Research Article

Joseph Antonelli* and Matthew Cefalu

Averaging causal estimators in high dimensions

https://doi.org/10.1515/jci-2019-0017
Received Jun 21, 2019; accepted Jul 06, 2020

Abstract: There has been increasing interest in recent years in the development of approaches to estimate causal effects when the number of potential confounders is prohibitively large. This growth in interest has led to a number of potential estimators one could use in this setting. Each of these estimators has different operating characteristics, and it is unlikely that one estimator will outperform all others across all possible scenarios. Coupling this with the fact that an analyst can never know which approach is best for their particular data, we propose a synthetic estimator that averages over a set of candidate estimators. Averaging is widely used in statistics for problems such as prediction, where there are many possible models, and averaging can improve performance and increase robustness to using incorrect models. We show that these ideas carry over into the estimation of causal effects in high-dimensional scenarios. We show theoretically that averaging provides robustness against choosing a bad model, and show empirically via simulation that the averaging estimator performs quite well, and in most cases nearly as well as the best among all possible candidate estimators. Finally, we illustrate these ideas in an environmental wide association study and see that averaging provides the largest benefit in the more difficult scenarios that have large numbers of confounders.

Keywords: High-dimensional statistics, Causal inference, Averaging estimators

2010 Mathematics Subject Classification: 62D20, 62P12, 62F99

1 Introduction

Estimating the causal effect of a treatment on an outcome with observational data relies on the assumption that the treatment assignment is unconfounded conditional on a set of pre-treatment covariates. There is a large literature on methods to adjust for confounding in order to satisfy this assumption, such as matching [1], propensity score weighting [2], stratification [3], outcome regression, and doubly robust estimators [4]. These approaches have all been shown to provide valid estimates of causal effects as long as the covariates necessary for the unconfoundedness assumption to hold are measured in the data. In recent years, the estimation of causal effects in high-dimensional data sets with large numbers of covariates have become increasingly common. A large number of measured covariates makes it more plausible that the covariates necessary for confounding adjustment have been collected; however, estimation becomes considerably more difficult with large numbers of parameters. Due to this trade off, methods have been developed to adjust for large numbers of covariates and produce consistent estimates of causal effects from observational data.

Much of the interest began as papers started to consider the problem of variable selection or model selection for causal estimation in low-dimensional settings [5–7]. Since then, a large number of manuscripts have been published that aim to estimate causal effects when the number of covariates is large relative to the sample size. Wilson and Reich [8] tailored the adaptive lasso [9] to effect estimation instead of prediction, by

***Corresponding Author: Joseph Antonelli:** University of Florida, Department of Statistics, United States of America; Email: jantonelli@ufl.edu
Matthew Cefalu: RAND Corporation; Email: mcefalu@rand.org

creating weights that involve both outcome model and treatment model coefficients. Other approaches have utilized penalization approaches to estimate causal effects in high-dimensions. Shortreed and Ertefaie [10] used the adaptive lasso to estimate the propensity score, but allowed the weights to depend on the outcome model coefficients. Ertefaie *et al.* [11] created a new penalized likelihood estimator that utilizes information from both the treatment and outcome to identify confounders, and then estimate causal effects conditional on this set of variables. The lasso [12] was used to estimate both propensity score and prognostic score models in [13] to obtain doubly robust estimates by matching on both estimated scores. Antonelli *et al.* [14] utilized a fully Bayesian approach to estimating treatment effects in high-dimensions that reduces finite sample bias while eliminating the impact of instrumental variables through informative prior distributions on variable inclusion parameters. Hahn *et al.* [15] utilized horseshoe priors on the coefficients from a re-parameterized likelihood to reduce shrinkage of variables that are associated with both the treatment and outcome. A large number of approaches have focused on obtaining uniformly valid inference of causal effects in high-dimensions [16–18]. Belloni *et al.* [16] used lasso models to perform variable selection on both the treament and outcome models, taking the union of the identified covariates from the two models, and then estimating effects conditional on the chosen model. Farrell [17] also utilized the lasso for a treatment and outcome model selection, and then fit post selection estimators to estimate the traditional doubly robust estimator. Chernozhukov *et al.* [18] estimated nuisance parameters such as treatment or outcome model parameters via penalization, and then solved the efficient score to estimate treatment effects. Similar ideas are used from the Bayesian perspective in [19], and it is shown that it leads to improved finite sample performance of frequentist metrics such as interval coverage. Other approaches have shown that $\sqrt{n}$ convergence of treatment effects can be achieved in high-dimensional settings. Athey *et al.* [20] combined high-dimensional outcome regression with balancing weights that eliminate any residual confounding remaining after the outcome regression. Propensity score estimation is improved for both high-dimensions and model misspecification using calibration weights in [21, 22]. Propensity scores were again used in high-dimensional scenarios, but are instead estimated directly by their ability to balance important covariates in [23]. Lastly, targeted maximum likelihood (TMLE, [24]) can be used in conjuction with high-dimensional models to produce estimates of treatment effects. Recent work has tailored TMLE to this scenario using collaborative TMLE [25, 26] or variance moderation [27].

These developments have led to a diverse and large set of estimators to use if one is faced with observational data that is high-dimensional. Choosing among this large set of estimators is a subjective choice, and it is not clear if, or when, any estimators are preferable to others. Given the complexity faced with high-dimensional data, it is highly unlikely that any one estimator will perform well universally.

In this manuscript, we leverage a large set of potential estimators to our advantage by introducing a simple, intuitive estimator that averages over all of the candidate estimators. We show both theoretically and empirically that this leads to robust performance for estimating treatment effects in high-dimensions. Through extensive simulation studies we show that averaging over all estimators generally leads to improved performance with respect to standard metrics such as bias, mean squared error (MSE), and coverage probabilities. We illustrate through a wide variety of data generating mechanisms that there is not any one individual estimator that performs well across all scenarios; however, the proposed averaged estimator performs nearly as well as the best estimator in the majority of cases considered, which shows the robustness that averaging provides for high-dimensional treatment effect estimation.

The idea of combining models, estimators, or tests has a rich history in the statistical literature. For example, model uncertainty can be accounted for through model averaging, which exists in both the Bayesian and frequentist framework [28–31]. Others have proposed combinations of estimators [32, 33], including in the field of small area estimation [34]. Omnibus tests for the global null hypothesis simultaneously test multiple hypotheses (e.g. ANOVA), and several methods for combining individual p-values from genome-wide association studies have been proposed [35]. These are just a few examples of combining models, estimators, or tests, but they all share a common goal: combining statistics for improved performance.

2 Methodology

Throughout, we assume that we have n i.i.d samples of $(\boldsymbol{X}_i, T_i, Y_i)$, where $\boldsymbol{X}_i$ is a p–dimensional vector of pre-treatment covariates, T_i is the treatment of interest, and Y_i is the outcome of interest that is measured after the treatment is initiated. We assume that T_i is a binary treatment as many of the existing estimators are only applicable to binary treatments. We do not discuss the rates at which p grows, but work in the high-dimensional regime where $p > n$ and grows with n. For simplicity, we restrict attention to the average treatment effect defined as ATE $= E(Y(1) - Y(0))$, where $Y_i(t)$ is the potential outcome we would have observed for subject i if they had been exposed to treatment $T_i = t$. We assume no unmeasured confounding and positivity, defined below:

Unconfoundedness: $Y(t) \perp\!\!\!\perp T|\boldsymbol{X}$ for t=0,1
Positivity: There exist $\delta \in (0, 1)$ such that $0 < \delta < P(T = 1|\boldsymbol{X}) < 1 - \delta < 1$ with probability 1,

where $P(T = 1|\boldsymbol{X})$ denotes the propensity score [2]. We also assume the stable unit treatment value assumption (SUTVA), which states that the treatment is well-defined and that the outcomes for one subject are not affected by the treatment status of other subjects. Further, in high-dimensional settings, many estimators rely on some form of sparsity, stating that unconfoundedness holds conditional of a subset of the covariates. The exact form of the sparsity condition varies by estimator and is not discussed in detail in this manuscript.

2.1 Averaged estimator

Consider constructing an averaged estimator from a set of M estimators for the ATE. Denote each estimator of the ATE by $\widehat{\theta}_m$ for $m = 1, \ldots, M$. Each estimator has a corresponding variance estimate, denoted by $\widehat{\sigma}_m^2$ for $m = 1, \ldots, M$. The averaged estimator takes the form:

$$\widehat{\theta}_A = \sum_{m=1}^{M} w_m \widehat{\theta}_m,$$

where $\boldsymbol{w} = (w_1, \ldots, w_M)$ is a weight vector. Ideally, we would estimate $\boldsymbol{w}$ by optimizing a pre-specified loss function. Following [33], one could find the weight vector that minimizes the mean squared error of $\widehat{\theta}_A$ conditional on $\boldsymbol{w}$, which corresponds to solving:

$$\underset{\boldsymbol{w}}{\operatorname{argmin}}\ \boldsymbol{w}^T \boldsymbol{V} \boldsymbol{w} + (\boldsymbol{w}^T \boldsymbol{B})^2,$$

where $\boldsymbol{B}$ and $\boldsymbol{V}$ are the bias vector and covariance matrix for the M estimators. Finding the weight vector that minimizes the mean squared error requires estimates of both $\boldsymbol{B}$ and $\boldsymbol{V}$. In the context of noncompliance in clinical trials, Antonelli *et al.* [33] were able to estimate these quantities. The bias was estimated by identifying an unbiased estimator among the candidate estimators, and the covariance was estimated using the bootstrap. In the case of observational studies, the bias of each estimator is unknown a priori and there do not exist methods for estimating it from observed data.

The covariance matrix $\boldsymbol{V}$ of the estimators is also problematic when there are high-dimensional confounders. The standard bootstrap is generally not valid for estimating sampling distributions of high-dimensional models such as the LASSO [36]. In the context of high-dimensional ATE estimation, it has been seen empirically that the bootstrap can lead to poor inferential properties such as incorrect standard error estimation [19]. Due to this, much of the existing work on estimating the ATE in high dimensions has focused on deriving asymptotic limiting distributions for each estimator. While these asymptotic results may provide variance estimates for individual estimators, estimation of $\boldsymbol{V}$ also requires the pairwise covariance between each estimator. It is extremely cumbersome, and perhaps not possible, to derive the covariance for each pair of estimators, and would require additional work whenever new candidate estimators are included. Due to the difficulty of estimating $\boldsymbol{B}$ and $\boldsymbol{V}$, we will proceed forward with a fixed weight vector $\boldsymbol{w} = (1/M, 1/M, \ldots, 1/M)$, such that each estimator gets equal weight. This provides a number of advantages over simply choosing a single estimator, which are detailed in the subsequent sections. The averaging provides a form of robustness ensuring that the averaged estimator does better in terms of bias and MSE than the worst of the M estimators. The empirical studies in Section 3 show that averaging not only provides protection against choosing the worst estimator, but the averaged estimator typically does close to, if not better, than the best of the M estimators. This is due to large decreases in the variance of the averaged estimator relative to any one estimator, particularly if the individual estimators are not strongly correlated. Note that unequal weights can easily be applied as long as they are chosen a priori and not chosen simply to provide a desired result. Further, if there is concern about one bad estimator, in particular one with large amounts of bias, then a trimmed average can be used. We explore this further in the simulation study of section 3 where we use a trimmed average estimator that removes the highest and lowest estimates from the averaged estimator.

2.2 Variance estimation

If the covariance matrix $\boldsymbol{V}$ of the M estimators is known, the variance of the averaged estimator can be written as:

$$\mathrm{var}(\widehat{\theta}_A) = \frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M} V_{ij},$$

where V_{ij} is the (i, j) element of the matrix $\boldsymbol{V}$. Exploring this quantity provides insight in how the averaged estimator can greatly reduce variance. In one extreme case where all the estimators are perfectly correlated, then the variance of the averaged estimator is equal to the average of the pairwise products of the standard deviations for each estimator, implying that its variance is no larger than the highest variance estimator. In a different extreme case where all of the estimators are uncorrelated, the variance of the averaged estimator is $\frac{1}{M^2}\sum_{i=1}^{M} V_{ii}$. This implies that the variance of the averaged estimator is an order of magnitude $(1/M)$ smaller than the average of the individual estimator variances. Even bigger gains are possible if the estimators are negatively correlated, but this is very unlikely to happen in practice. In our experience, most estimators are positively, but not perfectly, correlated with each other, meaning that the truth lies between the two extremes discussed. Nonetheless, the variance of the averaged estimator will be no worse than the highest variance estimator, and may be a substantial improvement over the individual estimators.

Estimating the variance of the averaged estimator is problematic due to the underlying high-dimensional estimators. A plug-in estimate of the variance based on the expression above is not feasible, because it is very difficult to produce the necessary covariances in high-dimensions. We can, however, estimate the individual estimator variances $\widehat{\sigma}_m^2$. Since the correlation between estimators is bounded between -1 and 1, we can bound the variance of the averaged estimator as:

$$\mathrm{var}(\widehat{\theta}_A) = \frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M} V_{ij} = \frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M} \mathrm{cor}(\widehat{\theta}_i, \widehat{\theta}_j)\sqrt{V_{ii}}\sqrt{V_{jj}} \leq \frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M} \sqrt{V_{ii}}\sqrt{V_{jj}}.$$

Therefore, we propose a conservative estimate of the variance as:

$$\widehat{\sigma}_A^2 = \frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M} \widehat{\sigma}_i\widehat{\sigma}_j.$$

If the estimators are highly, positively correlated, this variance estimate will not be very conservative. If the estimators are close to independent, then this variance estimator is very conservative. It is possible that being conservative in this setting is beneficial for two reasons. First, the estimates of standard errors $\widehat{\sigma}_m$ frequently rely on asymptotic approximations, which might not hold, and could be underestimates of the true variances.

Second, there is frequently bias in estimating treatment effects in high-dimensions due to penalization or variable selection, and conservative estimates of the variance may lead to better interval coverage. Section 3 provides empirical evidence of the performance of this variance estimation strategy.

2.3 Robustness to choosing the worst model

The relative performance of the averaged estimator to the candidate estimators is highly data dependent. However, we show that the averaged estimator protects against choosing a very bad estimator among the class of estimators. Intuitively, this is because the averaged estimator only assigns a weight of $1/M$ to any one estimator, including the worst estimator with respect to a given metric; therefore, the averaged estimator is guaranteed to perform no worse than this worst case estimator. The following result shows this for both bias and mean squared error:

Result 1: If each estimator is assigned equal weight, and we let $\widehat{\theta}^{*}$ and $\widehat{\theta}^{**}$ be the worst estimators in terms of bias and MSE, respectively, then

$$|E(\widehat{\theta}_A - \theta)| \leq |E(\widehat{\theta}^{*} - \theta)| \text{ and } E((\widehat{\theta}_A - \theta)^2) \leq E((\widehat{\theta}^{**} - \theta)^2)$$

A proof of this result can be found in the Appendix. While this result only guarantees that the averaged estimator does better than the worst estimator, empirically we have seen that these are typically very conservative bounds. In simulation studies in Section 3 we illustrate that the averaged estimator not only performs better than the worst estimators, but frequently performs as well as the best individual estimator in terms of MSE.

3 Simulation study

We empirically investigate the ability of the averaged estimator to estimate treatment effects in high-dimensional settings. We focus on 10 potential estimators that are either available in an R package, or straightforward to implement. An R package to implement these estimators along with the averaged estimator can be found at https://github.com/jantonelli111/AveragingCausalHD. Computation times for the estimators can also be found in Section B of the Appendix. The estimators that enter into the averaged estimator are listed below:

1. The double post selection approach (Double PS, [16])
2. The approximate residual debiasing approach (Debiasing, [20])
3. The doubly robust lasso approach (DR-Lasso, [17])
4. Doubly robust matching (DRME, [13])
5. High-dimensional confounding with spike-and-slab priors (HDC, [14])
6. Bayesian doubly robust estimator (DR-Bayes, [19])
7. Targeted maximum likelihood with lasso models (TMLE, [24])
8. Targeted maximum likelihood with an initial screening step followed by unpenalized linear models (TMLE-screen)
9. Double machine learning with lasso models (DML, [18])
10. Double machine learning with post selection estimators (DML-PS)

We do not go into the details of the respective approaches or how they are fit, as the aim of this work is to show the performance of the averaged estimator when there are a large number of estimators to average over, each with varying operating characteristics that are unknown to the user beforehand. For this reason we will only highlight the best, median, and worst estimator among the individual estimators with respect to any particular metric. We will also show results for two averaged estimators: The averaged estimator with equal weights for each estimator, and a trimmed average that removes the estimators with the highest and lowest estimated causal effects within a particular data set. For mean squared error, the best estimator is the one with the smallest MSE, the worst is the one with the largest MSE, and the median is the estimator with the median MSE among the individual estimators. For ease of discussion, the MSE for each simulation scenario is normalized so that the best estimator has MSE equal to 1. For interval coverage, the worst estimator is the one with the lowest coverage, and then the best estimator is the one with the coverage closest to 0.95 without being 1. We want to avoid an estimator with a coverage probability of 1 as it is overly conservative. The median estimator is the estimator with the median coverage probability among the individual estimators.

The exact details of the simulation scenarios can be found in Section C of the Appendix, though we will briefly summarize them here. In each simulation scenario we draw covariates from a multivariate normal distribution with correlation 0.3 between all covariates. We will set $n = 150$ and $p = 300$ for all simulation scenarios with the exception of scenario 2, which will have $n = 300$ subjects. The first simulation scenario represents a data generating mechanism that has a sparse, linear model for both the treatment and outcome. The second simulation scenario is one in which the treatment model is dense and all covariates are associated with the treatment. The third scenario looks at a setting where the outcome model is not sparse. The fourth setting is one in which the outcome model is nonlinear and therefore many of the approaches will be misspecified. The final simulation setting is one in which there is treatment effect heterogeneity, which also can cause misspecification of some of the approaches that assume homogeneous effects. These are chosen to capture a wide range of data generating mechanisms that might occur in practice. We simulated five additional scenarios, and the results can be found in Section C of the Appendix. The results in the additional simulation scenarios closely match those seen in the main manuscript.

3.1 Results

Figures 1 and 2 show the results averaged over 1000 simulations for both MSE and interval coverage. Both averaging estimators performed favorably in terms of MSE and interval coverage. The highest MSE estimator in each scenario had significantly higher MSE than the averaging estimators. The averaging estimators had lower MSE than the median MSE estimator in every scenario and were able to achieve MSE that is close to the optimal choice of the individual estimators. They had slightly lower MSE than the individual estimator with the lowest MSE in Scenario 5. The averaging estimators achieved confidence interval coverage near the nominal 0.95 level in all of the simulation scenarios, and significantly outperformed both the worst estimator and the median estimator. For the averaged estimator, the specific interval coverages in each simulation were 0.904, 0.975, 0.964, 0.987, and 0.931, respectively.

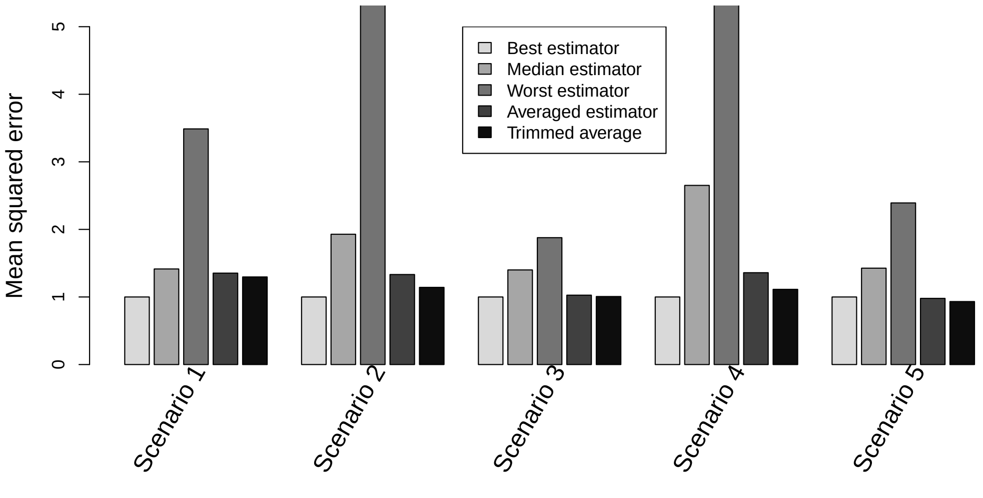

Figure 1: Mean squared error for the estimator with the lowest MSE, highest MSE, median MSE, the averaged estimator, and the trimmed average estimator.

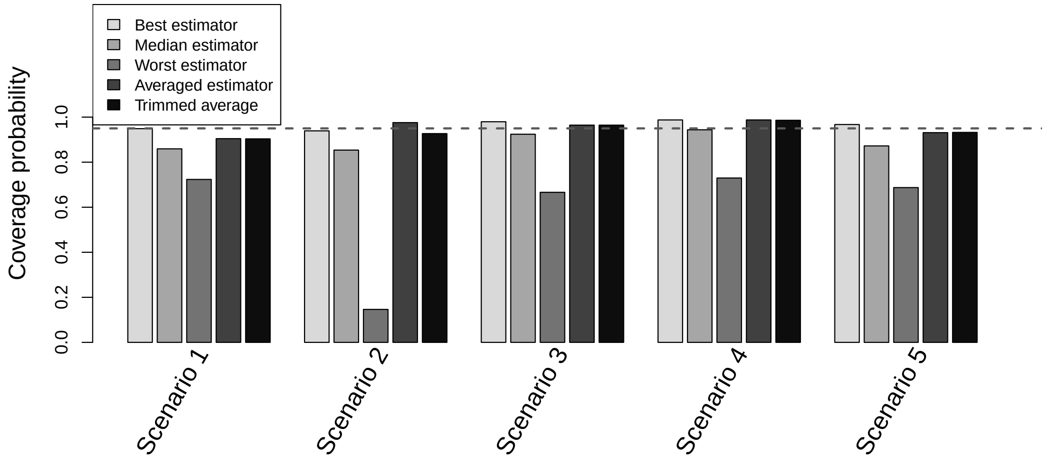

Figure 2: Interval coverage for the estimator with the best interval coverage, median interval coverage, worst interval coverage, the averaged estimator, and for the trimmed average estimator.

Overall, the results confirmed the theoretical results that the averaged estimator does better than the worst individual estimator, though more importantly, the simulation showed that averaging estimators tend to perform more closely to the best individual estimator (at least in the scenarios considered). It should be noted that the individual estimator that performed the best or worst in each scenario was different, and therefore there is not one estimator outperforming the averaged estimator across all scenarios. Further, the results showed that the conservative estimate of the variance is not leading to overly conservative confidence intervals for the averaging estimators. The trimmed average estimator did not provide drastically different results

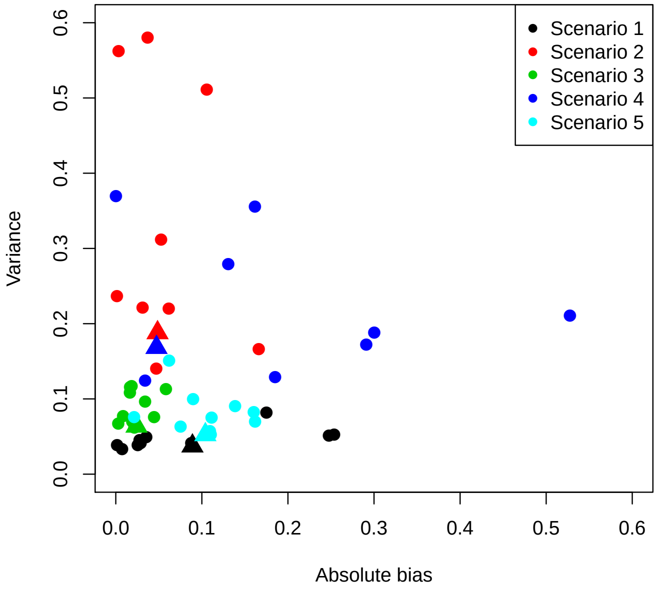

Figure 3: Absolute bias and variance for each of the estimators, over all scenarios looked at. The averaged estimator is denoted by large triangles, while the other estimators are denoted by small circles.

from the averaging estimator that assigns equal weights to all of the estimators, though it did tend to reduce the MSE slightly.

To assess the impact of bias and variance of the averaged estimator, Figure 3 shows the absolute bias and variance for each individual estimator across all the simulation scenarios considered. The circles represent the individual estimators while the triangles represent the averaged estimator for each scenario. The averaged estimator had small variance relative to the individual estimators, which is expected given the efficiency gains that are possible as indicated in Section 2.2. In terms of bias, the averaged estimator is guaranteed to have exactly the average bias of all of the existing estimators. This can lead to reductions in absolute bias in cases where certain estimators have positive bias while others have negative bias, which was observed in Scenario 4. In other cases, such as Scenarios 2 and 5, the direction of the bias for the individual estimators was the same, and the absolute bias of the averaged estimator was simply the average of the absolute biases for the individual estimators.

4 Application to NHANES data

Environmental wide association studies (EWAS) measure a variety of chemicals and toxins that humans are exposed to in an effort to understand how they affect the biological processes in the human body [37, 38]. The National Health and Nutrition Examination Survey (NHANES), is a cross-sectional data source made publicly available by the Centers for Disease Control and Prevention (CDC), and we aim to estimate the effects of a number of environmental exposures on three different outcomes: HDL cholesterol levels, LDL cholesterol levels, and triglyceride levels in the NHANES study. We use data found in [39], which contains a large number of potential confounders. Participants both filled out questionnaires regarding their health status, as well as received clinical and laboratory tests that contain information on environmental factors such as pollutants, allergens, bacterial/viral organisms, chemical toxicants, and nutrients. In previous work [40], the environmental agents were separated into different groups containing similar agents that might affect similar biological pathways. We look at the effects of 22 different environmental agent groups on the three outcomes, leading to 66 different analyses. Each exposure is defined as the average exposure level across all agents within the same grouping. As most of the approaches considered only consider binary treatments, we dichotomized the continuous exposure by classifying each exposure as being above or below the median exposure level. Each of the environmental agents was measured in a different subset of the NHANES data, leading to different populations, sample sizes, and covariate dimensions for each of the 22 different exposures. Some of the data sets have very small p/n ratios that do not require high-dimensional techniques, while others have p/n ratios between 0.3-0.4 and require some form of dimension reduction. In the following sections we show how the averaged estimator performs in comparison with the individual estimators, and how this varies as a function of the p/n ratio.

4.1 Agreement of statistical decisions

First, we evaluate whether the averaged estimator would lead to different statistical inferences. In particular, we are interested in testing the null hypothesis of no treatment effect, and evaluate if the different estimators lead to different conclusions of statistical significance (i.e., is the null hypothesis rejected). Figure 4 shows the results of the hypothesis tests for any data set in which there was disagreement among the estimators. Each row corresponds to a particular estimator, while each column corresponds to a particular data set. The averaged estimator generally sides with the majority of the individual estimators in terms of rejecting or failing to reject the null hypothesis, although this is not always the case. In the sixth column, 7 of the 10 estimators failed to reject the null hypothesis of no treatment effect, but the averaged estimator leads to a rejection of the null hypothesis. It is important to emphasize that Figure 4 is limited to cases where different estimators lead to different conclusions, which represents only 30% of analyses performed.

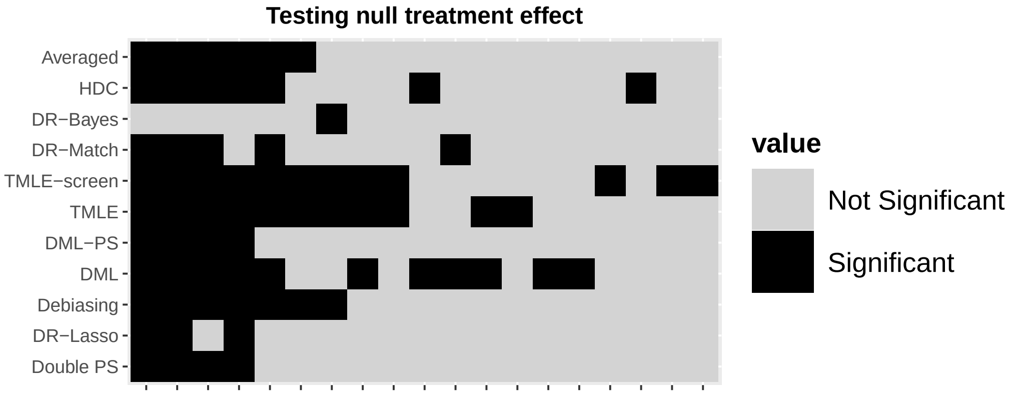

Figure 4: Plot of whether or not each estimator rejected a particular null hypothesis. The results only show data sets in which there was some disagreement among the 11 estimators in terms of the final hypothesis test.

4.2 When does averaging matter?

Next, we evaluated the potential impact that averaging can have on estimation within the NHANES data set, and attempted to identify where it is the most useful. Assuming that each estimator has only a small amount of bias, the main advantage of averaging is the efficiency gains one can achieve. When the individual estimators are highly correlated, there is very little to be gained (though also very little to be lost) by using averaging. On the other hand, when estimators are weakly correlated with each other, the averaged estimator gains more efficiency. The left panel of Figure 5 shows the correlation matrix of the 10 estimators across the data sets examined. The correlation is generally very high, with the majority of the values on the off-diagonal above 0.9. The right panel, however, depicts a somewhat different story when we restricted attention to the data sets with p/n ratios above 0.25. In these settings, there is less correlation between the estimators, with some of the correlations as low as 0.4, and many more in the 0.6 to 0.8 range. This illustrates that the potential for

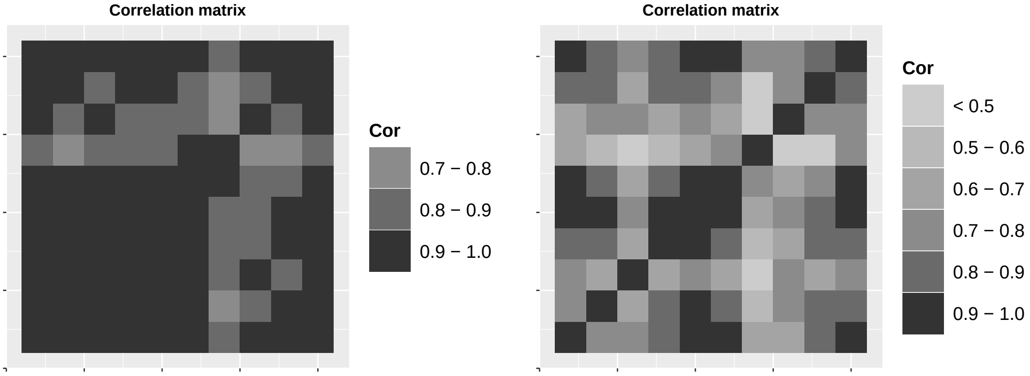

Figure 5: Correlation matrix across the 10 estimators over the analyses in the NHANES data. The right panel only shows this correlation for those data sets with a p/n ratio that is above 0.25

variance reduction by averaging estimators is greater in the the more difficult setting with a larger number of covariates relative to the sample size.

5 Discussion

We presented a simple approach to estimate the average treatment effect in high-dimensional settings that leverages many of the estimators that have been developed in recent years. As more approaches are developed, it becomes increasingly difficult for an analyst to choose an approach for their particular study. Further, there is likely no best approach in all scenarios, and the extent to which each estimator performs is highly dependent on the unknown data generating process. Averaging over all of the estimators provides an ideal choice when presented with so many estimators. Not only does the averaged estimator guarantee improved performance over the worst estimator in terms of bias and MSE, but our empirical results indicate that it can perform much more closely to the best of the individual estimators without knowing beforehand which estimators are best. At times, it can even outperform the best of the individual estimators with respect to MSE or interval coverage. Further, we derived a conservative estimate of the variance of the averaged estimator that leads to reasonable confidence interval coverage.

A potential criticism of the averaged estimator is that each individual estimator relies on its own set of assumptions, and therefore the averaged estimator assumes all of these assumptions in order to achieve unbiased inference. While this is generally true, this in some ways shows the strength of the averaged estimator. All of these assumptions are untestable, and we don’t know, and cannot test, which estimators’ assumptions are being violated. By averaging over all estimators, we mitigate the impact that gross misspecification of one of these assumptions could have if we simply chose the one, bad estimator. An analyst could alternatively choose their preferred estimator or the one with the minimum estimated variance among all candidate estimators. While this may work in certain scenarios, it can perform extremely poorly if the chosen estimator has a high bias. Further, simply choosing the estimator with the smallest estimated variance will likely lead to anti-conservative confidence intervals. The averaged estimator removes the burden of this decision from the analyst and protects against an estimator with high bias, high variance, or low interval coverage. We believe that including as many estimators as possible into the list of candidate estimators is ideal unless some estimators are known a priori to have high bias or variance. Including more estimators reduces the impact that a small number of bad estimators can have, while increasing the efficiency of the averaged estimator. Empirically we explored the possibility of using a trimmed average estimator that removes extreme estimators from the candidate list. Trimmed estimators such as this can fix the issue of one or two estimators having extremely large bias or variance, and can still lead to improvements in efficiency and interval coverage of the estimators.

While we believe that more individual estimators can lead to improved performance, it is important to stress that individual estimators should only be used if properly tuned and implemented. There are many subjective decisions that go into each of the individual estimators, and the performance of the averaged estimator will be improved if each of the individual estimators are implemented carefully. If an analyst is not comfortable with certain estimators and their respective implementation, then these should be removed from the averaged estimator. Further, there is a trade-off between the number of estimators and both the computational and implementation burden of utilizing more approaches. As more individual estimators are introduced, there is a greater burden on the analyst in terms of implementation effort and computation time required to fit all of the estimators, and this should be considered by the analyst before using the averaged estimator. We tried to alleviate these concerns by providing a library of estimators in our software package that are tuned to provide individual estimates of the causal effect, which works well across a wide range of data generating mechanisms that we have evaluated. Nonetheless, if an analyst is able to carefully tune individual estimators within the averaged estimator, that should improve performance of the averaged estimator.

Software

An R package for implementing the averaging estimator can be found at https://github.com/jantonelli111/AveragingCausalHD.

References

[1] D. B. Rubin and N. Thomas, "Matching using estimated propensity scores: relating theory to practice," *Biometrics*, pp. 249–264, 1996.

[2] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[3] P. R. Rosenbaum and D. B. Rubin, "Reducing bias in observational studies using subclassification on the propensity score," *Journal of the American statistical Association*, vol. 79, no. 387, pp. 516–524, 1984.

[4] H. Bang and J. M. Robins, "Doubly robust estimation in missing data and causal inference models," *Biometrics*, vol. 61, no. 4, pp. 962–973, 2005.

[5] C. M. Crainiceanu, F. Dominici, and G. Parmigiani, "Adjustment uncertainty in effect estimation," *Biometrika*, vol. 95, no. 3, pp. 635–651, 2008.

[6] S. Vansteelandt, M. Bekaert, and G. Claeskens, "On model selection and model misspecification in causal inference," *Statistical methods in medical research*, vol. 21, no. 1, pp. 7–30, 2012.

[7] C. Wang, G. Parmigiani, and F. Dominici, "Bayesian effect estimation accounting for adjustment uncertainty," *Biometrics*, vol. 68, no. 3, pp. 661–671, 2012.

[8] A. Wilson and B. J. Reich, "Confounder selection via penalized credible regions," *Biometrics*, vol. 70, no. 4, pp. 852–861, 2014.

[9] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American statistical association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[10] S. M. Shortreed and A. Ertefaie, "Outcome-adaptive lasso: Variable selection for causal inference," *Biometrics*, 2017.

[11] A. Ertefaie, M. Asgharian, and D. A. Stephens, "Variable selection in causal inference using a simultaneous penalization method," *Journal of Causal Inference*, vol. 6, no. 1, 2018.

[12] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[13] J. Antonelli, M. Cefalu, N. Palmer, and D. Agniel, "Doubly robust matching estimators for high dimensional confounding adjustment," *Biometrics*, 2016.

[14] J. Antonelli, G. Parmigiani, F. Dominici, *et al.*, "High-dimensional confounding adjustment using continuous spike and slab priors," *Bayesian Analysis*, vol. 14, no. 3, pp. 825–848, 2019.

[15] P. R. Hahn, C. Carvalho, and D. Puelz, "Bayesian regularized regression for treatment effect estimation from observational data," *Available at SSRN*, 2016.

[16] A. Belloni, V. Chernozhukov, and C. Hansen, "Inference on treatment effects after selection among high-dimensional controls," *The Review of Economic Studies*, vol. 81, no. 2, pp. 608–650, 2014.

[17] M. H. Farrell, "Robust inference on average treatment effects with possibly more covariates than observations," *Journal of Econometrics*, vol. 189, no. 1, pp. 1–23, 2015.

[18] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins, "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, vol. 21, no. 1, pp. C1–C68, 2018.

[19] J. Antonelli and F. Dominici, "A bayesian semiparametric framework for causal inference in high-dimensional data," *arXiv preprint arXiv:1805.04899*, 2018.

[20] S. Athey, G. W. Imbens, and S. Wager, "Approximate residual balancing: debiased inference of average treatment effects in high dimensions," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 80, no. 4, pp. 597–623, 2018.

[21] Z. Tan, "Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data," *arXiv preprint arXiv:1710.08074*, 2017.

[22] Z. Tan, "Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data," *arXiv preprint arXiv:1801.09817*, 2018.

[23] Y. Ning, S. Peng, and K. Imai, "Robust estimation of causal effects via high-dimensional covariate balancing propensity score," *arXiv preprint arXiv:1812.08683*, 2018.

[24] M. J. Van Der Laan and D. Rubin, "Targeted maximum likelihood learning," *The International Journal of Biostatistics*, vol. 2, no. 1, 2006.

[25] C. Ju, S. Gruber, S. D. Lendle, A. Chambaz, J. M. Franklin, R. Wyss, S. Schneeweiss, and M. J. van der Laan, “Scalable collaborative targeted learning for high-dimensional data,” *Statistical methods in medical research*, vol. 28, no. 2, pp. 532–554, 2019.

[26] C. Ju, R. Wyss, J. M. Franklin, S. Schneeweiss, J. Häggström, and M. J. van der Laan, “Collaborative-controlled lasso for constructing propensity score-based estimators in high-dimensional data,” *Statistical methods in medical research*, vol. 28, no. 4, pp. 1044–1063, 2019.

[27] N. S. Hejazi, S. Kherad-Pajouh, M. J. van der Laan, and A. E. Hubbard, “Variance stabilization of targeted estimators of causal parameters in high-dimensional settings,” *arXiv preprint arXiv:1710.05451*, 2017.

[28] D. Draper, “Assessment and propagation of model uncertainty,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 45–70, 1995.

[29] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky, “Bayesian model averaging: a tutorial,” *Statistical science*, pp. 382–401, 1999.

[30] N. L. Hjort and G. Claeskens, “Frequentist model average estimators,” *Journal of the American Statistical Association*, vol. 98, no. 464, pp. 879–899, 2003.

[31] B. E. Hansen, “Least squares model averaging,” *Econometrica*, vol. 75, no. 4, pp. 1175–1189, 2007.

[32] R. C. Mittelhammer and G. G. Judge, “Combining estimators to improve structural model estimation and inference under quadratic loss,” *Journal of econometrics*, vol. 128, no. 1, pp. 1–29, 2005.

[33] J. Antonelli, B. Han, and M. Cefalu, “A synthetic estimator for the efficacy of clinical trials with all-or-nothing compliance,” *Statistics in medicine*, vol. 36, no. 29, pp. 4604–4615, 2017.

[34] M. Ghosh, J. Rao, *et al.*, “Small area estimation: an appraisal,” *Statistical science*, vol. 9, no. 1, pp. 55–76, 1994.

[35] C.-W. Chen and H.-C. Yang, “OPATs: Omnibus P-value association tests,” *Briefings in Bioinformatics*, vol. 20, pp. 1–14, 07 2017.

[36] A. Chatterjee and S. Lahiri, “Asymptotic properties of the residual bootstrap for lasso estimators,” *Proceedings of the American Mathematical Society*, vol. 138, no. 12, pp. 4497–4509, 2010.

[37] C. P. Wild, “Complementing the genome with an “exposome”: the outstanding challenge of environmental exposure measurement in molecular epidemiology,” *Cancer Epidemiology Biomarkers & Prevention*, vol. 14, no. 8, pp. 1847–1850, 2005.

[38] C. J. Patel and J. P. Ioannidis, “Studying the elusive environment in large scale,” *Jama*, vol. 311, no. 21, pp. 2173–2174, 2014.

[39] A. Wilson, C. Zigler, C. Patel, and F. Dominici, “Model-averaged confounder adjustment for estimating multivariate exposure effects with linear regression,” *Biometrics*, 2018.

[40] C. J. Patel, M. R. Cullen, J. P. Ioannidis, and A. J. Butte, “Systematic evaluation of environmental factors: persistent pollutants and nutrients correlated with serum lipid levels,” *International journal of epidemiology*, vol. 41, no. 3, pp. 828–843, 2012.

A Proof of result 1

First we will show the result for bias:

$$\begin{aligned}
\left|E\left(\frac{1}{M}\sum_{m=1}^{M}\widehat{\theta}_m\right)-\theta\right| &= \left|\frac{1}{M}\sum_{m=1}^{M}\left(E(\widehat{\theta}_m)-\theta\right)\right| \\
&\leq \frac{1}{M}\sum_{m=1}^{M}\left|E(\widehat{\theta}_m)-\theta\right| \\
&\leq \frac{1}{M}\sum_{m=1}^{M}\left|E(\widehat{\theta}^{\star})-\theta\right| \\
&= \left|E(\widehat{\theta}^{\star})-\theta\right|
\end{aligned}$$

Now for the MSE:

$$\begin{aligned}
E\left[\left(\frac{1}{M}\sum_{m=1}^{M}\widehat{\theta}_m-\theta\right)^2\right] &= E\left[\frac{1}{M^2}\left(\sum_{m=1}^{M}(\widehat{\theta}_m-\theta)\right)^2\right] \\
&= \frac{1}{M^2}E\left[\sum_{m=1}^{M}(\widehat{\theta}_m-\theta)^2+\sum_{i\neq j}(\widehat{\theta}_i-\theta)(\widehat{\theta}_j-\theta)\right] \\
&\leq \frac{1}{M^2}E\left[\sum_{m=1}^{M}(\widehat{\theta}_m-\theta)^2+\sum_{i\neq j}|\widehat{\theta}_i-\theta||\widehat{\theta}_j-\theta|\right] \\
&\leq \frac{1}{M^2}\left[M\star E\left((\widehat{\theta}^{\star\star}-\theta)^2\right)+\sum_{i\neq j}\sqrt{E\left((\widehat{\theta}_i-\theta)^2\right)E\left((\widehat{\theta}_j-\theta)^2\right)}\right] \\
&\leq \frac{1}{M^2}\left[M\star E\left((\widehat{\theta}^{\star\star}-\theta)^2\right)+\sum_{i\neq j}E\left((\widehat{\theta}^{\star\star}-\theta)^2\right)\right] \\
&= E\left((\widehat{\theta}^{\star\star}-\theta)^2\right)
\end{aligned}$$

B Computational details for each estimator

Here we illustrate the computation time of the 10 individual estimators used in the manuscript when the dimension of the data is $n = 150$ and $p = 300$. The computation time (in minutes) can be found in Figure B1. All of the approaches have computation times smaller than 10 seconds with the exception of DR-Bayes and HDC, which is unsurprising given that both of these approaches require MCMC chains to be run for high-dimensional Bayesian models. If computational resources and time are a priority, then the averaged estimator can be fit without including these Bayesian approaches into the library of estimators.

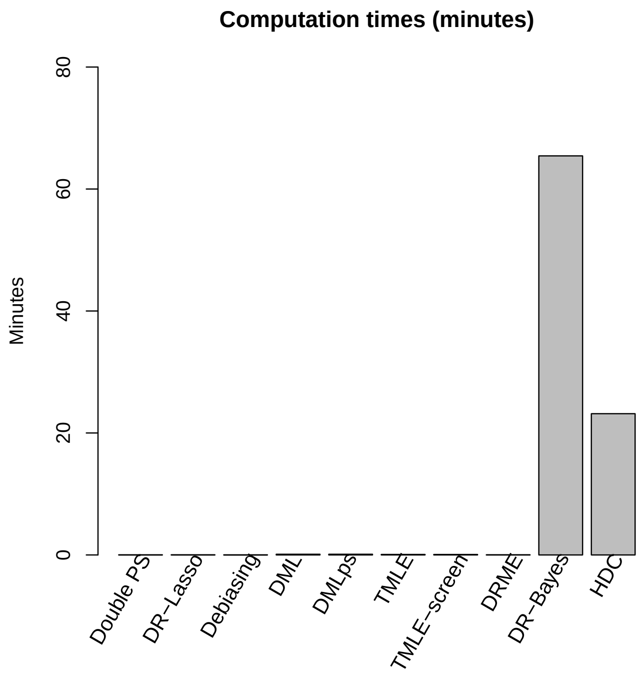

Figure B1: Computation time for individual estimators

C Data generating mechanisms for simulation scenarios and additional simulation results

Here we will detail the exact data generating processes for each of the simulation scenarios considered in the paper. In addition, we will provide five new simulation scenarios and present the results for them as well.

Scenario 1: For the first simulation scenario, we generated data under the following models for the treatment and the outcome processes:

$$\begin{aligned} Y_i &= T_i + \boldsymbol{X}_i\boldsymbol{\beta} + \epsilon_i \\ P(T_i = 1) &= \Phi(\boldsymbol{X}_i\gamma) \\ \boldsymbol{X}_i &\sim \text{MVN}(\mathbf{0}_p, \boldsymbol{\Sigma}_X) \\ \epsilon_i &\sim \mathcal{N}(0, 1), \end{aligned} \tag{C1}$$

where $\boldsymbol{\Sigma}_X$ is an exchangeable correlation matrix with correlation 0.3, $n = 150$, $p = 300$, $\boldsymbol{\beta} = (0.75, 1, 0.6, -0.8, -0.7, \mathbf{0}_{p-5})$, and $\gamma = (0.15, 0.2, 0, 0, -0.4, \mathbf{0}_{p-5})$.

Scenario 2: This simulation is based on a simulation from [20] that has a dense treatment model. Define 20 clusters, $\{\mathbf{c_1}, \ldots, \mathbf{c_{20}}\}$ where $\mathbf{c_k} \sim \mathcal{N}(0, I_p)$. For each subject, we randomly sample one of these 20 vectors, which we call $\mathbf{C}_i$. Then we draw the vector of covariates for subject i, denoted by $\mathbf{X}_i$, from a multivariate normal distribution centered at $\mathbf{C}_i$ with the identity matrix as the covariance. We set $T_i = 1$ with probability 0.1 for the first 10 clusters, and $T_i = 1$ with probability 0.9 for the remaining clusters. Finally, we generate data from the outcome model defined as $Y_i = 10T_i + \boldsymbol{X\beta} + \epsilon_i$, where $\boldsymbol{\beta} \propto (1, \frac{1}{\sqrt{2}}, \ldots, \frac{1}{\sqrt{p}})$ and is normalized such that $||\boldsymbol{\beta}||_2^2 = 18$. Lastly, $n = 300$ and $p = 300$.

Scenario 3: This scenario uses the same structure as in equation C1 and simulation scenario 1, with the exception of $\boldsymbol{\beta}$ and γ. In this case, we set $\boldsymbol{\beta} = (0.2, 0.3, 0.15, -0.15, 0.15, -0.15, \ldots 0.15, -0.15, \mathbf{0}_{p-74})$, and $\gamma = (0.6, -0.5, \mathbf{0}_{p-2})$.

Scenario 4: This scenario uses the same simulation structure as scenario 3, except the outcome model is chosen to be nonlinear. Now we generate the outcomes from a normal distribution with variance 1 and mean given by

$$E(Y_i|T_i, \boldsymbol{X}_i) = -2 + T_i + 0.5X_{1i} + 0.5X_{2i}^2 + 0.8X_{2i}^3 + 0.3e^{X_{3i}}$$

Scenario 5: This scenario uses the same simulation structure as simulation 1 except for the treatment and outcome models. In this setting, we have $P(T_i = 1) = \Phi(-0.7 + \boldsymbol{X}\gamma)$, where $\gamma = (0.2, 0, 0.4, -0.5, \mathbf{0}_{p-4})$ and we generate the outcome with a variance of 1 and mean given by

$$E(Y_i|T_i, \boldsymbol{X}_i) = -2 + T_i + 0.5X_{1i} + 0.8X_{2i} + 0.4X_{3i} + 0.5T_iX_{3i}$$

Additional scenario 1: This scenario uses the same simulation structure as the first simulation scenario of the main manuscript, however, both $\boldsymbol{\beta}$ and γ are $\mathbf{0}_p$ and there is no confounding.

Additional scenario 2: This scenario uses the same simulation structure as the first simulation scenario of the main manuscript, however,

$$\boldsymbol{\beta} = (0.9, 0.9, 0.2, 0.2, 0, 0, 0.9, 0.9, \mathbf{0}_{p-8}),$$

and

$$\gamma = (0.4, 0.9, -0.5, -0.7, -0.3, 0.6, \mathbf{0}_{p-6}).$$

Additional scenario 3: This scenario is the same as the second simulation scenario of the main manuscript, except that $T_i = 1$ with probability 0.25 and 0.75 for the first 10 and remaining clusters, respectively. Further, we set $\boldsymbol{\beta} \propto (1, \frac{1}{2}, \ldots, \frac{1}{p})$.

Additional scenario 4: This scenario is the same as the first simulation scenario of the main manuscript except the treatment model is nonlinear. In particular the treatment probabilities are given by

$$P(T_i = 1) = \text{expit}(0.5X_{1i}^3 + 0.3X_{1i}^2 - 0.3X_{2i}^4 + 0.4 \star X_{3i}^2)$$

Additional scenario 5: This scenario is the same as the fourth additional simulation scenario except that the outcome model is also nonlinear. Specifically, the outcome is drawn from a normal distribution with variance 1 and mean given by

$$E(Y_i|T_i, \boldsymbol{X}_i) = -2 + T_i + 0.5X_{1i} + 0.5X_{2i}^2 + 0.8X_{2i}^3 + 0.3e^{X_{3i}}$$

The first additional simulation scenario looks at a scenario with no confounding among the covariates. The second additional scenario looks at a different sparse data generating process for both the treatment and outcome models. The third additional scenario looks at a different scenario where the treatment model is not sparse. The fourth scenario looks at a case where the treatment model is nonlinear and therefore most approaches will be misspecified. The fifth additional scenario looks at a case where both the treatment and outcome models are nonlinear. The results from these additional simulations can be found in Figures C1 and C2. In every one of the scenarios looked at the averaged estimators do far better than the worst estimator and better than the median estimator. In some cases, the averaged estimators do far better than the median estimator and perform as well or better than the best individual estimator. The averaged estimator also obtains the nominal 95% coverage rate in every simulation scenario except for the fifth scenario. The only scenario in which the averaged estimators were substantially outperformed by the best individual estimator was scenario 5. In this scenario there was only one estimator that was unbiased and had nominal confidence interval coverages. The remaining estimators all had substantial biases and low interval coverages. Overall these results confirm the results from the manuscript. The averaged estimator, whether it is trimmed or not, shows performance that is much closer to the best individual estimator than either the median or worst individual estimator. This leads to low MSE values and interval coverages that are much closer to the nominal rate when compared with the majority of the individual estimators.

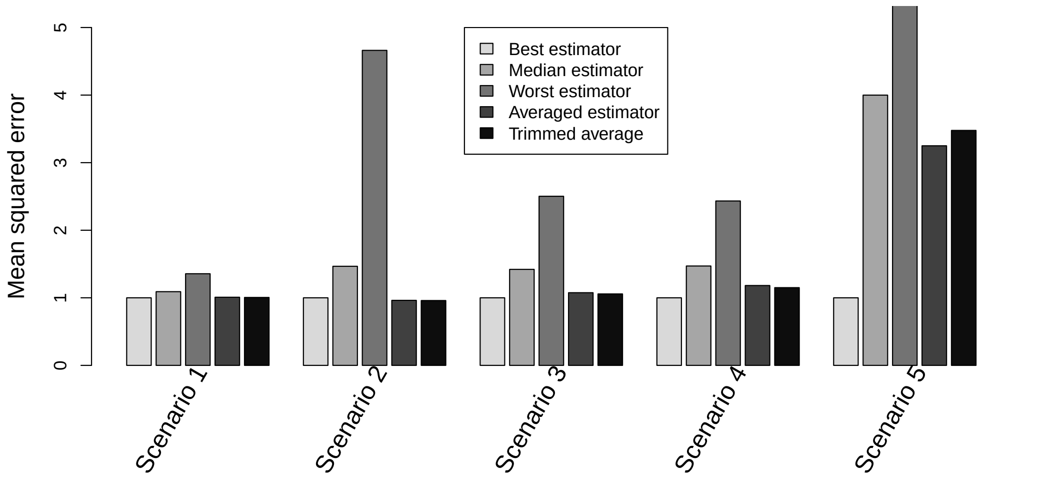

Figure C1: Mean squared error for the estimator with the lowest MSE, highest MSE, median MSE, the averaged estimator, and the trimmed average estimator. Note that these results are for the five additional simulation results of the appendix.

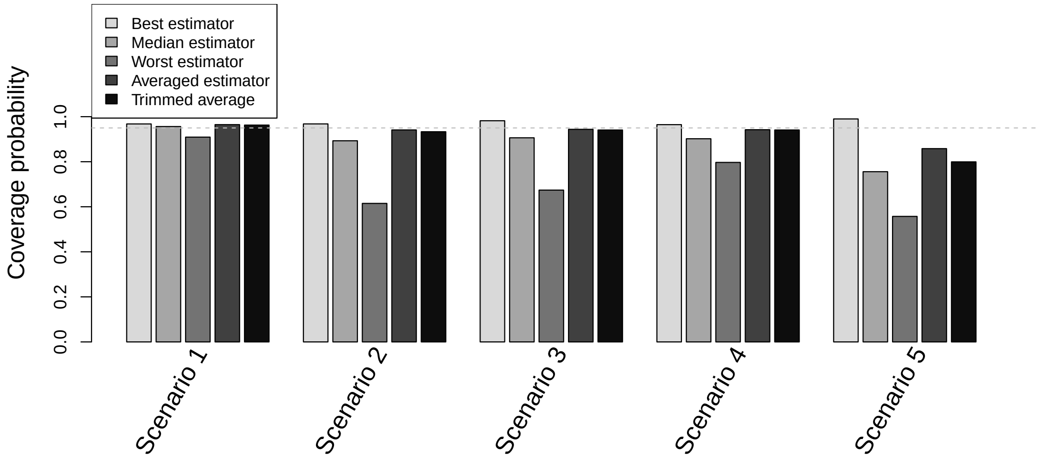

Figure C2: Interval coverage for the estimator with the best interval coverage, median interval coverage, worst interval coverage, the averaged estimator, and for the trimmed average estimator. Note that these results are for the five additional simulation results of the appendix.