Statistica Sinica: Supplement to Kernel Balancing A flexible non-parametric weighting procedure for estimating causal effects

Chad Hazlett
Departments of Statistics & Political Science,
University of California Los Angeles

May 29, 2019

S1 Choice of Discrepancy Measure

A method is needed to find the weight vector w such that $\frac{1}{N_1}\mathbf{K}_t\mathbf{1}_{N_1} = \mathbf{K}_c w$, while constraining the weights to be non-negative and sum to one. It is also desirable to do this with minimal variation in the weights, by some measure, and in particular to avoid large weights. Two natural candidates for this are empirical likelihood (Owen, 1988), and entropy balancing (Hainmueller, 2012), both special cases of Cressie-Read divergence from a uniform distribution (Cressie and Read, 1984). Other approaches such as those that explicitly minimize the variation in weights for a given degree of imbalance (e.g. Zubizarreta, 2015) may be valuable as well. In the `kbal`, I utilize entropy balancing, which seeks to satisfy these conditions while maximizing the Shannon entropy, $\sum_i w_i log(w_i)$, implied by the weights, which is also (proportional to) the Kullback divergence entropy between the distribution of weights and a uniform distribution. See Hainmueller (2012) and references therein for further discussion.

S2 Equivalence of K-imbalance and smoothed multivariate density imbalance

Recall that the choice the optimization procedure chooses the number of projections of $\mathbf{K}$ that must be balanced while seeking to minimize overall imbalance on $\mathbf{K}$. Minimizing an imbalance measure of the form $a||\overline{k_t} - \sum_{i:D=0} w_i k_i||$ for some norm $||\cdot||$ is natural given the goal of mean balance on $\mathbf{K}$. Such a norm also provides a measure of continuous multivariate imbalance. Setting a to $\frac{1}{\sqrt{2\pi b}}$ to obtain $||\frac{1}{N_1\sqrt{2\pi b}}\mathbf{K_t}^\top\mathbf{1}_{N_1} - \frac{1}{\sqrt{2\pi b}}\mathbf{K_c}^\top w||$ we see this equals $||\hat{p}_{D=1}(\mathbf{X}) - \hat{p}_{w,D=0}(\mathbf{X})||$, a norm on the difference between the smoothed density estimators for the treated and (weighted) controls, evaluated

at each observation in the dataset. Hence, norms of the form $||\overline{k_t} - \sum_{i:D=0} w_i k_i||$ are especially useful to minimize during optimization, as done in the selection of r here, because they both minimize imbalance in $\mathbf{K}$ and a reasonable measure of "multivariate imbalance", i.e. a norm over the different in multivariate densities for the treated and control.

When interpreted as a difference between estimated densities, the L_1 version of this norm described above is very much analogous to the L_1 metric used in Coarsened Exact Matching (Iacus et al., 2011), but without requiring coarsening in order to construct discrete bins in the covariates space.

S3 Unbiasedness for SATT

Theorem 1 states that the weighted difference in means estimator using kernel balancing weights is unbiased for the sample average treatment effect on the treated (SATT) and the (population) ATT.

The SATT is similar to the ATT, but computes the average differences between the treatment and non-treatment potential outcome of the treated units actually sampled, rather than the expectation over the population distribution for the treated. The SATT is thus a more natural immediate target for an estimator.

$$SATT = \frac{1}{N_1} \sum_{i:D_i=1} Y_{1i} - \frac{1}{N_0} \sum_{i:D_i=0} Y_{0i} \tag{S3.1}$$

Recall that the $\widehat{DIM}_w$ is defined as $\frac{1}{N_1} Y_{1i} - \sum_{D=0} w_i Y_{0i}$. Recall also that under the assumption $\mathbb{E}[Y_{0i}|X_i] = \phi(X_i)^\top \theta$ (Assumption 2), $Y_{0i} = \phi(X_i)^\top \theta + \epsilon_i$ for $\mathbb{E}[\epsilon_i|X_i] = 0$.

Hence the error of the $\widehat{DIM}_w$ estimate for the SATT is then

$$\widehat{DIM}_w - SATT = \frac{1}{N_1}\sum_{i:D_i=1} Y_{0i} - \sum_{D_i=0} w_i Y_{0i} \tag{S3.2}$$

$$= \frac{1}{N_1}\sum_{i:D_i=1} \left(\phi(X_i)^\top\theta + \epsilon_i\right) - \sum_{i:D_i=0} w_i \left(\phi(X_i)^\top\theta + \epsilon_i\right) \tag{S3.3}$$

$$= \theta^\top \frac{1}{N_1}\sum_{i:D_i=1} \phi(X_i) + \frac{1}{N_1}\sum_{i:D_i=1} \epsilon_i - \theta^\top \sum_{i:D_i=0} w_i\phi(X_i) - \sum_{i:D_i=0} w_i\epsilon_i \tag{S3.4}$$

$$= \theta^\top \left(\frac{1}{N_1}\sum_{i:D_i=1} \phi(X_i) - \sum_{i:D_i=0} w_i\phi(X_i)\right) + \frac{1}{N_1}\sum_{i:D_i=1} \epsilon_i - \sum_{i:D_i=0} w_i\epsilon_i \tag{S3.5}$$

$$= 0 + \frac{1}{N_1}\sum_{i:D_i=1} \epsilon_i - \sum_{i:D_i=0} w_i\epsilon_i \tag{S3.6}$$

The bias is the expectation of this quantity,

$$bias = \mathbb{E}\left[\widehat{DIM}_w - SATT\right] \tag{S3.7}$$

$$= \mathbb{E}\left[\frac{1}{N_1}\sum_{i:D_i=1} \epsilon_i - \sum_{i:D_i=0} w_i\epsilon_i\right] = 0 \tag{S3.8}$$

S3.1 Remarks

Note that $\mathbb{E}[SATT] = ATT$, and so unbiasedness of $\widehat{DIM}_w$ for the SATT also implies unbiasedness for the ATT.

The assumption that $\mathbb{E}[Y_{0i}|X_i] = \phi(X_i)^\top\theta$ is innocuous as $N \to \infty$, because the universal representation property of the Gaussian kernel ensures that the space of functions spanned by $\phi(X_i)^\top\theta$, which has representation $f(x_i) = \sum_j \alpha_j k(X_j, X_i)$, includes all continuous function. However, in finite samples the quality of the approximation is limited. Imagine the superposition of Gaussians view of

this functions space: with too few observations, there are limits to the shapes that can be built by placing Gaussians at each observation and rescaling them. Even though highly non-linear, non-additive functions can still be well modeled with relatively small samples (see Hainmueller and Hazlett, 2014), we may still wish to know how finite samples behave in terms of potential bias. Suppose that in truth, $\mathbb{E}[Y_{0i}|X_i] = \phi(X_i)^\top\theta + h(X_i) + \epsilon_i$, where $h(X_i)$ is the misspecification error, an additive component that cannot be captured by $\phi(X_i)^\top\theta$ using the sample available and by definition orthogonal to the span of $\phi(X_i)$. In this case, the difference between $\widehat{DIM}_w$ and the SATT becomes

$$\widehat{DIM}_w - SATT = \frac{1}{N_1}\sum_{i:D_i=1}\epsilon_i - \sum_{i:D_i=0} w_i\epsilon_i + \frac{1}{N_1}\sum_{i:D_i=1} h(X_i) - \sum_{i:D_i=0} w_i h(X_i) \tag{S3.9}$$

Notice that bias due to misspecification occurs only if $h(X_i)$ has different means for the treated and controls (after weighting). That is, even if in a small sample $\mathbb{E}[Y_{0i}|X_i]$ cannot be well approximated, this is only problematic if the misspecification error, $h(X_i)$ is correlated with the treatment assignment after adjusting for differences on the other covariates through weighting. This is analogous to the biased caused by omitted variables in regression models.

S4 Illustration of worst-case bound on bias

Next, a simulation is useful to illustrate the effectiveness of the worst-case bound. This involves five steps:

1. Randomly draw a function from the space of functions corresponding to a Gaussian kernel: first choose a set of 100 "knots", $Z_j \sim Unif(0,1)$ for $j \in 1, ..., 100$, then choose a kernel function

$k(\cdot,\cdot)$ (Gaussian with $b = 0.1$), and randomly draw the c vector according to $c_i \sim N(0,1)$ for $i \in 1, ..., 100$. Together these quantities characterize a chosen function in the RKHS of kernel k, $f(\cdot) = \sum_j c_j k(\cdot, Z_j)$.

2. For $N = 25$ and $N = 200$ (see below), randomly draw covariate data $X_i \sim Unif(0,1)$. Generate the values of Y_{0i} under the function drawn above, i.e. $Y_{0i} = \sum_j k(X_i, Z_i)$.

3. Draw treatment status values, D_i, with the probability of treatment relating to the values of $Y_i(0)$. Specifically let $\tilde{y}$ be the centered and standardized $Y_i(0)$, then let $D_i = 1$ with probability $\frac{1}{1+exp(-\tilde{y})}$ and 0 otherwise.

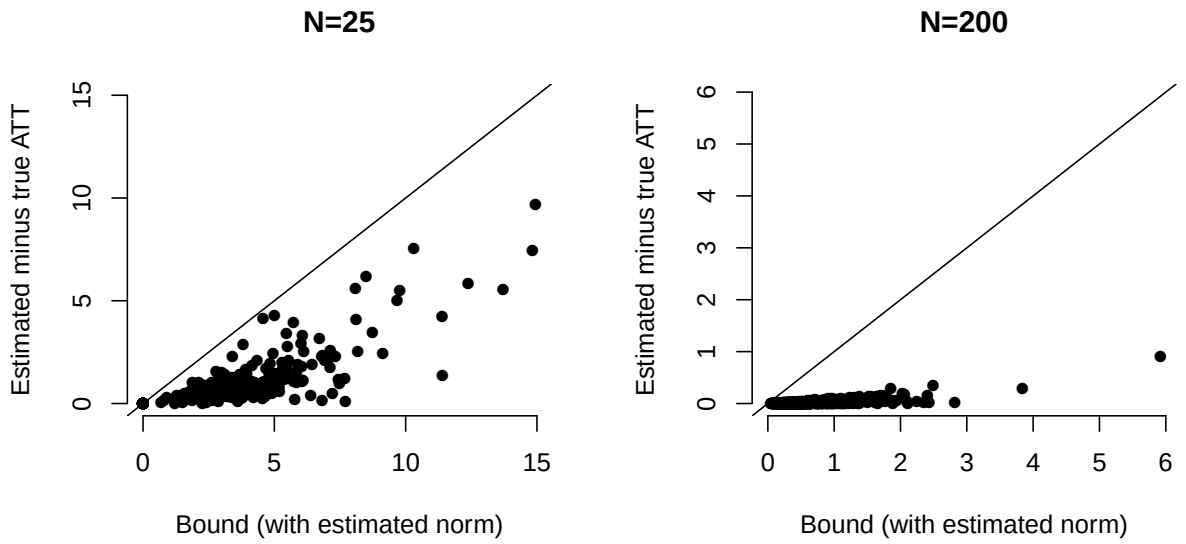
4. Choose treatment effect size, $TE = 1$. Construct the observed outcome, $Y_i = Y(0) + D * TE$.

5. The proposed method is then used to estimate the treatment effect. The estimated treatment effect minus the true one gives the bias due only to incomplete balance, as there is no noise added to the outcome and treatment effects are constant here.

This whole procedure is iterated 1000 times and the bound is computed for each iteration, with an estimate of the Hilbert norm (γ) obtained by using regularized kernel regression with the same kernel as used for kernel balancing (using the `KRLS` package for `R`).[1] Figure 1 shows the result, with N of just 25, and then with N of 200. Only with extremely small sample sizes (i.e. $N = 25$) does this bias come near the bound, and no bias was greater than the bound would allow. As the sample size grows, the bound becomes increasingly conservative. This is to be expected: the bias hits the bound only when

[1]The `KRLS` package uses generalized (leave-one-out) cross-validation to choose the regularization parameter, λ. This in turn determines the Hilbert space norm of the selected regression function, which can be computed using the empirical value of $\hat{c}^\top \mathbf{K} \hat{c}$, providing a reasonable choice of γ that corresponds well to the observed data.

Figure 1: Simulation Illustrating Behavior of Bias Bound



The worst-case bound on the error in the ATT estimate due to balancing on selected eigenvectors $\mathbf{K}$ rather than the entire matrix. Each of 200 iterations draws a new function at random from the RKHS, a new set of data at which to evaluate it, and a new estimate of the ATT using the method proposed here. The horizontal axis shows the estimated bound on this error from each simulation. The vertical axis shows the actual error. The required Hilbert norm, γ, is estimated using `KRLS`, in which leave-one-out cross validation is used to determine the appropriate complexity of the function. *Left:* All errors are less than the estimated bound, however with the very small sample size of 25, some errors come near the bound. *Right:* With $N = 200$, the bound on the error becomes highly conservative.

the coefficients in the eigenvector space (d) happen to be perfectly aligned with the imbalances on the eigenvectors. Such an alignment becomes increasingly unlikely as the sample size, and thus number of eigenvectors, grows.[2]

S5 Balance on an approximation of K

The main text focuses on minimizing the worst-case bias due to imperfect balancing as the rationale for achieving balance on eigenvectors with larger eigenvalues. A closely related justification begins by seeking a lower-dimensional approximation of $\mathbf{K}$ that is most similar in some respect.

[2]Future work could fruitfully derive less extreme values, such as the expectation of this bias. This however would likely require further assumptions over the probability distribution of functions in the ball of the Reproducing Kernel Hilbert Space carved out by γ.

Suppose we have rank-r approximation to $\mathbf{K}$, $\widetilde{\mathbf{K}}^{(r)}$. We might seek the $\widetilde{\mathbf{K}}^{(r)}$ closest to $\mathbf{K}$ in the Frobenius norm, i.e. minimizing

$$||\mathbf{K} - \widetilde{\mathbf{K}}^{(r)}||_{\mathcal{F}} = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N} |\mathbf{K}_{i,j} - \widetilde{\mathbf{K}}^{(r)}_{i,j}|^2}$$

. Alternatively, and closely related to *biasbound*, recall that our aim in achieving mean balance on $\mathbf{K}$ is to ensure that any linear projection $\mathbf{K}c$ for some $N \times 1$ vector c has equal means in the two groups. In choosing a rank r approximation, we thus want to ensure that for c of a particular size $||c||$, $\widetilde{\mathbf{K}}^{(r)}c$ and $\mathbf{K}c$ are as close as possible. Thus, it is desirable to minimize the operator 2-norm, $||\mathbf{K} - \widetilde{\mathbf{K}}^{(r)}||_2 = sup\ \frac{||\mathbf{K}c - \widetilde{\mathbf{K}}^{(r)}c||_2}{||c||_2}$. Among all rank r matrices, the choice of $\widetilde{\mathbf{K}}^{(r)}$ minimizing both $||\mathbf{K} - \widetilde{\mathbf{K}}^{(r)}||_2$ and $||\mathbf{K} - \widetilde{\mathbf{K}}^{(r)}||_{\mathcal{F}}$ is given by principal components analysis (PCA; Eckart and Young, 1936). Since PCA constructs $\widetilde{\mathbf{K}}^{(r)}$ as a linear projection of the first r eigenvalues $\mathbf{K}$, we need not actually work with the projected approximation $\widetilde{\mathbf{K}}^{(r)}$ – we can simply work directly with the eigenvectors themselves. That is, the eigenvectors of $\mathbf{K}$ (which we obtain here using singular value decomposition rather than PCA here) provides a new set of bases, and we will seek balance on the first r of them (as ordered by the corresponding eigenvalues). This provides an N by r matrix of orthonormal bases for which we attempt to make the control group have the same mean for the treated by weighting.

S6 Balance in $\mathbb{E}[\phi(X_i)]$ implies balance in $\mathbb{E}[Y_{0i}]$

The main text focuses principally on SATT estimation, and the implications of obtaining balance on $\phi(X_i)$ in the finite sample. However working with populations instead, we note that obtaining

$\mathbb{E}[\phi(X_i)|D_i = 1] = \mathbb{E}_w[\phi(X_i)|D_i = 0]$ also implies $\mathbb{E}[Y_{0i}|D_i = 1] = \mathbb{E}_w[Y_{0i}|D_i = 0]$, where $\mathbb{E}_w[\cdot]$ designates an expectation taken over the w-weighted distribution of X:

$$\mathbb{E}[Y_{0i}|D = 1] = \mathbb{E}_x\left[\mathbb{E}[Y_{0i}|X, D = 1]\right] \tag{S6.10}$$

$$= \theta^\top \int \phi(x)p(x|D = 1)dx \tag{S6.11}$$

$$= \theta^\top \mathbb{E}[\phi(x)|D = 1] \tag{S6.12}$$

$$\mathbb{E}_w[Y_{0i}|D = 0] = \mathbb{E}_{w,x}\left[\mathbb{E}[Y_{0i}|X, D = 0]\right] \tag{S6.13}$$

$$= \theta^\top \int \phi(x)wp(x|D = 0)dx \tag{S6.14}$$

$$= \theta^\top \mathbb{E}_w[\phi(x)|D = 0] \tag{S6.15}$$

Hence when balance of $\phi(X_i)$ for the treated and controls holds in expectations, we will have $\mathbb{E}[Y_{0i}|D_i = 1] = \mathbb{E}_w[Y_{0i}|D_i = 0]$, allowing a (weighted) difference in means to unbiasedly estimate the ATT.

S7 Proof of proposition 1

Proposition 1 states that for a density estimator for the treated, $\hat{f}_{X|D=1}$, and for the (weighted) controls, $\hat{f}_{X|D=0,w}$, both constructed with kernel k with scale b, the choice of weights that ensures mean balance in the kernel matrix $\mathbf{K}$ also ensures $\hat{f}_{X|D=1} = \hat{f}_{X|D=0,w}$ at every location in $\mathcal{X}$ at which an observation is located.

As detailed in the main text, the expression $\frac{1}{N_1\sqrt{2\pi b}}K_t\mathbf{1}_{N_1}$ places a multivariate standard normal density over each *treated* observation, sums these to construct a smooth density estimator at all points in $\mathcal{X}$, and evaluates the height of that joint density estimate at each of the points found in the dataset. Likewise, $\frac{1}{N_0\sqrt{2\pi b}}K_c\mathbf{1}_{N_0}$ estimates the density of the control units and returns its evaluated height at every datapoint in the dataset.

To reweight the controls would be to say that some units originally observed should be made more or less likely. This is achieved by changing the numerator of each weight $\frac{1}{N_0\sqrt{2\pi b}}$ to some non-negative value other than 1. Letting the weights sum to 1 (rather than N_0), the reweighted density of the controls would be evaluated at each point in the dataset according to $\frac{1}{\sqrt{2\pi b}}K_c w$, for vector of weights w. If weights are selected so that this equals the density of the treated:

$$\begin{aligned}
\frac{1}{N_1\sqrt{2\pi b}}\mathbf{K_t}\mathbf{1}_{\{N_1\}} &= \frac{1}{\sqrt{2\pi b}}\mathbf{K_c}w \\
\frac{1}{N_1}\mathbf{K_t}\mathbf{1}_{\{N_1\}} &= \mathbf{K_c}w \\
\overline{K_t} &= \mathbf{K_c}w \\
\overline{K_t} &= \overline{K_c(w)}
\end{aligned} \tag{S7.16}$$

where the final line is the definition of mean balance in $\mathbf{K}$. Thus, the weights that achieve mean balance in $\mathbf{K}$ are precisely the right weights to achieve equivalence of the measured multivariate densities for the treated and controls at all points in the dataset.

S8 Derivation of $\phi(X_i)$ for Gaussian Kernel

While the functions linear in $\phi(X_i)$ corresponding to a Gaussian kernel can more easily be understood as those that can be formed by superposing Gaussian kernels over the observations, one may also explicitly construct features $\phi(X_i)$ consistent with the requirement that $K(X_i, X_j) = \langle \phi(X_i).\phi(X_j) \rangle$ for the standard inner-product. One simple approach is, setting $b = .5$ for convenience, yields:

$$k(X_i, X_j) = exp(-||X_i - X_j||^2/1) \tag{S8.17}$$

$$= exp(-X_i^2)exp(-X_j^2)exp(2X_i, X_j) \tag{S8.18}$$

$$= exp(-X_i^2)exp(-X_j^2)\sum_{d=0}^{\infty} \frac{2^d X_i^d X_j^d}{d!} \tag{S8.19}$$

where the last line follows by a Taylor series expansion of $exp(2X_iX_j)$. Finally the division of terms can be completed, as:

$$k(X_i, X_j) = \sum_{d=0}^{\infty} \sqrt{\frac{2^d}{d!}} exp(-X_i^2 X_i^d) \sqrt{\frac{2^d}{d!}} exp(-X_i^2 X_i^d) \tag{S8.20}$$

This is simply an inner product of two infinite-dimensional vectors of the form

$$\phi(X_i) = \left[\sqrt{\frac{2^0}{0!}} exp(-X_i^2 X_i^0), \;\; \sqrt{\frac{2^1}{1!}} exp(-X_i^2 X_i^1), \;\; ..., \;\; \sqrt{\frac{2^\infty}{\infty!}} exp(-X_i^2 X_i^\infty) \right] \tag{S8.21}$$

Figure 2 considers a one dimensional covariate, X, and shows what value each of the first 5 of these

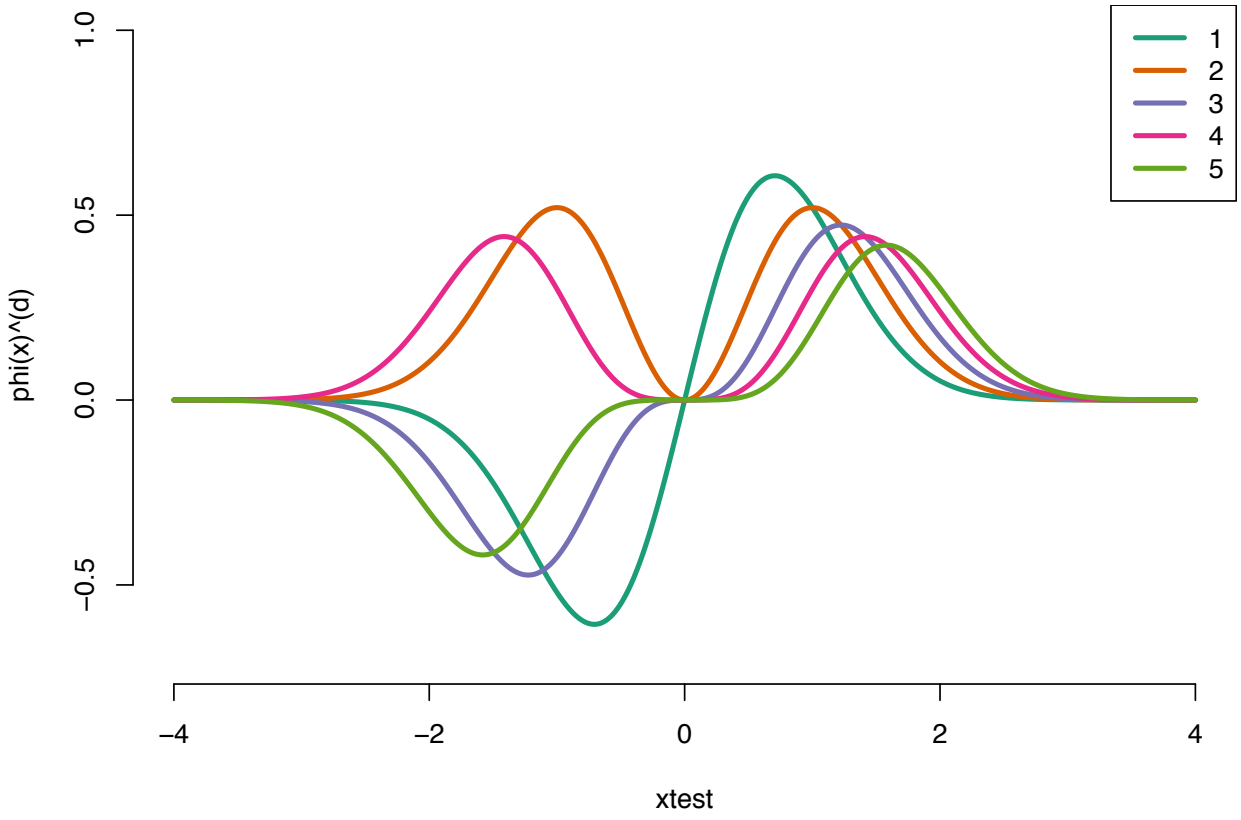
features would have at various values of X.

S9 Density Equalization Illustration

This example visualized the density estimates produced internally by kernel balancing using linear combinations of $\mathbf{K}$ as described above. Suppose X contains 200 observations from a standard normal distribution. Units are assigned to treatment with probability $1/(1 + exp(2 - 2X))$, which produces approximately 2 control units for each treated unit. Figure 3 shows the resulting density plots, using density estimates provided by `kbal` in which the density of the treated is given by $\frac{1}{N_1\sqrt{2\pi b}}\mathbf{K_t}\mathbf{1}_{N_1}$ and the density of the controls is given by $\frac{1}{N_0\sqrt{2\pi b}}\mathbf{K_c}\mathbf{1}_{N_0}$. As shown, the density estimates for the treated at each observations X position (black squares) is initially very different from the density estimates for the controls taken at each observation (black circles). After weighting, however, the new density of the controls as measured at each observation (red `x`) matches that of the treated almost exactly.

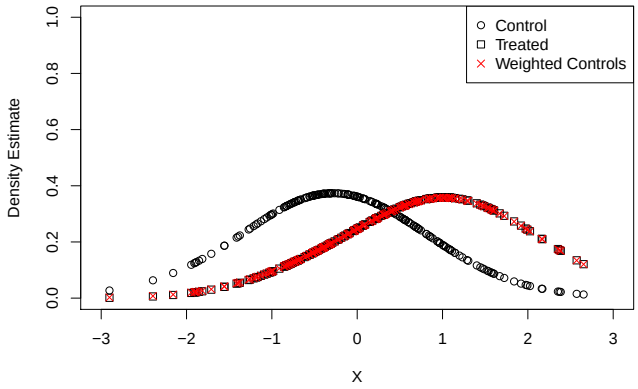
Note that in multidimensional examples, the density becomes more difficult to visualize across each dimension, but it is still straightforward to compute and to think about the pointwise density estimates for the treated or control as measured at each observation's X value. In contrast to binning approaches such as CEM, equalizing density functions continuously in this way avoids difficult or arbitrary binning decisions, is tolerant of high dimensional data, and smoothly matches the densities in a continous fashion, resolving the within-bin discrepancies implied by CEM.

Figure 2: First five values of $\phi(X)$ at varying values of X



Explicit view of $\phi(X_i)$ for one choice of $\phi(X_i)$ consistent with $K(X_i, X_j) = \langle \phi(X_i), \phi(X_j) \rangle$ for a Gaussian kernel K as described in Equation S8.

Figure 3: Density-Equalizing Property of Kernel Balancing



Plot showing the density-equalization property of kernel balancing. For 200 observations of $X \sim N(0,1)$, treatment is assigned according to $Pr(treatment) = 1/(1 + exp(2 - 2X))$, producing approximately two control units for each treated unit. Black squares indicate the density of the treated, as evaluated at each observation's location in the dataset (and given the choice of kernel and b). Black circles indicate the density of (unweighted) controls. The treated and control are seen to be drawn from different distributions, owing to the treatment assignment process. Red x's show the new density of the controls, after weighting by `kbal`. The reweighted density is nearly indistinguishable from the density of the treated, owing to the density equalization property of kernel balancing.

S10 Inverse Propensity Score Weights as Multivariate Density Equalization

It is useful to show more explicitly the role played by inverse propensity score weights in estimating the ATT, as this leads to an appreciation of how these weights relate to multivariate density equalization, and the sense in which they are equivalent to the kernel balancing weights despite flowing from different initial goals.

Under Assumption 1, the ATT can be re-written:

$$ATT = \mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 1] \tag{S10.22}$$

$$= \int \mathbb{E}[Y_{1i}|D_i = 1, x]p(x|D_i = 1)dx - \int \mathbb{E}[Y_{0i}|D_i = 1, x]p(x|D_i = 1)dx \tag{S10.23}$$

$$= \int \mathbb{E}[Y_{1i}|D_i = 1, x]p(x|D_i = 1)dx - \int \mathbb{E}[Y_{0i}|D_i = 0, x]p(x|D_i = 1)dx \tag{S10.24}$$

Expression S10.24 is identifiable in the sense that we only require treatment potential outcomes from the treated units, and non-treatment potential outcomes from the non-treated units. However, it remains problematic because it requires averaging outcomes from control units over the distribution of X for the treated, $p(x|D_i = 1)$, which is not the distribution of the control units in the sample. Specifically, the difference in means estimand,

$$\begin{aligned} \text{DIM} &= \mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0] && \text{(S10.25)} \\ &= \int \mathbb{E}[Y_{1i}|D_i = 1, x]p(x|D_i = 1)dx - \int \mathbb{E}[Y_{0i}|D_i = 1, x]p(x|D_i = 0)dx && \text{(S10.26)} \end{aligned}$$

differs from the ATT in its second term, because it averages over the outcomes of non-treated units at their natural density in X, $p(x|D_i = 0)$. To address this, consider a weighted difference in means estimand,

$$\begin{aligned} \text{DIM}_w &= \mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}_w[Y_{0i}|D_i = 0] && \text{(S10.27)} \\ &= \int \mathbb{E}[Y_{1i}|D_i = 1, x]p(x|D_i = 1)dx - \int w_i\mathbb{E}[Y_{0i}|D_i = 1, x]p(x|D_i = 0)dx && \text{(S10.28)} \end{aligned}$$

where w_i is a function of X that allows us to upweight or downweight control units. The difference between expression S10.24 and S10.26 can be resolved by choosing weights on the control units,

$$w_i = \frac{p(x|D_i = 1)}{p(x|D_i = 0)} \tag{S10.29}$$

Through Bayes theorem, we can replace the class densities in this expression with more familiar

propensity scores to obtain $w_i = \frac{p(D_i=1|x)p(D_i=0)}{p(D_i=0)|x)p(D_i=1)}$. Since $D_i = 0$ for all units given weights, this is $w_i = \frac{p(D_i)}{p(D_i|X_i)} \frac{1-p(D_i|X_i)}{1-p(D_i)}$. These are the stabilized inverse propensity scores one would apply to the control units to estimate the ATT. If properly estimated, these would ensure that the whole distribution of X for the control units is adjusted to equal the distribution among the treated.

Note that in the form S10.29, it becomes clear that were we to adjust the sample to make treated and control groups have the same distribution of covariates, these weights would become constant and thus unnecessary. This is achieved, insofar as the smoothed multivariate densities on which kernel balancing obtains balance are reasonable approximations of the true densities. In this sense, kernel balancing achieves the goals of inverse propensity score weighting, but has the advantage of avoiding any functional form assumption or direct estimation of the propensity score.

S11 Optional Trimming of the Treated

In some cases, balance can be greatly improved with less variable (and thus more efficient) weights if the most difficult-to-match treated units are trimmed. In estimating an ATT, control units in areas with very low density of treated units can always be down-weighted (or dropped if the weight goes to zero), but treated units in areas unpopulated by control units pose a greater problem. These areas may prevent any suitable weighting solution, or may place extremely large (and thus ineffecient) weights on a small set of controls.

While estimates drawn from samples in which the treated are trimmed no longer represent the ATT with respect to the original population, they can be considered a local or sample average treatment effect within the remaining population. King et al. (2011) refer similarly to a "feasible sample average

treatment effect on the treated" (FSATT), based on only the treated units for which sufficiently close matches can be found. In any case, the discarded units can be characterized to learn how the inferential population has changed.

However, even when the investigator is willing to change the population of interest by trimming the treated, it is not always clear on what basis trimming should be done. In kernel balancing, trimming of the treated can be (optionally) employed by using the multivariate density interpretation given above. Specifically, the density estimators at all points is constructed using the kernel matrix. Then, treated units are trimmed if $\frac{p_{X|D=1}(x_i)}{p_{X|D=0}(x_i)}$ exceeds the parameter *trimratio*. The value of *trimratio* can be set by the investigator based on qualitative considerations, inspection of the typical ratio of densities, a willingness to trim up to a certain percent of the sample, or performance on L_1. Whatever approach is taken to determine a suitable level of *trimratio*, `kbal` produces a list of the trimmed units, which the investigator can examine to determine how the inferential population has changed.

S12 Detailed Choice of Kernel

Using the kernel as defined by 7 for some choice of b, any continuous function $\mathbb{E}[Y_{0i}|X_i]$ can be consistently estimated by functions linear in $\phi(X_i)$. However, some kernel choices work better than others in a sample of limited size. Accordingly, in machine learning applications utilizing kernels, it is common to consider details of the kernel definition that may improve the ability to fit the target function linearly in $\phi(X_i)$ (or equivalently, the columns of $\mathbf{K}$) when the sample size is limited.

The first consideration of this type is how X is scaled and rotated. If some variables in X_i have variances orders of magnitude larger than others, the columns of $\mathbf{K}$ will reflect mostly distances on the

highest-variance variables, providing little information on distances among the smaller variables. This is unproblematic as the sample size grows to infinity – the superposition of Gaussians will still allow flexible modeling of the target functions in the limit. But in a small sample, it limits the quality of fit. It is thus common to utilize a Gaussian kernel that computes the Euclidean distance over variables that have been rescaled to have the same variance. This also has the benefit of making the results invariant to any unit-of-measure decisions. Kernel balancing utilizes this approach. Beyond this, some investigators also wish to make the results invariant to rotation, utilizing a Mahalanobis distance rather than Euclidean distance in the Gaussian kernel. This is left as an option in kernel balancing as implemented here.

Second, b must be chosen. Since mean balance on Y_{0i} is the primary goal, not density estimation or equalization, the choice of the kernel and b should be made accordingly. While it is tempting to think of b as the usual bandwidth that must be carefully selected in density estimation procedures, here the choice of parameter b is understood first as a feature-extraction decision that determines the construction of $\phi(X_i)$ and thus $\mathbf{K}$. It determines how close two points X_i and X_j need to be in order to have highly similar rows K_i and K_j. The choice of b also has implications in terms of a bias-variance tradeoff and feasibility: if b is too large, mean balance is easier to achieve and the weights will typically have lower variance, the resulting balance is less precise (and the corresponding smoothed densities more "blurred"). At the extreme as b approaches infinity, $\mathbf{K}$ approaches a matrix with 1 in every position, indicating that all observations are alike and nothing needs to be done to obtain balance. In the opposite extreme, as b becomes very small, $\mathbf{K}$ begins to approximate the identity matrix. In this case, the algorithm will not converge as balance cannot be attained. (The possibility of trimming away treated units that are difficult to match under small b is discussed in Supplement S11). The interesting

cases lie in between these extremes, where choices can be made to "blur" the features more or less in order to make balance easier to achieve on more dimensions of $\mathbf{K}$. Note that standard matching and weighting methods typically involve a bias-variance tradeoff as well, though it may be implicit or difficult to manipulate directly. For example, in matching, the number of control units matched to each treated unit, as well as the choice of caliper, and of course the choice of how many covariates to match on all have implications for the bias-variance tradeoff. In Coarsened Exact Matching Iacus et al. (2011), the size of the bins used to coarsen each covariates have direct bias-variance implications. King et al., 2017 usefully discusses the related "balance vs. sample-size frontier". Kallus (2016) discusses the bias-variance tradeoff in optimal matching procedures and the assumptions under which a mean squared error criterion is minimized by various procedures. In related weighting methods such as Hainmueller (2012), there is no direct control of the bias-variance tradeoff except implicitly through the set of covariates one is seeking (exact) mean balance on[3] Likewise, if one uses a propensity score model to choose inverse propensity score weights, for example, the bias-variance implications of those models are difficult to control.

Because there is no "right answer" as to what b should be, I provide here three guidelines for transparent reporting. First, a useful reporting standard would be to provide results at $b = dim(X)$, while also showing results at other choices for robustness. The reason to choose $b = dim(X)$ (the default value used above) is that the square of $\mathbb{E}[||X_i - X_j||]$, used in the exponent of the kernel calculation (7) scales with $dim(X)$. Choosing b proportional to $dim(X)$ thus ensures a relatively sound scaling of the data, such that some observations appear to be closer together, some further apart, and some

[3]In principle, setting the tolerance or stopping point for convergenge of the algorithm, or other procedures, could be added to allow a measure of control.

in-between, regardless of $dim(X)$. A similar logic has been proposed for regression technique using a Gaussian kernel (see e.g. Hainmueller and Hazlett, 2014; Schölkopf and Smola, 2002). The constant of proportionality remains open to debate, but the choice of $b = dim(X)$ has offered good performance in most cases (though higher values tend to perform more reliably when balance is very difficult to achieve, as in the National Supported Work example). Second, the degree to which weights become large and uneven should be reported. I propose the quantity *min90*, which is the minimum number of control units that are required to account for 90% of the total weight among the controls. For example, if *min90*=20, 90% of the total weight of the controls comes from just the 20 most heavily-weighted observations. This gives the user a sense of how many control units are effectively being used. The National Supported Work example reported above demonstrates this. Third, investigators may wish to present their results across a range of b values to ensure this choice is not consequential in a given application (see King et al., 2017 for a related proposition regarding matching estimators and the "balance/sample-size" tradeoff). Should the results vary across b values, inspecting L_1 and the concentration of weights (e.g. through $min90$) can be helpful for understanding the bias-variance consequences of a given choice.

Fortunately, the choice of b is easy to address in many cases because a wide range of b values often allow large improvements in L_1 paired with stable ATT estimates. Following the recommendation above to show estimates at $b = 1p$ and across a rage of b values, Supplement S13 shows ATT estimates for kernel balancing with the standard covariate set in the National Supported Work empirical benchmark (Section 4). While there is some variation in estimates when b is small ($2p$ or less), the estimates stabilize above that to values of over $50p$. Moreover, despite the potential for a bias-variance tradeoff, when good balance can be achieved on even the smaller b values without resorting to extreme weights,

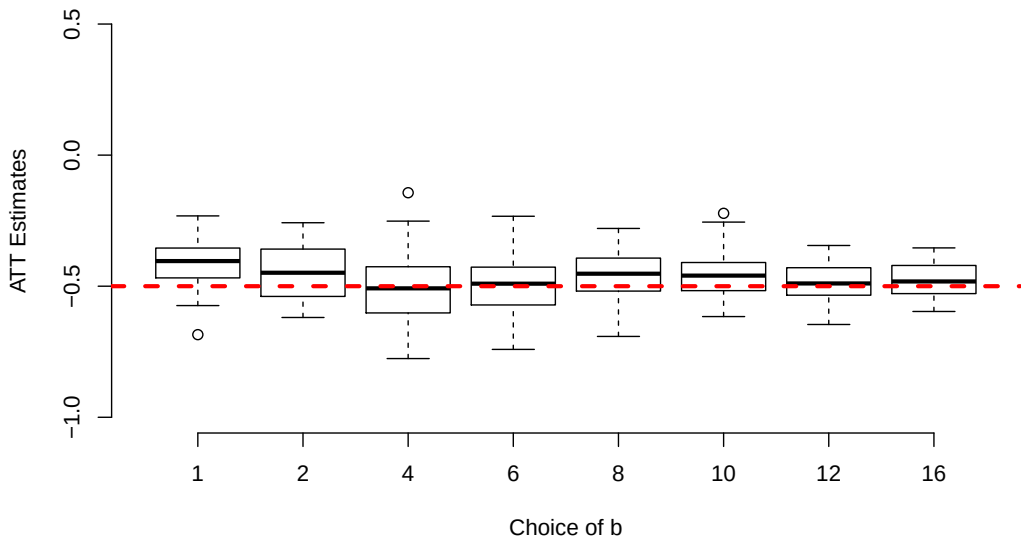
then the variability of ATT estimates (i.e. under resampling) can remain stable across a wide range of b values. Supplement S13 also shows boxplots of ATT estimates for the simulated example (Section 3.1), showing both low bias and stable variance across the range of attempted b values.

S13 Stability Across b in simulation and empirical example

As described in the text, kernel balancing is the only method of those attempted that approaches unbiasedness in estimating the simulated effect of peacekeeping when a non-linear function of the covariates was confounding. The only parameter that must be chosen by the user is b, though `kbal` provides a default of $b = dim(X)$. In Figure 4, we see that this result is largely insensitive to the choice of b ranging from one-quarter to four times the default. If anything, ATT estimates improve with b somewhat above the default, though setting b larger can come at the cost of more extreme weights in some natural datasets where overlap in the covariate distributions may not be as good.

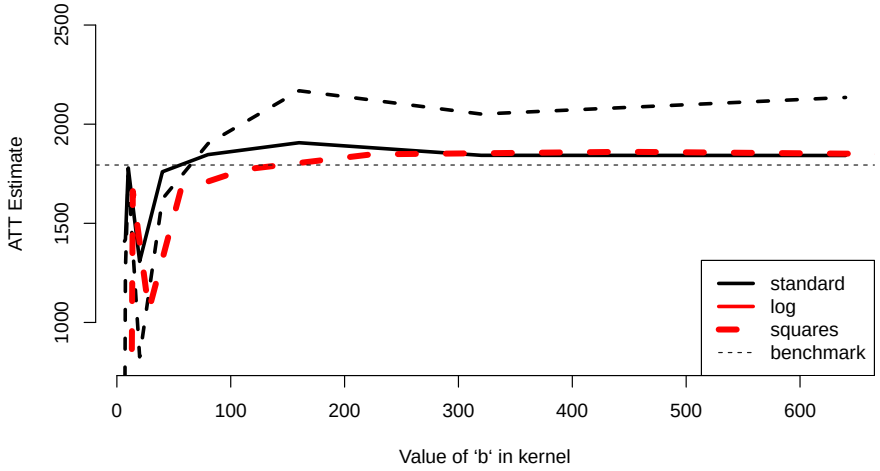
We also examine sensitivity of results to the choice of b in the National Supported Work benchmark example. As described in the text, kernel balancing is used in three ways: with the original set of 10 covariates used previously (*standard*), with a set of covariates that replaces the income variables with their logs (*log*), and with a set that, for the three continuous variables, includes their squares (*squares*). At the defaults values of b (the number of covariates), each estimate does well as shown in the text. Figure 5 shows the results for all three specifications over a wide range of b values. While there is some instability at low levels of b – due to the difficult of achieving balance in this example – past that point, the ATT estimates at the chosen weights are extremely stable and accurate. The guidelines suggested here is that investigators routinely report their estimates across a range of b estimates to avoid selective

Figure 4: Simulation: sensitivity to choice of p



Boxplot illustrating distribution of average treatment effect on the treated (ATT) estimates using *kernel balancing*, as the bandwith parameter b is varied. At each value of b, 50 simulations are used, each drawing a separate dataset from the same data generating process. The default choice of b is $dim(X) = 4$, with results here shown from one-quarter to four times that value. The actual (population) ATT is -0.5, indicated by the dashed line. Results show very low bias at all values of b.

Figure 5: Stability of National Supported Work estimates to choice of b



Stability of ATT estimates for *kbal* across different choices of b in the National Supported Work Example. The *standard* estimates employed all 10 of the covariates provides in the Dehejia & Wahba dataset Dehejia and Wahba (1999) and typically used for this task. The *log* specification makes the choice to replace the income variables with their logs (plus one), and the *squares* model add squared terms for the continuous variables. While there is some instability at small choices of b, the results are remarkable stable and close to the benchmark at higher values of b.

reporting of results.

S14 Additional Example: Are Democracies Inferior Counterinsurgents?

Decades of research in international relations has argued that democracies are poor counterinsurgents (see Lyall, 2010 for a review). Democracies, as the argument goes, are (1) sensitive to public backlash against wars that get more costly in blood or treasure than originally expected, (2) are unable to control the media in order to supress this backlash, and (3) often respect international prohibitions on brutal tactics that may be needed to obtain a quick victory. Each of these makes them more

prone to withdrawal from countinsurgency operations, which often become long and bloody wars of attrition. Empirical work on this question was significantly advanced by Lyall (2010), who points out that previous work (1) often examined only democracies rather, than a universe of cases with variation on polity type, and (2) did little to overcome the non-random assignment of democracy, and particular, the selection effects by which democracies may choose to fight different types of counterinsurgencies than non-democracies.

Lyall (2010) overcomes these shortcomings by constructing a dataset covering the period of 1800-2005, in which the polity type of the countinsurgent regimes vary. Matching is then used to adjust for observable differences between the conflicts selected by democracies and non-democracies, using one-to-one nearest neighbor matching on a series of covariates. These covariates are: a dummy for whether the counterinsurgent is an occupier (*occupier*), a measure of support and sanctuary for insurgents from neighboring countries (*support*), a measure of state power (*power*), mechanization of the military (*mechanized*), *elevation*, *distance* from the state capital to the war zone, a dummy for whether a state is in the first two years of independence (*new state*), a *cold war* dummy, the number of *languages* spoken in the country, and the *year* in which the conflict began.

In a battery of analyses with varying modeling approaches, Lyall (2010) finds that democracy, measured as a polity score of at least 7 in the specifications replicated here, has no relationship to success or failure in counter insurgency, either in the raw data or in the matched sample.

While the credibility of this estimate as a causal quantity depends on the absence of unobserved confounders, we can nevertheless assess whether the procedures used to adjust for observed covariates were sufficient, or whether an inability to achieve mean balance on some functions of the covariates may

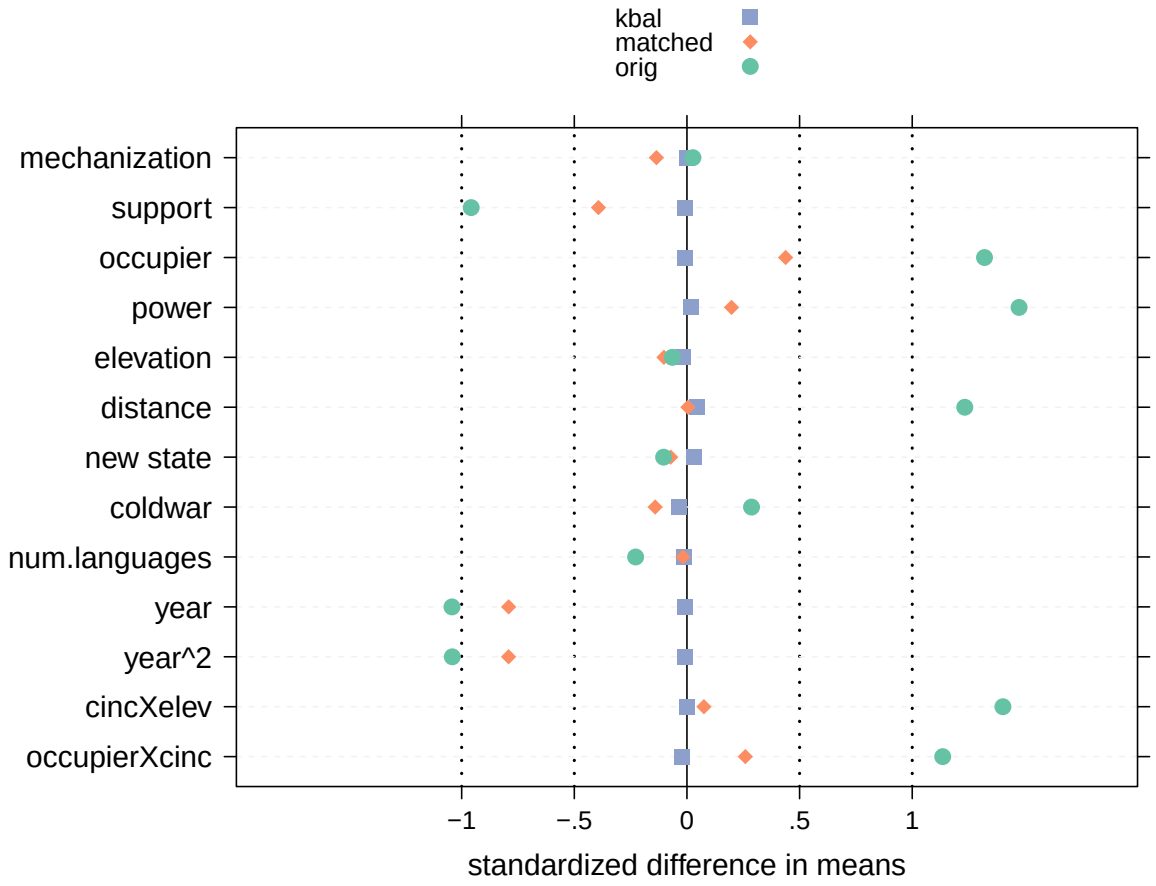
have led to bias even in the absence of unobserved confounders.

Here I reexamine these findings using the post-1945 portion of the data, which includes 35 counterinsurgencies by democracies and 100 by non-democracies, and is used in many of the analyses in Lyall (2010). The 1945 period is the only one with complete data on the covariates used for balancing here, but is also the period in which the logic of democratic vulnerability is expected to be most relevant.

First, I assess balance. As shown in Figure 6, numerous covariates are badly imbalanced in the original dataset (circles), where imbalance is measured on the x-axis by the standardized difference in means. This balance improves somewhat under matching (diamonds), but improves far more under kernel balancing (squares). Note that imbalance is shown both on the variables used in the matching/weighting algorithms (the first ten covariates up to and including *year*), as well as several others that were not explicitly included in the balancing procedure: $year^2$, and two multiplicative interactions that were particularly predicted of treatment status in the original data. Kernel balancing produces good balance on both the included covariates, and functions of them.

Next, I use the matched and weighted data to estimate the effect of democracy on counterinsurgency success. For this, I simply use linear probability models (LPM) to regress a dummy for victory (1) or defeat (0) on covariates according to five different specifications. While Lyall (2010) used a number of other approaches, including logistic regression, some of these models suffer "separation" under the specifications attempted here. This causes observations and variables to effectively drop out of the analysis, producing variability in effect estimates that are due only to this artefact of logistic regression and not due to any meaningful change in the relationship among the variables. Linear models do not suffer this problem, and provide a well defined approximation to the conditional expectation function,

Figure 6: Balance: Democracies vs. Non-democracies and the Counterinsurgencies they Fight



Balance in post-1945 sample of Lyall (2010). Imbalance, measured as the difference in means divided by the standard deviation, is shown on the x- axis. Democracies (treated) and non-democracies (controls) vary widely on numerous covariates. The matched sample (diamonds) shows somewhat improved balance over the original sample, but imbalances remain on numerous characteristics. Balance is considerably improved by kernel balancing (squares). The rows at or above *year* show imbalance on characteristics explicitly included in the balancing procedures. Those below *year* show imbalance on characteristics not explicitly included.

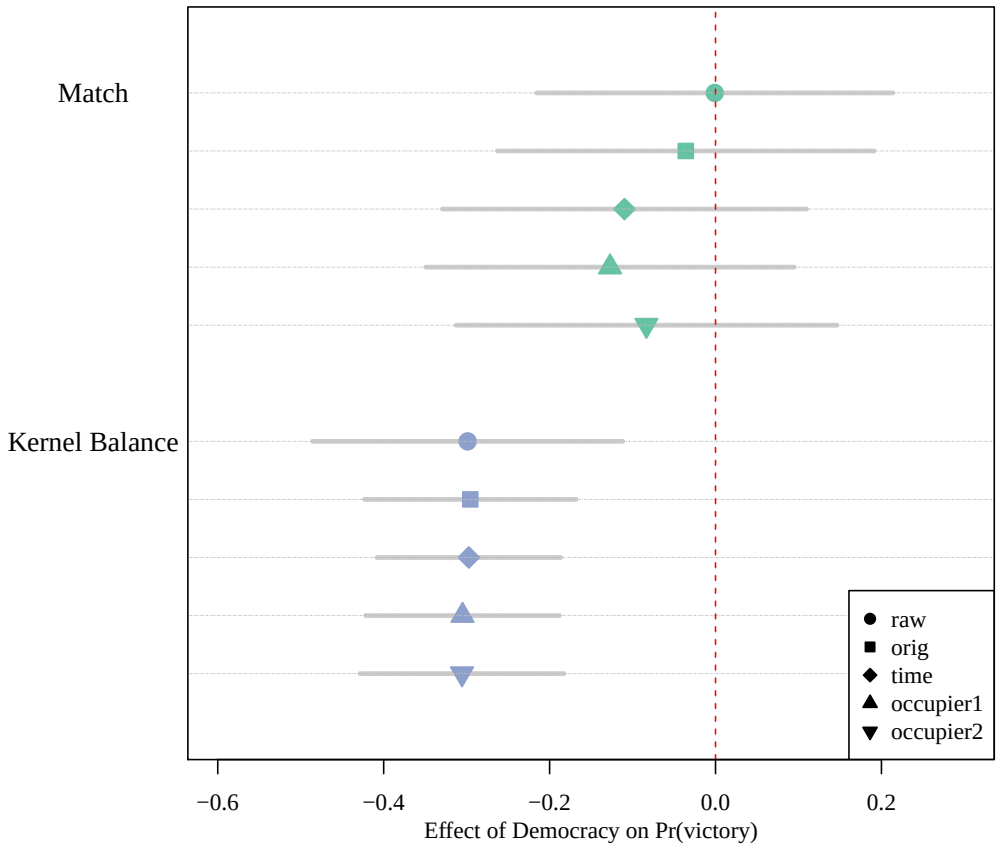
allowing valid estimation of the changing probability of victory associated with changes in the treatement variable, *democracy*. The first three specifications used are (1) *raw* regresses the outcome directly on *democracy* without covariates (and is equivalent to difference-in-means);(2) *orig* uses the same covariates as Lyall (2010), which are all those variables balanced on except for *year*, (3) *time* reincludes *year* as well as $year^2$ to flexibly model the effects of time. The final two models, *occupier1* (4) and *occupier2* (5), add flexibility by including interactions of *occupier* with other variables in the model. These interactions were chosen because analysis with KRLS revealed that interactions with *occupier* were particularly predictive of the outcome.

Figure 7 shows results for the matched and kernel balanced samples with 95% confidence intervals. Under matching, the effect varies considerably depending on the choice of model. No estimate is significantly different from zero, however. In stark contrast, kernel balancing producing estimates that are essentially invariant to the choice of model. Each kernel balancing estimate is between -0.26 and -0.27, indicating that democracy is associated with a 26 to 27 percentage point lower probability of success in fighting counterinsurgencies. This is a very large effect, both statistically and substantively, given that the overall success rate is only 33% in the post-1945 sample.

S14.1 Are democracies more selective?

One puzzle regarding the claim that democracies are inferior counterinsurgents has been why democracies, whatever their weaknesses as counerinsurgents, are not also better able to "select into" conflicts they are more likely to win. The same qualities that are theorized to make democracies more susceptible to defeat against insurgents – public accountability and media freedoms – might also push democracies

Figure 7: Effect of Democracy on Counterinsurgency Success



Effect of democracy on counterinsurgency success in post-1945 sample of Lyall (2010) using matching or kernel balancing for pre-processing followed by five different estimation procedures. Under matching, effect estimates remain highly variable, but none are significantly difference from zero. Kernel balancing shows remarkably stable estimates over the five estimation procedures, even when no covariates are included (*raw*). Results from kernel balancing are consistently in the -0.26 to -0.27 range and significantly different from zero, indicating that democracy is associated with a substantively large deficit in the ability to win counterinsurgencies.

to more carefully select what counterinsurgency operations they engage in.

The findings suggest that such a selection may occur. Specifically, the naive effect estimate obtained by a simple difference in mean probability of victory (on the unweighted sample) is -0.10 ($p = 0.13$). Recall that this difference in means can be decomposed,

$$\begin{aligned}\mathbb{E}[Y(1)|D=1] - \mathbb{E}[Y(0)|D=0] &= \mathbb{E}[Y(1)|D=1] - \mathbb{E}[Y(0)|D=1] + \mathbb{E}[Y(0)|D=1] - \mathbb{E}[Y(0)|D=0] \\ &= ATT + [\mathbb{E}[Y(0)|D=1] - \mathbb{E}[Y(0)|D=0]]\end{aligned}$$

That is, the naive difference in means is the average treatment effect on the treated (had they fought in the same types of cases), plus a selection effect indicating how democracies and non-democracies differ in their probabilities of victory based only on fighting different types of cases (i.e. in the absence of any effect of democracy). Since we know the ATT estimate and the raw difference in means, we can estimate the selection effect to be about 17 percentage points more likely to end in victory. While simple, this decomposition suggests that democracies do choose counterinsurgencies somewhat "wisely", but are also less likely to win a given a counterinsurgency once this selection is accounted for.

References

Cressie, N. and Read, T. R. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 440–464.

Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448):1053–1062.

Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46.

Hainmueller, J. and Hazlett, C. (2014). Kernel regularized least squares: Reducing misspecification bias with a flexible and interpretable machine learning approach. *Political Analysis*, 22(2):143–168.

Iacus, S. M., King, G., and Porro, G. (2011). Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association*, 106(493):345–361.

Kallus, N. (2016). Generalized optimal matching methods for causal inference. *arXiv preprint arXiv:1612.08321*.

King, G., Lucas, C., and Nielsen, R. A. (2017). The balance-sample size frontier in matching methods for causal inference. *American Journal of Political Science*, 61(2):473–489.

King, G., Nielsen, R., Coberley, C., Pope, J. E., and Wells, A. (2011). Comparative effectiveness of matching methods for causal inference. *Unpublished manuscript*, 15.

Lyall, J. (2010). Do democracies make inferior counterinsurgents? reassessing democracy's impact on war outcomes and duration. *International Organization*, 64(01):167–192.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249.

Schölkopf, B. and Smola, A. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. the MIT Press.

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922.