The Annals of Applied Statistics
2019, Vol. 13, No. 3, 1464–1483
https://doi.org/10.1214/19-AOAS1241

THE CLASSIFICATION PERMUTATION TEST: A FLEXIBLE APPROACH TO TESTING FOR COVARIATE IMBALANCE IN OBSERVATIONAL STUDIES

BY JOHANN GAGNON-BARTSCH AND YOTAM SHEM-TOV

University of Michigan and University of California, Berkeley

The gold standard for identifying causal relationships is a randomized controlled experiment. In many applications in the social sciences and medicine, the researcher does not control the assignment mechanism and instead may rely upon natural experiments or matching methods as a substitute to experimental randomization. The standard testable implication of random assignment is covariate balance between the treated and control units. Covariate balance is commonly used to validate the claim of as good as random assignment. We propose a new nonparametric test of covariate balance. Our Classification Permutation Test (CPT) is based on a combination of classification methods (e.g., random forests) with Fisherian permutation inference. We revisit four real data examples and present Monte Carlo power simulations to demonstrate the applicability of the CPT relative to other nonparametric tests of equality of multivariate distributions.

1. Introduction. Many applications in the social sciences, economics, biostatistics and medicine argue for "as good as random" assignment of units to treatment regimes. Examples include natural experiments, regression discontinuity designs and matching. To support a claim of as good as random assignment, researchers typically demonstrate that the observed covariates are balanced between treatment and control units. Typically it is required to show that pre-treatment characteristics cannot predict future treatment status.

This paper develops a nonparametric test that formalizes the question of whether the covariates can predict treatment status. The test makes use of classification methods and permutation inference, and we name it the Classification Permutation Test (CPT). The CPT trains a classifier (e.g., logistic regression, random forests) to distinguish treated units from control units. Then, using permutation inference, the CPT tests whether the classifier is in fact able to predict treated units from control units more accurately than would be expected by chance.

The CPT may be viewed as a test for equality of multivariate distributions. Because the CPT employs permutation inference, it tests the sharp null that treatment assignment is entirely independent of the covariates, as opposed to testing only whether the covariates are balanced on average. That is, the CPT tests whether the joint distribution of the covariates is the same in both the treatment and control

Received March 2018; revised November 2018.
Key words and phrases. Balance, matching, observational study, natural experiment.

groups. Several other nonparametric tests for equality of multivariate distributions have been proposed in the past. Rosenbaum (2005) developed the Cross-Match test which compares two multivariate distributions using a matching algorithm. First, the observations are matched into pairs, using a distance metric computed from the covariates (treatment status is ignored). The Cross-Match test statistic is then the number of matched pairs containing one observation from the treatment group and one from the control group; high values of the test statistic imply covariate balance, and for low values the null hypothesis of random assignment is rejected. Applications and extensions of the Cross-Match test are described in Heller, Rosenbaum and Small (2010) and Heller et al. (2010). Székely and Rizzo (2009a, 2009b) developed the energy test, another nonparametric test for equality of multivariate distributions. Still other methods include Hansen and Bowers (2008), Heller, Heller and Gorfine (2013), Cattaneo, Frandsen and Titiunik (2015), Chen and Small (2016), Ludwig, Mullainathan and Spiess (2017), Gretton et al. (2012), Romano (1989), and Taskinen, Oja and Randles (2005).

The CPT offers several practical advantages to researchers. First, the CPT can be used as a complementary analysis tool to the classic balance table. A common method of examining whether treatment and control groups are comparable in observable characteristics is to use a balance table that reports the mean of each characteristic in each of the groups and the p-value from a t-test (or Wilcoxon Rank Sum Test) for inference. Balance tables are highly informative but they also suffer from a multiple testing problem. The CPT, however, provides a single joint test for balance in all the characteristics at once, and can therefore complement a balance table by providing an overall measure of imbalance. Second, the CPT can be quite sensitive. Using both simulated and real data applications, we find that the CPT is often able to detect covariate imbalance where existing nonparametric methods do not. Third, the CPT can be used also in testing whether a continuous treatment (e.g., dose) is assigned at random relative to observable characteristics. For a continuous treatment, the classification problem is replaced by a regression problem. In contrast, other existing methods only compare observable characteristics across 2 or K discrete treatment groups. We demonstrate the use of the CPT in this context using data from Green and Winik (2010). Fourth, the CPT has a clear and intuitive interpretation. The test statistic is a direct measure of the ability of the covariates to predict treatment assignment. One advantage of this is that, if covariate imbalance is detected, regular methods of covariate importance (e.g., variable importance plots) can be used to assess which of the covariates or their interactions are causing the imbalances. Moreover, the CPT relates equality of multivariate distributions to the propensity score (Rosenbaum and Rubin (1983)). Rejection of the null hypothesis implies the covariates are predictive of treatment assignment, and can therefore be directly interpreted as a difference in the the distribution of the propensity score across the treatment and control groups. Finally, the CPT offers considerable flexibility, in that it can be used with any classifier. Thus, for example,

if the covariates are high dimensional, a classifier can be chosen that is appropriate to that setting, for example, the elastic net (Zou and Hastie (2005)).

The paper is organized as follows. Section 2 discusses the method, Section 3 examines the performance of the CPT on simulated data and Section 4 looks at four real-life data applications. Section 5 concludes.

2. Method. Suppose there are n units, indexed by i, and m treatment groups. For each unit there is a treatment assignment T_i and a vector of observed covariates $Z_i \in \mathbb{R}^p$. Presumably there is an outcome variable as well, but it is irrelevant for our purposes. For discrete treatments, in which each unit is assigned to one of m treatment groups, T_i is a vector in $\{0,1\}^m$ such that $T_{ik} = 1$ if unit i is in treatment group k and $T_{ik} = 0$ otherwise. For continuous treatments such as drug dose, $T_i \in \mathbb{R}$. We focus first on discrete treatments, and minor modifications to allow for continuous treatments are discussed below.

Let T be the $n \times m$ matrix whose ith row is T_i and let Z be the $n \times p$ matrix whose ith row is Z_i. We wish to test whether $T \perp\!\!\!\perp Z$ or whether treatment assignment is independent of the observed covariates. The CPT proceeds as follows. First, we train a classifier to predict T from Z. The classifier can be anything—logistic regression, a random forest, K-nearest neighbors, etc. We only require that the classifier provide us with a $n \times m$ matrix $\hat{T}$ of "predicted" treatment assignments such that $\hat{T}_i \in \{0,1\}^m$ and $\sum_{k=1}^m \hat{T}_{ik} = 1$ for all i. We then define the *in-sample classification accuracy rate* R as

$$R = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{m} T_{ik} \hat{T}_{ik} \tag{2.1}$$

and use R as our test statistic; intuitively, R should be high only if Z is predictive of T, implying that Z and T are not independent.

To determine statistical significance, we use permutation inference. We randomly permute the rows of T (but not Z) B times. Each time we retrain the classifier and recalculate the classification accuracy rate, which we denote $R_b^\star$, where $1 \le b \le B$. We then calculate our p-value as

$$P = \frac{1}{B} \sum_{b=1}^{B} I\{R \ge R_b^\star\},$$

where $I\{R \ge R_b^\star\}$ is the indicator function for whether $R \ge R_b^\star$.

A few comments: (1) Because B is finite, P is only an approximation to the true permutation test p-value; however the approximation can be made arbitrarily good by increasing B. (2) Ignoring the previous comment, the CPT is guaranteed to control the type-I error rate under the sharp null $T \perp\!\!\!\perp Z$, even in finite samples, no matter what classifier we use, because of the fact we use permutation inference. (3) In particular, the CPT will properly control the type-I error rate despite the fact

that we use the in-sample classification accuracy rate R as the test statistic. Overfitting may occur, causing R to be quite high, perhaps misleadingly so. However, overfitting would cause the $R_b^\star$ to be high as well; thus, any overfitting problem is also manifested in the null distribution, and thereby effectively accounted for. (4) The choice of classifier does affect the power of the test; the CPT will only have power if the classifier is able to distinguish the distribution of the covariates in one treatment group from the distribution of the covariates in the other treatment groups. In this paper we focus primarily on random forests and logistic regression with all pairwise interaction terms included in the design matrix. We select these classifiers because they are able to detect differences in the joint distribution of the covariates, as opposed to merely differences in the marginal distributions.

In addition to the CPT as it is described above, we also consider various modifications. One modification is to consider alternative metrics of classification accuracy. Many classification algorithms provide not only predicted assignments $\hat{T}_{ik} \in \{0, 1\}$ but also estimated "probabilities" of assignment $\hat{p}_{ik} \in [0, 1]$, where $\sum_{k=1}^{m} \hat{p}_{ik} = 1$ for all i. In other words, the $\hat{p}_{ik}$ are not constrained to be 0 or 1, and may therefore carry more information than the $\hat{T}_{ik}$. We may therefore replace the accuracy rate R in (2.1) with the alternative test statistic

$$S = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{m} T_{ik} \log(\hat{p}_{ik}) \tag{2.2}$$

which is commonly referred to as the logarithmic score; previous work suggests that this test statistic should have greater power (Lerch et al. (2017), Gneiting and Raftery (2007)).

A second modification allows for continuous treatment assignments, rather than discrete treatment groups. In this case, $T_i \in \mathbb{R}$ and we use a regression algorithm to predict T_i rather than a classification algorithm (thus $\hat{T}_i \in \mathbb{R}$ as well). Mean squared error may be used as the accuracy metric. We consider such a scenario in Section 4.2.

A third modification is to allow for multiple classifiers, which may increase the power of the test over a wider range of alternatives. We consider two ways of incorporating multiple classifiers. The first is to simply construct an ensemble classifier out of the constituent classifiers; the ensemble classifier may then be used in the CPT like any other classifier. One way to construct an ensemble classifier is to average the assignment probabilities of the constituent classifiers. More specifically, given q constituent classifiers, indexed by c, with assignment probabilities $\hat{p}_{ikc}$, we define the ensemble classifier probabilities as $\hat{p}_{ik} = \sum_{c=1}^{q} \hat{p}_{ikc}/q$. Other approaches are possible, such as having the constituent classifiers vote. The second way of incorporating multiple classifiers is to run the CPT separately for each of the classifiers, obtaining a p-value from each test, and then combine the individual p-values into an overall p-value. Our approach is to combine the p-values using Fisher's method, that is, we define $\chi^2 = -2 \sum_{c=1}^{q} \ln(P_c)$ where P_c is the p-value

of the cth test. However, because the individual tests are not independent, we do not obtain our overall p-value by comparing our χ^2 statistic to the standard χ^2_{2q} distribution. Rather, we again use permutation inference, ensuring the validity of the test. Note that Tang, Chen and Alekseyenko (2016) use a similar permutation-based approach to combine p-values from multiple tests, but take the minimum p-value (over all tests) instead of using Fisher's method.

A final modification of the CPT is considered in Section 4.4, where we consider a scenario in which the experimental units are blocked. We implement a variant of the CPT in which we permute treatment assignment only within blocks.

3. Simulations. We use simple simulations to study the power of the CPT, the Cross-Match test (Rosenbaum (2005)), the energy test (Székely and Rizzo (2009a, 2009b)) and Hotelling's T-test.

In the first simulation we generate $n = 100$ observations; 50 are in treatment and 50 in control. For each observation i we generate a vector Z_i of $p = 3$ covariates. In the control group, the covariates are drawn from a $N(0, I_{3\times 3})$ distribution, and in the treatment group from a $N(\mu, I_{3\times 3})$ distribution, where $\mu = (\beta, \beta, 0)$ and β is a specified parameter. Thus the first two covariates are predictive of treatment assignment, and the third is just noise. Note, importantly, that imbalance of the covariates between the treatment and control groups is apparent in the marginal distributions of the covariates. We refer to this as the "marginal imbalance" simulation.

The second simulation is similar to the first simulation, except that the covariates in the treatment group are drawn from a $N(0, \Sigma_\rho)$ distribution where

$$\Sigma_\rho \equiv \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}$$

and ρ is a specified parameter. Note that in this case, the marginal distributions of the covariates are identical between the treatment and control groups. The only difference is the correlation. We refer to this as the "marginal balance" simulation.

We run each of the two simulations 400 times at each of several different values of the parameters (β or ρ, respectively). In each simulation run, we apply Hotelling's T-test, the Cross-Match test, the energy test and several variants of the CPT: (1) logistic regression classifier; (2) logistic regression classifier with all two-way interactions included in the design matrix ("logistic2" for short); (3) random forest classifier; (4) ensemble classifier, with (1)–(3) as constituent classifiers; (5) a combined result, combining the p-values of (1)–(4) using Fisher's method as described in Section 2. We then estimate the power of the methods by calculating the fraction of times they reject (α-level 0.05).

Some details on implementation: All code is in R, and scripts to reproduce the analysis are provided in Supplement B (Gagnon-Bartsch and Shem-Tov (2019));

the CPT itself is implemented in the R package `cpt`, available on CRAN. Within the `cpt` package, the logistic regression classifiers are implemented using the `multinom` function in the `nnet` package (in order to allow more than two treatment groups). The random forests are implemented using the `randomForest` package; importantly, in the case of the random forest classifier, the out-of-bag estimates are used. When running the CPT we use $B = 500$ permutations. The test statistic is a slightly modified version of the logarithmic score given in (2.2); because $\hat{p}_{ik}$ can sometimes be 0, a constant value of 0.0001 is added to the $\hat{p}_{ik}$ before taking the logarithm. The `cpt` package offers a choice of ensemble classifiers; here we use the default option, which averages the $\hat{p}_{ik}$ of the constituent classifiers, as described in Section 2.

Results are shown in the top two plots of Figure 1. As expected, Hotelling's T-test performs very well when there is marginal imbalance. Notably, the CPT using logistic regression (without interactions) appears to perform just as well as Hotelling's T-test in this scenario. However, both of these tests perform poorly when the marginal distributions are balanced.

The other variants of the the CPT are sensitive to imbalances in both the marginal and joint distributions, and perform well in both simulations. In the "marginal balance" simulation, the "logistic2" variant of the CPT performs best. Notably, the "combined" method performs nearly as well as the best performing method in both the "marginal balance" and "marginal imbalance" scenarios.

Like the CPT, the energy and Cross-Match tests both test the null hypothesis that treatment assignment is independent of the covariates, and are sensitive to imbalances in both marginal and joint distributions. The energy test appears to be very sensitive to imbalances in the marginal distributions, but considerably less sensitive to imbalances that are only apparent in the joint distribution. The Cross-Match test appears to be moderately sensitive to both types of imbalance.

As noted previously, with the CPT we have used the logistic score as the test statistic. Figure 9 in Supplement A (Gagnon-Bartsch and Shem-Tov (2019)) shows results using the classification accuracy rate as the test statistic. The logistic score performs best, and is our recommended choice. We use this test statistic in the remainder of our examples in this paper.

Next, we run two additional simulations to investigate the performance of these tests in a high dimensional setting. The first simulation is similar to the marginal imbalance simulation, except that instead of having two predictive covariates and one noise covariate, we have two predictive covariates and 48 noise covariates. We refer to this as the "50 covariates" simulation. The fourth simulation is similar but with 998 noise covariates, and we refer to this as the "1000 covariates" simulation. In these simulations we replace the logistic regression variants of the CPT with one that uses the elastic net (Zou and Hastie (2005)). This variant is also implemented within the `cpt` R package, using the `cv.glmnet` function. Hotelling's test and the Cross-Match test are not run in the 1000 covariates simulation, since the number of covariates is greater than the number of samples.

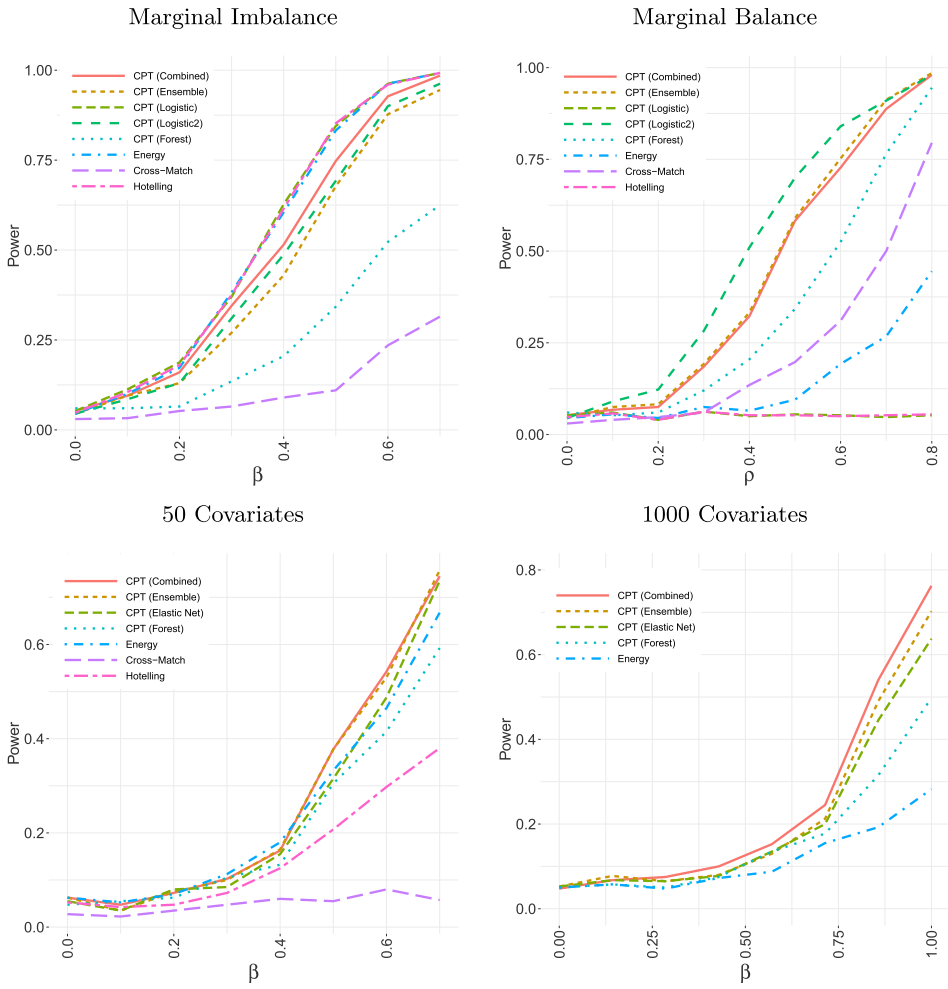

FIG. 1. *Estimated power of the CPT, Cross-Match test, energy test and Hotelling's T-test on simulated data.*

Results are shown in the bottom two plots of Figure 1. Interestingly, the energy test and all variants of the CPT outperform Hotelling's test in the "50 covariates" simulation, demonstrating the value of these methods in even moderately high dimensional settings. The Cross-Match test has comparatively low power. Among the variants of the CPT, the "combined" and "ensemble" methods perform the best. In the "1000 covariates" simulation, all variants of the CPT outperform the energy test. Encouragingly, the "combined" and "ensemble" variants once again have the highest power. This suggests that in practice, it is not critical to choose the one "right" classifier for a given dataset, and that a collection of classifiers may perform as well or better than any individual classifier.

One important weakness of the CPT concerns computational efficiency. With $n = 100$ samples and $p = 3$ covariates, the run times of the Cross-Match, energy and Hotelling tests are on the order of 10^{-2} seconds or less on a modern desktop PC. The runtime of the CPT with a logistic regression classifier however is a full second, and roughly 20 seconds with a random forest. Moreover, for $n = 500$ and $p = 3$ the runtime of the random forest CPT increases to two minutes; for $n = 100$ and $p = 1000$ it is 10 minutes, and for $n = 500$ and $p = 1000$ it is approximately one hour. Thus for very large datasets, it may be necessary to choose a more computationally efficient classifier.

4. Applications.

4.1. *Indiscriminate violence in Chechnya.* Lyall (2009) investigates the effect of indiscriminate violence, specifically the bombing of villages in Chechnya, on insurgent attacks. Villages are the unit of analysis, the treatment variable is bombing status, the outcome of interest is the number insurgent attacks, and covariates include population, elevation, distance to neighboring village, etc. One of the identification strategies is a matching procedure that yields almost perfectly balanced treatment and control groups in all the marginal distributions; see Table 1. Lyall also presents a balance table similar to Table 1 and uses it to support the claim of covariate balance.

The CPT finds significant evidence of covariate imbalance between the treatment and control groups, however, when using either a random forest classifier or logistic regression with two-way interactions. Figure 2 shows the distribution of the CPT test statistic under the null and the observed value of the test statistic. The null hypothesis of random assignment is clearly rejected. This illustrates that

TABLE 1
Covariate balance between treatment and control villages in Lyall (2009)

			P-value		
	Ave. Treat	Ave. Control	*T*-test	Wilcoxon	KS
Log-Population	7.830	7.759	0.699	0.952	0.569
Poverty	2.321	2.239	0.245	0.301	0.988
Tariqa	0.050	0.057	0.804	0.805	1.000
Log-Elevation	5.834	5.766	0.424	0.651	0.260
Isolation	3.767	3.836	0.802	0.656	0.569
Log distance to Neighbor	0.896	0.882	0.854	0.839	0.569
Garrison	0.258	0.283	0.615	0.615	1.000
Rebel	0.585	0.522	0.261	0.260	0.912

Notes: The table shows balance on each covariate separately. p-values are calculated with the t-test, Wilcoxon rank sum test and Kolmogorov–Smirnov (KS) test. The table replicates parts of Table 1 in Lyall (2009).

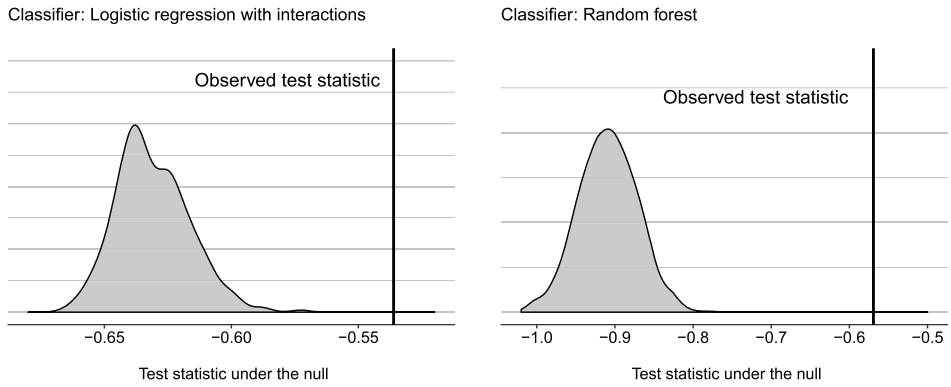

FIG. 2. *Distribution of the CPT test statistic under the null hypothesis*. Notes: *The figure shows the distribution of the CPT test statistic under the null hypothesis of random treatment assignment, and the vertical bar shows the observed test statistic. Results are shown for both logistic regression with all two-way interactions, and for random forests.*

assessing balance using only the marginal distributions of the covariates is not sufficient as imbalances can be hidden in the joint distribution. Similarly, an Hotelling T-test finds no statistically significant difference ($P = 0.81$), again because the imbalances are in the joint distribution of Z, while the marginal distributions of Z are balanced. The Cross-Match test is sensitive to differences in the joint distribution and rejects the null of random assignment ($P < 0.0001$). In principle, the energy test is also sensitive to differences in the joint distribution (this is confirmed in the simulation studies), but interestingly, in this example the energy test does not reject the null ($P = 0.28$).

One practical advantage of the CPT is that the classifier used to perform the test can often also be used to investigate which variables are responsible for any imbalance. For example, if one uses the CPT with random forests and detects an imbalance, one could also use a variable importance plot to explore which covariates are most responsible for the imbalance. One could also fit and plot a classification tree to further investigate the nature of the imbalance, and in particular which interactions are important. Figure 3 shows a variable importance plot and also a plot of a tree. The variable importance plot suggests that log-Population, log-Elevation and log-Distance to neighbor are the most predictive of treatment assignment. The tree suggests specifically that the joint distribution of elevation and distance to nearest village is imbalanced, especially when interacted with population size. Figure 10 in Supplement A (Gagnon-Bartsch and Shem-Tov (2019)) presents contour plots which show the joint distribution of these covariates in greater detail. These figures complement each other and re-enforce our conclusion of imbalance in the joint distribution of the covariates.

Similarly, when using the CPT with logistic regression, one could inspect the results of the regression to investigate the nature of any imbalance. Table 2 reports

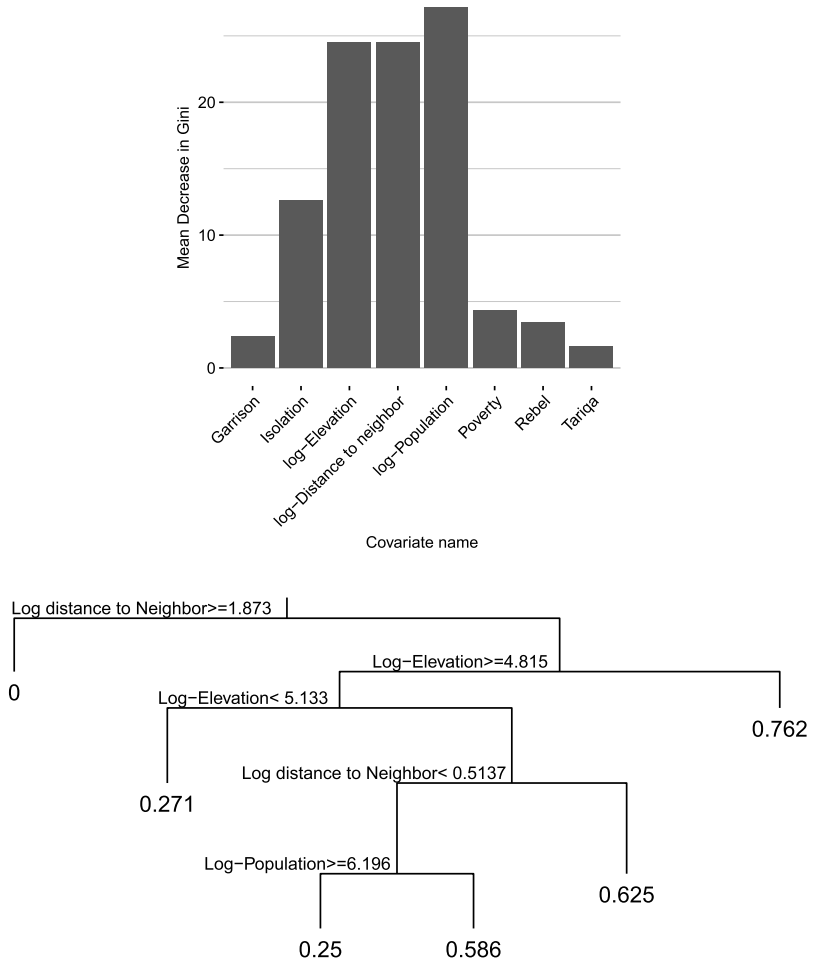

FIG. 3. *Exploring the imbalanced covariates. The top panel shows a variable importance plot. The bottom panel shows an individual tree.*

regression results that show that although each of the marginal distributions of the above covariates are not predictive, the interactions are predictive. Figure 3 and Table 2 complement the CPT's results presented in Figure 2.

4.2. *Random assignment of defendants to judge calendars.* Green and Winik (2010) studied the effect of incarceration length and probation length on recidivism. They argue that defendants are assigned in a quasi-random procedure to different judge calendars, and that judges vary in punishment propensities. The data consist of a sample of 1003 felony drug defendants that are assumed to be randomly allocated between nine different judge calendars. See Table 4 in Supplement A (Gagnon-Bartsch and Shem-Tov (2019)) for a list of all the observed

TABLE 2
Testing the impact of adding interactions to balance tests

	Dependent variable: Treatment group		
Log distance to Neighbor × Log-Elevation		0.687***	0.767***
		(0.233)	(0.265)
Log distance to Neighbor × Log-Elevation × Log-Population			0.118***
			(0.029)
Log-Elevation	0.282	−0.211	−0.764**
	(0.201)	(0.258)	(0.308)
Log-Population	0.095	0.155	−0.359**
	(0.091)	(0.095)	(0.160)
Log distance to Neighbor	0.075	−4.003***	−9.590***
	(0.188)	(1.395)	(2.179)
Constant	−2.444	0.191	7.193***
	(1.693)	(1.899)	(2.685)
Observations	318	318	318

Note: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

covariates. In this example we test Green and Winik's claim of random assignment. We also show that a simple analysis of propensity scores could be misleading, and that logistic regression p-values can be anti-conservative. Finally, we demonstrate how the CPT can be used to test for covariate balance across multiple (greater than two) treatment groups, and also to cases in which the treatment of interest is continuous.

One intuitive method to check for covariate balance that we have not yet discussed is to plot fitted propensity score ($e(Z)$) values. However, this method can be sensitive to over-fitting issues. Consider a binary indicator for whether defendant i was assigned to one specific judge calendar. (Below, we look at judge calendar 1; the choice of calendar 1 is arbitrary.) We fit $e(Z)$ using logistic regression and in Figure 4 we plot the fitted values $\hat{e}(Z)$ separately for the treated (assigned to judge calendar 1) and the controls (assigned to any other calendar). There appears to be some imbalance in the estimated propensity scores, especially when interactions are included. However, this apparent imbalance could either be the result of real imbalances between the treated and control units, or of over-fitting of the logistic regression model to the observed data.

We used the CPT to test this apparent imbalance. Since the CPT re-estimates the logistic regression in each permutation it accounts for any over-fitting. The CPT does not find any difference in the observable characteristics between defendants assigned to judge calendar 1 and the other defendants. The CPT p-value using a logistic regression classifier with all two-way interactions is 0.25.

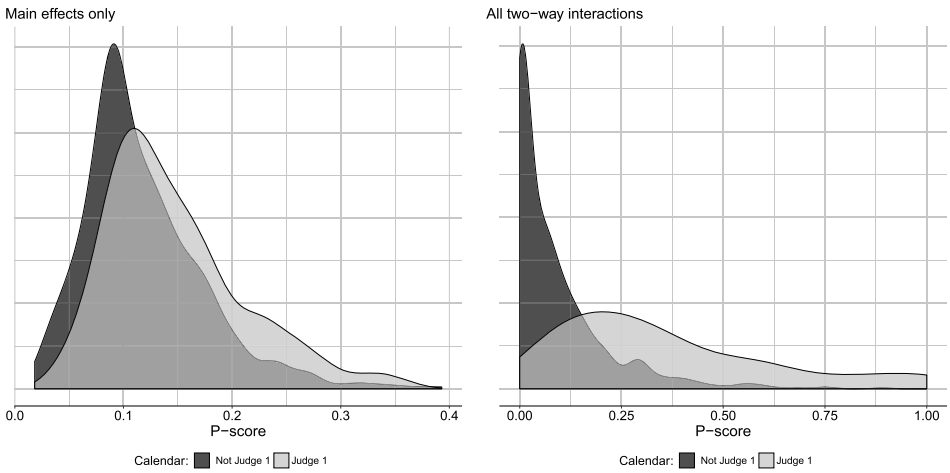

FIG. 4. *The distribution of the estimated propensity score using both a main effects model and a model with all two-way interactions.*

The standard likelihood ratio test (LRT) from a logistic regression is a common alternative to the CPT or other permutation based tests. We fit a logistic regression (both with and without interactions) for each judge calendar and computed the LRT p-values under the null hypothesis that all coefficients are zero. Results are shown in Table 3. Several of the p-values appear significant when the interactions are included; for example, the p-value for judge calendar 1 is 0.001. However, these p-values are not reliable as the LRT can be anti-conservative in finite samples. To demonstrate this we randomly permuted the treatment indicator and re-computed the p-values 1000 times in order to estimate the actual type-I error rate when the

TABLE 3
The Likelihood Ratio Test p-values and Type-I error rates for each judge calendar

	Main effects only (20 coefficients)		All two-way interactions (169 coefficients)	
Judge calendar	p-value	Type-I	p-value	Type-I
1	0.062	0.069	0.001	0.766
2	0.088	0.051	0.000	0.791
3	0.329	0.060	0.001	0.744
4	0.834	0.073	0.000	0.765
5	0.269	0.057	0.002	0.775
6	0.760	0.062	0.163	0.850
7	0.122	0.071	0.000	0.791
8	0.859	0.065	0.058	0.818
9	0.505	0.061	0.000	0.781

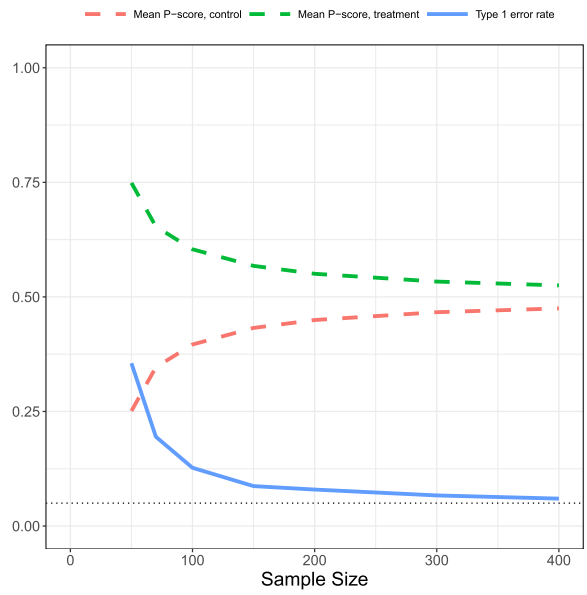

FIG. 5. *Small sample logistic regression simulations*: *Overfitting of propensity scores and anti–conservative bias of the likelihood ratio test*. *Simulation details*: *Sample size varies but the number of covariates is fixed at* 20. *For all sample sizes*, *half of the units are in treatment and half in control*. *All covariates are IID standard normal*; *there is no imbalance between treatment and control*. *We ran* 10,000 *simulations at each sample size*. *The blue line shows the true type*-1 *error rate over the* 10,000 *simulations when the nominal rate is* 0.05. *The green and red dashed lines show the mean fitted propensity scores in the treatment and control groups*, *respectively*.

nominal rate is 5%; results are in Table 3. Figure 11 in Supplement A (Gagnon-Bartsch and Shem-Tov (2019)) shows the entire distribution of the LRT p-values under this scenario in which the null hypothesis of random assignment is correct. It is clear that the finite sample distribution of the LRT p-value has incorrect Type-I error rates. To investigate whether this might be due to some unusual aspect of the distribution of the covariates in this particular dataset, we also performed simple simulation analyses with IID standard normal covariates. Results are in Figure 5. The over-fitting problem of the LRT in finite samples has been previously documented in the literature (Hansen and Bowers (2008)).

Many empirical studies that assume the random assignment of individuals to judges (or examiners) also assume that the judges differ in their degrees of harshness. In Green and Winik (2010), for example, the judges differ in their likelihood to sentence an individual to a term of imprisonment. The judges' "punishment propensities" are then used as a continuous treatment variable (or instrument). The punishment propensities must be estimated from the data; a common approach in the literature is to use the leave-one-out (henceforth LOO) mean sentence length of each judge (Doyle (2007), Doyle (2008), Aizer and Doyle (2015),

Bhuller et al. (2016), Dobbie, Goldin and Yang (2016), Stevenson (2016)). Specifically, suppose defendant i is assigned to judge calendar $j(i)$, where $j(i)$ is a mapping of defendants to judge calendars, $j(i) : \{1, \ldots, N\} \to \{1, \ldots, 9\}$. Let $I[j(l) = j(i)] \in \{0, 1\}$ be an indicator for whether defendants i and l are assigned to the same calendar. Let S_i denote the sentence length of defendant i. The LOO mean, which is used as the estimated punishment propensity, is defined as

$$T_i^{\mathrm{LOO}} \equiv \frac{1}{n_j(i) - 1} \cdot \sum_{l \neq i}^{N} S_l \cdot I\big[j(l) = j(i)\big] \quad \text{where } n_j(i) \equiv \sum_{l}^{N} I\big[j(l) = j(i)\big].$$

Importantly, it is assumed that T_i^{LOO} is independent of the observed characteristics of defendant i.

We can use a modified version of the CPT to test the assumption that T_i^{LOO} is independent of the observed covariates Z_i. As before, we will attempt to predict T_i^{LOO} using Z_i. However, since T_i^{LOO} is a continuous variable, the prediction problem is now a regression problem rather than a classification problem. We choose to use a random forest for the regression, and use the root mean squared prediction error as the test statistic. Figure 6 shows the permutation null distribution of the RMSE as well as the observed value. It is clear from the figure that we cannot reject the null of random assignment. Note that neither the Energy test nor the Cross-Match test can be used in this example, as they do not support continuous treatment variables.

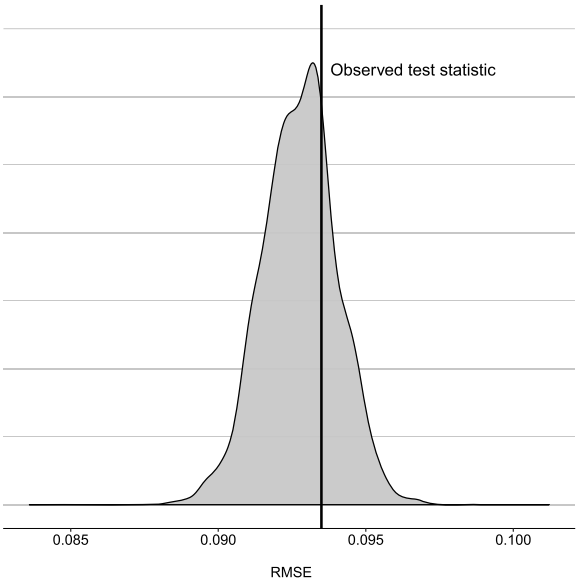

FIG. 6. *The distribution of the random forest prediction RMSE under the null hypothesis of random assignment.*

Finally, to further test the claims of Green and Winik, we ran a variant of the CPT in which we regard each judge as a separate treatment, and test for imbalance across the nine treatment groups simultaneously. Specifically, we used both a multinomial logistic regression without interactions and a random forest classifier to classify each observation as coming from one of the nine calendars. The p-values are 0.30 (logistic) and 0.22 (forest). We also ran the energy test, which allows for multiple treatment groups; the p-value was 0.46. Note that the Cross-Match test could not be used, because it can only be used when there are two groups.

4.3. *MPs for sale.* Eggers and Hainmueller (2009) (henceforth EH) studied the effect of membership in the UK parliament on personal wealth. EH use a regression discontinuity design (RDD) in which candidates for parliament who just barely won an election are compared to candidates who just barely lost. It is common for applied researchers to argue that the observations just above and just below a RDD threshold are roughly comparable, with similar distributions of covariates (Caughey and Sekhon (2011)). Importantly, the distributions of covariates above and below the threshold are not assumed to be identical, but merely similar (for any finite window size). The aim of this data application is to use a RDD design, in which we expect only a small degree of covariate imbalance, to test the sensitivity of different methods to detect this imbalance. Note that this application is *not* meant to argue that the CPT should be used to validate a RDD, but rather only to use real empirical data to benchmark the power of the CPT relative to other nonparametric tests.

In Figure 7 we compare the Energy test, Cross-Match test and the CPT over a grid of different window sizes. The horizontal axis reports the number of observations included in each window and the vertical axis reports the p-value of each of the tests. The results show that the CPT is able to detect imbalances at substantially smaller sample sizes than the Energy and Cross-Match tests. Even with only 34 observations, the CPT finds significant ($P < 0.01$) differences.

Note that in Figure 7 we use a random forest as the classifier, because logistic regression with all two-way interactions has more parameters than observations. This highlights another practical advantage of the CPT, specifically that it can be used even on very high dimensional data when used in conjunction with an appropriate high dimensional classification algorithm such as a random forest.

4.4. *The effect of community college on educational attainment.* Rouse (1995) studied the educational attainment of students who started in a two-year college to that of students at a four-year college. Heller, Rosenbaum and Small (2010) used this data to demonstrate the use of the Cross-Match test for testing imbalance between multivariate distributions. We use this data to demonstrate methodological issues in conducting inference after matching, and not to make any inference or

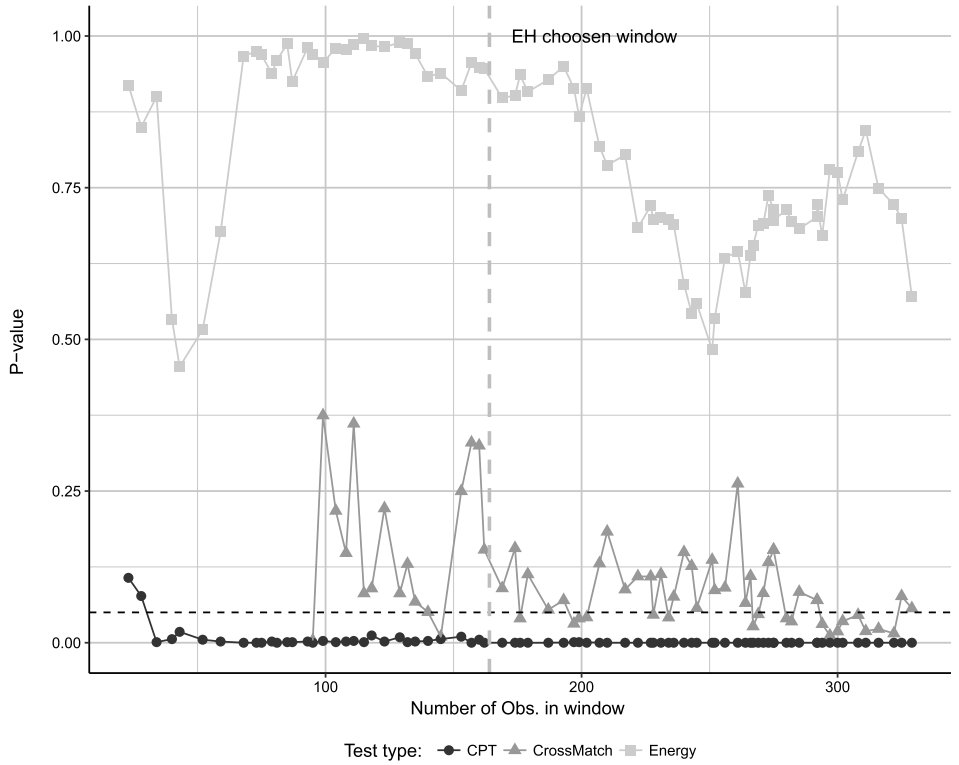

FIG. 7. *P-values of each of the multivariate balance test at different window sizes*. Notes: *For smaller window sizes the Cross-Match test statistic is not well defined* (*and the R function does not run*). *In the main RD treatment effect estimation*, *EH used several different window sizes*, *depending on the specification*, *containing between* 164 *to* 223 *observations*; *see Table* 4 *in EH*.

analysis on the effects of two-year college on educational attainment relative to four-year college.

In a matching design it is common to use Fisherian inference after conducting the matching procedure; see Rosenbaum (2010). A key question is whether after matching the researcher should imagine that units have been randomized within matched blocks, or whether the units have been randomized ignoring the block structure. In other words, in the hypothetical experiment that the matching design is meant to mimic, is the hypothetical experimental design a block-randomized design, or is it one of complete randomization? In this example we show it is essential to specify the treatment assignment model, because the two may lead to opposite conclusions when conducting balance diagnostics. This issue is especially important when using permutation inference as the researcher is required to permute the labels of treated and control units at the level in which treatment was assigned.

In Rouse's data, prior to conducting matching there is clear imbalance in the observable characteristics of students who started at a two-year college and those who

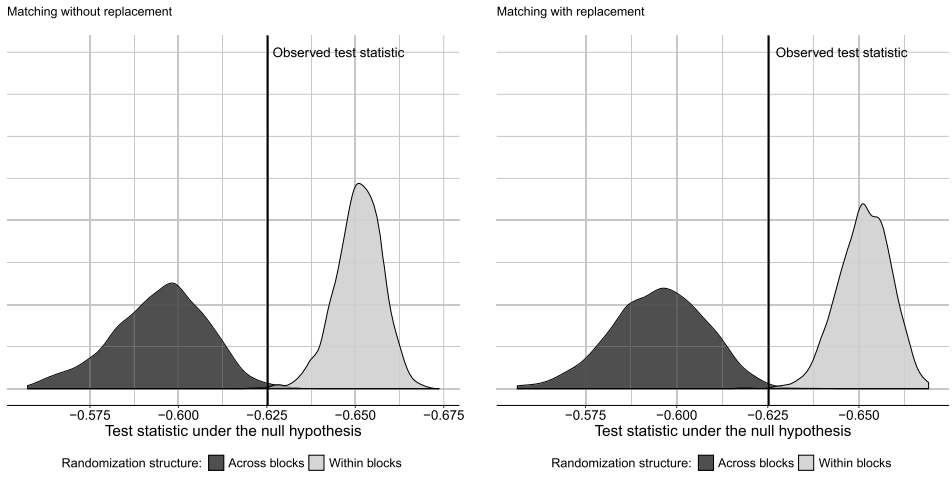

FIG. 8. *The distribution of the test statistic under the null according to randomization within blocks and across blocks for matching designs with and without replacement*. Notes: *The difference between the left and right panels is whether the matching was done with replacement or without, and as can be seen from the figure the matching procedure has no effect on our conclusions concerning within versus across block randomization.*

started at a four-year college (see Figure 12 in Supplement A (Gagnon-Bartsch and Shem-Tov (2019))). After matching, with or without replacement, the balance tables comparing the treated (two-year) and control (four-year) units show the groups are comparable in the observed characteristics and suggest the matching procedure worked well. To test whether there is imbalance in the joint distribution of the covariates we use two variants of the CPT: one in which treatment labels are permuted within blocks (matched pairs), and another in which treatment labels are permuted across all units. Figure 8 shows the results, and yields opposite conclusions depending on the randomization structure that is used. When the randomization structure is across blocks the observed test statistic is to the left of the null distribution, implying more balance than would have been likely under random assignment. When the randomization structure is within blocks the observed test statistic is to the right of the null distribution, implying the covariates can predict the treatment assignment better than under random assignment. The difference between the left and right plots in Figure 8 is the matching method—with or without replacement—and as can be seen the matching procedure has no effect on our discussion of within versus across block randomization.

5. Discussion. The CPT combines classification methods with Fisherian permutation inference. We illustrate the power of the method relative to existing procedures using simulations as well as four real data applications. We hope the CPT will illustrate the gains of using machine learning tools for the construction of pow-

erful new test statistics, and Fisherian inference for conducting hypothesis testing and inference.

This paper emphasizes the importance of examining the joint distribution in addition to the marginal distributions. We also emphasize the advantage of the CPT as an omnibus test that does not suffer from multiple testing concerns. However, importantly, the CPT is *not* a substitute for standard methods such as a balance table that test for differences in the means of each pre-treatment characteristic separately. Rather, the CPT is targeted to complement a balance table and provide a summary measure of the covariates' imbalance.

When using the CPT one must first select a classifier. Flexibility in the selection of the classifier has the advantage of allowing one to choose a classifier that is appropriate to a given dataset. An especially helpful guide to selecting a classifier can be found in Section 10.7 of Hastie, Tibshirani and Friedman (2009), and in particular Table 10.1. Key considerations include dimensionality, the presence of mixed datatypes (quantitative and categorical), presence of outliers or skewed distributions, and computational efficiency. In addition, the CPT allows for combining multiple classifiers, each of which may be more or less sensitive to specific types of imbalance (e.g., in the marginal or joint distributions). This may be a particularly attractive option in practice, since it lessens the need to select one "right" classifier. Moreover, as noted in the simulations, an ensemble classifier may perform better than any individual classifier. The primary drawback to this approach is computational efficiency. Indeed, for very large sample sizes, computational efficiency may be the overriding concern when selecting a classifier.

While flexibility in the choice of classifier is a strength of the CPT, it also opens the door to data snooping. In the context of testing for covariate imbalance, we feel that data snooping is less of a concern than it is in many other contexts. Rather, we feel that failing to detect a true imbalance is a greater concern. Nonetheless, to minimize concerns of data snooping, we recommend the following. First, that as a default, researchers report the result of a "combined" classifier (as described in Sections 2 and 3) consisting of two constituent classifiers: (1) a random forest, and (2) either ordinary logistic regression (if $p < n$) or the elastic net regularized logistic regression (if $p \geq n$). As demonstrated in the simulations, this combination of classifiers should be sensitive to a wide range of alternatives. Second, we recommend that researchers simply report the results of any other variants they run.

Finally, the CPT is also flexible in that it can be easily generalized. As we show in the empirical applications, the CPT can be applied to paired or blocked designs. In addition, it can accommodate discrete treatments with multiple levels as well as continuous treatments. This flexibility, combined with exact finite sample inference, allows researchers to verify random assignment to treatment in a variety of situations. The four empirical applications aim to illustrate the applicability and implementation of the method to different situations that arise in applied research.

Acknowledgements. The authors gratefully acknowledge Isaiah Andrews, Peter Aronow, David Card, Peng Ding, Avi Feller, Brian Graham, Ben Hansen, Patrick Kline, Edward Miguel, Demian Pouzo, Jasjeet Sekhon, Juliana Londoño Vélez, and participants at the UC Berkeley Development Lunch Seminar, the Political Economy Lunch Seminar, and the 2017 Atlantic Causal Inference Conference for helpful discussions and comments.

SUPPLEMENTARY MATERIAL

Supplement A: Online appendix (DOI: 10.1214/19-AOAS1241SUPPA; .pdf). Supplementary figures and tables referenced in the main text.

Supplement B: Code and data (DOI: 10.1214/19-AOAS1241SUPPB; .zip). R scripts and data to reproduce the analyses in this paper. Note that this code and data are also available on GitHub (username `johanngb`) and on Docker Hub (username `johanngb`), and that the R package is available on CRAN (package name `cpt`).

REFERENCES

AIZER, A. and DOYLE, J. J. (2015). Juvenile incarceration, human capital, and future crime: Evidence from randomly assigned judges. *Q. J. Econ.* **130** 759–803.

BHULLER, M., DAHL, G., LOKEN, K. and MOGSTAD, M. (2016). Incarceration, recidivism and employment. NBER 22648.

CATTANEO, M., FRANDSEN, B. and TITIUNIK, R. (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the U.S. Senate. *J. Causal Inference* **3** 1–24.

CAUGHEY, D. and SEKHON, J. (2011). Elections and the regression discontinuity design: Lessons from close US House races. *Polit. Anal.* **19** 385–408.

CHEN, H. and SMALL, D. (2016). New multivariate tests for assessing covariate balance in matched observational studies. Preprint. Available at arXiv:1609.03686.

DOBBIE, W., GOLDIN, J. and YANG, C. (2016). The effects of pre-trial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. Working Paper No. 22511, National Bureau of Economic Research.

DOYLE, J. (2007). Child protection and child outcomes: Measuring the effects of foster care. *Am. Econ. Rev.* **97** 1583–1610.

DOYLE, J. (2008). Child protection and adult crime: Using investigator assignment to estimate causal effects of foster care. *J. Polit. Econ.* **116** 1201–1205.

EGGERS, A. and HAINMUELLER, J. (2009). MPs for sale? Returns to office in postwar British politics. *Am. Polit. Sci. Rev.* **103** 513–533.

GAGNON-BARTSCH, J. and SHEM-TOV, Y. (2019). Supplement to "The classification permutation test: A flexible approach to testing for covariate imbalance in observational studies." DOI:10.1214/19-AOAS1241SUPPA, DOI:10.1214/19-AOAS1241SUPPB.

GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548

GREEN, D. P. and WINIK, D. (2010). Using random judge assignment to estimate the effect of incarceration and probation on recidivism among drug offenders. *Criminology* **48** 357–387.

GRETTON, A., BORGWARDT, K. M., RASCH, M. J., SCHÖLKOPF, B. and SMOLA, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res.* **13** 723–773. MR2913716

HANSEN, B. B. and BOWERS, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statist. Sci.* **23** 219–236. MR2516821

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2722294

HELLER, R., HELLER, Y. and GORFINE, M. (2013). A consistent multivariate test of association based on ranks of distances. *Biometrika* **100** 503–510. MR3068450

HELLER, R., ROSENBAUM, P. R. and SMALL, D. S. (2010). Using the cross-match test to appraise covariate balance in matched pairs. *Amer. Statist.* **64** 299–309. MR2758561

HELLER, R., JENSEN, S. T., ROSENBAUM, P. R. and SMALL, D. S. (2010). Sensitivity analysis for the cross-match test, with applications in genomics. *J. Amer. Statist. Assoc.* **105** 1005–1013. MR2752596

LERCH, S., THORARINSDOTTIR, T. L., RAVAZZOLO, F. and GNEITING, T. (2017). Forecaster's dilemma: Extreme events and forecast evaluation. *Statist. Sci.* **32** 106–127. MR3634309

LUDWIG, J., MULLAINATHAN, S. and SPIESS, J. (2017). Machine learning tests for effects on multiple outcomes. Preprint. Available at arXiv:1707.01473.

LYALL, J. (2009). Does indiscriminate violence incite insurgent attacks?: Evidence from Chechnya. *J. Confl. Resolut.* **53** 331–362.

ROMANO, J. P. (1989). Bootstrap and randomization tests of some nonparametric hypotheses. *Ann. Statist.* **17** 141–159. MR0981441

ROSENBAUM, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 515–530. MR2168202

ROSENBAUM, P. R. (2010). *Design of Observational Studies*. *Springer Series in Statistics*. Springer, New York. MR2561612

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974

ROUSE, C. (1995). Democratization or diversion? The effect of community colleges on educational attainment. *J. Bus. Econom. Statist.* **300** 217–224.

STEVENSON, M. (2016). Distortion of justice: How the inability to pay bail affects case outcomes. Working paper. Available at http://www.econ.pitt.edu/sites/default/files/Stevenson.jmp2016.pdf.

SZÉKELY, G. J. and RIZZO, M. L. (2009a). Brownian distance covariance. *Ann. Appl. Stat.* **3** 1236–1265. MR2752127

SZÉKELY, G. J. and RIZZO, M. L. (2009b). Rejoinder: Brownian distance covariance. *Ann. Appl. Stat.* **3** 1303–1308. MR2752135

TANG, Z.-Z., CHEN, G. and ALEKSEYENKO, A. V. (2016). PERMANOVA-S: Association test for microbial community composition that accommodates confounders and multiple distances. *Bioinformatics* **32** 2618–2625.

TASKINEN, S., OJA, H. and RANDLES, R. H. (2005). Multivariate nonparametric tests of independence. *J. Amer. Statist. Assoc.* **100** 916–925. MR2201019

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327

DEPARTMENT OF STATISTICS
UNIVERSITY OF MICHIGAN
ANN ARBOR, MICHIGAN 48109
USA
E-MAIL: johanngb@umich.edu

DEPARTMENT OF ECONOMICS
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: shemtov@berkeley.edu