Multivariate Matching Methods That Are Monotonic Imbalance Bounding

Stefano M. IACUS, Gary KING, and Giuseppe PORRO

We introduce a new "Monotonic Imbalance Bounding" (MIB) class of matching methods for causal inference with a surprisingly large number of attractive statistical 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 analyze in more detail a member of this class, called Coarsened Exact Matching, whose properties we analyze from this new perspective. We offer a variety of analytical results and numerical simulations that demonstrate how members of the MIB class can dramatically improve inferences relative to EPBR-based matching methods.

KEY WORDS: Causal inference; EPBR; Matching.

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 pretreatment 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 pretreatment variables.

Once a matched dataset 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 pretreatment differences. The advantage of matching is that inferences from better balanced datasets 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 $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ as a k-dimensional data set, where each X_j is a column vector of the observed values of pretreatment variable j for the n observations. That is, $\mathbf{X} = [X_{ij}, i = 1, \ldots, n, j = 1, \ldots, k]$. 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, that is, $d(\mathbf{X}_i, \mathbf{X}_l) \simeq 0$. 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 (1976b) 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 article, 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.

Stefano M. Iacus is Associate Professor, Department of Economics, Business and Statistics, University of Milan, Via Conservatorio 7, I-20124 Milan, Italy (E-mail: *stefano.iacus@unimi.it*). Gary King is the Albert J. Weatherhead University Professor, Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge MA 02138 (E-mail: *king@harvard.edu*). Giuseppe Porro is Associate Professor, Department of Economics and Statistics, University of Trieste, P.le Europa 1, I-34127 Trieste, Italy (E-mail: *giuseppe.porro@econ.units.it*). Open source software for R and Stata to implement the methods described herein is available at *http:// gking.harvard.edu/cem*; the cem algorithm is also available via the R package MatchIt. 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.

© 2011 American Statistical Association
Journal of the American Statistical Association
March 2011, Vol. 106, No. 493, Theory and Methods
DOI: 10.1198/jasa.2011.tm09599

Classes of methods define properties which all matching methods within the class must posses. Some methods may also belong to more than one class. For a review of many existing methods and their advantages and disadvantages, see Ho et al. (2007).

2. CLASSES OF MATCHING METHODS

In this section, we summarize the existing EPBR class of matching methods, introduce our new MIB class, discuss example methods within each class along with various comparisons, and show how MIB is able to explicitly bound model dependence, a longstanding goal of matching methods.

2.1 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 $M_\mathcal{T} \subseteq \mathcal{T}$ and $M_\mathcal{C} \subseteq \mathcal{C}$ 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 $\bar{\mathbf{X}}_{m_T} = \frac{1}{m_T}\sum_{i\in M_\mathcal{T}} \mathbf{X}_i$, and $\bar{\mathbf{X}}_{m_C} = \frac{1}{m_C}\sum_{i\in M_\mathcal{C}} \mathbf{X}_i$ be the vector of sample means for the matched data only.

EPBR requires all treated units to be matched, that is, $m_T = n_T$ (thus $M_\mathcal{T} = \mathcal{T}$), 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 1976a]. An EPBR matching solution satisfies

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

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

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

The right-hand side of (2) is the average mean-imbalance in the population that gives rise to the original data, and the left-hand 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, $X_j \times X_k$, and so forth). 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).

2.2 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). In addition, achieving in-sample balance is an excellent way to achieve expected balance as well. Let $\bar{X}_{n_T,j}$, $\bar{X}_{n_C,j}$ and $\bar{X}_{m_T,j}$, $\bar{X}_{m_C,j}$ denote the prematch and postmatch sample means, for variable X_j, $j=1,\ldots,k$, 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

$$\big|\bar{X}_{m_T,j} - \bar{X}_{m_C,j}\big| \leq \delta_j, \qquad j=1,\ldots,k, \tag{3}$$

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.

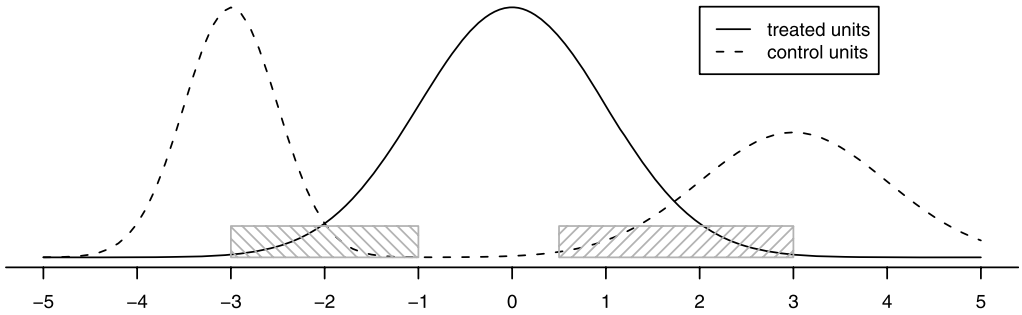

Figure 1. An example of a covariate for which minimizing mean-imbalance may be harmful. The example also shows that increasing mean-imbalance for this variable under MIB can be used to match more relevant features of the distributions (such as the shaded areas), without hurting mean-imbalance on other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

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 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 are most similar (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 $\mathcal{X}_{n_T} = [(X_{i1}, \ldots, X_{ik}), i \in \mathcal{T}]$ 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

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

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, that is, $f(\mathcal{X}_{m_T}) = \bar{\mathbf{X}}_{m_T}$ and $f(\mathcal{X}_{m_C}) = \bar{\mathbf{X}}_{m_C}$, $\delta = \gamma D(f(\mathcal{X}_{n_T}), f(\mathcal{X}_{n_C}))$, 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), that is, $|\bar{X}_{m_T,j} - \bar{X}_{m_C,j}| \le \delta_j$. Similarly, if $f_j(\cdot)$ is the sample variance, the kth centered moment, the qth quantile, and so forth.
- 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 $|\mathrm{Cov}(X_j, X_k)| \le \delta_{jk}$.
- In Section 3 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, and so forth). Let $\pi, \pi' \in \mathbb{R}_+^k$ be two nonnegative 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 $J \subseteq \{1, \ldots, k\}$, for which $\pi_j < \pi'_j$, $j \in J$. For a given function $f(\cdot)$ and a distance $D(\cdot,\cdot)$ we denote by $\gamma_{f,D}(\cdot) : \mathbb{R}_+^k \to \mathbb{R}_+$ a monotonically increasing function of its argument, that is, if $\pi \preceq \pi'$ then $\gamma_{f,D}(\pi) \le \gamma_{f,D}(\pi')$. Then our last step gives the definition of the new class:

Definition 3 [Monotonic Imbalance Bounding (MIB)]. A matching method is *Monotonic Imbalance Bounding* on the function of the data $f(\cdot)$ with respect to a distance $D(\cdot,\cdot)$, or simply MIB(f, D), if for some monotonically increasing function $\gamma_{f,D}(\cdot)$ and any $\pi \in \mathbb{R}_+^k$ we have that

$$D\big(f\big(\mathcal{X}_{m_T(\pi)}\big), f\big(\mathcal{X}_{m_C(\pi)}\big)\big) \le \gamma_{f,D}(\pi). \tag{5}$$

MIB is then a class of matching methods which produces subsets $\mathcal{X}_{m_T}$ and $\mathcal{X}_{m_C}$, where $m_T = m_T(\pi)$ and $m_C = m_C(\pi)$ on the basis of a given vector $\pi = (\pi_1, \pi_2, \ldots, \pi_k)$ of tuning parameters (such as a caliper), one for each covariate. As a result, the number of matched units is a function of the tuning parameter and is not fixed ex ante. In contrast, the function $\gamma_{f,D}$, once f and D are specified, depends only on the tuning parameter π, but not on the sample size m_T or m_C; indeed, it represents a bound, or the worst situation for a given value of the tuning parameter.

A crucial implication of the MIB property for practical data analysis is the following. Suppose that for a matching method

in the MIB class (such as the one we introduce in the Section 3), such that for each variable $j = 1, \ldots, k$, we have $f(x_1, \ldots, x_j) = f_j(x_j)$ (e.g., the empirical mean of X_j) and a function $\gamma_{f_j,D}(\pi_1, \ldots, \pi_k) = \gamma_j(\pi_j), j = 1, \ldots, k$. Then, we can write the system of inequalities

$$\begin{cases} D\big(f_1\big(\mathcal{X}_{m_T(\pi)}\big), f_1\big(\mathcal{X}_{m_C(\pi)}\big)\big) \le \gamma_1(\pi_1) \\ \qquad\qquad\qquad\qquad\vdots \\ D\big(f_k\big(\mathcal{X}_{m_T(\pi)}\big), f_k\big(\mathcal{X}_{m_C(\pi)}\big)\big) \le \gamma_k(\pi_k). \end{cases}$$

Now suppose a researcher changes only a single tuning parameter, for example for the first variable: that is, we take a new vector $\pi' = (\pi_1 - \epsilon, \pi_2, \ldots, \pi_k)$, with $\epsilon > 0$. The above system of inequalities still holds, that is, all inequalities from 2 to k remain unchanged and only the first one changes to $D(f_1(\mathcal{X}_{m_T(\pi)}), f_1(\mathcal{X}_{m_C(\pi)})) \le \gamma_1(\pi_1 - \epsilon) \le \gamma_1(\pi_1)$.

This means that relaxation of one tuning parameter for one variable controls monotonically the imbalance measures by (D, f_j), without altering the maximal imbalance on the 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), five 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.

2.3 Examples and Comparisons

Well known matching methods within the (potentially) EPBR class include nearest neighbor matching based on a propensity score or Mahalanobis distance. These methods are not MIB, because they fix the number of matched observations (m_T, m_C) ex ante (usually to twice the number of treated units) rather than, as in MIB methods, letting the number of matched units be the result of the user setting tuning parameters. 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 when the distance between the treated and control groups includes a tuning parameter for each variable. (Cochran and Rubin 1973, p. 420, also recognized that tight calipers control all linear and nonlinear imbalance under certain circumstances.)

More generally, let $X_i = (X_{i1}, X_{i2}, \ldots, X_{ik})$ and $X_h = (X_{h1}, X_{h2}, \ldots, X_{hk})$ be any two vectors of covariates for two sample units i and h. Let $d_j(X_i, X_h) = d_j(X_{ij}, X_{hj})$ define the distance for covariate j $(j = 1, \ldots, k)$. Then, the caliper distance between X_i and X_h is $d(X_i, X_h) = \max_{j=1,\ldots,k} \mathbf{1}_{\{d_j(X_i,X_h) \ge \epsilon_j\}}$, where $\mathbf{1}_A$ is an indicator function for set A and $\epsilon_1, \ldots, \epsilon_k$ are tuning parameters. So when $d(X_i, X_h) = 0$, X_i and X_h are close, and if $d(X_i, X_h) = 1$ units are far apart (i.e., unmatchable, which could also be expressed by redefining the latter as ∞). The simplest choice to complete the definition is the caliper distance, $d_j(X_i, X_h) = d_j(X_{ij}, X_{hj}) = |X_{ij} - X_{hj}|$ (see Rosenbaum 2002, chap. X) but any other one-dimensional distance will also be MIB provided the tuning parameter ϵ_j is on the scale of covariate X_j and is defined for all j. In this case, nearest neighbor matching with caliper or full optimal matching is MIB.

Another member of the MIB class is coarsened exact matching, which applies exact matching after each variable is separately coarsened (CEM is detailed in Section 3). Numerous other diverse MIB methods can be constructed by applying non-MIB methods within CEM's coarsened strata or within variable-by-variable calipers. For one example, we can coarsen with very wide bins, apply CEM, and then use the propensity score, or Mahalanobis distance within each bin, to further prune observations. The resulting methods are all MIB.

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 generation process, EPBR means that the degree of imbalance changes for all variables at the same time by proportionally the same amount; MIB, without extra assumptions on the data, means that changing one variable's tuning parameter 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 chooses the maximal imbalance ex ante and produces a matched sample size ex post. The latter would generally be preferred in observational analyses, where data is typically plentiful but is not under control of the investigator, and so reducing bias rather than inefficiency is the main focus.

In real datasets 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. Since the objective function used for estimating the propensity score differs from the analytical goal of finding balance, applied 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, by restricting only one tuning parameter per variable, imbalance in the means, other moments, comoments, interactions, nonlinearities, and the full multivariate distribution of the treated and control groups may be improved, without hurting maximum imbalance on other variables and regardless of the data type. The actual level of balance achieved by MIB methods can of course be better than the maximum level set ex ante. Moreover, as in CEM with the level of coarsening, setting this single tuning parameter per variable is typically a choice users are able to make based on their knowledge of the substantive problem and data.

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 can be thought of as approximating the *randomized block* experimental design, where values of the treatment variable are assigned within strata defined by the covariates. (Of course, methods from either class can be modified to create randomized block designs.) Although both designs are unbiased, randomized block designs have exact multivariate balance in each dataset on all observed covariates, whereas complete randomization designs are balanced only on average across random treatment assignments in different experiments, with no guarantees for the one experiment being run. Randomized block designs, as a result, are considerably more efficient, powerful, and robust, regardless whether one is estimating in-sample or population quantities (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 too far 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 proceeded by a separate method for eliminating these risky observations. One way to eliminate extreme counterfactuals is to discard control units that 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 a distant risky extrapolations, often even without a separate step.

2.4 MIB Properties: Bounding Model Dependence and Estimation Error

A key motivating factor in matching is to reduce model dependence (Ho et al. 2007). However, the relationship has never been proven directly for EPBR methods, or any other aside from exact matching. Thus, we contribute here a proof that the maximum degree of model dependence can be controlled by the tuning parameters for MIB methods. We also go a step farther and show how the same parameters also bound causal effect estimation error.

Model Dependence. At the unit level, exact matching estimates the counterfactual $Y_i(0) \equiv g_0(\mathbf{X}_i)$ for treated unit i with the value of Y of the control unit j such that $\mathbf{X}_j = \mathbf{X}_i$. If exact matching is not feasible, then we use a model m_ℓ to span the remaining imbalance to $Y_i(0)$ with control units close to the treated units, that is, using matched data such as $\hat{Y}(0) \equiv m_\ell(\tilde{\mathbf{X}}_j)$, where $\tilde{\mathbf{X}}_j$ is the vector of covariates for the control units close to treated i. Model dependence is how much $m_\ell(\tilde{\mathbf{X}}_j)$ varies as a function of the model m_ℓ for a given vector of covariates $\tilde{\mathbf{X}}_j$. We restrict the attention to the set of competing Lipschitz models.

Definition 4 (Competing models). Let m_ℓ $(\ell = 1, 2, \ldots)$ be statistical models for Y. For example, $m_\ell(x)$ may be a model for $E(Y|X = x)$. Then we consider the following class:

$$\mathcal{M}_h = \big\{ m_\ell : |m_\ell(x) - m_\ell(y)| \le K_\ell d(x, y) \text{ and } |m_i(x) - m_k(x)| \le h, i \ne k, x \in \Xi \big\}$$

with exogenous choices of a small prescribed nonnegative value for h and $0 < K_\ell < \infty$ and $\Xi = \Xi_1 \times \cdots \times \Xi_k$, where Ξ_j is the empirical support of variable X_j. Here $d(x, y)$ is some distance on the space of covariates Ξ.

In $\mathcal{M}_h$, the Lipschitz constants K_ℓ are proper constants of the models m_ℓ and, given the specification of m_ℓ, need not be estimated. The class $\mathcal{M}_h$ represents competing models which fit the observed data about as well, or in other words do not yield very different predictions for the same observed values $\tilde{\mathbf{X}}$; if this were not the case, we could rule out a model based on the data alone.

In this framework, for any two models $m_1, m_2 \in \mathcal{M}_h$, we define *model dependence* as $|m_1(\tilde{\mathbf{X}}_i) - m_2(\tilde{\mathbf{X}}_i)|$ (King and Zeng 2007). This leads to the key result for MIB methods with respect to $D(x, y) = d(x, y)$ and $f(x) = x$:

$$\begin{aligned}
&|m_1(\tilde{\mathbf{X}}_j) - m_2(\tilde{\mathbf{X}}_j)| \\
&\quad = |m_1(\tilde{\mathbf{X}}_j) \pm m_1(\mathbf{X}_i) \pm m_2(\mathbf{X}_i) - m_2(\tilde{\mathbf{X}}_j)| \\
&\quad \le |m_1(\mathbf{X}_i) - m_1(\tilde{\mathbf{X}}_j)| + |m_2(\mathbf{X}_i) - m_2(\tilde{\mathbf{X}}_j)| \\
&\qquad + |m_1(\mathbf{X}_i) - m_2(\mathbf{X}_i)| \\
&\quad \le (K_1 + K_2) d(\mathbf{X}_i, \tilde{\mathbf{X}}_j) + h \le (K_1 + K_2)\gamma(\pi) + h.
\end{aligned}$$

Thus, the degree of model dependence is directly bounded by the choice of π [via $\gamma(\pi)$] of the MIB method.

Estimation Error. To show how MIB bounds estimation error, we need to recognize that under certain specific conditions matching changes the estimand (Section 6.1 discusses when this may be desirable and how to avoid it when not). For example, consider an MIB method which produces one-to-one matching of treatment to control. We denote by $\mathbf{X}_i$ the values of the covariates for treated unit i and by $\tilde{\mathbf{X}}_j$ the values for control unit j matched with this treated unit. If m_T is the total number of treated units matched, then estimand can be defined as

$$\tau_{m_T} = \frac{1}{m_T} \sum_{i \in M_T} (Y_i(1) - Y_i(0)).$$

Then, an estimator of τ_{m_T} is given by

$$\hat{\tau}_{m_T} = \frac{1}{m_T} \sum_{i \in M_T} (Y_i(1) - \hat{Y}_i(0)) = \frac{1}{m_T} \sum_{i \in M_T} (Y_i(1) - g(\tilde{\mathbf{X}}_j, 0)),$$

where $\hat{Y}_i(0) = g(\tilde{\mathbf{X}}_j, 0)$ and $\tilde{\mathbf{X}}_j$ is the vector of covariate values for control unit j matched with treated unit i.

Proposition 1. Let $d(x, y)$ be a distance from $\Xi \times \Xi$ to R_+ and let $g(x, 0)$ be differentiable with bounded partial derivates, that is, $|\frac{\partial}{\partial_{x_i}} g(x, 0)| \le K_i$, for some $0 < K_i < \infty$, $i = 1, \ldots, k$. Then, for an MIB method with respect to $D(x, y) = d(x, y)$ and $f(x) = x$ we have that $|\tau_{m_T} - \hat{\tau}_{m_T}| \le \gamma(\pi)\mathcal{K} + o(\gamma(\pi))$ with $\mathcal{K} = \sum_{h=1}^{k} K_h$.

Proof. Taylor expansion of $g(\tilde{\mathbf{X}}_j, 0)$ around $\mathbf{X}_i$, gives

$$g(\tilde{\mathbf{X}}_j, 0) - g(\mathbf{X}_i, 0) = d(\tilde{\mathbf{X}}_j, \mathbf{X}_i) \cdot \sum_{h=1}^{k} \frac{\partial}{\partial_{x_h}} g(x, 0)\Big|_{x=\mathbf{X}_i} + o(d(\tilde{\mathbf{X}}_j, \mathbf{X}_i)).$$

We can decompose $\hat{\tau}_{m_T}$ as follows:

$$\begin{aligned}\hat{\tau}_{m_T} &= \frac{1}{m_T} \sum_{i \in M_T} (Y_i(1) - \hat{Y}_i(0) \pm Y_i(0)) \\ &= \frac{1}{m_T} \sum_{i \in M_T} (Y_i(1) - g(\tilde{\mathbf{X}}_j, 0) \pm g(\mathbf{X}_i, 0)) \\ &= \tau_{m_T} - \frac{1}{m_T} \sum_{i \in M_T} \left(d(\tilde{\mathbf{X}}_j, \mathbf{X}_i) \cdot \sum_{h=1}^{k} \frac{\partial}{\partial_{x_h}} g(x, 0)\Big|_{x=\mathbf{X}_i} \right. \\ &\quad \left. + o(d(\tilde{\mathbf{X}}_j, \mathbf{X}_i)) \right).\end{aligned}$$

Therefore, we have the statement of the proposition.

In some cases, the requirement of partial boundedness of derivates may be too stringent or the constants K_i can be too large. In this situation, a direct proof can still be obtained by imposing a weaker local Lipschitz condition directly on $g(x, 0)$, that is, $|g(x, 0) - g(y, 0)| \leq K_x d(x, y)$ for any y in a neighborhood of x. As a result, differentiability of $g(\cdot, 0)$ is not required. In all cases, the constants K_i [or K_x for locally Lipschitz $g(\cdot, 0)$] interact with $\gamma(\pi)$ and thus the bound may be sufficiently tight because, for a MIB method, $\gamma(\pi)$ can be controlled monotonically, while for a given $g(\cdot, 0)$ all the constants are given. The linear $g(\cdot, 0)$ usually adopted in the literature, or the example functions in Figure 5 below, are all locally Lipschitz functions.

The above results shows that, an MIB method bounds the error $|\tau_{m_T} - \hat{\tau}_{m_T}|$ as an explicit function of the vector of tuning parameters π.

3. COARSENED EXACT MATCHING AS AN MIB METHOD

In order to clarify how the MIB class of matching methods works in practical applications, we now introduce one member of the MIB class of matching methods that comes from the diverse set of approaches based on subclassification (also known as "stratification" or "intersection" methods). We call this particular method CEM for "Coarsened Exact Matching" (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, and so forth). (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, that is, $\mathcal{C}^s$ and $m_C^s = \#\mathcal{C}^s$. The number of matched units are, respectively for treated and controls, $m_T = \sum_{s \in S} m_T^s$ and $m_C = \sum_{s \in S} m_C^s$. 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, seven-point partisan identification scales are recoded as Democrat, Independent, and Republican; Likert issue questions as agree, neutral, and disagree; and multiparty vote returns as winners and losers. Many use a small number of categories to represent religion, occupation, U.S. Security and Exchange Commission industry codes, 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.

Although we find that data analysts have little trouble making coarsening choices on the basis of their substantive information, we have also developed a series of automated coarsening methods, such as those which automatically choose bin widths for histograms and other more sophisticated approaches. These are available in easy-to-use software that accompanies this article for R and Stata at *http://gking.harvard.edu/cem*.

CEM as an MIB Method

We show here that CEM is a member of the MIB class with respect to the mean, the centered absolute kth moment, and the empirical and weighted quantiles (all proofs are in the Appendix). 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 $\bar{X}^w_{m_T,j} = \frac{1}{m_T}\sum_{i\in\mathcal{T}} X_{ij}w_i$ and $\bar{X}^w_{m_C,j} = \frac{1}{m_C}\sum_{i\in C} X_{ij}w_i$.

Proposition 2. For $j = 1,\ldots,k$, $|\bar{X}^w_{m_T,j} - \bar{X}^w_{m_C,j}| \le \epsilon_j$.

Let R_j be the range of variable X_j and let $\theta_j = \max_{\epsilon_j \ge \epsilon_j^*}(\lceil R_j/\epsilon_j\rceil)$, 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, for example, the lowest value of ϵ_j which generates at most n_T nonempty intervals in CEM.

Proposition 3. Let $k \ge 1$ and consider the centered absolute kth moment for variable X_j for the treated and control units as $\bar{\mu}^k_{T,j} = \frac{1}{m_T}\sum_{s\in\mathcal{S}}\sum_{i\in\mathcal{T}^s}|X_{ij} - \bar{X}^w_{m_T,j}|^k w_i$ and $\bar{\mu}^k_{C,j} = \frac{1}{m_C}\sum_{s\in\mathcal{S}}\sum_{i\in\mathcal{C}^s}|X_{ij} - \bar{X}^w_{m_C,j}|^k w_i$. Then, $|\bar{\mu}^k_{T,j} - \bar{\mu}^k_{C,j}| \le \epsilon_j^k(\theta_j + 1)^k$, $j = 1,\ldots,k$, and $\epsilon_j \ge \epsilon_j^*$.

Proposition 4. Assume one-to-one matching. Denote by $X^q_{m_T,j}$ the qth empirical quantile of the distribution of the treated units for covariate X_j, and similarly $X^q_{m_C,j}$. Then, $|X^q_{m_T,j} - X^q_{m_C,j}| \le \epsilon_j$ for $j = 1,\ldots,k$.

Define the weighted empirical distribution functions for treated group as $F^w_{m_T,j}(x) = \sum_{X_{ij}\le x, i\in\mathcal{T}}\frac{w_i}{m_T}$ and for the control group as $F^w_{m_C,j}(x) = \sum_{X_{ij}\le x, i\in\mathcal{C}}\frac{w_i}{m_C}$. Define the qth quantile of the weighted distribution $X^{q,w}_{m_T,j}$ as the first observation in the sample such that $F^w_{m_T,j}(x) \ge q$ and similarly for $X^{q,w}_{m_C,j}$.

Proposition 5. Assume that the support of variable X_j is cut on subintervals of exact length ϵ_j. Then $|X^{q,w}_{m_T,j} - X^{q,w}_{m_C,j}| \le \epsilon_j$ for $j = 1,\ldots,k$.

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. One worry is the curse of dimensionality, which in this context means that the number of hyper-rectangles, and thus the number of possible strata $\#C(X_1)\times\cdots\times\#C(X_k)$, 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 either. 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, the sparseness that occurs in multidimensional space turns out not to be much of an issue in practice. First, real datasets have far more highly correlated data structures than the independent draws in the example above, and so CEM in practice tends to produce a reasonable number of matches. This has been our overwhelming experience in the numerous data sets we have analyzed. We studied this further by attempting to sample from the set of social science causal analyses in progress in many fields by broadcasting an offer (through blogs and listservs) help in making causal analyses in return for a look at their data (and a promise not to scoop anyone). We will report on this analysis in more detail in a later article but note here that, for almost every dataset we studied, CEM produced an inferentially useful number of matched observations and also generated substantially better balance for a given number of matches than methods in the EPBR class.

Finally, 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, the number of cells that contain only (unmatched) treated units goes to zero exponentially fast with the number of treated units n_T in the sample, if the number of control units n_C grows at rate $n_C = O(n_T^{1/r})$, with $r \ge k$ and k the number of continuous pretreatment covariates.

On CEM's Computational Efficiency

Although the number of empty CEM strata in real data tends to be small, the total number of cells to be explored in order to determine in which strata we have a match is exponentially large, for example, m^k where k is the number of covariates and m is the average number of intervals on which the support of each covariate has been cut. So, for example, with $k = 20$ and $m = 6$, the number of strata is huge: $6^{20} = 3.656158\cdot 10^{15}$. Fortunately, CEM produces at most n strata if all observations fall into different CEM strata. The implementation can then ignore this space and focus on at most n strata. Thus, the coarsening represents the original covariate values X_j with integers 1 to θ_j, leading an observation to be a feature vector of integers such as $(X_{i1}, X_{i2}, X_{i3}, \ldots, X_{i98}, X_{i99}, X_{i100}) = (1, 3, 5, \ldots, 3, 2, 7)$. The algorithm we implement then represents all the information in the feature vector for each observation with a string like "$1*3*5*\cdots*3*2*7$." The result is that CEM matching has the same complexity of a simple frequency tabulation, which is of order n. This algorithm can even be easily implemented on very large databases using SQL-like queries.

CEM and Propensity Scores

One way to convey how MIB generalizes the EPBR class of methods, and thus includes its benefits, is to note that a high quality nonparametric estimate of the propensity score is available from CEM results by merely calculating the proportion of units within each stratum that are treated. Indeed, this estimator would typically balance better than that from the usual logit model (which optimizes an objective function based on fit rather than balance) or various unwieldly ad hoc alternatives that iterate between balance checking and tweaking and rerunning the logit model. Of course, with CEM, estimating the propensity score is not needed, since CEM determines the matches first, but this point helps convey how using CEM can include the essential advantages of the propensity score.

4. REDUCING IMBALANCE

We now define and introduce a measure of imbalance between the treated and control groups. We then demonstrate how CEM outperforms (potentially) EPBR methods even in data generated to meet EPBR assumptions.

4.1 Definition and Measurement

Most matching methods were designed to reduce imbalance in the *mean* of each pretreatment variable between the treated and control groups, for which the balance metric is simply the difference in means for each covariate. (One 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, mean imbalance is only part of the goal as it does not necessarily represent the desired multidimensional imbalance between the treated and control groups.

We thus now introduce a new and more encompassing imbalance measure aimed at representing the distance between the *multivariate* empirical distributions of the treated and control units of the pretreatment covariates. Our measure is intended to be widely applicable across datasets and (postmatching) analysis methods, even though a better measure may well be available in situations where a researcher is restricted to a specific analysis method.

The proposed measure uses the L^1 norm to measure the distance between the multivariate histograms. To build this measure, we obtain the two multidimensional histograms by direct cross-tabulation of the covariates in the treated and control groups, given a choice of bins for each variable. Let $H(X_1)$ denote the set of distinct values generated by the bins chosen for variable X_1, that is, 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 $H(X_1) \times \cdots \times H(X_k) = H(\mathbf{X}) = H$.

Let f and g be the relative empirical frequency distributions for the treated and control units, respectively. Observation weights may exists as output of a matching method. For the raw data, the weights are equal to one for all observations in the sample. Let $f_{\ell_1 \cdots \ell_k}$ 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 $g_{\ell_1 \cdots \ell_k}$.

Definition 5. The multivariate imbalance measure is

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

We denote the measure by $\mathcal{L}_1(f, g; H) = \mathcal{L}_1(H)$ to stress its dependence on the choice of multidimensional bins H, the definition of which we take up below. [Sometimes, when we compare two methods MET1 and MET2 we will write $\mathcal{L}_1(MET1; H)$ and $\mathcal{L}_1(MET2; H)$.] An important property of this measure is that the typically numerous empty cells do not affect $\mathcal{L}_1(H)$, and so the summation in (6) has at most only n nonzero terms. The relative frequencies also control for what may be different sample sizes for the treated and control groups.

The L_1 measure offers an intuitive interpretation, for any given set of bins: If the two empirical distributions are completely separated (up to H), then $\mathcal{L}_1 = 1$; if the distributions exactly coincide, then $\mathcal{L}_1 = 0$. In all other cases, $\mathcal{L}_1 \in (0, 1)$. If say $\mathcal{L}_1 = 0.6$, then 40% of the area under 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 bins H must remain fixed. See also Racine and Li (2009).

Choosing a Bin Definition H for $\mathcal{L}_1(H)$. Although the definition of $\mathcal{L}_1(H)$ is intuitive, it depends on the apparently arbitrary choice of the bins H: Like the bandwidth in nonparametric density estimation, bins too small provide exact separation in multidimensional space [$\mathcal{L}_1(H) = 1$] and bins too large cannot discriminate [$\mathcal{L}_1(H) = 0$]. Thus, analogous to the purpose of ROC curves in avoiding the choice of the differential costs of misclassification, we now develop a single definition for H to represent all possible bin choices.

To begin, we study the $\mathcal{L}_1$-*profile*, which is our name for the distribution of $\mathcal{L}_1(H)$ in the set of all possible bin definitions $H \in \mathcal{H}$. We study this distribution by drawing 250 random samples from the $\mathcal{L}_1$-*profile* for data from Lalonde (1986), a commonly used benchmark dataset in the matching literature (the 10 variables included are not central to our illustration, and so we refer the interested reader to the original article or the replication file that accompanies our article). For each randomly drawn value of H, we calculate $\mathcal{L}_1(H)$ based on the raw data and on each of three matching methods: (1) nearest neighbors based on the propensity score calculated via logit model (PSC); (2) a optimal matching (MAHO) solution based on the Mahalanobis distance (see Rosenbaum 2002); and (3) a CEM solution with fixed coarsening (10 intervals for each continuous variable and no coarsening for the categorical variables).

The left panel of Figure 2 plots $\mathcal{L}_1(H)$ vertically by the randomly chosen bin H horizontally, with bins sorted by the $\mathcal{L}_1(H)$ value based on the raw data. For this reason the raw data (the red line) is monotonically increasing. For any given H (indicated by a point on the horizontal axis), the method with the lowest imbalance is preferred. The problem we address here is that different bin definitions can lead to different rankings among the methods. Fortunately, however, in these data the rank order imbalance reduction of the methods remains stable across almost the entire range of H values, aside from small random perturbations. For almost any value of H on the horizontal axis, the largest imbalance reduction is generated by CEM, then propensity score matching and Mahalanobis distance matching. CEM thus essentially dominates the other methods, in that regardless of the definition of the bins used in defining $\mathcal{L}_1(H)$ its matched data sets have the lowest imbalance. In these data, propensity score matching and Mahalanobis matching are slightly better than the raw data. The approximate invariance portrayed in this figure is worth checking for in specific applications, but we find it to be an extremely common empirical regularity across almost all the data sets we have analyzed. For another view of the same result, the right panel of Figure 2 represents the cumulative empirical distribution functions (ecdf) of the values

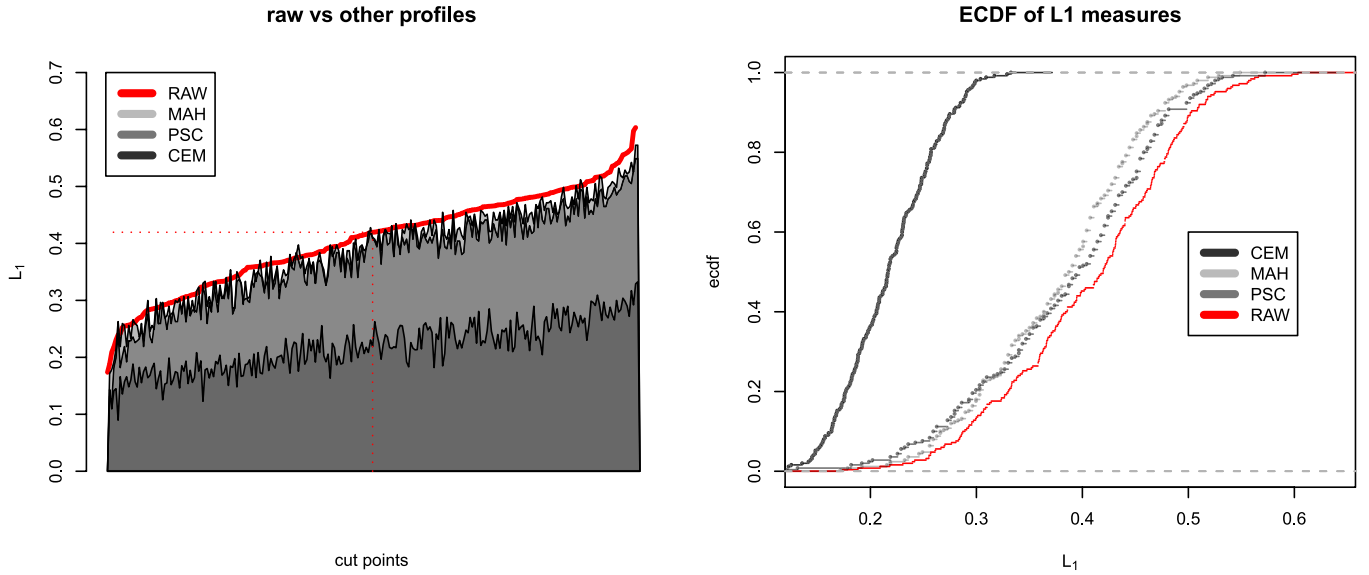

Figure 2. The $\mathcal{L}_1$-profile of the raw ("RAW") data (red line) compared to propensity score ("PSC"), Mahalanobis ("MAHO"), and Coarsened Exact Matching ("CEM") matched data. MAHO and PSC overlaps in the left panel. The left panel plots the L_1 profile by different bin choices sorted by L_1 for the raw data; the right panel plots the empirical cumulative distribution functions of the same set of $\mathcal{L}_1$ values.

of $\mathcal{L}_1(H)$ over the set $\mathcal{H}$: The method with the rightmost ecdf produces the highest levels of imbalance.

Thus, we have shown that the bin definition H in our imbalance measure (6) is often unimportant. We thus propose fixing it to a specific value to produce a final imbalance measure recommended for general use. The specific value we recommend is the set of bins $\bar{H}$ which corresponds to the median value of $\mathcal{L}_1$ on the profile of the raw data, $\bar{H} = \bar{H}^{\text{RAW}}$. We denote this value and new measure by $\bar{\mathcal{L}}_1 \equiv \mathcal{L}_1(\bar{H})$. In Figure 2, $\bar{\mathcal{L}}_1$ is 0.43 for Mahalanobis matching, 0.41 for propensity score matching, and 0.26 for CEM.

Is Balancing on the Means Enough?. 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 $p_i = P(T = 1|X_{i1}, X_{i2}, \ldots, X_{ik}) = 1/[1 + \exp\{-\beta_0 - \sum_{j=1}^{k} \beta_j X_{ij}\}]$ 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 $\bar{X}^w_{T,j} = \sum_{i\in\mathcal{T}} X_{ij} w_i / \sum_{i\in\mathcal{T}} w_i$ and $\bar{X}^w_{C,j} = \sum_{i\in\mathcal{C}} X_{ij} w_i / \sum_{i\in\mathcal{C}} w_i$. Then, it is well known that (without matching) $\bar{X}^w_{T,j} = \bar{X}^w_{C,j}$.

Although this weighting guarantees the elimination of all mean imbalance, the multidimensional distribution of the data may be still highly imbalanced. We consider again the same data as before. The value of the median on the $\mathcal{L}_1$-profile for the data is equal to $\bar{\mathcal{L}}_1 = 0.54$. The univariate ($I_1$) and global ($\bar{\mathcal{L}}_1$) 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, an almost perfect (weighted) match on the difference in means for all variables, but the overall global imbalance is equal to $\bar{\mathcal{L}}_1 = 0.53$, which is almost the same as the original data. However, after matching the raw data with CEM (which we do by coarsening the continuous variables into eight 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 $\bar{\mathcal{L}}_1 = 0.34$.

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 vari-

Table 1. Differences in means for each variable and global imbalance measure ($\bar{\mathcal{L}}_1$) on raw data from Lalonde (1986), after propensity score weighting, and following CEM matching. Variable names are as in Lalonde's original dataset. The propensity score is estimated by a logit model; CEM coarsens the continuous variables into eight categories

Variable	Raw data	Pscore weighting	CEM
Age	0.18	−0.00	0.11
Education	0.19	−0.00	−0.02
Black	0.00	0.00	0.00
Married	0.01	−0.00	0.00
Nodegree	−0.08	0.00	0.00
re74	−101.49	0.00	93.85
re75	39.42	0.00	36.12
Hispanic	−0.02	0.00	0.00
u74	−0.02	−0.00	0.00
u75	−0.05	−0.00	0.00
$\bar{\mathcal{L}}1$	0.54	0.53	0.34

able at a time. CEM, as an MIB method, and $\bar{\mathcal{L}}_1$ as a measure of imbalance, provide a simple way around these problems.

4.2 CEM versus 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 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.

The propensity score model is estimated as usual including all the main effects. For PSC we use a 1-nearest neighbor method without replacement while we denote by MAH 1-nearest neighbor matching, also without replacement, on the Mahalanobis distance and MAHO optimal matching (see Rosenbaum 2002) on the Mahalanobis distance. Due to the fact that CEM drops treated units, we also compute a second version of PSC where it is forced to further match on the subsample of treated and control units selected by the CEM algorithm (PSC2).

We begin by replicating an experiment proposed Gu and Rosenbaum (1993). This involves drawing two independent multivariate normal datasets: $\mathbf{X}_T \sim \mathrm{N}_5(\mu_T, \Sigma)$ and $\mathbf{X}_C \sim \mathrm{N}_5(\mu_C, \Sigma)$, 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 $n_T = 1000$ treated units from $\mathbf{X}_T$ and $n_C = 3000$ control units from $\mathbf{X}_C$. For CEM, we coarsen each covariate into 8 intervals of equal length. MAH, MAHO, and PSC match $m_T = 1000$ treated units against $m_C = 1000$ control units, whereas CEM selects both treated and control units and in turn PSC2 select $m_T = m_C$ depending on m_T, which is output from CEM in each simulation.

The properties of EPBR imply that MAH and PSC matching will optimally minimize expected mean imbalance (Rosenbaum and Rubin 1985b) in these data when all treated units are matched. In contrast, CEM is designed to reduce local multivariate imbalance, that is, the maximum distance between each treated unit and the corresponding matched control units. In addition to the global imbalance, $\bar{\mathcal{L}}_1$, we compute for each variable the global difference in means between the treated and control groups (I_1) and the average absolute difference in units stratum by stratum for CEM, and unit by unit for the other methods (I_2).

See Table 2. Overall, we find that CEM is substantially better than the other methods in terms of the difference in means, as well as local and global imbalance. Since one may argue that this effect is due to the fact that only CEM drops treated units, we also consider the performance of PSC2, but the conclusion here remains unchanged. PSC2 seems to benefit more from being combined with CEM. Thus, 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.

Table 2. Imbalance in means I_1 (top panel) and local imbalance I_2 (bottom panel) remaining after matching, for each variable listed, $X_1, \ldots, X_5$. Also reported are the number of treated m_T and control m_C units remaining after the match (top) and the multivariate $\mathcal{L}_1$ measure of imbalance (bottom, rightmost column). Results are averaged over 1000 replications, with $n_T = 1000$, $n_C = 3000$

Difference in means I_1

	X_1	X_2	X_3	X_4	X_5	m_T	m_C
Initial imb.	1.00	1.00	1.00	1.00	1.00	1000	3000
CEM	0.04	0.04	0.08	0.06	0.07	772	1851
MAH	1.00	1.00	1.00	1.00	1.00	1000	1000
MAHO	0.45	0.45	0.45	0.45	0.45	1000	1000
PSC	0.32	0.32	0.32	0.32	0.32	1000	1000
PSC2	0.14	0.15	0.16	0.14	0.16	772	772

Local imbalance I_2

	X_1	X_2	X_3	X_4	X_5	$\bar{\mathcal{L}}_1$
Initial						0.50
CEM	0.44	0.28	0.18	0.22	0.21	0.21
MAH	1.05	1.00	1.00	1.00	1.00	0.52
MAHO	2.64	1.54	1.11	1.54	1.11	0.34
PSC	2.67	1.43	0.88	1.43	0.88	0.31
PSC2	2.41	1.30	0.80	1.31	0.81	0.26

5. AN EMPIRICAL ANALYSIS: THE EFFECT OF HAVING A DAUGHTER ON CONGRESSIONAL REPRESENTATION

To illustrate how CEM works in practice, we replicate a recent article published in the leading journal in economics that seeks to explain U.S. Congressional decision making, a central concern to political scientists, using the effect of children on their parents, a longstanding subject of study among sociologists (Washington 2008). In this article, Washington showed that members of the U.S. House of Representatives who have a daughter (rather than a son) vote more liberally, especially on reproductive rights issues. Via a linear regression, Washington estimates this effect while controling for several discrete variables (race, gender, the number of children, political party, and religion) and continuous variables (seniority and its square, age and its square, and Democratic vote share). The outcome variable is a score given to each of 430 members of the 105th Congress (1997–1998) measuring agreement with the National Organization for Women (NOW). These scores range from 0 (no support) to 100 (complete support). (Washington reports that the NOW data were not available for other years.)

Conditional on Washington's regression specification, she finds that each additional girl causes a legislator to vote 2.23 points more liberally (with a standard error of 1.02). Of course, this is only one of many plausible specifications and so we study the degree of model dependence in these results. We do this by running all 656 possible linear regressions from the set of all main effects from her original set of covariates, all interactions up to order three, and all squared terms for the continuous variables. For each linear regression specification, we estimate the treatment effect, and our summary of model dependence is the variation across these estimates. We portray this variation by

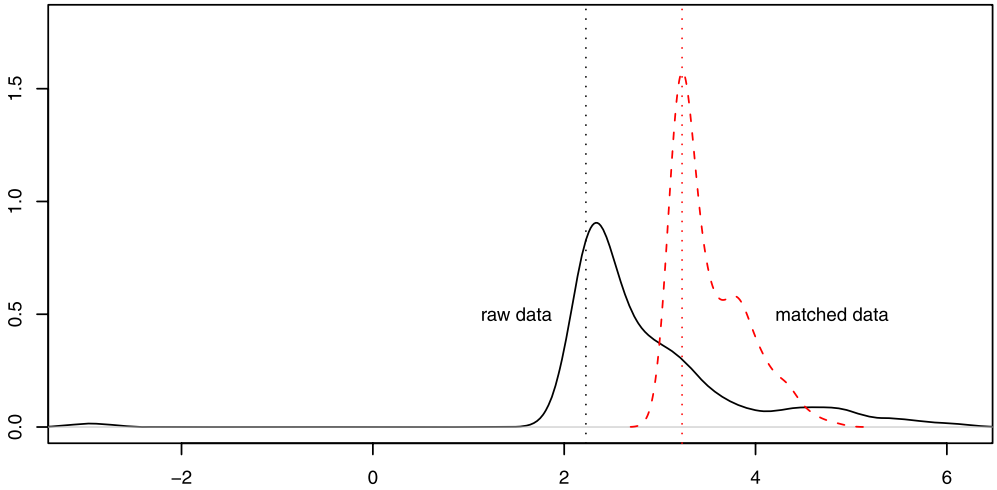

Figure 3. Quantifying Model Dependence: Density plots illustrating the variation in treatment effect estimates across different specifications of analysis models with (red, dashed) and without (black, solid) CEM matching. Dotted vertical lines give the point estimates using the specification chosen by Washington (2008). The online version of this figure is in color.

plotting a histogram (density estimate) of these results. See the solid line in Figure 3, where the estimated causal effects have a huge and substantively unacceptable range of $[-3.03, 6.11]$. By choosing only one specification (as most articles do), the regression approach can produce what seems like "evidence" for anything from a large negative result that rejects the hypothesis to a large positive result that supports it.

We then preprocess the data with CEM (coarsening each of the three continuous variables in 10 equal sized intervals). In this dataset, imbalance is modestly reduced from $\mathcal{L}_1 = 0.818$ with $n = 430$ in the raw data to $\mathcal{L}_1 = 0.762$ with $n = 175$. Yet, when we rerun all the same regression specifications on the CEM matched data, the degree of model dependence is substantially reduced, with a range of treatment effects now of only $[3.04, 4.80]$. Unlike the model dependence from Washington's analyses, every point within this range supports her scientific hypothesis. We can see this more clearly in Figure 3 by comparing the distribution of the treatment effects across different regression specifications run on the raw data (in black) to the distribution based on the CEM matched data (in dashed red). Even ignoring the outliers for the raw data, the variation in estimated effects after matching is much smaller.

As a result of matching, the point estimate of the causal effect has increased by a substantial 45%, from 2.2 to 3.2, and the new, larger estimate has much less model dependence and so should be regarded as considerably more reliable. We conclude that the effect Washington hypothesized, even if not demonstrated in her article, is even larger than her regression model indicated.

However, this more reliable causal effect is an average over the causal effects for only a subset of observations and so is making an inference about a different estimand. Most social science analyses with either observational or experimental data are based on convenience samples, with statistical inference conditional on the data analyzed, but we should nevertheless clarify precisely what quantity is being estimated, especially with matching analyses that preprocess the data and select a quantity of interest in the process. This includes CEM, as well as propensity score or Mahalanobis distance matching with calipers.

In other words, all matching analyses should routinely identify for readers the subset of observations in the matched set and summarize their values on the covariates. This step is rarely done in the literature, but as we show here it is easy to do and advisable. We do this in Figure 4 with a version of a "parallel plot." Each line a parallel plot corresponds to an observation in the dataset, where the vertical axis indicates the range of the values of each of the variables, and the horizontal axis is a variable list.

In the present example of a parallel plot in Figure 4, we have colored the lines representing legislators (in the treated group, i.e., with daughters) for whom matches were found in blue and those who were not unmatched in red. The results indicate that the matched treated units—which define the estimand—are a sample of legislators with a higher concentration of young, white, male Republicans than the full sample of representatives with daughters in the 105th Congress. In particular, the representatives with daughters in the matched set are 10 percentage points more male, an average of 4 years younger in age and 2.5 in seniority, 12 percentage points more white, and 24 percentage points more Republican. (Other differences are relatively minor.) Thus, because of who happened to be elected to congress in the 1996 election, we are able to make a relatively solid inference about this particular subset of legislators. For this group, it happens that the causal effect turns out to be reasonably sized and consistent with the hypothesis. However, due to the lack of an appropriate control groups, reliable inferences without model dependence happens not to be available for other groups of legislators with daughters.

6. REDUCING CAUSAL EFFECT ESTIMATION ERROR

In order to avoid inducing selection bias, statisticians suggest ignoring the outcome variable while choosing a matching procedure and focusing primarily on reducing imbalance in the covariates (as we did in Section 4). In this section, we go a step further and switch focus to reducing estimation error in the causal quantity of interest.

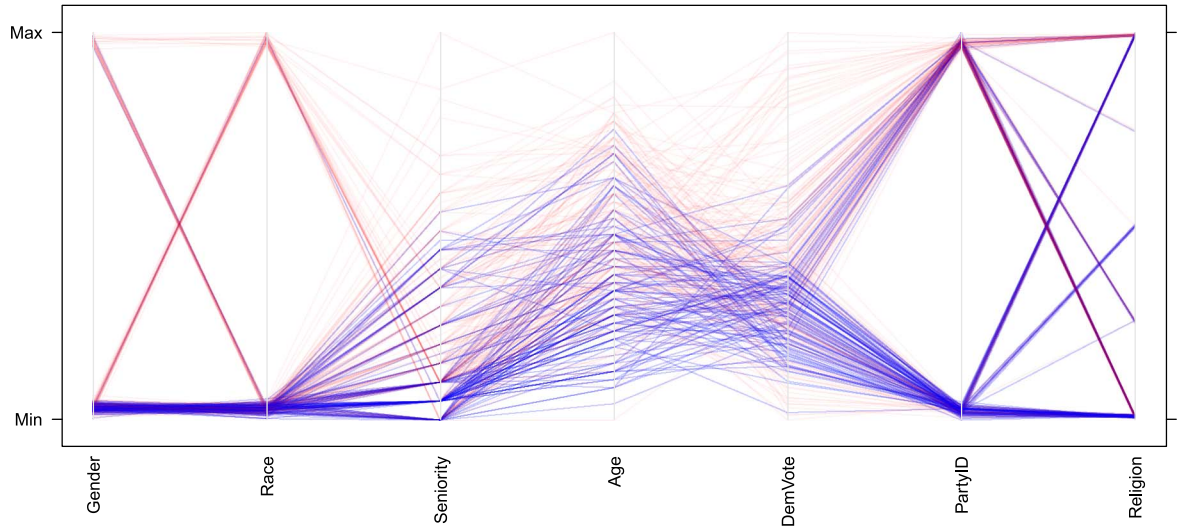

Figure 4. Parallel plot for the matched (blue) and unmatched (red) legislators. Each line describes a member of congress, with ranges for each variable: Gender (male, female), Race (white, other), Seniority (1–47 years), Age (26–87 years), DemVote (0.26–0.94), PartyID (Republican, Democrat), and Religion (protestant, none, catholic, other christian, other religion).

6.1 Definitions and Estimation

A crucial issue in causal inference is identifying the precise quantity to be estimated. This is an issue in observational data, which is typically based on convenience samples and may include whatever relevant data happen to be available. 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 units only, and no attempt is made to infer formally to a broader population.

We therefore define two quantities of interest: The sample average treatment effect on the treated SATT $= \frac{1}{n_T}\sum_{i\in\mathcal{T}}[Y_i(1) - Y_i(0)]$ and the population average treatment effect on the treated PATT $= E_{i\in\mathcal{T}^*}[Y_i(1) - Y_i(0)]$, where $\mathcal{T}$ and $\mathcal{T}^*$ are the sets of treated units in the sample and population, respectively.

SATT and PATT are especially convenient definitions for matching methods which prune (only) control units from a dataset and so do not change the estimand. In especially difficult datasets, 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, high levels of model dependence can result. 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 or PATT defined for the subset of treated units that have good matches among the pool of controls (for later use we denote as the "local" SATT or PATT). 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.

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 2.3). Second, run CEM on these preprocessed 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 (or similarly for PATT) separately in the two groups as follows. For the m_T treated units, suppose there exist m_C acceptable counterfactuals (as defined by the coarsening in CEM say), and so we can reliably estimate this "local SATT," $\hat{\tau}_{m_T}$, 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 SATT 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:

$$\hat{\tau}_{n_T} = \frac{\hat{\tau}_{m_T} \cdot m_T + \hat{\tau}_{n_T - m_T} \cdot (n_T - m_T)}{n_T}. \tag{7}$$

This procedure keeps the overall quantity of interest, SATT (or analogously PATT), fixed and isolates the model dependent piece of the estimator so it can be studied separately and its effects on the overall estimate 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.

6.2 CEM versus Propensity Score Matching

We now compare CEM with the standard use of propensity score matching. We focus on PATT rather than SATT to give the advantage to PSM as an EPBR method. (The results strongly favor CEM, but would even more in estimating SATT.) For simplicity, we use for our estimator the simple difference in means between matched treated and control groups, with weights for the matched units

$$\hat{\tau}_k = \sum_{i \in M_T} Y_i w_i - \sum_{j \in M_C} Y_j w_j,$$

where M_T and M_C are, respectively, the sets of treated and control units matched and Y_i is the observed outcome on the units and w_i are the weights of the different matching methods [for clarity of our simulation setup, we do not use the method in Equation (7)]. We run three separate experiments and evaluate results in terms of root mean square error (RMSE) for both PATT and the local PATT.

One-Dimensional Case. We begin with a population of $N_T = 5000$ treated units, with covariate X drawn from N(1, 1) and $N_C = 5000$ control units with X drawn from N(5, 1), which fits the EPBR data requirements. Denote the outcome variable as Y, and write the potential outcomes as $Y(t) = g_k(x, t)$ for $t = 0, 1$, where g_k is defined in the six ways indicated in Figure 5. The solid lines represent different choices for $g_k(x, 1)$ and the dotted lines represent each $g_k(x, 0)$; the vertical distance for a given value of x in each graph is a treatment effect. These functions represent constant, linear, and diverse nonlinear treatment effects. We construct the true outcome by applying the $g_k(x, t)$ to a sample drawn from the given population. We create the observed Y as the truth plus Gaussian noise, N(0, 0.3). We randomly sample from the original population of the treated units $n_T = 200$ units and from the population of the control units $n_C = 400$ units.

We generate 1000 random datasets; for each sample, we attempt to estimate PATT τ_k and the local PATT τ_k^m, and evaluate the RMSE for different matching methods. The first method (which we denote PS0) includes a propensity score estimated with a (main effects) logit model and with matching without replacement via nearest neighbors applied to the estimated score. We also estimate a set of CEM methods, with coarsening generated by progressively cutting the support of $\mathbf{X}$ from 2–11 equally sized intervals. We denote these matching solutions C2, C3, ..., C11. For PS0 $m_T = m_C = n_T$ and for CEM m_T and m_C are functions of the CEM solution. The weights w_i for PS0 are $w_i = 1/n_T$ and for CEM are the usual CEM weights.

Figure 6 reports the ratio of the RSME (based on PATT) between PS0 and the different CEM methods (on the horizontal axis) and for each of the six datasets (separate lines, numbered according to g_i). When the plotted points are below the dotted horizontal line drawn at a ratio of 1, the RMSE is better for the corresponding CEM method than for PS0. Thus, in this simulation, CEM has lower RMSE for 59 of the 60 experiments, and is approximately tied in the last (at the top left of the graph).

We also offer in Figure 7 the results when the target quantity of interest is changed from τ_k into the local PATT, τ_k^m, using the treated units matched by each CEM method. In this case, we report the absolute value of the RMSE because there is no propensity score or other base line to compare the different methods. This graph shows that CEM does not suffer as it coarsens more and drops more units; in fact, RMSE drops even as the number of observations (in parentheses beneath each label on the horizontal axis) declines. This makes sense, of course, because the variance is a function not only of n but also of the heterogeneity, which is reduced by matching. Thus, in these experiments, CEM has lower RMSE than PS0, even though the data were drawn to follow EPBR's requirements.

Multidimensional Gaussian Case, With Propensity Score Model Selection. We now consider Gaussian data with five covariates. We compare CEM with the standard propensity score estimated by the usual logit model ("PS0") and a more

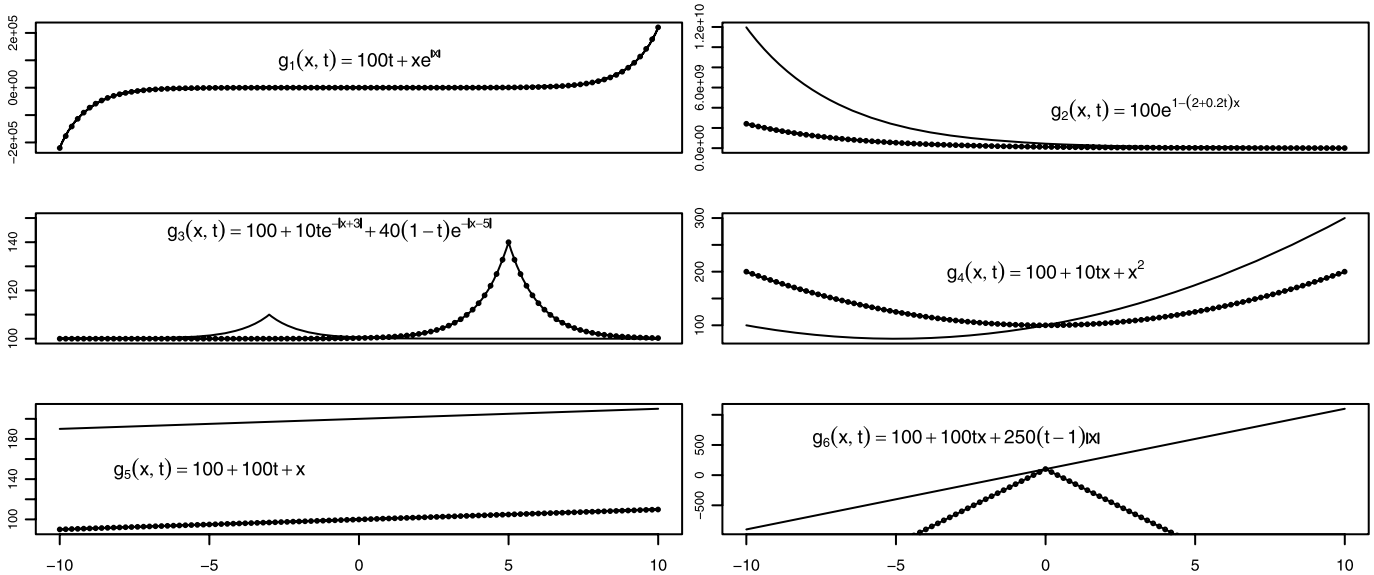

Figure 5. Graphs of x horizontally by $g_k(x, t)$ vertically, for lines $k = 1, \ldots, 6$. In each, the functions generating the treated $t = 1$ (solid lines) and control $t = 0$ (dotted lines) groups appear.

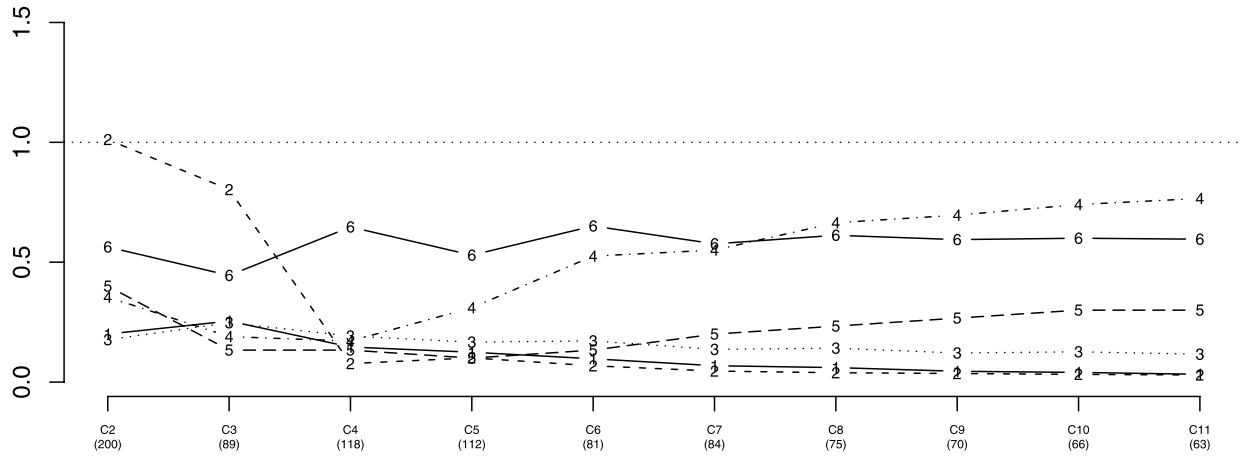

Figure 6. Ratio of RMSE of PS0 and 10 CEM methods for PATT. One dimensional case. Average values over 1000 Monte Carlo replications. In the graphs: $g_i = i$. A value less than 1 means the CEM method is preferable to PS0 in terms of RMSE. On the x axis Ck corresponds to a CEM solution with support of x cut into k equal sized intervals (and below in parentheses the number of treated matched units). Total number of treated units in the original sample is 200.

appropriate propensity score model optimized according to balance rather than fit Imbens and Rubin (2010, chap. 13) ("PS1"). For CEM, we consider a group of 50 different random coarsenings. For each variable X_i, $i = 1, \ldots, 5$, we cut the support of X_i by a random number of equispaced cutpoints selected from the uniform discrete distribution $U([3, 7])$. (Unlike our previous one dimensional simulation, the coarsenings here cannot be ordered, and so we order results for CEM according to the number of matched treated units.)

We draw five covariates with $N_T = N_C = 5000$ from $\mathrm{N}(\mathbf{0}, \mathbf{I})$ for the treated units and $\mathrm{N}(\mathbf{2}, \mathbf{I})$ for the control units, where $\mathbf{I}$ is the 5×5 identity matrix, $\mathbf{0} = (0, 0, 0, 0, 0)'$ and $\mathbf{2} = (2, 2, 2, 2, 2)'$. This again fits EPBR's data requirements. We use the following diverse multivariate $g_k(x, t)$ functions in the same way our previous simulation: $g_1(x, t) = 100 \cdot t + t \cdot x_1 \cdot e^{|x_2-2|} + \log(10+x_3) + 100 \cdot (1-t) \cdot x_2 \cdot e^{|x_4+2|} + x_5^2 + x_3 \cdot x_4 \cdot x_5$, $g_2(x, t) = 100 \cdot t + \sum_{i=1}^5 x_i$, $g_3(x, t) = 100 \cdot t + \sum_{i=1}^5 x_i + \sum_{i=1}^5 x_i^2$, $g_4(x, t) = 100 + 100 \cdot t \cdot \sum_{i=1}^5 x_i + 250 \cdot (t-1) \cdot \sum_{i=1}^5 x_i^2$, and $g_5(x, t) = 100 + 100 \cdot t \cdot \sum_{i=1}^5 x_i + 250 \cdot (t-1) \cdot \sum_{i=1}^5 x_i$. For each simulation, we randomly draw $n_T = 500$ treated units and $n_C = 1000$ control units.

Figure 8 reports the ratio of the RSME for PS0 to each of the different CEM methods, for PATT as the target quantity of interest. In this plot, CEMk corresponds to the kth CEM solution based on the 50 different coarsenings and the numerical labels on the lines correspond to function g_i, $i = 1, \ldots, 6$. The results indicate that CEM dominates the propensity score methods, as all experiments have lower RMSE than all the propensity score solutions.

Figure 9 reports the absolute value of the RMSE for the different methods when the target quantity of interest is changed to the local PATT defined by the treated units matched by CEM and PS1. In the top left corner of the plot, PS(i) is PS1 for function g_i, $i = 1, \ldots, 6$. Again we can see that the RMSE does not increase as the number of matched observations drops.

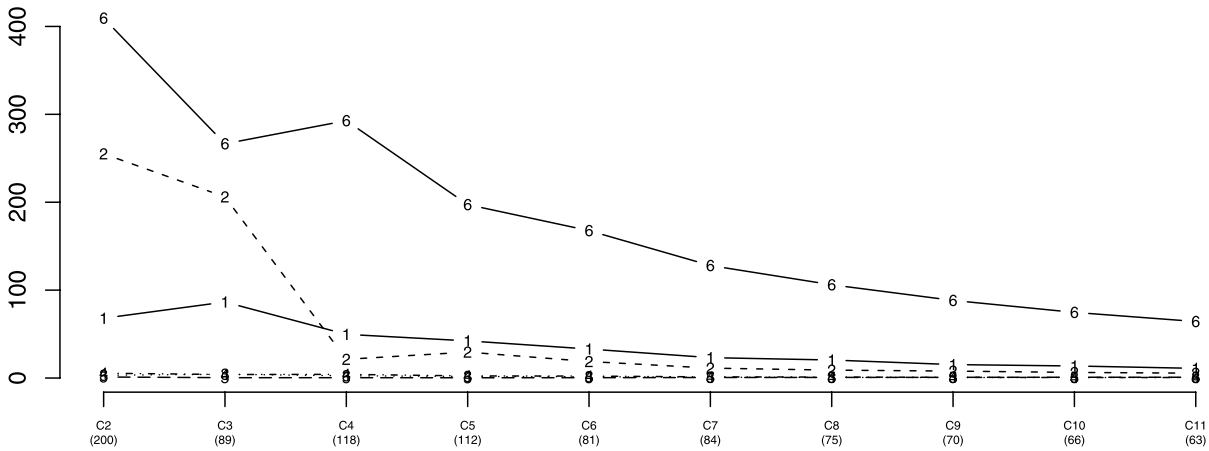

Figure 7. Absolute RMSE for different CEM solutions, for the local PATT. One-dimensional case. Average values over 1000 Monte Carlo replications. In the graphs: $g_i = i$. On the x axis Ck corresponds to a CEM solution with support of x cut into $k+2$ equal sized intervals (and below in parentheses the number of treated matched units). Lower values are better. Number of treated units in the original sample is 200.

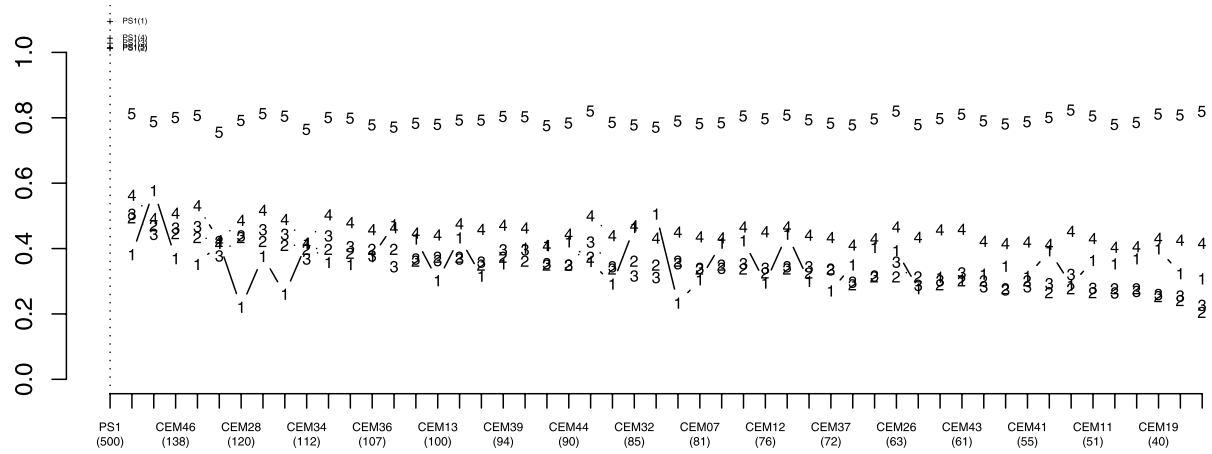

Figure 8. Ratio of RMSE of PS0 and CEM for PATT. Five-dimensional case. Average values over 1000 Monte Carlo replications for $g_i = i$. The x axis CEMk corresponds to the kth CEM solution (ordered by the number of treated matched units, given in parentheses). PS1(i) denotes the ratio of RMSE for PS0 to PS1 for g_i. A value less than 1 means that the given matching solution is preferable to PS0 in terms of RMSE. Number of treated units in the sample prior to matching: 500.

Lalonde Data. In this final analysis, we use the data from Lalonde (1986) as our population (with $N_T = 297$ and $N_C = 425$). From these data, we randomly sample $n_T = 150$ and $n_C = 300$ observations from the treated and control groups. We generate the outcome variable via $g_k(x,t)$ functions: $g_1(x,t) = 1000 + 2000 \cdot t \cdot \texttt{age} + \texttt{re74} + \log(1 + \texttt{re75}) + \texttt{black} \cdot \texttt{re75}^2$, $g_2(x,t) = 1000 + 2000 \cdot t + \texttt{age} + \texttt{re74} + \texttt{re75}^2$, $g_3(x,t) = 1000 + 2000 \cdot t + \texttt{age} + \texttt{re74} + \texttt{re75} + \texttt{black} + \texttt{education}$, and $g_4(x,t) = 1000 + 2000 \cdot t + \texttt{age} + \texttt{re74} + \texttt{re75} + \texttt{black} + \texttt{education} + \texttt{hispanic} + \texttt{nodegree} + \texttt{married}$.

Figure 10 reports the ratio of the RSME between PS0 and the other matching methods, for PATT as the target quantity of interest, while Figure 11 reports the absolute value of RMSE for target quantity of interest the local PATT. The notation in the figures is analogous that in the previous section. The results are also similar, in that CEM again clearly outperform the other methods, with lower RMSE and error that does not increase as matching becomes more stringent, leading to smaller matched datasets.

7. 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 a variety of desirable properties that should be of considerable use to applied researchers. We offer Coarsened Exact Matching as one such example, and demonstrate how it generates matching solutions that are better balanced and estimates of the causal quantity of interest that have lower root mean square error than methods under the older existing class, such as based on propensity scores, Mahalanobis distance, nearest neighbors, and optimal matching.

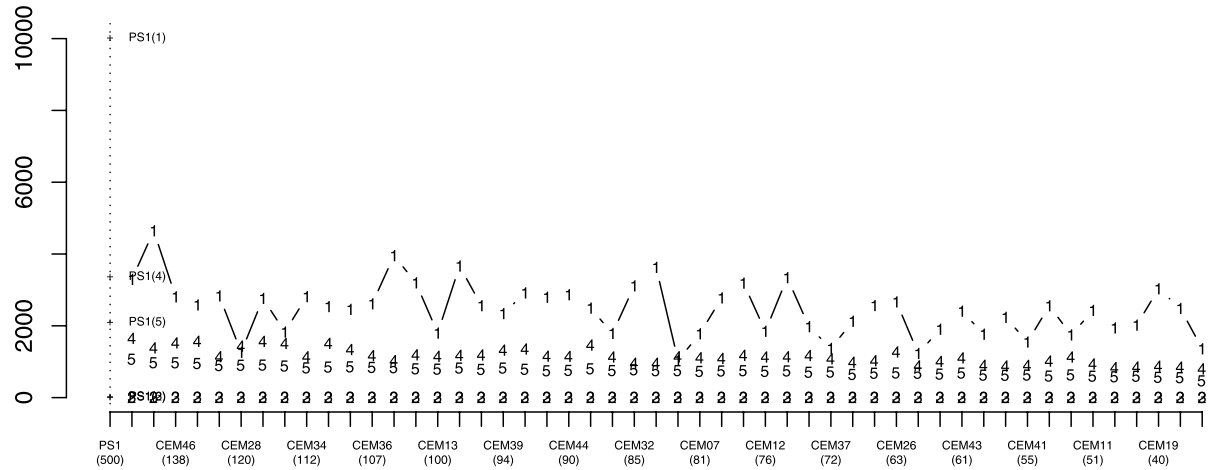

Figure 9. Absolute RMSE for different matching solutions, for PATT. Five-dimensional case. Average values over 1000 Monte Carlo replications for lines labeled as $g_i = i$. On the x axis CEMk corresponds to the kth CEM solution (see text) (with the number of matched treated units in parentheses). PS1(i) denotes the RMSE for PS1 for g_i. Lower values are better. Total number of treated units in the sample: 500.

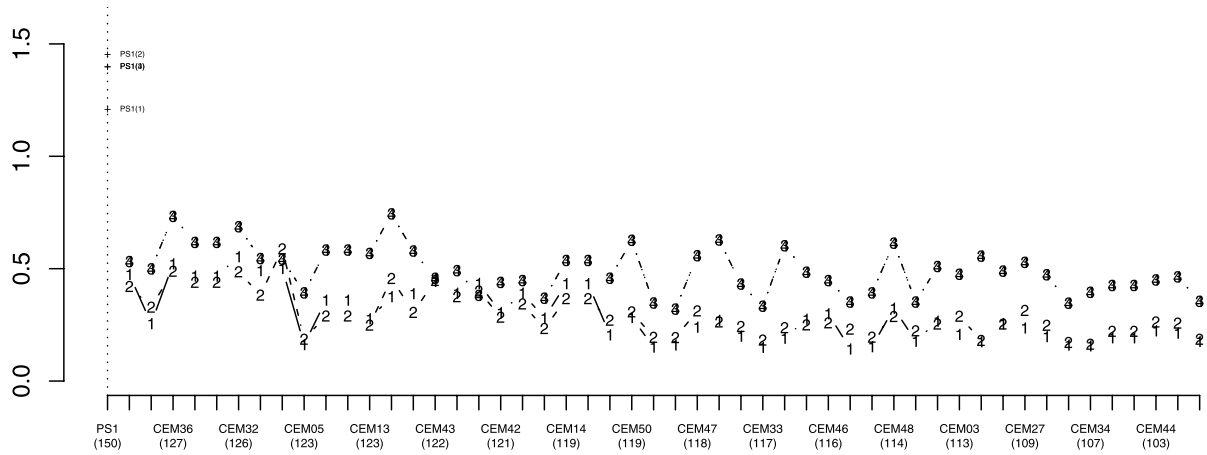

Figure 10. Ratio of RMSE of PS0 to CEM for PATT from the Lalonde data. Average values over 1000 Monte Carlo replications for lines labeled as $g_i = i$. On the x axis, CEMk corresponds to the kth CEM solution (see text) (with the number of matched treated units in parentheses). PS1(i) is the ratio of RMSE for PS0 to PS1 for function g_i. A value less than 1 means that the matching solution is preferable to PS0 in terms of RMSE. Total number of treated units in the sample: 150.

APPENDIX: PROOFS OF PROPOSITIONS IN SECTION 3

Proof of Proposition 2

Let us introduce the means by strata: $\bar{X}_{m^s_{T,j}} = \frac{1}{m^s_T}\sum_{i\in\mathcal{T}^s} X_{ij}$, $\bar{X}_{m^s_{C,j}} = \frac{1}{m^s_C}\sum_{i\in\mathcal{C}^s} X_{ij}$. Then $\bar{X}^w_{m_T,j} = \frac{1}{m_T}\sum_{i\in\mathcal{T}} X_{ij}w_i = \frac{1}{m_T} \times \sum_{s\in\mathcal{S}}\sum_{i\in\mathcal{T}_s} X_{ij} = \frac{1}{m_T}\sum_{s\in\mathcal{S}} m^s_T \bar{X}_{m^s_{T,j}}$ and $\bar{X}^w_{m_C,j} = \frac{1}{m_C}\sum_{i\in C} X_{ij} \times w_i = \frac{1}{m_C}\sum_{s\in\mathcal{S}}\sum_{i\in C_s} X_{ij}\frac{m_C}{m_T}\frac{m^s_T}{m^s_C} = \frac{1}{m_T}\sum_{s\in\mathcal{S}} m^s_T \bar{X}_{m^s_{C,j}}$. Hence, given that the mean is internal, in each stratum observations are at most far as ϵ_j; thus, $|\bar{X}^w_{m_T,j} - \bar{X}^w_{m_C,j}| \le \sum_{s\in\mathcal{S}}\frac{m^s_T}{m_T}|\bar{X}_{m^s_{T,j}} - \bar{X}_{m^s_{C,j}}| \le \sum_{s\in\mathcal{S}}\frac{m^s_T}{m_T}\epsilon_j = \epsilon_j$.

Proof of Proposition 3

We first rewrite $\bar{\mu}^k_{C,j}$

$$\bar{\mu}^k_{C,j} = \frac{1}{m_C}\sum_{s\in\mathcal{S}}\sum_{i\in\mathcal{C}^s}|X_{ij} - \bar{X}^w_{m_C,j}|^k w_i$$

$$\le \frac{1}{m_C}\sum_{s\in\mathcal{S}}\sum_{i\in\mathcal{C}^s}\left(|X_{ij} - \bar{X}^w_{m_T,j}| + |\bar{X}^w_{m_T,j} - \bar{X}^w_{m_C,j}|\right)^k w_i$$

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

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

by Proposition 2 we can write

$$\left(|X_{ij} - \bar{X}^w_{m_T,j}| + |\bar{X}^w_{m_T,j} - \bar{X}^w_{m_C,j}|\right)^k \le \sum_{h=0}^{k}\binom{k}{h}|X_{ij} - \bar{X}^w_{m_T,j}|^h \epsilon_j^{k-h} \le \epsilon_j^k \sum_{h=0}^{k}\binom{k}{h}|R_j|^h \epsilon_j^{-h}$$

$$= \epsilon_j^k \sum_{h=0}^{k}\binom{k}{h}\left|\frac{R_j}{\epsilon_j}\right|^h \le \epsilon_j^k \sum_{h=0}^{k}\binom{k}{h}\theta_j{}^h 1^{k-h} = \epsilon_j^k(\theta_j + 1)^k.$$

Therefore, $\bar{\mu}^k_{C,j} \le \epsilon_j^k(\theta_j + 1)^k \frac{1}{m_C}\sum_{s\in\mathcal{S}}\sum_{i\in\mathcal{C}^s} w_i = \epsilon_j^k(\theta_j + 1)^k$ because $\frac{1}{m_C}\sum_{s\in\mathcal{S}}\sum_{i\in\mathcal{C}^s} w_i = \frac{1}{m_C}\sum_{s\in\mathcal{S}}\sum_{i\in\mathcal{C}^s}\frac{m_C}{m_T}\frac{m^s_T}{m^s_C} = \frac{1}{m_T} \times \sum_{s\in\mathcal{S}} m^s_C\frac{m^s_T}{m^s_C} = 1$. Since $\frac{1}{m_T}\sum_{s\in\mathcal{S}}\sum_{i\in\mathcal{T}^s} w_i = 1$. The same bound

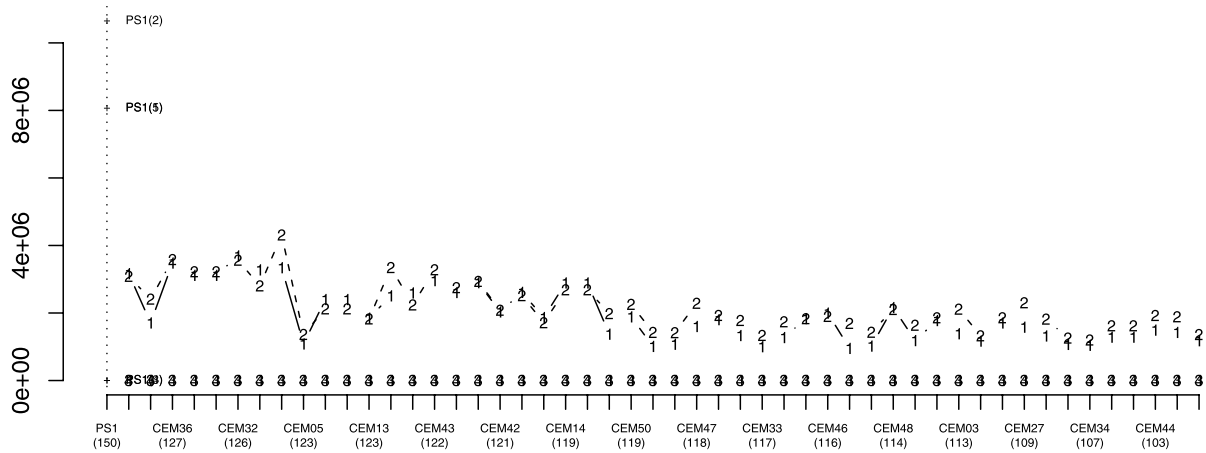

Figure 11. Absolute RMSE for different matching solutions, for the local PATT from the Lalonde data. Values averaged over 1000 Monte Carlo replications, with lines labeled as $g_i = i$. On the x axis CEMk corresponds to the kth CEM solution (see text) (with the number of matched treated units in parentheses). PS1(i) denotes the RMSE for PS1 for g_i. Lower values are better. Total number of treated units in the sample: 150.

exists for $\bar{\mu}^k_{T,j}$, so their absolute difference is $|\bar{\mu}^k_{T,j} - \bar{\mu}^k_{C,j}| \leq \epsilon_j^k(\theta_j + 1)^k$.

Proof of Proposition 4

Consider the qth empirical quantiles of the distribution of the treated and control units, $X^q_{m_T,j}$ and $X^q_{m_C,j}$. That is, $X^q_{m_T,j}$ is the qth ordered observation of the subsample of m_T matched treated units, and similarly for $X^q_{m_C,j}$. 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, $|X^q_{m_T,j} - X^q_{m_C,j}| < \epsilon_j$.

Proof of Proposition 5

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 $F^w_{m_T,j}(a_1) = F^w_{m_C,j}(a_1) = 0$. Then $F^w_{m_T,j}(b_1) = m_T^{s=1}/m_T$ 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

$$F^w_{m_C}(b_1) = \frac{m_C^{s=1}}{m_C} \cdot \frac{m_C}{m_T} \frac{m_T^{s=1}}{m_C^{s=1}} = \frac{m_T^{s=1}}{m_T}.$$

Thus, for each stratum, $F^w_{m_T,j}(b_s) = m^s_T/m_T = F^w_{m_C,j}(b_s)$, 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.

[Received October 2009. Revised December 2010.]

REFERENCES

Abadie, A., and Imbens, G. (2009), "A Martingale Representation for Matching Estimators," IZA Discussion Paper 4073, IZA, available at *http://ftp.iza.org/dp4073.pdf*. [351]

Box, G. E. P., Hunger, W. G., and Hunter, J. S. (1978), *Statistics for Experimenters*, New York: Wiley-Interscience. [349]

Cochran, W. G. (1968), "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies," *Biometrics*, 24, 295–313. [350]

Cochran, W. G., and Rubin, D. B. (1973), "Controlling Bias in Observational Studies: A Review," *Sankhya: The Indian Journal of Statistics, Ser. A*, 35, 417–466. [348]

Gu, X. S., and Rosenbaum, P. R. (1993), "Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms," *Journal of Computational and Graphical Statistics*, 2, 405–420. [354]

Hirano, K., Imbens, G. W., and Ridder, G. (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71 (4), 1161–1189. [346]

Ho, D., Imai, K., King, G., and Stuart, E. (2007), "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference," *Political Analysis*, 15, 199–236. Available at *http://gking.harvard.edu/files/abs/matchp-abs.shtml*. [345,346,349]

Iacus, S. M., and Porro, G. (2009), "Random Recursive Partitioning: A Matching Method for the Estimation of the Average Treatment Effect," *Journal of Applied Econometrics*, 24, 163–185. [349]

Imai, K., King, G., and Nall, C. (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. Available at *http://gking.harvard.edu/files/abs/cluster-abs.shtml*. [349]

Imai, K., King, G., and Stuart, E. (2008), "Misunderstandings Among Experimentalists and Observationalists About Causal Inference," *Journal of the Royal Statistical Society, Ser. A*, 171, 481–502. Available at *http://gking.harvard.edu/files/abs/matchse-abs.shtml*. [349]

Imbens, G., and Rubin, D. (2010), "Causal Inference," unpublished manuscript, Harvard University. [358]

King, G., and Zeng, L. (2007), "When Can History Be Our Guide? The Pitfalls of Counterfactual Inference," *International Studies Quarterly* (March), 183–210. Available at *http://gking.harvard.edu/files/abs/counterf-abs.shtml*. [349]

Lalonde, R. (1986), "Evaluating the Econometric Evaluations of Training Programs," *American Economic Review*, 76, 604–620. [352,353,359]

Racine, J. S., and Li, Q. (2009), "Efficient Estimation of Average Treatment Effects With Mixed Categorical and Continuous Data," *Journal of Business & Economic Statistics*, 27 (2), 203–223. [352]

Rosenbaum, P. R. (2002), *Observational Studies* (2nd ed.), New York: Springer Verlag. [348,352,354]

Rosenbaum, P. R., and Rubin, D. B. (1985a), "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score," *The American Statistician*, 39, 33–38. [346]

——— (1985b), "The Bias due to Incomplete Matching," *Biometrics*, 41 (1), 103–116. [354]

Rubin, D. B. (1976a), "Multivariate Matching Methods That Are Equal Percent Bias Reducing, I: Some Examples," *Biometrics*, 32 (1), 109–120. [346]

——— (1976b), "Multivariate Matching Methods That Are Equally Percent Bias Reducing, II: Maximums on Bias Reduction for Fixed Sampled Sizes," *Biometrics*, 32, 121–132. [345]

Rubin, D. B., and Stuart, E. A. (2006), "Affinely Invariant Matching Methods With Discriminant Mixtures of Proportional Ellipsoidally Symmetric Distributions," *The Annals of Statistics*, 34 (4), 1814–1826. [346]

Rubin, D. B., and Thomas, N. (1992), "Affinely Invariant Matching Methods With Ellipsoidal Distributions," *The Annals of Statistics*, 20 (2), 1079–1093. [346]

——— (1996), "Matching Using Estimated Propensity Scores, Relating Theory to Practice," *Biometrics*, 52, 249–264. [346]

Washington, E. L. (2008), "Female Socialization: How Daughters Affect Their Legislator Fathers' Voting on Woman's Issues," *American Economic Review*, 98 (1), 311–332. [354,355]