Robust Nonparametric Testing for Causal Inference in Observational Studies

Md. Noor-E-Alam
Department of Mechanical and Industrial Engineering, Northeastern University,
360 Huntington Avenue, Boston MA, 02115 E-mail: md.alam@neu.edu

Cynthia Rudin
MIT CSAIL and Sloan School of Management, Massachusetts Institute of Technology,
Cambridge, MA 02139 E-mail: rudin@mit.edu

Abstract We consider the decision problem of making causal conclusions from observational data. Typically, using standard matched pairs techniques, there is a source of uncertainty that is not usually quantified, namely the uncertainty due to the choice of the experimenter: two different reasonable experimenters can easily have opposite results. In this work we present an alternative to the standard nonparametric hypothesis tests, where our tests are robust to the choice of experimenter. In particular, these tests provide the maximum and minimum P-value associated with the set of reasonable assignments of matched pairs. We create robust versions of the sign test, the Wilcoxon signed rank test, the Kolmogorov-Smirnov test, and the Wilcoxon rank sum test (also called the Mann-Whitney U test).

keywords: causal inference, observational studies, nonparametric hypothesis test, matched pairs design, discrete optimization, integer programming.

1 Introduction

When conducting matched pairs tests using observational data, experimenters often use different algorithms to create match assignments. This subjectivity in the assignments can lead to severe problems with uncertainty. In particular, one experimenter (acting in good faith) can have the opposite result as another equally-reasonable experimenter. These inconsistencies can be severely problematic when hypothesis tests lead to important policy decisions, for instance, for drug safety decisions, pollution/environmental decisions, and affirmative action. What if the result that was acted on was simply a fluke of the way the match assignment was constructed? This could be severely problematic. What we would like to know is that no matter which reasonable experimenter conducted the test, that the result would be the same. This means the test should be robust to all reasonable experimenters, where *reasonable* should be defined precisely.

We provide a new set of computationally-aided hypothesis tests in this work that quantify the uncertainty due to the way the matches are constructed. To conduct these tests, we need to define what it means to be a reasonable experimenter. For instance, we could say an experimenter is reasonable if s/he considers a treatment-control pair as a possibility if all of the covariates are within a predefined caliper distance ϵ. We could also create assignments using covariate balance or propensity matching constraints. In this way, the set of reasonable experimenters is defined as a set of constraints for the matching procedure. Within that set of constraints, our tests find match assignments corresponding to the largest and smallest P-values that are possible. In this way, we attain the full range of results that the set of reasonable experimenters would attain. This range of P-values tells us something important about the

causal effect of the treatment. For instance, if the highest possible P-value is below 0.05, we can robustly say that the result is significant.

This work considers several important and popular hypothesis tests, namely the sign test, the Wilcoxon signed rank test, the Kolmogorov-Smirnov test, and the Wilcoxon rank sum test (also the Mann-Whitney U test). In ongoing work (Noor-E-Alam and Rudin 2015), we consider basic tests such as the two sample matched pair z-test and chi-squared test. The non-parametric tests pose challenges that are not present for the basic tests. In particular, non-parametric tests often involve placing the pairs in rank order, which is much more complex in terms of mathematical programming than for the basic tests. Our problems are naturally nonlinear, but we transform them into sequences of integer *linear* programs, leading to computationally practical solutions.

Let us provide formal definitions for our framework. We use the same setup and exposition as Noor-E-Alam and Rudin (2015) (in progress) to introduce the concepts of matching, and then delve into the formulations for the new nonparametric tests.

2 Matching for Robust Tests

This work concerns the potential outcomes framework (see Holland 1986, Rubin 1974). In our notation X represents a vector of covariates for an individual, and Y is an outcome variable that depends on X and whether the patient was treated, $Y(1, X)$ is the random variable for the outcome of a treated individual with covariates X, and $Y(0, X)$ is the outcome for an untreated individual. We make the classical SUTVA assumption (the treatment status of any unit does not affect the potential outcomes of the other units), and assume conditional ignorability, which is that Y is independent of treatment T given X. We also assume unconfoundedness. Our goal is to determine whether we can reject the claim that certain standard quantities are zero, such as:

$$\begin{aligned}\text{ATT}: &\ \mathbb{E}_{(X|T=1),(Y|T=1,X)}[Y(1,X) - Y(0,X)|T=1]\\ \text{ATE}: &\ \mathbb{E}_{(X,Y)}[Y(1,X) - Y(0,X)],\end{aligned}$$

though in this work, rather than the mean, we consider the median treatment effect on the treated and the median treatment effect on the population, and other distributions of the statistics of the treatment effect distribution. The distribution of X is different for the treatment and control groups, because there is a bias as to who receives the treatment. The distribution of $Y|0, X$ and $Y|1, X$ will be different if the treatment has an effect. To conduct our hypothesis test we have observations:

$$\begin{aligned}&(\mathbf{x}_1^t, y_1^t), ..., (\mathbf{x}_T^t, y_T^t)\\ &(\mathbf{x}_1^c, y_1^c), ..., (\mathbf{x}_C^c, y_C^c).\end{aligned}$$

A matching operator determines which control is assigned to which treatment. For simplicity, we write an assignment operator that assigns at most one control to each treatment without replacement, though we loosen this definition later.

Definition: A matching operator $\Omega : \{1, ..., T\} \to \{1, ..., C, \emptyset\}$ obeys the following: if $i \neq k$ and $\Omega(i) \neq \emptyset$ then $\Omega(i) \neq \Omega(k)$. That is, no two treatment units i and k are assigned to the same control unit. We define the size of the matching to be $|\Omega| = \sum_i \mathbf{1}_{[\Omega(i)\neq\emptyset]}$. The set of all matching operators is $\mathcal{A}$ =set of all assignments $\{\Omega\}$.

2.1 Traditional and Current Matching Procedures

Traditional and current matching procedures perform the following two steps, where the matching procedure (denoted "Algorithm" or "Alg" below) is usually not denoted explicitly. (For what follows, we need to denote this explicitly.) The test statistic is a function of the matching procedure.

Classical Procedure.

$$\begin{aligned}\text{Compute match assignment}: \Omega_{\text{Alg}} &= \text{Algorithm}(\{\mathbf{x}_i^t\}_{i=1}^T, \{\mathbf{x}_i^c\}_{i=1}^C)\\ \text{Compute test statistic}: z_{\Omega_{\text{Alg}}}(\{y_i^t\}_{i=1}^T, \{y_i^c\}_{i=1}^C), &\text{ and P-value}(z_{\Omega_{\text{Alg}}}).\end{aligned}$$

Examples of the matching part of the Classical Procedure could include the following:

Matching Example 1:

$$\Omega_{\text{Alg}} = \text{Greedy Matching Algorithm}(\{(\mathbf{x}_i^t)\}_{i=1}^T, \{(\mathbf{x}_i^c)\}_{i=1}^C),$$

where the first treatment observation $\mathbf{x}_1^t$ is matched to its nearest control, then the second treatment observation is matched, and so on. In this case, the test statistic and P-value for the ATT problem would depend on the greedy procedure. This means the P-value would be very sensitive to the particular experimenter and setup – it would even depend on the order in which the data were recorded.

Matching Example 2: Let us consider the classical personnel assignment problem, which is approximately or exactly solved by several software packages:

$$\begin{aligned}\Omega_{\text{Alg}} &\in \text{Optimal Assignment Matching Algorithm}(\{(\mathbf{x}_i^t)\}_{i=1}^T, \{(\mathbf{x}_i^c)\}_{i=1}^C), \\ &= \text{argmin}_{\Omega\in\mathcal{A}} \left(\sum_{i=1}^{T} \text{dist}(\mathbf{x}_i^t, \mathbf{x}_{\Omega(i)}^c) \right),\end{aligned}$$

where every treatment case must be matched to one or more controls for the ATT problem. (That is, $\text{dist}(\mathbf{x}_i^t, \mathbf{x}_\emptyset^c) = \infty$.) In this case the test statistic and P-value depend on the choice of distance measure and the result of this optimization procedure. If there are multiple optimal solutions to the matching problem, or close-to-optimal solutions, they would not be considered in the P-value calculation, and thus the result could depend upon the experimenter and/or the computer memory layout or specific instantiation of the algorithm. It could also depend heavily on outliers that are difficult to match. See Rosenbaum (1989) for a review of the assignment problem.

Matching Example 3: Let us consider a more flexible matching procedure, which is a simple version of Problem 1 of Rosenbaum (2012). Rosenbaum's work points out that some treated subjects may be too extreme to match. His matching algorithm makes three optimal decisions at once: (i) the number of treated subjects to match, (ii) the identity of the treated subjects to match, and (iii) the identity of the controls with whom they are paired.

$$\Omega_{\text{Alg}} \in \text{argmin}_{\Omega\in\mathcal{A}} \left(\sum_{i=1}^{|\Omega|} \text{dist}(\mathbf{x}_i^t, \mathbf{x}_{\Omega(i)}^c) : |\Omega| \geq N \right),$$

where $|\Omega|$ is the number of matches, which we want to be sufficiently large. This aims at a distribution of "marginal patients" for which there is an overlap in density between treatment and control patients, rather than either the density of the ATT or ATE problems. Rosenbaum (2012) also discusses the importance of covariate balance. In our formulations, we assume additional constraints for covariate balance would be included. A different approach is taken by Ho et al. (2007) who remove points outside of the overlap as a pre-processing step. In this work one could use Ho et al. (2007) as a pre-processing step, but our formulations encompass the full setting of Rosenbaum (2012).

Again in this example, a single assignment is chosen, and no assignments that are slightly suboptimal are considered. It does not consider, for instance, all solutions for which the distance measure used in the procedure is slightly different, all solutions for which the solution is slightly suboptimal, or all solutions for which N is varied within a reasonable range. There could be thousands or millions of assignments that are slightly different than the one chosen by this procedure, all leading to possibly different P-values. If all of these P-values were below (for instance) 0.05, the result would be more robust than if only some of them were.

Before we move on, let us give an example of the test statistic computation with the extra notation to make the matching explicit.

Test Statistic Computation Example: We consider the 2-sample matched pair z-test.

$$\bar{d}_{\Omega_{\text{Alg}}} = \frac{1}{|\Omega_{\text{Alg}}|} \sum_{i=1}^{|\Omega_{\text{Alg}}|} (y_i^t - y^c_{\Omega_{\text{Alg}}(i)})$$

$$\sigma^2_{\Omega_{\text{Alg}}} \approx s^2_{\Omega_{\text{Alg}}} = \frac{1}{|\Omega_{\text{Alg}}| - 1} \sum_{i=1}^{|\Omega_{\text{Alg}}|} (y_i^t - y^c_{\Omega_{\text{Alg}}(i)} - \bar{d}_{\Omega_{\text{Alg}}})$$

$$z_{\Omega_{\text{Alg}}} = \frac{\bar{d}_{\Omega_{\text{Alg}}}\sqrt{n}}{\sigma_{\Omega_{\text{Alg}}}}, \quad \text{P-value}_{\Omega_{\text{Alg}}} = 1 - \Phi(z_{\Omega_{\text{Alg}}}).$$

Notation for this problem does not usually include the explicit dependence on the assignment algorithm, masking its contribution to the uncertainty in the whole procedure. The choice of Algorithm is left to the experimenter, which means $z_{\Omega_{\text{Alg}}}$ depends on arbitrarily chosen aspects like the order of the data, and P-value$_{\Omega_{\text{Alg}}}$ suffers the same fate. Partly because of this, one cannot truly study its (finite sample) properties. Rosenbaum (2012) warns that one cannot look at many sets of pairs and pick the one providing the conclusion that the experimenter desires. However, the experimenter's assignment algorithm may be biased for their desired conclusion (or away from that conclusion) without the experimenter knowing it. The uncertainty in the assignment procedure is clearly *not* necessarily a random source of bias.

Moreover:

1. In the classical paradigm, the experimenter does not usually consider what bias is caused by a particular optimization method for the assignment.
2. We have no way of quantifying the uncertainty that comes from the matching procedure if the algorithm is chosen arbitrarily by the experimenter, as is usual practice. We must quantify this in order for the result to be robust to the choice of experimenter.
3. It is not our desire to place a probability distribution over the ways that human experimenters choose assignment algorithms. This is not interesting. Instead, what is interesting is the range of results that reasonable experimenters might obtain.
4. What we will propose is not equivalent to taking random subsamples of data and repeating the experiment. That procedure could yield trivial (and disastrous) results in the case where one calculates the maximum and minimum over test statistics. This range would grow infinitely large as the number of observations increases. In contrast, our range would generally decrease as the number of observations grows. We assume all observations are available to be matched, and we robustify only over the assignment procedure.

2.2 Proposed Approach

First we define a *set of good assignments* as $\mathcal{A}_{\text{good}} \subseteq \mathcal{A}$ where all $\Omega \in \mathcal{A}_{\text{good}}$ obey constraints besides those in $\mathcal{A}$ such as (for example):

- (Calipers) When $\Omega(i) \neq \emptyset$ then $\text{dist}(\mathbf{x}_i^t, \mathbf{x}^c_{\Omega(i)}) \leq \epsilon$.
- (Covariate balance, mean of chosen treatment units close to mean of full treatment group) $\forall$ covariates p, using notation $\bar{x}_p^t = \frac{1}{T}\sum_{t=1}^T x_{ip}^t$,

$$\frac{1}{|\Omega|} \sum_{\{i:\Omega(i)\neq\emptyset\}} x_{ip}^t - \bar{x}_p^t \leq \epsilon_p \text{ and}$$

$$\bar{x}_p^t - \frac{1}{|\Omega|} \sum_{\{i:\Omega(i)\neq\emptyset\}} x_{ip}^t \leq \epsilon_p.$$

- (Covariate balance, mean of chosen treatment units similar to mean of full sample) $\forall$ covariates p, using notation:

$$\bar{x}_p^{tc} := \frac{1}{T+C}\left(\sum_{t=1}^T x_{ip}^t + \sum_{c=1}^C x_{ip}^c\right), \tag{1}$$

$$\frac{1}{|\Omega|}\sum_{\{i:\Omega(i)\neq\emptyset\}} x_{ip}^t - \bar{x}_p^{tc} \leq \epsilon_p, \text{ and vice versa}$$
$$\bar{x}_p^{tc} - \frac{1}{|\Omega|}\sum_{\{i:\Omega(i)\neq\emptyset\}} x_{ip}^t \leq \epsilon_p.$$

– (Maximizing the fitness of the matches) One can optimize a measure of user-defined fitness for the assignment, and then constrain $\mathcal{A}_{\text{good}}$ to include all other feasible assignments at or near that fitness level, by including the following constraints:

$$\text{Fitness}(\Omega, \{x_i^t\}_{i=1}^T, \{x_i^c\}_{i=1}^C) \geq \text{Maxfit} - \epsilon,$$

where Maxfit is precomputed,

$$\text{Maxfit} = \max_{\Omega'\in\mathcal{A}} \text{Fitness}(\Omega', \{x_i^t\}_{i=1}^T, \{x_i^c\}_{i=1}^C).$$

If one desires the range of results for all maximally fit pairs and no other pairs, ϵ can be set to 0.

In our examples we use calipers mostly for ease of notation, but replacing constraints with those above (or other constraints) is trivially simple using MIP software.

$\mathcal{A}_{\text{good}}$ intuitively represents the set of assignments arising from all reasonable matching methods. The procedure is as follows:

Robust Procedure. Define $\mathcal{A}_{\text{good}}$. Compute and return the maximum and minimum P-value over the set $\mathcal{A}_{\text{good}}$, the maximum and minimum value of the test statistic, and the maximum and minimum value of the estimated treatment effect.

For the matched pair z-test we would write:

$$\text{P-value}_{\max} = \max_{\Omega\in\mathcal{A}_{\text{good}}} 1 - \Phi(z_\Omega) = 1 - \Phi\left(\min_{\Omega\in\mathcal{A}_{\text{good}}} z_\Omega\right)$$

and

$$\text{P-value}_{\min} = \min_{\Omega\in\mathcal{A}_{\text{good}}} 1 - \Phi(z_\Omega) = 1 - \Phi\left(\max_{\Omega\in\mathcal{A}_{\text{good}}} z_\Omega\right).$$

The two P-values quantify the possible uncertainty due to the matching procedure. If all reasonable matching algorithms for the Classical Procedure produce an $\Omega_{\text{Alg}} \in \mathcal{A}_{\text{good}}$, this would imply:

$$\text{For all } \Omega_{\text{Alg}} \in \mathcal{A}_{\text{good}}, \text{P-value}_{\min} \leq \text{P-value}_{\Omega_{\text{Alg}}} \leq \text{P-value}_{\max}.$$

Thus, computing $\text{P-value}_{\min}$ and $\text{P-value}_{\max}$ would be robust to the choices of human behavior that influence the algorithm for choosing Ω_{Alg}.

The procedure also returns the range of z-scores, $\min_{\Omega\in\mathcal{A}_{\text{good}}} z_\Omega$ and $\max_{\Omega\in\mathcal{A}_{\text{good}}} z_\Omega$.

Extending our reasoning from bullet 3 in Section 2.1,

– We do not want to consider how likely a certain assignment Ω is to appear. This would involve modeling human behavior of the experimenter. We do not want to place a distribution over the choice over algorithms that an experimenter would choose. As Morgan and Winship (2007) note, there is no clear guidance on the choice of matching procedure. We do not presuppose a distribution over these procedures.
– We do not want to consider statistics of the set of $\mathcal{A}_{\text{good}}$, such as the average P-value over the set of $\mathcal{A}_{\text{good}}$. This is simply a special case of the previous point, where we assume that all good assignments are equally likely to be chosen, which is clearly not true in practice.

A remark on the covariate balance constraints: as noted in many other works, by choosing constraints on Ω and/or an objective that favors using more treatment points, the estimates of causal effect will generally be less biased for the ATT. One can similarly alter constraints or the objective to reduce bias on ATE. Many other works speak to the benefits of better matching procedures and reduction of bias (e.g., Rosenbaum 1989).

A remark on the uses of this procedure: the techniques proposed here are appropriate for observational data, and not appropriate for studies where the matching is done prior to data collection, for the reason

that one generally would not want to match, collect data, then afterwards rethink the assignments among the collected data.

In what follows, we provide special cases of the Robust Procedure for various specific hypothesis tests. The goal is to provide conclusions that are robust to the class of experimenters within $\mathcal{A}_{\text{good}}$.

In what follows we create robust versions of the sign test, Wilcoxon signed rank test, Wilcoxon-Mann-Whitney U test, and the Kolmogorov-Smirnoff test, and demonstrate them afterwards in severcal case studies.

3 A Robust Sign Test for Matched Pairs

Let us consider the upper one-sided sign test, which is a test on the population medians. To conduct the matched pairs sign test in the classical way, we would first construct the matches. The test statistic would be the number of treatment-control pairs for which the outcome for the treatment observation was higher than that of the control observation. (see, for instance, the textbook of Tamhane and Dunlop 2000).

Let us formulate this test within the framework of mathematical programming. From each pair of observations, we are given treatment effect d_{ij}, which is the difference between the treatment and control outcomes for pair ij, that is, d_{ij}=$T_i - C_j$. These are all precomputed for each i and j in the treatment and control groups, and should be considered as input to the method. In order to consider the pairs which have positive d_{ij}'s, we define binary matrix $\boldsymbol{\theta}$, with elements θ_{ij} where if $d_{ij} > 0$ then $\theta_{ij} = 1$, otherwise $\theta_{ij} = 0$. Matrix $\boldsymbol{\theta}$ is also an input to the method.

We would like to know whether the median of the differences between treatment and control outcomes for the matched pairs is significantly greater than 0, meaning that the treatment is estimated to have an effect (assuming SUTVA and ignorability). We consider the hypothesis H_0 : (median difference = 0) versus the alternative H_1 : (median difference > 0). In what follows, S_+ is the number of positive differences, meaning $\sum_{(i,j):i \text{ and } j \text{ are a pair}} \theta_{ij}$, and n is the total number of pairs. The z-score is:

$$z = \frac{S_+ - n/2 - 1/2}{\sqrt{n/4}},$$

and the P-value is P-value = $\Phi(z)$. Since $\Phi(z)$ is monotonically increasing in z, maximizing z is the same as maximizing the value of $\Phi(z)$. Also maximizing S_+ is the same as maximizing z, as z is monotonically increasing in S_+.

Our algorithm will choose matched pairs in order to maximize and minimize S_+. The largest value of the objective will be called $S_{+,\max}$ where $S_{+,\max} = \max_{\mathbf{a}\in\{\mathbf{pairs}\}} S_+(\mathbf{a})$. Here $\mathbf{a}$ is a binary matrix where entry a_{ij} is 1 when i and j are a pair. A simple adaptation would minimize S_+ to obtain $S_{+,\min}$ where $S_{+,\min} = \min_{\mathbf{a}\in\{\mathbf{pairs}\}} S_+(\mathbf{a})$. The quantities we aim to report for each number of pairs n are the ranges of S_+, that is $S_{+,\max}$ and $S_{+,\min}$, the z-scores and P-values, that is,

$$[S_{+,\min}, S_{+,\max}], [z_{\min}, z_{\max}], \text{ and } [\Phi(z_{\min}), \Phi(z_{\max})].$$

Thus, the input parameters are as follows:

n is the total numbers of matched pairs we will find. We loop over all possible values of n, until the solution becomes infeasible as we run out of matches to make.

Q is the set of all observations in the treatment group, indexed by i

R is the set of all observations in the control group, indexed by j

T_i is the the outcome of a treated observation i in the treatment group

C_j is the the outcome of a control observation j in the control group

θ_{ij} is a binary precomputed parameter, which is 1 whenever $T_i > C_j$, 0 otherwise

s_{ij} is the ijth element of a matrix. It takes value 1 if the covariates of treated observation i and control observation j are similar enough to be a possible matched pair, otherwise 0.

There is only one matrix of decision variables, namely:

a_{ij} is a binary variable that is 1 if i and j are in the same pair, otherwise 0.

Optimization for S_+ can then be formulated as follows.

$$\text{Maximize/Minimize}_{\mathbf{a}} \quad S_+(\mathbf{a}) = \sum_{i \in Q} \sum_{j \in R} a_{ij} \theta_{ij}$$

subject to:

$$\sum_{i \in Q} \sum_{j \in R} a_{ij} = n \qquad \text{(Choose } n \text{ pairs)} \tag{2}$$

$$\sum_{i \in Q} a_{ij} \leq 1 \qquad \forall j \quad \text{(Choose at most one treatment observation)} \tag{3}$$

$$\sum_{j \in R} a_{ij} \leq 1 \qquad \forall i \quad \text{(Choose at most one control observation)} \tag{4}$$

$$a_{ij} \leq s_{ij} \qquad \forall i, j \quad \text{(Choose only pairs that are allowed)} \tag{5}$$

$$a_{ij} \in \{0, 1\} \qquad \forall i, j \quad \text{(Defines binary variable } a_{ij}\text{)} \tag{6}$$

(optional covariate balance constraints).

In the objective function $\sum_{i \in Q} \sum_{j \in R} a_{ij} \theta_{ij}$ the a_{ij} term ensures that only differences between the pairs i, j that we are selecting are used for calculating S_+. The first constraint ensures that we choose exactly n pairs (we loop over all values of n where feasible solutions exist), the second and third constraints ensure that only one treatment and one control observation are selected for each pair, and the fourth constraint is an if-then constraint, stating that we are only allowed to choose pairs i, j for which $s_{ij} = 1$. Recall that the s_{ij}'s were determined in advance, and they encode whether i's covariates are close enough to j's covariates to be chosen as a possible pair. The last constraint defines binary variables a_{ij}.

The sign test is a weak test in the sense that it requires very few distributional assumptions. This reduces the power of the test. On the other hand, it is more widely applicable to different types of datasets because it makes few assumptions.

4 Robust Wilcoxon Signed Rank Test

The Wilcoxon Signed Rank Test is a more powerful test on the median than the sign test, but with a stronger assumption: it assumes that the population distribution is symmetric. We define the following parameters and decision variables to formulate the model:

T_i is the the outcome of a treated observation i in the treatment group

C_j is the the outcome of a control observation j in the control group

s_{ij} is the ijth element of a matrix. It takes value 1 if the covariates of treated observation i and control observation j are similar enough to be a possible matched pair, otherwise 0 δ_{ij} is a binary precomputed parameter, which is equal to 1 if T_i is greater than C_j, 0 otherwise g_{ijkl} is a binary precomputed parameter, which is equal to 1 if η_{ij} is greater than η_{kl}, where $\eta_{ij} = |T_i - C_j|$, 0 otherwise

There are three matrices of decision variables, namely:

a_{ij} is a binary variable that is 1 if i and j are in the same pair, otherwise 0

z_{ijkl} is a binary variable that is 1 if g_{ijkl}, a_{ij} and a_{kl} all are equal to 1, 0 otherwise. Intuitively z_{ijkl} is 1 only when ij and kl are both being used as pairs and when ij's absolute difference is larger than kl's absolute difference.

h_{ij} is an integer variable whose value is the rank of pair ij. It is the count of kl pairs ranked beneath pair ij (plus one, so that the lowest rank is 1 rather than 0).

To compute the test statistic in the traditional way, one would compute absolute differences η_{ij} and rank order them. The test statistic is the sum of ranks of the positive differences.
The formulation is as follows :

$$\text{Maximize/Minimize}_{\mathbf{a},\mathbf{z},\mathbf{h}} \quad w_+(\mathbf{a}, \mathbf{z}, \mathbf{h}) = \sum_{i \in Q} \sum_{j \in R} h_{ij} \delta_{ij}$$

subject to:

$$g_{ijkl} + a_{ij} + a_{kl} - 2 \leq z_{ijkl} \quad \forall i,j,k,l \quad (z_{ijkl} \text{ takes a value of 1 only if } a_{ij},\ a_{kl} \text{ and } g_{ijkl} \text{ are 1}) \quad (7)$$

$$z_{ijkl} \leq a_{ij} \quad \forall i,j,k,l \quad (z_{ijkl} \text{ is allowed to take a value of 1 only if } a_{ij} \text{ is 1}) \quad (8)$$

$$z_{ijkl} \leq a_{kl} \quad \forall i,j,k,l \quad (z_{ijkl} \text{ is allowed to take a value of 1 only if } a_{kl} \text{ is 1}) \quad (9)$$

$$z_{ijkl} \leq g_{ijkl} \quad \forall i,j,k,l \quad (z_{ijkl} \text{ is allowed to take a value of 1 only if } g_{ijkl} \text{ is 1}) \quad (10)$$

$$a_{ij} \leq s_{ij} \quad \forall i,j \quad (\text{Choose only pairs that are allowed}) \quad (11)$$

$$\sum_{i \in Q} \sum_{j \in R} a_{ij} = n \quad (\text{Choose } n \text{ pairs}) \quad (12)$$

$$\sum_{i \in Q} a_{ij} \leq 1 \quad \forall j \quad (\text{Choose at most one treatment observation}) \quad (13)$$

$$\sum_{j \in R} a_{ij} \leq 1 \quad \forall i \quad (\text{Choose at most one control observation}) \quad (14)$$

$$a_{ij} \in \{0,1\} \quad \forall i,j \quad (\text{Defines binary variable } a_{ij}) \quad (15)$$

$$h_{ij} = \sum_{k \in Q} \sum_{l \in R} z_{ijkl} + a_{ij} \quad \forall i,j \quad (\text{Defines } h_{ij}). \quad (16)$$

(optional covariate balance constraints).

Equations (7) to (10) are used to ensure that z_{ijkl}=1 when g_{ijkl}, a_{ij} and a_{kl} all are one. Equation (11) is used to maintain covariates constraint such that only when the precomputed parameter s_{ij} is 1 then a_{ij} is allowed to take a value of 1. Constraints (12)-(14) are the same constraints as in the sign test, to make sure we have n pairs with one treatment and control observation in each pair. Equation (15) defines binary variables a_{ij}. Equation (16) is used to calculate rank h_{ij} for each pair ij. If the pair ij is not being used then a_{ij} will be 0, which means z_{ijkl} will be 0 by Constraint (8). This means the only ij pairs that have positive heights are those that are being used as matched pairs. The objective will only add up heights of pairs ij for which the absolute differences are positive, which was precomputed as δ_{ij}.

The z-score is computed from the optimal value of w_+ through the following formula, and the pvalue is computed as usual.

$$z = \frac{w_+(\mathbf{a}, \mathbf{z}, \mathbf{h}) - n(n+1)/4 - 1/2}{\sqrt{n(n+1)(2n+1)/24}}.$$

5 Robust Wilcoxon-Mann-Whitney Test (Wilcoxon Rank Sum Test)

The Wilcoxon-Mann-Whitney test (which is the same as the Wilcoxon Rank Sum test or the Mann-Whitney U test) is for two independent samples, and is used to show that observations from one population tend to be (stochastically) larger than from another population. Like the signed rank test, it requires us to sum ranks, but this time we sum the ranks of the observations rather than the ranks of pairs since we are working with independent samples. When applying this to observational data with covariates, we will choose samples from the treatment and control population that are similar to each other so that the test is meaningful. This requires us to select a subset of the treatment and control observations, so again we are in the robust setting. However, we are not constructing one-to-one matches as in the previous two hypothesis tests. Because of this, we chose to use covariate balance constraints rather than calipers for our experiments.

The way we were able to formulate this problem differs notationally from the previous two hypothesis tests. We define the following parameters and decision variables to formulate the model:

V is the set of all observations in the treatment and control group, indexed by i

t_i is precomputed binary parameter which is 1 if unit i is in the treatment group

y_i is the outcome of unit i

X_{ik} is the k^{th} covariate value for unit i

g_{ij} is 1 if outcome y_i is greater than outcome y_j, 0 otherwise

There are three matrices of decision variables, namely:

u_i is 1 if we use unit i

z_{ij} is a binary variable that is 1 if g_{ij}, u_i and u_j all are equal to 1, 0 otherwise
h_i is the rank of observation i.
The formulation is:

$$\text{Maximize/Minimize}_{\mathbf{u},\mathbf{z},\mathbf{h}} \quad w(\mathbf{u},\mathbf{z},\mathbf{h}) = \sum_{i \in V} h_i(1-t_i)$$

subject to:

$$g_{ij} + u_i + u_j - 2 \leq z_{ij} \qquad \forall i,j \quad (z_{ij} \text{ takes a value of 1 if } u_i, u_j \text{ and } g_{ij} \text{ are 1}) \tag{17}$$

$$z_{ij} \leq u_i \qquad \forall i,j \quad (z_{ij} \text{ is allowed to take a value of 1 when } u_i \text{ is 1}) \tag{18}$$

$$z_{ij} \leq u_j \qquad \forall i,j \quad (z_{ij} \text{ is allowed to take a value of 1 when } u_j \text{ is 1}) \tag{19}$$

$$z_{ij} \leq g_{ij} \qquad \forall i,j \quad (z_{ij} \text{ is allowed to take a value of 1 when } g_{ij} \text{ is 1}) \tag{20}$$

$$u_i \in \{0,1\} \qquad \forall i \quad (\text{Defines binary variable } u_i) \tag{21}$$

$$h_i = \sum_{j \in V} z_{ij} + u_i \qquad \forall i \quad (\text{Defines } h_i \text{ as the rank of observation } i) \tag{22}$$

$$\sum_{i \in V} u_i t_i = b \qquad (\text{total number of treatment units is } b) \tag{23}$$

$$\sum_{i \in V} u_i(1-t_i) = c \qquad (\text{total number of control units is } c). \tag{24}$$

$$\gamma \leq \frac{\sum_{i \in V} X_{ik} u_i t_i}{b} - \frac{\sum_{i \in V} X_{ik} u_i (1-t_i)}{c} \leq \beta \quad \forall k \quad (\text{covariate balance}) \tag{25}$$

Equations (17) to (20) are used to ensure that z_{ij}=1 when g_{ij}, u_i and u_j all are one. Equation (21) defines binary variables u_i. Equation (22) is used to calculate rank h_i for each observation i. Equations (23) and (24) constrain the sizes of the samples, and the sizes of the samples are used in the covariate balance constraint (25), which states that the average of each covariate in the treatment group is not too different from the average of each covariate in the control group. Usually we choose $\gamma = -\beta$ to be a small constant. One could construct many variations on the covariate balance equations. We chose simply to have the averages of each of the covariates match, which was a nonlinear constraint that we split into two loops over values for sizes b and c of the treatment and control groups.

The test statistic is converted to a z-score through the usual formulas:

$$z = \frac{u_1 - bc/2 - 1/2}{\sqrt{bc(b+c+1)/12}}, \text{ where } u_1 = w(\mathbf{u},\mathbf{z}) - \frac{b(b+1)}{2}, \tag{26}$$

where b and c are the same values we chose for the number of points in the treatment and control groups within the formulation.

6 Robust Kolmogorov Smirnov Test

The two-sample Kolmogorov Smirnov (K-S) test compares the distributions of two independent samples. The K-S test statistic is the maximum difference between the empirical cumulative distribution functions of the two samples, $D = \max |F_C(O_i) - F_T(O_i)|$, where O_i denotes the i^{th} possible value of the outcome. Because of this, the K-S test can detect differences between two distributions beyond simply the median. In our experiments below, the K-S test was able to robustly detect a difference in outcomes that other tests were not able to detect.

We will use that fact that empirical cumulative distribution functions are piecewise constant, and that they change only at each outcome value. This means we need only check the differences between the two empirical CDFs at each of the outcome values. For convenience and troubleshooting, we usually reorder the observations prior to the start of the computation so that the i^{th} observation has the i^{th} largest outcome, though this is not necessary.

We define the following parameters and decision variables to formulate the optimization problem for the test statistic:

V is the set of all observations in the treatment and control group, indexed by i
t_i is a precomputed parameter that is 1 if unit i is in the treatment group

y_i is the (real-valued) outcome of unit i
X_{ik} is the k^{th} covariate value for unit i
g_{ij} is 1 if outcome y_i is greater than outcome y_j, 0 otherwise
z_{ij} is a binary parameter that is 1 if g_{ij} and t_i are equal to 1, 0 otherwise
$\bar{z}_{ij}$ is a binary parameter that is 1 if g_{ij} is equal to 1 and t_i is equal to 0, 0 otherwise
There are seven matrices of decision variables, namely:
u_i is 1 if we use unit i
d_i is the difference between the empirical cumulative distribution function of treatment and control groups at unit i's outcome value
$\bar{d}_i$ is d_i when i is used, 0 otherwise
δ_i is the absolute value of $\bar{d}_i$
ψ_i is a binary variable used to linearize the absolute function δ_i in the formulation
$\bar{\psi}_i$ is another binary variable used to linearize absolute function δ_i
θ_i is a binary variable used to linearize $\max_i \delta_i$.
The formulation is :

$$\text{Maximize/Minimize}_{\mathbf{u},\mathbf{d},\bar{\mathbf{d}},\boldsymbol{\psi},\bar{\boldsymbol{\psi}},\boldsymbol{\theta},\boldsymbol{\delta}} \quad D(\mathbf{u},\mathbf{d},\bar{\mathbf{d}},\boldsymbol{\psi},\bar{\boldsymbol{\psi}},\boldsymbol{\theta},\boldsymbol{\delta})$$

subject to:

$$d_i = \sum_{j\in V}\left(\frac{u_j z_{ij}}{b} - \frac{u_j \bar{z}_{ij}}{c}\right) \qquad \forall i \quad \text{(Calculates } d_i\text{)} \tag{27}$$

$$u_i \geq \bar{d}_i \qquad \forall i \quad \text{(Confirms that } \bar{d}_i \text{ is 0 if } u_i \text{ is 0)} \tag{28}$$

$$\bar{d}_i - d_i \leq (1-u_i) \qquad \forall i \quad \text{(Confirms that } \bar{d}_i = d_i \text{ when } u_i \text{ is 1)} \tag{29}$$

$$d_i - \bar{d}_i \leq (1-u_i) \qquad \forall i \quad \text{(Confirms that } \bar{d}_i = d_i \text{ when } u_i \text{ is 1)} \tag{30}$$

$$\delta_i \geq \bar{d}_i \qquad \forall i \quad \text{(defines } \delta_i \text{ as absolute value of } d_i\text{)} \tag{31}$$

$$\delta_i \geq -\bar{d}_i \qquad \forall i \quad \text{(defines } \delta_i \text{ as absolute value of } d_i\text{)} \tag{32}$$

$$\delta_i \leq \bar{d}_i + (1-\psi_i) \qquad \forall i \quad \text{(defines } \delta_i \text{ as absolute value of } d_i\text{)} \tag{33}$$

$$-\delta_i \leq \bar{d}_i + (1-\bar{\psi}_i) \qquad \forall i \quad \text{(defines } \delta_i \text{ as absolute value of } d_i\text{)} \tag{34}$$

$$\sum_{i\in V} \psi_i + \bar{\psi}_i = 1 \qquad \forall i \quad \text{(defines } \delta_i \text{ as absolute value of } d_i\text{)} \tag{35}$$

$$D \geq \delta_i \qquad \forall i \quad (D \text{ takes the maximum value of } \delta_i \text{)} \tag{36}$$

$$D \leq \delta_i + (1-\theta_i) \qquad \forall i \quad (D \text{ takes the maximum value of } \delta_i \text{ when } \theta_i = 1) \tag{37}$$

$$\sum_{i\in V} \theta_i = 1 \qquad \forall i \quad \text{(Max occurs when } \theta_i = 1\text{)} \tag{38}$$

$$\gamma \leq \frac{\sum_{i\in V} X_{ik} u_i t_i}{b} - \frac{\sum_{i\in V} X_{ik} u_i (1-t_i)}{c} \leq \beta \qquad \forall k \quad \text{(covariate balance)} \tag{39}$$

$$\sum_{i\in V} u_i t_i = b \qquad \text{(total number of treatment units is } b\text{)} \tag{40}$$

$$\sum_{i\in V} u_i (1-t_i) = c \qquad \text{(total number of control units is } c\text{)} \tag{41}$$

$$\psi_i \in \{0,1\} \qquad \forall i \quad \text{(binary variables } \psi_i\text{)} \tag{42}$$

$$\bar{\psi}_i \in \{0,1\} \qquad \forall i \quad \text{(binary variables } \bar{\psi}_i\text{)} \tag{43}$$

$$\theta_i \in \{0,1\} \qquad \forall i \quad \text{(binary variables } \theta_i\text{).} \tag{44}$$

Equation (27) calculates d_i. Equations (28) to (30) are used to ensure that $\bar{d}_i = d_i$ when observation i is used, otherwise $\bar{d}_i = 0$. Equations (31) to (35) are used to linearize the absolute function in d_i. Equations (36) to (38) are used to linearize the max function. Equations (39) maintains covariate balance, and (40) and (41) are the sizes of the treatment and control samples. The P-value can be computed directly from the test statistic as usual.

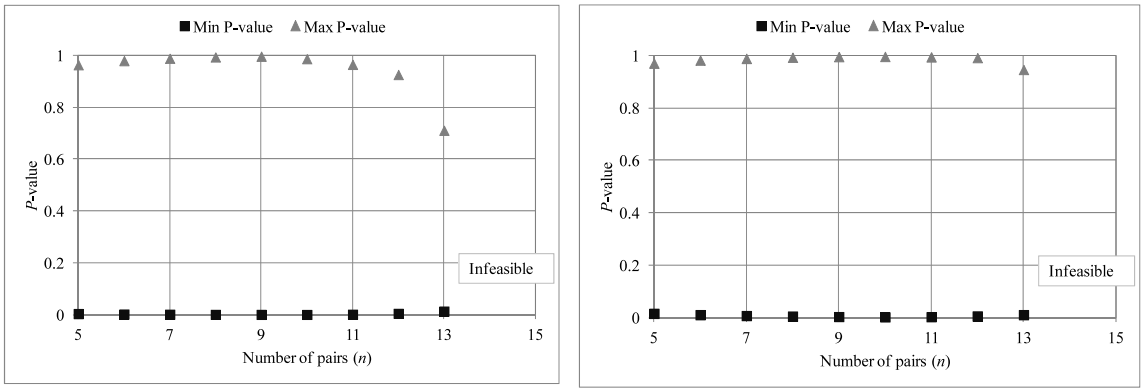

Fig. 1 (Left) Sign test maximum and minimum P-values for different n for bikeshare data with 50 control and treatment points. Each point on the plot corresponds to a set of matched pairs that comes from a solution to a mixed-integer optimization problem. The treatment is mist (vs. clear) weather. The result is inconclusive. (Right) Wilcoxon signed rank test maximum and minimum P-values for different n for bikeshare data, with similar results.

7 Case Studies

Our optimization models have been implemented in AMPL (Fourer et al. 2002), and solved with the solver CPLEX (ILOG 2007). The following two case studies demonstrate our proposed algorithms. Both datasets are publicly available for the purpose of reproducibility. Our code will be made available upon request.

7.1 Case Study 1

Capital Bikeshare is a bike rental company operating in Washington DC. We have two years (2011-2012) of bike sharing data (see Fanaee-T and Gama 2014), comprised of 3,807,587 rental records from 731 days. Each day's data had detailed feature information including the following: Season (Spring; Summer; Fall; Winter), Year (2011; 2012), Workday (No; Yes), Temperature, Humidity, Wind speed. We aimed to test whether misty weather or snow had an effect on the number of bikes rented. We grouped days into those with clear weather (Clear, Few clouds, Partly cloudy, and Cloudy), misty weather (Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist), and snowy weather (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds). The outcome is the number of rental bikes that day. For the one-to-one matched pairs test (the sign test and signed rank test), days were considered to be possible matches if covariates season, year and workday were the same, and the differences in temperature, humidity and wind speed are less or equal to 2, 5 and 5 for clear vs. misty weather experiment, respectively for treated unit i and control unit j, 0 otherwise; these are the days for which $s_{ij} = 1$. In the covariate balance constraint, we have set $b = c$, $\gamma = -1$ and $\beta = 20$ ($\beta = 2$ for clear vs. snowy weather).

Figure 1 (left) shows the sign test results for the bikeshare data, and Figure 1 (right) shows the signed rank test results. The hypothesis is that misty weather influences bike rentals. On the horizontal axis of each plot is the number of matched pairs n, and the two curves on each plot result from the maximization and minimization of the P-value at each n. The result from both tests is inconclusive, where we showed it was possible to obtain almost any P-value along the full range of n where a feasible set of matches exists. In other words, it is possible for a reasonable experimenter to obtain a P-value of almost 0 and another experimenter to obtain a P-value of almost 1. This is the type of uncertainty one would want to know about when conducting matched pairs tests on observational data. The sign test and signed rank tests consider only the median, so let us conduct the other tests to compare other aspects of the distributions.

Figure 2 (left) shows the Wilcoxon-Mann-Whitney U test results, indicating that as the number of points b/c is increased, the maximum and minimum P-values tend to converge to around $p = 0.4$, which in this case means we cannot reject the null hypothesis that misty weather does not cause a decrease

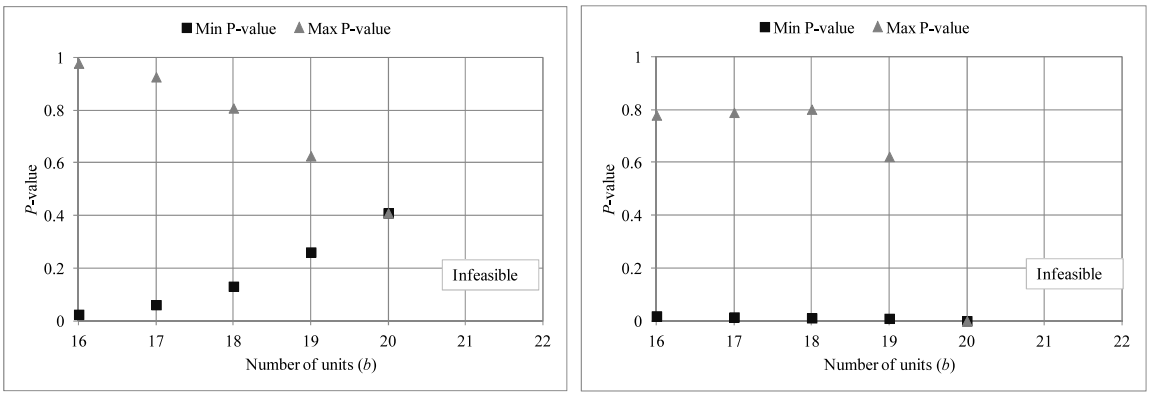

Fig. 2 (Left) Wilcoxon-Mann-Whitney U test maximum and minimum P-values for different b for bikeshare data with 20 control and treatment points. Each point on the plot corresponds to a set of treatment and control samples that are chosen as the solution to a mixed-integer optimization problem. The treatment is mist (vs. clear) weather. The result is that we cannot reject the null hypothesis that the misty weather affects the median number of bikes rented. (Right) Kolmogorov Smirnov Test maximum and minimum P-values for different b for bikeshare data

.

in the number of rental bikes. Figure 2 (right) shows the Kolmogorov-Smirnov test results. The plot illustrates why considering only the largest number of matched pairs (as traditionally done) leaves out a considerable source of uncertainty. If one chooses only the largest number of pairs, there is only one solution and the null hypothesis is rejected even in the most pessimistic case. If one gives the matching algorithm slightly more flexibility by allowing it to choose fewer matches, the range of possible P-value becomes much larger, explicitly showing the uncertainty.

To try to understand why the KS test would reject for larger numbers of pairs, and why the KS test might have different results from the U test, sign test, and signed rank test, we considered the histograms of outcomes for the control and treatment groups provided in Figure 3. The KS test considers differences in CDF, whereas the other tests consider much coarser statistics of the distribution, like differences in median; from the figures, it is clear that the medians of the two distributions are similar, however, the CDF's are very different which could easily explain the apparent discrepancy in results between the tests. Once the number of pairs required is sufficiently large, there would be several treatment outcomes within the low 800-1000 bikes range. The treatment outcomes within this range do not have many corresponding control outcomes (the control histogram values are low within that range). This is possibly why the medians of the distributions could be similar yet the CDF's could be different.

Since mist did not seem to conclusively influence the number of bikes rented in Washington DC, we considered whether snow would have an influence. Figure 4 shows results for the Wilcoxon-Mann-Whitney U test (left) and Kolmogorov-Smirnov test (right). These tests clearly reject the null hypothesis that snowy weather does not affect the number of bikes rented.

7.2 Case Study 2

In this case study we used cardiotocography data from the study of Ayres-de Campos et al. (2000). This dataset was derived from fetal cardiotocograms (CTGs). For each fetus, the available covariates are : AC - number of accelerations per second, FM - number of fetal movements per second, UC - number of uterine contractions per second, DL - number of light decelerations per second, ASTV - percentage of time with abnormal short term variability, MSTV - mean value of short term variability, ALTV - percentage of time with abnormal long term variability, MLTV - mean value of long term variability. The outcome is fetal heart rate and the "treatment" is whether the fetus has a pathologic fetal state, which is measured after birth. Pathologic state is not a true treatment in that there may be no intervention (similarly to issues with testing on gender). The test estimates whether poor fetal state affects the heart rate, all else being equal, controlling for all other covariates. For the one-to-one matched pairs test (the sign test and signed rank test, whose results were inconclusive and shown in Figure 5), babies were considered to be

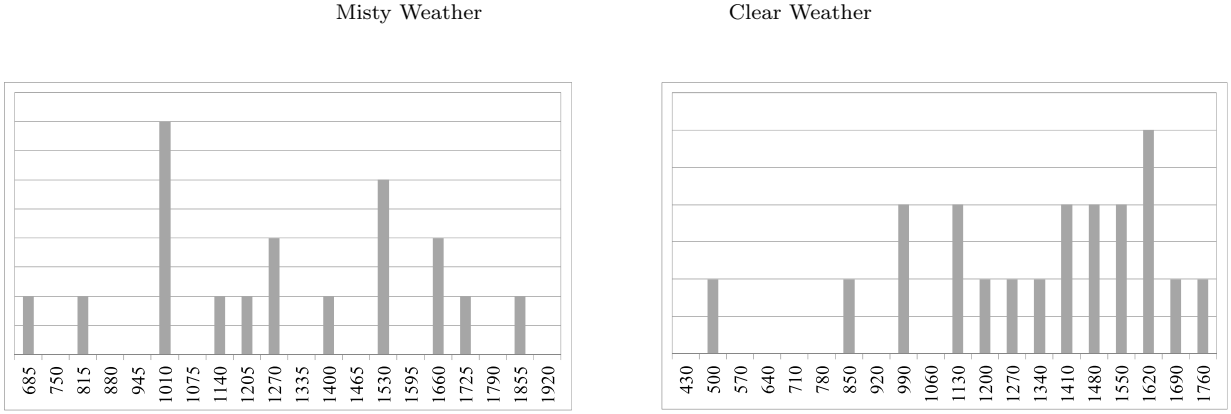

Fig. 3 Histograms of treatment and control outcomes (histogram of number of bikes rented per day in misty weather and in clear weather).

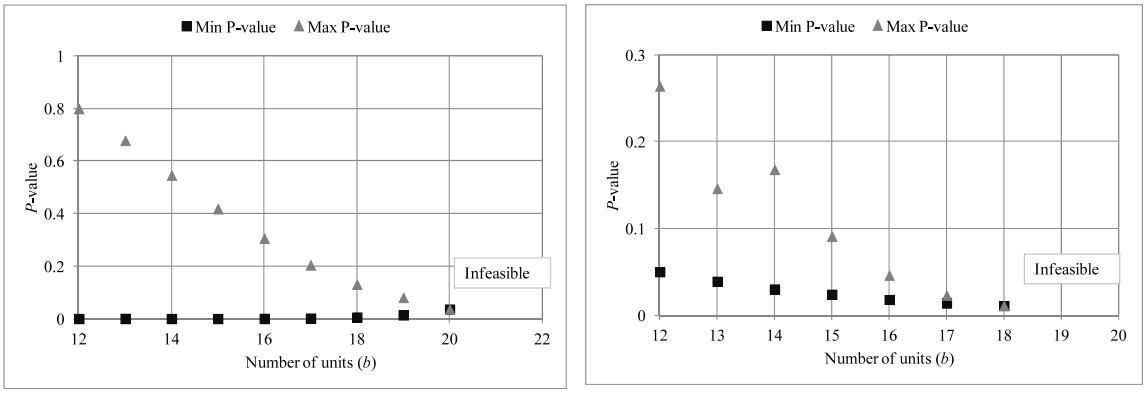

Fig. 4 (Left) Wilcoxon-Mann-Whitney U test maximum and minimum P-values for different b for bikeshare data with 20 control and treatment points. Each point on the plot corresponds to a set of treatment and control samples that are chosen as the solution to a mixed-integer optimization problem. The treatment is snowy (vs. clear) weather. The result is that we can reject the null hypothesis that the snow weather affects the median number of bikes rented. (Right) Kolmogorov Smirnov Test maximum and minimum P-values for different b for bikeshare data.

possible matches if the differences in AC, FM, UC, DL, ASTV, MSTV, ALTV and MLTV are less or equal to 5, 300, 5, 7, 14.14, 5, 7 and 7, respectively for treated unit i and control unit j, 0 otherwise; these are the babies for which $s_{ij} = 1$. Similar to the previous experiments, we have set $b = c$, $\gamma = -100$ and $\beta = 100$ in the covariate balance constraint.

For the Wilcoxon-Mann-Whitney U test shown in Figure 6 (left), the result was inconclusive, meaning it is possible that the distribution of fetal heart rates for the fetuses in poor health is not stochastically different than the distribution of heart rates for the healthy fetuses. The Kolmogorov-Smirnov test (Figure 6, right) robustly indicated that we cannot reject the null hypothesis that the CDF's for the two samples are the same. However, in both cases it seems that if more treatment points were available, the results are less likely to be significant. Viewing trends in the range of significance levels as the amount of data increases can potentially be very useful, as exemplified by both the left and right of Figure 6. Rather than simply reporting one significance level in the range, we can see a much fuller picture of the tests we are conducting.

8 Background Literature and Discussion

Observational data is now pervasive, and it is becoming much more common for companies, medical practitioners, policy makers, and scientists (particularly social scientists like economists and sociologists)

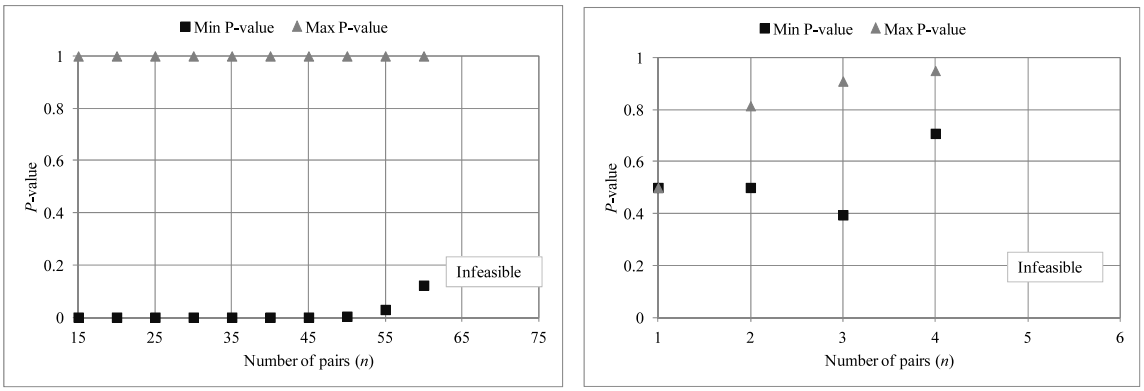

Fig. 5 (Left) Sign test maximum and minimum P-values for different n for full cardiotocography dataset. Each point on the plot corresponds to a set of matched pairs that comes from a solution to a mixed-integer optimization problem. The result is inconclusive. (Right) Similar results for the Wilcoxon signed rank test. This particular test was difficult to solve because it considers ranks of pairs, so it was only able to produce a small number of feasible solutions.

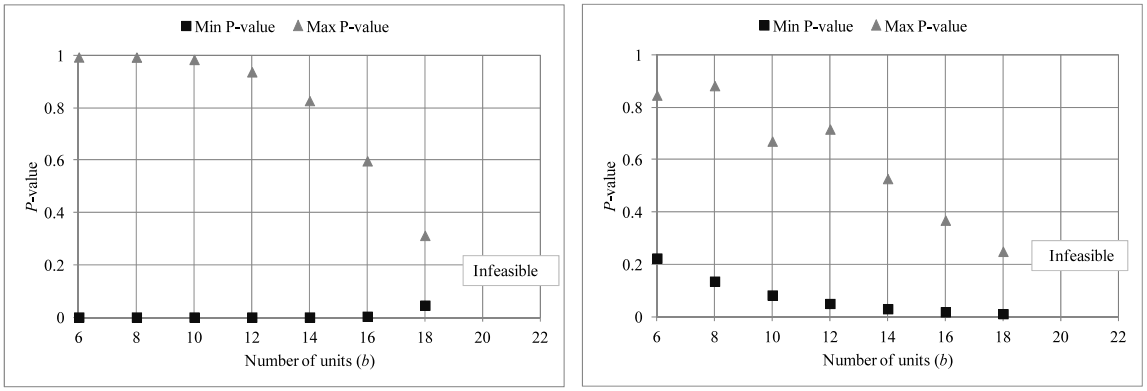

Fig. 6 (Left) Wilcoxon-Mann-Whitney U test optimum P-values for different b for cardiotocography data with 20 control and treatment points. Each point on the plot corresponds to a set of treatment and control samples that are chosen as the solution to a mixed-integer optimization problem. (Right) Similar results for Kolmogorov Smirnov test.

to infer causal effects and make decisions based on these observational data. Matched pairs techniques are used very often, because it helps to reduce confounding, which is caused by the treatment and control populations being different. Treatment and control populations can be very different – for instance, consider the question of whether blood thinners reduce the occurrence of stroke. It might clearly be observed that people *on* blood thinners actually have *more* strokes than people not on blood thinners, but this is misleading. In fact, the people who are given blood thinners are precisely those who are at high risk for stroke. However, if we were to match each person with blood thinners to an almost-identical person without blood thinners, then differences between the outcomes of the two populations are not simply due to bias and thus become meaningful.

Observational studies can allow us to investigate treatment effects when it is not feasible to assign units to treatment and control groups randomly (Rosenbaum 2010). Because the treatment and control populations are often different, matching is commonly done in order to remove the confounding effects of the observed variables (Stuart 2010). Rosenbaum (1989) first proposed to solve the matched pair design problem as an optimal matching problem using network flow theory, as opposed to greedy heuristics. He formulated the optimal matched pair problem as a minimum cost flow problem and compared its performance with the greedy approach for choosing pairs. In some observational studies, it is required to consider full matching, where some matched sets have one control unit and one or more treated unit(s), while other matched sets have one or more control unit(s) and a single treated unit. For such studies, Hansen (2004), Hansen and Olsen Klopfer (2006) developed an equivalent network flow algorithm.

The first line of work to formulate the problem of constructing matched pairs as a mixed integer program is that of Zubizarreta (2012), Zubizarreta et al. (2013, 2014), which provided a substantially broader perspective on the possible ways to construct matches. In particular, Zubizarreta (2012), Zubizarreta et al. (2013, 2014) noted that integer programming would be able to accommodate very complex constraints, such as balance constraints, that would be very difficult to do within the network flow setting. They used interesting objective functions and constraints to construct their matched pairs. They optimized combinations of quality measures for the matches, such as the sum of distances between treatment and control. Tam Cho et al. (2013) also aimed to construct treatment and control sets based on discrete optimization, where balance criteria was optimized, but they used simulated annealing and did not form the problem as a mixed-integer program. All past works on optimization for forming matched pairs construct the matches only from the covariates and not from the outcomes. This is different than our work, where we advocate using the outcomes.

Network flow is limiting, in that it cannot handle complicated constraints. Zubizarreta (2012) provides an important step further than the optimal matching techniques described above, showing that if one is willing to solve an integer program, then a lot of flexibility can be gained in how the matches are constructed; we can then use the full flexibility of ILP to encode logical constraints that affect the matching. In that study, an ILP-based optimal matching method for kidney failure after surgery was described, where the objective is either to minimize both the total sum of distances between treatment and control, and a weighted sum of specific measures of covariate imbalance, or the total sum of distances while constraining the measures of imbalance to be less than or equal to certain tolerances. (These constraints can also be added to the formulations in the present work if desired.) Zubizarreta et al. (2013) used ILP-based matched pair techniques for late preterm birth outcomes, while Zubizarreta et al. (2014) used such techniques to assess effectiveness of for-profit and not-for-profit high schools in Chile. Tam Cho et al. (2013) proposed a discrete optimization model for causal inference problems, where a measure of balance is maximized to find a subset of the treatment group and a subset of the control group, and a simulated annealing algorithm was used for optimization.

9 Conclusions

In an age where important decisions are made based on conclusions from data, we cannot afford to overlook important sources of uncertainty. This work handles one of those sources, namely the uncertainty associated with the procedure for assigning the matches. These robust tests provide a much clearer perspective on what outcomes can be produced. They allow us to know not only that *one* set of matched pairs produced a certain result, but that *all* reasonable sets of matched pairs produced that result. These tests do this in a way that avoids making behavioral assumptions about the way experimenters choose matches.

The contributions of this paper are: (i) This work quantifies an important and overlooked form of uncertainty that can have profound effects on observational studies. (ii) We provide causal inference researchers with a capability they did not have before. Prior to this work it would have been almost impossible to compute the reasonable range of matched pairs test results. (iii) A set of efficient mixed-integer programming formulations that are all *linear* in the decision variables, even though they encode complicated quantities like ranks and cumulative distribution function values.

We have clearly shown in experiments that a single matching procedure can fail miserably to show the necessary information for making policy decisions, whereas results from these new robust tests provide a more complete perspective.

Acknowledgments

The authors express their gratitude to the Natural Sciences and Engineering Research Council of Canada (NSERC) for their financial support of this research.

References

Ayres-de Campos, D., J. Bernardes, A. Garrido, J. Marques-de Sá, L. Pereira-Leite. 2000. Sisporto 2.0: a program for automated analysis of cardiotocograms. *J Matern Fetal Med.* **9**(5) 311–318.

Chen, D.-S., R.G. Batson, Y. Dang. 2011. *Applied Integer Programming: Modeling and Solution*. Wiley.

Fanaee-T, Hadi, Joao Gama. 2014. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence* **2**(2-3) 113–127. URL `http://dx.doi.org/10.1007/s13748-013-0040-3`.

Fourer, R., D.M. Gay, B.W. Kernighan. 2002. Ampl: A modeling language for mathematical programming. Duxbury Press, Cole Publishing Co.

Hansen, Ben B. 2004. Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association* **99** 609–618.

Hansen, Ben B., S. Olsen Klopfer. 2006. Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics* **15**(3) 609–627.

Ho, Daniel E., Kosuke Imai, Gary King, Elizabeth A. Stuart. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* **15**(3 (Summer)) 199–236.

Holland, Paul W. 1986. Statistics and causal inference. *Journal of the American Statistical Association* **81**(396) 945–960.

ILOG. 2007. Cplex 11.0 user's manual. ILOG, Inc.

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

Noor-E-Alam, M., A. Mah, J. Doucette. 2012. Integer linear programming models for grid-based light post location problem. *European Journal of Operational Research* **222**(1) 17–30.

Noor-E-Alam, M., C. Rudin. 2015. Robust testing for causal inference in natural experiments. *Working paper* .

Rosenbaum, Paul R. 2012. Optimal matching of an optimally chosen subset in observational studies. *Journal of Computational and Graphical Statistics* **21** 57–71.

Rosenbaum, P.R. 1989. Optimal matching for observational studies. *Journal of the American Statistical Association* **84** 1024–1032.

Rosenbaum, P.R. 2010. *Design of Observational Studies*. Springer, New York.

Rubin, Donald B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **5**(66) 688–701.

Stuart, E.A. 2010. Matching methods for causal inference: A review and a look forward. *Statistical Science* **25**(1) 1–21.

Tam Cho, W.K., J.J. Sauppe, A.G. Nikolaev, S.H. Jacobson, E.C. Sewell. 2013. An optimization approach for making causal inferences. *Statistica Neerlandica* **67**(2) 211–226.

Tamhane, A.C., D.D. Dunlop. 2000. *Statistics and Data Analysis.*. Prentice Hall, New Jersey.

Winston, W.L., M. Venkataramanan. 2003. *Introduction to Mathematical Programming,(4th ed.)*. Thomson (Chapter 9).

Wolsey, L.A. 1998. *Integer Programming*. Wiley-Interscience, Series in Discrete Mathematics and Optimization, Toronto.

Zubizarreta, J.R. 2012. Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association* **107**(500) 1360–1371.

Zubizarreta, J.R., R.D. Paredes, P.R. Rosenbaum. 2014. Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in chile. *The Annals of Applied Statistics* **8**(1) 204–231.

Zubizarreta, J.R., D.S. Small, N.K. Goyal, S. Lorch, P.R. Rosenbaum. 2013. Stronger instruments via integer programming in an observational study of late preterm birth outcomes. *Annals of Applied Statistics* **7**(1) 25–50.

Supplement A: Integer Linear Programming Basics

Integer linear programming (ILP) techniques have become practical for many large-scale problems over the past decade, due to a combination of increased processor speeds and better ILP solvers. Any type of logical condition can be encoded as linear constraints in an ILP formulation with binary or integer variables. Consider two binary variables $x \in \{0, 1\}$ and $y \in \{0, 1\}$. The logical condition "if $y = 0$ then $x = 0$" can be simply encoded as

$$x \leq y.$$

Note that this condition imposes no condition on x when $y = 1$. Translating if-then constraints into linear constraints can sometimes be more complicated; suppose, we would like to encode the logical condition that if a function $f(w)$ is greater than 0, then another function $g(w)$ is greater or equal to 0. We can use the following two linear equations to do this, where θ is a binary variable and M is a positive number that is larger than the maximum values of both f and g:

$$\begin{aligned} -g(w) &\leq M\theta \\ f(w) &\leq M(1-\theta). \end{aligned}$$

In order for $f(w)$ to be positive, then θ must be 0, in which case, $g(w)$ is then restricted to be positive. If $f(w)$ is negative, θ must be 0, in which case no restriction is placed on the sign of $g(w)$. (See for instance the textbook of Winston and Venkataramanan (2003), for more examples of if-then constraints).

ILP can capture other types of logical statements as well. Suppose we would like to incorporate a restriction such that the integer variable S_i takes a value of K only if $i = t$, and 0 otherwise. The following four if-then constraints can be used to express this statement, where λ_1 and λ_2 are binary variables:

$$\begin{aligned} &\lambda_1 = 1 \text{ if } i + 1 > t \\ &\lambda_2 = 1 \text{ if } t + 1 > i \\ &S_i = k \text{ if } \lambda_1 + \lambda_2 > 1 \\ &S_i = 0 \text{ if } \lambda_1 + \lambda_2 < 2. \end{aligned}$$

Each of these if-then constraints (4)-(7) can be converted to a set of equivalent linear equations, similar to what we described above. (See also Noor-E-Alam et al. (2012) and Winston and Venkataramanan (2003)).

There is no known polynomial-time algorithm for solving ILP problems as they are generally NP-hard, but they can be solved in practice by a number of well-known techniques (Wolsey 1998). The LP relaxation of an ILP provides bounds on the optimal solution, where the LP relaxation of an ILP is where the integer constraints are relaxed and the variables are allowed to take non-integer (real) values. For instance, if we are solving a maximization problem, the solution of the LP relaxation can serve as an upper bound, since it solves a problem with a larger feasible region, and thus attains a value at least as high as that of the more restricted integer program. ILP solvers use branch-and-bound or cutting plane algorithms combined with other heuristics, and are useful for cases where the optimal integer optimal solution is not attained by the LP relaxation. The branch-and-bound algorithms often use LP relaxation and semi-relaxed problems as subroutines to obtain upper bounds and lower bounds, in order to determine how to traverse the branch-and-bound search tree (Chen et al. 2011, Wolsey 1998). The most popular ILP solvers such as CPLEX, Gurobi and MINTO each have different versions of branch-and-bound techniques with cutting plane algorithms and problem-specific heuristics.