LOOP with External Data Sets: Simulation

Ed Wu* Johann Gagnon-Bartsch*

June 26, 2018

1. Simulations

In this section, we examine the performance of LOOP-E using simulation. In particular, we consider how LOOP-E performs when varying sample size, the predictive power of the covariates, and the predictive power of the external predictions. Consider a randomized experiment in which there are N subjects. The potential outcomes and covariates are generated from the following linear model:

$$\begin{aligned} a_i &= 2Z_{1,i} + Z_{2,i} + \delta_i \\ c_i &= \frac{a_i}{\sigma_a} \\ t_i &= c_i + 3 \end{aligned}$$

where $\delta_i \sim \mathrm{N}(0, \sigma^2_{gen})$, $Z_{ij} \sim \mathrm{Unif}(0, 10)$, and $\sigma^2_a = \mathrm{Var}(a_i) = \frac{500}{12} + \sigma^2_{gen}$. By generating our potential outcomes as above, we have defined our generative model such that the control potential outcomes have unit variance. We can alternatively write the observed outcome as:

$$Y_i = 3T_i + \frac{2}{\sigma_a} Z_{1,i} + \frac{1}{\sigma_a} Z_{2,i} + \epsilon_i$$

where $\epsilon_i \sim \mathrm{N}(0, \sigma^2_{gen}/\sigma^2_a)$.

For each observation, we also simulate external predictions $\tilde{t}_i$ and $\tilde{c}_i$ for t_i and c_i by taking the true t_i or c_i and adding a normally distributed noise term with mean 0 and standard deviation σ_{ext}.

To reiterate, we wish to consider variations in sample size, the predictive power of the covariates, and the predictive power of the external predictions. Sample size is directly indexed with N. We can index the predictive power of the covariates using the control-side R^2:

$$R^2_c = 1 - \frac{\sigma^2_{gen}}{\sigma^2_a}.$$

*Department of Statistics, University of Michigan, Ann Arbor, MI.

Similarly, the predictive power of our external prediction $\tilde{c}_i$ is

$$R_p^2 = 1 - \frac{\sigma_{ext}^2}{\text{Var}(\tilde{c}_i)} = 1 - \frac{\sigma_{ext}^2}{1 + \sigma_{ext}^2}.$$

Given a desired R_c^2 and R_p^2, we can calculate the corresponding values of σ_{gen}^2 and σ_{ext}^2 (note that σ_{ext}^2 and σ_{gen}^2 characterize the distribution of the external predictions and potential outcomes for a given set of covariates):

$$\sigma_{gen}^2 = \frac{1 - R_c^2}{R_c^2} \times \frac{500}{12}$$
$$\sigma_{ext}^2 = \frac{1 - R_p^2}{R_p^2}.$$

We perform three sets of simulations, in which we hold two of N, R_c^2, and R_p^2 constant and vary the third. For each set of simulations, we compare the following methods:

1. LOOP: uses the LOOP estimator including only the covariates Z_1 and Z_2
2. LOOP with External Predictions: uses the LOOP estimator with external predictions as a covariate (in addition to Z_1 and Z_2)
3. LOOP OLS: uses the LOOP estimator, with OLS as the imputation method. Only includes the external predictions as a covariate
4. LOOP-E: interpolates between the previous two methods

We use the following simulation procedure. For a given set of N, R_c^2, and R_p^2, we perform $k = 1000$ trials. For each trial, we generate a set of potential outcomes, a treatment assignment vector, and external predictions for t_i and c_i. We then produce an estimate of the variance of each method for that treatment assignment vector. Next, we average the estimated variance across the k trials. Finally, we plot the average nominal variance for each method relative to the variance of the simple difference estimator.

1.1. Varying Sample Size

For this simulation, we hold the predictive power of the covariates and external predictions constant and vary the sample size $N = 30, 40, 50, 75, 100, 150, 200$. We consider four scenarios: (1) $R_p^2 = 0.25, R_c^2 = 0.25$; (2) $R_p^2 = 0.75, R_c^2 = 0.25$; (3) $R_p^2 = 0.25, R_c^2 = 0.75$; and (4) $R_p^2 = 0.75, R_c^2 = 0.75$:

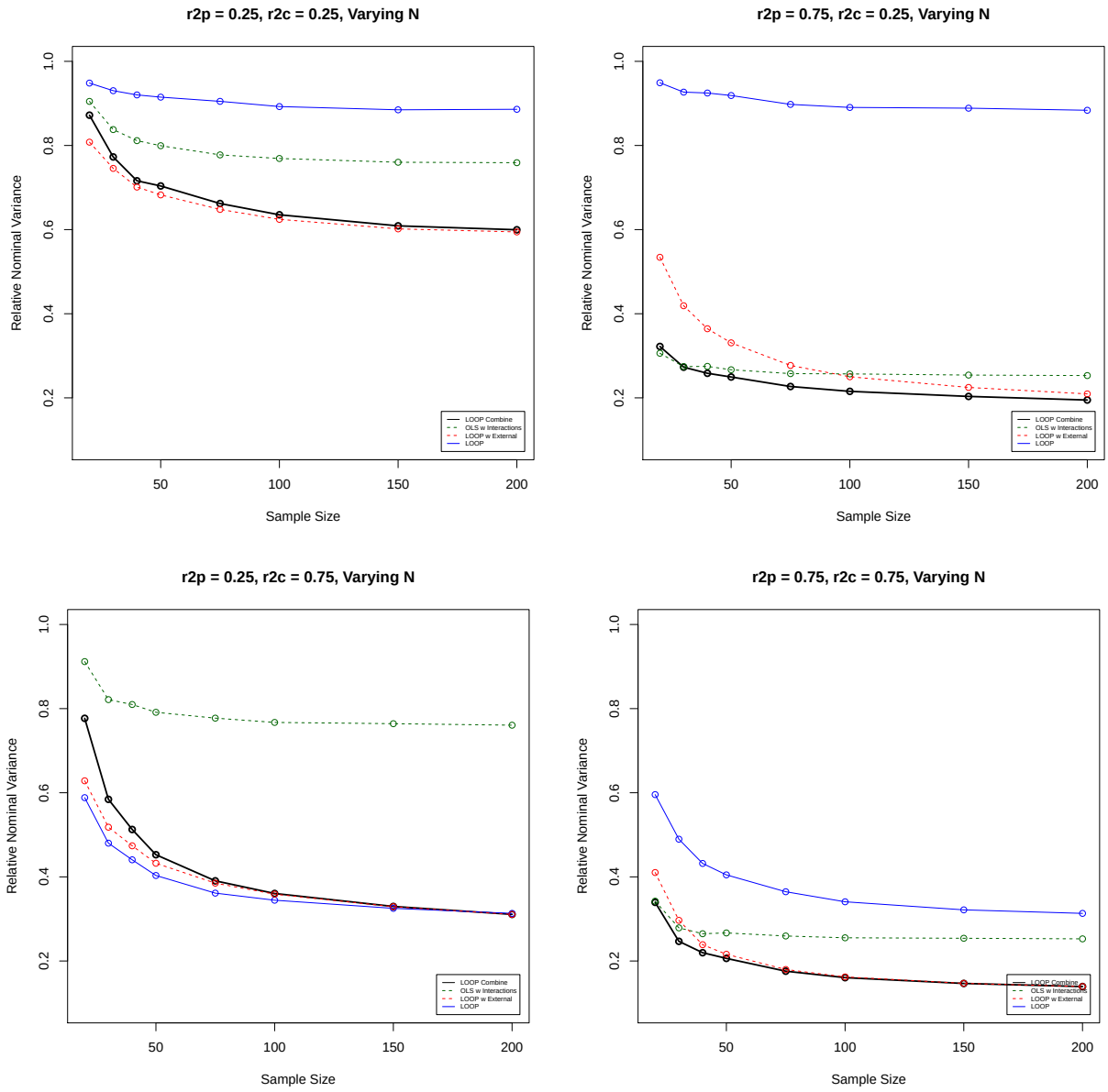

Figure 1: Top Left: $R_p^2 = 0.25, R_c^2 = 0.25$; Top Right: $R_p^2 = 0.75, R_c^2 = 0.25$; Bottom Left: $R_p^2 = 0.25, R_c^2 = 0.75$; Bottom Right: $R_p^2 = 0.75, R_c^2 = 0.75$

As we can see, when the predictive power is low for both the covariates and the external predictions ($R_p^2 = 0.25, R_c^2 = 0.25$), LOOP is outperformed by LOOP-E. Similarly, when the predictive power is high for both, LOOP-E outperforms LOOP. Even when R_p^2 is low and R_c^2 is high, the performance of LOOP-E quickly converges to the performance of LOOP. Finally we observe that LOOP-E does well at tracking the better performing component (and generally outperforms both components, when the components perform similarly to each other).

1.2. Varying Predictive Power of External Prediction

For this simulation, we hold the predictive power of the covariates and sample size constant and vary $R_p^2 = 0.05, 0.15, ..., 0.85, 0.95$. We consider four scenarios: (1) $N = 30, R_c^2 = 0.25$; (2) $N = 30, R_c^2 = 0.75$; (3) $N = 60, R_c^2 = 0.25$; and (4) $N = 60, R_c^2 = 0.75$:

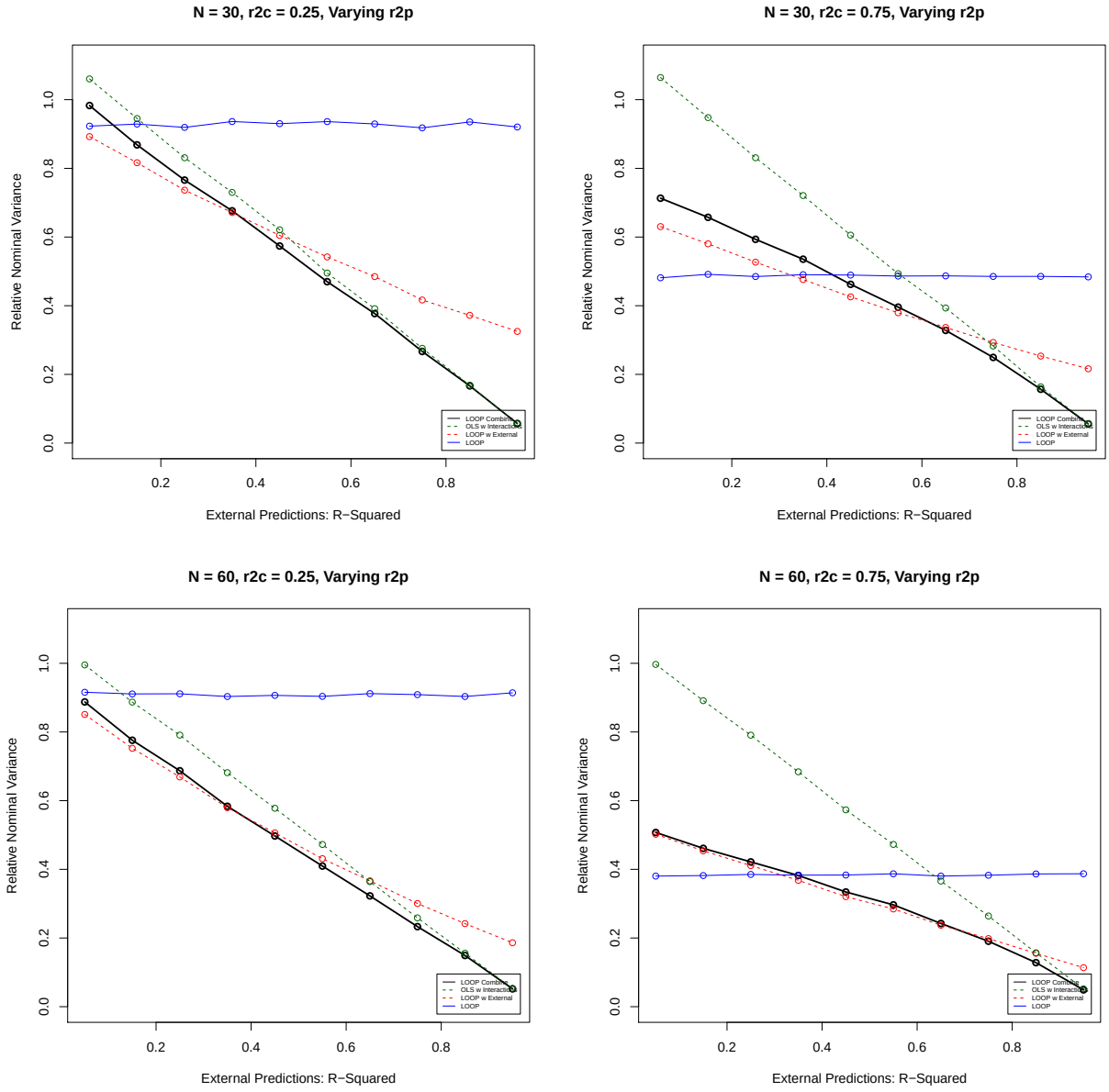

Figure 2: Top Left: $N = 30, R_c^2 = 0.25$; Top Right: $N = 30, R_c^2 = 0.75$; Bottom Left: $N = 60, R_c^2 = 0.25$; Bottom Right: $N = 60, R_c^2 = 0.75$

Once again, we observe that LOOP-E tends to perform at least as well as either component. This is particularly true for $N = 60$, where LOOP-E closely follows (or drops below) the lower of the component lines. As expected, the three methods that incorporate the exter-

nal predictions all improve as R_p^2 increases, while the performance of LOOP stays constant. We can see that LOOP-E is outperformed by LOOP when only when R_p^2 is much lower than R_c^2.

1.3. Varying Predictive Power of Covariates

For this simulation, we hold the predictive power of the external predictions and sample size constant and vary $R_c^2 = 0.05, 0.15, ..., 0.85, 0.95$. We consider four scenarios: (1) $N = 30, R_p^2 = 0.25$; (2) $N = 30, R_p^2 = 0.75$; (3) $N = 60, R_p^2 = 0.25$; and (4) $N = 60, R_p^2 = 0.75$:

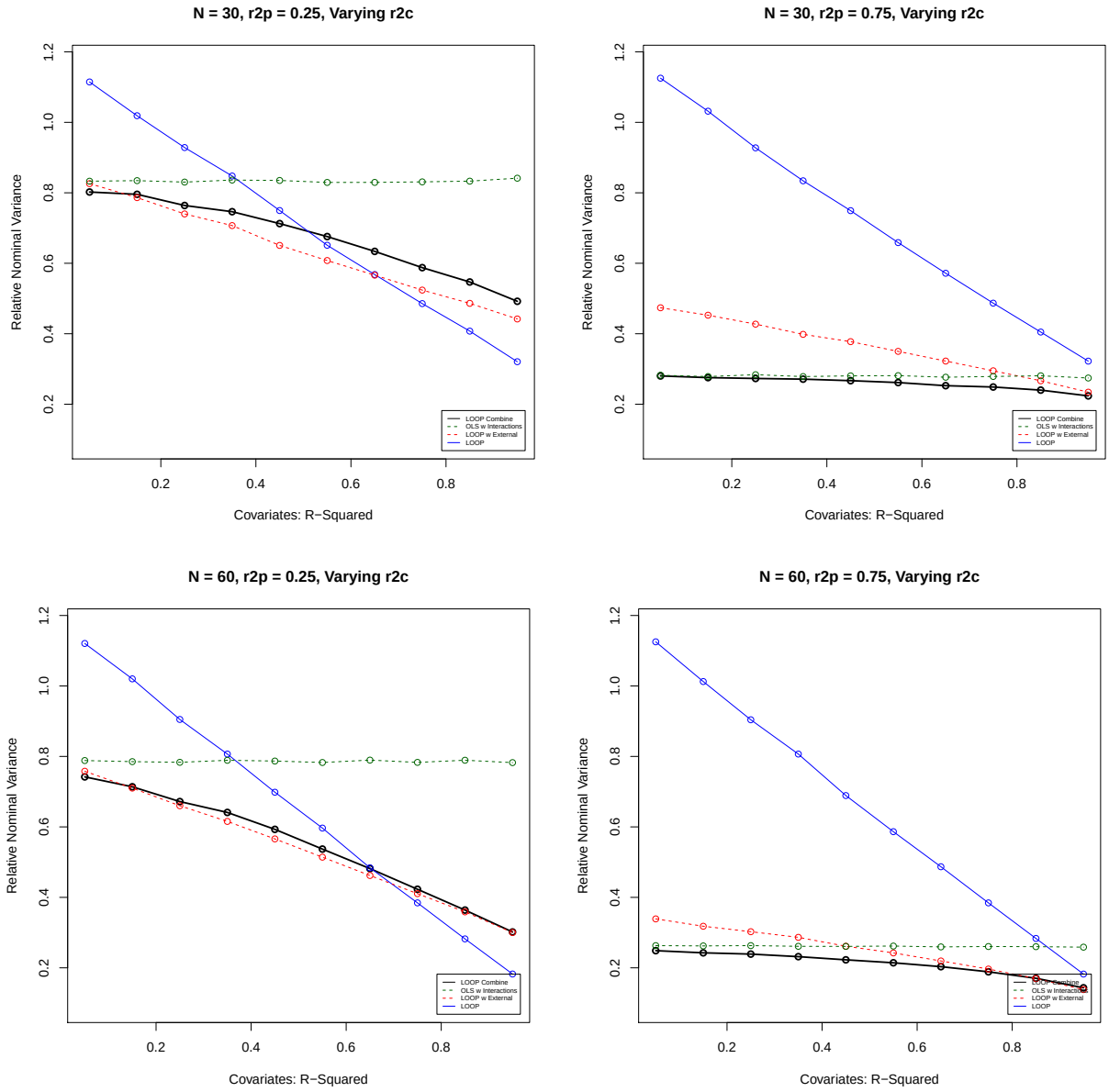

Figure 3: Top Left: $N = 30, R_p^2 = 0.25$; Top Right $N = 30, R_p^2 = 0.75$; Bottom Left $N = 60, R_p^2 = 0.25$; Bottom Right: $N = 60, R_p^2 = 0.75$

The performance of LOOP OLS stays constant, as the imputation method only incorporates the external predictions. The remaining methods all improve as R_c^2 increases. As before, we can see that LOOP-E tracks the better performing component well (especially when $N = 60$) and is only outperformed by LOOP when R_c^2 is much higher than R_p^2.