Bernoulli **31**(4), 2025, 2545–2568
https://doi.org/10.3150/24-BEJ1782

Treatment effect estimation with efficient data aggregation

SNIGDHA PANIGRAHI[1,a], JINGSHEN WANG[2,b] and XUMING HE[3,c]

[1]*Department of Statistics, University of Michigan, Ann Arbor, USA,* [a]*psnigdha@umich.edu*
[2]*Division of Biostatistics, UC Berkeley, Berkeley, USA,* [b]*jingshenwang@berkeley.edu*
[3]*Department of Statistics and Data Science, Washington University in St. Louis, Saint Louis, USA,* [c]*hex@wustl.edu*

Data aggregation, also known as meta analysis, is widely used to combine knowledge on parameters shared in common (e.g., average treatment effect) between multiple studies. In this paper, we introduce an attractive data aggregation scheme that pools summary statistics from various existing studies. Our scheme informs the design of new validation studies and yields unbiased estimators for the shared parameters. In our setup, each existing study applies a LASSO regression to select a parsimonious model from a large set of covariates. It is well known that post-hoc estimators, in the selected model, tend to be biased. We show that a novel technique called *data carving* yields us a new unbiased estimator by aggregating simple summary statistics from all existing studies. The proposed estimator has two key features: (a) we make the fullest possible use of data, from all studies, without the risk of bias from model selection; (b) we enjoy the added benefit of individual data privacy, because raw data from these studies need not be shared or stored for efficient estimation.

Keywords: Conditional inference; debiased estimation; data carving; LASSO; meta analysis

1. Introduction

Aggregating knowledge across studies is a popular practice to pool multiple findings, increase precision of shared results, and plan future studies. In this paper, we focus on the problem of treatment effect evaluation after adjusting for potential confounders from various independent, existing studies. Existing studies in our setup can be pilot studies, preliminary analyses, or previously reported results, all evaluating the same treatment effect. Our paper introduces a data aggregation scheme which informs an efficient design for follow-up validation studies, and yields us estimators for the treatment effects by pooling summary statistics from the existing studies.

In the context of high-dimensional data, the number of potential measured confounders might be very large. Estimation of treatment effects in the presence of many possible confounders usually calls for a model selection procedure in any individual study. Two practical considerations gain prominence in this setup. First, full data from the existing studies, e.g., raw data for the confounders, might be unavailable, due to either practical limitations in data-sharing, or due to legitimate privacy and confidentiality concerns (see Wolfson et al., 2010, Cai, Liu and Xia, 2022, for example). Second, if model selection is performed on each existing study, post-hoc estimation of treatment effects in the selected model tends to be biased. We thus ask: (i) whether and how data from two or more studies can be aggregated without the risk of selection bias, and (ii) whether the former goal can be achieved by using only summary statistics that are sufficient for data aggregation.

To address the questions raised above, we focus on the very widely used LASSO regression (Tibshirani, 1996), which is simply called the LASSO. By placing a penalty on the absolute value of regression coefficients, the LASSO shrinks many of them to be exactly zero and leads to a more parsimonious model. In each existing study, when the LASSO is used to select a set of covariates as confounders, the

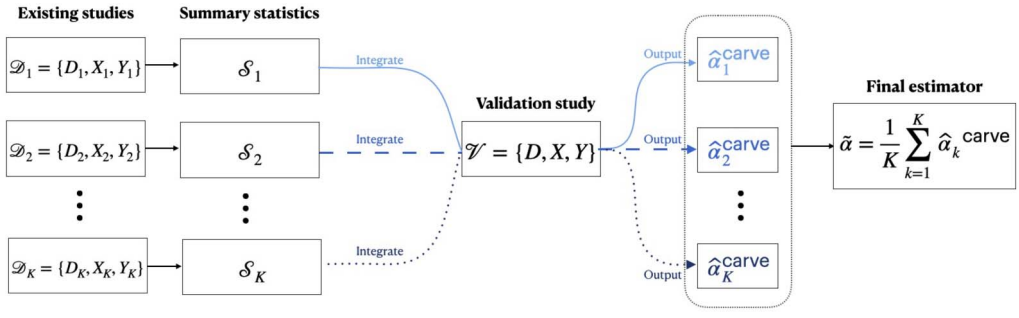

Figure 1. Schematic depiction of efficient and privacy-preserving data aggregation. The exact form for the summary statistics in our scheme and our estimator $\widetilde{\alpha}$ is provided in Section 2.

usual summary statistics—the first two moments of data required to conduct regression in the selected model—are no longer sufficient for obtaining unbiased estimators through data aggregation. It, however, turns out that the summary statistics required by our aggregation scheme assume a fairly simple and compressed form. An immediate by-product of this finding is a new guideline for what needs to be reported in publications for an efficient and unbiased estimation of shared parameters in models selected by the LASSO.

While data aggregation from existing studies can be performed on their own, follow-up validation studies are either necessitated by a regulatory requirement, or pursued to report findings with higher statistical power. Researchers in science and public health usually find it more cost-efficient to design and carry out new validation studies by relying on a synthesis of prior research. This practice allows researchers to focus on a smaller set of confounding factors, and reduce variability of estimators from any single study at the same time. For example, research into cost-efficient trial designs has led to the development of methodology for seamless phase II/III designs in clinical trials. Some of these approaches have been reviewed by Stallard and Todd (2011). A more recent Lancet article by Park et al. (2021) highlights not just the need for well designed clinical trials, but also emphasizes the need that these trials are structured as per a master protocol in a coordinated and collaborative manner. Back to our problem setup, the validation study $\mathcal{V}$ collects data on a subset of potential confounders that were selected by any of the already existing studies.

We showcase how a recent technique, namely *data carving*, can be used to efficiently aggregate summary statistics from each existing study for an unbiased estimation of treatment effects. Data carving allows us to re-use information from data that has not been been fully exploited by the LASSO at the time of model selection. A schematic depiction of our aggregation scheme is provided in Figure 1. Suppose that we have K existing studies. We apply data carving to aggregate summary statistics from existing study k with data from the validation study $\mathcal{V}$. This gives us our new estimator for treatment effects $\widehat{\alpha}_k^{\text{carve}}$, which we call the carved estimator. A simple averaging of the K carved estimators allows us to pool information from the existing studies, and yields us our final aggregated estimator $\widetilde{\alpha}$.

1.1. Related work

A simple estimator in our problem setup is the split-and-aggregate estimator. The split-and-aggregate estimator is obtained by first (i) refitting the selected model from each existing study on the validation dataset, followed by (ii) averaging the K split-based estimators. Because the data used for model selection is not re-used at the time of estimation, the resulting estimator does not suffer from the issue of the

selection bias. However, this estimator suffers from a loss of efficiency as it discards all data from the existing studies. This loss in efficiency is especially noticeable if the validation study is small relative to the existing studies.

Inspired by recent efforts to re-use information from model selection (Fithian, Sun and Taylor, 2014, Dwork et al., 2015, Tian, 2020), the carved estimator makes the fullest possible use of data for estimating treatment effects. Data carving, as introduced in Fithian, Sun and Taylor (2014), resembles data-splitting in the selection step, because model selection is carried out on only a subset of the full data. But, carved estimators, unlike those based on data splitting, use the full data instead of the held-out portion alone. More precisely, carved estimators are obtained by conditioning on the observed selection event to remove bias from the model selection step. We point out previous work such as Panigrahi, Taylor and Weinstein (2021), Panigrahi (2023), Schultheiss, Renaux and Bühlmann (2021), which deploys data carving to obtain interval estimators for selected regression coefficients. More generally, papers by Lee et al. (2016), Tian and Taylor (2018), Panigrahi and Taylor (2018), Gao, Bien and Witten (2024), Panigrahi et al. (2023), Chen and Bien (2020), Panigrahi and Taylor (2023), Panigrahi, MacDonald and Kessler (2023) have used conditioning as an effective technique to offer post-selection inference through valid p-values, confidence intervals, confidence regions, and credible intervals in various problems.

More sophisticated proposals by Tang, Zhou and Song (2020), Lee et al. (2017), and Maity, Sun and Banerjee (2019) aggregate debiased LASSO estimators from distributed datasets. In the same spirit, Lin and Zeng (2010) and Cai, Liu and Xia (2022) provide a more efficient, inversely variance-weighted debiased LASSO estimator to accommodate cross-study heterogeneity in similar settings. The proposed debiased procedures in these papers offer a valid way out to mitigate model selection bias in high-dimensional settings. But, the debiasing correction involves estimating and inverting a high-dimensional information matrix. Moreover, the reduction in bias comes at a price in statistical efficiency. Especially, this price is large when the treatment assignment is correlated with noise variables in the model. For example, Wang, He and Xu (2020) have previously illustrated this phenomenon on large observational studies. Not only does our new estimator avoid estimation (and storage) of high-dimensional matrices for removing bias, but also delivers higher statistical efficiency than existing estimators in the situation described above.

1.2. Organization of paper

The paper is organized as follows. Section 2 formally states our problem setup and introduces the main ideas behind our new estimator through a first example. In Section 3, we outline the steps to obtain our carved estimator through an aggregation of summary statistics, and develop the asymptotic theory for the new estimator. Numerical experiments that support our asymptotic results are presented in Section 4. In Section 5, we highlight the efficacy of our data aggregation and estimation scheme on synthetic data generated from a clinical trial. Section 6 concludes our work with some summarizing remarks. Proofs of our technical results are collected in the Supplementary Material (Panigrahi, Wang and He, 2025).

2. Methodology

2.1. Problem setup

Basic setup. We consider K independent, existing studies in our setup. Each study measures n_k independently and identically distributed triples: (i) $Y_k \in \mathbb{R}^{n_k \times 1}$ is the observed outcome, (ii) $X_k \in \mathbb{R}^{n_k \times p_k}$

represents the high-dimensional covariates where p_k is allowed to be larger than n_k, and (iii) $D_k \in \mathbb{R}^{n_k \times s}$ is the $\mathbb{R}^s$-valued treatment variable. We denote the data from these studies by

$$\mathcal{D}_k = \{D_k, X_k, Y_k\},$$

for $k \in \{1, 2, \cdots, K\}$. For any k, we use E_k to represent an index set of covariates with cardinality less than p_k. For any vector or matrix A and an index set E, let A_E be the sub-vector or sub-matrix containing the columns of A indexed by the set E. We use the symbol 1_k and 0_k to represent a $\mathbb{R}^k$-valued vector of all ones and all zeroes, respectively, and use $\|v\|$ to represent the ℓ_2 norm of a vector unless specified otherwise.

We begin with a simple model to describe the data in our studies. Suppose that $\varepsilon_k \in \mathbb{R}^{n_k \times 1}$ are random errors that satisfy

$$\mathbb{E}(\varepsilon_k | D_k, X_k) = 0_{n_k}, \quad \mathrm{Cov}(\varepsilon_k | D_k, X_k) = \sigma^2 I, \quad k = 1, 2, \ldots, K,$$

where I is the identity matrix. We assume that the observations in our studies are drawn from the population model

$$Y_k = D_k \alpha + X_{k,E_0} \beta_{E_0} + \varepsilon_k, \tag{1}$$

where $\alpha \in R^s$ measures the treatment effect, E_0 indicates the covariates with nonzero effects, and β_{E_0} measures the impact of these covariates.

We note that in our manuscript, the term "treatment effect" denotes the coefficients of shared variables across multiple studies. We note that this does not necessarily refer to causal effects, which typically require additional identification conditions, such as assumptions of unconfoundedness. This terminology aligns with the existing high-dimensional inference literature, as in Cattaneo, Jansson and Newey (2018), Belloni, Chernozhukov and Kato (2019), Wang, He and Xu (2020), where regression coefficients are often labeled as treatment effects. Additionally, when the unconfoundedness assumption is met, our model can be adopted for estimating causal effects, an approach often termed regression adjustment in causal inference literature; see Imbens and Rubin (2015). Additional adjustments incorporating propensity scores can also be appealing to consider to avoid the model selection bias; see Belloni, Chernozhukov and Hansen (2014), Farrell (2015). One compelling aspect of high-dimensional confounder adjustment in causal inference is its ability to include a large number of potential confounders, which potentially mitigates the concern raised by unmeasured confounders. Later in this section, we show that our introduced scheme generalizes to a variety of linear models, such as models with cross-study heterogeneity.

We can think of the first column of X_k as a vector of all 1's to represent the intercept in the population model (1). In practice, we center all the data vectors, and simply assume that Y_k sums to zero, and so does every column of D_k and X_k without a loss of generality. Each existing study collects data for an extensive collection of covariates, and the treatment assignment is randomized conditional on these covariates. We further assume that the set of observed covariates, X_k, in each study include the ones indexed by E_0, that is, the covariates with nonzero effects in the population model are measured in each study. This is a rather strong assumption, requiring that the set of unobserved confounders is balanced in all the studies. Unobserved confounders in more general settings certainly deserve further attention.

Because our existing studies begin with a large number of covariates, we focus on the case where each study has used the LASSO to select a subset of potential confounders. Denote by E_k the active set of covariates selected by the LASSO in existing study k, and let $q_k = |E_k|$. We assume that $E_0 \subset E_k$, which holds with probability tending to one as $n_k \to \infty$ under appropriate conditions. Define

$$E = \cup_{k=1}^K E_k$$

with cardinality equal to q.

Besides the existing studies, we also consider a validation study $\mathcal{V} = \{D, X, Y\}$, which is independent of $\mathcal{D}_k$ for all $k = 1, \ldots, K$. The data in the validation study $\mathcal{V}$ contains n measurements for the same outcome $Y \in \mathbb{R}^{n\times 1}$ and treatment variables $D \in \mathbb{R}^{n\times s}$ as our existing studies. Furthermore, $\mathcal{V}$ contains matched measurements for the potential confounders in the subset E, that are now represented by $X = X_E \in \mathbb{R}^{n\times q}$. Clearly, the findings in our K already existing studies provide a guideline for what covariates need to be included as possible confounders in the follow-up study. We assume that the validation data is drawn from the same population model as (1), i.e.,

$$Y = D\alpha + X_{E_0}\beta_{E_0} + \varepsilon,$$

where ε is centered at 0_n and has covariance $\sigma^2 I$ given the treatment variable and covariates.

Fixing some more notation for the remaining paper, let $N_k = n + n_k$, and let $r_k = \frac{n_k}{N_k}$ in the rest of the paper. In each existing study, we use the LASSO to estimate

$$\widehat{\gamma}_k^{(\mathrm{L})} = \begin{pmatrix} \widehat{\alpha}_k^{(\mathrm{L})} \\ \widehat{\beta}_k^{(\mathrm{L})} \end{pmatrix} = \underset{\alpha\in\mathbb{R}^s, \beta\in\mathbb{R}^{p_k}}{\arg\min} \left\{ \frac{1}{2r_k\sqrt{N_k}} ||Y_k - D_k\alpha - X_k\beta||^2 + ||\Lambda_k\beta||_1 \right\}, \tag{2}$$

$\widehat{\alpha}_k^{(\mathrm{L})} \in \mathbb{R}^s$, $\widehat{\beta}_k^{(\mathrm{L})} \in \mathbb{R}^{q_k}$ for $k = 1, 2, \cdots, K$, where $\Lambda_k \in R^{p_k\times p_k}$ is a diagonal matrix with the tuning parameters in the LASSO penalty in the diagonal entries. Denote by

$$\widehat{\gamma}_k = \begin{pmatrix} \widehat{\alpha}_k \\ \widehat{\beta}_k \end{pmatrix} \tag{3}$$

the least squares estimator, based on $\begin{bmatrix} X'_{k,E_k} & X'_{E_k} \end{bmatrix}' \in \mathbb{R}^{N_k\times q_k}$, and $\begin{pmatrix} Y'_k & Y' \end{pmatrix}' \in \mathbb{R}^{N_k\times 1}$ after pooling observations from the validation study and existing study k.

Some extensions. The model in (1) assumes that β_{E_0}, the effect of the true confounders, is the same across all the studies. Now we consider a more general setup with cross-study heterogeneity, and show that we can still work with our basic model with the same form as (1). To see this, suppose that data in our existing studies are drawn from the model

$$Y_k = D_k\alpha + X_{k,E_{0,k}}\beta_{E_{0,k}} + \varepsilon_k, \tag{4}$$

and data in our validation study are generated from the model

$$Y_0 = D_0\alpha + X_{0,E_{0,0}}\beta_{E_{0,0}} + \varepsilon_0, \tag{5}$$

where $\mathbb{E}(\varepsilon_k|D_k, X_k) = 0_{n_k}$, $\mathrm{Var}(\varepsilon_k|D_k, X_k) = \sigma^2 I$, for $k = 0, \ldots, K$. Let $\delta_{i,k}$ be a dummy variable that assumes the value 1 if observation i is in study k, and is 0 otherwise. Note that $\sum_{k=0}^K \delta_{i,k} = 1$ for any observation i. Thus, for a variable $v_i \in \mathbb{R}^t$, let

$$\delta_{i,k}v_i = \begin{cases} v_i & \text{if observation } i \in \text{ study } k, \\ 0_t & \text{otherwise.} \end{cases}$$

Given the above generating models, we can describe observation i in our data, for the validation study as well as the K existing studies, through the unified model

$$y_i = D'_i\alpha + Z'_{i,E_0}\gamma_{E_0} + \varepsilon_i,$$

where

$$\gamma_{E_0} = \left(\beta'_{E_{0,0}} \quad \beta'_{E_{0,1}} \quad \cdots \quad \beta'_{E_{0,K}}\right)', \text{ and}$$

$$Z_{i,E_0} = \left[(\delta_{i,0}X_{i,E_{0,0}})' \quad (\delta_{i,1}X_{i,E_{0,1}})' \quad \cdots \quad (\delta_{i,K}X_{i,E_{0,K}})'\right]'.$$

Because we can write the unified model in the same form as the model in (1), we stick to our basic model for ease of presentation and revisit the extensions after introducing our proposal.

2.2. Debiasing through data carving: A first example

We provide a first example to illustrate how we utilize data carving to construct an unbiased estimator of treatment effects. Suppose that we have an existing study, i.e., $K = 1$. We let $\mathcal{D}_1$ contain a real valued covariate $X_1 \in \mathbb{R}$ and a treatment variable D_1, i.e., $p_1 = 1$ and $s = 1$. Both X_1 and D_1 in this example are standardized to have mean zero and variance one with correlation ρ. Consider observing n_1 and n observations in $\mathcal{D}_1$ and $\mathcal{V}$ respectively, from the population model (1) with $\beta_{E_0} = 0$, i.e. $E_0 = \varnothing$, and with noise variance $\sigma^2 = 1$. For simplicity, we let $n = n_1$, which means that $N_1 = 2n_1$, and $r_1 = \frac{1}{2}$ in this example.

Suppose that we select the full model $E_1 = \{1\}$ after solving the LASSO on $\mathcal{D}_1$. Then, we fit a Gaussian model with the selected covariate X_1, i.e., we model the real-valued outcome variable as independent realizations from the conditional distribution

$$\mathrm{y} \mid \mathrm{d}, \mathrm{x}_1 \sim N(\alpha \mathrm{d} + \beta \mathrm{x}_1, 1). \tag{6}$$

Note that this Gaussian model, which we fit to the pooled data $\mathcal{D}_1 \cup \mathcal{V}$, may only serve as a working model and may not exactly match with the population model in (1).

Define

$$\gamma_1 = \begin{pmatrix}\alpha & \beta\end{pmatrix}'.$$

Consistent with our notations, let $\widehat{\gamma}_1$ be the least squares estimator for γ_1 using the pooled observations in $\mathcal{D}_1$ and $\mathcal{V}$. Note that the likelihood in the selected model (6) is

$$\ell_{N_1}(\gamma_1; \widehat{\gamma}_1) \propto \exp\left(-\frac{N_1}{2}(\widehat{\gamma}_1 - \gamma_1)'\widehat{\Sigma}_1(\widehat{\gamma}_1 - \gamma_1)\right),$$

if we ignored the impact of model selection bias.

To construct our carved estimator, we introduce a new variable which we call the randomization variable. Denoted as

$$\omega_1 = \begin{pmatrix}\omega_{1,1} & \omega_{1,2}\end{pmatrix}' \in \mathbb{R}^2,$$

the randomization variable that takes the value

$$\frac{\partial}{\partial \gamma_1}\left(\frac{1}{2\sqrt{N_1}}\|Y_1 - \alpha D_1 - \beta X_1\|^2 + \frac{1}{2\sqrt{N_1}}\|Y - \alpha D - \beta X\|^2 - \frac{1}{2r_1\sqrt{N_1}}\|Y_1 - \alpha D_1 - \beta X_1\|^2\right)\Bigg|_{\widehat{\gamma}_1^{(L)}},$$

where $r_1 = \frac{1}{2}$. Observe that there is a simple equivalence between the Karush–Kuhn–Tucker (K.K.T.) mapping for (2) and

$$\underset{\alpha,\beta}{\text{minimize}} \ \frac{1}{2\sqrt{N_1}}\|Y - \alpha D - \beta X\|^2 + \frac{1}{2\sqrt{N_1}}\|Y_1 - \alpha D_1 - \beta X_1\|^2 - \omega_{1,1}\alpha - \omega_{1,2}\beta + \lambda|\beta|, \tag{7}$$

which is a randomized version of the LASSO on the pooled data $\mathcal{D}_1 \cup \mathcal{V}$. Under the selected model in (6) and ignoring the impact of model selection, we have as $N_1 \to \infty$:

1. $\sqrt{N_1}(\widehat{\gamma}_1 - \gamma_1)$ and ω_1 jointly follow an Gaussian distribution such that ω_1 is centered at 0 with the covariance $\mathbb{E}[\widehat{\Sigma}_1]$
2. ω_1 is independent of $\sqrt{N_1}(\widehat{\gamma}_1 - \gamma_1) \sim N_2(0_2, (\mathbb{E}[\widehat{\Sigma}_1])^{-1})$.

Suppose for now, the limiting distributional properties listed under 1 and 2 hold exactly with

$$\mathbb{E}[\widehat{\Sigma}_1] = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

Our carved estimator can be constructed in two main steps. We eliminate bias from model selection by conditioning on the selection event

$$\left\{\widehat{E}_1 = E_1, \operatorname{sign}(\widehat{\beta}_1^{(L)}) = s_{E_1}\right\}. \tag{8}$$

In Step 1, we recognize that

$$\widehat{\alpha}_1^{\text{initial}} = \widehat{\alpha}_1 - (1-\rho^2)^{-1}\frac{1}{\sqrt{N_1}} \cdot (\omega_{1,1} - \rho\omega_{1,2})$$

is an unbiased, initial estimator that is independent of the selection through LASSO. This leads us to observe that $\widehat{\alpha}_1^{\text{initial}}$ is also unbiased for α_1 conditional on the event in (8).

In Step 2, we improve upon the initial estimator through Rao-Blackwellization by conditioning further upon the complete sufficient statistic for γ_1. In our case, this statistic is $\widehat{\gamma}_1$, because of the basic fact that conditioning preserves the complete sufficient statistic, and that the sufficient statistic in the usual likelihood is the least squares estimator. Please see Fithian, Sun and Taylor (2014) for a formal proof of this fact. Thus, our carved estimator is given by

$$\widehat{\alpha}_1^{\text{carve}} = \mathbb{E}\left[\widehat{\alpha}_1^{\text{initial}} \,\middle|\, \widehat{\gamma}_1, \widehat{E}_1 = E_1, \operatorname{sign}(\widehat{\beta}_1^{(L)}) = s_{E_1}\right], \tag{9}$$

where we have conditioned upon $\widehat{\gamma}_1$ alongside the event of selection in (8). Proposition 1.1 in the Supplementary Material provides details of both steps and a final expression for the carved estimator. Because

$$\mathbb{E}\left[\widehat{\alpha}_1^{\text{carve}} \,\middle|\, \widehat{E}_1 = E_1, \operatorname{sign}(\widehat{\beta}_1^{(L)}) = s_{E_1}\right] = \alpha,$$

we also have $\mathbb{E}\left[\widehat{\alpha}_1^{\text{carve}}\right] = \alpha$.

A few remarks are in order. First, the estimator in (9) can be written as

$$\widehat{\alpha}_1^{\text{carve}} = \widehat{\alpha}_1 - (1-\rho^2)^{-1}\frac{1}{\sqrt{N_1}} \cdot \mathbb{E}\left[\omega_{1,1} - \rho\omega_{1,2} \,\middle|\, \widehat{\gamma}_1, \widehat{E}_1 = E_1, \operatorname{sign}(\widehat{\beta}_1^{(L)}) = s_{E_1}\right],$$

In particular, we note that $\widehat{\alpha}_1^{\text{carve}}$ applies an additive *debiasing correction* to the simple least squares estimator $\widehat{\alpha}_1$. The final expression for the carved estimator, as stated in Proposition 1.1, is obtained

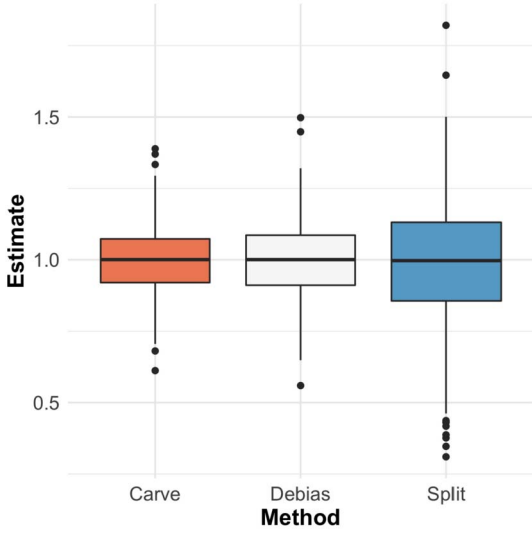

Figure 2. Box-plots of the parameter estimates in the simulation experiment for Section 2.

by evaluating the conditional expectation in the debiasing correction. Second, to construct the carved estimator, we used: (i) the least squares estimator in the selected model, (ii) the tuning parameter λ, and (iii) the sign of the LASSO estimator for the selected covariate s_{E_1}, which depend on only some simple summary statistics from the existing study.

Next we extend our main ideas from this simple example to the more general setup. Before proceeding, we provide some numerical evidence for the effectiveness of our new estimator. We depict in Figure 2 the result of a simulation for $n = 100$, $n_1 = 50$, and $1,000$ Monte Carlo samples. We compare our carved estimator against the two popular alternatives that were discussed earlier in the introduction, (1) the estimator based on data splitting, which simply uses the validation data $\mathcal{V}$ for estimation, and (2) the debiased LASSO estimator based on a one-step bias correction to the LASSO estimator. We elaborate on the generative scheme for the simulation as well as the implementation details for the other two alternatives in Section 4. Note that all the three estimators are centered around the true value $\alpha_1 = 1$. Especially, our new estimator removes the effect of selection bias, and has smallest variability among the three estimators.

3. Carved estimator

3.1. Privacy-preserving aggregation

We turn to our general setup, and outline our scheme to aggregate summary statistics from existing studies with data carving.

We begin by specifying the summary statistics involved in the construct of our estimator:

1. *Summary from model selection in existing studies*: the support set of the LASSO estimator E_k, the penalty weights Λ_{E_k}, and the signs of the (nonzero) LASSO estimator for the selected covariates $s_{E_k} = \text{sign}(\widehat{\beta}_k^{(\text{L})})$.

2. *Summary based on the first two moments from data in existing studies*: the sample covariance matrices from the existing study k, based on the selected model E_k and the observed sample of size n_k, which are equal to

$$\widehat{\xi}_k = \begin{bmatrix} \widehat{\xi}_{k,01} \\ \widehat{\xi}_{k,02} \end{bmatrix} = \frac{1}{n_k} \begin{bmatrix} D_k' Y_k \\ X_{k,E_k}' Y_k \end{bmatrix},$$

$$\widehat{\Xi}_k = \begin{bmatrix} \widehat{\Xi}_{k,11} & \widehat{\Xi}_{k,12} \\ \widehat{\Xi}_{k,21} & \widehat{\Xi}_{k,22} \end{bmatrix} = \frac{1}{n_k} \begin{bmatrix} D_k' D_k & D_k' X_{k,E_k} \\ X_{k,E_k}' D_k & X_{k,E_k}' X_{k,E_k} \end{bmatrix}.$$

Note that $\widehat{\xi}_k$ has two blocks, one s-dimensional and the other q_k-dimensional. In a similar fashion, $\widehat{\Xi}_k$ is partitioned along the same dimensional structure.

First, we compute the least squares estimator

$$\widehat{\gamma}_k = \begin{pmatrix} \widehat{\alpha}_k \\ \widehat{\beta}_k \end{pmatrix} = \begin{pmatrix} n_k \widehat{\Xi}_{k,11} + D'D & n_k \widehat{\Xi}_{k,12} + D' X_{E_k} \\ n_k \widehat{\Xi}_{k,12} + X_{E_k}' D & n_k \widehat{\Xi}_{k,22} + X_{E_k}' X_{E_k} \end{pmatrix}^{-1} \begin{pmatrix} n_k \widehat{\xi}_{k,01} + D'Y \\ n_k \widehat{\xi}_{k,02} + X_{E_k}' Y \end{pmatrix}. \tag{10}$$

For a fixed vector $u \in \mathbb{R}^k$ and for $v \in \mathbb{R}^k$, let $\log(1 + \frac{1}{uv})$ be a $\mathbb{R}^k$-valued vector whose j-th component equals

$$\log\left(1 + \frac{1}{u_j v_j}\right).$$

Denote by

$$\widehat{\Sigma}_k = \frac{1}{N_k} \left(n_k \widehat{\Xi}_k + \begin{pmatrix} D & X_{E_k} \end{pmatrix}' \begin{pmatrix} D & X_{E_k} \end{pmatrix} \right), \tag{11}$$

the sample covariance matrix based on the selected covariates. We solve $\widehat{z}_k = \begin{pmatrix} \widehat{z}_{k,1}' & \widehat{z}_{k,2}' \end{pmatrix}'$ from the following convex optimization problem:

$$\begin{aligned} \underset{z_k \in \mathbb{R}^{s+q_k}}{\operatorname{arginf}} \Bigg\{ & (1 - r_k)^{-1} r_k \frac{1}{2} \left(\sqrt{N_k} z_k - \sqrt{N_k} \widehat{\gamma}_k + \widehat{\Sigma}_k^{-1} \begin{pmatrix} 0' & (\Lambda_{E_k} s_{E_k})' \end{pmatrix} \right)' \widehat{\Sigma}_k \\ & \left(\sqrt{N_k} z_k - \sqrt{N_k} \widehat{\gamma}_k + \widehat{\Sigma}_k^{-1} \begin{pmatrix} 0' & (\Lambda_{E_k} s_{E_k})' \end{pmatrix} \right) + B_{s_{E_k}} \left(\sqrt{N_k} z_{k,2} \right) \Bigg\}, \end{aligned} \tag{12}$$

where

$$B_{s_{E_k}} \left(\sqrt{N_k} z_{k,2} \right) = 1_k' \log\left(1 + \frac{1}{s_{E_k} \sqrt{N_k} z_{k,2}} \right).$$

Our carved estimator pools data from the existing study k with the validation study, and is given by

$$\widehat{\alpha}_k^{\text{carve}} = \widehat{\alpha}_k + (1 - r_k)^{-1} r_k \left(\widehat{\alpha}_k - \frac{1}{\sqrt{N_k}} e_1' \widehat{\Sigma}_k^{-1} \begin{pmatrix} 0' & (\Lambda_{E_k} s_{E_k})' \end{pmatrix}' - e_1' \widehat{z}_k \right). \tag{13}$$

Finally, we propose the aggregated estimator by taking a simple average of the K carved estimators as

$$\widetilde{\alpha} = \frac{1}{K}\sum_{k=1}^{K} \widehat{\alpha}_k^{\text{carve}}. \tag{14}$$

We conclude the section with some observations about our estimator.

Remark 3.1. If the sample size for each existing study, n_k, varies with k, we can replace the simple average in (14) by a weighted average, where the weights are proportional to N_k.

Remark 3.2. Let

$$Q = \max_{k\in\{1,\ldots,K\}} |E_k|$$

denote the maximum size (cardinality) for the selected covariate sets across our existing studies. Then, we note that the amount of information required for constructing the estimator in (14) is only $O(Q^2)$.

Remark 3.3. We revisit the population models in (4) and (5) where we considered cross-study heterogeneity. We observe that our estimator easily adapts to the following two scenarios. First, say $E_{0,0} \subseteq \cap_{k=1}^{K} E_{0,k}$. In this case, we can proceed as prescribed above. Second, suppose that $E_{0,0} \subseteq \cup_{k=1}^{K} E_{0,k}$. In this case, we assume that the union of the selected variables $E = \cup_{k=1}^{K} E_k$ is collected in all the existing studies. We can now form our carved estimator by fitting the union model E to the pooled data from existing study k and the validation study.

3.2. Our debiasing approach

In this section, we show how the debiasing approach in the first example can be generalized to our problem. In line with the first example, our carved estimator $\widehat{\alpha}_k^{\text{carve}}$ is obtained by conditioning on the selection event

$$\left\{\widehat{E}_k = E_k,\ \text{sign}(\widehat{\beta}_k^{(L)}) = s_{E_k}\right\}, \tag{15}$$

which we observe after solving (2) on data $\mathcal{D}_k$. As before, we work with the selected model which assumes that data for the outcome variable are drawn as independent realizations from the conditional distribution

$$\mathrm{y} \mid \mathrm{d}, \mathrm{x}_{E_k} \sim N(\mathrm{d}'\alpha + \mathrm{x}'_{E_k}\beta_{E_k}, \sigma^2).$$

We proceed by fitting the selected model to $\mathcal{D}_k \cup \mathcal{V}$. Since we work under a fixed σ^2 setting, we will assume $\sigma^2 = 1$ for the sake of simplicity.

To develop the theory, we let

$$\begin{aligned}\omega_k &= \left((\omega_{k,1})' \quad (\omega_{k,2})'\right)';\ \omega_{k,1} \in \mathbb{R}^s, \omega_{k,2} \in \mathbb{R}^{p_k} \\ &= \frac{\partial}{\partial\gamma}\left(\frac{1}{2\sqrt{N_k}}\|Y_k - D_k\alpha - X_k\beta\|^2 + \frac{1}{2\sqrt{N_k}}\|Y - D\alpha - X\beta\|^2\right. \\ &\qquad \left. - \frac{1}{2r_k\sqrt{N_k}}\|Y_k - D_k\alpha - X_k\beta\|^2\right)\Bigg|_{\widehat{\gamma}_k^{(L)}},\end{aligned} \tag{16}$$

be our randomization variable, which is based on the LASSO solution $\widehat{\gamma}_k^{(L)}$. Let E_k^c denote indices of covariates that are not selected by the LASSO in existing study k. In addition, we will also consider the statistic

$$\widehat{\Gamma}_k = \frac{1}{N_k}\left\{X'_{k,E_k^c}(Y_k - D_k\widehat{\alpha}_k - X_{k,E_k}\widehat{\beta}_k) + X'_{E_k^c}(Y - D\widehat{\alpha}_k - X_{E_k}\widehat{\beta}_k)\right\}.$$

Note, we need not observe the covariates that were not selected by the LASSO in the validation study. The statistics defined above only serve to provide a theoretical justification for our debiasing term.

Lemma 3.4 notes that the variables $\widehat{\gamma}_k$, Γ_k and ω_k admit an asymptotic linear representation in the selected model, if we ignored the impact of model selection and treated E_k as a fixed set. Formally, this implies that these variables are distributed as Gaussian variables in the limit as the sample size $N_k \to \infty$.

Lemma 3.4. *Let* $\gamma_k = \begin{pmatrix}\alpha' & \beta'_{E_k}\end{pmatrix}'$ *for a fixed subset* E_k. *Then, the following assertion holds:*

$$\begin{pmatrix}\sqrt{N_k}(\widehat{\gamma}_k - \gamma_k)' & \sqrt{N_k}\widehat{\Gamma}'_k & \omega'_k\end{pmatrix}' = \sqrt{N_k}\bar{T}_{N_k} + R_{N_k},$$

where (i) $\bar{T}_{N_k}$ *is the average of* N_k *i.i.d. variables with mean equal to* $0_{2p_k+2s_k}$, *(ii)* $R_{N_k} = o_p(1)$, *(iii)* $\widehat{\gamma}_k$, $\widehat{\Gamma}_k$ *and* ω_k *are asymptotically independent, and (iv)* ω_k *is asymptotically centered at* 0 *with the covariance* $(1-r_k)r_k^{-1}\mathbb{E}[\mathfrak{G}_k]$, *where* $\mathfrak{G}_k = N_k{}^{-1}\left\{\begin{pmatrix}D & X\end{pmatrix}'\begin{pmatrix}D & X\end{pmatrix} + \begin{pmatrix}D_k & X_k\end{pmatrix}'\begin{pmatrix}D_k & X_k\end{pmatrix}\right\}$.

Suppose that we condition this limiting Gaussian distribution on the selection event in (15). This gives us an asymptotic conditional distribution. Say that we denote this distribution by $\widetilde{L}_{N_k}$. Next, Proposition 3.5 derives a debiasing correction from this asymptotic conditional distribution as was done in the first example. Before stating this result, we let

$$\mu_{\mathcal{H}_0}(\alpha_0, \Sigma_0)$$

denote the first moment of a Gaussian variable with mean α_0 and covariance Σ_0 that is truncated to the region $\mathcal{H}_0$. Further, let $\Delta(\eta_0)$ be the log-partition function for the related truncated Gaussian density at the natural parameter

$$\eta_0 = \Sigma_0^{-1}\alpha_0.$$

Finally, for $\widehat{\Sigma}_k$ defined in (11), let $\Sigma_k = \mathbb{E}[\widehat{\Sigma}_k]$ be the population covariance.

Proposition 3.5. *Define the half-space*

$$\mathcal{H} = \left\{z = (z_1, z_2),\ z_1 \in \mathbb{R}^s,\ z_2 \in \mathbb{R}^{q_k} : sign(z_2) = s_{E_k}\right\}.$$

Then, the estimator

$$\widehat{\alpha}_k + (1-r_k)^{-1}r_k\left(\widehat{\alpha}_k - \frac{1}{\sqrt{N_k}}\cdot e_1'\Sigma_k^{-1}\zeta_k - \frac{1}{\sqrt{N_k}}\cdot e_1'\mu_{\mathcal{H}}\left(\sqrt{N_k}\widehat{\gamma}_k - \Sigma_k^{-1}\zeta_k, r_k^{-1}(1-r_k)\Sigma_k^{-1}\right)\right)$$

is unbiased for α *with respect to* $\widetilde{L}_{N_k}$.

The estimator in Theorem 3.5, however, does not admit direct computations due to lack of readily available expressions for the moments of a truncated Gaussian variable. To this end, we use the

plug-in estimator $\widehat{\Sigma}_k$ and the inverse of $\widehat{\Sigma}_k$ for estimating Σ_k and its inverse, respectively, and apply a Laplace-type approximation to facilitate a manageable calculation for this estimator. Deferring a rigorous justification of asymptotic unbiasedness to the next discussion, we complete the steps to numerically approximate our estimator through an easy-to-solve optimization. For

$$\eta_0 = (1-r_k)^{-1} r_k \widehat{\Sigma}_k (\sqrt{N_k}\widehat{\gamma}_k - \widehat{\Sigma}_k^{-1}\zeta_k),$$

a Laplace approximation grants us the following

$$\begin{aligned}
\Delta(\eta_0) &\approx (1-r_k)^{-1} r_k \frac{1}{2} (\sqrt{N_k}\widehat{\gamma}_k - \widehat{\Sigma}_k^{-1}\zeta_k)' \widehat{\Sigma}_k (\sqrt{N_k}\widehat{\gamma}_k - \widehat{\Sigma}_k^{-1}\zeta_k) \\
&- \inf_{\text{sign}(z_{k,2}) = s_{E_k}} (1-r_k)^{-1} r_k \frac{1}{2} (\sqrt{N_k} z_k - \sqrt{N_k}\widehat{\gamma}_k + \widehat{\Sigma}_k^{-1}\zeta_k)' \widehat{\Sigma}_k (\sqrt{N_k} z_k - \sqrt{N_k}\widehat{\gamma}_k + \widehat{\Sigma}_k^{-1}\zeta_k) \\
&= \sup_{\text{sign}(z_{k,2}) = s_{E_k}} (1-r_k)^{-1} r_k (\sqrt{N_k} z_k)' \widehat{\Sigma}_k (\sqrt{N_k}\widehat{\gamma}_k - \widehat{\Sigma}_k^{-1}\zeta_k) \\
&\qquad - (1-r_k)^{-1} r_k \frac{1}{2} (\sqrt{N_k} z_k)' \widehat{\Sigma}_k \sqrt{N_k} z_k.
\end{aligned}$$

We replace the constrained optimization with the unconstrained version through a logarithmic barrier penalty to obtain

$$\begin{aligned}
\Delta(\eta_0) \approx \sup_{z_k} \; & (1-r_k)^{-1} r_k (\sqrt{N_k} z_k)' \widehat{\Sigma}_k (\sqrt{N_k}\widehat{\gamma}_k - \widehat{\Sigma}_k^{-1}\zeta_k) \\
& - (1-r_k)^{-1} r_k \frac{1}{2} (\sqrt{N_k} z_k)' \widehat{\Sigma}_k \sqrt{N_k} z_k - B_{s_{E_k}} (\sqrt{N_k} z_{k,2}).
\end{aligned}$$

Applying the above approximation to our debiasing correction in Proposition 3.5 allows us to write

$$\begin{aligned}
\mu_{\mathcal{H}} \left(\sqrt{N_k}\widehat{\gamma}_k - \widehat{\Sigma}_k^{-1}\zeta_k, r_k^{-1}(1-r_k)\widehat{\Sigma}_k^{-1} \right) &= \nabla \Delta ((1-r_k)^{-1} r_k \widehat{\Sigma}_k (\sqrt{N_k}\widehat{\gamma}_k - \widehat{\Sigma}_k^{-1}\zeta_k)) \\
&\approx \sqrt{N_k}\widehat{z}_k,
\end{aligned}$$

where $\widehat{z}_k$ is the solution in (12). This gives us the expression of our carved estimator in (13).

3.3. Asymptotic properties

Our main result in the section, Theorem 3.10, establishes the asymptotic unbiasedness of our carved estimator.

Suppose that we have N i.i.d. realizations of the response $\mathrm{y} \in \mathbb{R}$, the covariates for all existing studies $\mathrm{x} \in \mathbb{R}^{p_1 + \cdots + p_k}$ and the treatment $\mathrm{d} \in \mathbb{R}^s$, such that y is drawn from the model in (1) with

$$\left(\alpha' \quad \beta_{E_0}' \right)' = \left((\alpha^{(N)})' \quad (\beta_{E_0}^{(N)})' \right)'.$$

Each of our existing studies, in particular, observes N_k samples for the outcome and treatment along with a subset of covariates that are seen across all the studies. For the remaining section, we consider the following parameters

$$\sqrt{N} \left((\alpha^{(N)})' \quad (\beta_{E_0}^{(N)})' \right)' = a_N \left((\alpha_0)' \quad (\beta_{0,E_0})' \right) \tag{17}$$

in our generation scheme, such that α_0 and β_{0,E_0} are constants and $a_N = o(\sqrt{N})$. We state the regularity and moment conditions on the data-generating mechanism below.

To start with, we impose asymptotic screening as a requirement on the model selection in the existing studies. This requirement is necessary to ensure that the bias from missing a true covariate is only negligible with a growing sample size. The LASSO is an important example that satisfies the asymptotic screening requirement under conditions discussed at length by Meinshausen and Bühlmann (2006), Bickel, Ritov and Tsybakov (2009). Similar conditions are required by estimators which are aggregated over selected models, see for example the work by Schultheiss, Renaux and Bühlmann (2021), Fei and Li (2021).

Condition 3.6. *Assume that* $\lim_{n_k \to \infty} \mathbb{P}(E_0 \subseteq E_k) \to 1$, *for* $k = 1, 2, \ldots, K$.

The second and third conditions below allow us to control the bias from the use of a Laplace-type approximation for our debiasing correction.

Note that the first part of Condition 3.7 assumes the existence of an exponential moment in a neighborhood of zero. This is required to justify a Laplace-type approximation for probabilities involving the mean of N i.i.d. variables, which in our case is $\bar{T}_{N_k}$ by setting $N = N_k$ (please refer to Lemma 3.4). The second part of this condition is required to extend this approximation to asymptotically linear variables with an added $o_p(1)$ error term, which we see in the left-hand side of the representation in Lemma 3.4.

Next, we discuss the necessity of Condition 3.8. The probability of selection can be written approximately as the probability of a polyhedron:

$$\mathbb{P}\left[\widehat{E}_k = E_k,\ \operatorname{sign}(\widehat{\beta}_k^{(L)}) = s_{E_k}\right] = \mathbb{P}\left[A\left(\sqrt{N_k}\widehat{\gamma}_k' \quad \sqrt{N_k}\widehat{\Gamma}_k' \quad \omega_k'\right)' + o_p(1) \leq b\right],$$

where A and b are fixed matrices. We refer interested readers to Proposition 4.2 of Panigrahi, Taylor and Weinstein (2021) for this approximate characterization of the selection event. The condition in 3.8 allows us to replace the log-probability of selection with the log-probability of the polyhedron that is determined by matrices A and b. In other words, this condition allows us to ignore the $o_p(1)$ remainder in the characterization for our selection event.

Condition 3.7. *Fix*

$$\mathrm{V} = (\mathrm{y} - \alpha'\mathrm{d} - \beta_{E_0}'\mathrm{x}_{E_0}) \cdot \left(\mathrm{d}' \quad \mathrm{x}'\right)',$$

for a fixed subcollection of our covariates containing E_0. *We assume that*

$$\mathbb{E}\left[\exp\left(\eta\|\mathrm{V}\|\right)\right] < \infty$$

for some $\eta \in \mathbb{R}^+$. *For the observed set of covariates* E_k *in the existing study* k, *consider the linearizable representation in Lemma 3.4. We assume that when* $a_{N_k} \to \infty$, *the remainder term* R_{N_k} *satisfies:*

$$\lim_{N_k \to \infty} \frac{1}{a_{N_k}^2} \cdot \log \mathbb{P}\left[a_{N_k}^{-1}\|R_{N_k}\| > \epsilon\right] = -\infty \qquad \textit{for every } \epsilon > 0.$$

Condition 3.8. *We assume that when* $a_{N_k} \to \infty$, *the probability of selection satisfies*

$$\lim \frac{1}{a_{N_k}^2}\left\{\log \mathbb{P}\left[\widehat{E}_k = E_k,\ sign(\widehat{\beta}_k^{(L)}) = s_{E_k}\right] - \log \mathbb{P}\left[A\left(\sqrt{N_k}\widehat{\gamma}_k' \quad \sqrt{N_k}\widehat{\Gamma}_k' \quad \omega_k'\right)' \leq b\right]\right\} = 0.$$

A few comments are in order now. We note that both conditions for our theory require a particular parameterization with (17). However, it is worth pointing out that we can construct our estimator without having to access the sequence a_N for $N = N_k$. In practical terms, this means that we do not need to know the underlying parameterization of the data generating process for estimation. We also remark that the construct of our estimator can be applied to a wider range of penalties that exhibit a similar approximate polyhedral geometry with model selection, using the relevant matrices A and b. Our concluding remarks include a discussion on this topic.

Our final condition bounds the deviation of the sample estimate $\widehat{\Sigma}_k$ from the population parameter $\mathbb{E}[\widehat{\Sigma}_k]$.

Condition 3.9. *For the selected set E_k, we assume that $\mathbb{E}[\|\widehat{\Sigma}_k - \Sigma_k\|_{op}] = O(1)$, where $\|M\|_{op}$ denotes the operator norm of the matrix M.*

For the remaining section, consider the parameters in (17) by fixing $N = N_k$.

Theorem 3.10. *Let $\widehat{\alpha}_k^{carve}$ be defined according to (13). Under the conditions in 3.6, 3.7, 3.8 and 3.9, we have*

$$\mathbb{E}\left[\|\sqrt{N_k}(\widehat{\alpha}_k^{carve} - \alpha)\|^2\right] \leq a_{N_k}^2 C,$$

where C is a constant.

The asymptotic variance of our estimator $\sqrt{N_k}\widehat{\alpha}_k^{\text{carve}}$, based on the observed Fisher information matrix, is

$$\widehat{V}_{E_k}^{\text{carve}} = e_1'\left((1-r_k)^{-1}\widehat{\Sigma}_k^{-1} - (1-r_k)^{-2}r_k^2\left((1-r_k)^{-1}r_k\widehat{\Sigma}_k + \nabla^2 B_{s_{E_k}}(\sqrt{N_k}\widehat{z}_k)\right)^{-1}\right)e_1.$$

The expression for the estimated variance is detailed out in Proposition A.1, under Section 1. Lemma 3.11 then quantifies the relative gain in variance of our estimator over splitting for each study k. Here, $\widehat{V}_{E_k}^{\text{split}}$ denotes the estimated variance of the least squares estimate when refitted on the validation data alone. Let $\widehat{V}_{j,E_k}^{\text{carve}}$ and $\widehat{V}_{j,E_k}^{\text{split}}$ be the j-th diagonal entry of the corresponding s-dimensional matrices, respectively.

Lemma 3.11. *Let r_k be the ratio $\frac{n_k}{N_k}$, and let B_{max} be the maximum value of the $(\mathbb{R}^+)^{s+q_k}$-valued vector*

$$(s_{E_k}\sqrt{N_k}\widehat{z}_k)^{-2} - (1 + s_{E_k}\sqrt{N_k}\widehat{z}_k)^{-2},$$

and let λ_{min} be the smallest eigen value of $\widehat{\Sigma}_k$. Then, the following holds for $j \in \{1, \ldots, K\}$:

$$\left(\widehat{V}_{j,E_k}^{split}\right)^{-1}\left(\widehat{V}_{j,E_k}^{split} - \widehat{V}_{j,E_k}^{carve}\right) \geq (1-r_k)^{-1}r_k^2\left((1-r_k)^{-1}r_k + B_{max}\lambda_{min}^{-1}\right)^{-1}.$$

Clearly, our carved estimator $\widehat{\alpha}_k^{\text{carve}}$ dominates the split estimators in variance. The variance of the averaged carved estimator, however, involves correlations between the refitted estimators from the selected models on any pair of existing studies in addition to the variances from each study. Heuristically, the correlation between a pair of split estimators, using the validation data alone, is expected to be larger than the carved counterpart. This is because the carved estimator also uses information from the

independent samples in this pair of existing studies. Consider, for instance, the situation when each study reports the same model

$$E_1 = E_2 = \cdots = E_k.$$

Let the bound on the right-hand side of Lemma 3.11 be equal to B_k. Comparing the averaged carved estimator in (14) with the split estimator, the difference in the two variances is seen to be bounded below by

$$\frac{1}{K^2}\left(\sum_{k=1}^{K}\frac{B_k}{1-B_k}\widehat{V}_{E_k}^{\text{carve}} + 2\sum_{k_1<k_2}\left(\frac{1}{\sqrt{(1-B_{k_1})(1-B_{k_2})}}-1\right)\sqrt{\widehat{V}_{E_{k_1}}^{\text{carve}}}\sqrt{\widehat{V}_{E_{k_2}}^{\text{carve}}}\right).$$

Note, this bound is strictly greater than 0 since $B_k < r_k < 1$ for $k \in 1, \ldots, K$.

Our simulation studies in the next section empirically support the gain in efficiency that we claim for our carved estimator.

Remark 3.12. In this section, we introduced a new way to remove bias from the refitted least squares estimator in a selected model and aggregate summary statistics from existing studies for unbiased estimation of treatment effects. More generally, our new estimator extends to the class of generalized linear models (GLMs). In Supplementary Material 2, we develop an extension of our proposal to the widely used logistic regression model, which is a concrete instance in the class of GLMs.

4. Simulation studies

We conduct simulation studies to evaluate the finite-sample performance of our new estimator. The outcome variables, in the existing and validation studies, are generated through simple linear models as follows

$$\begin{aligned} Y_k &= \alpha D_k + X_{k,E_0}\beta_{E_0} + \varepsilon_k, \quad \varepsilon_k \sim N(0, \sigma_\varepsilon^2 I_{n_k}), \quad \text{for } k = 1, 2, \ldots, K, \\ Y &= \alpha D + X_{E_0}\beta_{E_0} + \varepsilon, \quad \varepsilon \sim N(0, \sigma_\varepsilon^2 I_n). \end{aligned} \tag{18}$$

We set the coefficients $\alpha = 1$ and $\beta_{E_0} = (1.5, 1, 1, 1, 1)'$, i.e., for each study k, X_{k,E_0} is the first 4 columns of X_k. The dimension of the covariates varies with k as $p_k = 400 + 20k$ and $p = 500$. Our sample sizes for the existing studies are $n_1 = \ldots = n_K = 100$, where $K \in \{2, 3, 5, 10\}$; the validation study has $n = 50$ samples. Our data is generated under two signal-to-noise-ratio values by using two values of $\sigma_\varepsilon \in \{2, 4\}$.

We summarize the performance of our carved estimator in three different settings.

1. Setting I. In the first setting, there is a moderate degree of correlation between the treatment assignment and the covariates. We generate (D_k, X_k) and (D, X) from a multivariate normal distribution $N(0, \Sigma)$ where $\Sigma = \left(\Sigma_{jl}\right)_{j,l=1}^{p_k+1}$ and $\Sigma_{jl} = 0.5^{|j-l|}$.
2. Setting II. The second setting generates highly correlated treatment and covariates by first generating

$$D = X'_{M_0}\gamma_{M_0} + \nu, \ \nu \sim N(0, \sigma_\nu^2 I_n), \quad D_k = X'_{k,M_0}\gamma_{M_0} + \nu_k, \ \nu_k \sim N(0, \sigma_\nu^2 I_{n_k}), \tag{19}$$

followed by drawing the response according to (18). In the model (19) for generating the treatment variable, we fix $M_0 = \{5, 6\}$, $\gamma_{M_0} = (1, 1)'$, and fix the noise variance $\sigma_\nu^2 = 0.1$.

3. Setting III. In the final setting, we vary the coefficients of our covariates for generating the data in the existing and validation studies. We simulate the shift in covariates across the existing and validation data as follows. We draw our data according to the generating scheme described for Setting II, except now we set the value of covariate coefficients $\beta_{E_0} = (1.5, 1, 1, 1, 0)'$ in model (18) to draw the response in our validation study. That is, the coefficient of one of the confounders has changed between the existing studies and the validation study in the follow-up stage.

We compare the performance of our carved estimator with the split-and-aggregate estimate and the aggregated debiased LASSO estimator. As remarked early on, an off-the-shelf option available for estimation is a simple split-and-aggregate estimator. This estimator is obtained by refitting the selected model E_k on the validation data $\mathcal{V}$ and then averaging the resulting K split estimators. Another legitimate estimator is the aggregated debiased LASSO estimator that can be viewed as the proposal in Cai, Liu and Xia (2022) when customized to our problem setup. To be specific, the aggregated estimator is formed by averaging the $K + 1$ debiased LASSO estimators from the existing studies $\mathcal{D}_k$ and the validation dataset $\mathcal{V}$. Given that the debiased LASSO and the post-double selection estimates are asymptotically equivalent (see Wang, He and Xu (2020) for example), we implement the debiased LASSO estimator via the `R` package `hdm` for its computational speed. We found that in a small number of cases, the package `hdm` produced numerically unstable results, and therefore, we report only the median bias and the median squared errors, instead of the averages that would be heavily against the debiased LASSO estimator due to a few unstable cases. We report both the mean and median bias and squared errors for our carved estimator and the split-and-aggregate estimator. The comparison of our estimator with debiased LASSO should be made with respect to the corresponding medians, even though we provided mean for the carved and the split estimators for a relative comparison between these two estimators.

The cells in Tables 1, 2, and 3 summarize our findings in the three different settings. In all the tables, our proposed estimator is indicated as "Carved", while the split-and-aggregate estimator is called "Split" and the aggregated debiased LASSO estimator is called "Debiased". Based on Table 1, we observe that when the noise level is low and the treatment variable is not highly correlated with the covariates, the selected model includes all the confounding variables with a high probability, and therefore, both the carved estimator and the split-and-aggregate estimator have little bias in estimating α. When the noise level is higher, the two estimators tend to have slightly more bias than the debiased LASSO estimator. But, the bias of the estimators is still dominated by their variance in the mean squared errors. On the other hand, the carved estimator has smaller mean-squared error than the split estimator in all the considered cases, which is expected since the carved estimator integrates information from the existing studies with higher efficiency. We note that the aggregated debiased LASSO estimator loses

		Bias					MSE				
σ_ε	K	Carved		Split		Debiased	Carved		Split		Debiased
4	3	0.107	0.095	0.119	0.103	−0.051	0.109	0.042	0.248	0.103	0.044
4	5	0.117	0.122	0.123	0.109	−0.028	0.065	0.032	0.157	0.077	0.024
4	10	0.127	0.125	0.124	0.124	−0.046	0.043	0.021	0.083	0.043	0.014
2	3	−0.003	0.005	0.010	−0.001	−0.076	0.017	0.006	0.042	0.017	0.013
2	5	0.002	−0.002	0.012	0.009	−0.061	0.010	0.004	0.024	0.009	0.008
2	10	0.004	0.003	0.010	0.011	−0.051	0.004	0.002	0.012	0.006	0.005

Table 1. Simulation results under Setting I. The two columns under the Carved and Split estimators report the mean and median bias and squared errors respectively. The cells in the single column under Debiased report the median bias and squared errors, due to reasons indicated in the description.

		Bias					MSE				
σ_ε	K	Carved		Split		Debiased	Carved		Split		Debiased
4	3	0.066	0.065	0.088	0.089	−0.081	0.024	0.011	0.056	0.026	0.059
4	5	0.066	0.072	0.069	0.063	−0.083	0.016	0.007	0.033	0.013	0.045
4	10	0.070	0.072	0.079	0.079	−0.079	0.011	0.006	0.022	0.009	0.016
2	3	0.016	0.012	0.026	0.023	−0.089	0.005	0.002	0.012	0.005	0.020
2	5	0.016	0.018	0.019	0.017	−0.075	0.005	0.001	0.007	0.003	0.013
2	10	0.016	0.015	0.020	0.021	−0.061	0.002	0.001	0.004	0.002	0.006

Table 2. Simulation results under Setting II. The two columns under the Carved and Split estimators report the mean and median bias and squared errors respectively. The cells in the single column under Debiased report the median bias and squared errors.

much efficiency relative to the aggregated carved estimator when the treatment variable is highly correlated with some of the covariates, which is shown in Tables 2 and 3. The empirical comparisons here are in line with what we expect from our theoretical investigation.

5. Synthetic study on clinical trial data

We demonstrate the efficacy of our aggregation and estimation scheme on synthetic data generated from the Afya II clinical trial. Note that our paper proposes a new design to collect data for conducting validation studies and pooling data from already existing studies to estimate treatment effects. Therefore, collecting data that aligns with our recommendation for new study designs and verifying the effectiveness of our proposal is beyond the scope of this paper. Instead, we evaluate the efficiency gains of our estimator through a "realistic" simulation study that very closely resembles the Afya II clinical trial data. In our synthetic case study, we assume the following scenario. Two previous clinical trial studies have already been conducted to assess the effectiveness of an intervention, and we seek to design a new trial to confirm the results of the earlier clinical trials.

The main goal of the Ayfa II clinical trial was to test the effectiveness of a cash incentive on enhancing HIV patient adherence to antiretroviral (ARV) therapy and on achieving HIV viral suppression. The actual trial enrolled a total of 534 HIV-positive patients, who were randomly assigned to one of three arms: a control arm, a treatment arm with a cash incentive amount of approximately 5 US dollars, and a treatment arm with a cash incentive amount of approximately 11 US dollars. The primary outcome was determined by measuring the HIV viral load in patient plasma specimens 12 months after the treatment

		Bias					MSE				
σ_ε	K	Carved		Split		Debiased	Carved		Split		Debiased
4	3	0.102	0.103	−0.179	−0.164	−0.096	0.053	0.020	0.084	0.036	0.067
4	5	0.110	0.112	−0.170	−0.167	−0.091	0.042	0.015	0.067	0.033	0.042
4	10	0.092	0.100	−0.163	−0.162	−0.092	0.023	0.011	0.042	0.027	0.026
2	3	0.028	0.024	−0.156	−0.151	−0.079	0.008	0.003	0.042	0.024	0.018
2	5	0.024	0.034	−0.156	−0.159	−0.075	0.016	0.002	0.034	0.025	0.013
2	10	0.029	0.029	−0.152	−0.151	−0.067	0.003	0.001	0.038	0.023	0.008

Table 3. Simulation results under Setting III. The two columns under the Carved and Split estimators report the mean and median bias and squared errors respectively. The cells in the single column under Debiased report the median bias and squared errors.

Bias			MSE		
Carved	Split	Debiased	Carved	Split	Debiased
0.001	−0.011	−0.062	0.058	0.154	0.068

Table 4. Results on 500 synthetic datasets modeled along the Afya II clinical trial.

assignment. During this study, the trial collected patient status information, including age, sex, marital status (single, married/cohabitating, divorced/separated, widowed, or unknown), weight (measured in kilograms), WHO clinical stage (ranging from 1 to 4), pregnancy status, nutrition status (normal, obesity, or malnourished), family planning ID (with a total of 5 possible statuses), ARV status (start ARV, continue, change, or stop), TB screening, TB status, and initial CD4 count. Sociodemographic survey data was collected at the end of the study. Overall, the trial data measured 92 covariates for the 534 enrolled patients.

In our case study, we focus on the treatment arm where patients received a cash incentive of 11 US dollars. This reduces the sample size to 360 patients. First, we use the LASSO to regress the outcome (viral load) against the treatment indicator variable and the collected covariates on the trial data. The LASSO selects a total of 12 covariates. We record the estimated sparse coefficients as $\widehat{\beta}_{\texttt{Lasso}}$. Using covariate measurements from the actual trial, we generate the viral load outcome from a linear model

$$Y = 10 \cdot D + X\widehat{\beta}_{\texttt{Lasso}} + \varepsilon,$$

where $\varepsilon \in \mathbb{R}^{360}$ is a multivariate Gaussian random variable with mean zero and identity covariance matrix. Finally, we randomly divide the synthetic dataset into three folds, treating two of them as data collected from existing studies and the remaining one as data from our validation study. We repeat this entire process 500 times by generating data from the above linear model. This gives us 500 synthetic datasets that are closely modeled along the actual Afya II clinical trial. During estimation, we use only the selected covariates from any of the two existing studies as part of our validation data.

In line with the preceding section, Table 4 reports the comparison between our estimator, and the split-and-aggregate estimator and the aggregated debiased LASSO estimator. As done before, we compare their bias and mean squared errors. We see that the carved estimator, based on aggregating only summary statistics from the two existing studies, achieves the lowest bias and smallest mean squared error. More specifically, the improvement in mean squared error over the debiased LASSO estimator is roughly around 15%, and is a significant 60% over the split-and-aggregate estimator. This case study on clinical trial data re-emphasizes that our aggregation scheme can be used to design validation studies from a synthesis of existing studies, and that we can estimate the common treatment effects without sacrificing statistical efficiency.

6. Concluding remarks

In this paper, we develop a new scheme for estimating common treatment effects from a synthesis of prior studies. Our scheme applies ideas from data carving to: (i) use already existing data for an unbiased estimation of treatment effects in selected models, and (ii) aggregate summary statistics from existing studies when the LASSO is used in each individual study to select potential confounding variables. As a result, we have laid out a data aggregation and validation analysis protocol that gives us efficient unbiased estimators while preserving the privacy of individual records. The summary statistics required by our estimator include the first two moments in each selected model along with some compressed information from the model selection. Our estimator, unlike the debiased LASSO estimator, does not require us to store, report, or estimate the inverse of a high-dimensional matrix. As shown

by our simulation studies, the introduced estimator can not just mitigate bias from model selection, but also yield higher statistical efficiency than popular alternatives including off-the-shelf alternatives such as splitting.

In our paper, we focus on the LASSO or the ℓ_1-penalty. However, the proposed debiasing correction's form, and consequently the estimator, can be expanded to a larger class of penalties that result in an approximate polyhedral geometry with model selection. For this class of penalties, our primary construct relies on the asymptotic linear representation in the selected model, which we presented in Lemma 3.4. The only difference would be in the summary statistics required for aggregation, which might vary depending on the specific penalty used and might be different from those needed for model selection with the LASSO. Therefore, we believe that our approach will find applications beyond the LASSO penalty for model selection.

Interval estimation, which requires us to characterize the distribution of the aggregated estimator, is a challenging question and a natural next goal. We hope to address the problem of interval estimation with our proposed scheme in future work.

Appendix

We provide a proof of our main theoretical result in this section.

We introduce some important convex functions that we will need to prove Theorem 3.10. Let

$$D_{N_k}(\widetilde{\gamma}) = \frac{1}{2}\widetilde{\gamma}'\widehat{\Sigma}_k\widetilde{\gamma} + \inf_{\widetilde{z}}\Big\{\frac{1}{2}(1-r_k)^{-1}r_k(\widetilde{z}-\widetilde{\gamma}+\frac{1}{a_{N_k}}\widehat{\Sigma}_k^{-1}\zeta_k)'\widehat{\Sigma}_k(\widetilde{z}-\widetilde{\gamma}+\frac{1}{a_{N_k}}\widehat{\Sigma}_k^{-1}\zeta_k) + \frac{1}{(a_{N_k})^2}B_{s_{E_k}}(a_{N_k}\widetilde{z})\Big\},$$

and let

$$\widetilde{\mathcal{P}}_{N_k}(\eta_0) = \sup_{\widetilde{\gamma}}\widetilde{\gamma}'\eta_0 - D_{N_k}(\widetilde{\gamma}), \tag{20}$$

which is the convex conjugate for D_{N_k}. Similarly, let

$$\check{D}_{N_k}(\widetilde{\gamma}) = \frac{1}{2}\widetilde{\gamma}'\Sigma_k\widetilde{\gamma} + \inf_{\widetilde{z}}\Big\{\frac{1}{2}(1-r_k)^{-1}r_k(\widetilde{z}-\widetilde{\gamma}+\frac{1}{a_{N_k}}\Sigma_k^{-1}\zeta_k)'\Sigma_k(\widetilde{z}-\widetilde{\gamma}+\frac{1}{a_{N_k}}\Sigma_k^{-1}\zeta_k) + \frac{1}{(a_{N_k})^2}B_{s_{E_k}}(a_{N_k}\widetilde{z})\Big\},$$

and let $\check{\mathcal{P}}_{N_k}$ be the corresponding convex conjugate.

Proof of Theorem 3.10. First, we note that our carved estimator is equal to

$$\frac{1}{a_{N_k}}\sqrt{N_k}\widehat{\alpha}_k^{\text{carve}} = e_1'\widehat{\Sigma}_k^{-1}\nabla D_{N_k}\left(\frac{1}{a_{N_k}}\sqrt{N_k}\widehat{\gamma}_k\right), \tag{21}$$

where ∇D_{N_k} is the gradient of D_{N_k}. This is because

$$\nabla D_{N_k}(\widetilde{\gamma}) = \widehat{\Sigma}_k\widetilde{\gamma} - (1-r_k)^{-1}r_k\widehat{\Sigma}_k\left(z^*(\widetilde{\gamma}) - \widetilde{\gamma} + \frac{1}{a_{N_k}}\widehat{\Sigma}_k^{-1}\zeta_k\right),$$

where

$$z^*(\widetilde{\gamma}) = \operatorname{arginf}_z \Big\{ \frac{1}{2}(1-r_k)^{-1} r_k \Big(\widetilde{z} - \widetilde{\gamma} + \frac{1}{a_{N_k}} \widehat{\Sigma}_k^{-1} \zeta_k \Big)' \widehat{\Sigma}_k \Big(\widetilde{z} - \widetilde{\gamma} + \frac{1}{a_{N_k}} \widehat{\Sigma}_k^{-1} \zeta_k \Big) + \frac{1}{(a_{N_k})^2} B_{s_{E_k}}(a_{N_k} \widetilde{z}) \Big\},$$

and $z^*\left(\frac{1}{a_{N_k}} \sqrt{N_k} \widehat{\gamma}_k\right) = \frac{1}{a_{N_k}} \sqrt{N_k} \widehat{z}_k$, where $\widehat{z}_k$ is defined in (12).

Also, define

$$\widetilde{\mathcal{P}}_{0,N_k}(\eta_0) = \frac{1}{(a_{N_k})^2} \log \mathbb{P}\left[\widehat{E}_k = E_k,\ \operatorname{sign}(\widehat{\beta}_k^{(L)}) = s_{E_k} \right] + \frac{1}{2} \eta_0' \Sigma_k^{-1} \eta_0,$$

where $\eta_0 = \Sigma_k \gamma_0$. Let $\lambda_{\max}$ and $\lambda_{\min}$ be the largest and smallest eigen-value of $\widehat{\Sigma}_k$. Now, we have

$$\begin{aligned}
&\mathbb{E}\left[\Big\| \frac{1}{a_{N_k}} \sqrt{N_k} (\widehat{\alpha}_k^{\text{carve}} - \alpha) \Big\|^2 \Big| \widehat{E}_k = E_k, \operatorname{sign}(\widehat{\beta}_k^{(L)}) = s_{E_k} \right] \\
&\leq \frac{1}{\lambda_{\min}^2(\widehat{\Sigma}_k)} \mathbb{E}\left[\Big\| \nabla \widetilde{\mathcal{P}}_{N_k}^{-1} \Big(\frac{1}{a_{N_k}} \sqrt{N_k} \widehat{\gamma}_k \Big) - \widehat{\Sigma}_k \gamma_0 \Big\|^2 \Big| \widehat{E}_k = E_k, \operatorname{sign}(\widehat{\beta}_k^{(L)}) = s_{E_k} \right] \\
&\leq \frac{C^2}{\lambda_{\min}^2(\widehat{\Sigma}_k)} \mathbb{E}\left[\Big\| \frac{1}{a_{N_k}} \sqrt{N_k} \widehat{\gamma}_k - \nabla \widetilde{\mathcal{P}}_{N_k}(\widehat{\Sigma}_k \gamma_0) \Big\|^2 \Big| \widehat{E}_k = E_k, \operatorname{sign}(\widehat{\beta}_k^{(L)}) = s_{E_k} \right] \\
&\leq \frac{C^2}{\lambda_{\min}^2(\widehat{\Sigma}_k)} \Bigg\{ \mathbb{E}\left[\Big\| \frac{1}{a_{N_k}} \sqrt{N_k} \widehat{\gamma}_k - \nabla \widetilde{\mathcal{P}}_{0,N_k}(\Sigma_k \gamma_0) \Big\|^2 \Big| \widehat{E}_k = E_k, \operatorname{sign}(\widehat{\beta}_k^{(L)}) = s_{E_k} \right] \\
&\qquad + \mathbb{E}\left[\Big\| \nabla \breve{\mathcal{P}}_{N_k}(\Sigma_k \gamma_0) - \nabla \widetilde{\mathcal{P}}_{N_k}(\widehat{\Sigma}_k \gamma_0) \Big\|^2 \Big| \widehat{E}_k = E_k, \operatorname{sign}(\widehat{\beta}_k^{(L)}) = s_{E_k} \right] \\
&\qquad + \Big\| \nabla \breve{\mathcal{P}}_{N_k}(\Sigma_k \gamma_0) - \nabla \widetilde{\mathcal{P}}_{0,N_k}(\Sigma_k \gamma_0) \Big\|^2 \Bigg\}.
\end{aligned}$$

The second display uses (20) and (21) to deduce:

$$e_1' \widehat{\Sigma}_k^{-1} \nabla \widetilde{\mathcal{P}}_{N_k}^{-1} \Big(\frac{1}{a_{N_k}} \sqrt{N_k} \widehat{\gamma}_k \Big) = e_1' \widehat{\Sigma}_k^{-1} \nabla D_{N_k} \Big(\frac{1}{a_{N_k}} \sqrt{N_k} \widehat{\gamma}_k \Big) = \frac{1}{a_{N_k}} \sqrt{N_k} \widehat{\alpha}_k^{\text{carve}}.$$

The third display uses the fact that $\nabla \widetilde{\mathcal{P}}_{N_k}^{-1}$ is Lipschitz with constant C where $C = (1-r_k)^{-1} \lambda_{\max}(\widehat{\Sigma}_k)$. The last display follows after using the standard triangle inequality for decomposing the preceding expectation. For the decomposition shown in the last display, it is easy note that the expectation in the first term is $O(1)$. Observe that $\nabla \widetilde{\mathcal{P}}_{N_k}(\cdot)$, in the second display, is Lipschitz, because it is the gradient of the conjugate of a strongly convex function. Using the fact that $\widehat{\Sigma}_k - \Sigma_k = o_p(1)$ in our i.i.d. framework, jointly with the condition in Assumption 3.9, we deduce that the expected value of the second term is $O(1)$. By utilizing the convexity of $\widetilde{\mathcal{P}}_{0,N_k}$ and the pointwise convergence demonstrated in Proposition 1.3 (in the Supplementary Material), we observe that the third term converges to 0 as $a_{N_k} \to \infty$, and otherwise it is $O(1)$. This completes our proof. □

We turn to deriving a feasible value for the observed Fisher information matrix in order to estimate the variance of the carved estimator. To this end, the limiting value for the probability of selection in Proposition 1.3 (included in the Supplementary Material) gives rise to an approximate log-partition function in terms of $\eta_0 = \Sigma_k \gamma_0$. Note that η_0 is the natural parameter in the asymptotic distribution of the refitted least squares estimation after conditioning on the selection event.

Letting

$$\sqrt{N}_k \eta = a_{N_k} \eta_0,$$

observe that the approximate log-partition function based on Proposition 1.3 is given by:

$$\begin{aligned}
&\frac{a_{N_k}^2}{2}\eta_0' \Sigma_k^{-1}\eta_0 - a_{N_k}^2 \cdot \inf_{\widetilde{\gamma},\widetilde{z}} \Big\{ \frac{1}{2}(\widetilde{\gamma} - \Sigma_k^{-1}\eta_0)' \Sigma_k (\widetilde{\gamma} - \Sigma_k^{-1}\eta_0) \\
&\quad + \frac{1}{2}(1-r_k)^{-1} r_k (\widetilde{z} - \widetilde{\gamma} + \frac{1}{a_{N_k}}\Sigma_k^{-1}\zeta_k)' \Sigma_k (\widetilde{z} - \widetilde{\gamma} + \frac{1}{a_{N_k}}\Sigma_k^{-1}\zeta_k) + \frac{1}{(a_{N_k})^2} B_{s_{E_k}}(a_{N_k} z) \Big\} \\
&= \frac{N_k}{2}\eta' \Sigma_k^{-1}\eta - \inf_{\gamma,z} \Big\{ \frac{1}{2}(\sqrt{N_k}\gamma - \sqrt{N_k}\Sigma_k^{-1}\eta)' \Sigma_k (\sqrt{N_k}\gamma - \sqrt{N_k}\Sigma_k^{-1}\eta) \\
&+ \frac{1}{2}(1-r_k)^{-1} r_k (\sqrt{N_k} z - \sqrt{N_k}\gamma + \Sigma_k^{-1}\zeta_k)' \Sigma_k (\sqrt{N_k} z - \sqrt{N_k}\gamma + \Sigma_k^{-1}\zeta_k) \\
&\qquad + B_{s_{E_k}}(\sqrt{N_k} z) \Big\}.
\end{aligned} \tag{22}$$

The last display is derived by reparameterizing $a_{N_k}\widetilde{\gamma} = \sqrt{N_k}\gamma$, $a_{N_k}\widetilde{z} = \sqrt{N_k} z$. Denote by $\mathcal{P}_{N_k}(\eta)$ the approximate log-partition function in the last display, which we use to obtain an operational expression for the observed Fisher information matrix next. We plug in the sample estimate for the unknown covariance Σ_k to estimate the value of the observed Fisher information matrix.

Proposition A.1 outlines the derivation of this estimator. The symbol e_1 in the claim denotes a $q_k + s$ dimensional block diagonal matrix, where the first s-dimensional matrix along the diagonal is the identity matrix with s columns and the second q_k-dimensional matrix is a matrix of all zeros.

Proposition A.1. *Based on the approximate value for the log-partition function in* (22)*, the inverse of the observed Fisher information matrix assumes the following expression:*

$$e_1' \left((1-r_k)^{-1} \widehat{\Sigma}_k^{-1} - (1-r_k)^{-2} r_k^2 \left((1-r_k)^{-1} r_k \widehat{\Sigma}_k + \nabla^2 B_{s_{E_k}}(\sqrt{N_k}\widehat{z}_k) \right)^{-1} \right) e_1.$$

Proof. Using the approximate expression for the log-partition function in (22), the observed Fisher information submatrix is equal to

$$e_1' \widehat{\Sigma}_k \nabla^2 \mathcal{P}_{N_k}(\widehat{\Sigma}_k \widehat{\gamma}_k^{\text{carve}}) \widehat{\Sigma}_k e_1 = e_1' \widehat{\Sigma}_k \nabla \gamma^* \widehat{\Sigma}_k e_1$$

where γ^* equals

$$\begin{aligned}
&\arg\sup_{\gamma} \sqrt{N_k}\gamma' \widehat{\Sigma}_k \sqrt{N_k} \widehat{\gamma}_k^{\text{carve}} - \frac{1}{2}\sqrt{N_k}\gamma' \widehat{\Sigma}_k \sqrt{N_k}\gamma \\
&\quad - \inf_{z} \Big\{ \frac{1}{2}(1-r_k)^{-1} r_k (\sqrt{N_k} z - \sqrt{N_k}\gamma + \widehat{\Sigma}_k^{-1}\zeta_k)' \widehat{\Sigma}_k (\sqrt{N_k} z - \sqrt{N_k}\gamma + \widehat{\Sigma}_k^{-1}\zeta_k) \\
&\qquad + B_{s_{E_k}}(\sqrt{N_k} z) \Big\}.
\end{aligned} \tag{23}$$

Solving (23) gives us

$$\widehat{\Sigma}_k\sqrt{N_k}\widehat{\gamma}_k^{\text{carve}} = \widehat{\Sigma}_k\sqrt{N_k}\gamma^* + (1-r_k)^{-1}r_k\widehat{\Sigma}_k\left(\sqrt{N_k}\gamma^* - \widehat{\Sigma}_k^{-1}\zeta_k - \sqrt{N_k}\widehat{z}_k\right),$$

where

$$(1-r_k)^{-1}r_k\widehat{\Sigma}_k\left(\sqrt{N_k}\widehat{z}_k - \sqrt{N_k}\gamma^* + \widehat{\Sigma}_k^{-1}\zeta_k\right) + \nabla B_{s_{E_k}}(\sqrt{N_k}\widehat{z}_k) = 0.$$

The former equations yields $\gamma^* = \widehat{\gamma}_k$, using the expression for our carved estimator. Taking further derivatives, we obtain

$$\nabla_{\gamma^*}\widehat{z}_k = \left((1-r_k)^{-1}r_k\widehat{\Sigma}_k + \nabla^2 B_{s_{E_k}}(\sqrt{N_k}\widehat{z}_k)\right)^{-1}(1-r_k)^{-1}r_k\widehat{\Sigma}_k,$$

$$\begin{aligned}\nabla\gamma^* &= \left((1-r_k)^{-1}\widehat{\Sigma}_k - (1-r_k)^{-1}r_k\widehat{\Sigma}_k\nabla_{\gamma^*}\widehat{z}_k\right)^{-1}\\ &= \left((1-r_k)^{-1}\widehat{\Sigma}_k - (1-r_k)^{-1}r_k\widehat{\Sigma}_k\left((1-r_k)^{-1}r_k\widehat{\Sigma}_k + \nabla^2 B_{s_{E_k}}(\sqrt{N_k}\widehat{z}_k)\right)^{-1}(1-r_k)^{-1}r_k\widehat{\Sigma}_k\right)^{-1}.\end{aligned}$$

Then, the inverse for the Fisher information matrix is given by:

$$(1-r_k)^{-1}\widehat{\Sigma}_k^{-1} - (1-r_k)^{-2}r_k^2\left((1-r_k)^{-1}r_k\widehat{\Sigma}_k + \nabla^2 B_{s_{E_k}}(\sqrt{N_k}\widehat{z}_k)\right)^{-1},$$

which directly leads to our claim in the Proposition. □

Acknowledgments

The authors thank the reviewers and the Associate Editor for their constructive remarks.

Funding

The first author is supported by NSF Grants DMS-1951980 and DMS-2113342. The second author is supported in part by NSF Grants DMS-2015325, DMS-2220537, and DMS-2239047. The third author is supported in part by NSF Grants DMS-1914496 and DMS-1951980.

Supplementary Material

Supplement to "Treatment effect estimation with efficient data aggregation" (DOI: 10.3150/24-BEJ1782SUPP; .pdf). The supplementary material includes additional results supporting the main technical findings, as well as a derivation of the carved estimator in our first example. It also contains an instance where we extend our proposal to logistic regression.

References

Belloni, A., Chernozhukov, V. and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.* **81** 608–650. MR3207983 https://doi.org/10.1093/restud/rdt044

Belloni, A., Chernozhukov, V. and Kato, K. (2019). Valid post-selection inference in high-dimensional approximately sparse quantile regression models. *J. Amer. Statist. Assoc.* **114** 749–758. MR3963177 https://doi.org/10.1080/01621459.2018.1442339

Bickel, P.J., Ritov, Y. and Tsybakov, A.B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469 https://doi.org/10.1214/08-AOS620

Cai, T., Liu, M. and Xia, Y. (2022). Individual data protected integrative regression analysis of high-dimensional heterogeneous data. *J. Amer. Statist. Assoc.* **117** 2105–2119. MR4528492 https://doi.org/10.1080/01621459.2021.1904958

Cattaneo, M.D., Jansson, M. and Newey, W.K. (2018). Inference in linear regression models with many covariates and heteroscedasticity. *J. Amer. Statist. Assoc.* **113** 1350–1361. MR3862362 https://doi.org/10.1080/01621459.2017.1328360

Chen, S. and Bien, J. (2020). Valid inference corrected for outlier removal. *J. Comput. Graph. Statist.* **29** 323–334. MR4116045 https://doi.org/10.1080/10618600.2019.1660180

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O. and Roth, A. (2015). Generalization in adaptive data analysis and holdout reuse. In *Adv. Neural. Inf. Process. Syst.*

Farrell, M.H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *J. Econometrics* **189** 1–23. MR3397349 https://doi.org/10.1016/j.jeconom.2015.06.017

Fei, Z. and Li, Y. (2021). Estimation and inference for high dimensional generalized linear models: A splitting and smoothing approach. *J. Mach. Learn. Res.* **22** Paper No. 58, 32. MR4253751

Fithian, W., Sun, D. and Taylor, J. (2014). Optimal inference after model selection. arXiv preprint. arXiv:1410.2597.

Gao, L.L., Bien, J. and Witten, D. (2024). Selective inference for hierarchical clustering. *J. Amer. Statist. Assoc.* **119** 332–342. MR4713896 https://doi.org/10.1080/01621459.2022.2116331

Imbens, G.W. and Rubin, D.B. (2015). *Causal Inference—for Statistics, Social, and Biomedical Sciences: An Introduction*. New York: Cambridge Univ. Press. MR3309951 https://doi.org/10.1017/CBO9781139025751

Lee, J.D., Sun, D.L., Sun, Y. and Taylor, J.E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44** 907–927. MR3485948 https://doi.org/10.1214/15-AOS1371

Lee, J.D., Liu, Q., Sun, Y. and Taylor, J.E. (2017). Communication-efficient sparse regression. *J. Mach. Learn. Res.* **18** Paper No. 5, 30. MR3625709

Lin, D.Y. and Zeng, D. (2010). On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika* **97** 321–332. MR2650741 https://doi.org/10.1093/biomet/asq006

Maity, S., Sun, Y. and Banerjee, M. (2019). Communication-Efficient Integrative Regression in High-Dimensions. arXiv preprint arXiv:1912.11928.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363 https://doi.org/10.1214/009053606000000281

Panigrahi, S. (2023). Carving model-free inference. *Ann. Statist.* **51** 2318–2341. MR4682699 https://doi.org/10.1214/23-aos2318

Panigrahi, S., MacDonald, P.W. and Kessler, D. (2023). Approximate post-selective inference for regression with the group LASSO. *J. Mach. Learn. Res.* **24** Paper No. [79], 49. MR4582501

Panigrahi, S. and Taylor, J. (2018). Scalable methods for Bayesian selective inference. *Electron. J. Stat.* **12** 2355–2400. MR3832095 https://doi.org/10.1214/18-EJS1452

Panigrahi, S. and Taylor, J. (2023). Approximate selective inference via maximum likelihood. *J. Amer. Statist. Assoc.* **118** 2810–2820. MR4681622 https://doi.org/10.1080/01621459.2022.2081575

Panigrahi, S., Taylor, J. and Weinstein, A. (2021). Integrative methods for post-selection inference under convex constraints. *Ann. Statist.* **49** 2803–2824. MR4338384 https://doi.org/10.1214/21-aos2057

Panigrahi, S., Wang, J. and He, X. (2025). Supplement to "Treatment effect estimation with efficient data aggregation." https://doi.org/10.3150/24-BEJ1782SUPP

Panigrahi, S., Mohammed, S., Rao, A. and Baladandayuthapani, V. (2023). Integrative Bayesian models using post-selective inference: A case study in radiogenomics. *Biometrics* **79** 1801–1813. MR4643957 https://doi.org/10.1111/biom.13740

Park, J.J.H., Mogg, R., Smith, G.E., Nakimuli-Mpungu, E., Jehan, F., Rayner, C.R., Condo, J., Decloedt, E.H., Nachega, J.B. and Reis, G. (2021). How COVID-19 has fundamentally changed clinical research in global health. *Lancet Glob. Health* **9** e711–e720.

Schultheiss, C., Renaux, C. and Bühlmann, P. (2021). Multicarving for high-dimensional post-selection inference. *Electron. J. Stat.* **15** 1695–1742. MR4255311 https://doi.org/10.1214/21-ejs1825

Stallard, N. and Todd, S. (2011). Seamless phase II/III designs. *Stat. Methods Med. Res.* **20** 623–634. MR2882895 https://doi.org/10.1177/0962280210379035

Tang, L., Zhou, L. and Song, P.X.-K. (2020). Distributed simultaneous inference in generalized linear models via confidence distribution. *J. Multivariate Anal.* **176** 104567, 13. MR4040151 https://doi.org/10.1016/j.jmva.2019.104567

Tian, X. (2020). Prediction error after model search. *Ann. Statist.* **48** 763–784. MR4102675 https://doi.org/10.1214/19-AOS1818

Tian, X. and Taylor, J. (2018). Selective inference with a randomized response. *Ann. Statist.* **46** 679–710. MR3782381 https://doi.org/10.1214/17-AOS1564

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

Wang, J., He, X. and Xu, G. (2020). Debiased inference on treatment effect in a high-dimensional model. *J. Amer. Statist. Assoc.* **115** 442–454. MR4078474 https://doi.org/10.1080/01621459.2018.1558062

Wolfson, M., Wallace, S.E., Masca, N., Rowe, G., Sheehan, N.A., Ferretti, V., LaFlamme, P., Tobin, M.D., Macleod, J. and Little, J. (2010). DataSHIELD: Resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data. *Int. J. Epidemiol.* **39** 1372–1382.

Received May 2023 and revised May 2024