[‡] Giuseppe Porro[§]

October 8, 2009

Abstract

We introduce a new “Monotonic Imbalance Bounding” (MIB) class of matching methods for causal inference that satisfies several important in-sample properties. MIB generalizes and extends in several new directions the only existing class, “Equal Percent Bias Reducing” (EPBR), which is designed to satisfy weaker properties and only in expectation. We also offer strategies to obtain specific members of the MIB class, and present a member of this class, called Coarsened Exact Matching, whose properties we analyze from this new perspective.

*Open source R and Stata software to implement the methods described herein (called CEM) is available at <http://gking.harvard.edu/cem>; the cem algorithm is also available via the R package MatchIt (which has an easy-to-use front end). Thanks to Erich Battistin, Nathaniel Beck, Matt Blackwell, Andy Eggers, Adam Glynn, Justin Grimmer, Jens Hainmueller, Ben Hansen, Kosuke Imai, Guido Imbens, Fabrizia Mealli, Walter Mebane, Clayton Nall, Enrico Rettore, Jamie Robins, Don Rubin, Jas Sekhon, Jeff Smith, Kevin Quinn, and Chris Winship for helpful comments.

[†]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

A defining characteristic of observational data is that the investigator does not control the data generation process. The resulting impossibility of random treatment assignment thus reduces attempts to achieve valid causal inference to the process of selecting treatment and control groups that are as balanced as possible with respect to available pre-treatment variables. One venerable but increasingly popular method of achieving balance is through matching, where each of the treated units is matched to one or more control units as similar as possible with respect to the given set of pre-treatment variables.

Once a matched data set is selected, the causal effect is estimated by a simple difference in means of the outcome variable for the treated and control groups, assuming ignorability holds, or by modeling any remaining pre-treatment differences. The advantage of matching is that inferences from better balanced data sets will be less model dependent (Ho et al., 2007).

Consider a sample of n units, a subset of a population of N units, where $n \leq N$. For unit i , denote T_i as the treatment variable, where $T_i = 1$ if unit i receives treatment (and so is a member of the “treated” group) and $T_i = 0$ if not (and is therefore a member of the “control” group). The outcome variable is Y , where $Y_i(0)$ is the “potential outcome” for observation i if the unit does not receive treatment and $Y_i(1)$ is the potential outcome if the (same) unit receives treatment. For each observed unit, only one potential outcome is observed, $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$, which means that $Y_i(0)$ is unobserved if i receives treatment and $Y_i(1)$ is unobserved if i does not receive treatment. Without loss of generality, when we refer to unit i , we assume it is treated so that $Y_i(1)$ is observed while $Y_i(0)$ is unobserved and thus estimated by matching it with one or more units from a given reservoir of the control units.

Denote $\mathbf{X} = (X_1, X_2, \dots, X_k)$ as a k -dimensional data set, where each X_j is a column vector of the observed values of pre-treatment variable j for the n observations. That is, $\mathbf{X} = [X_{ij}, i = 1, \dots, n, j = 1, \dots, k]$. We denote by $\mathcal{T} = \{i : T_i = 1\}$ the set of

indexes for the treated units and by $n_T = \#\mathcal{T}$ the number of treated units; similarly $\mathcal{C} = \{i : T_i = 0\}$, $n_C = \#\mathcal{C}$ for the control units, with $n_T + n_C = n$. Given a treated unit $i \in \mathcal{T}$ with its vector of covariates \mathbf{X}_i , the aim of matching is to discover a control unit $l \in \mathcal{C}$ with covariates \mathbf{X}_l such that, the dissimilarity between \mathbf{X}_i and \mathbf{X}_l is very small in some metric, i.e. $d(\mathbf{X}_i, \mathbf{X}_l) \simeq 0$. A special case is the exact matching algorithm where, for each treated unit i , a control unit l is selected such that $d(\mathbf{X}_i, \mathbf{X}_l) = 0$, with d of full rank (i.e., if $d(a, b) = 0$ if and only if $a = b$).

The literature includes many methods of selecting matches, but only a single rigorous class of methods has been characterized, the so-called Equal Percent Bias Reducing (EPBR) methods. In introducing EPBR, Rubin (1976c) recognized the need for more general classes: “Even though nonlinear functions of X deserve study. . . , it seems reasonable to begin study of multivariate matching methods in the simpler linear case and then extend that work to the more complex nonlinear case. In that sense then, EPBR matching methods are the simplest multivariate starting point.” The introduction of the EPBR class has led to highly productive and, in recent years, fast growing literatures on the theory and application of matching methods. Yet, in the more than three decades since Rubin’s original call for continuing from this “starting point” to develop more general classes of matching models, none have appeared in the literature. We take up this call here and introduce a new class, which we denote Monotonic Imbalance Bounding (MIB) methods. This new class of methods generalize EPBR in a variety of useful ways.

In this paper, we review EPBR, introduce MIB, discuss several specific matching methods within the new class, and illustrate their advantages for empirical analysis. Throughout, we distinguish between *classes of methods* and specific *methods* (or algorithms) within a class that can be used in applications. Classes of methods define properties which all matching methods within the class must possess. Some methods may also belong to more than one class.

2 The Equal Percent Bias Reducing Class

Let $\mu_t \equiv E(\mathbf{X}|T = t)$, $t = 0, 1$, be a vector of expected values and denote by m_T and m_C the number of treated and control units matched by some matching method. Let $M_T \subseteq \mathcal{T}$ and $M_C \subseteq \mathcal{C}$ be the sets of indexes of the matched units in the two groups. Let $\bar{\mathbf{X}}_{n_T} = \frac{1}{n_T} \sum_{i \in \mathcal{T}} \mathbf{X}_i$, and $\bar{\mathbf{X}}_{n_C} = \frac{1}{n_C} \sum_{i \in \mathcal{C}} \mathbf{X}_i$ be the vector of sample means of the observed data and $\bar{\mathbf{X}}_{m_T} = \frac{1}{m_T} \sum_{i \in M_T} \mathbf{X}_i$, and $\bar{\mathbf{X}}_{m_C} = \frac{1}{m_C} \sum_{i \in M_C} \mathbf{X}_i$ be the vector of sample means for the matched data only.

EPBR requires all treated units to be matched, i.e. $m_T = n_T$ (thus $M_T = \mathcal{T}$), but allows for the possibility that only $m_C \leq n_C$ control units are matched, where m_C is chosen ex ante.

Definition 1 (Equal Percent Bias Reducing (EPBR); Rubin (1976b)). *An EPBR matching solution satisfies*

$$E(\bar{\mathbf{X}}_{m_T} - \bar{\mathbf{X}}_{m_C}) = \gamma(\mu_1 - \mu_0), \quad (1)$$

where $0 < \gamma < 1$ is a scalar.

A condition of EPBR is that the number of matched control units be fixed ex ante (Rubin, 1976a, p.110) and the particular value of γ be calculated ex post, which we emphasize by writing $\gamma \equiv \gamma(m_C)$. (The term ‘‘bias’’ in EPBR violates standard statistical usage and refers instead to the equality across variables in the reduction in covariate imbalance.) If the realized value of \mathbf{X} is a random sample, then (1) can be expressed as

$$E(\bar{\mathbf{X}}_{m_T} - \bar{\mathbf{X}}_{m_C}) = \gamma E(\bar{\mathbf{X}}_{n_T} - \bar{\mathbf{X}}_{n_C}). \quad (2)$$

The right side of (2) is the average mean-imbalance in the population that gives rise to the original data, and the left side is the average mean-imbalance in the population subsample of matched units. The EPBR property implies that improving balance in the difference in means on one variable also improves it on all others (and their linear combinations) by a proportional amount, which is why γ is assumed to be a scalar. EPBR is a

relevant property only if one assumes that the function which links the covariates and the outcome is equally sensitive to all components (for example a linear function), or if the analyst scales the covariates so this is the case.

EPBR attempts to improve only *mean* imbalance (or main effects in \mathbf{X}) and says nothing about other moments, interactions, or nonlinear relationships (except inasmuch as one includes in \mathbf{X} specifically chosen terms like X_j^2 , $X_j \times X_k$, etc.). Rubin and Thomas (1992) give some specialized conditions which can generate the maximum level of imbalance reduction possible for any EPBR matching method. Although this result does not indicate which method will achieve the maximum, it may provide useful guidance about how well the search is going.

No method of matching satisfies EPBR without data restrictions. To address these issues, Rosenbaum and Rubin (1985a) suggest considering special conditions where controlling the means enables one to control all expected differences between the multivariate treated and control population distributions, which is the ultimate goal of matching. The most general version of these assumptions now require:

- (a) X is drawn randomly from a specified population \mathbf{X} ,
- (b) The population distribution for \mathbf{X} is an ellipsoidally symmetric density (Rubin and Thomas, 1992) or a discriminant mixture of proportional ellipsoidally symmetric densities (Rubin and Stuart, 2006), and
- (c) The matching algorithm applied is invariant to affine transformations of \mathbf{X} .

With these conditions, there is no risk of decreasing any type of expected imbalance in some variables while increasing it in others. Checking balance in this situation involves checking only the difference in means between the treated and control groups for only one (and indeed, any one) covariate.

Although the requirement (c) can be satisfied (e.g., by propensity score matching, unweighted Mahalanobis matching, discriminate matching), assumptions (a) and (b) rarely hold (and are almost never known to hold) in observational data. Rubin and Thomas

(1996) give some simulated examples where certain violations of these conditions still yield the desired properties for propensity score and Mahalanobis matching, but the practical problem of improving balance on one variable leading to a reduction in balance on others is very common in real applications in many fields. Of course, these matching methods are only *potentially EPBR*, since to apply them to real data requires the additional assumptions (a) and (b).

3 The Monotonic Imbalance Bounding Class

We build our new class of matching methods in six steps, by generalizing and modifying the definition of EPBR. First, we drop any assumptions about the data, such as conditions (a) and (b). Second, we focus on the actual *in-sample* imbalance, as compared to EPBR's goal of increasing *expected* balance. Of course, efficiency of the ultimate causal quantity of interest is a function of in-sample, not expected, balance, and so this can be important (and it explains otherwise counterintuitive results about EPBR methods, such as that matching on the estimated propensity score is more efficient than the true score, see Hirano, Imbens and Ridder 2003). Let $\bar{X}_{n_T,j}$, $\bar{X}_{n_C,j}$ and $\bar{X}_{m_T,j}$, $\bar{X}_{m_C,j}$ denote the pre-match and post-match sample means, for variable X_j , $j = 1, \dots, k$, for the subsamples of treated and control units. Then, third, we replace the equality in (2) by an inequality, and focus on the variable-by-variable relationship $|\bar{X}_{m_T,j} - \bar{X}_{m_C,j}| \leq \gamma_j |\bar{X}_{n_T,j} - \bar{X}_{n_C,j}|$ which we rewrite as

$$|\bar{X}_{m_T,j} - \bar{X}_{m_C,j}| \leq \delta_j, \quad j = 1, \dots, k, \quad (3)$$

where $\delta_j = \gamma_j |\bar{X}_{n_T,j} - \bar{X}_{n_C,j}|$. Fourth, we require δ_j to be chosen ex ante and let m_T and m_C to be determined by the matching algorithm instead of the reverse as under EPBR.

Equation (3) states that the *maximum imbalance* between treated and matched control units, as measured by the absolute difference in means for variable X_j , is bounded from above by the constant δ_j . Analogous to EPBR, one would usually prefer the situation

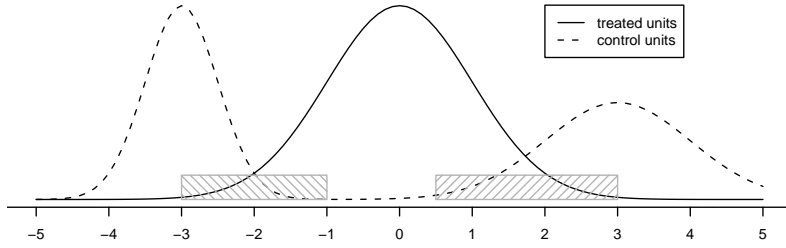


Figure 1: An example of a covariate for which minimizing mean-imbalance may be harmful. The example also shows that increasing mean-imbalance for this variable under MIB can be used to match more relevant features of the distributions (such as the shaded areas), without hurting mean-imbalance on other variables. This would be impossible under EPBR.

when the bound on imbalance is reduced due to matching, $\gamma_j = \delta_j / |\bar{X}_{n_T, j} - \bar{X}_{n_C, j}| < 1$, although this is not (yet) guaranteed by a method in this class.

To motivate the next change, consider data where the subsample of treated units has a unimodal distribution with a sample mean zero, and the control group has a bimodal distribution with almost zero empirical mean (see Figure 1). Then, reducing the difference in means in these data with a matching algorithm will be difficult. Instead, one would prefer locally good matches taken from where distributions overlap the most (see the two shaded boxes). Using these regions containing good matches may increase the mean imbalance by construction, but overall balance between the groups will greatly improve.

Thus, fifth, we generalize (3) from mean imbalance to a general measure of imbalance. Denote by $\mathcal{X}_{n_T} = [(X_{i1}, \dots, X_{ik}), i \in \mathcal{T}]$ the subset of the rows of treated units, and similarly for \mathcal{X}_{n_C} , \mathcal{X}_{m_T} and \mathcal{X}_{m_C} . We also replace the difference in means with a generic distance $D(\cdot, \cdot)$. Further, instead of the empirical means, we make use of a generic function of the sample, say $f(\cdot)$. This function may take as argument one variable X_j at time, or more, for example if we want to consider covariances. This leads us to the intermediate definition:

Definition 2 (Imbalance Bounding (IB)). *A matching method is Imbalance Bounding on the function of the data $f(\cdot)$ with respect to a distance $D(\cdot, \cdot)$, or simply $IB(f, D)$, if*

$$D(f(\mathcal{X}_{m_T}), f(\mathcal{X}_{m_C})) \leq \delta \tag{4}$$

where $\delta > 0$ is a scalar.

In a sense, EPBR is a version of IB if we take $D(x, y) = E(x - y)$, $f(\cdot)$ the sample mean, i.e. $f(\mathcal{X}_{m_T}) = \bar{X}_{m_T}$ and $f(\mathcal{X}_{m_C}) = \bar{X}_{m_C}$, $\delta = \gamma D(f(\mathcal{X}_{n_T}), f(\mathcal{X}_{n_C}))$, the inequality replaces the equality, and $\gamma < 1$. Although quite abstract, IB becomes natural when $f(\cdot)$ and $D(\cdot, \cdot)$ are specified. Assume $f(\cdot) = f_j(\cdot)$ is a function solely of the marginal empirical distribution of X_j . Then consider the following special cases:

- Let $D(x, y) = |x - y|$ and $f_j(\mathcal{X})$ denote the sample mean for the variable X_j of the observations in the subset \mathcal{X} . Then, (4) becomes (3), i.e. $|\bar{X}_{m_T, j} - \bar{X}_{m_C, j}| \leq \delta_j$. Similarly, if $f_j(\cdot)$ is the sample variance, the k -th centered moment, the q -th quantile, etc.
- If $f_j(\cdot)$ is the empirical distribution function of X_j , and $D(\cdot, \cdot)$, the sup-norm distance, then (4) is just the Kolmogorov distance, and if a nontrivial bound δ_j exists, then an IB methods would control the distance between the full distributions of the treated and control groups.
- Let $D(x, y) = |x|$ and $f(\cdot) = f_{jk}(\cdot)$ is the covariance of X_j and X_k and $\delta = \delta_{jk}$; then $|\text{Cov}(X_j, X_k)| \leq \delta_{jk}$.
- In Section 5 we introduce a global measure of multivariate imbalance denoted \mathcal{L}_1 in (6), which is also a version of $D(f(\cdot), f(\cdot))$.

To introduce our final step, we need some additional notation. As in Definition 2, let f be any function of the empirical distribution of covariate X_j of the data (such as the mean, variance, quantile, histogram, etc). Let $\pi, \pi' \in \mathbb{R}_+^k$ be two non-negative k -dimensional vectors and let the notation $\pi \preceq \pi'$ require that the two vectors π and π' be equal on all indexes except for a subset $J \subseteq \{1, \dots, k\}$, for which $\pi_j < \pi'_j$, $j \in J$. For a given function $f(\cdot)$ and a distance $D(\cdot, \cdot)$ we denote by $\gamma_{f, D}(\cdot) : \mathbb{R}_+^k \rightarrow \mathbb{R}_+$ a monotonically increasing function of its argument, i.e. if $\pi \preceq \pi'$ then $\gamma_{f, D}(\pi) \leq \gamma_{f, D}(\pi')$. Then our last step gives the definition of the new class:

remaining variables. This property is especially useful if we conceptualize the maximum imbalance in a variable as the maximal measurement error one can tolerate. For example, for many applications, we can probably tolerate an imbalance of 2 pounds in weighting people (since individuals can vary this much over the course of a day), 5 years of difference in age (for middle ages), or a year or two of education not near the threshold of graduation from high school, college, etc. Once these thresholds are set, an MIB method guarantees that no matter how much other variables imbalance is adjusted, these maxima will not change.

4 Examples and Comparisons

Well-known matching methods within the (potentially) EPBR class include nearest neighbor matching based on propensity scores or Mahalanobis distance. These methods are not MIB, because the number of matched observations (m_T, m_C) must be an outcome of the method rather than of a tuning parameter. These and other nearest neighbor matching methods applied with a scalar caliper, even when (m_T, m_C) is an outcome of the method, are not MIB because the dimension of the tuning parameter π in the definition has to be k in order to have separability as in (5). Caliper matching as defined in Cochran and Rubin (1973) is not MIB because of the orthogonalization and overlapping regions; without orthogonalization, it is MIB if applied variable by variable. (Cochran and Rubin (1973)[[p.420] also recognized that tight calipers control all linear and nonlinear imbalance under certain circumstances.) Coarsened exact matching (CEM), where exact matching is applied after each variable is separately coarsened (see Section 5), is MIB. Non-MIB methods can usually be made MIB if they operate within CEM's coarsened strata, so long as the coarsened strata take precedence in determining matches.

Both EPBR and MIB classes are designed to avoid, in different ways, the problem of making balance worse on some variables while trying to improve it for others, a serious practical problem in real applications. With additional assumptions about the data gener-

ation process, EPBR means that the degree of imbalance changes for all variables at the same time by the same amount; MIB, without extra assumptions on the data, means that changing one variable's imbalance does not affect the maximum imbalance for the others.

Neither class can guarantee both a bound on the level of imbalance and, at the same time, a prescribed number of matched observations. In EPBR methods, the user chooses the matched sample size ex ante and computes balance ex post, whereas in MIB methods the user choose the maximal imbalance ex ante and produces a matched sample size ex post.

In real data sets that do not necessarily meet EPBR's assumptions, no results are guaranteed from potentially EPBR methods and so balance may be reduced for some or all variables. Thus, methods that are potentially EPBR require verifying ex post that balance has improved. For example, in propensity score matching, the functional form of the regression of T on X must be correct, but the only way to verify this is to check balance ex post. In practical applications, researchers commonly find that substantial tweaking is required to avoid degrading mean balance on at least some variables, and other types of balance are rarely checked or reported.

Under MIB, imbalance in the means, other moments, co-moments, interactions, nonlinearities, and the full multivariate distribution of the treated and control groups are improved, without hurting maximum imbalance on other variables and regardless of the data type. The actual level of balance achieved by MIB methods can be better than the maximum level set ex ante, but only the bound is guaranteed.

In practice, MIB methods may sometimes generate too few matched observations, which indicates that either the maximum imbalance levels chosen are too restrictive (e.g., too stringent a caliper), or that the data set cannot be used to make inferences without high levels of model dependence. In observational data, analyzing counterfactuals too far from the data to make reliable inferences is a constant concern and so MIB's property of sometimes producing no matched observations can also be considered an important advantage.

By attempting to reduce expected imbalance, potentially EPBR methods attempt to approximate with observational data the classic *complete randomization* experimental design, with each unit randomly assigned a value of the treatment variable. In contrast, MIB methods approximate the *randomized block* experimental design, where values of the treatment variable are assigned within strata defined by the covariates. Although both are unbiased, randomized block designs have perfect balance in each data set on all observed covariates, whereas complete randomization designs are balanced only on average across experiments, with no guarantees for the one experiment being run. Randomized block designs, as a result, are considerably more efficient, powerful, and robust (see Box, Hunger and Hunter 1978, p.103 and Imai, King and Stuart 2008); in an application by Imai, King and Nall (2009), complete randomization gives standard errors as much as six times larger than the corresponding randomized block design.

Finally, a consensus recommendation of the matching literature is that units from the control group outside the range of the data of the treated group should be discarded as they lead to unacceptable levels of model dependence. This means that the application of potentially EPBR methods must be preceded by a separate method for eliminating these risky observations. One way to eliminate extreme counterfactuals is to discard control units which fall outside the convex-hull (King and Zeng, 2007) or the hyper-rectangle (Iacus and Porro, 2009) delimited by the empirical distribution of the treated units. Unfortunately, these and other two-step matching approaches are not even potentially EPBR. In contrast, MIB methods which eliminate this extrapolation region (sometimes even without a separate step) are easy to construct.

5 Coarsened Exact Matching as an MIB Method

We introduce here a specific member of the MIB class of matching methods that comes from the diverse set of approaches based on subclassification (aka “stratification” or “intersection” methods). We call this particular method CEM for “Coarsened Exact Match-

ing” (or “Cochran Exact Matching” since the first formal analysis of any subclassification-based method appeared in Cochran 1968).

Definition CEM requires three steps: (1) Coarsen each of the original variables in \mathbf{X} as much as the analyst is willing into, say, $C(\mathbf{X})$ (e.g., years of education might be coarsened into grade school, high school, college, graduate school, etc.). (2) Apply exact matching to $C(\mathbf{X})$, which involves sorting the observations into strata, say $s \in \mathcal{S}$, each with unique values of $C(\mathbf{X})$. (3) Strata containing only control units are discarded; strata with treated and control units are retained; and strata with only treated units are used with extrapolated values of the control units or discarded if the analyst is willing to narrow the quantity of interest to the remaining set of treated units for which a counterfactual has been properly identified and estimated.

Denote by \mathcal{T}^s the treated units in stratum s , with count $m_T^s = \#\mathcal{T}^s$, and similarly for the control units, i.e. \mathcal{C}^s and $m_C^s = \#\mathcal{C}^s$. The number of matched units are, respectively for treated and controls, $m_T = \bigcup_{s \in \mathcal{S}} m_T^s$ and $m_C = \bigcup_{s \in \mathcal{S}} m_C^s$. Then for subsequent analysis, assign each matched unit i in stratum s , the following *CEM-weights* $w_i = 1$, if $i \in \mathcal{T}^s$ and $w_i = m_C/m_T \cdot m_T^s/m_C^s$, if $i \in \mathcal{C}^s$, with unmatched units receiving weight $w_i = 0$.

Coarsening Choices Because coarsening is so closely related to the substance of the problem being analyzed and works variable-by-variable, data analysts understand how to decide how much each variable can be coarsened without losing crucial information. Indeed, even before the analyst obtains the data, the quantities being measured are typically coarsened to some degree. Variables like gender or the presence of war coarsen away enormous heterogeneity within the given categories. Data analysts also recognize that many measures include some degree of noise and, in their ongoing efforts to find a signal, often voluntarily coarsen the data themselves. For example, 7-point partisan identification scales are recoded as Democrat, independent, and Republican; Likert issue questions as agree, neutral, and disagree; and multi-party vote returns as winners and losers. Many use a small number of categories to represent religion, occupation, U.S. Security and Ex-

change Commission industry codes, and international classification of disease codes, and many others. Indeed, epidemiologists routinely dichotomize all their covariates on the theory that grouping bias is much less of a problem than getting the functional form right. Although coarsening in CEM is safer than at the analysis stage, the two procedures are similar in spirit since the discarded information in both is thought to be relatively unimportant — small enough with CEM to trust to statistical modeling.

For continuous variables, coarsening can cut the range of the variable X_j into equal intervals of length ϵ_j . If the substance of the problem suggests different interval lengths, we use ϵ_j to denote the maximum length. For categorical variables, coarsening may correspond to grouping different levels of the variable.

CEM as an MIB method We prove here that CEM is a member of the MIB class with respect to the mean, the centered absolute k^{th} moment, and the empirical and weighted quantiles. Other similar properties can be proved along these lines as well. Beginning with Definition 3, let $D(x, y) = |x - y|$, $\pi_j = \epsilon_j$, $\gamma_j = \gamma_j(\epsilon_j)$ be a function of ϵ_j , and the function $f(\cdot)$ vary for the different propositions. Changing ϵ_j for one variable then does not affect the imbalance on the other variables.

Denote the weighted mean for the treated and control units respectively as $\bar{X}_{m_T, j}^w = \frac{1}{m_T} \sum_{i \in T} X_{ij} w_i$ and $\bar{X}_{m_C, j}^w = \frac{1}{m_C} \sum_{i \in C} X_{ij} w_i$.

Proposition 1. For $j = 1, \dots, k$, $|\bar{X}_{m_T, j}^w - \bar{X}_{m_C, j}^w| \leq \epsilon_j$.

Proof of Proposition 1. Let us introduce the means by strata: $\bar{X}_{m_T^s, j} = \frac{1}{m_T^s} \sum_{i \in T^s} X_{ij}$, $\bar{X}_{m_C^s, j} = \frac{1}{m_C^s} \sum_{i \in C^s} X_{ij}$. Then $\bar{X}_{m_T, j}^w = \frac{1}{m_T} \sum_{i \in T} X_{ij} w_i = \frac{1}{m_T} \sum_{s \in S} \sum_{i \in T^s} X_{ij} = \frac{1}{m_T} \sum_{s \in S} m_T^s \bar{X}_{m_T^s, j}$ and $\bar{X}_{m_C, j}^w = \frac{1}{m_C} \sum_{i \in C} X_{ij} w_i = \frac{1}{m_C} \sum_{s \in S} \sum_{i \in C^s} X_{ij} \frac{m_C}{m_T} \frac{m_T^s}{m_C^s} = \frac{1}{m_T} \sum_{s \in S} m_T^s \bar{X}_{m_C^s, j}$. Hence, given that the mean is internal, in each stratum observations are at most far as ϵ_j ; thus, $|\bar{X}_{m_T, j}^w - \bar{X}_{m_C, j}^w| \leq \sum_{s \in S} \frac{m_T^s}{m_T} |\bar{X}_{m_T^s, j} - \bar{X}_{m_C^s, j}| \leq \sum_{s \in S} \frac{m_T^s}{m_T} \epsilon_j = \epsilon_j$. \square

Let R_j be the range of variable X_j and let $\theta_j = \max_{\epsilon_j \geq \epsilon_j^*} (\lceil R_j / \epsilon_j \rceil)$, where $\lceil x \rceil$ is the first integer greater or equal to x . In the definition of θ_j , ϵ_j^* is any reasonable strictly positive value, e.g. the lowest value of ϵ_j which generates at most n_T non empty intervals in CEM.

Proposition 2. Let $k \geq 1$ and consider the centered absolute k -th moment for variable X_j for the treated and control units as $\bar{\mu}_{T,j}^k = \frac{1}{m_T} \sum_{s \in \mathcal{S}} \sum_{i \in \mathcal{T}^s} |X_{ij} - \bar{X}_{m_T,j}^w|^k w_i$ and $\bar{\mu}_{C,j}^k = \frac{1}{m_C} \sum_{s \in \mathcal{S}} \sum_{i \in \mathcal{C}^s} |X_{ij} - \bar{X}_{m_C,j}^w|^k w_i$. Then, $|\bar{\mu}_{T,j}^k - \bar{\mu}_{C,j}^k| \leq \epsilon_j^k (\theta_j + 1)^k$, $j = 1, \dots, k$, and $\epsilon_j \geq \epsilon_j^*$.

Proof of Proposition 2. We first rewrite $\bar{\mu}_{C,j}^k$

$$\bar{\mu}_{C,j}^k = \frac{1}{m_C} \sum_{s \in \mathcal{S}} \sum_{i \in \mathcal{C}^s} |X_{ij} - \bar{X}_{m_C,j}^w|^k w_i \leq \frac{1}{m_C} \sum_{s \in \mathcal{S}} \sum_{i \in \mathcal{C}^s} (|X_{ij} - \bar{X}_{m_T,j}^w| + |\bar{X}_{m_T,j}^w - \bar{X}_{m_C,j}^w|)^k w_i.$$

and then apply the binomial expansion to the inner term of the summation

$$(|X_{ij} - \bar{X}_{m_T,j}^w| + |\bar{X}_{m_T,j}^w - \bar{X}_{m_C,j}^w|)^k = \sum_{h=0}^k \binom{k}{h} |X_{ij} - \bar{X}_{m_T,j}^w|^h |\bar{X}_{m_T,j}^w - \bar{X}_{m_C,j}^w|^{k-h}$$

by Proposition 1 we can write

$$\begin{aligned} (|X_{ij} - \bar{X}_{m_T,j}^w| + |\bar{X}_{m_T,j}^w - \bar{X}_{m_C,j}^w|)^k &\leq \sum_{h=0}^k \binom{k}{h} |X_{ij} - \bar{X}_{m_T,j}^w|^h \epsilon_j^{k-h} \\ &\leq \epsilon_j^k \sum_{h=0}^k \binom{k}{h} |R_j|^h \epsilon_j^{-h} = \epsilon_j^k \sum_{h=0}^k \binom{k}{h} \left| \frac{R_j}{\epsilon_j} \right|^h \leq \epsilon_j^k \sum_{h=0}^k \binom{k}{h} \theta_j^h 1^{k-h} = \epsilon_j^k (\theta_j + 1)^k \end{aligned}$$

Therefore, $\bar{\mu}_{C,j}^k \leq \epsilon_j^k (\theta_j + 1)^k \frac{1}{m_C} \sum_{s \in \mathcal{S}} \sum_{i \in \mathcal{C}^s} w_i = \epsilon_j^k (\theta_j + 1)^k$ because $\frac{1}{m_C} \sum_{s \in \mathcal{S}} \sum_{i \in \mathcal{C}^s} w_i = \frac{1}{m_C} \sum_{s \in \mathcal{S}} \sum_{i \in \mathcal{C}^s} \frac{m_C}{m_T} \frac{m_T^s}{m_C^s} = \frac{1}{m_T} \sum_{s \in \mathcal{S}} m_C^s \frac{m_T^s}{m_C^s} = 1$. Since $\frac{1}{m_T} \sum_{s \in \mathcal{S}} \sum_{i \in \mathcal{T}^s} w_i = 1$. The same bound exists for $\bar{\mu}_{T,j}^k$, so their absolute difference is $|\bar{\mu}_{T,j}^k - \bar{\mu}_{C,j}^k| \leq \epsilon_j^k (\theta_j + 1)^k$. \square

Proposition 3. Assume one-to-one matching. Denote by $X_{m_T,j}^q$ the q^{th} empirical quantile of the distribution of the treated units for covariate X_j , and similarly $X_{m_C,j}^q$. Then, $|X_{m_T,j}^q - X_{m_C,j}^q| \leq \epsilon_j$ for $j = 1, \dots, k$.

Proof of Proposition 3. Consider the q^{th} empirical quantiles of the distribution of the treated and control units, $X_{m_T,j}^q$ and $X_{m_C,j}^q$. That is, $X_{m_T,j}^q$ is the q^{th} ordered observation of the subsample of m_T matched treated units, and similarly for $X_{m_C,j}^q$. In one-to-one matching, the first treated observation is matched against the first control observation in the first

stratum and, in general, the corresponding quantiles belong to the same strata. Therefore,

$$|X_{m_T,j}^q - X_{m_C,j}^q| < \epsilon_j. \quad \square$$

Define the weighted empirical distribution functions for treated group as $F_{m_T,j}^w(x) = \sum_{X_{ij} \leq x, i \in T} \frac{w_i}{m_T}$ and for the control group as $F_{m_C,j}^w(x) = \sum_{X_{ij} \leq x, i \in C} \frac{w_i}{m_C}$. Define the q -th quantile of the weighted distribution $X_{m_T,j}^{q,w}$ as the first observation in the sample such that $F_{m_T,j}^w(x) \geq q$ and similarly for $X_{m_C,j}^{q,w}$.

Proposition 4. *Assume that the support of variable X_j is cut on subintervals of exact length ϵ_j . Then $|X_{m_T,j}^{q,w} - X_{m_C,j}^{q,w}| \leq \epsilon_j$ for $j = 1, \dots, k$.*

Proof of Proposition 4. Consider the generic stratum $[a_s, b_s]$, $s \in \mathcal{S}$, where a_s is the left-most cut-point of the discretization and $b_s = a_s + \epsilon_j$. For simplicity, take $s = 1$, so that $F_{m_T,j}^w(a_1) = F_{m_C,j}^w(a_1) = 0$. Then $F_{m_T,j}^w(b_1) = m_T^{s=1}/m_T$ because there are at most $m_T^{s=1}$ treated units less than or equal to b_1 . Similarly, for the weighted distribution of the control units we have

$$F_{m_C}^w(b_1) = \frac{m_C^{s=1}}{m_C} \cdot \frac{m_C}{m_T} \frac{m_T^{s=1}}{m_C^{s=1}} = \frac{m_T^{s=1}}{m_T}$$

Thus, for each stratum, $F_{m_T,j}^w(b_s) = m_T^s/m_T = F_{m_C,j}^w(b_s)$, and hence the difference between weighted empirical distribution functions at the end points of each stratum $[a_s, b_s]$ is always zero. Therefore, the weighted quantiles of the same order for treated and control units always belong to the same stratum and hence the difference between them is at most ϵ_j . \square

On Filling CEM Strata A problem may occur with MIB methods if too many treated units are discarded. This can be fixed of course by adjusting the choice of maximum imbalance, but it is reasonable to ask how often this problem occurs for a “reasonable” choice in real data. The worry for MIB methods is curse of dimensionality, which in this context means that the number of hyper-rectangles, and thus the number of possible strata $\#C(X_1) \times \dots \times \#C(X_k)$, is typically very large. For example, suppose \mathbf{X} is composed of 10,000 observations on 20 variables drawn from independent normal densities. Since

20-dimensional space is enormous, odds are that no treated unit will be anywhere near any control unit. In this situation, even very coarse bins under CEM will likely produce no matches. For example, with only two bins for each variable, the 10,000 observations would need to be sorted into 2^{20} possible strata, in which case the probability would be extremely small of many stratum winding up with both a treated and control unit.

Although EPBR methods fix the number of matches ex ante (on the hope that imbalance would be reduced on average across experiments), no EPBR matching method would provide much help in making inferences from these data. The fact that in these data CEM would likely produce very few matches may be regarded as a disadvantage, since some estimate may still be desired no matter how model dependent, it is better regarded as an advantage in real applications, since no method of matching will help produce high levels of local balance in this situation.

Fortunately, for two reasons, this problem turns out not to be much of an issue in practice. First, and most importantly, real data sets have much more highly correlated data structures than independent draws in the simulation above, and so CEM in practice tends to produce reasonable numbers of matches. This has certainly been our overwhelming experience in the numerous data sets we have analyzed.

And second, if the reservoir of control units is sufficiently large, it is possible to derive, following the proof of Proposition 1 in Abadie and Imbens (2009), an exponential bound on the probability that the number of CEM strata with unmatched treated units remains positive. In particular, at rate $n_C = O(n_T^{1/r})$, with $r \geq k$, where k is the number of continuous pre-treatment covariates, the number of cells that contain only (unmatched) treated units goes to zero with the number of treated units n_T in the sample, if the number of control units n_C grows appropriately.

An Illustration of Multivariate Imbalance Reduction Most matching methods were designed to reduce imbalance in the *mean* of each pre-treatment variable between the treated and control groups. (A notable exception is the full optimal matching algorithm,

Rosenbaum (2002), which is designed to minimize functions such as the average of the local distances among each matched treated and control units, although these methods are not MIB because of their use of a scalar imbalance metric.) Of course, reducing mean imbalance does not necessarily reduce the full multidimensional imbalance between the treated and control groups. We thus now complement Section 5’s proofs and show we can control imbalance for each variable X_j via coarsening; we do this by directly measuring the distance between the full multidimensional histograms of the populations of the treated and control units. Multidimensional histograms are obtained by cross tabulation of the coarsened pre-treatment variables. Let $H(X_1)$ be the set of distinct values generated by the coarsening on variable X_1 , i.e., the set of intervals into which the support of variable X_1 has been cut. Then, the multidimensional histogram is constructed from the set of cells generated by the Cartesian product $H(X_1) \times \cdots \times H(X_k) = H(\mathbf{X})$. There is no universal way to define a proper coarsening for the propose of balance assessment, so one practical option is to use a very fine coarsening. A crucial point is that this coarsening should be different and finer than the coarsening used in CEM, although it otherwise is not be related to or based on coarsening in CEM.

The proposed measure is then the (possibly weighted) distance between two multidimensional histograms measured by the L^1 norm. Let f and g be the relative empirical frequency distributions for two the treated and control units. Let $f_{\ell_1 \dots \ell_k}$ be the relative frequency for observations belonging to the cell with coordinates $\ell_1 \cdots \ell_k$ of the multivariate cross-tabulation, and similarly for $g_{\ell_1 \dots \ell_k}$.

Definition 4. *The multivariate imbalance measure is*

$$\mathcal{L}_1(f, g) = \frac{1}{2} \sum_{\ell_1 \dots \ell_k \in H(\mathbf{X})} |f_{\ell_1 \dots \ell_k} - g_{\ell_1 \dots \ell_k}|. \quad (6)$$

An important property of this measure is that the typically numerous empty cells do not affect $\mathcal{L}_1(f, g)$, and so the summation in (6) has at most n nonzero terms. The relative frequencies also control for what may be different sample sizes for the treated and control

groups. If the two distributions of data are completely separated (up to the fine coarsening of the histogram), then $\mathcal{L}_1 = 1$; if the two distributions overlaps exactly, then $\mathcal{L}_1 = 0$. In all other cases, $\mathcal{L}_1 \in (0, 1)$. For a given coarsening $H(\mathbf{X})$, the values of \mathcal{L}_1 provide useful relative information in making comparisons. Indeed, if say $\mathcal{L}_1 = 0.6$, then only 40% of the density of the two histograms overlap. Let f^m and g^m denote the distributions of the matched treated and control units corresponding to the distributions f, g of the original unmatched data. Then a good matching method will result in matched sets such that $\mathcal{L}_1(f^m, g^m) \leq \mathcal{L}_1(f, g)$. Of course, to make coherent matching comparisons, the coarsening $H(\mathbf{X})$ must remain fixed.

Although the point is simple mathematically, a large empirical literature suggests that it may be worth clarifying why controlling for one dimensional distributions is not enough to control the global imbalance of the joint distribution (outside the special cases such as multivariate Gaussians). Indeed, let $p_i = P(T = 1 | X_{i1}, X_{i2}, \dots, X_{ik}) = 1/[1 + \exp\{-\beta_0 - \sum_{j=1}^k \beta_j X_{ij}\}]$ be the logistic model for the propensity score. And let \hat{p}_i be the propensity score estimated by maximum likelihood. Set $w_i = 1 - \hat{p}_i$, for $i \in \mathcal{T}$ and $w_i = \hat{p}_i$ for $i \in \mathcal{C}$.

Matching in some way based on this propensity score in arbitrary data has no known theoretical properties (and does not perform well in these data), and so for clarification we switch to propensity score weighting, which is simpler in this situation. Denote the weighted means for treated and control units as $\bar{X}_{T,j}^w = \sum_{i \in \mathcal{T}} X_{ij} w_i / \sum_{i \in \mathcal{T}} w_i$ and $\bar{X}_{C,j}^w = \sum_{i \in \mathcal{C}} X_{ij} w_i / \sum_{i \in \mathcal{C}} w_i$. Then, it is well known that $\bar{X}_{T,j}^w = \bar{X}_{C,j}^w$.

Although this weighting guarantees the elimination of all mean imbalance, the multi-dimensional distribution of the data may be still highly imbalanced. A numerical example illustrates this fact. We use the Lalonde (1986) data, a commonly used example in the matching literature. The role of the variables are not relevant to our illustration, so we do not describe the data but the interested reader can refer to the original paper. The multidimensional imbalance on the raw data is equal to $\mathcal{L}_1 = 0.735$ (where we calculate \mathcal{L}_1 based on 20 intervals for the four continuous variables and no coarsening for the

variable	raw data	pscore weighting	CEM
age	0.18	0.00	0.19
education	0.19	0.00	0.01
re74	-101.49	0.00	7.20
re75	39.42	0.00	12.21
nodegree	-0.08	0.00	0.00
black	0.00	0.00	0.00
married	0.01	0.00	0.00
hispanic	-0.02	0.00	0.00
u74	-0.02	0.00	0.00
u75	-0.05	0.00	0.00
\mathcal{L}_1	0.735	0.730	0.599

Table 1: Differences in means for each variable (I_1) and global imbalance measure (\mathcal{L}_1) on raw data from Lalonde (1986), after propensity score weighting, and following CEM matching. Variable names are as in Lalonde’s original data set.

six categorical variables). The univariate (I_1) and global (\mathcal{L}_1) imbalance measures are given in Table 1 for the raw data, propensity score weighting, and CEM. After applying propensity score weighting (see middle column) we get, as expected, a perfect (weighted) match on the difference in means for all variables, but the overall global imbalance is equal to $\mathcal{L}_1 = 0.730$, which is almost the same as the original data, i.e. 99.3% of the original imbalance value. However, after matching the raw data with CEM (which we do by coarsening the four variables into 10 intervals), the data are more balanced because CEM pruned observations that would have led to large extrapolations. This can be seen in the last line of the table which gives the global imbalance, which has now been substantially reduced to $\mathcal{L}_1 = 0.599$, i.e. 81.5% of the original imbalance.

This example thus shows that simple weighting can reduce or eliminate mean imbalance without improving global multivariate imbalance. The same of course holds for any matching algorithm designed to improve imbalance computed one variable at a time. CEM, as an MIB method, and \mathcal{L}_1 as a measure of imbalance, provides a simple way around these problems.

6 MIB vs. EPBR Methods under EPBR-Compliant Data

We now simulate data best suited for EPBR methods and compare CEM, an MIB matching method, to the propensity score (PSC) and Mahalanobis distance (MAH) matching from the EPBR class of methods. We show that the MIB properties of CEM (in particular, the in-sample multivariate imbalance reduction) enables CEM to outperform EPBR methods even in data generated to optimize EPBR performance.

We begin by replicating Gu and Rosenbaum (1993). This involves drawing two independent multivariate normal data sets: $\mathbf{X}_T \sim N_5(\mu_T, \Sigma)$ and $\mathbf{X}_C \sim N_5(\mu_C, \Sigma)$, with common variances $(6, 2, 1, 2, 1)$ and covariances, $(2, 1, 0.4, -1, -0.2, 1, -0.4, 0.2, 0.4, 1)$, and means vectors $\mu_T = (0, 0, 0, 0, 0)$ and $\mu_C = (1, 1, 1, 1, 1)$. We randomly sample $n_T = 1,000$ treated units from \mathbf{X}_T and $n_C = r \cdot n_T$ control units from \mathbf{X}_C with $r = 1, 3$. For CEM, we coarsen each covariate into 8 intervals of equal length. We also allow PSC and MAH the advantage of matching with replacement, in order to help them avoid trivial solutions. MAH and PSC thus match $m_T = 1,000$ treated units against a variable number m_C of control units, whereas CEM selects both treated and control units.

In these data, the properties of EPBR imply that MAH and PSC matching will optimally minimize expected mean imbalance (Rosenbaum and Rubin, 1985*b*). In contrast, CEM is designed to reduce local multivariate imbalance, that is, the maximum distance between a treated unit and the corresponding matched control units. We can measure these with \mathcal{L}_1 overall, and the average of the difference in means between treated and control units stratum by stratum for each variable, which we denote I_2 . (For \mathcal{L}_1 we divided each covariate into 11 equally spaced intervals to evaluate the k -dimensional histogram.)

Overall, we find that CEM is as good as the other methods in terms of the difference in means (I_2), for which these other methods were designed, but CEM is superior in matching all other local and multivariate aspects of the treated and control distributions, as measured by the average local imbalance I_2 and multivariate \mathcal{L}_1 .

These results can be seen in Table 2 which reports results for 1,000 (top two panels)

and 3,000 (bottom two panels) control units. The table also reports, I_1 , I_2 and \mathcal{L}_1 . The table show that MAH is systematically worse than PSC and CEM in terms of I_1 . As would be expected when there is more to the data than just the mean, CEM is better than PSC on the first two covariates (which have much larger variances) whereas the contrary is true for the remaining covariates. Of course, all these differences are relatively small, and so from that perspective we could reasonably conclude that they have about the same performance.

However, in terms of local imbalance measured by I_2 , CEM considerably outperforms PSC and MAH on all covariates. So in terms of I_2 , CEM dominates MAH which in turn dominates PSC. The same ordering is produced by \mathcal{L}_1 . Imbalance reduction as measured by \mathcal{L}_1 (i.e., compared to the raw data) is very small for MAH and PSC and quite large for CEM. This means that CEM is indeed greatly reducing the distance between the two k -dimensional distributions of treated and control units. Since the two EPBR methods in these data are known to be optimal only in expectation, the additional advantage of CEM is coming from MIB's in-sample multivariate imbalance reduction property.

Other regularities emerges from this analysis as well: all methods perform about as well as the reservoir of control units (drawn from the same population) grows. MAH matching and CEM agree on the fact that not all the control units are good counterfactuals, and the numbers of control units selected does not differ drastically across methods.

7 Estimating the Causal Effect

A crucial issue in causal inference is identifying the precise quantity of interest to be estimated. This is an issue in observational data, which is often based on convenience samples and may include whatever relevant data happen to be available. However, the same issue applies to most randomized medical experiments, for example, since they are also based on convenience samples (such as patients who happen to show up at a research hospital). In these situations, the target causal effect is typically defined for the observed

Simulation 1: $n_T = 1,000, n_C = 1,000$.

	Difference in means I_1					m_T	m_C
	X_1	X_2	X_3	X_4	X_5		
Raw	1.00	1.00	1.00	1.00	1.00	1000	1000
CEM	0.04	0.02	0.06	0.06	0.04	341	340
MAH	0.20	0.20	0.20	0.20	0.20	1000	408
PSC	0.11	0.06	0.03	0.06	0.03	1000	616

	Local imbalance I_2					\mathcal{L}_1
	X_1	X_2	X_3	X_4	X_5	
Raw						1.24
CEM	0.42	0.26	0.17	0.22	0.19	0.78
MAH	0.56	0.36	0.29	0.36	0.29	1.13
PSC	2.38	1.25	0.74	1.25	0.74	1.18

Simulation 2: $n_T = 1,000, n_C = 3000$.

	Difference in means I_1					m_T	m_C
	X_1	X_2	X_3	X_4	X_5		
Raw	1.00	1.00	1.00	1.00	1.00	1000	3000
CEM	0.04	0.02	0.05	0.06	0.04	513	921
MAH	0.14	0.14	0.14	0.14	0.14	1000	625
PSC	0.07	0.04	0.02	0.04	0.02	1000	2157

	Local imbalance I_2					\mathcal{L}_1
	X_1	X_2	X_3	X_4	X_5	
Raw						1.17
CEM	0.38	0.24	0.16	0.21	0.17	0.75
MAH	0.51	0.32	0.25	0.32	0.25	0.89
PSC	2.40	1.26	0.75	1.26	0.75	0.99

Table 2: Imbalance in means (I_1) and average local imbalance (I_2) remaining after matching for each variable listed, X_1, \dots, X_5 , for the raw data (Raw), Coarsened Exact Matching (CEM), Mahalanobis Distance matching (MAH), and propensity score matching (PSC). Also reported are the number of treated m_T and control m_C units matched and multivariate imbalance, \mathcal{L}_1 . Results are averaged over 5,000 replications, with $n_T = 1,000, n_C = 1,000$ (top panel) and $n_C = 3,000$ (bottom panel).

units only, and no attempt is made to formally infer to a broader population.

One example of a quantity of interest defined for the sample data is the causal effect averaged over all the treated units, the sample average treatment effect on the treated: $SATT = \frac{1}{n_T} \sum_{i \in T} \{Y_i(1) - Y_i(0)\}$. SATT is an especially convenient definition for matching methods which prune (only) control units from a data set and so do not change the

estimand. In especially difficult data sets, however, some treated units may have no reasonable match among the available pool of control units. These treated units are easy to identify in MIB methods such as CEM, since matches are only made when they meet the ex ante specified level of permissible imbalance; under EPBR methods, all treated units are matched, no matter how deficient the set of available controls and so a separate analytical method must be applied to identify these units.

When reasonable control units do not exist for one or more treated units, SATT cannot be estimated without high levels of model dependence. In this situation, the analyst can choose to (a) create virtual controls for the unmatched treated units via extrapolation and modeling assumptions, (b) conclude that the data include insufficient information to estimate the target causal effect and give up, or (c) change the quantity of interest to the SATT defined for the subset of treated units that have good matches among the pool of controls. Since the data are deficient to the research question posed, all three options are likely to be unsatisfying, (a) because of model dependence, (b) because we learn nothing, and (c) because this is not the quantity we originally sought; although each of these options can be reasonable in some circumstances.

Although no better solution to the problem can be constructed, we offer here a way to think about this problem more broadly by combining all these options together. This process requires four steps. First, preprocess the data to remove the worst potential matches (and thus the most strained counterfactuals) from the set of available control units. This can be done easily using the convex hull or the hyper-rectangle approaches (see Section 4). Second, run CEM on these pre-processed data without the extreme counterfactuals and obtain $m_T \leq n_T$ treated units matched with $m_C \leq n_C$ control units. Third, use these results to split the entire set of treated units in the two groups of m_T matched and $n_T - m_T$ unmatched individuals.

Fourth, compute the SATT separately in the two groups as follows. For the m_T treated units, there exist m_C acceptable counterfactuals (as defined by the coarsening in CEM say), and so we can reliably estimate this “local SATT,” say $\hat{\tau}_{m_T}$, using only this subset of

treated units. Then, for the rest of the treated units, either extrapolate the model estimated on the matched units to obtain virtual counterfactuals for the unmatched treated units or consider all the unmatched units as a single CEM stratum and estimate the ATT locally. In either case, denote this estimate by $\hat{\tau}_{n_T - m_T}$.

Finally, calculate the overall SATT estimate $\hat{\tau}_{n_T}$ as the weighted mean of the two estimates:

$$\hat{\tau}_{n_T} = \frac{\hat{\tau}_{m_T} \cdot m_T + \hat{\tau}_{n_T - m_T} \cdot (n_T - m_T)}{n_T}.$$

This procedure keeps the overall quantity of interest, SATT, fixed and isolates the model dependent piece of the estimator so it can be studied separately and its effects on SATT isolated. In practice, analysts might wish to present $\hat{\tau}_{n_T}$, which is necessarily model dependent, as well as $\hat{\tau}_{m_T}$, which is well estimated (and not model dependent) but is based on only a subset of treated units.

8 Concluding Remarks

We offer a new class of matching methods that generalizes the only existing class proposed. This new monotonic imbalance bounding class enables the creation of methods that are easy to apply and which we show possess properties that should be of considerable interest to applied researchers. We offer Coarsened Exact Matching as one such example.

References

- Abadie, Alberto and Guido Imbens. 2009. “A Martingale Representation for Matching Estimators.” IZA Discussion Papers number 4073. <http://ftp.iza.org/dp4073.pdf>.
- Box, George E.P., William G. Hunter and J. Stuart Hunter. 1978. *Statistics for Experimenters*. New York: Wiley-Interscience.
- Cochran, William G. 1968. “The effectiveness of adjustment by subclassification in removing bias in observational studies.” *Biometrics* 24:295–313.
- Cochran, William G. and Donald B. Rubin. 1973. “Controlling bias in observational studies: A review.” *Sankhya: The Indian Journal of Statistics, Series A* 35, Part 4:417–466.

- Gu, X.S. and Paul R. Rosenbaum. 1993. "Comparison of multivariate matching methods: structures, distances, and algorithms." *Journal of Computational and Graphical Statistics* 2:405–420.
- Hirano, Keisuke, Guido W. Imbens and Geert Ridder. 2003. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica* 71(4, July):1161–1189.
- Ho, Daniel, Kosuke Imai, Gary King and Elizabeth Stuart. 2007. "Matching as Non-parametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15:199–236. <http://gking.harvard.edu/files/abs/matchp-abs.shtml>.
- Iacus, Stefano M. and Giuseppe Porro. 2009. "Random Recursive Partitioning: a matching method for the estimation of the average treatment effect." *Journal of Applied Econometrics* 24:163–185.
- Imai, Kosuke, Gary King and Clayton Nall. 2009. "The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation." *Statistical Science* 24(1):29–53. <http://gking.harvard.edu/files/abs/cluster-abs.shtml>.
- Imai, Kosuke, Gary King and Elizabeth Stuart. 2008. "Misunderstandings Among Experimentalists and Observationalists about Causal Inference." *Journal of the Royal Statistical Society, Series A* 171, part 2:481–502. <http://gking.harvard.edu/files/abs/matchse-abs.shtml>.
- King, Gary and Langche Zeng. 2007. "When Can History Be Our Guide? The Pitfalls of Counterfactual Inference." *International Studies Quarterly* (March):183–210. <http://gking.harvard.edu/files/abs/counterf-abs.shtml>.
- Lalonde, Robert. 1986. "Evaluating the Econometric Evaluations of Training Programs." *American Economic Review* 76:604–620.
- Rosenbaum, Paul R. 2002. *Observational Studies, 2nd Edition*. New York, NY: Springer Verlag.
- Rosenbaum, Paul R. and Donald B. Rubin. 1985a. "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score." *The American Statistician* 39:33–38.
- Rosenbaum, P.R. and D.B. Rubin. 1985b. "The Bias Due to Incomplete Matching." *Biometrics* 41(1):103–116.
- Rubin, Donald. 1976a. "Inference and Missing Data." *Biometrika* 63:581–592.
- Rubin, Donald B. 1976b. "Multivariate Matching Methods That are Equal Percent Bias Reducing, I: Some Examples." *Biometrics* 32(1):109–120.
- Rubin, Donald B. 1976c. "Multivariate Matching Methods that are Equally Percent Bias Reducing, II: Maximums on Bias Reduction for Fixed Sample Sizes." *Biometrics* 32:121–132.
- Rubin, Donald B. and Elizabeth A. Stuart. 2006. "Affinely invariant matching methods with discriminant mixtures of proportional ellipsoidally symmetric distributions." *Annals of Statistics* 34(4):1814–1826.
- Rubin, Donald B. and Neal Thomas. 1992. "Affinely Invariant Matching methods with Ellipsoidal Distributions." *Annals of Statistics* 20(2):1079–1093.
- Rubin, Donald B. and Neal Thomas. 1996. "Matching Using Estimated Propensity Scores, Relating Theory to Practice." *Biometrics* 52:249–264.