Alternative Balance Metrics for Bias Reduction in Matching Methods for Causal Inference*

Jasjeet S. Sekhon†

Version: 1.2 (00:38)

*I thank Alberto Abadie, Jake Bowers, Henry Brady, Alexis Diamond, Jens Hainmueller, Ben Hansen, Guido Imbens, Kosuke Imai, Gary King, Walter Mebane, Donald Rubin and Jonathan Wand for many helpful discussions. Matching software used in this research note is available at `http://sekhon.berkeley.edu/matching`.

†Associate Professor, Travers Department of Political Science, `sekhon@berkeley.edu`, `http://sekhon.berkeley.edu/`, Survey Research Center, 2538 Channing Way, UC Berkeley, Berkeley, CA, 94720.

Abstract

Sekhon (2006; 2004a) and Diamond and Sekhon (2005) propose a matching method, called Genetic Matching, which algorithmically maximizes the balance of covariates between treatment and control observations via a genetic search algorithm (Sekhon and Mebane 1998). The method is neutral as to what measures of balance one wishes to optimize. By default, cumulative probability distribution functions of a variety of standardized statistics are used as balance metrics and are optimized without limit. The statistics are not used to conduct formal hypothesis tests, because no measure of balance is a monotonic function of bias in the estimand of interest and because we wish to maximize balance. Descriptive measures of discrepancy generally ignore key information related to bias which is captured by probability distribution functions of standardized test statistics. For example, using several descriptive metrics, one is unable reliably to recover the experimental benchmark in a testbed dataset for matching estimators (Dehejia and Wahba 1999). And these metrics, unlike those based on optimized distribution functions, perform poorly in a series of Monte Carlo sampling experiments just as one would expect given their properties.

Matching has become an increasingly popular method of causal inference in many fields including statistics (Rubin 2006; Rosenbaum 2002), medicine (Christakis and Iwashyna 2003; Rubin 1997), economics (Abadie and Imbens 2006; Galiani, Gertler, and Schargrodsky 2005; Dehejia and Wahba 2002, 1999), political science (Bowers and Hansen 2005; Herron and Wand 2006; Imai 2005; Sekhon 2004b), sociology (Morgan and Harding 2006; Diprete and Engelhardt 2004; Winship and Morgan 1999; Smith 1997) and even law (Rubin 2001). There is, however, no consensus on how exactly matching ought to be done, how to measure the success of the matching procedure, and whether or not matching estimators are sufficiently robust to misspecification so as to be useful in practice (Arceneaux, Gerber, and Green 2006; Heckman, Ichimura, Smith, and Todd 1998).

Sekhon (2006; 2004a) and Diamond and Sekhon (2005) propose a matching algorithm, called Genetic Matching (GenMatch), which maximizes the balance of observed covariates between treated and control groups. GenMatch is a generalization of propensity score and Mahalanobis distance matching, and it has been used by a variety of researchers (e.g., Brady and Hui 2006; Gilligan and Sergenti 2006; Gordon and Huber 2006; Herron and Wand 2006; Morgan and Harding 2006; Lenz and Ladd 2006; Park 2006; Raessler and Rubin 2005). The algorithm uses a genetic algorithm (Mebane and Sekhon 1998; Sekhon and Mebane 1998) to optimize balance as much as possible given the data. The method is nonparametric and does not depend on knowing or estimating the propensity score, but the method is improved when a propensity score is incorporated. Diamond and Sekhon (2005) use this algorithm to show that the long running debate between Dehejia and Wahba (2002; 1997; 1999; Dehejia 2005) and Smith and Todd (2005b,a, 2001) is largely a result of researchers using models which do not produce good balance—even if some of the models get close by chance to the experimental benchmark of interest. They show that Genetic Matching is able to quickly find good balance and reliably recover the experimental benchmark.

The Genetic Matching algorithm is neutral as to what metric is used to evaluate balance. The usual practice of using hypothesis tests to determine if observed covariates are balanced

between matched and treated groups is seriously flawed in part because we are strictly interested in the current sample and not in inferences about a population or superpopulation when evaluating balance and because we wish to maximize balance without limit. However, it is useful to optimize cumulative probability distribution functions of standardized test statistics in the context of Genetic Matching as long as one does so without limit.[1] Ho, Imai, King, and Stuart (2007) go further and argue that instead of using standardized quantities (such as t-statistics), researchers should use difference of means directly.[2] Ho et al. (2007) caution against the use of hypothesis tests because such tests lose power with reductions in sample size, are sensitive to changes in the ratio between treated and control groups and can be made to look better by increasing variance when the mean gap remains the same or even when it is increased.

Although these concerns are important and all too often ignored by researchers, they do not lead to the conclusion that one should optimize descriptive metrics instead of probability distribution functions of standardized test statistics. The probability distributions associated with standardized statistics do change with the sample size, and this is an issue which can make it problematic to compare tests before and after matching. As Lee (2006) discusses, we need to better distinguish between balance tests that are conducted before and after matching. They simply serve different purposes. But this problem may be avoided simply by making sure that the ratio of treated to control units and the sample size are held constant across the matched datasets being compared as is the case with Genetic Matching.[3]

There are serious and open issues in the matching literature on the question of how to think about matching when observations are dropped. For example, what is the estimand after observations are dropped? The situation is clear enough when we restrict attention to the average treatment effect for the treated units (ATT) instead of the average treatment

[1] Moreover, after GenMatch optimization, the p-values from balance tests cannot be interpreted as true probabilities because of standard pre-test problems, but they remain useful measures of balance.

[2] Also see Imai, King, and Stuart (2006).

[3] GenMatch can be used in combination with a caliper but this raises a number of issues because different observations will be dropped as the algorithm progresses. Hence, the estimand will change arbitrarily which is an obvious problem.

effect (ATE) or we restrict attention to white females and drop Asian males because we cannot find adequate matches for them. But often we drop units based on the propensity score so the population we are making an inference about becomes unclear—e.g., some Asian males will be dropped but not others depending on the values of other covariates. In such cases what we are estimating becomes unclear. The difficulty of the problem can be seen by noting that if there are only two observations in a dataset which are perfectly matched—one in treatment and one in control—an algorithm which is allowed to drop observations without limit to maximize balance will drop all but these two.[4] Since such a wide search is not well posed, Genetic Matching does optimization after the analyst has fixed the data, chosen the estimand and the ratio of treated to control—i.e., whether this means, for example, estimating ATT with 1-to-1 matching with replacement or ATE with optimal matching (Hansen 2004; Gu and Rosenbaum 1993; Rosenbaum 1991, 1989). After these decisions, Genetic Matching can be used to optimize balance.

However, even with the sample size and ratio of treatment to control fixed, the cumulative probability distribution function of statistics such as a t-test can become uninformative if a matching algorithm simply increases variance and does not decrease the mean discrepancy. Therefore it is important to either normalize the t-statistic by an unchanging quantity such as the variance of the ATT observations and/or to also test to make sure that the variances are made equal.

The central problem is that no single metric will be equally sensitive to all departures of balance. The problem is more general than just that of inflated variances. Every metric can be fooled by making some other function of balance worse—often much worse. For example, if one only examines the simple difference of means, one can be similarly tricked. An algorithm could decrease the mean gap between variables but markedly increase the gap in some quantile of interest such as the maximum gap.

[4]One possibility is to explicitly write down a loss function which takes into account the tradeoff between internal and external validity in a study due to limited overlap in the distribution of covariates (Crump, Hotz, Imbens, and Mitnik 2006).

The most important point to keep in mind is that we are matching not simply to increase balance, but to reduce bias in a causal estimand we wish to estimate. In no case is this bias simply a function of the mean gap between the confounders. Even in the simplest case where the relationship between a single confounder the outcome is linear, the bias due to imperfect matching is a function of both the covariance between the confounder and treatment assignment *and* the variance of the confounder.

In the case where there are two confounders interesting tradeoffs emerge when deciding what the best set of matches is. For example, the set of matches which result in *higher* mean differences averaged over two different confounders may result in less bias even if the bias is linear. This can occur if the unstandardized difference in means in the first confounder is larger than the unstandardized difference in the second confounder, but the relative sizes of the standardized differences are reversed. In this case, in order to minimize bias, one may be better off being more concerned about the second variable than the first and hence using the matches with higher average mean difference but smaller standardized average difference.

Thus, measures which depend on at least the first two moments are to be preferred. And ideally we would like measures for all of the higher moments and measures which are multidimensional. Useful approximations for the latter measures are not feasible to compute unless one has vast amounts of data, but measures of the former are. By default GenMatch maximizes the Kolmogorov-Smirnov (KS) test (along with the paired t-test) without limit, but there are alternatives. The descriptive analogs to the KS test are functions of the empirical quantile-quantile plots which compare the empirical distributions of a given covariate between treatment and control groups—eQQ plots. And GenMatch is happy to optimize these instead.

Functions of eQQ plots such as the mean differences in the eQQ plots have, however, a serious drawback. These functions do not distinguish between departures from perfect balance which are all skewed positive versus those which are all skewed negative versus those which are half and half. Thus, for the same value of the mean eQQ discrepancy one could see

a plot where the variable is more commonly larger for treatment than control or where errors tend to go both ways and hence are more likely to cancel each other out. Descriptive metrics, such as mean eQQ or max eQQ, are invariant to these considerations. In contrast, tests that compare cumulative distributions such as the KS test are sensitive to these distributional features. In this sense, descriptive measures are leaving information on the table which is relevant for the bias in the estimand of interest. In other words, not all imbalances with the same mean gap in eQQ plots are of equal concern.

An example is always helpful, and top panel in Figure 1 plots bias in the causal estimate in the Dehejia and Wahba (1999) sample as a function of a hypothesis test metric. The balance metric is the smallest p-value across a large number of tests conducted for all univariate baseline covariates, as well as their first-order interactions and quadratic terms. This figure shows the relationship between the fitness value (the lowest p-value obtained, after Genetic Matching, from covariate-by-covariate paired t- and KS-tests across all covariates' interaction and quadratic terms) and the causal estimate. Each point represents an attempt at matching resulting in a balance test and an estimate of causal effect with estimates distributed above and below the target experimental result. Note that it is possible to get lucky and produce a reliable result even when balance has not been attained, but there is no reason to pick these matches when we don't know the experimental benchmark. Reliable results are obtained only at the highest fitness values and we know these fitness values not only when we don't know the experimental benchmark, but also when we don't even know the estimand of interest. The best fit has a p-value of 0.21 and it produces an estimate of \$1777.762 which is very close to the experimental benchmark of \$1794.34.

The bottom panel in Figure 1 plots the same set of matches as the previous plot but where the matches are sorted by the mean standardized eQQ across covariates. The eQQ metric is scaled by a multiplication by -1 so that the plot is more comparable with the previous one—i.e., the right most points have the best fit as measured by the metric. This metric performs poorly. The best fit has an estimate of \$923.54 while the experimental benchmark

is \$1794.34.[5] Other functions of the eQQ plots perform even worse, including the maximum average gap which is shown in the Monte Carlo sampling experiment discussed below.

Tables 1 and 2 present the results of two Monte Carlo sampling experiments which reinforce the results presented in Figure 1. The first experimental setup is designed so that the conditions of Rubin and Thomas (1992a,b) are satisfied. The baseline covariates are multivariate normal, the mapping between the baseline covariates and the outcome is linear, and all of the confounders are observed. Therefore, the Equal Percent Bias Reduction (EPBR) property holds and both the propensity score and Mahalanobis distance matching should perform relatively well. In the second experiment, conditions for EPBR are not met. It is a difficult matching problem designed to be as similar as possible to the Dehejia and Wahba sample. In particular, the baseline covariates do not have symmetric ellipsoidal distributions. The covariates are taken to be exactly those in the Dehejia and Wahba (1999) experimental sample of the LaLonde (1986) data. These covariates include discrete variables as well as semi-continuous variables with significant skew. And the mapping between the baseline covariates and the outcome is not linear. But all of the confounders are observed so a matching method *could* work well. The details of the experimental designs are described in the Appendix.

Table 1 presents results of the first Monte Carlo experiment. The first column of the table presents the mean estimate of a given estimator and the second column the root mean square error over 1000 Monte Carlo samples. The true estimate is \$0 in this experiment. The default (p-value based) GenMatch has, as expected, the lowest mean square error. The "Raw" estimate refers to the naive unadjusted ATE which is simply, in a given sample, the mean treatment outcome minus the mean control outcome. The raw bias is 60.4. The bias of Mahalanobis distance matching is -8.63, for the joint propensity score Mahalanobis distance estimator the bias is -5.96, and for propensity score matching the bias is -2.45.

[5]The set of matches in both figures was generated by randomly picking variables weights, by running Genetic Matching with the p-value balance metric and then by running Genetic Matching with the mean eQQ metric. Both plots contain the same points. They are only sorted differently. For details and software see Sekhon (2006, 2004a) and `http://sekhon.berkeley.edu/matching`.

Default GenMatch has a bias of -2.47, which is the second lowest, almost indistinguishable from the bias of propensity score matching. Consistent with the results of Rosenbaum and Rubin (1983), the estimator with the second lowest root means square error is the joint propensity score Mahalanobis distance estimator. The results of Genetic Matching runs with two different functions of eQQ are also presented. In the first case, GenMatch is asked to minimize the mean of the average standardized eQQ discrepancy across covariates (eQQ mean). In the second case, GenMatch is asked to minimize the maximum of the average standardized eQQ discrepancy across covariates (eQQ max).

The last two columns of Table 1 present the ratios of bias and root mean square error of a given estimator relative to GenMatch. Both of the eQQ GenMatch runs perform worse in terms of both bias and RMSE than default GenMatch. The mean eQQ run has 3.21 times the bias of default GenMatch and 1.25 times the RMSE. The max eQQ run has 3.51 times the bias and 1.29 times the RMSE. Both of the eQQ runs also perform worse in terms of bias than the propensity score model (more than three times the bias) and worse than the pscore model when it is combined with Mahalanobis distance.

The two standard multivariate matching methods, Mahalanobis distance and the joint propensity score plus Mahalanobis distance estimator, have significantly larger bias than default GenMatch: Mahalanobis distance has 3.5 times the bias and the joint estimator 2.4 times the bias. Propensity score matching, however, only has .993 times the bias of GenMatch. This is to be expected because as Abadie and Imbens (2006) prove, if one is matching on more than one continuous variable the bias is not $\sqrt{n}$ consistent. What **is** surprising is how close the GenMatch bias is to that of propensity score matching.

The Mahalanobis distance mean square error is 1.75 times as large as that of GenMatch, for the propensity score estimator it is 2.57 times as large, and for the joint estimator it is 1.49 times as large. Although the propensity score estimator has a slightly lower bias than GenMatch (its bias is 0.993 times that of GenMatch), its mean square error is 2.57 times as large. The GenMatch dominates all of the other multivariate matching methods both in

terms of bias and MSE, and dominates propensity score matching in terms of MSE.

Table 2 presents the results for the second Monte Carlo experiment. In this experiment the EPBR conditions do not hold and the propensity score is misspecified. Default GenMatch now clearly dominates all other estimators both in terms of means square error and bias. The other matching estimators have a bias which range from 10 times to 28 times as large as that of default GenMatch. And the RMSE of the other matching estimators ranges from 1.22 times to 2.84 times as large as that of GenMatch. In this case, unlike the previous Monte Carlo, the two eQQ GenMatch runs perform better than the standard methods, but as in the first experiment, they perform worse than the standard GenMatch run both in terms of bias (more than 10 times as large) and RMSE (about 1.2 to 1.27 times as large).

Default GenMatch is the only matching method which, across samples, produces a reliable estimate of the true effect. The true causal estimate is \$1000 and the average GenMatch bias is only \$25.6 (2.56%), with root mean square error of 455. The mean eQQ GenMatch run has a bias of \$261 (26.1%) and the max eQQ run has a bias of \$294 (29.4%).

Conclusion

The literature on matching, particularly the debate over the LaLonde (1986) data, makes clear the need to find algorithms which produce matched datasets with high levels of covariate balance. The fact that so many talented researchers over several years failed to produce a propensity score model which had a high degree of covariate balance is a cautionary tale. In situations like these, machine learning can come to the rescue. There is little reason for a human to try all of the multitude of models possible to achieve balance when a computer can do this more systematically and much faster. Even talented and well trained researchers need aid.

There is, however, no perfect measure of balance. No measure is equally sensitive to all departures from balance so it is important to test balance in a variety of ways. The software

package `Matching` (Sekhon 2006, 2004a) provides a large number of balance measures ranging from t-tests to various functions of the raw and standardized eQQ plots. The best choice of what metric to use is obviously dependent on the precise application. And the `GenMatch()` function in `Matching` allows the user to use an arbitrary function of balance or to select from a list of measures to optimize. `GenMatch` also allows the user to give more weight to some covariates than others. In this way, the researcher is able to bring her substantive knowledge to the problem. In general, however, I suspect that we do not know as often as we would like what the relative importance of the possible confounders is. The previous outcome (if observed) is often the most important covariate, but beyond that we are usually ignorant.

Finally, one may wish that balance tests should not make reference to speculative superpopulations and should be consistent with the design that treatment assignment is the only source of randomness. Hansen (2006) discusses permutation tests when matching is done without replacement, and Lee (2006) discusses such tests for when matching is done with replacement. These proposals are promising, and their performance should be carefully evaluated.

A Design of Monte Carlos

Two different Monte Carlo experiments are presented. In the first, the experimental conditions satisfy assumptions outlined in Rubin and Thomas (1992a). In this experiment, all of the confounders are observed, they are distributed following a normal distribution, and the mapping between X and Y is linear. The propensity score is reestimated in each Monte Carlo sample for efficiency reasons (Rosenbaum and Rubin 1983).

In the second Monte Carlo experiment, the assumptions required for EPBR are not satisfied. This experiment is a difficult case for matching. Some of the baseline variables are discrete and others contain point masses and skewed distributions. The propensity score is not correctly specified, and the mapping between X and Y is nonlinear. But all confounders

are observed. One thousand Monte Carlo samples are performed for both experiments.

For each Monte Carlo sample in Experiment 1, there are 50 treated observations and 100 control observations. There are three baseline covariates all of which are normally distributed with variance 1 and zero covariances. The baseline covariates for the treated observations all have means equal to zero and the covariates for the control group all have means equal to 0.2. The effect of treatment is zero and the outcome, Y, is generated as follows:

$$Y = X\beta + \epsilon$$

where $\epsilon \sim N(0, .5)$ and all of the β parameters are equal to 1.

For the second Monte Carlo sampling experiment, the distribution of covariates was chosen to make the setting as realistic as possible, with variables taken from the Dehejia and Wahba (1999) experimental sample of the LaLonde (1986) data. There are eight baseline variables, none of which have ellipsoidal distributions. They are age, years of education, real earnings in 1974, real earnings in 1975 and a series of indicator variables. The indicator variables are Black, Hispanic, married and high school diploma. The two earnings variables have large point masses at zero, have fat tails and are heavily skewed distributions. Given this, the EPBR property is unlikely to hold. In this simulation we assume a homogeneous treatment effect of \$1000. The equation that determines outcomes Y (fictional earnings) is:

$$Y = 1000\, T + .1 \exp\left[.7 \log(\text{re74} + .01) + .7 \log(\text{re75} + 0.01)\right] + \epsilon$$

where $\epsilon \sim N(0, 10)$, re74 is real earnings in 1974, re75 is real earnings in 1975 and T is the treatment indicator. The mapping from baseline covariates to Y is obviously nonlinear and only two of the baseline variables are directly related to Y.

The true propensity score for each observation, π_i, is defined by:

$$\pi_i = \text{logit}^{-1}\left[1 + .5\hat{\mu} + .01\ \text{age}^2 - .3\ \text{educ}^2 - .01 \log(\text{re74} + .01)^2 + .01 \log(\text{re75} + .01)^2\right]$$

where $\hat{\mu}$ equals the linear predictor obtained by estimating a logistic regression model, where the dependent variable is the actually observed treatment indicator in the Dehejia Wahba (1999) experimental sample of the LaLonde (1986) data. The true propensity score in the Monte Carlo experiment is a mix of the estimated propensity score in the Dehejia and Wahba sample plus extra variables in Equation 1, because we want to ensure that the propensity model estimated in the Monte Carlos samples would be badly misspecified. The linear predictor is:

$$\begin{aligned}
\hat{\mu} &= 1 + 1.428 \times 10^{-4}\text{age}^2 - 2.918 \times 10^{-3}\text{educ}^2 - .2275 \text{ black} + -.8276 \text{ Hisp} \\
&+ .2071 \text{ married} - .8232 \text{ nodegree} - 1.236 \times 10^{-9}\text{re74}^2 + 5.865 \times 10^{-10}\text{re75}^2 \\
&- .04328 \text{ u74} - .3804 \text{ u75}
\end{aligned}$$

where u74 is an indicator variable for if real earnings in 1974 are zero and u75 is an indicator variable for if real earnings in 1975 are zero.

In each Monte Carlo sample of this experiment, the propensity score is estimated using logistic regression and the following incorrect functional form:

$$\begin{aligned}
\hat{\mu}^* &= \alpha + \alpha_1 \text{ age} + \alpha_2 \text{ educ} + \alpha_3 \text{ black} + \alpha_4 \text{ Hisp} \\
&+ \alpha_5 \text{ married} + \alpha_6 \text{ nodegree} + \alpha_7 \text{ re74} + \alpha_8 \text{ re75} \\
&+ \alpha_9 \text{ u74} + \alpha_{10} \text{ u75}
\end{aligned}$$

References

Abadie, Alberto and Guido Imbens. 2006. "Large Sample Properties of Matching Estimators for Average Treatment Effects." *Econometrica* 74: 235–267.

Arceneaux, Kevin, Alan S. Gerber, and Donald P. Green. 2006. "Comparing Experimental and Matching Methods Using a Large-Scale Voter Mobilization Experiment." *Political*

Analysis 14: 37–62.

Bowers, Jake and Ben Hansen. 2005. "Attributing Effects to a Get-Out-The-Vote Campaign Using Full Matching and Randomization Inference." Working Paper.

Brady, Henry and Iris Hui. 2006. "Is it Worth Going the Extra Mile to Improve Causal Inference?"." Paper presented at the 23rd Annual Summer Meeting of the Society of Political Methodology.

Christakis, Nicholas A. and Theodore I. Iwashyna. 2003. "The Health Impact of Health Care on Families: A matched cohort study of hospice use by decedents and mortality outcomes in surviving, widowed spouses." *Social Science & Medicine* 57 (3): 465–475.

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

Dehejia, Rajeev. 2005. "Practical Propensity Score Matching: A Reply to Smith and Todd." *Journal of Econometrics* 125 (1–2): 355–364.

Dehejia, Rajeev and Sadek Wahba. 1997. "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs." Rejeev Dehejia, *Econometric Methods for Program Evaluation*. Ph.D. Dissertation, Harvard University, Chapter 1.

Dehejia, Rajeev and Sadek Wahba. 1999. "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94 (448): 1053–1062.

Dehejia, Rajeev H. and Sadek Wahba. 2002. "Propensity Score Matching Methods for Nonexperimental Causal Studies." *Review of Economics and Statistics* 84 (1): 151–161.

Diamond, Alexis and Jasjeet S. Sekhon. 2005. "Genetic Matching for Estimating Causal

Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies." Working Paper. See http://sekhon.berkeley.edu/papers/GenMatch.pdf.

Diprete, Thomas A. and Henriette Engelhardt. 2004. "Estimating Causal Effects With Matching Methods in the Presence and Absence of Bias Cancellation." *Sociological Methods & Research* 32 (4): 501–528.

Galiani, Sebastian, Paul Gertler, and Ernesto Schargrodsky. 2005. "Water for Life: The Impact of the Privatization of Water Services on Child Mortality." *Journal of Political Economy* 113 (1): 83–120.

Gilligan, Michael J. and Ernest J. Sergenti. 2006. "Evaluating UN Peacekeeping with Matching to Improve Causal Inference." Working paper.

Gordon, Sandy and Greg Huber. 2006. "The Effect of Electoral Competitiveness on Incumbent Behavior." Working paper.

Gu, X. and Paul Rosenbaum. 1993. "Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms." *Journal of Computational and Graphical Statistics* 2: 205–420.

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. 2006. "Appraising covariate balance after assignment to treatment by groups." University of Michigan, Technical Report #436.

Heckman, James J., Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66 (5): 1017–1098.

Herron, Michael C. and Jonathan Wand. 2006. "Assessing Partisan Bias in Voting Technology: The Case of the 2004 New Hampshire Recount." *Electoral Studies*. Forthcoming.

Ho, Daniel, Kosuke Imai, Gary King, and Elizabeth Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis*.

Imai, Kosuke. 2005. "Do Get-Out-The-Vote Calls Reduce Turnout? The Importance of Statistical Methods for Field Experiments." *American Political Science Review* 99 (2): 283–300.

Imai, Kosuke, Gary King, and Elizabeth A. Stuart. 2006. "The Balance Test Fallacy in Matching Methods for Causal Inference." Working Paper.

LaLonde, Robert. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76 (September): 604–20.

Lee, Wang-Sheng. 2006. "Propensity Score Matching and Variations on the Balancing Test." Working Paper.

Lenz, Gabriel S. and Jonathan McDonald Ladd. 2006. "Exploiting a Rare Shift in Communication Flows: Media Effects in the 1997 British Election." Working paper.

Mebane, Walter R. Jr. and Jasjeet S. Sekhon. 1998. "GENetic Optimization Using Derivatives (GENOUD)." Software Package. `http://sekhon.berkeley.edu/rgenoud/`.

Morgan, Stephen L. and David J. Harding. 2006. "Matching Estimators of Causal Effects: Prospects and Pitfalls in Theory and Practice." *Sociological Methods & Research* 35 (1): 3–60.

Park, Johng Hee. 2006. "Causal Effect of Information on Voting Behavior from a Natural Experiment: An Analysis of Candidate Blacklisting Campaign in 2000 South Korean National Assembly Election." Working paper.

Raessler, S. and D. B. Rubin. 2005. "Complications when using nonrandomized job training data to draw causal inferences." *Proceedings of the International Statistical Institute*.

Rosenbaum, Paul R. 1989. "Optimal Matching for Observational Studies." *Journal of the American Statistical Association* 84 (408): 1024–1032.

Rosenbaum, Paul R. 1991. "A Characterization of Optimal Designs for Observational Studies." *Journal of the Royal Statistical Society, Series B* 53 (3): 597–610.

Rosenbaum, Paul R. 2002. *Observational Studies*. New York: Springer-Verlag 2nd edition.

Rosenbaum, Paul R. and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55.

Rubin, Donald B. 1997. "Estimating Causal Effects from Large Data Sets Using Propensity Scores." *Annals of Internal Medicine* 127 (8S): 757–763.

Rubin, Donald B. 2001. "Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation." *Health Services & Outcomes Research Methodology* 2 (1): 169–188.

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

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

Rubin, Donald B. and Neal Thomas. 1992b. "Characterizing the Effect of Matching Using Linear Propensity Score Methods with Normal Distributions." *Biometrika* 79 (4): 797–809.

Sekhon, Jasjeet S. 2004a. "Matching: Multivariate and Propensity Score Matching with Balance Optimization." Software available at `http://sekhon.berkeley.edu/matching/`.

Sekhon, Jasjeet S. 2004b. "The Varying Role of Voter Information Across Democratic Societies." Working Paper.
URL `http://sekhon.berkeley.edu/papers/SekhonInformation.pdf`

Sekhon, Jasjeet S. 2006. "Algorithms for Multivariate and Propensity Score Matching with Balance Optimization via Genetic Search." Working Paper.

Sekhon, Jasjeet Singh and Walter R. Mebane, Jr. 1998. "Genetic Optimization Using Derivatives: Theory and Application to Nonlinear Models." *Political Analysis* 7: 189–203.

Smith, Herbert L. 1997. "Matching with Multiple Controls to Estimate Treatment Effects in Observational Studies." *Sociological Methodology* 27: 305–353.

Smith, Jeffrey and Petra Todd. 2005a. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* 125 (1–2): 305–353.

Smith, Jeffrey and Petra Todd. 2005b. "Rejoinder." *Journal of Econometrics* 125 (1–2): 365–375.

Smith, Jeffrey A. and Petra E. Todd. 2001. "Reconciling Conflicting Evidence on the Performance of Propensity Score Matching Methods." *AEA Papers and Proceedings* 91 (2): 112–118.

Winship, Christopher and Stephen Morgan. 1999. "The estimation of causal effects from observational data." *Annual Review of Sociology* 25: 659–707.

Table 1: Experimental Condition 1: Multivariate Normal Covariates

Estimator	Bias	RMSE	$\frac{\text{Bias}}{\text{Bias Genmatch}}$	$\frac{\text{RMSE}}{\text{RMSE Genmatch}}$
GenMatch	−2.47	13.1		
mean eQQ	−7.94	16.4	3.21	1.25
max eQQ	−8.66	16.9	3.51	1.29
Pscore	−2.45	21.0	.993	1.61
Mahalanobis (MH)	−8.63	17.3	3.50	1.33
Pscore + MH	−5.96	16.0	2.41	1.21
Raw	−60.4	68.6	24.6	5.31

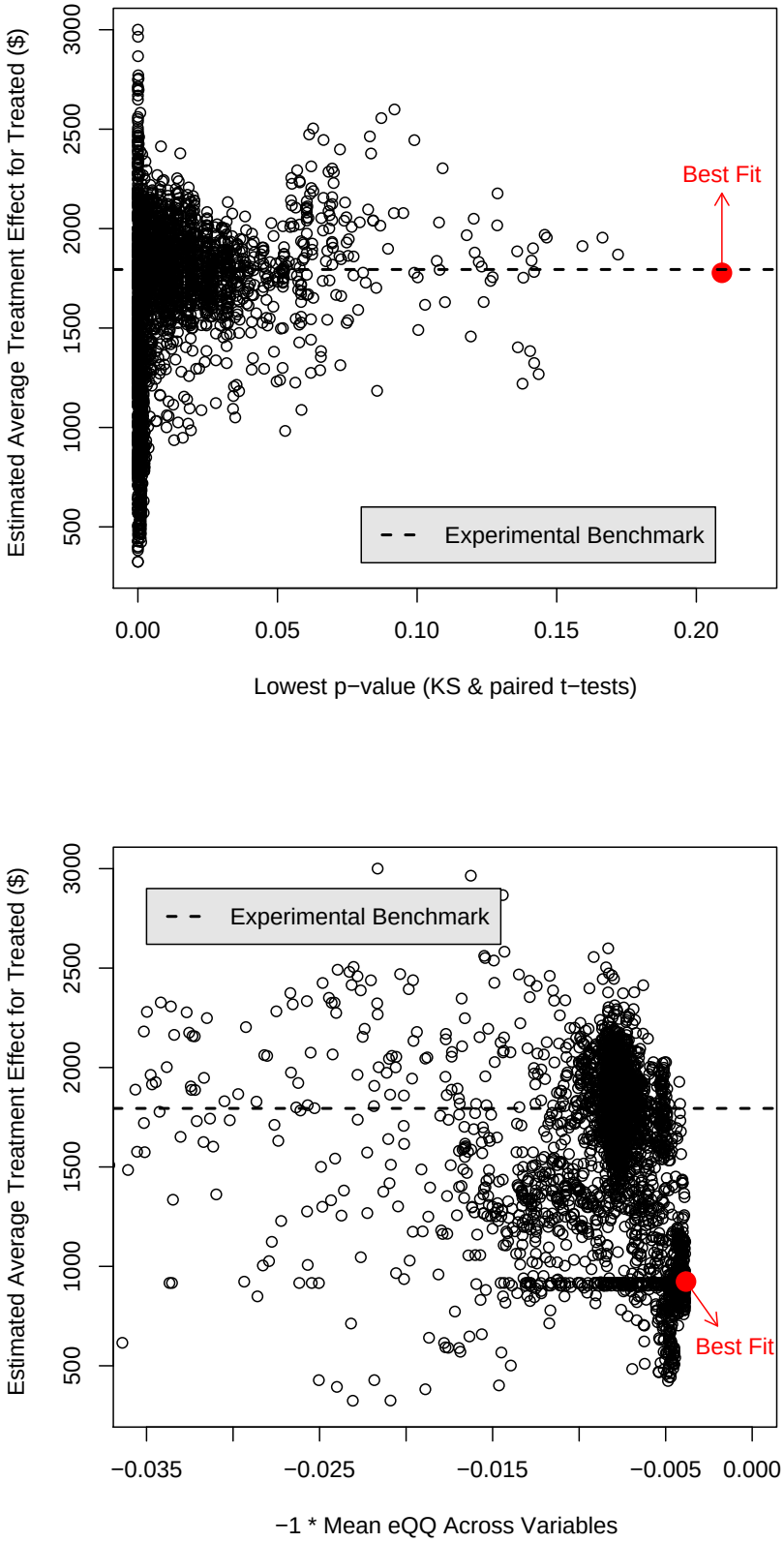
The true treatment effect is \$0. The last three columns present the ratios of bias and RMSE of a given estimator relative to GenMatch using the hypothesis test metric. The experiment is done under assumptions which satisfy the conditions outlined in Rubin and Thomas (1992a,b). In particular, the baseline covariates are multivariate normal. And the mapping between the baseline covariates and the outcome is linear.

Table 2: Experimental Condition 2: Distribution of LaLonde Covariates

Estimator	Bias	RMSE	$\frac{\text{Bias}}{\text{Bias Genmatch}}$	$\frac{\text{RMSE}}{\text{RMSE Genmatch}}$
GenMatch	25.6	455		
mean eQQ	261	553	10.1	1.22
max eQQ	294	580	11.5	1.27
Pscore	512	1294	20.0	2.84
Mahalanobis (MH)	−717	959	28.0	2.11
Pscore + MH	428	743	16.8	1.63
Raw	485	1611	19.0	3.54

The true treatment effect is $1000. The last two columns present the ratios of bias and RMSE of a given estimator relative to GenMatch using hypothesis test metric. The experiment is done under assumptions which do **not** satisfy the conditions outlined in Rubin and Thomas (1992a,b). In particular, the baseline covariates do not have symmetric ellipsoidal distributions. The covariates include discrete variables as well as semi-continuous variables with significant skew. And the mapping between the baseline covariates and the outcome is not linear.

Figure 1: Hypothesis Test vs. eQQ Metric



Each point represents a GenMatch evaluation. Estimates are accurate at high fitness values for the hypothesis test metric (top panel). Both metrics are scaled so that the right most points have the best fit.