The Balance-Sample Size Frontier in Matching Methods for Causal Inference: Supplementary Appendix

Gary King* Christopher Lucas† Richard Nielsen‡

March 22, 2016

Abstract

This is a supplementary appendix to Gary King, Christopher Lucas, and Richard Nielsen, "The Balance-Sample Size Frontier in Matching Methods for Causal Inference," forthcoming, *American Journal of Political Science*.

A Inspecting the Means of Pruned Observations

In our main paper, Figures 3 and 5 display scaled covariate means across the Lalonde (1986) and the Boyd, Epstein and Martin (2010) frontiers for observations *remaining* in the matched samples. However, a researcher might also be interested in a description of the pruned data; that is, the covariate means for observations removed from the sample. In this section, we create these plots for both illustrations using. To make this calculation, we use the `MatchingFrontier` software package (King, Lucas and Nielsen, 2015).

A.1 Job Training

As a companion to the analysis displayed in Section 6.1, Figure 1 presents covariate means for *pruned* observations at each point on the frontier. Note that this differs from Figure 3 in the main text, which presents the means for observations remaining in the matched sample, rather than those pruned. Importantly, this implies that all covariates included

*Albert J. Weatherhead III University Professor, Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge MA 02138; GaryKing.org, king@harvard.edu, (617) 500-7570.

†Ph.D. Candidate, Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge MA 02138; christopherlucas.org, clucas@fas.harvard.edu, (617) 982-2718.

‡Assistant Professor, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge MA 02139; www.mit.edu/∼rnielsen, rnielsen@mit.edu, (857) 998-8039.

in Figure 1 are untreated because the estimand in this example is the *sample average treatment effect on the treated*, for which treated units are not pruned.

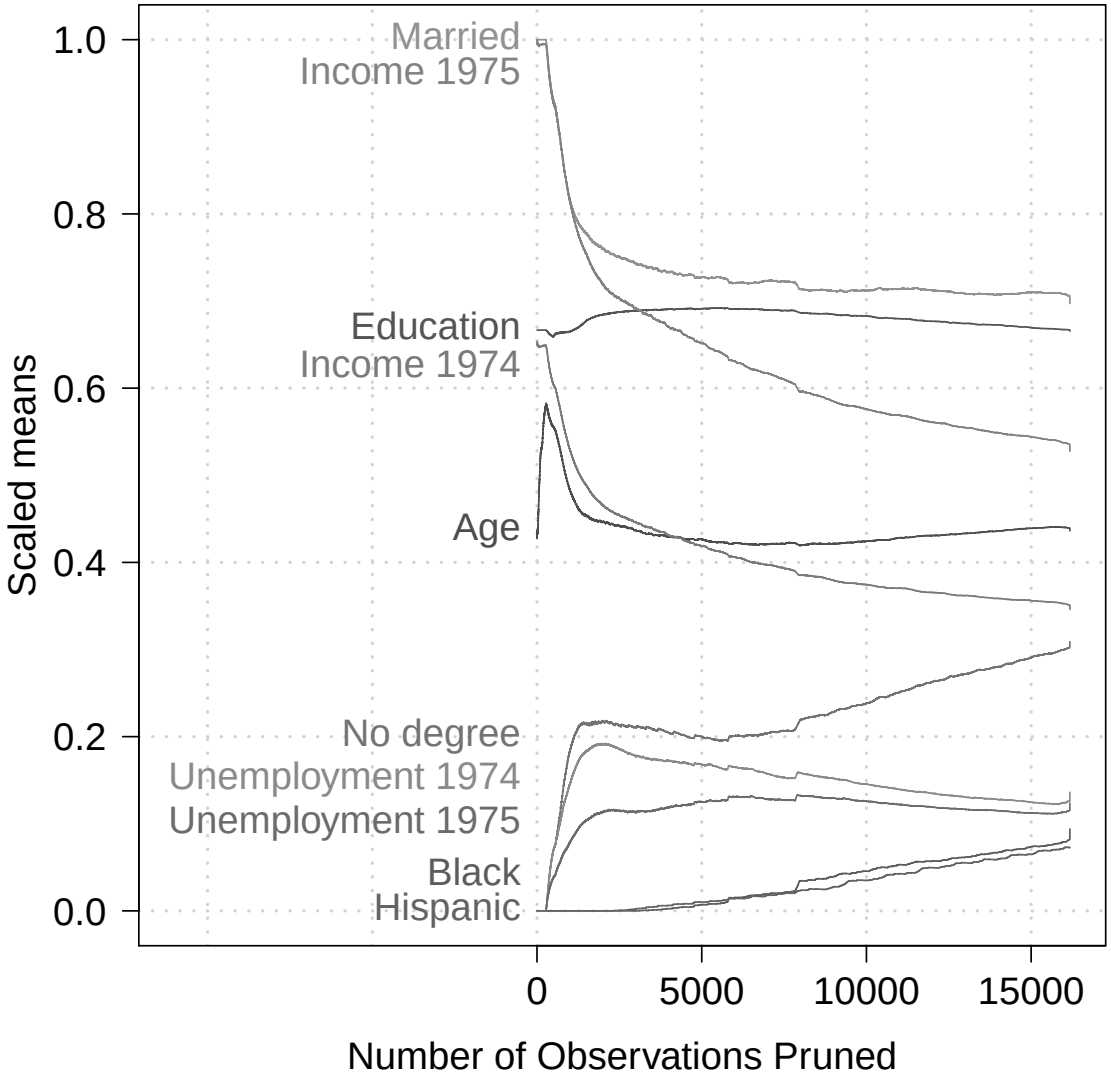

Figure 1: Covariate means for *pruned* observations on the Lalonde (1986) frontier.

Figure 1 reveals that early in the frontier, the algorithm prunes employed participants with relatively high incomes. This is clear from the means for variables `re74`, `re75`, `u74`, and `u75`, which are incomes in '74 and '75, along with indicators for unemployment. The intuition that eligibility for the original experimental study was contingent on income and employment status. Naturally, after supplementing the original experiment with observational data, the worst controls are those that would not have been eligible for the experiment — namely, those with very high incomes. This is reflected in the sharp

downward slope at the beginning of the frontier of means for pruned observations. The first hundred or so observations pruned are individuals with high income. However, as increasingly many observations are pruned, fewer bad matches remain and the means for pruned observations begin to level off.

A.2 Sex and Judging

We complement the analysis in Section 6.2 with Figure 2, which displays covariate means for pruned units.

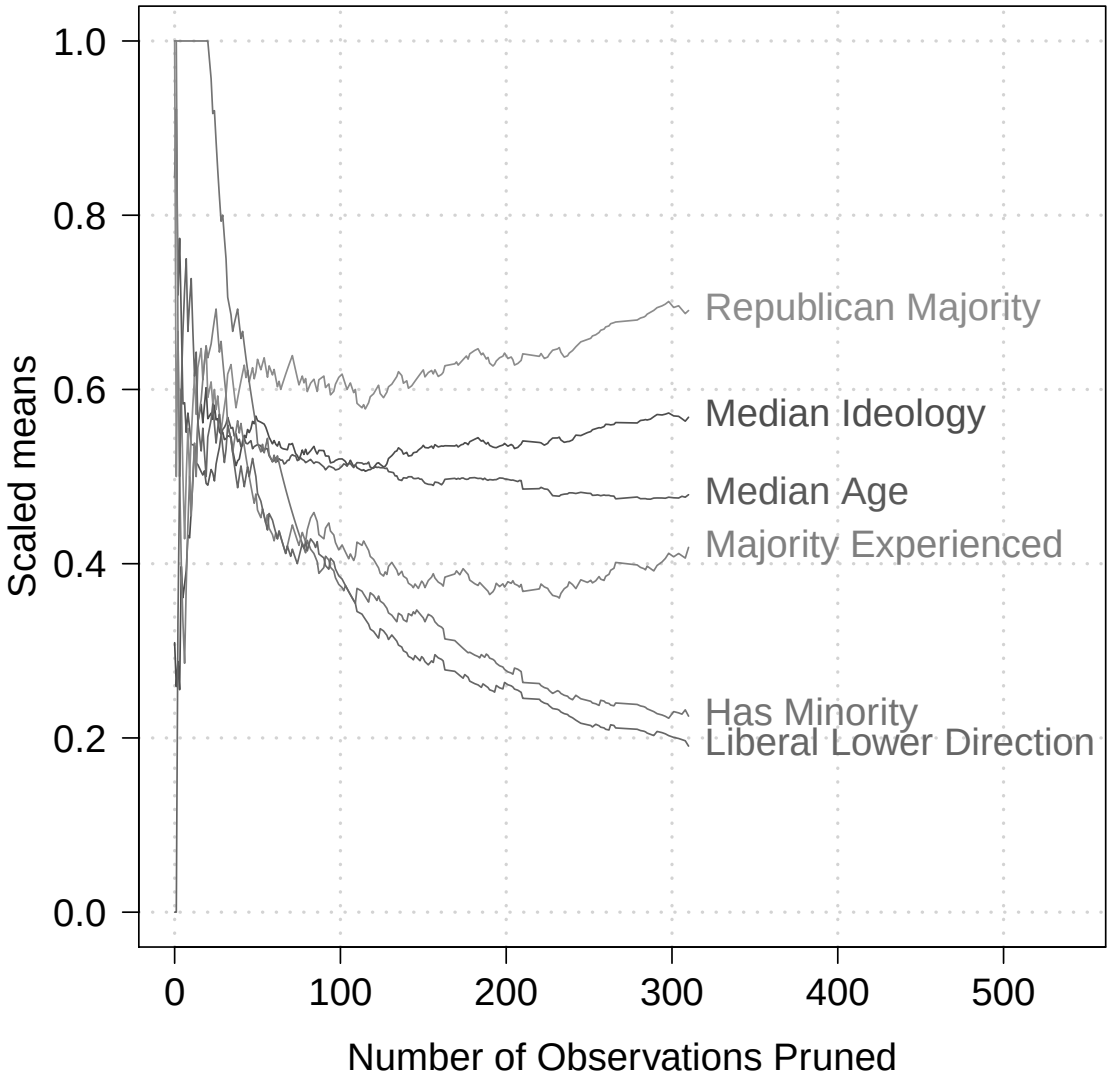

Figure 2: Covariate means for *pruned* observations on the Boyd, Epstein and Martin (2010) frontier.

As in the Lalonde (1986) example, Figure 2 provides intuition for the dimensions

most responsible for imbalance between the treated and control groups. In the Boyd, Epstein and Martin (2010) example, that dimension is the presence of another minority on the bench. The intuition here is that the appointment of justices to appellate courts is a political process; women are more likely to be appointed in districts where other minorities were also appointed, which may reflect a general commitment to representation on the bench.

B Simulation

In this section, we conduct a simple simulation to provide intuition about imbalance and sample size in a context where the true data generating process is exactly known.

B.1 Data Generating Process

Following King and Nielsen (2015), we conduct a simulation in which a completely randomized experiment is hidden within an imbalanced data set. Specifically, we construct an example with 10,000 observations, 5,000 of which are treated. Two covariates — X1 and X2 — are observed for all observations. For control units, both X1 and X2 are distributed Uniform(0,5), whereas they are both distributed Uniform(1,6) for treated observations. We generate the outcome Y as $Y_i = 2T_i + X1_i + X2_i + \epsilon_i$, where ϵ is $N(0, 1)$. Note that the true treatment effect is two.

B.2 Frontier and Estimates

Given these data, we calculate the AMI FSATT frontier, displayed in the left panel of Figure 3. The right panel of Figure 3 displays the treatment effects across this frontier, along with model dependence. In this example, we define model dependence as the range of point estimates for the treatment effect estimated by different model specifications. Thus, the red region represents the range of possible point estimates for the treatment effect given some point on the frontier. Points with wider ranges are those at which the researcher must know the true model in order to correctly estimate the treatment effect. In contrast, points with narrow ranges are those at which the researcher would correctly estimate the treatment effect, no matter the specified functional form.

It is possible to estimate the true effect (the number 2) at any point on the frontier where a reasonable number of data points remain. However, when imbalance is largest

(at the start of the frontier), it is possible to estimate an effect more than double the true effect. By contrast, when imbalance is low, a researcher would estimate the true effect no matter the specified form.

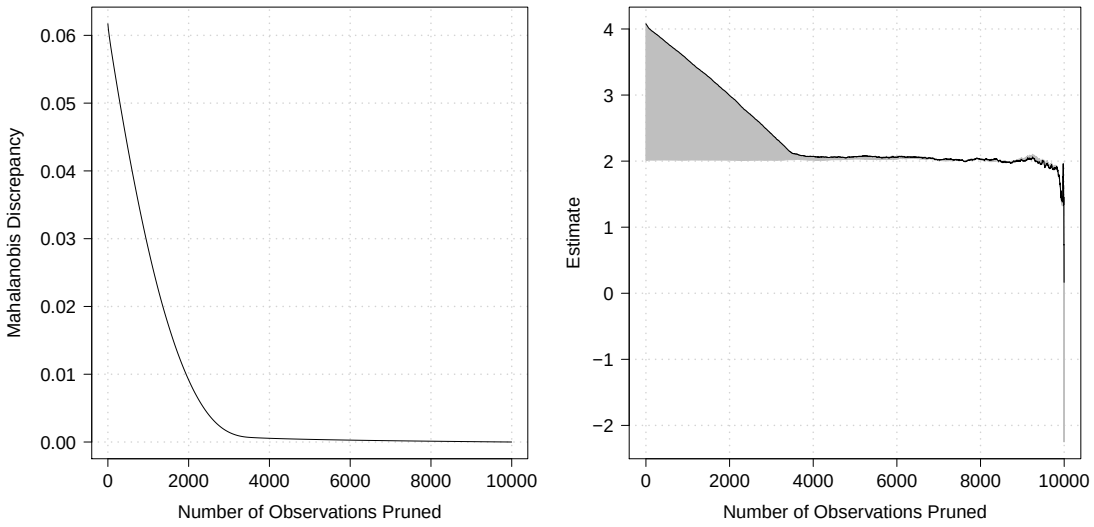

Figure 3: The left panel displays the imbalance frontier for the simulated data, while the right panel displays estimated treatment effects across it.

B.3 Covariate Means and Intuition

The two covariates on which we matched — X1 and X2 — are distributed Uniform(0,5) for control units and Uniform(1,6) for treated units. This implies a square region of common support, to which the matching algorithm subsets the data. To see this, represent the data in the two dimensional space of X1 and X2, as shown in Figure 4. Let (X1, X2) denote a coordinate within the space. The distribution of the data imply support over a two-dimensional square with corners at (1,1) and (5,5), which is observable in the figure.

As in previous analyses, we calculate covariate means for both pruned and remaining observations. However, in this illustration, we also calculate the *difference in means* between treated and control groups. Because this data set is relatively simple, we do not scale the covariate means (in real data scaling in ways that are substantively meaningful is important). The top left panel in Figure 5 displays the means for both covariates in the matched samples (i.e., the observations remaining), pooling across treated and control groups.

Note that pruning bad matches has relatively little influence on the means of these co-

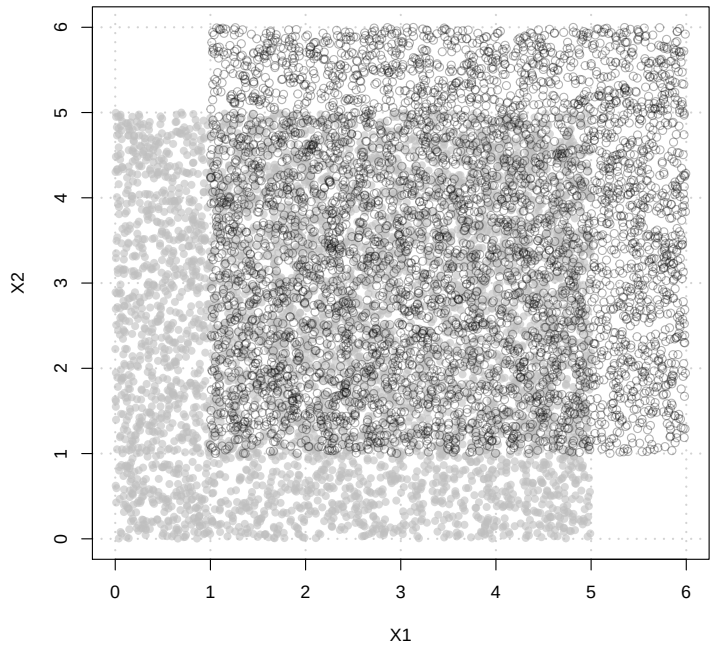

Figure 4: Distribution of the simulated data over covariates.

variates. The intuition for this lies in the *FSATT* estimand. Specifically, pooling treatment conditions, $E[X_1]$ and $E[X_2]$ both equal three, which is also the centerpoint of the square ranging from (1,1) to (5,5) over which we have common support. That is, the centerpoint (3,3) in this supported square is the centerpoint of the raw data. Because data are symmetrically ditributed around this point, pruning observations according to continuous distance preserves these expectations.

To further illustrate this point, the top right panel of Figure 5 displays the *difference in the means for treated and control* in the remaining matched sample. Note that the difference in the means of X_1 and X_2 before pruning is roughly equal to 1, which equals E[Uniform(1,6)] - E[Uniform(0,5)]. Because imbalance in both X1 and X2 is equal in expectation, imbalance decreases at roughly equal rates in these dimensions until approaching zero. As in the previous examples displayed in this appendix, we plot the means of pruned observations in the lower panel of Figure 5. As implied by the discussion of covariate means in the matched sample, the means of pruned observations do not change across the frontier.

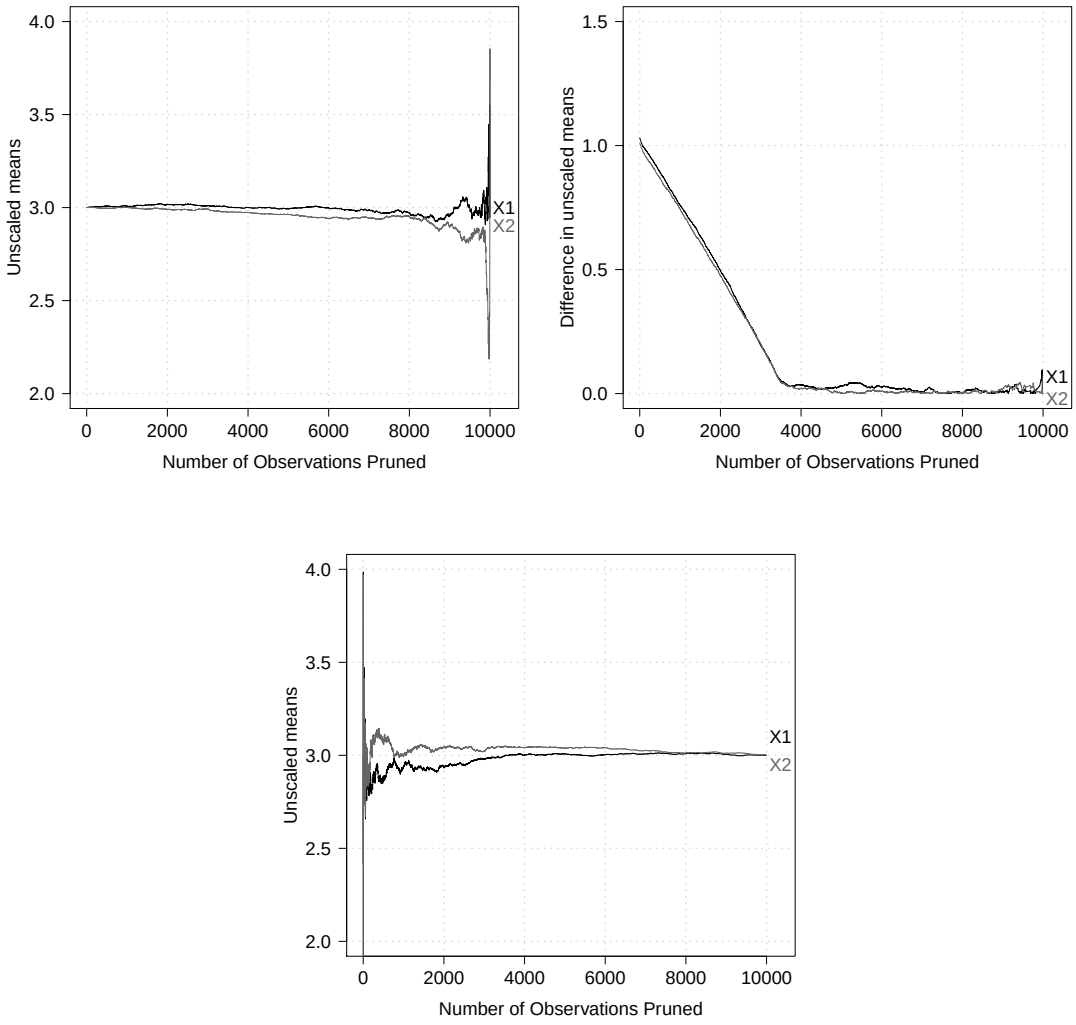

Figure 5: The top left panel displays covariate means across the frontier, the top right displays the difference in means for treated and control units, and the bottom displays the covariate means for pruned observations.

References

Boyd, Christina L, Lee Epstein and Andrew D Martin. 2010. "Untangling the causal effects of sex on judging." *American journal of political science* 54(2):389–411.

King, Gary, Christopher Lucas and Richard Nielsen. 2015. *MatchingFrontier: Automated Matching for Causal Inference*. R package version 2.0.0.

King, Gary and Richard Nielsen. 2015. "Why Propensity Scores Should Not Be Used for Matching." *Working* .

Lalonde, Robert. 1986. "Evaluating the Econometric Evaluations of Training Programs." *American Economic Review* 76:604–620.