Political Analysis Advance Access published August 23, 2011

doi:10.1093/pan/mpr013

Causal Inference without Balance Checking: Coarsened Exact Matching

Stefano M. Iacus
Department of Economics, Business and Statistics, University of Milan, Via Conservatorio 7, I-20124 Milan, Italy
e-mail: stefano.iacus@unimi.it

Gary King
Institute for Quantitative Social Science, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138
e-mail: king@harvard.edu (corresponding author)

Giuseppe Porro
Department of Economics and Statistics, University of Trieste, P.le Europa 1, I-34127 Trieste, Italy
e-mail: giuseppe.porro@econ.units.it.

We discuss a method for improving causal inferences called ''Coarsened Exact Matching'' (CEM), and the new ''Monotonic Imbalance Bounding'' (MIB) class of matching methods from which CEM is derived. We summarize what is known about CEM and MIB, derive and illustrate several new desirable statistical properties of CEM, and then propose a variety of useful extensions. We show that CEM possesses a wide range of statistical properties not available in most other matching methods but is at the same time exceptionally easy to comprehend and use. We focus on the connection between theoretical properties and practical applications. We also make available easy-to-use open source software for *R*, *Stata,* and *SPSS* that implement 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 might summarize as follows: information, information everywhere, nor a datum to trust (with apologies to Samuel Taylor Coleridge).

Matching is a nonparametric method of controlling for 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 mean 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 require controlling for **X** with a model (such as the same model that would have been used without matching), but the only inferences necessary are those relatively close to the data, leading to less model dependence and reduced statistical bias than without matching (Ho et al. 2007).

Authors' note: Open source R, Stata, and SPSS software to implement the methods described herein (called CEM) is available at http://gking.harvard.edu/cem; the CEM algorithm is also available via a standard interface offered in the R package MatchIt. Thanks to Erich Battistin, Nathaniel Beck, Matt Blackwell, Andy Eggers, Adam Glynn, Justin Grimmer, Jens Hainmueller, Ben Hansen, Kosuke Imai, Guido Imbens, Fabrizia Mealli, Walter Mebane, Clayton Nall, Enrico Rettore, Jamie Robins, Don Rubin, Jas Sekhon, Jeff Smith, Kevin Quinn, and Chris Winship for helpful comments. All information necessary to replicate the results in this paper appear in Iacus, King, and Porro (2011b).

The central dilemma means that model dependence and statistical bias are usually much bigger problems than large variances.[1] Unfortunately, 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 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 and model dependence.

Continually checking balance, rematching, and checking again until balance is improved on all variables is the best current practice with most 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, it is hard to blame applied researchers who might reasonably expect that a method touted for its ability to reduce imbalance might actually do so when used once.

The problem stems from the fact that widely used current methods, such as propensity score and Mahalanobis matching, are members of the class of matching methods known as ''equal percent bias reducing'' (EPBR), which does not guarantee any level of imbalance reduction in any given data set; its properties only hold on average across samples and even then only by assuming a set of normally unverifiable assumptions about the data generation process. In any application, a single use of these techniques can increase imbalance and model dependence by any amount.

To avoid these and other problems with EPBR methods, Iacus, King, and Porro (2011) introduce a new generalized class of matching methods known as ''Monotonic Imbalance Bounding'' (MIB). We discuss a particular member of the MIB class of matching methods that Iacus, King, and Porro (2011) call ''Coarsened Exact Matching'' (CEM). CEM works in sample and requires no assumptions about the data generation process (beyond the usual ignorability assumptions). More importantly, CEM and other MIB methods invert the process and thus guarantee that the imbalance between the matched treated and control groups will not be larger than the ex ante user choice. This level is chosen by the user on the basis of specific, intuitive substantive information, which they demonstrably have. (If you understand the tradeoffs in drawing a histogram, you will understand how to use this method.) With MIB methods, improvements in the bound on balance for one covariate can be studied and improved in isolation as it will have no effect on the maximum imbalance of each of the other covariates.

CEM-based causal estimates possess a large variety of other powerful statistical properties as well. Some of these are proven in Iacus, King, and Porro (2011) and others are demonstrated here for the first time. In a large variety of real and simulated data sets, including data that meet the assumptions made by EPBR methods, Iacus, King, and Porro (2009, 2011) and King et al. (2011) show that CEM dominates commonly used existing (EPBR and other) matching methods in its ability to reduce imbalance, model dependence, estimation error, bias, variance, mean square error, and other criteria. We summarize the properties of CEM here and then introduce a variety of extensions that make the method more widely applicable in practice.

[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.''

CEM can thus be thought of as an easy first line of defense in protecting users from the threats to validity in making causal inferences. The method can also used with other existing methods so that the combined method inherits the properties shown here apply to CEM. In what follows, we introduce our notation and setup (Section 2), describe CEM (Section 3), discuss the properties of CEM (Section 4), and extend CEM in various useful ways (Section 5). We then offer an empirical illustration to show how it works in practice (Section 6) and conclude with a discussion of what can go wrong when using this approach (Section 7). All data and code necessary to replicate the results in this paper appear in Iacus, King, and Porro (2011b).

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 postestimation matching, what to do when some treated units cannot be matched, a general characterization of error in estimating the target quantities, and how to measure imbalance.

2.1 *Notation*

Consider a sample of $n \leqslant N$ units drawn from a population of size N. Let T_i denote an indicator variable for unit i that takes on value 1 if unit i is a member of the ''treated'' group and 0 if i is a member of the ''control'' group. The observed outcome variable is $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$, 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, $Y_i(0)$ 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 pretreatment covariates. For this purpose, we denote $\mathbf{X} = (X_1, X_2, \ldots X_k)$ as a k-dimensional data set, where each X_j is a column vector of observed values of pretreatment variable j for the n sample observations (possibly drawn from a population, of size N). That is, $\mathbf{X} = [X_{ij}, i = 1, \ldots, n, j = 1, \ldots, k]$. Let $\mathcal{T} = \{i\colon T_i = 1\}$ be the set of indexes for the treated units and $n_T = \#\mathcal{T}$ be a count of the elements of this set; similarly $\mathcal{C} = \{i\colon T_i = 0\}$, $n_C = \#\mathcal{C}$ for the control units, with $n_T + n_C = n$. Let $\mathbf{X}_i = (X_{i1}, \ldots, X_{ik})$ be the vector of covariates for observation i. We denote by m_T and m_C the number of treated and control units matched by some method. Let $M_\mathcal{T} \subseteq \mathcal{T}$ and $\mathrm{M}_\mathcal{C} \subseteq \mathcal{C}$ be the sets of indexes of the matched units in the two groups.

2.2 *Quantities of Interest*

As usual, the treatment effect for unit i, $\mathrm{TE}_i = Y_i(1) - Y_i(0)$, is unobserved. Many relevant causal quantities of interest are averages of TE_i over different subsets of units and so must be estimated. The most common include the *sample* (SATT) and *population* (PATT) *average treatment effect on the treated*:

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

where $\mathcal{T}^*$ is the set of indexes of treated units in the whole population and $N_T = \#\mathcal{T}^*$ (see Imbens 2004; Morgan and Winship 2007).

Although SATT is a quantity of interest in and of itself, without regard to a population beyond the sample data, if the sample is randomly drawn from the relevant population, $E(\mathrm{SATT}) = \mathrm{PATT}$ (where the expected value operator averages over repeated samples).

2.3 *Simplifying Assumptions*

First, similar to the ''no omitted variable bias'' assumption in the social sciences, we make the standard ignorability assumption: conditional on $\mathbf{X}$, the treatment variable is independent of the potential outcomes: $T_i \perp\!\!\!\perp \{Y_i(0), Y_i(1)\} | \mathbf{X}$.

Second, matching-based estimators tend to focus on SATT (or PATT) so that if they choose to retain all treated units, and prune only control units, the target quantity of interest remains the same. Thus, for each

observation, $Y_i(1)$ is always observed, whereas $Y_i(0)$ is always estimated (by choosing values from the control units via some matching algorithm or applying some model). Section 2.6 discusses what to do when the analyst chooses to prune treated units when no reasonable match exists among the pool of available controls.

2.4 *Existing Matching Methods*

This section outlines the most commonly used 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 $\mathbf{X}_i$), 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 $\mathbf{X}_i$ 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 treated and control units outside the common empirical support of both groups since including them would require unreasonable extrapolation far from the data. The second step then matches the treated unit to some control observation $\tilde{X}$ that, if not exactly $\mathbf{X}$, 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, genetic algorithm, 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 must be 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 could use entirely different technologies for the first step, such as kernel density estimation (Heckman, Ichimura, and Todd 1997) or dropping control units outside the hyperrectangle (Iacus and Porro 2009) or convex hull (King and Zeng 2006) of the treated units. In practice, most published applications skip the first step entirely and instead match all treated units, which is not advisable. The method we introduce below avoids these problems by satisfying both steps simultaneously in the same algorithm.

2.5 *Postmatching 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 (such as some type of regression model or specially designed nonparametric methods; Abadie and Imbens 2007) can alternatively be used on the matched data set with the benefit of having a lower risk of model dependence (Ho et al. 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. 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 $\mathbf{X}_i$, with the treated unit-level estimated causal effects averaged over all treated units. For an example of an implementation of these approaches, see Iacus, King, and Porro (2009).

2.6 *When Matches for All Treated Units Do Not Exist*

When one or more treated units have no reasonable matches among the set of available controls, standard approaches will lead to unacceptably high levels of model dependence. In this situation, three options are available with any matching method. First, we can decide that the data include insufficient information to estimate the target causal effect and give up, producing no inference at all. Second, we can create controls by extrapolating from some given model, although leaving us with high levels of model dependence. Or finally, we can change the estimand to the *local SATT*, that is the treatment effect averaged over only the subset of treated units for which good matches exist among available controls. This third approach is often used in applications, such as when applying propensity score or Mahalanobis matching with calipers.

The recognized best practice in the literature currently is to eliminate the extrapolation region as a separate prior step and then to match. This procedure deletes treated units without good matches and so is a version of the third option of changing the estimand. This choice is reasonable so long as one is transparent about the choice and the consequences in terms of the new set of treated units over which the causal effect is defined (as, e.g., Crump et al. 2009). 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 (Imbens and Angrist 1994). 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).

However, most published applications of standard matching methods do not eliminate the extrapolation region and instead match at all costs regardless of whether reasonable matches exist among the control units. In these studies, analysts are effectively taking the second option and producing highly model dependent inferences, but without necessarily even knowing it.

We also offer here a more general way to think about this problem, following Iacus, King, and Porro (2011). Thus, we first partition the n_T treated units into the $m_T \leqslant n_T$ units, which can be matched well from the set of controls, and the $n_T - m_T$ units, which involve extreme counterfactuals (i.e., extrapolations) far from the treated units. (Unlike the matching method we introduce below, most standard methods require a separate prior step to accomplish this, such as the convex hull or hyperrectangle; see Section 2.4.) Then, we match the data in the first subset with acceptable controls to produce a "local SATT," say $\hat{\tau}_{m_T}$. Then, for the rest of the treated units, we extrapolate via some model estimated on the matched units to obtain virtual control units for the unmatched treated units and produce an (necessarily model dependent) estimate $\hat{\tau}_{n_T - m_T}$. Finally, we calculate the overall SATT estimate $\hat{\tau}_{n_T}$ as the weighted mean of the two estimates:

$$\hat{\tau}_{n_T} = \frac{\hat{\tau}_{m_T} \cdot m_T + \hat{\tau}_{n_T - m_T} \cdot (n_T - m_T)}{n_T}. \quad (1)$$

The result is that the SATT is fixed ex ante, with the two components separately estimated and clearly identified. In practice, analysts may wish to present all three or just the local SATT.

2.7 *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:

$$Y_i(0) = g_0(X_i) = g_0(X_{i1}, \ldots, X_{ik}), \quad (2)$$

where g_0 is an unknown function (cf. Imai, King, and Stuart 2008). If equation (2) 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 equation (2) as $\mathrm{TE}_i = Y_i(1) - Y_i(0) = \widehat{\mathrm{TE}}_i + \mathcal{E}_0(\tilde{\mathbf{X}}_i, \mathbf{X}_i)$, where $\mathcal{E}_0(\tilde{\mathbf{X}}_i, \mathbf{X}_i) \equiv g_0(\tilde{\mathbf{X}}_i) - g_0(\mathbf{X}_i) = \hat{\mathbf{Y}}_i(0) - Y_i(0)$ is the unit-level treatment effect error (not an expected value). Then, we aggregate this over treated units into $\mathrm{SATT} = \frac{1}{n_T} \sum_{i \in T} \mathrm{TE}_i = \widehat{\mathrm{SATT}} + \bar{\mathcal{E}}_{\mathrm{o}}$, where $\widehat{\mathrm{SATT}} = \sum_{1 \in T} \widehat{\mathrm{TE}}_i / n_T$, and the average estimation error is as follows:

$$\bar{\mathcal{E}}_{\mathrm{o}} \equiv \frac{1}{n_T} \sum_{i \in \mathcal{T}} \mathcal{E}_0(\tilde{\mathbf{X}}_i, \mathbf{X}_i) = \frac{1}{n_T} \sum_{i \in \mathcal{T}} [g_0(\tilde{\mathbf{X}}_i) - g_0(\mathbf{X}_i)]. \quad (3)$$

The ultimate goal of matching-based estimators is to reduce the absolute matching error, $|\bar{\mathcal{E}}_{\mathrm{o}}|$. This goal can be parsed into two (nonadditive) components (Imai, King, and Stuart 2008). 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 pretreatment covariates for the control group $p(\tilde{\mathbf{X}}|T = 0)$ and treated group $p(\mathbf{X}|T = 1)$ in some chosen metric (such as those discussed in Section 2.8). 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 equation (3), where the difference between $\tilde{\mathbf{X}}_i$ and $\mathbf{X}_i$ 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 $\left|\bar{\mathcal{E}}_{\mathrm{o}}\right|$ vanishes, no matter what g_0 is. When that lucky situation does not occur, the two components must be considered together.

2.8 *Measuring Imbalance*

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

$$I_1 = \left| \bar{\boldsymbol{X}}^w_{m_T, j} - \bar{\boldsymbol{X}}^w_{m_C, j} \right|, \quad j = 1, \ldots, k, \quad (4)$$

where $\bar{\boldsymbol{X}}^{w}_{m_T,j}$ and $\bar{\boldsymbol{X}}^{w}_{m_C,j}$ denote weighted means of variable X_j for the groups 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 is 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 the empirical quantile distributions (Rubin 2001; Austin and Mamdani 2006; Imai, King, and Stuart 2008). 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.

An alternative approach introduced in Iacus, King, and Porro (2011) is to measure the multivariate differences between $p(\mathbf{X}|T = 1)$ and $p(\tilde{\mathbf{X}}|T = 0)$ via an L_1 distance, fixing the bin size to that for the median L_1 for all possible binnings on the raw data. (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])

Let $H(X_1)$ be the set of distinct values generated by binning on variable X_1—the set of intervals into which the support of variable X_1 has been cut. Then, the multidimensional histogram is constructed from the set of cells generated by the Cartesian product $H(X_1)\times\ldots\times H(X_k) = H(\boldsymbol{X})$. Let f and g be the relative empirical frequency distributions for the treated and control groups. Let $f_{\ell_1\ldots\ell_k}$ be the relative frequency for observations belonging to the cell with coordinates $\ell_1,\ldots,\ell_k$ of the multivariate cross-tabulation of the treated units and $g_{\ell_1\ldots\ell_k}$ for the control units.

Definition 1 (Iacus, King, and Porro 2011).

The multivariate imbalance measure is

$$\mathcal{L}_1(f,g) = \frac{1}{2}\sum_{\ell_1\ldots\ell_k\in H(\mathbf{X})} \left|f_{\ell_1\ldots\ell_k} - g_{\ell_1\ldots\ell_k}\right|. \tag{5}$$

Thus, the typically huge number of empty cells do not affect $\mathcal{L}_1(f, g)$, and the summation in equation (5) never has more than n nonzero terms. The relative frequencies also control for potentially different sample sizes between the groups. Denote by f^m and g^m the empirical frequencies for matched treated and control groups corresponding to the unmatched f and g frequencies and use the same discretization for both the treated and the control units. Then, a good matching method will have $\mathcal{L}_1(f^m, g^m) \leqslant \mathcal{L}_1(f, g)$. The values of $\mathcal{L}_1$ are easily interpetable: if the two distributions of data are completely separated (up to the fine coarsening of the histogram), then $\mathcal{L}_1 = 1$; if the two distributions overlap exactly, then $\mathcal{L}_1 = 0$. In all other cases, $\mathcal{L}_1 \in (0, 1)$. The values of $\mathcal{L}_1$ provide useful *relative* information; if, for example, $\mathcal{L}_1 = 0.6$, then only 40% of the density of the two histograms overlap. This measure is relative because its meaning is conditional on the data set and chosen covariates.

[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 postmatching imbalance. Moreover, in practice, we use Iacus, King, and Porro's (2011) suggestion of a fixed bin width, computed by the median of all possible bin widths computed from the raw data.

3 Coarsened Exact Matching

The basic idea of CEM 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 and to prune unmatched units. Finally, the coarsened data are discarded and the original (uncoarsened) values of the matched data are retained.

Put differently, after coarsening, the CEM algorithm creates a set of strata, say $s \in \mathrm{S}$, each with same coarsened values of $\mathbf{X}$. Units in strata that contain at least one treated and one control unit are retained; units in the remaining strata are removed from this sample. We denote by $\mathcal{T}^s$ the treated units in stratum s and by $m_T^s = \#\mathcal{T}^s$ the number of treated units in the stratum, similarly for the control units, that is, $\mathcal{C}^s$and $m_C^s = \#\mathcal{C}^s$. The number of matched units are, respectively, for treated and controls, $m_T = \cup_{s\in S}\, m_T^s$ and $m_C = \cup_{s\in S}\, m_C^s$. To each matched unit i in stratum s, CEM assigns the following weights:

$$w_i = \begin{cases} 1, & i \in T^s \\ \frac{m_C}{m_T}\frac{m_T^s}{m_C^s}, & i \in C^s \end{cases}. \tag{6}$$

Unmatched units receive weight $w_i = 0$.

CEM therefore assigns to matching the task of eliminating all imbalances (i.e., differences between the treated and control groups) beyond some chosen level defined by the coarsening. Imbalances eliminated by CEM include all multivariate nonlinearities, interactions, moments, quantiles, comoments, 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.

Like exact matching, 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 5.2).

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 produces 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 multiparty 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 are still used at the analysis stage to estimate the causal effect, coarsening for CEM involves less onerous assumptions than that made by researchers who regularly 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 *Values of the Coarsened Variables*

We recommend that coarsened values be chosen in a 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 trade-off is normally worth it since these parameters will typically be well known to users (but see Section 5.2). We also 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 $R_j = \max_{i=1,\ldots,n} X_{ij} - \min_{i=1,\ldots,n} X_{ij}$. Then, choosing a default coarsening is equivalent to choosing the value ε_j for each variable, such that $0 < \varepsilon_j \leqslant R_j$, where $\varepsilon_j = R_j$ corresponds to all the observations grouped in a single interval, and $\varepsilon_j = 0$ corresponds to no coarsening. We denote by θ_j the number of nonempty intervals generated, that is, the number of distinct values after coarsening variable X_j. (If the problem requires different length size for each interval, as will often be the case in practice when choosing customized coarsenings, as we recommend, we denote by ε_j the maximal length for our proofs.)

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 may be a large figure. Similarly, smaller intervals may be useful for lower incomes and possibly with \$0 a logically distinct group. 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.

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, consider a survey question about religion that asks about the specific denomination, including say six Protestant denominations, three Jewish, one Catholic, and two Muslim. For this example, a reasonable choice for many 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. In any case, for ordinal or nonordinal variables, one can group different levels together. For example, most 7-point Likert scales have a prominent neutral category and so can often be reasonably coarsened into $\theta_j = 3$ groups as follows: {completely disagree, strongly disagree, disagree}, {neutral}, {agree, strongly agree, completely agree}.

4 Properties of CEM

We list here several attractive properties of CEM, in addition to its simplicity and ease of use. No other matching method satisfies more than a subset of these.

4.1 *An MIB Method*

As proven in Iacus, King, and Porro (2011), CEM is a member of the MIB class of matching methods. This result means, first, that when a researcher chooses a coarsening for a variable, the maximum degree to which that variable can be out of balance between the treated and control groups is also determined (the more coarsening, the more imbalance is allowed). The degree of imbalance may be less than the maximum, but we know for certain it cannot be more than this chosen level.

Second, the coarsening choice for any one variable can have no effect on the imbalance bound for any of the other variables. The result is that the arduous process in other methods of balance checking, tweaking, and repeatedly rerunning the matching procedure is eliminated with CEM, as is the uncertainty about whether the matching procedure will reduce imbalance or instead reduce imbalance on one variable and make it worse on others. You get what you want rather than getting what you get. Of course fixing imbalance ex ante in this way means that we learn the number of observations matched as a consequence of the procedure, rather than determining it as an input, but bias is more crucial than variance in observational data analyses and choosing both requires different types of procedures (see King et al., 2011). In addition, matching can sometimes even reduce variance by removing heterogeneity and model dependent inferences.

4.2 *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 the natural k-dimensional space in the metric of the original data to a (different) space defined by the propensity score or Mahalanobis distance metrics.

In contrast, CEM meets the congruence principle by operating in the space where $\mathbf{X}$ 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).

4.3 *Comparisons with Other Methods*

Whereas CEM uses simple, fixed, nonoverlapping intervals of local indifference, defined ex ante based on the metric of each variable one at a time, nearest neighbor caliper matching uses orthogonalization and a more complicated geometry of n_T overlapping hyperparallelepipeds centered around each treated data point (Cochran and Rubin 1973). 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 postgraduate 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 1 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 of subclassification (or ''stratification'') approaches, such as full matching, frequency matching, subclassification on the propensity score, and others. However, these other 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).

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. 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 most existing methods.[3]

4.4 *Automatic Restriction to Common Empirical Support*

As described in Section 2.4, other 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 a control unit by definition do not involve extrapolating beyond the data and so these observations will be included; otherwise, they 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; CEM could instead be used as a simple way to restrict data to common support to improve other matching methods.

4.5 *Approximate Invariance to Measurement Error*

Suppose T is ignorable conditional on unobserved pretreatment covariates $\mathbf{X}^* = \left(X_1^*, \ldots, X_k^*\right)$, but we match instead on $\mathbf{X}$, 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 $\mathbf{X}$ does not imply balance with respect to $\mathbf{X}^*$; the true propensity score based on $\mathbf{X}$ is not a balancing score for $\mathbf{X}^*$; and adjusting based on $\mathbf{X}$ instead of $\mathbf{X}^*$ will lead to biased estimates of the treatment effect (Battistin and Chesher 2004).

Under CEM, if measurement error is less than ε_j, $\varepsilon_j \geqslant \max(|\eta_j|)$, and it happens to respect the resulting strata boundaries, then CEM will produce the same preprocessed data set whether matching on $\mathbf{X}$ or on $\mathbf{X}^*$ 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 to measurement error (in the sense of Battistin and Chesher 2004) via a real data set described in Section 6.1. We do this by randomly perturbing the earnings variable by adding the Gaussian error $N(\mu = 1000, \sigma^2 = 1000^2)$ and replacing perturbed negative earnings with zero. We run 5000 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 both subsets of matched units before and after perturbation. To measure the sensitivity to perturbation, we calculate $K_T/\min\left(m_T, m_T'\right) \cdot 100\%$ and $K_C/\min\left(m_C, m_C'\right) \cdot 100\%$. For all methods but CEM, $m_T = m_T'$, whereas for all matching algorithms, $m_C \neq m_C'$. 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.

[3]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 in the equal percent bias reducting class of methods 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, 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 this existing method.

Table 1 Percentage of units present in matched sets both before and after perturbation, averaged over 5000 simulations, and computational time (for all methods but CEM, $K_T = 100\%$)

	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

4.6 *Bounding Model Dependence*

To make a causal inference, one must estimate the counterfactual potential outcome $Y_i(0)$ for each treated unit (i.e, the value that Y_i would take if T_i were 0 when it is in fact 1). To do this, we could use the value of the outcome variable for a different unit with a good match, say Y_j, such that $X_j \approx Y_i$. However, in the usual case where insufficient exact matches exist, a better estimate might be obtained by a model, such as some type of regression model: $\hat{Y}(0) \equiv m_\ell(\tilde{\mathbf{X}}_j)$, where $\tilde{\mathbf{X}}_j$ is the vector of covariates for the control units close to treated i and $m_\ell(\,\cdot\,)$ is one of many possible models. Model dependence is defined by how much $m_\ell(\tilde{\mathbf{X}}_j)$ varies as a function of the model m_ℓ for a given vector of covariates $\tilde{\mathbf{X}}_j$ (King and Zeng 2007). Unfortunately, in many situations, model dependence is remarkably large, so that apparently small and otherwise indefensible specification decisions in the regression can have large effects on causal estimates.

A key advantage of matching is that it should reduce model dependence. In other words, preprocessing data via matching ought to lead to different modeling choices having considerably less influence on the estimate of the causal quantity of interest than it would without matching. This relationship has been illustrated in real data by Ho et al. (2007), but it has never been proven mathematically for any previous method, EPBR or otherwise. In contrast, MIB methods have been shown to possess this property (Iacus, King, and Porro 2011), and since CEM is an MIB method, it too possesses this property.

In other words, by choosing the coarsening for each variable, a researcher also controls the bound on the degree of model dependence possible. Less coarsening directly lowers the maximum possible level of model dependence.

4.7 *Bounding the Average Treatment Effect Estimation Error*

Another attractive property of MIB matching methods, and one that distinguishes them from EPBR and other matching methods, is that their tuning parameters bound not only the model dependence used to estimate the causal effect but also the causal effect estimation error itself (from equation (3)). In particular, choosing CEM coarsening to be finer directly reduces the maximum possible estimation error.

To show this result for CEM, 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. We denote by $\Xi_{-j} = \Xi_1 \times \Xi_2 \times \ldots \times \Xi_{j-1} \times \Xi_{j+1} \times \ldots \times \Xi_k$, $x_{-j} = (x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k)$, and $g_0(x_j|x_{-j}) = g_0(x_1, x_2, \ldots, x_k)$.

Assumption 1 (Lipschitz behavior).

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,

$$\max_{x_{-j} \in \Xi_{-j}} \left| g_0(x'_j|x_{-j}) - g_0(x''_j|x_{-j}) \right| \leqslant L_j d_j(x'_j, x''_j),$$

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 the estimation of the treatment effect. This assumption easily fits essentially 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 a 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 0 otherwise. If, alternatively, $H \subset \mathcal{O}$, where $\mathcal{O}$ is the abstract set of ordered categories, the distance is $d(x, y) = |\text{rank}(x) - \text{rank}(y)|$, where rank($x$) is the rank/order of category x in $\mathcal{H}$.

Then, the definitions in Section 2.7 imply directly that the estimation error, $\bar{\mathcal{E}}_0 \equiv \text{SATT} - \widehat{\text{SATT}}$, is bounded from above and below by $|\bar{\mathcal{E}}_0|$, that is, $-|\bar{\mathcal{E}}_0| \leqslant \text{SATT} - \widehat{|\text{SATT}|} \leqslant |\bar{\mathcal{E}}_0|$and a consequence of Assumption 1 is that $|g_0(\mathbf{X}_i) - g_0(\tilde{\mathbf{X}}_i)| \leqslant \max_{j=1,\ldots,k} L_j \varepsilon_j$. Therefore, for the CEM algorithm, which keeps matched treated and control units for each covariate a maximum of ε_j apart, we conclude that

$$|\bar{\mathcal{E}}_0| \leqslant \max_{j=1,\ldots,k} L_j \varepsilon_j. \tag{7}$$

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 5.5.2.).

4.8 *The Number of Matched Units*

If too many treated units are discarded, inferences with CEM may be inefficient. This can be remedied by widening the degree of maximum imbalance. Of course, we might be concerned about the curse of dimensionality, where the number of possible strata from the cross-tabulation of the possible values of $\mathbf{X}$ is ordinarily huge. For example, suppose $\mathbf{X}$ is composed of 10,000 observations on 20 variables drawn from independent normal densities. Since 20-dimensional space is so large, no treated unit will likely be anywhere near any control unit. In this situation, even very coarse bins under CEM will likely produce no matches. For example, with only two bins for each variable, the 10,000 observations would need to be sorted into more than a million strata. In data like these, no matching method could do any good.

Fortunately, most real data sets have much more highly correlated data than the independent draws in the hypothetical example above, and so CEM, in practice, tends to produce reasonable numbers of matches. This has been our recurring experience in the numerous data sets we have analyzed with CEM. In addition, Iacus, King, and Porro (2011) show that if the number of control units is large enough, the number of cells with unmatched treated units goes to zero at a fixed and known rate. That is, in practice, if the data are useful for making causal inferences, CEM will normally produce a well-balanced data set with a reasonable number of observations.

4.9 *Computational Efficiency*

An attractive feature of CEM is that it is extremely efficient computationally, especially compared to some other matching methods. Indeed, each observation i with vector of covariates $\mathbf{X}_i$ is stored as a record containing only the coarsened values pasted one after the other in a single string. As a whole, for n observations, we have only n strings stored. So, the number of covariates do not affect the dimension of the coarsened data set (its length is always n) and finding observations in the same multidimensional cell has the same computational complexity of the tabulation of a distribution of n units (i.e., it is of order n). Thus, even if in principle one should search in the grid of an exponentially large number of cells, in practice, the search is only made on the nonempty cells, which are at most n. This is important because it means the method works out-of-the-box on huge databases using SQL-type queries without the need for statistical software or modeling. In addition, the computational efficiency and simplicity of this CEM procedure are much easier to completely automate.

4.10 *Empirical Properties*

CEM has been compared to most commonly used methods in a large number of real data sets. These include analyses of Food and Drug Administration drug approval times (Carpenter 2002), job training programs (Lalonde 1986), two large data sets evaluating disease management programs (King et al.

2011), and the effects of having a daughter on a member of congress' voting behavior (Washington 2008). We also extended our empirical experience by sampling from all social science causal analyses in progress in many fields by advertising help in making causal analyses in return for a look at researchers' data, promising to preserve authors rights of first publication (Iacus, King, and Porro 2011). In almost all these analyses, CEM generated matched data sets with lower imbalance and a larger sample size than other approaches.

Finally, what may be the most commonly used method presently, propensity score matching, was shown by King et al. (2011) to approximate random matching, thus increasing imbalance, in many circumstances. CEM does not have this damaging property.

5 Extensions of CEM

CEM is so simple that it is easy to extend in a variety of productive ways. We offer seven extensions here.

5.1 *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 (Imbens 2000; Lu et al. 2001; Imai and van Dyk 2004).

5.2 *Combining CEM with Other Methods*

CEM is one of the simplest methods with MIB properties (and the additional properties in Section 4) 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 4. Genetic matching as defined in Diamond and Sekhon (2005) is not MIB, but by choosing a variable-by-variable caliper, it would be; if it were run within CEM strata, 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 and the MIB property would hold.

If the user does not know enough about $\mathbf{X}$'s measurement to coarsen, then productive data analysis of any kind may be infeasible. But in some applications, we can partition $\mathbf{X}$ 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 or any 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.

5.3 *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, it is 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. Thus, we now offer two options to used both in the same analysis enabled by CEM; all our software implementations of CEM allow for multiply imputed data.

The simplest approach is to treat missing values as a discrete ''observed'' value and then to apply CEM with other coarsening used for the nonmissing 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 to feed multiply 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.

5.4 *Avoiding Consequences of Arbitrary 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 point is irrelevant for CEM's intended use, that is, when coarsening is chosen based on substantive criteria (such as a college diploma marking a distinct point in an otherwise continuous education variable), but can be a concern if coarsening is set more arbitrarily or automatically. In this situation, all the properties of CEM described in Section 4 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.

In this situation, we run 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. (Our software implements this automatically.) We then use 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.

5.5 *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.

5.5.1 *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 interquartile 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) $\Delta_{\text{fd}} = 2\big(Q_3 - Q_1\big)n^{-1/3}$. More recently, Shimazaki and Shinomoto (2007) developed an approach based on the Poisson sampling in time series analysis (in the attempt to recover spikes), which we find works well. Our software offers these approaches as options.

5.5.2 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 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 follows:

$$\hat{L_j} = \max_{i_1 \neq i_2 \in \mathcal{C}} \frac{|Y_{i_1}(0) - Y_{i_2}(0)|}{d_j\big(X_{i_1 j}, X_{i_2 j}\big)}. \tag{8}$$

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). Equation (7) 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 equation (8). 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 $\hat{\mathcal{E}}_0 = \max_j \hat{L_j}\varepsilon_j$ and use the terms $\hat{L_j}\varepsilon_j$ as a hints during matching about which covariate may give rise to the largest estimation errors or bias in estimating SATT. Although equation (8) uses the outcome variable, it only does so for control units (as in Hansen 2008), and so inducing selection bias is not a risk.

5.5.3 Inductive coarsening choices

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 recommend that users make these choices based on the substance, 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 (i.e., increase ε_j selectively) by using $\theta' = (\theta'_1, \ldots, \theta'_k)$ such that $\theta' \leqslant \theta$, that is, $\theta'_i = \theta_i$ for all i but a subset of indexes j such that $\theta'_j \leqslant \theta_j$. 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 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 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 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 2, that is, $\theta'_j = \theta_j - 2$, until $\theta'_j < 10$ and then by 1 or up to a user chosen minimally tolerable number of intervals, $\theta_j^{\min}$. (We also shift each intermediate solution as in Section 5.4.) We then repeat the procedure for pairs of variables, (θ_i, θ_j), triplets (θ_i, θ_j, θ_k), etc.[4]

To illustrate progressive coarsening, we make use of the Lalonde (1986) data (described below in Section 6.1). Although we recommend choosing ε on the basis of substantive knowledge of the variables, for our methodological purposes, we select ε via the Sturges automatic rule. We then relax each variable sequentially decreasing the number of intervals of the discretization used to coarsen the data.

It took 7.0 seconds to perform 30 CEM relaxations. Figure 1 summarizes the results, which 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 ($\theta_{\text{re74}}^{\min} = 6$, $\theta_{\text{re75}}^{\min} = 5$, $\theta_{\text{age}}^{\min} = 3$, $\theta_{\text{education}}^{\min} = 3$). 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 Fig. 1), 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.

5.6 *Blocking in Randomized Experiments*

Since "blocking" (i.e., prerandomization matching) in randomized experiments bests complete randomization with respect to bias, efficiency, power, and robustness, it should be used whenever feasible (Imai, King, and Stuart 2008; Imai, King, and Nall 2009). CEM provides an easy method of determining the blocks: After matching the coarsened pretreatment covariates $\mathbf{X}$ 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. Multicategory treatments in blocking are also easy to create with CEM by randomly assigning observations within each stratum to 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.

5.7 *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 4.4). Avoiding extreme model dependence is also an issue that applies to any type of counterfactual inference—including

[4]Combined with shifted coarsenings, an exhaustive procedure with greater than triplets is feasible only via parallel processing, which happens to be 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.

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 $\mathbf{X}$ for the counterfactual inference of interest and then run CEM on the augmented data set. Observations that fall in the same stratum as the pseudoobservation 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 procedure represents a small generalization (due to coarsening) of a point emphasized by Manski (1995), who would use $\varepsilon = 0$.

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.

6 CEM in Practice

We now offer an illustration of the operation of CEM based on simulations (Section 6.2) and real data (Section 6.3). We describe the data used in both sections first (Section 6.1).

6.1 *Data*

Data in this paper come from the National Supported Work Demonstration, a U.S. job training program (Lalonde 1986). 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. From this experiment, Lalonde (1986) created an experimental and an observational data set for further analysis. A cleaned subset of both data sets created by Dehejia and Wahba (2002), which we also analyze, have been widely used in the matching literature as a benchmark for evaluating methods (e.g., Imbens 2003; Smith and Todd 2005). The experimental data set includes 297 treated and 425 control units from the experiment, which, because of randomization, are easy to match. The observational data set combines the 297 treated units from the experiment with 2490 control units from an observational survey, the Panel Study of Income Dynamics. Lalonde et al. have shown that one cannot recover SATT from the observational data set in part because the data are highly imbalanced and relatively few good matches exist within the control group. We use the experimental data set in Sections 4.5 and 6.2 and the observational data set in Section 6.3.

Pretreatment variables in these data were measured for both participants and controls, and include age (age), years of education (education), marital status (married), lack of a high school diploma (nodegree), race (black and 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, and u75), some are categorical (age and education), and the earnings variables are continuous and highly skewed, with point masses at zero.

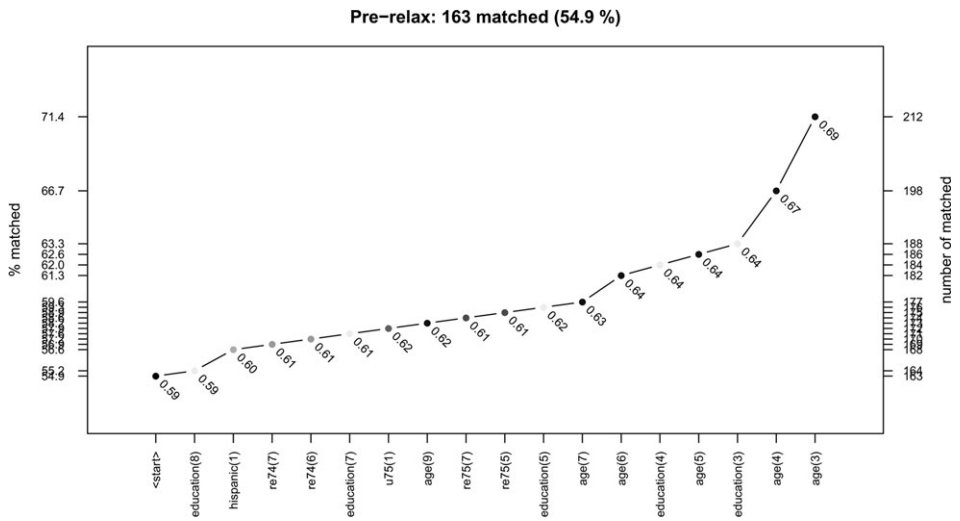

Fig. 1 Relaxation of each covariate.

6.2 *Simulations*

We now compare CEM to several other methods by using the data generation process chosen by Diamond and Sekhon (2005) to evaluate their algorithm. This involves using covariates chosen by Dehejia and Wahba (1999), a subset of the Lalonde data, setting the (homogeneous) treatment effect to \$1000, and generating Y via this highly nonlinear form:

$$Y = 1000 \cdot \mathrm{T} + 0.1 \cdot \exp(0.7 \cdot \log(\mathrm{re74} + 0.01) + 0.7 \cdot \log(\mathrm{re75} + 0.01)) + \varepsilon$$

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

$$e_i = \mathrm{logit}^{-1}\left\{1 + 0.5 \cdot \hat{\mu} + 0.01 \cdot \mathrm{age}^2 - 0.3 \cdot \mathrm{education}^2 - 0.01 \cdot \log(\mathrm{re74} + 0.01)^2 + 0.01 \cdot \log(\mathrm{re75} + 0.01)^2\right\},$$

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):

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

In each of 5000 replications from this process, we assign the treatment to observation i by sampling from the Bernoulli with parameter e_i, that is, $T_i \sim \mathrm{Bern}(e_i)$, so the number of prematch treated and control units in the sample varies over replications. We then compare SATT estimators based on the difference in

Table 2 Comparison of bias, standard deviation, root mean square error, computational speed (seconds), and the (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 5000 Monte Carlo replications. Also given are the number of treated and control units selected by each method

	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

means (RAW in Table 2), the nearest neighbor propensity score matching (PSC), the nearest neighbor Mahalanobis matching (MAH), genetic matching (GEN), and CEM using our automatically selected discretization.

Following 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 5000 Monte Carlo replications. We also report the average number of matched units, which is lower 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 2 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 99.8% of the 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) and is feasible with many more covariates and observations. 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 the different methods to estimate their own best estimand, keeping SATT constant, or letting it vary by also matching treated units—and reached similar conclusions. King et al. (2011) performed similar comparisons while also allowing PSM and MAH to work with calipers and showed that CEM still dominates to some extent even more strongly than the results here. Even though this section shows that CEM substantially outperforms other methods, it would be easy to outperform these results using other applications of CEM or the combined methods discussed in Section 5.2. The usual ''ping pong theorem'' qualifications certainly apply.

6.3 *Empirical Example*

We now present a step-by-step illustration of estimating a causal effect in data from Lalonde (1986).

6.3.1 Matching

To begin a CEM analysis, we first choose a reasonable coarsening for each variable. The more coarsening we allow, the more observations we will have, but larger the bound on model dependence and estimation error. For education, we divide years of education in classes corresponding to different levels: grade school (0–6), middle school (7–8), high school (9–12), college (13–16), and graduate school (>16). For age, we use standard labor force classes: (15–19), (20–24), (25–34), (35–44), (45–54), (55–64), and (>65) years.

The data includes two indicator variables to identify unemployment in 1974 and 1975 (u74 and u75); we include these or, since the unemployed have zero earnings, we instead equivalently add the interval [0, 1)

Table 3 Imbalance measured by difference in means between treated and control units on the original data (RAW), after CEM, propensity score matching (PSM), PSM with re74^2 in the model (PSM*), and CEM with five quantiles for re74 (CEM*)

	Age	*Education*	*re74*	*re75*	*u74*	*u75*	*Married*	*Nodegree*	*Black*	*Hispanic*
RAW	−10.22	−1.74	−15,857.75	−15,997.24	0.35	0.27	−0.70	0.43	0.55	0.06
PSM	−3.55	−0.34	−3706.94	−3243.38	0.21	0.15	−0.31	0.16	0.10	0.01
CEM	−0.43	−0.10	−1158.83	−1364.83	0.00	0.00	0.00	0.00	0.00	0.00
PSM*	−3.96	−0.31	−3809.26	−2959.22	0.20	0.13	−0.31	0.15	0.11	0.02
CEM*	−0.44	−0.13	−1046.94	−1140.66	0.00	0.00	0.00	0.00	−0.00	0.00

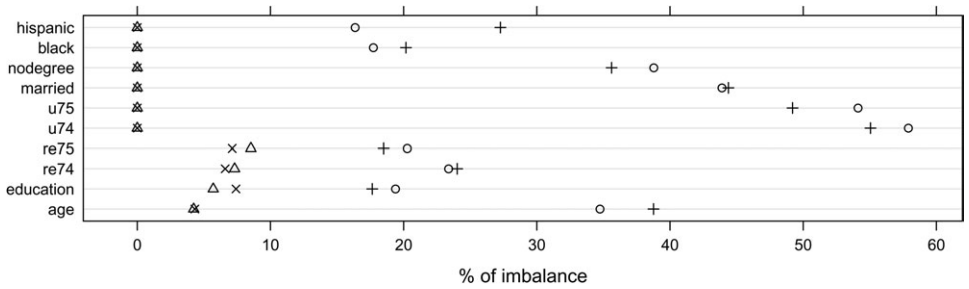

Fig. 2 Percentage of bias left after matching (relative to the RAW data at 100%), for CEM (△), CEM* (×), PSM* (+), and PSM (○). This figure offers an alternative visualization of Table 3.

dollars to the coarsening of re74 and re75. We divide earnings into quantiles, calculated on the distributions of positive earnings. The quantiles (25, 50, 75)% of re74 are (11,756, 18,925, 26,842). So finally, we obtain the following cutpoints for variable re74: [0–1), [1–11,756), [11,756–18,925), [18,925–26,842), and [26,842–137,149). For variable re75, we have [0–1), [1–11,069), [11,069–18,261), [18,261–26,855), and [26,855–156,653). The initial multidimensional imbalance is $\mathcal{L}_1 = 0.977$.

After running CEM with the above coarsening, we obtain $m_T = 176$ treated units matched with $m_C = 218$ control units, resulting in a multivariate imbalance of $\mathcal{L}_1 = 0.806$; this corresponds to a moderately sized 17.47% imbalance reduction.

We now do a parallel analysis using the most common approach to propensity score matching. This involves using nearest neighbor matching on the estimated propensity score, where the score comes from a logistic regression using all pretreatment variables and all treated units matched. This produces a multivariate imbalance of $\mathcal{L}_1 = 0.953$, which corresponds to a reduction of only 2.48% relative to the raw data and considerably smaller than CEM's result.

But what about the simpler univariate difference in means imbalance metric for which the propensity score was designed? As can be seen in the first three rows of Table 3, or the triangles in Figure 2, propensity score matching (PSM) reduces imbalance for each variable, but CEM reduces it more for every one.

In any matching method, the acceptable level of imbalance for a pretreatment variable depends on its expected effect on the outcome—with lower imbalance being more desirable for variables with larger expected effects. Unfortunately, changing imbalance in predictable ways is often impossible with methods, such as the propensity score, that use a scalar imbalance metric. The problem is complicated by the fact that the usual ways to assess model specification are irrelevant here; the only question is whether imbalance is changed as predicted. For example, suppose we wish to reduce imbalance further for re74 and decide to try adding the term re74^2 to the propensity score logit model specification. The result is that the in-sample fit (as measured by the AIC statistic) is improved, but row "PSM*" in Table 3, and the "+" symbol in Figure 2, shows that imbalance on re74 has increased. Indeed, the direction of change in imbalance is unpredictable for the other variables as well, as can be seen, for example, in the reduction in imbalance for variable re75.

Consider how much easier tightening the imbalance is with CEM. We can, for example, change the coarsening by splitting re74 into five quantiles instead of the previously used three. Row "CEM*" in Table 3, and the "×" symbol in Figure 2, shows that we vastly reduced the imbalance on this variable and others never exceed the maximum imbalance we specified ex ante. For example, although the imbalance slightly increases for age and education, this increase respects the MIB property: for ages, we tolerated a maximal imbalance of 5 years, so 0.01 increase of imbalance were deemed ex ante as irrelevant. Similarly for education.

One can improve propensity score matching by continuing to tweak and rerun the model, recheck imbalance, and rerun again, but because the imbalance results generally go in unpredictable directions, finding a specification that improves balance on all variables can often be challenging. In contrast, because it is a member of the MIB class of methods, CEM produce no surprises about imbalance, which makes data analysis far easier: reducing maximum imbalance on one variable never has any effect on the maximum imbalance specified for any of the other variables.

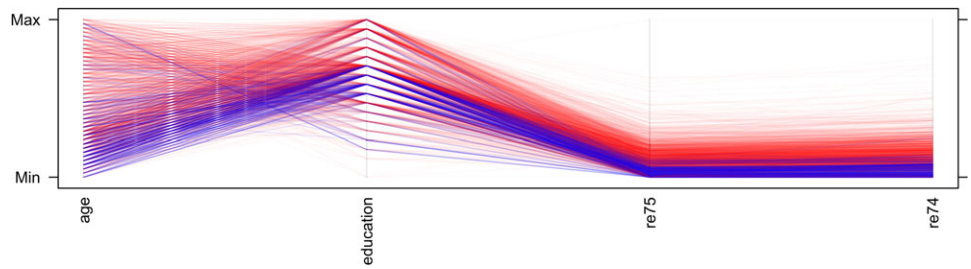

Fig. 3 Parallel plot of noncategorical variables for CEM matched units (blue) against unmatched units (red) in the Lelonde versus Panel Study of Income Dynamics data.

6.3.2 Estimation

Finally, we illustrate ways of estimating the causal effect under CEM. We begin with the local SATT, using the CEM-matched units via linear regression using all variables in the data set. The result is an estimate of $\hat{\tau}_{m_T} = -1223.7$ and leaves $n_T - m_T = 121$ unmatched treated units. So one option is to extrapolate the model estimated for the m_T matched units to the remaining unmatched treated units to estimate SATT for all treated units. This approach yields $\tilde{\tau}_{n_T} = -\$554.7$, which is quite far from $\hat{\tau}_{m_T}$.

Another approach is to estimate the local SATT on the unmatched $n_T - m_T = 121$ treated and control $n_C - m_C = 2272$ units only and then use formula 1 to obtain another estimate of the global SATT. Doing this, we obtain $\hat{\tau}_{n_T - m_T} = -1467.5$ and thus another estimate of the global SATT:

$$\hat{\tau}_{n_T} = \frac{176(-1223.7) + 121(-1467.5)}{297} = -1323.1.$$

A somewhat more conservative approach in the extrapolation region of unmatched treated units is to first prune control units outside the hyperrectangle of the subsample of treated units. This approach leaves us with 121 treated units and 43 control units. In this case, we get $\hat{\tau}'_{n_T - m_T} = -\$7494.20$ and

$$\hat{\tau}'_{n_T} = \frac{121(-1223.7) + 176(-7494.2)}{297} = -3778.4.$$

The differences in the estimates of the global SATT illustrate the unavoidable model dependence in the extrapolation region. For the local SATT, CEM enables one to produce a highly stable estimate that is relatively invariant to the estimation method.

6.3.3 Defining the quantity of interest

In practice, most methods of matching select by treated and controls and so produce a definition of the estimand as part of the matching procedure. Since matching is a method of preprocessing, and so the estimand is defined prior to estimation, all the usual statistical properties, such as bias and efficiency, still apply. However, understanding what quantity (which "local SATT") is being estimated is crucial. Methodologists recognize this fact but rarely do anything about it. We offer here a simple way of defining the estimand, via a parallel plot; see Figure 3. In a parallel plot, each observation in the original data set is represented by a single line that traces out its values on each of the continuous variables (horizontally) between its minimum and maximum values (vertically). We adapt this graphic for our purposes by coloring matched units in blue and unmatched units in red. The estimand is the average treatment effect for the people who were matched, which we can see from the figure are younger, have middling levels of education, and receive lower earnings than the full group. This is the group for whom a relatively stable, not model dependent inference is available. Applications ought to use a parallel plot like this, or some other approach, to characterizing the quantity being estimated.

7 Concluding Remarks on What Can Go Wrong

We conclude here with a discussion of what can go wrong in applying CEM and how to avoid these problems.

Choosing the coarsening (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.

References

Abadie, Alberto, and Javier Gardeazabal. 2003. The economic costs of conflict: A case study of the Basque Country. *American Economic Review* 93:113–32.

Abadie, Alberto, and Guido W. Imbens. 2007. Bias-corrected matching estimators for average treatment effects. Unpublished manuscript. 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* 25:2084–106.

Battistin, Erich, and Andrew Chesher. 2004. *The impact of measurement error on evaluation methods based on strong ignorability* . Working paper, Institute for Fiscal Studies, London.

Carpenter, Daniel Paul. 2002. Groups, the media, agency waiting costs, and FDA drug approval. *American Journal of Political Science* 46:490–505.

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* 4:417–66.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar Mitnik. 2009. Dealing with limited overlap in estimation of average treatment effects. *Biometrika* 96:187.

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* 94:1053–62.

———. 2002. Propensity score matching methods for non-experimental causal studies. *Review of Economics and Statistics* 84:151–61.

Diamond, Alexis, and Jasjeet Sekhon. 2005. *Genetic matching for estimating causal effects: A new method of achieving balance in observational studies*. Working paper, http://jsekhon.fas.harvard.edu/ (accessed 2005).

Freedman, David, and Persi Diaconis. 1981. On the histogram as a density estimator: L_2 theory. *Probability Theory and Related Fields* 57:453–76.

Galdo, Jose, Jeffrey Smith, and Dan Black. 2008. *Bandwidth selection and the estimation of treatment effects with unbalanced data*. Working paper, University of Michigan.

Girosi, Federico, and Gary King. 2008. *Demographic forecasting*. Princeton, NJ: Princeton University Press. Unpublished manuscript. http://gking.harvard.edu/files/smooth/ (accessed 2008).

Hansen, Ben. 2008. The prognostic analogy of the propensity score. *Biometrika* 95:481–88.

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* 64:605–54.

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 (accessed 2007).

Iacus, Stefano M., Gary King, and Giuseppe Porro. 2009. CEM: Coarsened Exact Matching Software. *Journal of Statistical Software* 30(9), http://gking.harvard.edu/cem.

———. 2011. Multivariate matching methods that are Monotonic Imbalance Bounding. *Journal of the American Statistical Association*. http://gking.harvard.edu/files/abs/cem-math-abs.shtml.

———. 2011b. Replication data for: Causal inference without balance checking: Coarsened Exact Matching. Murray Research Archive [distributor] V1 [version]. http://hdl.handle.net/1902.1/15601.

Iacus, Stefano M., and Giuseppe Porro. 2007. Missing data imputation, matching and other applications of random recursive partitioning. *Computational Statistics and Data Analysis* 52:773–89.

———. 2008. Invariant and metric free proximities for data matching: An R package. *Journal of Statistical Software* 25(11):1–22.

———. 2009. Random recursive partitioning: A matching method for the estimation of the average treatment effect. *Journal of Applied Econometrics* 24:163–85.

Imai, Kosuke, Gary King, and Clayton Nall. 2009. The essential role of pair matching in cluster-randomized experiments, with application to the Mexican universal health insurance evaluation. *Statistical Science* 24(1):29–53. http://gking.harvard.edu/files/abs/cluster-abs.shtml.

Imai, Kosuke, Gary King, and Elizabeth Stuart. 2008. Misunderstandings among experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A 171,* 2:481–502. http://gking.harvard.edu/files/abs/matchse-abs.shtml (accessed 2008).

Imai, Kosuke, and D. A. van Dyk. 2004. Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association* 99:854–66.

Imbens, Guido W. 2000. The role of the propensity score in estimating dose-response functions. *Biometrika* 87:706–10.

———. 2003. Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review* 96:126–32.

———. 2004. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics* 86:4–29.

Imbens, Guido W., and Joshua D. Angrist. 1994. Identification and estimation of local average treatment effects. *Econometrica* 62:467–75.

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* 95:49–69. http://gking.harvard.edu/files/abs/evil-abs.shtml (accessed 2001).

King, Gary, Richard Nielsen, Carter Coberley, James Pope, and Aaron Wells. 2011. *Comparative effectiveness of matching methods for causal inference*.

King, Gary, and Langche Zeng. 2006. The dangers of extreme counterfactuals. *Political Analysis* 14:131–59. http://gking.harvard.edu/files/abs/counterft-abs.shtml.

———. 2007. When can history be our guide? The pitfalls of counterfactual inference. *International Studies Quarterly* 51:183–210. http://gking.harvard.edu/files/abs/counterf-abs.shtml.

Lalonde, Robert. 1986. Evaluating the econometric evaluations of training programs. *American Economic Review* 76:604–20.

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* 96:1245–53.

Manski, Charles F. 1995. *Identification problems in the social sciences*. Cambridge, MA: Harvard University Press.

Mielke, Paul W., and Kenneth 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.

Rosenbaum, Paul R., Richard N. Ross, and Jeffrey 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* 102:75–83.

Rubin, Donald B. 1976. Inference and missing data. *Biometrika* 63:581–92.

———. 1987. *Multiple imputation for nonresponse in surveys*. New York: John Wiley.

———. 2001. Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services & Outcomes Research Methodology* 2:169–88.

———. 2006. *Matched sampling for causal effects*. Cambridge, UK: Cambridge University Press.

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

Shimazaki, Hideaki, and Shigeru Shinomoto. 2007. A method for selecting the bin size of a time histogram. *Neural Computation* 19:1503–27.

Smith, Jeffrey A., and Petra E. Todd. 2005. Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics* 125:305–53.

Washington, Ebonya L. 2008. Female socialization: How daughters affect their legislator fathers' voting on woman's issues. *American Economic Review* 98:311–32.