Matching for Causal Inference Without Balance Checking[1]

Stefano M. Iacus[2] Gary King[3] Giuseppe Porro[4]

October 15, 2008

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

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

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

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

Abstract

We address a major discrepancy in matching methods for causal inference in observational data. Since these data are typically plentiful, the goal of matching is to reduce bias and only secondarily to keep variance low. However, most matching methods seem designed for the opposite problem, guaranteeing sample size ex ante but limiting bias by controlling for covariates through reductions in the imbalance between treated and control groups only ex post and only sometimes. (The resulting practical difficulty may explain why many published applications do not check whether imbalance was reduced and so may not even be decreasing bias.) We introduce a new class of "Monotonic Imbalance Bounding" (MIB) matching methods that enables one to choose a fixed level of maximum imbalance, or to reduce maximum imbalance for one variable without changing it for the others. We then discuss a specific MIB method called "Coarsened Exact Matching" (CEM) which, unlike most existing approaches, also explicitly bounds through ex ante user choice both the degree of model dependence and the causal effect estimation error, eliminates the need for a separate procedure to restrict data to common support, meets the congruence principle, is approximately invariant to measurement error, works well with modern methods of imputation for missing data, is computationally efficient even with massive data sets, and is easy to understand and use. This method can improve causal inferences in a wide range of applications, and may be preferred for simplicity of use even when it is possible to design superior methods for particular problems. We also make available open source software for R and Stata which implements all our suggestions.

1 Introduction

Observational data are often inexpensive to collect, at least compared to randomized experiments, and so are typically in plentiful supply. However, key aspects of the data generation process — especially the treatment assignment mechanism — are unknown or ambiguous, and in any event are not controlled by the investigator. This generates the central dilemma of the field, which we summarize as: information, information everywhere, nor a datum to trust (with apologies to Samuel Taylor Coleridge).

Matching is a nonparametric method of controlling for some or all of the confounding influence of pretreatment control variables in observational data. The key goal of matching is to prune observations from the data so that the remaining data have better *balance* between the treated and control groups, meaning that the empirical distributions of the covariates (X) in the groups are more similar. Exactly balanced data means that controlling further for X is unnecessary (since it is unrelated to the treatment variable), and so a simple difference in means on the matched data can estimate the causal effect; approximately balanced data requires controlling for X with a model (such as the same model that would have been used without matching), but the only inferences necessary are only those relatively close to the data, leading to less model dependence and reduced statistical bias than without matching.

The central dilemma of matching in observational data means that model dependence and statistical bias are usually much bigger problems than large variances.[1] The key problem we address is that most matching methods seem designed for the opposite problem. They guarantee the matched sample size ex ante (thus fixing most aspects of the variance) and produce some level of reduction in imbalance between the treated and control groups (hence reducing bias and model dependence) only as a consequence and only sometimes. That is, the less important criterion is guaranteed by the procedure, and any success at achieving the most important criterion is uncertain and must be checked ex post. Because the methods are not designed to achieve the goal set out for them, numerous applications of matching methods fail the check and so need to be repeatedly tweaked and rerun.

This disconnect gives rise to the most difficult problem in real empirical applications of matching: In many observational data sets, finding a matching solution that improves balance between the treated and control groups is easy for most covariates, but the result often leaves balance "slightly" worse for some other variables at the same time. Thus, analysts are left with the nagging worry that all their "improvements" in applying matching may actually have increased bias.

Continually checking balance, rematching, and checking again until balance is improved on all variables is the best current practice with existing matching algorithms. The process needs to be repeated multiple times because any change in the matching algorithm may alter balance in unpredictable ways on any or all variables. Perhaps the difficulty in following best practices in this field explains why many applied articles do not measure or report levels of imbalance at all, and appear to run some chosen matching algorithm only once. Moreover, even when balance is checked and reported, at best a table comparing means in the treatment and control groups is included. Imbalance due to differences in variances, ranges, covariances, and higher order interactions are typically ignored. This of course is a real mistake, since any one application of most existing matching algorithms is not guaranteed (without balance checking) to do any good at all. Of course, its hard to blame applied researchers who might quite reasonably expect that a method touted for its ability to reduce imbalance might actually do so when used once.

We introduce a new Monotonic Imbalance Bounding (MIB) class of matching methods, and

[1] As Rubin (2006) writes, "First, since it is generally not wise to obtain a very precise estimate of a drastically wrong quantity, the investigator should be more concerned about having an estimate with small bias than one with small variance. Second, since in many observational studies the sample sizes are sufficiently large that sampling variances of estimators will be small, the sensitivity of estimators to biases is the dominant source of uncertainty."

discuss a simple and widely applicable method from the class, that inverts the process and thus guarantees a fixed level of balance ex ante between the treated and control groups. This level is chosen by users on the basis of specific, intuitive substantive information which they demonstrably have. (If you understand the trade-offs in drawing a histogram, you will understand how to use this method.) Improvements in the bound on balance for one covariate can be studied in isolation with our approach because they are known to have no effect on the maximum imbalance in each of the other covariates. We show that our method controls, up to specified levels, for all imbalances in central absolute moments, comoments, coskewness, interactions, nonlinearities, and other multidimensional distributional differences between treated and control groups, which most other methods do not address. In fact, the method controls not only covariate imbalance; it also explicitly controls the degree of model dependence and, more importantly, the size of estimation error (and statistical bias) in the causal quantity of interest. Although most matching methods attempt to approximate a classic experiment with complete randomization, our approach produces additional local balance (and the resulting efficiency) by attempting to approximate the superior randomized block experimental design.

Whereas most prior matching methods must be preceded by an entirely different algorithm limiting covariates to areas of common empirical support, our approach does this automatically as a natural part of the same matching algorithm. The method is approximately invariant to measurement error and the global multivariate differences between treated and control groups are controllable by easy-to-understand local decisions about specific variables and their measurement characteristics. The method avoids the troubling difficulty in existing matching methods of working with modern methods of imputation for missing data. The algorithm is fast and efficient, even with extremely large data sets, with speed scaling linearly with the number of variables. The same algorithm can be used for binary or multi-category treatments, and for pre-randomization blocking in experiments. With this paper, we make available free, open source, and easy-to-use software that implements these methods.

Our approach can improve causal inferences across a very wide range of applications, and thus is designed as an easy default choice or first line of defense in protecting users from the threats to validity in making causal inferences. The method is not necessarily optimal in every application and may be out-performed in specific cases by methods designed or tuned for specific data sets in ways we discuss, usually at the cost of more work designing special procedures. In what follows, we introduce our notation and setup (Section 2), describe the method we introduce (Section 3), characterize the new class of matching methods into which our method falls (Section 4), discuss the methods other properties (Section 5), and extend it in various ways (Section 6). We then show in simulated and real data how it works in practice (Section 7) and conclude with a discussion of what can go wrong when using this approach (Section 8).

2 Preliminaries

This section describes our setup. It includes our notation, definitions of our target quantities of interest, some simplifying assumptions, a brief summary of existing matching methods and post-estimation matching, a general characterization of error in estimating the target quantities, and how to measure imbalance.

2.1 Notation

Consider a sample of n units randomly drawn from a population of N units, where $n \leq N$. For unit i, denote T_i as an indicator variable with value $T_i = 1$ if unit i receives the 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 denoted 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, the observed outcome is $Y_i = T_i Y_i(1) + (1 - T_i)Y_i(0)$ and so $Y_i(0)$ is unobserved if i receives treatment and $Y_i(1)$ is unobserved if i does not receive treatment.

To compensate for the observational data problem where the treated and control groups are not necessarily identical before treatment (and, lacking random assignment, not the same on average), matching estimators attempt to control for pre-treatment covariates. For this purpose, we denote $X = (X_1, X_2, \ldots, X_k)$ as a k-dimensional data set, where each X_j is a column vector of observed values of pre-treatment variable j for the n sample observations (possibly drawn from a population, of size N). That is, $X = [X_{ij}, i = 1, \ldots, n, j = 1, \ldots, k]$.

2.2 Quantities of Interest

As usual, the treatment effect for unit i, $\text{TE}_i = Y_i(1) - Y_i(0)$, is unobserved. All relevant causal quantities of interest are functions of TE_i, for different groups of units, and so must be estimated. The most common include the *sample* (SATE) and *population* (PATE) *average treatment effect*:

$$\text{SATE} = \frac{1}{n}\sum_{i=1}^{n} \text{TE}_i \qquad \text{PATE} = \frac{1}{N}\sum_{i=1}^{N} \text{TE}_i,$$

and the *sample* (SATT) and *population* (PATT) *average treatment effect on the treated*:

$$\text{SATT} = \frac{1}{n_T}\sum_{i \in T} \text{TE}_i \qquad \text{PATT} = \frac{1}{N_T}\sum_{i \in T^*} \text{TE}_i,$$

where $n_T = \sum_{i=1}^{n} T_i$ and $T = \{1 \leq i \leq n : T_i = 1\}$ and $N_T = \sum_{i=1}^{N} T_i$ and $T^* = \{1 \leq i \leq N : T_i = 1\}$. See Imbens (2004); Morgan and Winship (2007).

Although SATE and SATT are quantities of interest in and of themselves, without regard to a population beyond the sample data, if the sample is randomly drawn from the relevant population, $E(\text{SATE}) = \text{PATE}$ and $E(\text{SATT}) = \text{PATT}$ (where the expected value operator averages over repeated samples). Separately, if T is randomly assigned, $E(\text{SATT}) = \text{SATE}$ (where the expected value operator here averages over different random assignments of T).

2.3 Simplifying Assumptions

First, we make the standard assumption, known as "no omitted variable bias" in the social sciences, "ignorability" in statistics, and "unconfounding" in epidemiology, that X is defined such that conditional on X, the treatment variable is independent of the potential outcomes: $P(T|X, Y(0), Y(1)) = P(T|X)$.

Second, as with most matching-based estimators, we focus on SATT (or PATT) so that, by retaining all treated units and matching on controls, the process of pruning observations does not necessarily change the target quantity of interest, as would not be the case for SATE or PATE (which prune both treated and control units). This convention in the methodological literature is a reasonable but practical decision, chosen because SATE or PATE are not amenable to straightforward matching-based estimation. Of course, the initial set of units in observational data is almost always arbitrary to some degree anyway. This decision implies that, for each observation, $Y_i(1)$ is always observed, while $Y_i(0)$ is always estimated (usually by choosing values from the control units via some matching algorithm or applying some model).

And finally, another common practical decision is to go another step and match both treated and control units. The result changes the estimand, which is not unreasonable so long as one is transparent about the choice and the consequences in terms of the new set of units over which the causal effect is defined. We thus also follow this convention and recommend it to users for applications (as, e.g., Crump et al., 2006), although all methods and results we discuss below also hold if we keep all treated units and thus retain a fixed target quantity of interest. The same

change in the quantity of interest is common in other methods for observational data, such as local average treatment effects and regression discontinuity designs. The practice is even similar to most randomized experiments which do not select subjects randomly, and so have an estimand that is also defined over a somewhat arbitrary set of units (such as patients who happen to show up at a hospital and agree to be enrolled in a study, or those who fit conditions researchers believe will demonstrate larger causal effects).

2.4 Overview of Existing Matching Methods

This section outlines the most commonly matching methods. To begin, *one-to-one exact matching* estimates the unobserved $Y_i(0)$, corresponding to each observed treated unit i (with outcome value Y_i and covariate values X_i), with the outcome value of a control unit (denoted $\tilde{Y}_\ell$ with covariate values $\tilde{X}_\ell$), chosen such that $\tilde{X}_\ell = X_i$. We denote the resulting estimate of $Y_i(0)$ as $\hat{Y}_i(0)$. To increase efficiency the alternative *exact matching* algorithm uses *all* control units that match each treated unit (i.e., all X_i such that $\tilde{X}_\ell = X_i$).

Unfortunately, in most real applications with covariates sufficiently rich to make ignorability assumptions plausible, insufficient units can be exactly matched. Thus, analysts must choose one of the existing *approximate matching* methods, the best practice for which involves two separate steps. The first step drops control units (and sometimes treated units) outside the common empirical support of both groups or in the sample region requiring extrapolation. The second step then matches the treated unit to some control observation $\tilde{X}$ that, if not exactly X, is close by some metric. The second step of most existing approximate matching procedures can be distinguished by the choice of metric. For example, nearest neighbor Mahalanobis matching chooses the closest control unit to each treated unit (among those within the common empirical support), using the Mahalanobis distance metric. For another example, nearest neighbor propensity score matching first summarizes the vector of covariate values for an observation by the scalar propensity score, which is the probability of treatment given the vector of covariates, estimated in some way, typically via a simple logit model. Then the closest control to each treated unit is used as a match, with the distance defined by the absolute difference between the two scalar propensity score values. Other options include optimal, subclassification, full genetic, and other procedures.) Since the second step in existing algorithms do not guarantee an improvement in balance except under specialized conditions, the degree of imbalance must be measured, the matching algorithm must be respecified, and imbalance checked again, etc., until a satisfactory solution is reached. (For example, the correct specification of the propensity score is not indicated by measures of fit, only by whether matching on it achieved balance.)

An additional problem for existing approximate matching methods is that most of the technologies used for matching in the second step are unhelpful for completing the first step. For example, the propensity score can be used to find the area of extrapolation only after we know that the correct propensity score model has been used. However, the only way to verify that the correct propensity score model has been specified is to check whether matching on it produces balance between the treated and control groups on the relevant covariates. But balance cannot be reliably checked until the region of extrapolation has been removed. To avoid this type of infinite regress, researchers use entirely different technologies for the first step, such as kernel density estimation (Heckman, Ichimura and Todd, 1997) or dropping control units outside the hyper-rectangle (Iacus and Porro, 2008, forthcoming) or convex hull (King and Zeng, 2006) of the treated units. The method we introduce below avoids these problems by satisfying both steps simultaneously in the same algorithm.

2.5 Post-Matching Estimation

Matching methods are data preprocessing algorithms, not statistical estimators. Thus, after preprocessing, some type of estimator must be applied to the data to make causal inferences. For

example, if one-to-one exact matching is used, then a simple difference in means between Y in the treated and control groups provides a fully nonparametric estimator of the causal effect. When the treated and control groups do not match exactly, the estimator will necessarily incorporate some modeling assumptions designed to span the remaining differences, and so results will be model-dependent to some degree (King and Zeng, 2007). Preprocessing via matching can greatly reduce the degree of modeling necessary and thus also the degree of model dependence (Ho et al., 2007).

Under a matching method that produces a one-to-one match (or in general any match that has a fixed positive number of treated and control units across strata), any analysis method that might have been appropriate without matching can alternatively be used on the matched data set with the benefit of having a lower risk of model dependence (Ho et al., 2007) including for example specially designed nonparametric methods (Abadie and Imbens, 2007).

When different numbers of control units are matched to each treated unit — or in general if different numbers of treated and control units appear in different strata, as in exact matching — the analysis model must weight or adjust for the different stratum sizes. In this situation, the simplest SATT estimator is a weighted difference in means between the treated and control groups, or equivalently a weighted linear regression of Y on T, (using weights defined in Appendix A). We can go further by trying to span the remaining imbalance via a weighted regression of Y on T and X. In either regression, the coefficient on T is our SATT estimate. Alternatively, to avoid the implicit constant treatment effect assumption of the regression approach, we can apply a statistical model within each stratum without weights and average the results across stratum with appropriate weights; when few observations exist within each stratum, a Bayesian, empirical Bayes, or random effects model can be applied in the same way. Finally, nonlinear (or linear) models may also be fit to all the data and used to predict, for each treated unit, the unobserved potential outcome under control $Y_i(0)$ given its observed covariate values X_i, with the treated unit-level estimated causal effects averaged over all treated units.

2.6 Quantifying Estimation Error

We derive the precise point of this balance checking here, as well as its connection to the real goal: accurate estimation of the causal effect. For simplicity, we analyze the case where the analysis method used after preprocessing is the simple difference in means. Begin by writing the unobserved potential outcome for each unit as

$$Y_i(0) = g_0(X_i) = g_0(X_{i1}, \ldots, X_{ik}). \tag{1}$$

where g_0 is an unknown function (cf. Imai, King and Stuart, 2008). If (1) included an error term that affects $Y_i(t)$ but is unrelated to T, it would be implied by the ignorability assumption. Our results would not be materially changed if it were included, except we would have to add expected values or probability limits. We omit it here for simplicity and because the concepts of repeated samples from the same data generation process, and samples that grow without limit, are forced analogies in many observational data sets.

We now decompose the unit-level treatment effect, TE_i, into the estimated treatment effect, $\widehat{\mathrm{TE}}_i = Y_i(1) - \hat{Y}_i(0)$, and the error in estimation. We do this by substituting into the definition of the true treatment effect $Y_i(1) = \widehat{\mathrm{TE}}_i + \hat{Y}_i(0)$ and using (1) as $\mathrm{TE}_i = Y_i(1) - Y_i(0) = \widehat{\mathrm{TE}}_i + \mathcal{E}_0(\tilde{X}_i, X_i)$, where $\mathcal{E}_0(\tilde{X}_i, X_i) \equiv g_0(\tilde{X}_i) - g_0(X_i) = \hat{Y}_i(0) - Y_i(0)$ is the unit level treatment effect error (not an expected value). Then we aggregate this over treated units into $\mathrm{SATT} = \frac{1}{n_T}\sum_{i\in T}\mathrm{TE}_i = \widehat{\mathrm{SATT}} + \bar{\mathcal{E}}_0$ where $\widehat{\mathrm{SATT}} = \sum_{i\in T}\widehat{\mathrm{TE}}_i/n_T$ and the average estimation error is

$$\bar{\mathcal{E}}_0 \equiv \frac{1}{n_T}\sum_{i\in T}\mathcal{E}_0(\tilde{X}_i, X_i) = \frac{1}{n_T}\sum_{i\in T}[g_0(\tilde{X}_i) - g_0(X_i)]. \tag{2}$$

The ultimate goal of matching-based estimators is to reduce the absolute matching error, $|\bar{\mathcal{E}}_0|$. This goal can be parsed into two (nonadditive) components. The first component of matching error is the *imbalance* between the control and treatment groups, or in other words the difference between the empirical distribution of the pre-treatment covariates for the control group $p(\tilde{X}|T=0)$ and treated group $p(X|T=1)$ in some chosen metric (such as those discussed in Section 2.7). The second component is the *importance* of each of the variables and their interactions in influencing Y given T. The two components are formalized in (2), where the difference between $\tilde{X}_i$ and X_i represents local imbalance for treated observation i and the unknown function g_0 represents the importance of different parts of the covariate space. If preprocessing results in exact matches between the treatment and control groups, imbalance is eliminated and $|\bar{\mathcal{E}}_0|$ vanishes, no matter what g_0 is. When that lucky situation does not occur, the two components must be considered together.

2.7 Measuring Imbalance

The goal of measuring imbalance is to summarize the difference between the multivariate *empirical* distribution of the pre-treatment covariates for the treated $p(X|T=1)$ and matched control $p(\tilde{X}|T=0)$ groups. Unfortunately, many matching applications do not check balance. Most of those which do check balance only compare the univariate absolute difference in means in the treated and control groups:

$$I_1^{(j)} = \left| \bar{X}^{(j)}_{m_T,w} - \bar{X}^{(j)}_{m_C,w} \right|, \quad j = 1, \ldots, k \tag{3}$$

where $\bar{X}^{(j)}_{m_T,w}$ and $\bar{X}^{(j)}_{m_C,w}$ denote weighted means of the group of m_T treated units and m_C control units matched, with weights appropriate to each matching method.

Sometimes researchers argue that only matching the mean is necessary because most analysis models used after or in place of matching (such as regression) only adjust for the mean. However, the purpose of matching is to reduce model dependence, and so it does not make sense to assume that the analysis model is correct, as implied by this argument; for model independent inferences, matching as much of the entire empirical distribution as possible must be the goal.

A few have measured imbalance in univariate moments, univariate density plots, propensity score summary statistics, or the average of the univariate differences between of the empirical quantile distributions (Austin and Mamdani, 2006; Imai, King and Stuart, 2008; Rubin, 2001). Except for the occasional discussion about using the differences in covariances, most researchers ignore all aspects of multivariate balance not represented in these simple variable-by-variable summaries. Unfortunately, improving on current practice by applying existing methods of comparing multivariate histograms — such as Pearson's χ^2, Fisher's G^2, or models for contingency tables — would typically work poorly because of the numerous zero cell values.

Our alternative idea is to measure the multivariate differences between $p(X|T=1)$ and $p(\tilde{X}|T=0)$ via an L_1-type distance. We first choose the number of bins for each continuous variable via standard automated univariate histogram methods and with categorical variables left as is (see Section 6.6.1). (If prior information indicates that some variables are more important than others in predicting the outcome, one might choose to use more bins for that variable. Either way, the bin sizes must be defined ex ante and not necessarily related to any matching method, including our proposal.[2]) Then, we cross-tabulate the discretized variables as $X_1 \times \cdots \times X_k$ for the treated and control groups separately, and record in each cell the k-dimensional relative frequency for the treated $f_{\ell_1 \ldots \ell_k}$ and control $g_{\ell_1 \ldots \ell_k}$ units, where the number of bins or levels of categorical variables ℓ_j may vary for each X_j. Then our measure of imbalance is the absolute difference over

[2] Although this initial choice poses all the usual issues and potential problems when choosing bins in drawing histograms, we use it only as a fixed reference to evaluate pre and post matching imbalance.

all the cell values:

$$\mathcal{L}_1(f,g) = \sum_{\ell_1 \cdots \ell_k} |f_{\ell_1 \cdots \ell_k} - g_{\ell_1 \cdots \ell_k}| \tag{4}$$

where the summation is over all cells of the multivariate histogram. An important property is that the typically numerous empty cells do not affect $\mathcal{L}_1(f,g)$, and so the summation in (4) has at most n nonzero terms. The relative frequencies also control for potentially different sample sizes between the treated and control groups. Equation 4 is similar to Cochran and Rubin (1973) except that we directly compute the empirical density and require no normality assumption.

Denote f^m and g^m as the empirical frequencies for treated and control units after the match corresponding to f and g before, and use the same discretization for both the treated and control units. Then a good matching method will result in matched sets such that $\mathcal{L}_1(f^m, g^m) \leq \mathcal{L}_1(f,g)$. We also sometimes use $\mathcal{L}_1^{(j)}$ for each covariate separately in order understand how the overall imbalance $\mathcal{L}_1$ projects onto each dimension.

Since the units of $\mathcal{L}_1$ are not on the scale of the original variables, we also recommend examining variable-level measures such as the first few moments and several quantiles of each variable, but this is primarily to get a feel for the data. We also propose a more sensitive variable-level measure for use in comparing matching solutions, as follows. (It is not defined in the original data and so cannot be used to study the reduction in imbalance.) Because any matching method produces a set of S strata, we define a measure of *local imbalance* as

$$I_2^{(j)} = \frac{1}{S} \sum_{s=1}^{S} \left| \bar{X}^{(j)}_{m_T^s} - \bar{X}^{(j)}_{m_C^s} \right|, \quad j = 1, \ldots, k \tag{5}$$

where $s = 1, \ldots, S$ are the strata generated by the matching method and $\bar{X}^{(j)}_{m_T^s}$ and $\bar{X}^{(j)}_{m_T^s}$ are the empirical means of variable X_j for the treated and control units in strata s and m_T^s and m_C^s are the numbers of treated and control units matched in stratum s.

3 Coarsened Exact Matching

Because the method we offer here is so simple, we describe it here, before characterizing the general class of MIB methods into which it falls. The method is also part of the diverse set of approaches based on subclassification (aka "stratification" or "intersection" methods). We call our particular method CEM for "Coarsened Exact Matching" (or "Cochran Exact Matching" since, although variants of it had already been used for decades, the first formal analysis of any subclassification-based method appeared in Cochran 1968). As we show, it has always been available, requires no complicated concepts, algorithms, or mathematics, and ameliorates a wide range of causal inference problems and can improve many existing methods.

The basic idea is to coarsen each variable by recoding so that substantively indistinguishable values are grouped and assigned the same numerical value (groups may be the same size or different sizes, depending on the substance of the problem). Then the "exact matching" algorithm is applied to the coarsened data to determine the matches. Finally, the coarsened data are discarded and the original (uncoarsened) values of the matched data are retained. This procedure therefore assigns to matching the task of eliminating all differences between the treated and control groups beyond some chosen level. Differences eliminated include all multivariate nonlinearities, interactions, moments, quantiles, and other distributional differences beyond the chosen level of coarsening. The remaining differences are thus all within small, coarsened strata and so are highly amenable to being spanned by a statistical model without risk of much model dependence.

CEM produces variable sized strata. If this is not convenient and enough data are available, users can produce a one-to-one match by randomly selecting the desired number of treated and control units from those within each stratum or apply an existing method within strata (see Section 6.3).

3.1 Coarsening Choices

Coarsening is almost intrinsic to the act of measurement. Even before the analyst obtains the data, the quantities being measured are typically coarsened to some degree. Just as a photograph taken with more powerful lenses produce more detail, so it is with better measurement devices of all kinds. Data analysts take what they can get, but recognize that whatever they get has likely been coarsened to some degree first. Variables like gender or the presence of war coarsen away enormous heterogeneity within the given categories.

But coarsening frequently does not stop once the analyst has the data. Data analysts recognize that many measures include some degree of noise and, in their ongoing efforts to find a signal amidst the noise, often voluntarily coarsen the data themselves. For example, political scientists often recode the 7-point partisan identification scale as Democrat, independent, and Republican; Likert issue questions into agree, neutral, and disagree; and multi-party vote returns into winners and losers. Many social scientists use a broad three or four category measure for religion, even when information is available for numerous specific denominations. Occupation is almost always coarsened into three or four categories. Economists and financial analysts commonly use highly coarsened versions of the U.S. Security and Exchange Commission industry codes for firms even though the same data source offers far more finely grained coding. Epidemiologists routinely dichotomize all their covariates on the theory that grouping bias is much less of a problem than getting the functional form right. Coarsening is also common for Polity II democratization scores, the International Classification of Disease codes, and numerous other variables.

Since the original values can still be used at the analysis stage to estimate the causal effect, coarsening for CEM involves less onerous assumptions than that made all the time by researchers who make the coarsening permanent. Of course, although coarsening in CEM is safer than at the analysis stage, the two procedures are similar in spirit since the coarsened information in both is thought to be relatively unimportant — small enough with CEM to trust to statistical modeling and in data analysis to ignore altogether.

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. The CEM procedure requires a coarsening operator and the values the operator produces, which we now introduce more formally.

3.2 The Coarsening Operator

Denote by Ξ_j the set on which variable X_j takes values, which may be the real line, the set of integers, or another abstract set (such as labels for nominal variables, ordered labels for ordinal variables, etc.), and let $\Xi = \Xi_1 \times \Xi_2 \times \cdots \times \Xi_k$ be the product space on which the data set X lives, i.e. $X \in \Xi$. Denote the number of distinct observed values of variable X_j as θ_j^*, where we collect the set of all these counts as $\theta^* = \{\theta_1^*, \ldots, \theta_k^*\}$. Whether X_j is categorical or continuous, X_j will never have more than n distinct values and so $\theta_j^* \leq n$. We also define a set $\Theta_j = \{1, \ldots, \theta_j^*\}$ ($j = 1, \ldots, k$), as well as the k-dimensional set of indexes $\Theta = \Theta_1 \times \Theta_2 \times \cdots \times \Theta_k$.

Define θ_j as the number of distinct values the user *chooses* variable X_j to have when coarsened, where $\theta_j \leq \theta_j^* \leq n$ for all j and $\theta = \{\theta_1, \ldots, \theta_k\}$. Then define the coarsening operator as $G_\theta(X) = G(X; \theta) : \Xi \times \Theta \mapsto \Xi$, where the amount of coarsening is determined by θ. Of course, if the number of distinct values for all variables is the same as the original data set then $G_{\theta^*}(X) = X$. Although written in matrix form, this operator works variable by variable, with the result being a copy of the original data set in which each value is recoded.

3.3 Values of the Coarsened Variables

We recommend that coarsened values be chosen in a completely customized way based on substantive knowledge of the measurement scale of each variable. The number of adjustable parameters

in CEM is thus at least k, but the tradeoff is normally worth it since these parameters will typically be well-known to users (but see Section 6.3).

We offer here reasonable operational defaults for continuous, nominal, and ordered variables, respectively, and some examples. For continuous variables, denote the range of X_j as $R_j = M_j - m_j$ where $M_j = \max_{i=1,\ldots,n} X_{ij}$ and $m_j = \min_{i=1,\ldots,n} X_{ij}$. Then, the user *chooses* ϵ such that $0 < \epsilon_j \leq R_j$. The case $\epsilon_j = R_j$ corresponds to all the observations grouped in a single interval, and, $\epsilon_j^* = R_j/\theta_j^*$ is the initial coarsening level. The relationship between the number of distinct values X_j is coarsened into by choosing θ_j, and the length of each interval ϵ, is

$$\theta_j \leq \min\left(\theta_j^*, \left\lceil \frac{R_j}{\epsilon_j} \right\rceil\right) \tag{6}$$

where $\lceil x \rceil$ is the ceiling function. For a fixed θ_j, the corresponding value of ϵ_j is such that $\epsilon_j \leq R_j/\theta_j$, if $1 \leq \theta_j < \theta_j^*$ and $\epsilon_j = \epsilon_j^*$ for $\theta_j = \theta_j^*$. But of course, in applications, we choose ϵ_j or θ_j by (6).

For example, denote $\theta = (\theta_1, \theta_2^*, \ldots, \theta_k^*)$ and let X_1 be a numeric variable. Define $G_\theta(X) = \tilde{X}$, where $\tilde{X}$ is a data set $\tilde{X} = (X_1', X_2, \ldots, X_k)$ with X_1' obtained from X_1 by grouping it into $\theta_1 < \theta_1^*$ intervals, each of length ϵ_j. If annual income is measured to the penny, then it is difficult to see objections to setting the ϵ_j interval length to be \$1.00. In most applications, however, the interval could be a good deal larger without any real loss of relevant information. For one, it could reasonably be set to the average uncertainty a respondent would likely have about his or her income or the daily variability in actual income. For the wealthy, this can be a large figure. For data with people of many different incomes, the user may wish to let ϵ_j vary with the value of the variable, presumably with larger values for larger incomes. Similarly, smaller intervals may be useful for lower incomes and possibly with \$0 a logically distinct group. In these situations, our proofs below change only slightly (replacing ϵ with its maximum).

The second category of variables are nominal, which we do not coarsen unless the user makes specific choices for how the coarsening would take place. For one example, imagine a survey question about religion that asks for about the specific denomination, including say 6 protestant denominations, 3 Jewish, 1 catholic, and 2 Muslim. For this example, a reasonable choice for some applied problems would be to coarsen to these broader categories. Of course, for some problems, where the differences among the denominations with the broad categories were of substantive importance, this would not be advisable. Similar examples would include the U.S. Security and Exchange Commission code for firms, which is published in a hierarchy designed for use by coarsening occupation codes, etc.

Our final variable type is ordered factors. Since most ordered variables are intended to be approximately interval valued, our default procedure is to treat them as such. We thus use our procedure for coarsening continuous variables and set θ_j to some smaller value than θ_j^*, such as $[\theta_j^*/2]$. Like any default, this is not universally applicable, and better choices may be available in some applications. For example, most 7-point Likert scales have a prominent neutral category and so can often best be coarsened into $\theta_j = 3$ groups as: {completely disagree, strongly disagree, disagree}, {neutral}, {agree, strongly agree, completely agree}.

4 Classes of Matching Methods

The matching literature includes many methods, but only a single class of methods has been characterized, the so-called Equal Percent Bias Reducing (EPBR) methods. In introducing EPBR, Rubin (1976*c*) wrote "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." Thus, in addition to EPBR, we describe a new class, called

Monotonic Imbalance Bounding (MIB) methods, which covers this multivariate nonlinear case and other features.

Each class of matching methods is designed to avoid, in different ways, the problem of making balance worse on some variables while trying to improve it for others — EPBR by changing the imbalance on all variables at the same time by the same amount, and MIB by changing one variable's imbalance while not affecting maximum imbalance on the others. In addition, whereas EPBR methods fix the matched sample size ex ante and balance is computed ex post, MIB methods fix the maximal imbalance ex ante and produce a matched sample size ex post. Satisfying EPBR requires a matching method with certain properties as well as data of a special type, whereas satisfying MIB requires a matching method with different properties but no restrictions on data types. CEM is the simplest method within the MIB class.

4.1 Equal Percent Bias Reducing Methods

Suppose X is the realized value of a random matrix $\boldsymbol{X}$ which is drawn from a density with expected values $\mu_t \equiv E(\boldsymbol{X}|T=t)$, (for $t=0,1$). Denote by n_T and n_C, respectively, the number of treated and control units in the original data. Let m_C denote the number of control units chosen *ex ante* to be remaining after matching from the n_C available control units, such that $m_C/n_C \leq 1$ (and as per Section 2.3, $m_T = n_T$.) Then:

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

$$E(\bar{\boldsymbol{X}}_{m_T} - \bar{\boldsymbol{X}}_{m_C}) = \gamma(\mu_1 - \mu_0), \tag{7}$$

where $\bar{\boldsymbol{X}}_{m_T} = \frac{1}{m_T}\sum_{i\in T\cap M}\boldsymbol{X}_i$ *and* $\bar{\boldsymbol{X}}_{m_C} = \frac{1}{m_C}\sum_{i\in C\cap M}\boldsymbol{X}_i$ *are random variables representing the sample means in the matched data set,* $M \subset (T \cup C)$ *is the subset of indexes of the matched treated and control units,* $\gamma \leq 1$ *is a scalar interpreted as the proportionate reduction in mean-imbalance, and* $\bar{\boldsymbol{X}}_{m_T}$*,* $\bar{\boldsymbol{X}}_{m_C}$*,* μ_0*, and* μ_1 *are* k*-dimensional vectors.*

A condition of EPBR is that the number of matched control units be fixed ex ante (Rubin, 1976*a*, 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 X is sampled randomly from its density, then (7) can be expressed as

$$E(\bar{\boldsymbol{X}}_{m_T} - \bar{\boldsymbol{X}}_{m_C}) = \gamma E(\bar{\boldsymbol{X}}_{n_T} - \bar{\boldsymbol{X}}_{n_C}) \tag{8}$$

where $\bar{\boldsymbol{X}}_{n_T} = \frac{1}{n_T}\sum_{i\in T}\boldsymbol{X}_j$ and $\bar{\boldsymbol{X}}_{n_C} = \frac{1}{n_C}\sum_{i\in C}\boldsymbol{X}_i$. The right side of (8) is the average mean-imbalance in the population that gives rise to the original data, and the left side is the average mean-imbalance in the population subsample of matched units. The EPBR property means that improving balance in the difference in means on one variable also improves it on *all* others and their *linear* combinations.

One limitation of EPBR is that it only controls the means of the covariates and says nothing about other moments, interactions, or nonlinear relationships. Another issue is that no method of matching meets EPBR in general. To address these issues, Rosenbaum and Rubin (1985*a*) consider special conditions where controlling the means enables one to control all expected differences between the multivariate treated and control population distributions. For this property to hold, we require the following additional conditions

(a) X is drawn randomly from a specified population $\boldsymbol{X}$,

(b) The population distribution for $\boldsymbol{X}$ 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 X.

Under these special 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. (Under some further conditions, Rubin and Thomas (1992) give the maximum level of imbalance reduction possible for an EPBR method. Although it is not known in general which EPBR method attains this bound, an estimate of the maximum value may provide useful guidance about how well the search is going.)

Rubin and Thomas (1996) give some simulated examples where certain violations of these conditions still yield the desired properties. The common violations of the condition that occur in practice are of course why the best researchers work (and why they have to work) so hard to try to improve balance on all variables, rather than checking just one. When observational data sets are not drawn randomly, condition (a) is violated. Only rare observational data sets in many fields are composed solely of continuous variables, and relatively few applications would entirely fit this particular class of densities, thus violating (b) (although see Rubin and Stuart (2006) on conditionally discriminant mixtures of proportional ellipsoidally symmetric distributions). Of course, researchers can ensure condition (c) by choosing an algorithm appropriately, such as based on nearest neighbor applications of the propensity score, Mahalanobis distance (without weights), or discriminant matching. Since these methods satisfy Definition 1 only in data where conditions (a) and (b) happen to hold, we describe the methods as *potentially EPBR*.

In practice, even if conditions (a)-(c) hold, EBPR compliant methods are more useful for applications that also satisfy two additional conditions:

(d) All covariates in X are equally important in their effect on the outcome Y; and

(e) Y is a linear function of X.

Without these additional conditions, we can reduce imbalance equally across all variables, but only with these conditions do we also know that this equal reduction in mean-imbalance translates into an equal reduction in bias in estimating the ultimate quantity of interest. The problem of course is that very few applications have equally important covariates and only a subset have linear functions.

4.2 Monotonic Imbalance Bounding Methods

We now introduce our alternative class of matching methods by generalizing and modifying EPBR in five steps. First, we drop EPBR's associated data conditions so that our class applies to all data types. Second, note that balancing only the expected value of the population distribution of the treated and control groups under (8), rather than the observed values, can lead to inefficient estimation (which explains the otherwise counterintutive result that matching on the estimated propensity score is more efficient than the true score; Hirano, Imbens and Ridder 2003). Thus, consider a slightly modified version of (8) where the random variables and expected values are replaced by their sample counterparts, and in addition (for later convenience) the equality is replaced by an inequality: $|\bar{X}_{m_T} - \bar{X}_{m_C}| \leq \gamma |\bar{X}_{n_T} - \bar{X}_{n_C}|$, which we write more simply as

$$|\bar{X}_{m_T} - \bar{X}_{m_C}| \leq \delta \tag{9}$$

where $\delta = \gamma |\bar{X}_{n_T} - \bar{X}_{n_C}|$. Equation (9) states that the *maximum imbalance* between treated and control units, as measured by the absolute difference in means, is bounded from above by some constant δ. Analogous to EPBR, one would usually prefer when the bound on imbalance is reduced due to matching, $\gamma = \delta / |\bar{X}_{n_T} - \bar{X}_{n_C}| < 1$, but dropping EPBR's associated conditions implies (for now) that this is not guaranteed.

Third, we generalize (9) to allow for any measure of imbalance, rather than merely the mean. Denote by $\mathcal{X}_{n_T}$ and $\mathcal{X}_{n_C}$ the subset of treated and control units in the original data and by $\mathcal{X}_{m_T}$ and $\mathcal{X}_{m_C}$ the subsets of treated and control units produced by the matching algorithm. Then, for clarity at this intermediate step, we define formally:

Definition 2 (Imbalance Bounding (IB))**.** *A matching method is Imbalance Bounding with respect to a function f and a distance D, or simply IB(f, D), if*

$$D\left(f\left(\mathcal{X}_{m_T}\right), f\left(\mathcal{X}_{m_C}\right)\right) \leq \delta \tag{10}$$

where δ is some constant.

Thus, EPBR is a version of IB where $D(x, y) = E(x-y)$, f is the sample mean of the marginal distribution of X_j (for $j = 1, \ldots, k$), $\delta = \gamma D\left(f\left(\mathcal{X}_{n_T}\right), f\left(\mathcal{X}_{n_C}\right)\right)$, the inequality replaces the equality, and $\gamma < 1$.

Although quite abstract, IB becomes natural when f and D 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, (10) becomes $|\bar{X}^{(j)}_{m_T,w} - \bar{X}^{(j)}_{m_C,w}| \leq \delta$, which is a bound on the imbalance as measured by I_1 of (3). The analogous result holds if $f_j(\cdot)$ is the sample variance, the k-th centered moment, the q-th quantile, etc.

- If $f_j(\cdot)$ is a univariate histogram of X_j, then consider the mean absolute difference of the frequencies between the distributions of treated and control units. This is equivalent to defining D as $\mathcal{L}_1^{(j)}$ in (4) one covariate at time, and so (10) represents a bound on the imbalance in the full one-dimensional distribution.

- If $D(x, y)$ is the average absolute difference over the strata and $f_j(\mathcal{X}) = \mathcal{X}_j$, with a slight abuse of notation, we obtain a bound on the measure of local imbalance $I_2^{(j)}$ from (5).

- Let $D(x, y) = |x|$ and $f(\cdot) = f_{jk}(\cdot)$ is the covariance of X_j and X_k and $\delta = \delta_{jk}$ we have $|\text{Cov}(X_j, X_k)| \leq \delta_{jk}$.

- The full global imbalance $\mathcal{L}_1$ for the k-dimensional distribution can also be obtained by taking D equal to (4) and $f(x) = x$.

A matching method can be IB for all, some of the above, or other different specifications of D and f. However, the bound δ in (10) is not always meaningful per se, because most matching methods bound some form of imbalance, so some comparison with the initial imbalance $D\left(f\left(\mathcal{X}_{n_T}\right), f\left(\mathcal{X}_{n_C}\right)\right)$ should be considered.

Fourth, IB methods may be of interest when $\delta / D\left(f\left(\mathcal{X}_{n_T}\right), f\left(\mathcal{X}_{n_C}\right)\right) \equiv \gamma < 1$, but IB does not require it to hold. To avoid this problem, we allow the maximum imbalance to be controlled *ex ante* and *monotonically* instead of being calculated after the match. When this is the case, one can fine choose or tune δ in order to guarantee that $\gamma < 1$. We now introduce this generalization.

Finally, consider the class of matching methods which produces subsets $\mathcal{X}^{\pi}_{m_T}$ and $\mathcal{X}^{\pi}_{m_C}$ on the basis of a given vector $\pi = (\pi_1, \pi_2, \ldots, \pi_k)$ of tuning parameters (such as ϵ in CEM or some sort of caliper), corresponding to the k covariates, such that $\pi_j > 0$ for $j = 1, \ldots, k$. As in Definition 2, let f be any function of the empirical distribution of covariate X_j of the data (such as the mean, variance, quantile, histogram, etc). Let π and π' be two k-dimensional vectors and let the notation $\pi' \prec \pi$ denote that the two vectors π and π' are equal on all components but one, which we denote j, for which $\pi'_j < \pi_j$ and analogously for $\pi' \succ \pi$.

Let $J = \{j_1, j_2, \ldots, j_{m_1}\}$ be a subset of $\{1, 2, \ldots, k\}$ covariates and $H = \{h_1, h_2, \ldots, h_{m_2}\}$ the complementary subset, with $m_1 + m_2 = k$, i.e.

$$J \cup H = \{1, \ldots, k\} \quad \text{and} \quad J \cap H = \emptyset\,. \tag{11}$$

Denote by $\gamma_J(\pi_J) = \gamma_{j_1 j_2, \ldots, j_m}(\pi_{j_1}, \pi_{j_2}, \pi_{j_{m_1}})$, $f_J(\mathcal{X}_J) = f_{j_1 j_2, \ldots, j_{m_1}}(X_{j_1}, X_{j_2}, \ldots, X_{j_{m_1}})$ and similarly for $\gamma_H(\pi_H)$ and $f_H(\mathcal{X}_H)$. Then we define:

Definition 3 (Monotonic Imbalance Bounding (MIB))**.** *A matching method is Monotonic Imbalance Bounding with respect to a vector function $f = (f_J, f_H)'$ and a distance D, or simply MIB(f, D), if for any J and H as in* (11) *there exists a monotonically increasing vector function $\gamma(\pi) = (\gamma_J(\pi_J), \gamma_H(\pi_H))'$ — i.e. if $\pi' \prec \pi$ then $\gamma(\pi') \leq \gamma(\pi)$ — such that*

$$\begin{aligned} D(f_J(\mathcal{X}^{\pi}_{J,m_T}), f_J(\mathcal{X}^{\pi}_{J,m_C})) &\leq \gamma_J(\pi_J) \\ D(f_H(\mathcal{X}^{\pi}_{H,m_T}), f_H(\mathcal{X}^{\pi}_{H,m_C})) &\leq \gamma_H(\pi_H) \end{aligned} \tag{12}$$

with $m_T = m_T(\pi)$ and $m_C = m_C(\pi)$.

(In CEM, $\pi = \epsilon$; in exact matching, $\pi = 0$.) Thus, the tuning parameter π of an MIB method bounds *ex ante* and *monotonically* the maximum imbalance in the difference of one or more features of the empirical distribution of treated and control units matched without altering the maximum imbalance in the complementary set of covariates and the number of treated and matched units is obtained *ex post* as a result of the match. An MIB method is an IB method, with the additional capabilities of controlling the inequality (10) ex ante, monotonically, and independently for a subset of covariates without hurting on the rest. The converse is not true: an EPBR method is not an MIB method because m_T and m_C are fixed ex-ante. A matching method can be MIB for all, some, or alternative specifications of D and f given above. (A special case of (12) involves separability in bounding selectively on each single variable, i.e. for $J = \{j\}$, $j = 1, \ldots, k$ and $H = \{1, 2, \ldots, j-1, j+1, \ldots, k\}$.)

4.3 Comparing EPBR and MIB

EPBR methods help with mean-imbalance in a linear context, whereas MIB methods help with all forms of linear and nonlinear multivariate imbalance. Methods can only be potentially EPBR given its associated data assumptions, whereas methods can be MIB, regardless of the data to which it is applied. When EPBR and its associated conditions hold, the number of matched units (m_T and m_C) are fixed ex ante and mean-imbalance is calculated only ex post, whereas under MIB all chosen forms of maximum imbalance are fixed ex ante and monotonically by a simple tuning parameter and the number of matched units are calculated ex post. (A class of methods could be constructed which enabled one to bound both imbalance and the number of matched units ex ante, but then many user choices would produce no results.) In EPBR, reducing imbalance on one variable reduces population imbalance on all other variables by the same amount; under MIB, one can reduce maximum in-sample imbalance on one variable without affecting the maximum multivariate in-sample imbalance on the others. (See also Rubin 1976*a*, p.112.)

Choosing what to balance on can be crucial. In some situations, matching only the mean is irrelevant, in which case EPBR methods can be hazardous. For example, Figure 1 portrays a covariate that is unimodal among treated units (see the solid line) but bimodal among control units (see the dashed line), but for which the mean of both groups is zero. For this covariate, mean-imbalance is irrelevant, since the distributions do not overlap in the region near the mean. Instead, this variable should be matched in the areas of common empirical support, indicated by the shaded areas. Applying an MIB method with these data enables one to increase the bound on mean-imbalance on this variable in order to match other more relevant features of the distributions, without hurting mean-imbalance or other types of imbalance on other variables.

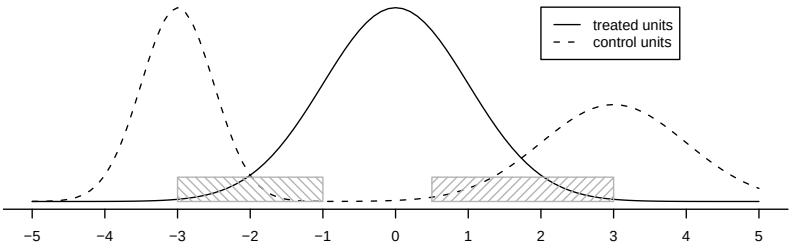

Figure 1: An example of a covariate for which minimizing mean-imbalance may be harmful. The example also shows that increasing mean-imbalance for this variable under MIB can be used to match more relevant features of the distributions (such as the shaded areas), without hurting mean-imbalance on other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

EPBR is defined for expected values of the covariates whereas MIB is defined for the observed values in each sample. One way to think about this difference is that potentially EPBR methods represent an attempt to approximate with observational data the classic *complete randomization* experimental design. In this design, observations are randomly selected, and each unit is randomly assigned a value of the treatment variable. In contrast, MIB methods like CEM attempt to approximate the *randomized block* experimental design, where values of the treatment variable are assigned within strata defined by the covariates. Randomized block designs have perfect balance in each data set on all observed covariates, whereas complete randomization designs are balanced only on average across experiments. Both are unbiased. Randomized block designs, as a result, are thus considerably more efficient, powerful, and robust than complete randomization designs (see Box, Hunger and Hunter 1978, p.103 and Imai, King and Stuart 2008); in an application by Imai, King and Nall (2008), complete randomization gives standard errors as much as six times larger than the corresponding randomized block design.[3]

Finally, we offer some examples. CEM is MIB, as we show in the next section. While matching exactly on the Mahalanobis distance or propensity score are potentially EPBR under specified distributional assumptions, they are not MIB. This can be seen because as EBPR methods they require the number of matched observations (m_T, m_C) to be fixed ex ante, while MIB requires that the number of matched observations be an outcome of the method rather than a tuning parameter. Nearest neighbor matching methods, including those based on Mahalanobis and propensity score metrics, are also not MIB, and these methods applied within 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 (12). Caliper matching as defined in Cochran and Rubin (1973) is not MIB because of the orthogonalization and overlapping regions; without orthogonalization, it is MIB if applied variable by variable (although applications of it typically violate the congruence principle; see Sections 5.1-5.2). For other MIB methods, see Section 6.3.

[3]The increased efficiency of MIB methods can be seen in CEM by its ability to match all aspects of the distribution of the treated and control units (greater than ϵ). That is, even if we somehow knew that in an observational study $E(\bar{\boldsymbol{X}}_{m_T} - \bar{\boldsymbol{X}}_{m_C}) = 0$, we could increase efficiency and reduce estimation error by continuing to improve the matching solution until the realized values were such that $\bar{X}_{m_T} - \bar{X}_{m_C} = 0$ and so that all other aspects of the empirical distributions of treateds and controls match as well as possible. For another example, suppose X is composed of 10,000 observations on 20 variables drawn jointly from independent normal densities. Since 20-dimensional space is enormous, odds are that no treated unit is anywhere near any control unit. Thus, some aspects of the empirical balance will almost surely be very poor, meaning that estimation error can be very large, even if the data generating process satisfies EPBR's conditions by being ellipsoidally symmetric. The only issue in trying to approximate a randomized block design with observational data is that observations may not be available for some strata, in which case the estimand may change or be inestimable.

4.4 CEM as an MIB Method

We prove here that CEM has the MIB property with respect to many important definitions of D and f from Section 4.2. To do this, we merely need to prove that CEM bounds important univariate and multivariate aspects of the difference between the the treated and control groups. For simplicity, but (unlike in EPBR) without loss of generality, suppose X is composed solely of continuous variables. Then, within each strata, the difference in means of the original uncoarsened variable X_j (for all j) is at most ϵ_j. Our results apply for one-to-one, j-to-k, and j_s-to-k_s matching for strata s. We discuss the one-to-one case here and the more complicated situations in Appendix A.

Setting ϵ in one-to-one CEM immediately implies a bound on the global difference in means for X_j between the treated and control groups. (That is, the strata-level difference in means, each of which is no longer than ϵ_j, averaged over strata is also bounded from above by ϵ_j.) Setting ϵ also bounds many other features of the global difference in distributions between the treated and control groups. We give evidence for this claim in two steps.

Local Imbalance Bounds For the difference in variance between the treated and control groups, denote the variance of z as $\bar{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z}_n)^2$, where $\bar{z}_n$ is the arithmetic mean of the z_i's. Then, we use the upper bound of the von Szökefalvi Nagy-Popoviciu inequality (Popoviciu, 1935): $(R^2/2n) \leq \bar{S}^2 \leq (R^2/4)$, where R is the range of z. For our problem, this result implies that the maximum variance between the treated and control units within each strata after application of CEM is $\epsilon_j^2/4$. Then it follows immediately that the difference in two variances, each bounded between 0 and $\epsilon_j^2/4$, is itself bounded from above by $\epsilon_j^2/4$.

The same approach applies to covariances: within each strata we know that $\bar{X}^j - X_{ij}$ is always bounded by ϵ_j, because the mean is internal, and the covariances are bounded in absolute value by the product of the coarsening levels: $|\mathrm{Cov}(X_j, X_k)| \leq [\bar{S}^2(X_j)\bar{S}^2(X_k)]^{1/2} \leq \frac{\epsilon_j \epsilon_k}{4}$, where $\bar{S}^2(X_j)$ and $\bar{S}^2(X_k)$ are the variances of variables X_j and X_k respectively. It is also easy to show that the k-th centered absolute moment, $E|X - \bar{X}|^k$, is bounded by ϵ^k and, as a result, so is the difference in all centered absolute moments between the treated and control groups. For the same reason, measures of joint variation like co-skewness, co-kurtosis and comoments are also bounded. And of course, the local imbalance I_2 in (5) is monotonically controlled by ϵ as well.

Global Imbalance Bounds We now use the results just derived to show how setting ϵ_j bounds the global differences. In CEM, θ_j is the number of strata, as in (6). We denote by w_i the weight for unit i where $w_i = (m_C/m_T)(m_T^s/m_C^s)$ if $i \in C_s$ and $w_i = 1$ if $i \in T_s$. In addition, m_C and m_T are the total numbers of controls and treated units matched and m_C^s and m_T^s are the number of control and treate) units matched in stratum s. The weighted mean for the control units is $\bar{X}_{m_C,j}^w = \frac{1}{m_C}\sum_{s=1}^{\theta_j}\sum_{i \in C_S} X_i w_i$ and similarly for the treated units.

Proposition 1. *CEM in MIB with respect to difference in means for each variable X_j, $j = 1, \ldots, k$.*

$$\left|\bar{X}_{m_T,j}^w - \bar{X}_{m_C,j}^w\right| \leq \epsilon_j$$

Proposition 2. *CEM is MIB with respect to the difference in the weighted centered moments for each variable X_j. Let $D(x, y) = |x - y|$,*

$$f_j(\mathcal{X}_{m_T}) = \frac{1}{m_T}\sum_{s=1}^{\theta_j}\sum_{i \in T_s}\left|X_{i,j}^T - \bar{X}_{m_T,j}^w\right|^k w_i$$

and

$$f_j(\mathcal{X}_{m_C}) = \frac{1}{m_C}\sum_{s=1}^{\theta_j}\sum_{i \in C_s}\left|X_{i,j} - \bar{X}_{m_C,j}^w\right|^k w_i\,.$$

Then,

$$D(f_j(\mathcal{X}_{m_T}), f_j(\mathcal{X}_{m_C})) \leq \epsilon_j^k(\theta_j^* + 1)^k, \quad j = 1, \ldots, k.$$

In the case of k_s-to-k_s matching, with k_s eventually varying in each strata, the bound is $\epsilon_j^k \left((\theta_j^* + 2)^k - \theta_j^{*k}\right)$.

(See the Appendix A for a proof.) In Proposition 2 the function $\gamma(\epsilon) = \epsilon_j^k((\theta_j^* + 2)^k - \theta_j^{*k})$ is monotonic with respect to the tuning parameter ϵ_j. Since ϵ_j is chosen ex ante and fixed, and θ_j^* is the maximal number of strata for variable X_j in a given data set, the inequality above establishes a bound on the difference in centered moment after the match, which decreases with ϵ_j.

A corollary of Proposition 2 is that, in the case of k_s-to-k_s matching, for a given sample and a fixed value of θ_j^*, a decreasing ϵ_j also decreases the bound on the difference of variances: $|S^2_{m_C,j} - S^2_{m_T,j}| \leq \epsilon_j^2(4 + 2\theta_j^*)$, for $j = 1, \ldots, k$.

A similar logic shows CEM is MIB with respect to the univariate quantiles:

Proposition 3. *Assume one-to-one matching. Let $D(x, y) = |x - y|$, $f(\mathcal{X}_{m_T}) = Q_{m_T,j}$, q^{th} denote the empirical quantile of the distribution of the treated units for covariate X_j, and similarly $f(\mathcal{X}_{m_C}) = Q_{m_C,j}$. Then, $D(f(\mathcal{X}_{m_T}), f(\mathcal{X}_{m_C})) = |Q_{m_T,j} - Q_{m_C,j}| \leq \epsilon_j$, for $j = 1, \ldots, k$. The same result holds for k_s-to-j_s matching with weights.*

(See the Appendix A for a proof.)

A final, but crucial, property of CEM is that the complete k-dimensional weighted histograms for the treated and control groups, with bins at subsume ϵ_j on each covariate X_j ($j = 1, \ldots, k$), are exactly equal. Similarly, so long as the automated method of computing histogram bin sizes in Section 2.7 uses bin sizes larger than ϵ, CEM will produce $\mathcal{L}_1(f, g) = 0$ (assuming the bin size is a multiple of ϵ; otherwise $\mathcal{L}_1$ will be approximately 0). This result shows that setting ϵ locally for each variable bounds all multivariate differences, for all levels of interaction, up to the chosen level.

5 Other Properties of Coarsened Exact Matching

The most important property of CEM is that it enables one to choose imbalance ex ante, on the scale of the variables one at a time, to be certain of the level of (global) balance you will get out at the end, and so that changes in balance on one variable do not affect maximum imbalance on others. Balance checking and uncertainty about what balance you will get is eliminated. You get what you want rather than getting what you get. Although of course fixing imbalance ex ante means that we learn the number of observations matched as a consequence, but bias is more crucial than variance in observational data analyses, and because matching can improve variance too by removing heterogeneity. We now discuss additional advantages of CEM, including a comparision to existing approaches.

5.1 Meeting the Congruence Principle

A crucial problem with many matching methods is that they operate on a metric different from the original data, and thus violate the *congruence principle*. This principle requires congruence between the data space and analysis space. Methods violating this principle lead to less robust inferences with suboptimal and highly counterintuitive properties (Mielke and Berry, 2007).

The violation of the congruence principle in propensity score and Mahalanobis distance matching methods is easy to see because both project the covariates from their natural k-dimensional space to a (different) one-dimensional quantity and match on that quantity: because different matching solutions can map into the same place on the one-dimensional projection, reducing imbalance on one variable will sometimes increase imbalance for others in unpredictable ways.

In contrast, CEM meets the congruence principle by operating in the space where X was created and its variables were measured, and regardless of whether the data are continuous, discrete,

or mixed. This is the space most understood by data producers and analysts and so the technique should also be easier to understand as well. Examples of other matching methods that meet the congruence principle include Iacus and Porro (2007, 2008).

5.2 Comparisons with Other Methods

Whereas CEM uses simple, fixed, non-overlapping intervals of local indifference, defined ex ante based on the metric of each variable one at a time, nearest neighbor caliper matching Cochran and Rubin (1973) uses orthogonalization and a more complicated geometry of n_T overlapping hyper-parallelepipeds centered around each treated data point. The result is not MIB and does not meet the congruence principle. If we modify the caliper approach by applying it to each variable separately without orthogonalization, it is MIB. For truly continuous variables, it also meets the congruence principle. However, a large fraction of variables used in the social sciences are discrete or mixed in complicated ways, in which case calipers (used separately or with other methods) violate the congruence principle. For example, CEM can make a variable like "years of education" respect important milestones, like high school, college, and post-graduate degrees, by appropriate coarsening into these categories. In contrast, caliper matching uses a different grouping for each treated unit (e.g., ± 5 years) that would inappropriately combine some units that span across these logical category boundaries, such as by matching a college dropout with a first year graduate student. For another example, the difference in income between Bill Gates and Warren Buffett is enormous in any one year; with CEM, we could group them together, whereas a caliper for income would likely leave them unmatched. Similar issues exist for lower levels of income (with different tax rate thresholds), age (at or near birth, puberty, legality, retirement, etc.), temperature (phase transitions), and numerous other variables.

CEM is related to a large number subclassification (or "stratification") approaches, such as full matching, frequency matching, subclassification on the propensity score, and others. These approaches are not MIB. By having the ability to set ϵ_j differently for each variable, CEM is also similar in spirit, although not methods, to various creative combinations of approaches, such as Rosenbaum, Ross and Silber (2007). The core of the algorithm in CEM was first studied formally in Cochran (1968), although we use it in different ways — such as by setting ϵ_j to substantively meaningful values related to the metric of each variable rather than a minimal and arbitrary number, using all available variables rather than only the major confounders, proving many different properties, assuming finite rather than infinite samples, and introducing a range of practical extensions.

Although CEM works by setting balance as desired and getting the number of matched units as a result, and most other methods work in reverse, obtaining similar results with different methods will often be possible when the specialized conditions required by previous methods hold (see Section 4.1). Under these conditions, however, CEM is still considerably easier to use and understand and faster in computational and human time. When these conditions do not at least approximately hold, CEM will usually be superior since balance will be guaranteed on all higher order moments and interactions on all variables, something not addressed by methods that are potentially EPBR unless their specialized data restrictions hold.

To illustrate, suppose we run optimal or nearest neighbor matching on the Mahalanobis or propensity score distance with a fixed number of matched control units, m_C. The result would be some level of average imbalance for each variable. If we use this imbalance to define ϵ_j and apply CEM, we would usually obtain a similar number for m_C as set ex ante. Similarly, consider a method and data that meet EPBR and its associated data requirements, and run it given some fixed number of control units m_C. Assume the maximum imbalance can be computed explicitly (Rubin, 1976*a*, Equation 2.2), and define γ as one minus this maximum imbalance. In most situations, we would expect that running CEM would produce a similar number of control units as fixed ex ante

by the EPBR method.

5.3 Automatic Restriction to Common Empirical Support

As described in Section 2.4, all existing approximate matching procedures require a separate step prior to matching, where the data are restricted to the region of common empirical support of the treated and control units. This eliminates the region where extrapolations beyond the limits of the data would be needed. In contrast, users of CEM require no separate step. All observations within a coarsened stratum for which we have both a treated and control unit by definition do not involve extrapolating beyond the data and so the observation will be included; otherwise, it will be removed. The process is easy, automatic, and no extra steps are required. Since applied researchers seem to remove extrapolation regions as infrequently as their scant efforts to check balance, CEM may enhance compliance with proper data analysis procedures; alternatively, CEM can also be used as a simple way to restrict data to common support to improve other matching methods.

5.4 Approximate Invariance to Measurement Error

Suppose T is ignorable conditional on unobserved pretreatment covariates X^*, and so we match instead on X, where $X_j = X_j^* + \eta_j$ given a vector of measurement errors η_j for each covariate j. Commonly used matching methods are directly affected by the degree of measurement error, even when other conditions they may impose hold, and even if $E(\eta_j) = 0$. In particular, balance with respect to X does not imply balance with respect to X^*; the true propensity score based on X is not a balancing score for X^*; and adjusting based on X instead of X^* will lead to biased estimates of the treatment effect (Battistin and Chesher, 2004).

Under CEM, if measurement error is less than ϵ_j, $\epsilon_j \geq \max(|\eta_j|)$, and it happens to respect the resulting strata boundaries, then CEM will produce the same preprocessed data set whether matching on X or on X^* and so is invariant to measurement error. If only the first condition holds, the second condition will hold for many observations under many conditions and so CEM will normally be approximately invariant to measurement error, even if not invariant.

We study sensitivity measurement error (in the sense of Battistin and Chesher 2004) via a real data set described in Section 7.1 and randomly perturb the earnings variable by adding with gaussian error $N(\mu = 1000, \sigma^2 = 1000^2)$ and replacing perturbed negative earnings with zero. We run 5,000 simulations and at each replication match before and after perturbation. Denote by m_T and m_C the number of matched units before perturbation, and m'_T and m'_C the number after perturbation. Then define K_T and K_C as the number of treated and control units present in the both subsets of matched units before and after the perturbation. To measure the sensitivity to perturbation, we calculate the percentages $K_T/\min(m_T, m'_T) \cdot 100\%$ and $K_C/\min(m_C, m'_C) \cdot 100\%$. For all methods but CEM, $m_T = m'_T$ while for all matching algorithms $m_C \neq m'_C$ in general. Table 1 shows that CEM is considerably closer to invariant (i.e., less sensitive) to measurement error. Mahalanobis matching (MAH) and genetic matching (GEN) preserve 80% of the total matched subset and propensity score matching (PSC) around 70%. In contrast, CEM preserves 95% of the treated units and 98% of the control units. Thus, to some extent, coarsening can overcome measurement error problems, at least for the (preprocessing) matching stage.

5.5 Bounding the Average Treatment Effect Estimation Error

We first introduce a slight constraint on the possible range of functions $g_0(\cdot)$ and then derive the theoretical bound. The following assumption restricts the sensitivity of $g_0(x_1, \ldots, x_k)$ to changes in its arguments: along each direction (i.e. along each x_j), g_0 behaves like a Lipschitz function. Following the notation of Section 3.2, denote by $\Xi_{-j} = \Xi_1 \times \Xi_2 \times \cdots \Xi_{j-1} \times \Xi_{j+1} \times \cdots \times \Xi_k$, $x_{-j} = (x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k)$ and $g_0(x_j|x_{-j}) = g_0(x_1, x_2, \ldots, x_k)$.

	CEM(K_T)	CEM(K_C)	PSC(K_C)	MAH(K_C)	GEN(K_C)
% Common Units	95.3	97.7	70.2	80.9	80.0
Seconds	0.07	0.07	0.08	0.15	126.64

Table 1: Percentage of units present in matched sets both before and after perturbation, averaged over 5,000 simulations, and computational time. (For all methods but CEM, $K_T = 100\%$.)

Assumption 1 (Lipschitz behaviour)**.** *For each variable j ($j = 1, \ldots, k$) there exists a constant L_j, $0 < L_j < \infty$, such that, for any values $x_j' \neq x_j''$ of x_j, $\max_{x_{-j} \in \Xi_{-j}} |g_0(x_j'|x_{-j}) - g_0(x_j''|x_{-j})| \leq L_j d_j(x_j', x_j'')$ where $d_j(\cdot, \cdot)$ is an appropriate distance for variable x_j.*

This assumption is very mild and only bounds g_0 from taking infinite values on finite sets. Given two values x_j' and x_j'' of the variable x_j, the maximum excursion of g_0, regardless of all possible values of the remaining variables x_i ($i \neq j$), is bounded by the distance between x_j' and x_j'' times some finite constant. This means that given finite variation in one variable, the function g_0 does not explode. If this assumption does not hold, g_0 could have strange properties, such that even arbitrarily small and otherwise irrelevant imbalance in the covariates could produce arbitrarily large estimation error in SATT. This assumption easily fits all functional forms used regularly in the social sciences.

Without loss of generality, we measure distance for numerical covariates as $d_j(x, y) = |x - y|$. For categorical variables we adopt the following definitions for convenience, and without loss of generality. Let X_j be categorical variable and H be the set of distinct values of X_j. Then, if $H \subset \mathcal{U}$, where $\mathcal{U}$ is an abstract set of unordered categories, define the distance as $d(x, y) = \mathbf{1}_{\{x \neq y\}}$, where $\mathbf{1}_A = 1$ for elements in set A and zero otherwise. If, alternatively, $H \subset \mathcal{O}$, where $\mathcal{O}$ is the abstract set of ordered categories, the distance is $d(x, y) = |\mathrm{rank}(x) - \mathrm{rank}(y)|$, where $\mathrm{rank}(x)$ is the rank/order of category x in $\mathcal{H}$.

Then, the definitions in Section 2.6 imply directly that the estimation error, $\bar{\mathcal{E}}_0 \equiv \mathrm{SATT} - \widehat{\mathrm{SATT}}$, is bounded from above and below by $|\bar{\mathcal{E}}_0|$, i.e., $-|\bar{\mathcal{E}}_0| \leq \mathrm{SATT} - \widehat{\mathrm{SATT}} \leq |\bar{\mathcal{E}}_0|$ and a consequence of Assumption 1 is that $|g_0(X_i) - g_0(\tilde{X}_i)| \leq \max_{j=1,\ldots,k} L_j \epsilon_j$. Therefore, for the CEM algorithm, which keeps matched treated and control units for each covariate a maximum of ϵ_j apart, we conclude that

$$|\bar{\mathcal{E}}_0| \leq \max_{j=1,\ldots,k} L_j \epsilon_j. \tag{13}$$

Thus, setting ϵ_j locally for each variable bounds the SATT estimation error, not merely the imbalance between treated and control groups. (We discuss how to estimate this in Section 6.6.2.)

5.6 Bounding Model Dependence

A key advantage of matching done well is that it should reduce model dependence. However, aside from exact matching the relationship has never been proven directly. Thus, we prove here that the maximum degree of model dependence is controlled by setting ϵ in CEM.

When exact matching is feasible, we estimate $Y_i(0) \equiv g_0(X_i)$ via observed values of Y_ℓ for which $\tilde{X}_\ell = X_i$. When its infeasible, we resort to using a parametric or nonparametric statistical model m_ℓ to span the remaining imbalance by estimating $Y_i(0) \equiv g_0(X_i)$ via model extrapolation or interpolation, conditional on the matched data set as $\hat{Y}(0) \equiv m_\ell(\tilde{X})$. Model dependence is how much $m_\ell(\tilde{X})$ varies as a function of the model m_ℓ. Without loss of generality, consider the case where X is a one dimensional numerical covariate. We restrict the attention to the set of competing Lipschitz models, as an analogy to Assumption (1), such that

Definition 4 (Competing models)**.** *Let m_ℓ, $\ell = 1, 2, \ldots$, be models estimated on the matched data*

$\tilde{X}$ and consider the following class

$$\mathcal{M}_h = \left\{ m_\ell : |m_\ell(x) - m_\ell(y)| \leq K_\ell |x - y|, \quad \textit{such that} \quad |m_i(\tilde{X}) - m_k(\tilde{X})| \leq h, i \neq k \right\}$$

with exogenous choices of a small prescribed nonnegative value for h and $0 < K_\ell < \infty$.

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 $\tilde{X}$; 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 $|m_1(\tilde{X}_i) - m_2(\tilde{X}_i)|$ (King and Zeng, 2007). This leads to our key result:

$$\begin{aligned} |m_1(\tilde{X}) - m_2(\tilde{X})| =& |m_1(\tilde{X}) \pm m_1(X) \pm m_2(X) - m_2(\tilde{X})| \\ \leq& |m_1(X) - m_1(\tilde{X})| + |m_2(X) - m_2(\tilde{X})| + |m_1(X) - m_2(X)| \\ \leq& (K_1 + K_2)|X - \tilde{X}| + h \leq (K_1 + K_2)\epsilon + h \end{aligned}$$

Thus, the degree of model dependence is directly bounded by the choice of ϵ in CEM.

5.7 Computational Efficiency

The most important computational efficiency of CEM is setting rather than checking balance. Even when multiple steps are required to ensure sufficient observations, the number of steps will normally be very few in comparison, and assessing the number of observations is trivial compared to ex post balance checks. In addition, each run is fast because it scales linearly in the number of variables, and is about as complex as a simple tabulation procedure.

Consider the most computationally difficult case of continuous covariates. Assume a given ϵ vector, which coarsens each variable X_j into $\theta_j \leq n$ intervals labeled with integers $(1, \ldots, \theta_j)$. This operation is of computational order n and so for k variables is of order kn. More specifically, coarsening produces a matrix of integers $G_\theta(X)$ in kn steps. Each row of $G_\theta(X)$ is collapsed into a character string, $S(G_\theta(X))$[4], which requires n additional operations. Finally, we tabule $S(G_\theta(X))$ in n additional steps. The total number of computations in CEM is thus $(k + 2)n$.

In contrast, Mahalanobis matching requires the inversion of a $k \times k$ matrix, with computational order k^3, and propensity score matching requires this inversion for each iteration of fitting the logistic or other regression. This is just for the matrix inversion step, without regard to the other matrix computations to obtain the final distance, which are of order n. For a large k and moderate n (e.g., $k = 35$ and $n = 1000$), merely the construction of the distance matrix is of computational order comparable to the whole CEM algorithm. In addition, once the distance matrix is available, the nearest neighbor algorithm must be applied to determine matches, which requires additional computational steps. Other methods require even heavier burdens.

Similarly, simple implementations of common methods require the storage of the entire $k \times k$ matrix, and a final $n \times n$ distance matrix, or its triangular version of length $n(n-1)/2$. In contrast, CEM requires an $n \times k$ data matrix of integers and a subsequent vector of strings of length n. Even for large data sets, these objects may be accessed sequentially on disk in CEM whereas with Mahalanobis or propensity score matching, continuous non-sequential access to the distance matrix must be maintained, most likely in RAM. These computational efficiency and

[4]For example, we transform the row $(1, 3, 6, 15, 1)$ into "1*3*6*15*1". This procedure may be further optimized by replacing the strings with fixed size binary allocations in 64 bit implementations, but we do not pursue this additional efficiency here.

memory requirements for distance based methods are approximately the same for caliper-based methods.

An example, with CEM programmed in R and the key parts of other methods programmed in C is given in Table 1. In larger applications — such as microarray analyses with k in the thousands and n in the dozens, or unstructured text analyses with k also in the thousands, but n ranging from the thousands to the millions — CEM should work well even though most prior approaches seem infeasible or at least not amenable to automation.

6 Extensions of Coarsened Exact Matching

6.1 Shifted Coarsenings

One seeming inconsistency with the basic CEM algorithm described in Section 3 is that it can be sensitive to changes in X smaller than ϵ near stratum boundaries even though it is insensitive to changes in X within strata. This does not matter if ϵ is set based on substantive criteria, but can be a concern if set without as much thought. In this situation, all the properties of CEM described in Section 5 still hold, but there may be an opportunity to increase the matched sample size a bit more, given the same chosen balance level, even without relaxing any assumptions.

Thus our software runs the basic CEM algorithm several times, each with a fixed value of ϵ, and thus a fixed stratum size, but with values of the cutpoints shifted together by different amounts. We then output the single coarsening solution that maximizes the remaining sample size. The number of shifted coarsenings and the size of each may be chosen by the user, but our default is to try only three since we find that the advantages of this procedure are small and additional improvements beyond this are not worth the computational time. Whichever choice the user makes, all the properties of the basic CEM method also apply to this slightly generalized algorithm.

6.2 Matching and Missing Data

When it comes to estimating causal effects in data with missing values, divergent messages are putting applied researchers in a difficult position. One message from methodologists writing on causal inference in observational data is that matching should be used to preprocess data prior to modeling. Another message is that missing data should not be listwise deleted, but should instead be treated via multiple imputation or another proper statistical approach (Rubin, 1987; King et al., 2001). Although most causal inference problems have some missing data, its not obvious how to apply matching while properly dealing with missing data. Indeed, we know of no matching software that allows missing data for anything other than listwise deletion prior to matching, and no missing data software that conducts or allows for matching. We offer two options here enabled by CEM.

The simplest approach is to treat missing values as a discrete "observed" value, and then to apply CEM with other coarsening used for the non-missing values. The default operation of our software uses this approach. In some situations, however, we might wish to customize this approach to the substance of the problem by coarsening the missing value with a specific observed value. For example, for survey questions on topics respondents may not be fully familiar with, the answers "no opinion" and "neutral" may convey similar or in some cases identical information, and so grouping for the purpose of matching may be a reasonable approach. Since the original values of these variables would still be passed to the analysis model, special procedures could still be utilized to distinguish between the effects of the two distinct answers.

Although this first approach to missing data and matching will work for many applications, it will be less useful when the occurrence of missing values are to some extent predictable from the observed values of other variables in complicated ways we do not necessarily foresee and include in our customized coarsening operator. Indeed, this is precisely what the "missing at

random" assumption common in multiple imputation models is designed for. Thus, an alternative is feed multiply imputed imputed data into a modified CEM algorithm. The modification works by first placing each missing value in whichever coarsened stratum a plurality of the individual imputations falls. (Alternatively, at some expense in terms of complication, the imputations could stay in separate strata and weights could be added.) Then the rest of the algorithm works as usual. The key here is that all the original uncoarsened variable values fed into CEM — in this case including the *multiple* uncoarsened imputed values for each missing value — are output from CEM as separately imputed matched data sets. Then, as usual with multiple imputation, each imputed matched data set is analyzed separately and the results combined. Thus, unlike with other matching procedures combined with imputation, multiple imputation followed by this modified CEM algorithm will produce proper uncertainty estimates.

6.3 Combining CEM with Other Methods

CEM is the simplest method with MIB properties (and those in Section 5) and so may have the widest applicability, but other improved methods could easily be developed for specific applications by applying existing approaches within each CEM stratum. For example, instead of retaining all units matched within each stratum and moving to the analysis stage, we could fine tune local (i.e., sub-ϵ) imbalance further by selecting or weighting units within each stratum via distance or other methods. Indeed, non-MIB methods can usually be made MIB if they operate within CEM strata, so long as the coarsened strata take precedence in determining matches. Thus, full and optimal matching are not MIB, but if applied within CEM strata would be MIB and would inherit the properties given in Section 5. Genetic matching as defined in Diamond and Sekhon (2005) is not MIB, but by choosing the variable-by-variable caliper option in GenMatch (Sekhon, 2008) it would be MIB, and when operating within CEM strata (as GenMatch now implements) it would be MIB and would also meet the congruence principle. Similarly, one could run the basic CEM algorithm and then use either a synthetic matching approach (Abadie and Gardeazabal, 2003), nonparametric adjustment (Abadie and Imbens, 2007), or weighted cross-validation (Galdo, Smith and Black, 2008) within each stratum.

If the user does not know enough about X's measurement to coarsen, then productive data analysis seems infeasible. But in some applications, we can partition X into two sets, only the first of which includes variables known to have an important effect on the outcome (such as in public health, age, sex, and a few diagnostic indicators). In this case, we may be willing to take good matches on any *subset* of the second set and to forgo the MIB property within this second set. To do this, we merely set ϵ artificially high for this second set, but small as usual for the first set, and then apply a non-MIB method within CEM strata. For example, because the relative importance of the variables is unknown, the propensity score other distance metric, if correctly specified, could be helpful. When the correct specification is unlikely, one can alternatively leave the remaining adjustment to the analysis stage, where analysts have more experience assessing model fit.

6.4 Multicategory Treatments

Under CEM, we set ϵ and then match the coarsened data, all without regard to the values of the treatment variable. This means that CEM works without modification for multicategory treatments: after the algorithm is applied, keep every stratum that contains all desired values of the treatment variable and discard the rest. This is a simple approach that can be easily used with or in place of more complicated approaches, such as based on generalizations of the propensity score (Imai and van Dyk, 2004; Lu et al., 2001; Imbens, 2000).

6.5 Blocking in Randomized Experiments

Since "blocking" (i.e., pre-randomization matching) in randomized experiments bests complete randomization with respect to bias, efficiency, power, and robustness, it should be used whenever

feasible (Imai, King and Nall, 2008; Imai, King and Stuart, 2008). Fortunately, CEM also works for blocking without modification: After matching the coarsened pre-treatment covariates X via CEM, create the treatment variable by randomly assigning one (or more) of the units within each stratum to receive treatment; the others are assigned to be control units. CEM also works with multicategory treatments in blocking by randomly assigning observations within each stratum each of the values of the treatment variable. Strata without sufficient observations to receive at least one possible value of each treatment and control condition are discarded.

6.6 Automating User Choices

As described in Section 3, we recommend that users of CEM choose ϵ based on their knowledge of the covariate measurement process and other substantive criteria such as the likely importance of different variables. Although we have shown that making these decisions is relatively easy and intuitive in most situations, users may sometimes want an automated procedure to orient them or to make fast calculations. We offer several such approaches here.

6.6.1 Histogram Bin Size Calculations

When automation is necessary because of the scale of the problem, or to provide some orientation as a starting point, we note here that choosing ϵ is very similar to the choice of the bin size in drawing histograms. Some classic measures of bin size are based on the range of the data, an underlying normal distribution, or the inter-quartile range. These are, respectively, known as Sturges, $\Delta_{\text{st}} = (x_{(n)} - x_{(1)})/(\log_2 n + 1)$, Scott, $\Delta_{\text{sc}} = 3.5\sqrt{\bar{s}_n^2}n^{-1/3}$ (Scott, 1992), and Freedman and Diaconis (1981) $\Delta_{\text{fd}} = 2(Q_3 - Q_1)n^{-1/3}$. More recently, Shimazaki and Shinomoto (2007) developed an approach based on Poisson sampling in time series analysis (in the attempt to recover spikes), which we find works well. Our software implements these approaches but also provides a way to specify non-constant bins for each variable, in which case the corresponding ϵ for our proofs is the maximal bin size.

6.6.2 Estimating the SATT Error Bound

Assumption 1 is a natural part of standard observational data analysis, but it gives no hint how big or small the L_j's are. In practice, they can take any finite value, but their ranking implies a rough order on the importance of each variable in affecting g_0. That means that some insight about the size of ϵ_j in CEM (or π_j in any MIB method) and its effect on the treatment effect may come from information about L_j. Thus, we note that L_j, for variable j ($j = 1, \ldots, k$), may be estimated from the data as:

$$\hat{L}_j = \max_{i_1 \neq i_2 \in C} \frac{|Y_{i_1}(0) - Y_{i_2}(0)|}{d_j(X_{i_1 j}, X_{i_2 j})}, \tag{14}$$

where $C = \{i : 1 \leq i \leq n \cap T_i = 0\}$. These $\hat{L}_j$ are estimates from below of the true L_j's, but they may still give insights about the relative importance of each variable on g_0 for the given data. Under additional assumptions on g_0, the estimators of the L_j may have better performance (e.g. g_0 is linear or well approximated by a Taylor expansion, etc.). Equation (13) is independent of the number of matched treated units m_T when L_j are known, but in general the L_j are not independent and can be estimated via (14). In such a case, the bound naturally depends on m_T. Thus, although knowing that CEM bounds SATT error is an attractive property in and of itself, we can go further and estimate the value of this bound with $\hat{\mathcal{E}}_0$ given as $\hat{\mathcal{E}}_0 = \max_j \hat{L}_j \epsilon_j$ and use the terms $\hat{L}_j \epsilon_j$ as a hints during matching about which covariate may give rise to the largest estimation errors or bias in estimating SATT. Although (14) uses the outcome variable, it only does so for control units (as in Hansen, 2008), and so inducing selection bias is not a risk.

6.6.3 Progressive Coarsening

Under CEM, setting balance by choosing ϵ may yield too few observations in some applications. Of course, this situation reveals a feature of the data, not a problem with the method, where the only real solution is to collect more data. In some circumstances, however, this situation may cause users to rethink their choices for ϵ and rerun CEM. Although we prefer users to make these choices explicitly, we offer here an automated procedure that may help in understanding data problems, identify the new types of data that would be most valuable to collect, or help them rethink their choices about ϵ.

Thus, we now study systematic ways to *relax* a CEM solution (that is increase ϵ_j selectively) by using θ' such that $\theta' \prec \theta$ (using notation from Section 4.2). When different relaxations or coarsenings, say θ' and θ'', lead to the same total numbers of matched units, $m_T(\theta') + m_C(\theta') = m_T(\theta'') + m_C(\theta'')$, then an automated procedure needs a way to choose among these solutions that are for our purposes equivalent. We discriminate among these by minimizing the $\mathcal{L}_1$ distance. Furthermore, although setting $\theta_j = 1$ is equivalent to dropping X_j from the match, we keep X_j with $\theta_j = 1$ to maintain comparability because the $\mathcal{L}_1$ distance depends on the number of covariates (as with any measure of dissimilarity in multidimensional histograms). In addition to keeping the number of covariates the same in this way, we also keep the bins of the multidimensional histogram used to calculate $\mathcal{L}_1$ the same.

With these requirements, we adopt a heuristic algorithm which we first describe conceptually, without regard to computer time, and then what we use in practice. Given the original user choice of θ, the algorithm relaxes each θ_j in increments of two, that is $\theta'_j = \theta_j - 2$, until $\theta'_j < 10$ and then by one or up to a user chosen minimally tolerable number of intervals, $\theta_j^{\min}$. (We also shift each intermediate solution as in Section 6.1.) We then repeat the procedure for pairs of variables, (θ_i, θ_j), triplets $(\theta_i, \theta_j, \theta_k)$, etc. Combined with shifted coarsenings, an exhaustive procedure with greater than triplets is feasible only via parallel processing, which happens to be is easy to implement with CEM. In practice, however, there no need to explore all these combinations of different coarsenings because even the basic application of CEM clearly reveals which data are well matched overall and also with respect to how the treated and control units differ in the multidimensional distribution. When we use this algorithm, we usually relax only one or two variables at a time. We then also use the MIB property of CEM to provide convenient graphical summaries of the results (see Section 7.1).

6.7 Avoiding the Dangers of Extreme Counterfactuals

In making causal inferences, the best current research practice is to eliminate extreme model dependence by discarding observations outside the region of common empirical support (see Section 5.3). Avoiding extreme model dependence is also an issue that applies to any type of counterfactual inference — including causal inferences, forecasts, and what if questions. Typically, scholars do this by eliminating data in the region requiring extrapolation, outside the convex hull of the data (King and Zeng, 2006). However, as is widely recognized, the hull may contain voids with little data nearby where estimation would be model dependent. Similarly, regions may exist just outside the hull, but near a lot of data just inside, for which a small extrapolation may be safe.

CEM can help avoid these problems as follows. First augment the covariate data set with a pseudo-observation that represents the values of X for the counterfactual inference of interest and then run CEM on the augmented data set. Observations that fall in the same stratum as the pseudo-observation can be used to make a relatively model-free inference about this counterfactual point, and so the number of such observations is a measure of the reliability of an inference about this counterfactual. This is thus a small generalization due to coarsening of a point emphasized by Manski (1995), who would use $\epsilon = 0$.

It may also be worth repeating this procedure after widening the definition of ϵ to include the

largest values you would be willing to extrapolate for your particular choice of dependent variable. For example, log-mortality for most causes of death is known to vary relatively smoothly with age (Girosi and King, 2008), and so extrapolating age by 10 or 20 years would normally not be very model dependent, except for the very young or very old. Thus, we might set ϵ_{age} in this way, even though it might normally be set much smaller for using the basic CEM algorithm where the goal would be to eliminate as much dependence on these types of assumptions as possible. This additional procedure is of course more hazardous because it involves assumptions about a specific outcome variable and because of interactions. For example, even if extrapolating age by 10 years is reasonable in one application, and extrapolating education by 4 years is also reasonable, evaluating a counterfactual that involved simultaneously extrapolating 10 years of age *and* 4 years of education beyond the data might well be unreasonable. Examples like these are much less likely to occur or matter if ϵ is defined as we do for CEM.

7 Coarsened Exact Matching in Practice

Although the main advantage of CEM compared to other approaches may be the way it closely connects the substance of each variable with the ultimate match, we compare the methods on other grounds here. We start in Section 7.1 by showing how CEM reduces imbalance and illustrating its MIB property via our progressive coarsening algorithm. We then show how CEM compares on imbalance, bias, root mean square error, and computational time in data that fits (in Section 7.2) and does not fit (in Section 7.3) EPBR's associated data conditions.

Results here are meant to show that CEM performs very well with minimal effort, using only our automated algorithms, and without optimizing CEM based on the substance of the variables (as we recommend for practice). Even though we show that CEM substantially outperforms other methods, it would be easy to outperform these results using CEM or the combined methods discussed in Section 6.3. The usual "ping pong theorem" qualifications certainly apply.

7.1 Empirical Evidence on Achieving Balance

Data Our data are from the National Supported Work (NSW) Demonstration, a U.S. job training program Lalonde (1986). Although a unique experimental target result is not easily defined due to the apparent failure of random treatment assignment, we use these data here to assess the degree to which CEM can balance the data relative to other methods. The program provided training to the participants for 12-18 months and helped them in finding a job. The goal of the program was to increase participants' earnings, and so 1978 earnings (`re78`) is the key outcome variable. Pretreatment variables were measured for both participants and controls, including age (`age`), years of education (`education`), marital status (`married`), lack of a high school diploma (`nodegree`), race (`black`, `hispanic`), indicator variables for unemployment in 1974 (`u74`) and 1975 (`u75`), and real earnings in 1974 (`re74`) and 1975 (`re75`). Some of these are dichotomous (`married`, `nodegree`, `black`, `hispanic`, `u74`, `u75`), some are categorical (`age` and `education`), and the earnings variables are continuous and highly skewed, with point masses at zero. The conditions for EPBR to hold are violated.

Basic Analysis Table 2 reports several measures of the degree of imbalance in the original unmatched data, including the difference in means I_1, the variable-by-variable distributional difference $\mathcal{L}_1^j$, and differences for the 25th, 50th, and 75th percentiles. The overall imbalance between the treated and control groups in the original data is the distance between multidimensional histograms, $\mathcal{L}_1 = 1.149$ (We discretize `re74`, `re75`, `age` and `education` according to intervals of size 5000, 5000, 5 and 1, respectively). Table entries that are exactly zero are left blank. An exact match returns 74 controls against 55 treated units.

We then match with CEM by defining ϵ for the variables `re74`, `re75`, `age` and `education` via Sturges formula which returns, respectively, the values $\epsilon_{re74} = 3957.1$, $\epsilon_{re75} = 3743.2$,

	$\mathcal{L}_1$	I_1	25%	50%	75%
age	0.01	0.18	1.00		1.00
education	0.20	0.19		1.00	1.00
black	0.00	0.00			
married	0.02	0.01			
nodegree	0.17	0.08	1.00		
re74		101.49		69.73	584.92
re75		39.42		294.18	660.69
hispanic	0.04	0.02			
u74	0.04	0.02			
u75	0.09	0.05			

Table 2: Measures of absolute imbalance in the distributions of treated and control units in the original data. The global imbalance is $\mathcal{L}_1 = 1.149$, with $n_T = 297$, $n_C = 425$. An entry of "0.00" indicates that at least the third decimal digit is nonzero; a blank entry denotes the number is exactly zero.

	$\mathcal{L}_1$	I_1	25%	50%	75%	I_2
age	-100.00	0.02	-2.63			0.03
education	-89.58	-1.40		-7.69	-7.69	0.00
black	-100.00	-0.13				
married	-100.00	-1.07				
nodegree	-100.00	-8.35	-100.00			
re74		-0.24		-0.18	-1.30	0.01
re75		-0.07		-0.16	-1.39	0.01
hispanic	-100.00	-1.87				
u74	-100.00	-2.01				
u75	-100.00	-4.51				

Table 3: Imbalance reduction due to CEM compared to the original data (in Table 2), as a percent of the range of each variable. Positive values for $\mathcal{L}_1$, I_1, and the three quantiles means imbalance increased (but by an amount less than the bound). The local imbalance, I_2, is also reported. Number of matched units: $m_T = 163$, $m_C = 222$. Global imbalance: $\mathcal{L}_1 = 0.34$.

$\epsilon_{age} = 3.8$ and $\epsilon_{edu} = 1.3$. The result gives 163 treated units matched with 222 control units, with an overall imbalance of $\mathcal{L}_1 = 0.62$. A summary of the percentage imbalance reduction due to CEM, as compared to the absolute imbalance in the original data (reported in Table 2), is given in Table 3.[5]

Although CEM only guarantees imbalance in the matched sample to be less than or equal to the bound set by the chosen ϵ, actual imbalance can be a good deal smaller for any one variable because of the effects on this variable on bounding other variables or because the observations within the chosen strata happen to contain better matches. We can see this if we rescale the ϵ's to the length of the support of each variable. For example, for variable `age` we have: $\epsilon_{age} = 3.8$ and $|R_{age}| = 38$, hence $\epsilon_{age}/|R_{age}| \cdot 100\% = 10$ which is considerably larger than the number 0.03 appearing in column I_2 of Table 3. Finally, we apply Mahalanobis nearest neighbor matching to

[5]Denote imb_0 the imbalance in the original data and imb the imbalance left after the match. Then, for column $\mathcal{L}_1$, Table 3 reports $imb/imb_0 \cdot 100\%$, columns I_1, 25%, 50% and 75% report $|imb - imb_0|/|R| \cdot 100\%$, and column I_2 gives $I_2/|R| \cdot 100\%$, where $|R|$ is the range of the corresponding variable.

	$\mathcal{L}_1$	I_1	25%	50%	75%	I_2
age	-100.00	1.41			-2.63	6.59
education	-31.37	-0.99				3.50
black	400.00	0.54				0.67
married	151.67	1.62				3.37
nodegree	-51.60	-4.31				6.73
re74		0.08		0.40		3.36
re75		0.24		-0.63	0.66	3.23
hispanic	-100.00	-1.87				
u74	-16.24	-0.33				2.36
u75	-70.13	-3.16				2.02

Table 4: Imbalance reduction due to Mahalanobis matching from the original data (in Table 2) as a percent of the range of each variable. Positive values for $\mathcal{L}_1$, I_1, and the three quantiles means imbalance increased. The local imbalance, I_2, is also reported. Global imbalance: $\mathcal{L}_1 = 1.06$, $m_T = 297$, $m_C = 297$.

the original data, which gives a global imbalance of $\mathcal{L}_1 = 1.06$. Other imbalance measures are given in Table 4.

With CEM, the $\mathcal{L}_1$ imbalance is greatly improved overall and for each variable and is always better than the corresponding value for Mahalanobis matching. The difference in global imbalance when comparing multidimensional histograms, as measured by $\mathcal{L}_1$, is almost unchanged by Mahalanobis matching but greatly improved as a result of CEM.

Progressive coarsening We now illustrate the progressive coarsening extension of CEM introduced in Section 6.6.3, which is useful when the number of matched units are fewer than desired, and you are willing to consider alternative values for ϵ. Although we recommend choosing ϵ on the basis of substantive knowledge of the variables, for our methodological purposes we begin this illustration by selecting ϵ via Sturges automatic rule. We then relax each variable sequentially decreasing the number of intervals of the discretization used to coarsen the data.

It took 3.2 seconds to perform 30 CEM relaxations. Figure 2 summarizes the results, which both makes it easy to choose new values of ϵ. The figure gives on the horizontal axis the name of the covariate relaxed (with the smaller number of intervals used for the discretization in parentheses). The corresponding percentage of treated units matched is reported on the left vertical axis with the absolute number on the right vertical axis. Each dot on the plot is labeled with the value of the $\mathcal{L}_1$ measure for that particular CEM solution. In this example, we chose minimal coarsenings to constrain the algorithm ($\theta^{\min}_{re74} = 6$, $\theta^{\min}_{re75} = 5$, $\theta^{\min}_{age} = 3$, $\theta^{\min}_{education} = 3$). The label "`<start>`" on the x-axis represents the starting point, and each successive change is listed to its right. The results are sorted in order from closest to this starting point, on the left, to the biggest increases in sample size on the right (as is typical, $\mathcal{L}_1$ increases with the matched sample size in these data). The MIB property of CEM can be seen by noting that multiple coarsenings for any one (color-coded) variable appears farther to the right as the number of coarsened strata decline.

From the largest vertical jumps (on the right side of Figure 2), it is clear that variable `age` is the most difficult variable for matching in these data, followed by `education`. Dots connected by horizontal lines on the figure reveal different solutions with the same number of matched units, some of which have different levels of imbalance, $\mathcal{L}_1$. In applications, we may also wish to consider joint relaxation of variables, but we do not pursue this here.

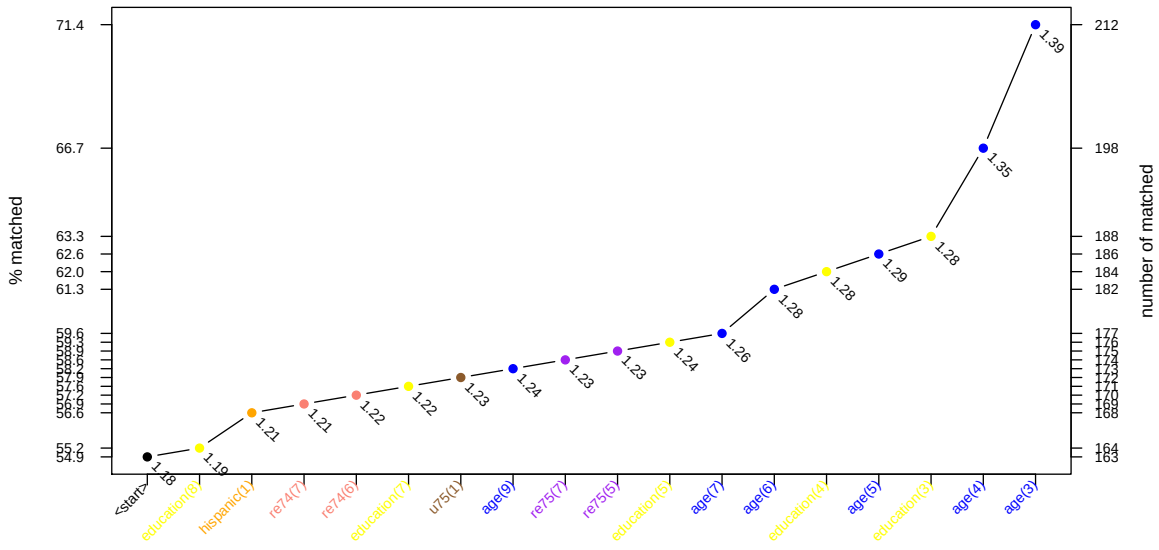

Figure 2: Relaxation of each covariate.

7.2 EPBR Data

We consider here EPBR-compliant data. First, we draw two data sets from multivariate normal distributions: $\mathbf{X}_T \sim N_5(\mu_T, \Sigma)$ and $\mathbf{X}_C \sim N_5(\mu_C, \Sigma)$, with with common variances $(6, 2, 1, 2, 1)$ and covariances, $(2, 1, 0.4, -1, -0.2, 1, -0.4, 0.2, 0.4, 1)$, and means vectors $\mu_T = (0, 0, 0, 0, 0)$ and $\mu_C = (1, 1, 1, 1, 1)$. We randomly sample $n_T = 1,000$ treated units from $\mathbf{X}_T$ and $n_C = r \cdot n_T$ control units from $\mathbf{X}_C$ with $r = 1, 2, 3$. In this experiment, we compare Mahalanobis (MAH) and propensity score (PSC) matching and CEM. For CEM, each covariate is coarsened into 8 intervals of equal length. We also allow PSC and MAH the advantage of matching with replacement, in order to help them avoid trivial solutions. That is, MAH and PSC match $m_T = 1,000$ treated units against a variable number m_C of control units, whereas CEM selects both treated and control units (see Section 2.3).

Mahalanobis and propensity score matching should perform optimally in the sense of minimizing the difference in means I_1 after the match on average (Rosenbaum and Rubin, 1985*b*). CEM is designed to constrain the local imbalance, that is, the maximum distance between a treated unit and the corresponding matched control units, which we can measure with $\mathcal{L}_1$ overall and I_2 for each variable. (See Section 2.7 for definitions; For $\mathcal{L}_1$ we divide each covariate into 11 equispaced intervals to evaluate the k-dimensional histogram.)

Overall, we find that CEM is as good as the other methods in terms of the difference in means (I_1) for which these other methods were designed, but CEM is clearly superior in matching all other local aspects of the treated and control distributions, as measured by I_2 and $\mathcal{L}_1$.

These results can be seen in Tables 5 and 6, which report results for 1,000 and 3,000 control units, respectively, with I_1 reported in the top panel and I_2 and $\mathcal{L}_1$ reported in the bottom panel of each table. The tables also show that MAH is systematically worse than PSC and CEM. As would be expected when there is more to the data than just the mean, CEM is better than PSC on the first two covariates (which have much larger variances) whereas the contrary is true for the remaining covariates. All these differences are relatively small. The tables also show that CEM is as fast or faster than the other methods computationally, and this is with CEM programmed in R and the others in native C.

In terms of local imbalance, measured by I_2 (in the bottom panel of each table), CEM is considerably better than PSC on all covariates. We can also see that PSC is consistently worse than MAH. So in terms of I_2, CEM clearly dominates MAH which in turn dominates PSC. The same ordering is produced by $\mathcal{L}_1$. Although $\mathcal{L}_1$ is not linear, and should primarily be used for comparisons, it is evident that the imbalance reduction, as measured by $\mathcal{L}_1$, is very small for MAH and PSC and quite large for CEM. This means that CEM is indeed greatly reducing the distance between the two k-dimensional distributions of treated and control units.

	X_1	X_2	X_3	X_4	X_5	m_T	m_C	Seconds
initial imb.	1.00	1.00	1.00	1.00	1.00	1000	1000	0.00
CEM	0.04	0.02	0.06	0.06	0.04	341	340	0.08
MAH	0.20	0.20	0.20	0.20	0.20	1000	408	0.28
PSC	0.11	0.06	0.03	0.06	0.03	1000	616	0.16

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

Table 5: 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). Results are averaged over 5,000 replications, with $n_T = 1,000$, $n_C = 1,000$. The initial global imbalance is $\mathcal{L}_1 = 1.24$. (Computational times are in seconds on a 2.00 GHz Intel Core 2 Duo machine.)

	X_1	X_2	X_3	X_4	X_5	m_T	m_C	Seconds
initial imb.	1.00	1.00	1.00	1.00	1.00	1000	3000	0.00
CEM	0.04	0.02	0.05	0.06	0.04	513	921	0.15
MAH	0.14	0.14	0.14	0.14	0.14	1000	625	0.60
PSC	0.07	0.04	0.02	0.04	0.02	1000	2157	0.40

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

Table 6: See caption to Table 5, which holds here except $n_C = 3000$ and the initial global imbalance $\mathcal{L}_1 = 1.17$.

Other regularities emerges from this analysis as well: all methods performs about as well as the reservoir of control units (drawn from the same population) grows. Mahalanobis matching and CEM agrees on the fact that not all the control units are good counterfactuals, and the numbers of control units selected do not differ drastically. These results are also consistent with how the methods were designed: PSC is designed to make means of the distributions closer but is not intended to make any other aspect of the distributions match well. MAH matching takes into account the local characteristics of the data and so it is a better measure of local closeness and thus matches more than just the mean. CEM is designed to look for the best counterfactuals locally and

match all aspects of the distributions. Only MAH behaves like a real EPBR method in reducing the initial imbalance by (almost) the same amount on each covariate in terms of I_1.

Thus, given data drawn to meet EPBR conditions, the optimal approach would be to choose a method based on the nature of the g_0 function (i.e. linear/non linear, smooth/non-smooth, etc.) and the relative importance of different covariates, if they are known, based on whether one needs matching of more than the mean. The conservative approach, which is appropriate when little is known about the real nature of g_0, seems to apply CEM, as it performs almost as the best EPBR method under EPBR conditions in terms of what EPBR methods are designed for, and performs much better for what CEM is designed for, at the cost of loosing some treated units. We show in the next section that this cost is not binding in this decision.[6].

7.3 Non-EPBR Data

We now evaluate CEM in data that violate the EPBR assumptions. To do this, we use the data generation process chosen by Diamond and Sekhon (2005) to evaluate their genetic matching algorithm. This involves using covariates chosen by Dehejia and Wahba (1999), a subset of the Lalonde data, setting the (homogeneous) treatment effect to \$1,000, and generating Y via this highly nonlinear form:

$$Y = 1000 \cdot \mathtt{T} + 0.1 \cdot \exp\left(0.7 \cdot \log(\mathtt{re74} + 0.01 + 0.7 \cdot \log(\mathtt{re75} + 0.01)\right) + \epsilon$$

where $\epsilon \sim N(0, 10)$. The value of the treatment variable is then assigned to each observation on the basis of a true propensity score e, given by

$$e_i = \text{logit}^{-1}\Big\{1 + 0.5 \cdot \hat{\mu} + 0.01 \cdot \mathtt{age}^2 - 0.3 \cdot \mathtt{education}^2 - 0.01 \cdot \log(\mathtt{re74} + 0.01)^2 + 0.01 \cdot \log(\mathtt{re75} + 0.01)^2\Big\}$$

where $\hat{\mu}$ is the linear predictor of the following misspecified logistic model used to estimate a propensity score (as in Dehejia and Wahba 1999):

$$\begin{aligned}\hat{\mu} = {} & 1 + 1.428 \cdot 10^{-4} \cdot \mathtt{age}^2 - 2.918 \cdot 10^{-3} \cdot \mathtt{educ}^2 - 0.2275 \cdot \mathtt{black} - 0.8276 \cdot \mathtt{hispanic} \\ & + 0.2071 \cdot \mathtt{married} - 0.8232 \cdot \mathtt{nodegree} - 1.236 \cdot 10^{-9} \cdot \mathtt{re74}^2 \\ & + 5.865 \cdot 10^{-10} \cdot \mathtt{re75}^2 - 0.04328 \cdot \mathtt{u74} - 0.3804 \cdot \mathtt{u75}\end{aligned}$$

In each of 5000 replications from this process, we assign the treatment to observation i by sampling from the Bernoulli with parameter e_i, i.e. $T_i \sim \text{Bern}(e_i)$, so the number of pre-match treated and control units in the sample varies over replications. We then compare SATT estimators based on the difference in means (RAW in Table 7), the nearest neighbor propensity score matching (PSC), the nearest neighbor Mahalanobis matching (MAH), Genetic Matching (GEN), and CEM using our automatically selected discretization.

As in Diamond and Sekhon (2005), we report results in terms of the bias ("BIAS"), standard deviation ("SD"), and root mean square error ("RMSE") of the SATT estimate over the 5,000 Monte Carlo replications. We also report the average number of matched units, which is lower

[6]We include genetic matching in our simulations in the next section, but skip it here because of unrealistic computational times. Each genetic matching run takes about 2.5 minutes with 1,000 controls and about 4–6 minutes for 3000, which would mean about 75 hours to complete 1,000 replications only for genetic matching, compared to a total of less than 2 hours for the total run time for all remaining methods taken together. We did run a small number of genetic matching runs and find that it selects the same number of control units as Mahalanobis matching, with CEM outperforming in terms of I_1. CEM is also better in terms of I_2 most of the time in each replication, although for some variables (X_3 and X_5), the methods provide the same level of local imbalance.

	BIAS	SD	RMSE	Treated	Controls	Seconds	$\mathcal{L}_1$
RAW	-423.72	1566.49	1622.63	151	293	0.00	1.28
MAH	784.80	737.93	1077.20	151	151	0.03	1.08
PSC	260.45	1025.83	1058.28	151	151	0.02	1.23
GEN	78.33	499.50	505.55	151	143	27.38	1.12
CEM	0.78	111.39	111.38	86	151	0.03	0.76

Table 7: Comparison of bias, standard deviation, root mean square error, computational speed (seconds) and the $\mathcal{L}_1$ measure of imbalance for the original data (RAW), Mahalanobis distance (MAH), propensity score matching (PSC), genetic matching (GEN), and CEM, with values averaged over 5,000 Monte Carlo Replications. Also given are the number of treated and control units selected by each method.

for CEM than for other methods, given the automated coarsening we chose (in practice of course, coarsening should be chosen based on the substance of the variables and so in general the number could be larger or smaller). Despite this, CEM dominates the other methods on each of the three evaluative criteria. Table 7 also gives results on computational speed and the $\mathcal{L}_1$ balance metric, which CEM also improves on.

Relative to the original data, Mahalanobis matching increases the absolute bias but reduces the variance, which nets out to reducing the RMSE by about a third. Propensity score matching reduces the variance (but less than Mahalanobis) and also the bias, which nets to about the same RMSE. Genetic matching reduces both bias and variance, resulting in about a two-thirds reduction in RMSE compared to the raw data. In contrast, CEM eliminates nearly all bias, and the vast majority of the variance, which nets to a 93% reduction in RMSE as compared to the original data. CEM (programmed in R) is also about 900 times faster than genetic matching (programmed mostly in C). Of course, each of these other methods have many potential uses, and the timing differences in particular do not matter much for smaller data sets, but at a minimum CEM would seem to be very widely applicable. (We ran other Monte Carlo experiments with more difficult, complicated, and heterogeneous data generation processes — and also allowed different methods to estimate their own best estimand, keeping SATT constant, and then letting it vary by also matching treated units — and reached similar conclusions.)

8 Concluding Remarks on What Can Go Wrong

Our main goal in this paper has been to introduce the new class of MIB matching methods for making causal inferences from observational data. We demonstrate the usefulness of this class of methods by developing CEM as the simplest method with MIB properties. We conclude here with a discussion of what can go wrong and how to avoid it.

Setting ϵ appropriately is the primary issue to consider when running CEM. If an element of ϵ is set too large, then information that might have been useful to produce better matches may be missed. This is an issue, but analysts have a second chance to avoid the consequences of this problem in the analysis after matching. Of course, the less precise the match, the more burden is put on getting the modeling assumptions correct in the analysis stage.

In contrast, if elements of ϵ are set too small, then too many observations may be discarded without a chance for compensation during the analysis stage. If they are set much too small, a solution may either be unavailable or lead to a low efficiency solution. One must also be careful allowing selection to occur on the treated units and to recognize and clarify for readers the new estimand. As we use CEM in practice, we tend to choose higher standards for what constitutes a match and thus are sometimes left in real observational data sets with fewer observations than

we might have otherwise, with the result being less covariate imbalance, less model dependence, and less resulting statistical bias. In many cases, smaller CEM matched data sets eliminate much heterogeneity, resulting also in causal estimates with smaller variances. With or without these lower variances, the additional bias reduction means that CEM-based estimates will normally have lower mean square error as well. Of course, if ϵ is set as high as you are comfortable with, and your matched data set is still too small, then no magical method will be able to fix this basic data inadequacy, and you will be left trying to model your way out of the problem or to collect more informative data.

When used properly with informative data, CEM can reduce model dependence and bias, and improve efficiency, across a wide range of potential applications. Even when it is possible to design a superior matching method specially for a particular data set, the simplicity of CEM will ordinarily still be far better than the commonly used parametric-only approaches. In these situations, users may opt for CEM, but they should be aware of the potential gain from delving more deeply into the increasingly sophisticated methodological literature in this area.

Finally, all the issues with matching in general may also go wrong with CEM. For example, CEM will not save you if an important covariate is not matched on, unless it is closely related to a variable that is matched on.

A Proofs of Propositions

To simplify the notation, we drop the index j everywhere in the proofs of this section. For the notation refer to Section 4.4.

Proof of Proposition 1. Let us introduce the means by strata

$$\bar{X}_{m_T^s} = \frac{1}{m_T^s} \sum_{i \in T_s} X_i \quad \text{and} \quad \bar{X}_{m_C^s} = \frac{1}{m_C^s} \sum_{i \in C_s} X_i$$

then

$$\bar{X}^w_{m_T} = \frac{1}{m_T} \sum_{i \in T} X_i w_i = \frac{1}{m_T} \sum_{s=1}^{\theta} \sum_{i \in T_s} X_i = \frac{1}{m_T} \sum_{s=1}^{\theta} m_T^s \bar{X}_{m_T^s}$$

and

$$\bar{X}^w_{m_C} = \frac{1}{m_C} \sum_{i \in C} X_i w_i = \frac{1}{m_C} \sum_{s=1}^{\theta} \sum_{i \in C_s} X_i \frac{m_C}{m_T} \frac{m_T^s}{m_C^s} = \frac{1}{m_T} \sum_{s=1}^{\theta} m_T^s \bar{X}_{m_C^s}$$

hence

$$\left| \bar{X}^w_{m_T} - \bar{X}^w_{m_C} \right| \leq \sum_{s=1}^{\theta} \frac{m_T^s}{m_T} \left| \bar{X}_{m_T^s} - \bar{X}_{m_C^s} \right| \leq \sum_{s=1}^{\theta} \frac{m_T^s}{m_T} \epsilon = \epsilon$$

□

Proof of Proposition 2. The k-th centered moment for variable X_j around the weighted mean for the control units can be written as follows

$$\frac{1}{m_C} \sum_{s=1}^{\theta} \sum_{i \in C_s} |X_i - \bar{X}^w_{m_c}|^k w_i = \frac{1}{m_C} \sum_{s=1}^{\theta} \sum_{i \in C_s} (|X_i - \bar{X}^w_{m_T}| + |\bar{X}^w_{m_T} - \bar{X}^w_{m_c}|)^k w_i$$

now we apply the binomial expansion $(a+b)^k = \sum_{h=0}^{k} \binom{k}{h} a^h b^{k-h}$ to the inner terms of the

summation

$$
\begin{aligned}
(|X_i - \bar{X}^w_{m_T}| + |\bar{X}^w_{m_T} - \bar{X}^w_{m_C}|)^k &= \sum_{h=0}^{k} |X_i - \bar{X}_{m_T}|^h |\bar{X}_{m_T} - \bar{X}_{m_C}|)^{k-h} \\
&\leq \sum_{h=0}^{k} |X_i - \bar{X}^w_{m_T}|^h \epsilon^{k-h} && \text{(by global bounds)} \\
&\leq \epsilon^k \sum_{h=0}^{k} |R|^h \epsilon^{k-h} = \epsilon^k \sum_{h=0}^{k} \left|\frac{R}{\epsilon}\right|^h && \text{(mean is internal)} \\
&= \epsilon^k \sum_{h=0}^{k} \theta^{*h} = \epsilon^k (\theta^* + 1)^k && \text{(by (6) and bin. exp.)}
\end{aligned}
$$

The same bound occurs for $|X_i - \bar{X}^w_{m_T}|$ when $i \in T_s$. Therefore,

$$
\frac{1}{m_C} \sum_{s=1}^{\theta} \sum_{i \in C_s} |X_i - \bar{X}^w_{m_c}|^k w_i \leq \epsilon^k (\theta^* + 1)^k \frac{1}{m_C} \sum_{s=1}^{\theta} \sum_{i \in C_s} w_i = \epsilon^k (\theta^* + 1)^k
$$

because

$$
\frac{1}{m_C} \sum_{s=1}^{\theta} \sum_{i \in C_s} w_i = \frac{1}{m_C} \sum_{s=1}^{\theta} \sum_{i \in C_s} \frac{m_C}{m_T} \frac{m^s_T}{m^s_C} = \frac{1}{m_T} \sum_{s=1}^{\theta} m^s_C \frac{m^s_T}{m^s_C} = 1
$$

notice also that $\frac{1}{m_T} \sum_{s=1}^{\theta} \sum_{i \in T_s} w_i = 1$. Hence, each centered k-th moment is bounded by the positive quantity $\epsilon^k (\theta^* + 1)^k$, so it their absolute difference

$$
\left| \frac{1}{m_C} \sum_{s=1}^{\theta} \sum_{i \in C_s} |X_i - \bar{X}^w_{m_C}|^k w_i - \frac{1}{m_T} \sum_{s=1}^{\theta} \sum_{i \in T_s} |X_i - \bar{X}^w_{m_T}|^k w_i \right| \leq \epsilon^k (\theta^* + 1)^k
$$

In the case of k-to-j matching, where k and j do not vary over the strata or for one-to-one matching, weights can be discarded and the result simplifies to

$$
\left| \frac{1}{m_C} \sum_{i \in C} |X_i - \bar{X}_{m_C}|^k - \frac{1}{m_T} \sum_{i \in T} |X_i - \bar{X}_{m_T}|^k \right| \leq \epsilon^k (\theta^* + 1)^k
$$

but it is possible to obtain a better bound as follows in the case of one-to-one or k_s-to-k_s matching. We focus on stratum s and denote by X^C_s the control units in that stratum and X^T_s the treated ones. The proof is presented for one-to-one matching, but the results also apply to k_s-to-k_s matching.

$$
\begin{aligned}
|X^C_s - \bar{X}_{m_C}|^k &= |X^C_s - X^T_s + X^T_s - \bar{X}_{m_T} + \bar{X}_{m_T} - \bar{X}_{m_C}|^k \\
&\leq (|X^C_s - X^T_s| + |X^T_s - \bar{X}_{m_T}| + |\bar{X}_{m_T} - \bar{X}_{m_C}|)^k \\
&\leq \left(|X^T_s - \bar{X}_{m_T}| + 2\epsilon_j\right)^k && \text{(by local bounds)} \\
&= |X^T_s - \bar{X}_{m_T}|^k + \sum_{h=0}^{k-1} \binom{k}{h} |X^T_s - \bar{X}_{m_T}|^h (2\epsilon)^{k-h} && \text{(by bin. exp.)} \\
&\leq |X^T_s - \bar{X}_{m_T}|^k + \epsilon^k \sum_{h=0}^{k-1} \binom{k}{h} \left|\frac{X^T_s - \bar{X}_{m_T}}{\epsilon}\right|^h 2^{k-h} \\
&\leq |X^T_s - \bar{X}_{m_T}|^k + \epsilon^k \sum_{h=0}^{k-1} \binom{k}{h} \theta^{*h} 2^{k-h} && \text{(by (6))} \\
&\leq |X^T_s - \bar{X}_{m_T}|^k + \epsilon^k ((\theta^* + 2)^k - \theta^{*k}) && \text{(by bin. exp.)}
\end{aligned}
$$

Hence

$$\frac{1}{m_T}\left(\sum_{s=1}^{\theta}|X_{s,j}^C-\bar{X}_{m_C,j}|^k-|X_{s,j}^T-\bar{X}_{m_T,j}|^k\right)\leq\epsilon^k\left((\theta^*+2)^k-\theta^{*k}\right)$$

The two bounds are equivalent for large k but for $k=2$ the above bound becomes linear in θ^*, i.e., $\epsilon^2\left((\theta^*+2)^2-\theta^{*2}\right)=\epsilon^2(4+2\theta^*)$. □

Proof of Proposition 3. Consider the q^{th} empirical quantiles of the distribution of the treated and control units, $Q_{m_T,j}$ and $Q_{m_C,j}$. That is, $Q_{m_T,j}$ is the q^{th} ordered observation of the subsample of m_T matched treated units, and similarly for $Q_{m_C,j}$. In one-to-one matching, the first treated observation is matched against the first control observation in the first strata, and in general, the corresponding quantiles belong to the same strata. Therefore, $|Q_{m_T,j}-Q_{m_C,j}|<\epsilon_j$.

A similar proof holds for the k_s-to-j_s matching. First define the weighted empirical distribution functions for treated $F_{m_T}^w(x)$ and control groups $F_{m_C}^w(x)$ as

$$F_{m_T}^w(x)=\sum_{x_i\leq x,i\in T}\frac{w_i}{m_T}\quad\text{and}\quad F_{m_C}^w(x)=\sum_{x_i\leq x,i\in C}\frac{w_i}{m_C}.$$

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

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

Thus, for each stratum, $F_{m_T}^w(b_s)=m_T^s/m_T=F_{m_C}^w(b_s)$, and hence the difference between weighted empirical distribution functions at the end points of each stratum $[a_s,b_s]$ is always zero. Therefore, if we define the *weighted empirical quantile* of size q as the first observation x such that $F(x)\geq q$ (where F is either $F_{m_C}^w$ or $F_{m_T}^w$), the weighted quantiles of the same order for treated and control units always belong to the same stratum and hence the difference between the quantile x for the treated and control units, given q is at most ϵ. □

References

Abadie, A. and J. Gardeazabal. 2003. "The Economic Costs of Conflict: A Case Study of the Basque Country." *American Economic Review* 93(1):113–132.

Abadie, Alberto and Guido W. Imbens. 2007. "Bias-Corrected Matching Estimators for Average Treatment Effects.". http://ksghome.harvard.edu/ aabadie/research.html.

Austin, Peter C. and Muhammad M. Mamdani. 2006. "A comparison of propensity score methods: A case-study estimating the effectiveness of post-AMI statin use." *Statistics in Medicine* 25:2084–2106.

Battistin, E. and A. Chesher. 2004. "The Impact of Measurement Error on Evaluation Methods Based on Strong Ignorability." *Institute for Fiscal Studies, London* .

Box, George E.P., William G. Hunger and J. Stuart Hunter. 1978. *Statistics for Experimenters*. New York: Wiley-Interscience.

Cochran, William G. 1968. "The effectiveness of adjustment by subclassification in removing bias in observational studies." *Biometrics* 24:295–313.

Cochran, William G. and Donald B. Rubin. 1973. "Controlling bias in observational studies: A review." *Sankhya: The Indian Journal of Statistics, Series A* 35, Part 4:417–466.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens and Oscar Mitnik. 2006. "Moving the Goalposts: Addressing Limited Overlap in Estimation of Average Treatment Effects by Changing the Estimand." Department of Economics, UC Berkeley.

Dehejia, Rajeev H. and Sadek Wahba. 1999. "Causal Effects in Nonexperimental Studies: Re-Evaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94(448):1053–62.

Diamond, Alexis and Jasjeet Sekhon. 2005. "Genetic Matching for Estimating Causal Effects: A New Method of Achieving Balance in Observational Studies." http://jsekhon.fas.harvard.edu/.

Freedman, D. and P. Diaconis. 1981. "On the histogram as a density estimator: L_2 theory." *Probability Theory and Related Fields* 57:453–476.

Galdo, Jose, Jeffrey Smith and Dan Black. 2008. "Bandwidth Selection and the Estimation of Treatment Effects with Unbalanced Data." University of Michigan.

Girosi, Federico and Gary King. 2008. *Demographic Forecasting*. Princeton: Princeton University Press. http://gking.harvard.edu/files/smooth/.

Hansen, Ben. 2008. "The Prognostic Analogy of the Propensity Score." *Biometrika* 95(2):481–488.

Heckman, James, H. Ichimura and P. Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program." *Review of Economic Studies* 64:605–654.

Hirano, Keisuke, Guido W. Imbens and Geert Ridder. 2003. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica* 71(4):1161–1189.

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

Iacus, Stefano M. and Giuseppe Porro. 2007. "Missing data imputation, matching and other applications of random recursive partitioning." *Computational Statistics and Data Analysis* 52(2):773–789.

Iacus, Stefano M. and Giuseppe Porro. 2008. "Invariant and Metric Free Proximities for Data Matching: An R Package." *Journal of Statistical Software* 25(11):1–22.

Iacus, Stefano M. and Giuseppe Porro. 2008, forthcoming. "Random recursive partitioning: a matching method for the estimation of the average treatment effect." *Journal of Applied Econometrics* .

Imai, K. and D.A. van Dyk. 2004. "Causal inference with general treatment regimes: Generalizing the propensity score." *Journal of the American Statistical Association* 99(467):854–866.

Imai, Kosuke, Gary King and Clayton Nall. 2008. "The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation." *Statistical Science* . (tentatively accepted), http://gking.harvard.edu/files/abs/cluster-abs.shtml.

Imai, Kosuke, Gary King and Elizabeth Stuart. 2008. "Misunderstandings Among Experimentalists and Observationalists about Causal Inference." *Journal of the Royal Statistical Society, Series A* 171, part 2:481–502. http://gking.harvard.edu/files/abs/matchse-abs.shtml.

Imbens, Guido W. 2000. "The Role of the Propensity Score in Estimating Dose-Response Functions." *Biometrika* 87(3):706–710.

Imbens, Guido W. 2004. "Nonparametric estimation of average treatment effects under exogeneity: a review." *Review of Economics and Statistics* 86(1):4–29.

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

King, Gary and Langche Zeng. 2006. "The Dangers of Extreme Counterfactuals." *Political Anal-*

ysis 14(2):131–159. http://gking.harvard.edu/files/abs/counterft-abs.shtml.

King, Gary and Langche Zeng. 2007. "When Can History Be Our Guide? The Pitfalls of Counterfactual Inference." *International Studies Quarterly* pp. 183–210. http://gking.harvard.edu/files/abs/counterf-abs.shtml.

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

Lu, Bo, Elaine Zanuto, Robert Hornik and Paul R. Rosenbaum. 2001. "Matching With Doses in an Observational Study of a Media Campaign Against Drug Abuse." *Journal of the American Statistical Association* 96(456):1245–1253.

Manski, Charles F. 1995. *Identification Problems in the Social Sciences*. Harvard University Press.

Mielke, P.W. and K.J. Berry. 2007. *Permutation Methods: A Distance Function Approach*. New York: Springer.

Morgan, Stephen L. and Christopher Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge: Cambridge University Press.

Popoviciu, T. 1935. "Sur Les Équations Algébriques Ayant Toutes Leurs Racines Réelles." *Mathematica* 9:129–145.

Rosenbaum, Paul R. and Donald B. Rubin. 1985*a*. "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score." *The American Statistician* 39:33–38.

Rosenbaum, P.R. and D.B. Rubin. 1985*b*. "The Bias Due to Incomplete Matching." *Biometrics* 41(1):103–116.

Rosenbaum, P.R., R.N. Ross and J.H. Silber. 2007. "Minimum Distance Matched Sampling With Fine Balance in an Observational Study of Treatment for Ovarian Cancer." *Journal of the American Statistical Association* 102(477):75–83.

Rubin, Donald. 1976*a*. "Inference and Missing Data." *Biometrika* 63:581–592.

Rubin, Donald B. 1976*b*. "Multivariate Matching Methods That are Equal Percent Bias Reducing, I: Some Examples." *Biometrics* 32(1):109–120.

Rubin, Donald B. 1976*c*. "Multivariate Matching Methods that are Equally Percent Bias Reducing, II: Maximums on Bias Reduction for Fixed Sampled Sizes." *Biometrics* 32:121–132.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.

Rubin, Donald B. 2001. "Using propensity scores to help design observational studies: Application to the tobacco litigation." *Heatlh Services & Outcomes Research Methodology* 2(3-4):169–188.

Rubin, Donald B. 2006. *Matched Sampling for Causal Effects*. Cambridge, England: Cambridge University Press.

Rubin, Donald B. and Elizabeth A. Stuart. 2006. "Affinely invariant matching methods with discriminant mixtures of proportional ellipsoidally symmetric distributions." *Annals of Statistics* 34(4):1814–1826.

Rubin, Donald B. and Neal Thomas. 1992. "Affinely Invariant Matching methods with Ellipsoidal Distributions." *Annals of Statistics* 20(2):1079–1093.

Rubin, Donald B. and Neal Thomas. 1996. "Matching Using Estimated Propensity Scores, Relating Theory to Practice." *Biometrics* 52:249–264.

Scott, D.W. 1992. *Multivariate density estimation. Theory, practice and visualization*. New York: John Wiley & Sons, Inc.

Sekhon, Jasjeet S. 2008. "Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The matching Package for R." *Journal of Statistical Software* .

Shimazaki, H. and S. Shinomoto. 2007. "A Method for Selecting the Bin Size of a Time Histogram." *Neural Computation* 19(6):1503–1527.