Improving Treatment Effect Estimators Through Experiment Splitting

Dominic Coey
Facebook
Menlo Park, California
coey@fb.com

Tom Cunningham
Facebook
Menlo Park, California
tomcunningham@fb.com

ABSTRACT

We present a method for implementing shrinkage of treatment effect estimators, and hence improving their precision, via experiment splitting. Experiment splitting reduces shrinkage to a standard prediction problem. The method makes minimal distributional assumptions, and allows for the degree of shrinkage in one metric to depend on other metrics. Using a dataset of 226 Facebook News Feed A/B tests, we show that a lasso estimator based on repeated experiment splitting has a 44% lower mean squared predictive error than the conventional, unshrunk treatment effect estimator, a 18% lower mean squared predictive error than the James-Stein shrinkage estimator, and would lead to substantially improved launch decisions over both.

CCS CONCEPTS

• **General and reference** → **Experimentation**; • **Mathematics of computing** → *Nonparametric statistics*; *Multivariate statistics*; • **Computing methodologies** → *Supervised learning by regression.*

KEYWORDS

Causal Inference; A/B Tests; Empirical Bayes Shrinkage; Sample Splitting; Experiment Meta-Analysis.

ACM Reference Format:
Dominic Coey and Tom Cunningham. 2019. Improving Treatment Effect Estimators Through Experiment Splitting. In *Proceedings of the 2019 World Wide Web Conference (WWW'19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3308558.3313452

1 INTRODUCTION

As the cost of experimentation decreases, researchers are increasingly able to collect data on multiple, related experiments, which test similar kinds of changes [6, 8, 15, 27, 30]. Technology companies may test multiple versions of the machine learning models determining what content is shown to users, where models may differ in their features, architecture, hyperparameter values, or the optimization algorithm used. This raises the prospect of being able to improve inferences for any given experiment, by using the information contained in the other, related experiments. In Bayesian terms, estimating a prior for treatment effect sizes based on previous experiments, and updating it given the data from the current experiment, may yield a better estimate of the true treatment effect than would be possible given the current experiment's data alone. This kind of *empirical Bayes* analysis has a rich history in statistics [11, 12, 19, 22, 31]. With the proliferation of A/B testing, it is increasingly relevant to the modern analysis of experimental data.

We propose a methodology for performing nonparametric, multivariate empirical Bayes shrinkage, which turns the shrinkage problem into a standard prediction problem. The basic idea is as follows: we split each experiment randomly into two subexperiments, each with their own test and control group. We regress the estimated treatment effect in the first subexperiment on the estimated treatment effect in the second subexperiment. Because the dependent variable is unbiased, this regression also estimates the conditional mean of the true treatment effect, given the data in the second subexperiment. In other words, it is an empirical Bayes estimator of the parameter of interest. Other variables likely to be predictive of the treatment effect can be added as covariates into this regression, including estimated treatment effects on auxiliary metrics in the second subexperiment. The prediction algorithm used may be a simple linear regression, but could also be more flexible and nonparametric, allowing for nonlinearities, interactions, and regularization of covariates.

In this methodology, the prior distribution of treatment effects is not explicitly estimated.[1] Instead, we directly estimate the conditional mean of the treatment effect given the observed data, using the experiment splits. Other experiments implicitly determine the prior, and hence the degree of shrinkage implied by this conditional mean. This raises the question of how to choose the fraction of data allocated to the first and the second subexperiments. We derive an expression for the mean squared error of the experiment splitting estimator with univariate least squares shrinkage, and characterize the limiting behavior of the optimal experiment splitting fraction as a function of the number of experiments and the variance of treatment effects and sampling error.

Using a dataset of 226 Facebook News Feed A/B tests, we evaluate the performance of experiment splitting estimators relative to the leading alternatives, including other empirical Bayes shrinkage estimators. We find that experiment splitting reduces mean squared prediction error by 44% compared to the conventional, unshrunk estimator, by 18% compared to James-Stein shrinkage, and by 13% compared to the "Tweedie" estimator. Launching all experiments with positive estimates results in a 40% higher estimated cumulative lift for the lasso than the unshrunk estimator. While experiment splitting generates some performance improvements from flexible,

[1] Other empirical Bayes methods which avoid estimating priors are surveyed in [18], including Robbins' formula [31], the local false-discovery rate [17], and Tweedie's formula [16].

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313452

nonparametric shrinkage and appropriate regularization of treatment effect predictors, we find the main gains are attributable to the multivariate shrinkage that experiment splitting straightforwardly enables. Estimated treatment effects in one metric may reveal information about the true treatment effect in another metric, and experiment splitting makes it easy to incorporate this covariance information.

We focus on shrinkage of treatment effect estimators across experiments, but the idea of using split sample regressions to generate better estimators applies more broadly. If the heterogeneity in treatment effects across user subgroups within an experiment is of interest, experiment splitting can be applied at the level of user subgroups within a single experiment. In non-experimental settings where multiple means from separate samples are being estimated, sample splitting can be used to determine the appropriate degree of shrinkage for each sample mean.

Relative to existing work on empirical Bayesian shrinkage, which we discuss in Section 2, our contribution is to formulate the shrinkage problem as a prediction problem to which standard machine learning algorithms can be applied. This leads to easy-to-implement algorithms for very general forms of shrinkage, which do not require parametric assumptions on the treatment effect or sampling error distributions, and which allow for incorporating information from a potentially high-dimensional vector of auxiliary metrics.

While using split sample regressions to shrink treatment effect estimates appears to be new to the literature, other applications of sample splitting for causal inference abound. Sample splitting has been proposed as a means for constructing valid confidence intervals for heterogeneous treatment effect estimates [3, 37], making the statistical analysis of random forests more tractable [4, 7], and reducing false positives from specification searches [2, 23]. A fast-growing body of work analyzes the properties of semiparametric "cross-fitting" estimators, in which high-dimensional nuisance parameters are estimated by machine learning algorithms on a subsample of the data, and their out-of-sample predictions used to estimate the causal parameters of interest [5, 13, 14, 29, 36, 38].

2 EMPIRICAL BAYESIAN SHRINKAGE

For each of N_E experiments, we have a test group and a control group, each made up of N_P people.[2] In experiment i, we observe test and control group participant outcomes, $(Y^t_{i,k})_{k=1}^{N_P}$ and $(Y^c_{i,k})_{k=1}^{N_P}$. Within each experiment, test and control outcomes are iid across people. The treatment effect in experiment i is defined as $\theta_i = E(Y^t_{i,k}) - E(Y^c_{i,k})$, and the θ_i are themselves iid across experiments. We denote by $\widehat{\theta}_i$ the difference of sample means treatment effect estimator: $\widehat{\theta}_i = \frac{1}{N_P}\sum_{i=k}^{N_P}(Y^t_{i,k} - Y^c_{i,k}) = \theta_i + \varepsilon_i$, where ε_i is sampling error, assumed independent of θ_i. We wish to compute the conditional mean of the treatment effects given the observed data. Of particular interest is the Bayes estimator $E(\theta_i|\widehat{\theta}_i)$, which minimizes mean squared error [31]. We briefly review existing approaches to this problem under various distributional assumptions on $(\theta_i, \varepsilon_i)$.

[2]The assumption that the test and control groups are of equal size is made only to simplify the exposition, and is inessential in what follows.

2.1 Normal Treatments, Normal Errors

Known variances. Consider the case where $\theta_i \sim N(0, \sigma^2_\theta)$, $Y^c_{i,k} \mid \theta_i \sim N(0, \sigma^2_Y)$ and $Y^t_{i,k} \mid \theta_i \sim N(\theta_i, \sigma^2_Y)$, for known variances σ^2_θ and σ^2_Y. The difference of sample means estimator has distribution $\widehat{\theta}_i \mid \theta_i \sim N(\theta_i, \sigma^2_\varepsilon)$, where $\sigma^2_\varepsilon = 2\sigma^2_Y/N_P$. By Bayes' rule the posterior distribution of the treatment effect given the data is $\theta_i \mid \widehat{\theta}_i \sim N(\alpha\widehat{\theta}_i, \alpha\sigma^2_\varepsilon)$, where $\alpha = \sigma^2_\theta/(\sigma^2_\theta + \sigma^2_\varepsilon)$. The Bayes estimator of θ_i is $E(\theta_i \mid \widehat{\theta}_i) = \alpha\widehat{\theta}_i$. This shrinks the difference of sample means estimator, $\widehat{\theta}_i$, towards the prior mean of the treatment effects, zero, by the coefficient α.[3] Thus with normal treatment effects and sampling error, optimal shrinkage is linear in the estimate $\widehat{\theta}_i$. The Bayes risk with squared loss for the difference of sample means $\widehat{\theta}_i$ is $\sum_{i=1}^{N_E} E(\widehat{\theta}_i - \theta_i)^2 = \sigma^2_\varepsilon N_E$, whereas for the Bayesian shrinkage estimator $\alpha\widehat{\theta}_i$, it is $\sum_{i=1}^{N_E} E(\alpha\widehat{\theta}_i - \theta_i)^2 = \alpha\sigma^2_\varepsilon N_E$ [17]. The gains from optimal shrinkage are large if the signal-to-noise ratio $\sigma^2_\theta/\sigma^2_\varepsilon$, and hence α, is small. When the sampling error variance σ^2_ε is very large, almost all of the variation in $\widehat{\theta}_i$ reflects sampling error, rather than variation in the underlying treatment effect θ_i, and the Bayes estimator offsets the effect of sampling error by aggressively shrinking $\widehat{\theta}_i$ towards the prior mean.

Unknown treatment effect variance. In practice σ^2_θ, and hence the desired shrinkage factor α, may be unknown. One may specify a prior distribution over σ^2_θ and conduct full Bayesian inference, computing the posterior of θ_i given the data and priors [34, 35]. An alternative is empirical Bayes, in which the shrinkage factor is estimated from the data. The celebrated James-Stein estimator is

$$\widehat{\theta}^{JS}_i = \left(1 - \frac{(N_E - 2)\sigma^2_\varepsilon}{\sum_{i=1}^{N_E}\widehat{\theta}^2_i}\right)\widehat{\theta}_i.$$

Although originally studied in a frequentist setting with fixed $\theta_i's$ [26, 32], it can be viewed as an empirical Bayes estimator, where $1 - (N_E - 2)\sigma^2_\varepsilon/\sum_{i=1}^{N_E}\widehat{\theta}^2_{i,2}$ is an unbiased estimator of the Bayes optimal shrinkage factor [17, 20, 33].

Unknown sampling error variance or prior mean. In experimental settings the sampling error variance σ^2_ε is typically unknown, but can be readily estimated. Letting $\widehat{\sigma}^2_Y$ denote the sample variance of all test and control outcomes across all people and experiments, we use the estimator $\widehat{\sigma}^2_\varepsilon = 2\widehat{\sigma}^2_Y/N_P$. When the prior mean is non-zero, the same shrinkage logic as above applies, with estimates being linearly shrunk towards the prior mean instead of towards zero. When this prior mean is unknown, it can be estimated as the sample mean across experiments of the observed $\widehat{\theta}_i$.

2.2 General Treatments, Normal Errors

The treatment effects θ_i need not be normally distributed, and are significantly non-normal in the data we study in Section 5. This implies that $E(\theta_i \mid \widehat{\theta}_i)$ need not be linear in $\widehat{\theta}_i$, unlike in the normal case. Let f denote the marginal density of $\widehat{\theta}_i$. Then *Tweedie's formula* ([10, 16]) for the posterior mean of θ_i is:

[3]The fact that the Bayes optimal estimator of θ_i involves shrinking $\widehat{\theta}_i$ towards zero is consistent with $\widehat{\theta}_i$ being unbiased for θ. The former observation is that $E(\theta_i \mid \widehat{\theta}_i) \leq \widehat{\theta}_i$; the latter that $E(\widehat{\theta}_i \mid \theta_i) = \theta_i$.

$$E(\theta_i \mid \widehat{\theta}_i) = \theta_i + \sigma_\varepsilon^2 \frac{d \log f(\widehat{\theta}_i)}{d\theta_i}. \quad (1)$$

Because the $\widehat{\theta}_i$ are observed, unlike the θ_i, we can directly estimate the density f from the data, and use this estimate in equation (1) to construct an estimator of $E(\theta_i \mid \widehat{\theta}_i)$. Even if the treatment effects are non-normal, normality may still be a reasonable assumption for the sampling error. If N_P is large and the assumptions of the central limit theorem hold, ε_i will be close to normal.

3 EXPERIMENT SPLITTING

Experiment splitting takes a different approach to estimating treatment effects. The main observation is that multiple independent estimates of the same treatment effect can be obtained from a single experiment, by simply partitioning the experiment into subexperiments. In brief, by randomly splitting each experiment into two subexperiments, we can estimate the conditional mean of the treatment effect estimate in the second subexperiment given the treatment effect estimate in the first. This conditional mean is what is required for optimally shrinking our estimates from the second subexperiment.

In more detail, we split both the test group and the control group into two subgroups, with a fraction γ of people selected uniformly at random for the first subexperiment (up to integer constraints), and the remainder assigned to the second subexperiment. Without loss of generality, let $(Y^t_{i,k})_{k=1}^{\lfloor \gamma N_P \rfloor}$ and $(Y^c_{i,k})_{k=1}^{\lfloor \gamma N_P \rfloor}$ be the test and control outcomes of the people in the first subexperiment, and similarly $(Y^t_{i,k})_{k=\lfloor \gamma N_P \rfloor+1}^{N_P}$ and $(Y^c_{i,k})_{k=\lfloor \gamma N_P \rfloor+1}^{N_P}$ for people in the second subexperiment. In each subexperiment, we construct the difference of means treatment effect estimator: $\widehat{\theta}_{i,1} = \frac{1}{\lfloor \gamma N_P \rfloor} \sum_{k=1}^{\lfloor \gamma N_P \rfloor} (Y^t_{i,k} - Y^c_{i,k})$ and $\widehat{\theta}_{i,2} = \frac{1}{N_P - \lfloor \gamma N_P \rfloor} \sum_{k=\lfloor \gamma N_P \rfloor+1}^{N_P} (Y^t_{i,k} - Y^c_{i,k})$. It follows that $E(\widehat{\theta}_{i,2} \mid \widehat{\theta}_{i,1}) = E(\theta_i \mid \widehat{\theta}_{i,1})$, as $\widehat{\theta}_{i,2}$ equals θ_i plus a mean zero error term independent of $\widehat{\theta}_{i,1}$. Moreover, $E(\widehat{\theta}_{i,2} \mid \widehat{\theta}_{i,1})$, and hence $E(\theta_i \mid \widehat{\theta}_{i,1})$, can be estimated by standard regression or prediction techniques. This conditional mean is the Bayes estimator of θ_i given the data in the first subexperiment. Thus by having a set of "hold-out" subexperiments, we can estimate the optimal level of shrinkage. This requires no assumptions on the marginal distributions of θ_i or ε_i, beyond the mild regularity conditions for estimating a conditional mean.

Experiment splitting yields an estimate of $E(\theta_i \mid \widehat{\theta}_{i,1})$, not $E(\theta_i \mid \widehat{\theta}_i)$. The conditioning is on the difference in means in only the first subexperiment, not the entire data. In principle not conditioning on all data might impair estimator performance, to the extent that more conventional estimators dominate experiment splitting. The results of Section 5 indicate to the contrary that experiment splitting estimators can substantially outperform unshrunk, James-Stein and Tweedie estimators, in real-world A/B tests.

This framework also naturally extends to accommodate covariates. Letting X denote a vector of variables available based on data from the first subexperiment, we can estimate $E(\widehat{\theta}_{i,2} \mid X_i)$. In contrast to conventional experiment meta-analysis, X_i includes an estimate of the outcome of interest formed from the data in experiment i, $\widehat{\theta}_{i,1}$. In addition, the covariate vector X_i may include treatment effects on other metrics in the same experiment, or experimental metadata, all of which may be useful for predicting θ_i. Thus the experiment splitting estimator's performance may be improved, if auxiliary metrics or experiment metadata are available that are predictive of the outcome of interest. The conditional expectation can be estimated by standard prediction algorithms like the lasso, gradient boosted decision trees, or neural nets. Thus experiment splitting allows us to treat nonparametric multivariate shrinkage as a simple prediction problem. Unlike classical approaches to the problem of multivariate shrinkage (e.g. [21]), it is unnecessary to estimate the full covariance matrix of all metrics' treatment effects.[4] The covariate vector may even be high-dimensional, as this can be accommodated by common supervised learning algorithms.

Generating confidence intervals around the estimate of $E(\theta_i \mid X_i)$ poses no special difficulty for the experiment splitting estimator. The only requirement is that valid confidence intervals can be calculated for whatever statistical procedure is applied for estimating the conditional mean.

3.1 Selecting the Experiment Splitting Fraction

Because of the relation between experiment splitting and cross-validation described in Section 3.4, we also call the first subexperiment (used to construct the covariates in the experiment splitting regression) the "training" subexperiment, and the second subexperiment (used to construct the outcomes) the "validation" subexperiment. The relative amount of data in training and validation subexperiments affects the properties of the experiment splitting estimator. If most of the data is assigned to the training subexperiment, then the sampling error in $\widehat{\theta}_{i,1}$ is relatively small; if most of the data is assigned to the validation subexperiment, then the sampling error in $\widehat{\theta}_{i,2}$ is relatively small. The optimal splitting fraction strikes a balance so that $\widehat{E}(\theta_i \mid \widehat{\theta}_{i,1})$, the estimator of the conditional mean $E(\theta_i \mid \widehat{\theta}_{i,1})$, is as close as possible on average to θ_i.

We present two strategies for estimating the out-of-sample error rate for selecting the optimal experiment splitting fraction. The first is cross-validation. This requires splitting the original experiment into *three*: two subexperiments corresponding to the training and validation data, and a third, "test" subexperiment, which is used as a out-of-sample, independent measure of the truth. The difference between a model's predictions—which are a function of the training and validation data—and the estimates of the true effect from the test data, can be used to estimate the model's performance. We perform the random assignment of data into training, validation and tests sets $S > 1$ times and average the results, to reduce the variance in the estimated performance metric due to the choice of splits. This entire procedure can be performed for various choices of γ (which determines the relative sizes of the training and validation subexperiments), and the γ yielding the lowest prediction error can be selected. Algorithm 1 summarizes this procedure, given $\gamma \in \{\gamma_1, \gamma_2 \ldots, \gamma_G\}$, and with half of the overall data allocated to the test subexperiment.

[4] In addition [21] assume that the metrics' sampling error is independent across metrics and homoskedastic. These assumptions—especially independence across metrics—are hard to justify in typical A/B tests.

Algorithm 1 Cross-Validation for Optimal Splitting

(1) For splitting fraction indices $g = 1$ to G:
  (a) For splitting simulations $s = 1$ to S:
    (i) For experiments $i = 1$ to N_E, randomly split experiment i into the training, validation and test subexperiments, each with a proportion $0.5\gamma_g, 0.5(1-\gamma_g)$ and 0.5 of the data, respectively.
    (ii) Apply the experiment splitting estimator to the training and validation subexperiments, and obtain predictions of the true effect in each experiment.
    (iii) Calculate the mean squared difference between these predictions and the estimated effects in the test subexperiment. Denote this err_{gs}.
  (b) Calculate the mean squared prediction error over splitting simulations s, $\text{MSPE}_g := \frac{1}{S}\sum_{s=1}^{S}\text{err}_{gs}$.
(2) Return γ_{g^*}, where $g^* \in \arg\min_g \text{MSPE}_g$.

The second strategy for selecting the experiment splitting fraction involves minimizing an unbiased risk estimate. This is analogous to covariance penalty approaches to estimating prediction error, including Mallows' C_p [28], the Akaike Information Criterion (AIC) [1], and Stein's unbiased risk estimate [34]. Unlike cross-validation, the unbiased risk estimate we develop applies only to multivariate OLS shrinkage, but does not require holding out a test subexperiment, is computationally simpler, and yields insight into how the optimal splitting fraction depends on sample sizes and variance parameters. Given a vector of experiment covariates X_i of length l, the treatment effect estimator for experiment i is $X_i'\widehat{\beta}$, where $\widehat{\beta} = \arg\min_\beta \sum_i(\widehat{\theta}_{i,2} - X_i'\widehat{\beta})^2$. The unshrunk subexperiment treatment effect estimates are written as $\widehat{\theta}_{i,j} = \theta_i + \varepsilon_{i,j}$, for $j = 1, 2$. We assume that the error $\varepsilon_{i,j}$ is independent across i and j, and that $\varepsilon_{i,2}$ is independent of X_i. Let $\sigma^2_{\varepsilon_j}$ denote the variance of $\varepsilon_{i,j}$. We also assume that the $N_E \times l$ regressor matrix $X = (X_1, X_2, \ldots, X_{N_E})'$ has rank l with probability 1. The following proposition forms the basis of our unbiased risk estimator. All proofs are in the appendix.

Proposition 1. *The mean squared error of the multivariate linear shrinkage estimators* $(X_i'\widehat{\beta})_{i=1}^{N_E}$ *is*

$$\sum_{i=1}^{N_E} E(\theta_i - X_i'\widehat{\beta})^2 = \left(\sum_{i=1}^{N_E} E(\widehat{\theta}_{i,2} - X_i'\widehat{\beta})^2\right) - (N_E - 2l)\sigma^2_{\varepsilon_2}.$$

Proposition 1 shows that to estimate risk, we simply adjust the sum of squared residuals by a term which depends on N_E, the number of experiments, and l, the number of regressors. As l increases, the in-sample model fit, measured by the sum of squared residuals, increasingly understates the true error. Let $\widehat{\sigma}^2_{\varepsilon_2}$ denote an unbiased estimator of $\sigma^2_{\varepsilon_2}$, which can be easily calculated from the individual-level outcomes in the second subexperiment. Then $\left(\sum_{i=1}^{N_E}(\widehat{\theta}_{i,2} - X_i'\widehat{\beta})^2\right) - (N_E - 2l)\widehat{\sigma}^2_{\varepsilon_2}$ is unbiased for the risk of the experiment splitting estimator. This risk estimate can be calculated for different experiment splitting fractions, in order to select the estimated risk-minimizing fraction.

3.2 Determinants of the Optimal Split

In the case of univariate OLS, we regress the estimated treatment effect in the validation subexperiment on a constant and the estimated treatment effect in the training subexperiment. Our shrinkage estimator for experiment i is $\widehat{\beta}_0 + \widehat{\beta}_1\widehat{\theta}_{i,1}$, where $(\widehat{\beta}_0, \widehat{\beta}_1) = \arg\min_{\beta_0,\beta_1}\sum_i(\widehat{\theta}_{i,2} - \beta_0 - \beta_1\widehat{\theta}_{i,1})^2$. With relatively more data in the training subexperiment, the variance of $\widehat{\theta}_{i,1}$ decreases. But little data in the validation subexperiment means that the $\widehat{\theta}_{i,2}$ are imprecisely estimated, increasing the variance of $(\widehat{\beta}_0, \widehat{\beta}_1)$. We show how the optimal splitting fraction that balances these two effects depends on the number of experiments, and the variance of treatment effects and sampling error. The following result gives an explicit expression for risk of univariate OLS shrinkage in terms of the sample size and variance parameters.

Corollary 1. *The mean squared error of the univariate linear shrinkage estimators* $(\widehat{\beta}_0 + \widehat{\beta}_1\widehat{\theta}_{i,1})_{i=1}^{N_E}$ *is*

$$\sum_{i=1}^{N_E} E(\theta_i - \widehat{\beta}_0 - \widehat{\beta}_1\widehat{\theta}_{i,1})^2 = \frac{(N_E - 2)\sigma^2_\theta\sigma^2_{\varepsilon_1}}{\sigma^2_\theta + \sigma^2_{\varepsilon_1}} + \sigma^2_{\varepsilon_2}.$$

The next result derives the limiting behavior of the optimal split. For simplicity we treat the sample sizes and the experiment splitting fraction as real-valued variables, so the necessary derivatives are defined. We require $0 < \underline{\gamma} \le \gamma \le \overline{\gamma} < 1$ for some $\underline{\gamma}$ and $\overline{\gamma}$ close to zero and one, so that no subexperiment has zero observations.

Proposition 2. *Assume* $Var(Y_i^t|\theta_i) = Var(Y_i^c|\theta_i) = \sigma^2_Y$ *for all experiments* i, *and* $N_E \ge 3$. *Define* $\sigma^2_\varepsilon = \frac{2\sigma^2_Y}{N_P}$, *and let* $\gamma^* \in [\underline{\gamma}, \overline{\gamma}]$ *minimize mean squared error from univariate linear shrinkage. Then* $\lim_{N_E\to\infty}\gamma^* = \overline{\gamma}$, $\lim_{\sigma^2_\theta\to 0}\gamma^* = \lim_{\sigma^2_\varepsilon\to\infty}\gamma^* = \underline{\gamma}$, *and* $\lim_{\sigma^2_\theta\to\infty}\gamma^* = \lim_{\sigma^2_\varepsilon\to 0}\gamma^* = \frac{\sqrt{N_E-2}}{1+\sqrt{N_E-2}}$.

Of particular note is the limiting behavior of γ^* as N_E grows. The optimal split puts increasingly more data in the training subexperiment, as any noise in the dependent variable of the splitting regression can be compensated for with a sufficiently large number of experiments N_E. Figure 1 shows the mean squared error from univariate linear shrinkage, for varying values of σ^2_θ and N_E, for σ^2_ε fixed at 1, and normal treatment effects and sampling errors. Consistent with Proposition 2, the optimal γ is increasing in σ^2_θ and N_E. The figure illustrates the asymmetry in risk caused by very small vs. very large values of γ. As γ approaches 0, the signal in $\widehat{\theta}_{i,1}$ vanishes, $\widehat{\beta}_1$ approaches 0, and experiment splitting approaches the bounded risk of the complete shrinkage case. As γ approaches 1, risk is unbounded: extreme imprecision in the dependent variable can cause extreme errors in the shrinkage estimator.

When $\sigma^2_\theta = 0.1$ and $N_E = 100$, there is no interior solution for γ^*, and risk is strictly increasing in γ. This implies that experiment splitting is futile if the signal-to-noise ratio is too small. For large enough sampling error variance, complete shrinkage—the limiting case as γ approaches 0—is always superior to experiment splitting. The suboptimality involved in performing experiment splitting nonetheless is small, however, even for moderate values of γ.

Figure 2 shows γ^*, the optimal fraction of data to allocate to the training subexperiment, as a function of the variance of treatment

Figure 1: Experiment Splitting Fractions and RMSE

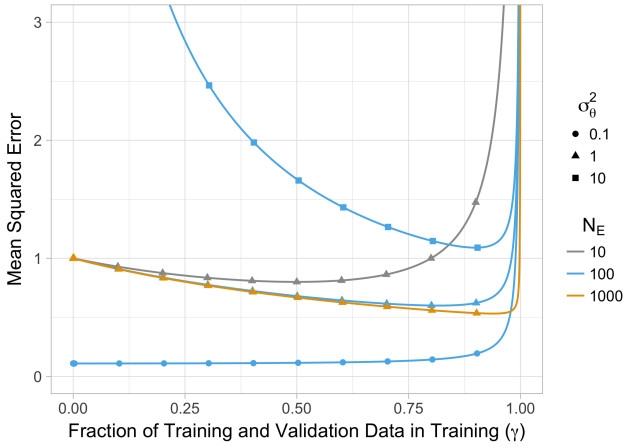
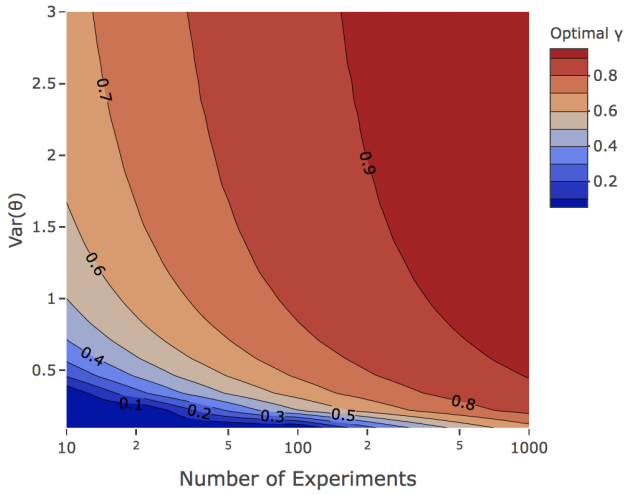


Figure 2: Optimal Experiment Splitting Fraction



effects, σ^2_θ, and the number of experiments, N_E. The sampling error σ^2_ε is again fixed at one. With N_E between 100 and 200 and a signal-to-noise ratio between around one and three, the training subexperiment should contain about 80% to 90% of the data.

3.3 Repeated Experiment Splitting

Our focus until now has been on generating predictions from datasets in which each experiment is split *once*.[5] An alternative is repeated splitting, where we randomly split each experiment into training and validation subexperiments $R > 1$ times. This is appealing, as averaging over repeated splits reduces the "excess" variance due to the random choice of user assignments to subexperiments. This averaging can be understood as a form of Rao-Blackwellization. The experiment splitting estimator is a function of the training and validation assignment, and the individual-level data. Conditioning on the individual-level data—a sufficient statistic—by averaging possible training and validation splits will improve the mean-squared error of the experiment splitting estimator. In the empirical results of Section 5, we verify that this is indeed the case.

[5]This is true even in the cross-validation procedure of Algorithm 1: the additional split there is used to evaluate the quality of the model predictions, not to generate the predictions themselves.

Algorithm 2 Repeated Experiment Splitting

(1) For experiment splitting repetitions $r = 1$ to R:
  (a) For $i = 1$ to N_E, randomly split experiment i into two subexperiments, with a fraction γ of data in the training subexperiment and $1 - \gamma$ in the validation subexperiment. Let $\widehat{\theta}^{(r)}_{i,1}$ and $\widehat{\theta}^{(r)}_{i,2}$ denote the unshrunk estimated treatment effects in those subexperiments.
  (b) Given the set of N_E pairs $\left(\widehat{\theta}^{(r)}_{i,1}, \widehat{\theta}^{(r)}_{i,2}\right)_{i=1}^{N_E}$, estimate the conditional mean $E\left(\widehat{\theta}^{(r)}_{i,2} \mid \widehat{\theta}^{(r)}_{i,1}\right)$. Denote this estimator by $\widehat{m}^{(r)}(\cdot)$.
  (c) For $i = 1$ to N_E, compute the shrinkage estimates $s^{(r)}_{i,1} := \widehat{m}^{(r)}\left(\widehat{\theta}^{(r)}_{i,1}\right)$.
(2) For $i = 1$ to N_E, return the repeated experiment splitting shrinkage estimator for experiment i, given by $\frac{1}{R}\sum_{r=1}^{R} s^{(r)}_{i,1}$.

Algorithm 2 describes the repeated experiment splitting algorithm. The case of $R = 1$ reduces to the original, single repetition experiment splitting estimator. The validity of this procedure for $R > 1$ follows immediately from the validity of the $R = 1$ case. The conditional mean $E\left(\widehat{\theta}^{(r)}_{i,2} \mid \widehat{\theta}^{(r)}_{i,1}\right)$ does not depend on the repetition r, since the data is identically distributed across repetitions. Denote this conditional mean function by $m(\cdot)$. For each repetition r, as long as our estimator $\widehat{m}^{(r)}(\cdot)$ is consistent for $m(\cdot)$, the average of these estimators over r will also be consistent. If the estimates $\widehat{m}^{(r)}(\cdot)$ were independent across r, setting the number of repetitions R to be sufficiently high would ensure arbitrarily precise estimates of m. In practice the $\widehat{m}^{(r)}(\cdot)$ are correlated as they are all functions of the same dataset, and this will not be the case.

For cross-validation of the splitting fraction, we nest this algorithm in the cross-validation meta-algorithm, Algorithm 1. The inner loop of Algorithm 1 calls for applying an experiment splitting estimator given a splitting fraction γ_g. In this setting, implementing that step itself requires splitting the data R times.

3.4 Connection with Cross-Validation

Experiment splitting may appear similar to the practice of cross-validating predictive models. The researcher splits experiments and estimates the optimal degree of shrinkage, based on predicting outcomes in the "held-out" subexperiments. Similarly, in cross-validation, the researcher splits data and estimates the optimal degree of regularization based on predicted outcomes in the held-out data. This is more than an analogy. Experiment splitting with univariate OLS shrinkage and a prior treatment effects mean of zero

is equivalent to cross-validation of a particular ridge regression. This connection holds for any fixed splitting fraction γ, and is unrelated to the use of cross-validation for selecting the splitting fraction as described in Section 3.1.

In experiment splitting with a prior treatment effect mean of zero, the univariate linear regression problem is $\min_\beta \sum_i (\widehat{\theta}_{i,2} - \beta\widehat{\theta}_{i,1})^2$. Define $\Delta_{i,k} = Y^t_{i,k} - Y^c_{i,k}$. That is, we arbitrarily pair up test and control participants into N_P pairs, and denote each pair's difference in outcomes by $\Delta_{i,k}$. Note that $E(\Delta_{i,k}) = \theta_i$ for all i. Regress $\Delta_{i,k}$ on N_E indicator variables, one corresponding to each experiment, on the sample of people in the first subexperiment:

$$\Delta_{i,k} = \sum_{j=1}^{N_E} \alpha_j \mathbb{I}(i = j) + \epsilon_{i,k},$$

for $i = 1, 2, \ldots, N_E$ and $k = 1, 2, \ldots \lfloor \gamma N_P \rfloor$. Let $\widehat{\alpha}_{OLS,j}$ denote the OLS estimates of the α_j. Each $\widehat{\alpha}_{OLS,j}$ will simply be the average of the dependent variable for observations corresponding to that experiment, and so we recover the difference in sample means estimator: $\widehat{\alpha}_{OLS,j} = \widehat{\theta}_{j,1}$ for all experiments j. Letting $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_{N_E})'$, the ridge regression problem for a given regularization parameter λ is

$$\min_\alpha \left\{ \left(\Delta_{i,k} - \sum_{j=1}^{N_E} \alpha_j \mathbb{I}(i = j) \right)^2 + \lambda \alpha' \alpha \right\}. \tag{2}$$

Denoting the design matrix for this problem by X and the dependent variable by y, the solution is $\widehat{\alpha}_{Ridge} = (X'X + \lambda I)^{-1} X'y$. Since X is orthogonal, this implies $\widehat{\alpha}_{Ridge,j} = \frac{\lfloor \gamma N_P \rfloor}{\lfloor \gamma N_P \rfloor + \lambda} \widehat{\theta}_{j,1}$ for all j. Thus ridge regression shrinks all the raw estimates to zero by the same relative amount, where the shrinkage factor depends on λ.

Choosing λ by cross-validation, i.e. minimizing out-of-sample squared error in predictions on the data from the second subexperiments, amounts to solving:

$$\min_\lambda \sum_{i=1}^{N_E} \sum_{k=\lfloor \gamma N_P \rfloor + 1}^{N_P} \left(\Delta_{i,k} - \frac{\lfloor \gamma N_P \rfloor}{\lfloor \gamma N_P \rfloor + \lambda} \widehat{\theta}_{i,1} \right)^2.$$

Because $\widehat{\theta}_{i,2} = \frac{1}{N_P - \lfloor \gamma N_P \rfloor} \sum_{k=\lfloor \gamma N_P \rfloor + 1}^{N_P} \Delta_{i,k}$, this is equivalent to

$$\min_\lambda \sum_{i=1}^{N_E} \left(\widehat{\theta}_{i,2} - \frac{\lfloor \gamma N_P \rfloor}{\lfloor \gamma N_P \rfloor + \lambda} \widehat{\theta}_{i,1} \right)^2,$$

which in turn is equivalent to the original experiment splitting problem, $\min_\beta \sum_{i=1}^{N_E} (\widehat{\theta}_{i,2} - \beta\widehat{\theta}_{i,1})^2$. It follows that the experiment splitting gives the same estimators as cross-validated ridge regression. Similarly, the repeated experiment splitting procedure described in Algorithm 2 is equivalent to a variant of k-fold cross-validation, where for each fold, each experiment is randomly split into training and validation datasets. This connection motivates the naming convention of calling the first subexperiment the training subexperiment and the second the validation subexperiment, as equation (2) is estimated on the first dataset, and its tuning parameter λ is chosen by cross-validation against the second dataset.

Section 2.1 notes that with normal sampling error and treatment effects, the optimal shrinkage is linear. The Bayesian interpretation of ridge regression makes the same assumptions of normality on the treatment effects and the sampling error, so it is natural that it also implies linear shrinkage. Other penalties on the α parameter in equation (2) correspond to non-normal priors on treatment effects, and hence to fitting different, nonlinear regression models in the experiment splitting problem. The L_1 lasso penalty, for example, is equivalent to solving the experiment splitting problem with the soft-thresholding operator $\min_\alpha \sum_{i=1}^{N_E} \left(\widehat{\theta}_{i,2} - \text{sign}(\widehat{\theta}_{i,1})(|\widehat{\theta}_{i,1}| - \alpha)^+ \right)^2$ (see [24] for discussion of the relation between the lasso and the soft-thresholding operator).

Relative to the experiment-level regression of $\widehat{\theta}_{i,2}$ on $\widehat{\theta}_{i,1}$, the individual-level regression in equation (2) requires specifying the form of regularization term (and implicitly, the prior distribution of treatment effects). Inappropriate choices of regularization could lead to suboptimal forms of shrinkage. The L_2 penalty in ridge regression, implying linear shrinkage, may be inappropriate when the distribution of treatment effects is highly non-normal. The experiment-level regression of $\widehat{\theta}_{i,2}$ on $\widehat{\theta}_{i,1}$ requires no such assumptions, and will asymptotically recover the optimal shrinkage with any consistent nonparametric regression technique.

4 EXTENSIONS

4.1 Shrinkage Across Subgroups

Our focus has been on experiment splitting regressions where each data point in the regression corresponds to a separate experiment. We may also wish to perform shrinkage across non-overlapping subgroups of people within an experiment, either instead of, or in addition to, shrinkage across experiments. The experiment splitting methodology extends directly to this setting: instead of i indexing experiments, let i index subgroups (or subgroup by experiment pairs), and the same methodology applies without modification. Subgroup metrics which may be correlated with the magnitude of treatment effects or their sampling error variance, including demographics and subgroup size, are natural covariates to include in the experiment splitting regression.

4.2 Predicting Using Pooled Estimates

A variation on experiment splitting is to estimate $E(\theta_i \mid \widehat{\theta}_{i,1})$ as usual, but to generate predictions of θ_i by evaluating this conditional mean estimate $\widehat{m}$ at the *pooled* estimate $\widehat{\theta}_i$, instead of the subexperiment estimate $\widehat{\theta}_{i,1}$. Intuitively, evaluating $\widehat{m}$ at an estimator with the same mean and a lower variance should improve performance. We find this indeed results in a non-trivial decrease in prediction error in our data, but only if repeated experiment splitting is not already being used. A heuristic explanation can be given in the case of univariate OLS experiment splitting. For simplicity of exposition we assume the prior mean of θ_i to be zero. Let $\mathcal{D}$ denote all experimental data, and $\mathcal{A}$ the random assignments of people to subexperiments. Conditional on the data and averaging over repeated splits, the prediction from repeated experiment splitting as described in Algorithm 2 is $E_{\mathcal{A}}(\widehat{\beta}_1 \widehat{\theta}_{i,1} \mid \mathcal{D}) + O(1/R)$. The average prediction from repeated experiment splitting, evaluating the estimated conditional mean at the pooled mean $\widehat{\theta}_i$, is $E_{\mathcal{A}}(\widehat{\beta}_1 \mid \mathcal{D})\widehat{\theta}_i$. Under standard regularity conditions, for any fixed experiment i, $E_{\mathcal{A}}(\widehat{\beta}_1 \widehat{\theta}_{i,1} \mid \mathcal{D}) = E_{\mathcal{A}}(\widehat{\beta}_1 \mid \mathcal{D}) E_{\mathcal{A}}(\widehat{\theta}_{i,1} \mid \mathcal{D}) + O(1/N_E)$, as the

influence any single experiment has on the estimated regression parameter $\widehat{\beta}_1$ is $O(1/N_E)$. But $E_{\mathcal{A}}(\widehat{\theta}_{i,1} \mid \mathcal{D}) = \widehat{\theta}_i$, so with repeated experiment splitting, the difference between predictions using pooled estimates and predictions using subexperiment estimates must be of order $O(1/R) + O(1/N_E)$.

It may appear that the reason for this lack of performance improvement is that the function $\widehat{m}$, designed as it is for optimally shrinking the subexperiment estimate $\widehat{\theta}_{i,1}$, results in "overshrinking" the more accurate, pooled estimate $\widehat{\theta}_i$. A feasible procedure for transforming an estimate of $E(\theta_i \mid \widehat{\theta}_{i,1})$ into an estimate of $E(\theta_i \mid \widehat{\theta}_i)$ would help avoid this issue. One could estimate the former function by experiment splitting, transform it to an estimate of the latter function, and evaluate the latter function at the pooled estimate, $\widehat{\theta}_i$. It appears that performing this transformation in the general, multi-metric, non-normal case would effectively require estimating the full joint distribution of treatment effects and sampling error.[6] This would void the benefits that shrinkage via experiment splitting has in simplicity and ease of implementation, so we do not pursue this extension here.

5 EMPIRICAL RESULTS

5.1 Data

We study a set of 226 Facebook News Feed A/B tests conducted in 2018.[7] Each A/B test caused a change in the order in which stories appear in News Feed, with the goal of showing users more relevant content first. We analyze a single day's worth of data from each test. The outcome metric we study is the percentage change in posts per person to News Feed in the test group relative to the control group. In addition, for each test we track the percentage change in 23 auxiliary metrics, including likes, comments and link clicks. These will be used as covariates in our regressions, allowing us to make more precise inferences about the percentage change in posts per person. People in each test are randomly assigned with equal probability to the test and control groups. The sample size in each test varies between 10.0 and 10.9 million people.[8]

Figure 3 shows the distributions across experiments of the percentage change in posts, and its standard error. We calculate standard errors using the delta method. The percentage change in posts is non-normal: the Shapiro-Wilk test rejects the null hypothesis of normality with a p-value below 0.01%. This remains true after removing the experiments with a greater than 2% change in absolute value. This suggests that linear shrinkage, as implied by normal treatment effects and normal sampling error (and hence normal observed outcomes), is likely suboptimal.

Figure 4 shows the scatterplot of the percentage change in posts, for the training and validation subexperiments. There are two sets of points. The first is for $\gamma = 0.2$ (a 20%-80% split of data into

Figure 3: Distributions of Estimates and Standard Errors

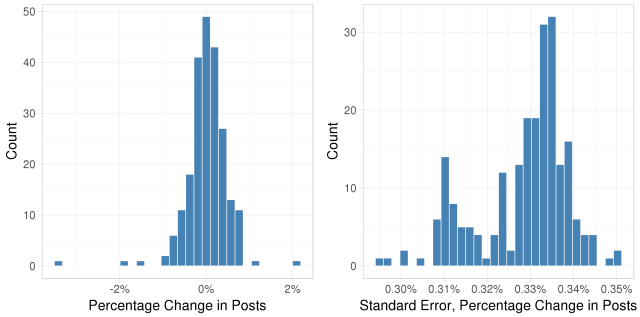


Figure 4: Percentage Change in Posts, $\gamma = 0.2, 0.8$

training and validation sets), and the second is for $\gamma = 0.8$. For each set of points, we plot the OLS line of best fit. The dashed gray line is the 45-degree line. For large N_E, the OLS slope coefficient is approximately $Cov(\widehat{\theta}_{i,2}, \widehat{\theta}_{i,1})/Var(\widehat{\theta}_{i,1}) = \sigma_\theta^2/(\sigma_\theta^2 + \sigma_{\varepsilon_1}^2)$, which is decreasing in $\sigma_{\varepsilon_1}^2$ and less than one.[9] Consistent with this, in Figure 4, increasing γ from 0.2 to 0.8 increases the slope of the line of best fit, but the slope remains less than than one.

5.2 Estimators

The estimators we consider fall into two classes. The first are the experiment splitting estimators, which fit predictive models for the percentage change in posts in the validation subexperiments, given the data from the training subexperiments. The second are the pooling estimators, which combine the data from the training

[6] The special case of a single metric with normal treatment effects and sampling error is straightforward. With a fraction γ of data in the training subexperiment, if $E(\theta_i \mid \widehat{\theta}_{i,1}) = \beta_1 \widehat{\theta}_{i,1}$, then it can be shown that $E(\theta_i \mid \widehat{\theta}_i) = \frac{\beta_1}{\lambda(1-\beta_1)+\beta_1} \widehat{\theta}_i$.

[7] This is a convenience sample, and is not intended to give a representative picture of the effects of News Feed tests generally.

[8] Estimates from smaller experiments should be shrunk more than estimates from larger experiments, all else equal. In practice the variation in sample sizes in this dataset is relatively small, and the gains from including sample size as an additional covariate in our predictive models were minimal.

[9] If $\gamma = 0.5$, the training and validation datasets have the same distributions, and it might be supposed that, by symmetry, the line of best fit through the scatterplot must have a slope of approximately one. This intuition is incorrect, as the formula for the slope coefficient shows.

and validation experiments. In either case we measure estimator performance by calculating the squared distance between estimator predictions and the percentage change in posts in the test subexperiments, on average across all experiments, as in Algorithm 1.

Experiment Splitting Estimators. The experiment splitting estimators all predict the percentage change in posts for the validation subexperiments. The estimators differ in the prediction algorithm used, and the covariates included. The covariates are either the percentage change in posts for the training subexperiments, or the percentage changes in all metrics in the training subexperiments. The predicted values from the estimated models provide the estimated treatment effect for each experiment. We study the following experiment splitting estimators.[10]

- *Univariate OLS*: Linear regression on a constant and the percentage change in posts in the training subexperiments.
- *Univariate Loess*: Locally quadratic regression, using the same regressors as univariate OLS.
- *Multivariate OLS*: Linear regression, using the percentage change in all metrics (posts, and all 23 auxiliary metrics) in the training subexperiments as predictors.
- *Lasso*: Lasso, using the same predictors as multivariate OLS.

In addition, we study the repeated experiment splitting versions of the above estimators, with R, the number of times the training and validation data is repeatedly split, equal to 10. Larger values of R result in only minimal improvements in estimator performance. The tuning parameters required for the loess and lasso models are selected by experiment-level cross-validation on the union of the training and validation data.

Pooling Estimators. The pooling estimators combine training and validation data. Unlike experiment splitting, which uses the validation observations to construct the outcomes and the training observations to construct the predictors, the pooling estimators make no distinction in which sample an observation is from.

- *Unshrunk*: The percentage change in posts in the combined training and validation test group, relative to the combined training and validation control group.
- *James-Stein*: The James-Stein type method of moments estimator described in [9], which allows for heteroskedastic sampling error.
- *Tweedie*: The nonparametric estimator based on equation (1), using Lindsey's method to estimate the density, as in [16].

Figure 5 plots the unshrunk treatment effect estimates against the unshrunk, James-Stein, lasso without repeated splits, and lasso with repeated splits estimators. The splitting estimators are calculated with $\gamma = 0.2$. By construction, the unshrunk points lie on the 45-degree line. The other estimators tend to shrink the unshrunk estimates towards the overall mean, but do so in different ways. James-Stein is an almost linear function of the unshrunk estimates, with a slope of less than one.[11] The lasso estimators are not smooth functions of the unshrunk estimator, reflecting the influence of auxiliary metrics. They sometimes shrink more and sometimes less

[10] Random forests and gradient boosted decision trees perform worse than multivariate OLS and lasso, and are omitted from the results.

[11] The James-Stein estimates are not exactly linear in the unshrunk estimates, because of the adjustment for heteroskedasticity, as in [9].

Figure 5: Unshrunk vs. Other Estimators, $\gamma = 0.2$

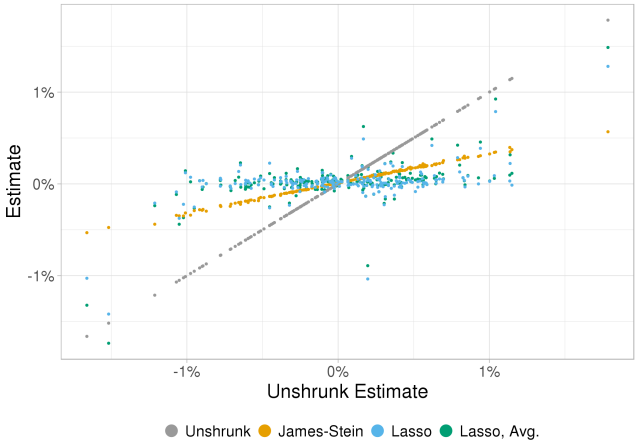
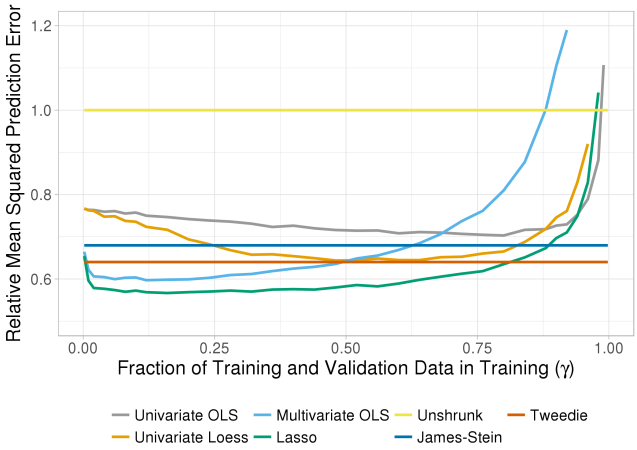


Figure 6: Pooled and Single Split Estimators, Error



than the James-Stein estimates, but on average both have larger variance than James-Stein.

Figure 6 shows the prediction errors for the basic experiment splitting estimators, and Figure 7 shows the prediction errors for the experiment splitting estimators with averaging over repeated splits. Both figures also show the three pooling estimators. The prediction errors are calculated as in Algorithm 1, with the number of splitting simulations S set to 100. The unshrunk estimator performs quite poorly. James-Stein shrinkage reduces mean squared prediction error (MSPE) by 32% relative to the unshrunk case. The non-normal distribution of effect sizes in Figure 3 suggests that Tweedie's estimator, which can adapt to a non-normal distribution of the truth, θ_i, should outperform James-Stein. This is indeed the case, with Tweedie having a 6% lower MSPE than James-Stein.

The unaveraged estimators, with the exception of univariate OLS, all perform better than James-Stein for an appropriately chosen

Figure 7: Pooled and Repeated Split Estimators, Error

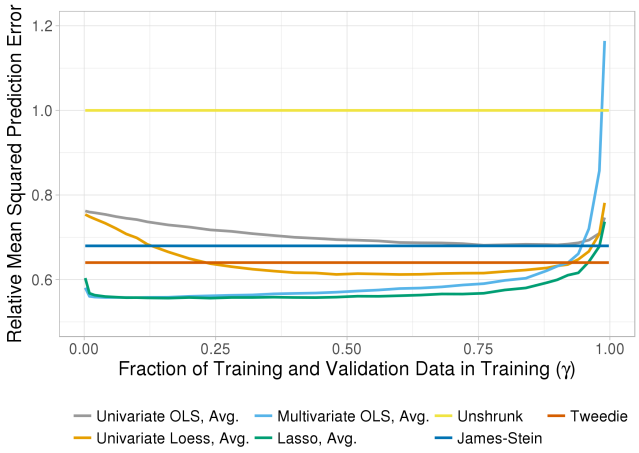


Table 1: Relative MSPE and Standard Errors, $\gamma = 0.24$

Estimator	Relative MSPE	Std. Error
Unshrunk	1.000	0.009
James-Stein	0.680	0.006
Tweedie	0.640	0.006
Univariate OLS	0.739	0.006
Univariate OLS, Avg.	0.717	0.006
Univariate Loess	0.682	0.006
Univariate Loess, Avg.	0.638	0.006
Multivariate OLS	0.603	0.006
Multivariate OLS, Avg.	0.562	0.005
Lasso	0.570	0.005
Lasso, Avg.	0.556	0.005

value of γ. Of particular note are the multivariate models, multivariate OLS and lasso. One might expect the comovements in auxiliary metrics to be informative: intuitively, if a experiment shows a large movement on the metric of interest, but not on auxiliary metrics which tend to covary with the metric of interest, we might infer that the large movement is likely a false positive due to sampling error. Indeed, the multivariate OLS and lasso estimators perform well. At their optimal splitting fractions they improve on the unshrunk estimator by 40% and 43%, and on Tweedie by 7% and 11%. This implies that our inferences about changes in posts can be much improved by taking into account how *other* metrics have changed, because of the correlation between the true treatment effects for posts and for other metrics. Thus experimental analyses which analyze one metric in isolation, without taking into account the comovements in other, related metrics may be quite suboptimal.

The MSPEs of the estimators which average over repeated splits are uniformly lower than their unaveraged counterparts. Repeated splitting reduces the minimum MSPE of multivariate OLS and lasso by 7% and 2%. The best performing estimator is the lasso averaged over repeated splits. It improves on the unshrunk estimator's MSPE

Table 2: Cumulative Treatment Effects, $\gamma = 0.24$

Estimator	Top 1%	Top 5%	Top 10%	All Positive
Unshrunk	0.81%	2.63%	4.36%	11.20%
Lasso, Avg.	2.34%	6.27%	6.83%	15.66%

by 44%, on James-Stein by 18%, and on Tweedie by 13%, and its minimum MSPE is reached at $\gamma = 0.24$. Table 1 shows all estimators' MSPE relative to the unshrunk estimator at $\gamma = 0.24$, along with standard errors summarizing statistical uncertainty caused by the finite number of splitting simulations, $S = 100$.

Figure 6 also shows that choosing precisely the right γ for the multivariate repeated experiment splitting estimators is not necessary in order to obtain large performance improvements relative to the pooling estimators. This is especially true for the repeated lasso splitting estimator, as both averaging and the regularization of the lasso penalty have the effect of reducing the sensitivity of MSPE to the splitting fraction, γ. The repeated splitting lasso estimator has a roughly constant error for $\gamma \in [0.05, 0.75]$. By contrast the performance of the univariate OLS and loess estimators are considerably more sensitive with respect to γ, and their optimal γ is larger.

Some distinctive patterns in Figures 6 and 7 are the relatively flat MSPE curves for the lasso, and the observation that repeated splitting does not much decrease the lasso's minimal RMPSE. These features are consistent with a data generating process in which lifts in the most important covariates are precisely estimated, for any reasonable sample size. This implies that for γ above a small threshold, those covariate values will change little. If the error in the outcome variable is not too large (γ is not too close to one), the lasso gives high quality predictions, which will vary little both as a function of γ, and across repeated experiment splits.

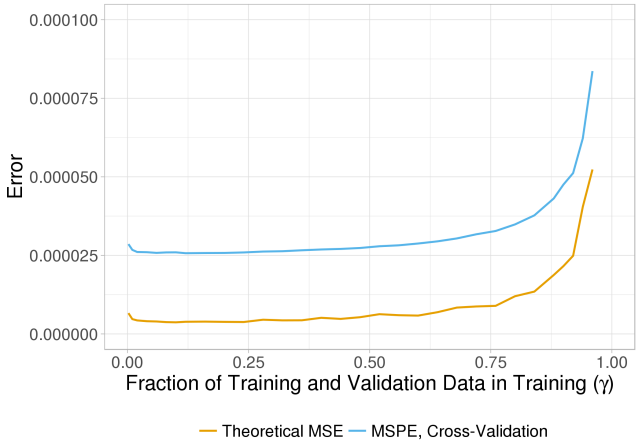
The purpose of A/B tests is typically to identify and launch the best performing treatments. Consequently MSPE may not be the most relevant evaluation criterion for practitioners. Instead, the key desideratum for an estimator may be whether the experiments it ranks highest do indeed have relatively large treatment effects. We can assess this by selecting the most promising experiments, as ranked by a given estimator, and evaluating their performance on the out-of-sample, test subexperiments. Table 2 shows that repeated lasso, the best performing estimator by MSPE, also performs well according to this criterion. The top 10% of experiments, as ranked by the unshrunk estimator, have a cumulative treatment effect of 4.36%.[12] The corresponding figure for the lasso is 57% higher, at 6.83%. The relative difference between the lasso and unshrunk estimators are even more pronounced in the far right tail, as the columns for the top 5% and top 1% of experiments show. Launching all estimates with a positive estimated treatment effect gives an estimated cumulative lift of 11.20% for the unshrunk estimator, while the figure for the lasso is 40% higher, at 15.66%.[13]

Proposition 1 allows us to estimate the mean squared error of the multivariate OLS experiment splitting estimator as a function of γ.

[12]The James-Stein estimator performs very similarly to the unshrunk estimator, as it results in almost the same ordering of experiments.

[13]The repeated lasso outperforms all other experiment splitting estimates according to all criteria in Table 2 except "Top 10%", where it is second to repeated multivariate OLS, which has a marginally higher cumulative treatment effect of 6.84%.

Figure 8: Multivariate OLS Error Estimates



This can be compared to the prediction error of multivariate OLS as computed by the cross-validation procedure of Algorithm 1. The latter is larger than the former for all γ, as it measures prediction error relative to the *estimated effects* in the test subexperiments, rather than error relative to the truth. However, Figure 8 indicates that estimating the optimal γ according to Proposition 1 performs comparably to cross-validation.

6 CONCLUSION

Inference on a particular metric and experiment can generally be improved by borrowing strength both across other experiments and other metrics. Experiment splitting provides a easily implementable methodology for realizing these performance gains, by transforming the problem of developing an appropriate shrinkage estimator into a simple prediction problem. In our sample of Facebook News Feed A/B tests, experiment splitting outperforms unshrunk estimators and conventional shrinkage estimators, especially after incorporating information from comovements in other metrics related to the outcome of interest. While our focus has been on average treatment effects in the context of experimentation, analogous "shrinkage via sample splitting" techniques are likely to be useful in causal inference more generally, with potential applications to estimating quantile treatment effects, local average treatment effects given instrumental variables, and average treatment effects under unconfoundedness.

ACKNOWLEDGMENTS

The authors are grateful to Peter Aronow, Susan Athey, Eytan Bakshy, Matt Goldman, Alex Peysakhovich, Jas Sekhon, Sean Taylor, Stefan Wager and Ark Zhang for useful comments and discussions.

APPENDIX

Proof of Proposition 1.

We have $\sum_{i=1}^{N_E} E(\theta_i - X_i'\widehat{\beta})^2 = N_E\sigma_{\varepsilon_2}^2 + \sum_{i=1}^{N_E} E(\widehat{\theta}_{i,2} - X_i'\widehat{\beta})^2 - 2\sum_{i=1}^{N_E} E\left(\varepsilon_{i,2}(\widehat{\theta}_{i,2} - X_i'\widehat{\beta})\right)$. Collect the $\widehat{\theta}_{i,2}$ and $\varepsilon_{i,2}$ observations across experiments i into the N_E-vectors $\widehat{\boldsymbol{\theta}}_2$ and $\boldsymbol{\varepsilon}_2$. Define the matrix $M = I - X(X'X)^{-1}X'$, with elements m_{ij}. Then

$$\begin{aligned} E\sum_{i=1}^{N_E} \varepsilon_{i,2}(\widehat{\theta}_{i,2} - X_i'\widehat{\beta}) &= E(\boldsymbol{\varepsilon}_2' M \widehat{\boldsymbol{\theta}}_2) \\ &= E\left\{E\left(\sum_i m_{ii}\varepsilon_{i,2}\widehat{\theta}_{i,2} \mid M\right)\right\} \\ &= E\left\{\sum_i m_{ii}\sigma_{\varepsilon_2}^2\right\} \\ &= \sigma_{\varepsilon_2}^2 E(\mathrm{tr}(M)) \\ &= (N_E - l)\sigma_{\varepsilon_2}^2. \end{aligned}$$

Hence $\sum_{i=1}^{N_E} E(\theta_i - X_i'\widehat{\beta})^2 = \left(\sum_{i=1}^{N_E} E(\widehat{\theta}_{i,2} - X_i'\widehat{\beta})^2\right) - (N_E - 2l)\sigma_{\varepsilon_2}^2$.

Proof of Corollary 1.

From standard results on the relation between the variance of regression residuals and the variance of regression errors (e.g. [25]),

$$\sum_{i=1}^{N_E} E(\widehat{\theta}_{i,2} - \widehat{\beta}_0 - \widehat{\beta}_1\widehat{\theta}_{i,1})^2 = (N_E - 2)E(\widehat{\theta}_{i,2} - \beta_0 - \beta_1\widehat{\theta}_{i,1})^2.$$

Further,

$$\begin{aligned} E(\widehat{\theta}_{i,2} - \beta_0 - \beta_1\widehat{\theta}_{i,1})^2 &= Var(\widehat{\theta}_{i,2}) - \frac{Cov^2(\widehat{\theta}_{i,2}, \widehat{\theta}_{i,1})}{Var(\widehat{\theta}_{i,1})} \\ &= \sigma_\theta^2 + \sigma_{\varepsilon_2}^2 - \frac{\sigma_\theta^4}{\sigma_\theta^2 + \sigma_{\varepsilon_1}^2}. \end{aligned} \tag{3}$$

The first equality follows from $\beta_1 = Cov(\widehat{\theta}_{i,2}, \widehat{\theta}_{i,1})Var(\widehat{\theta}_{i,1})^{-1}$, and the second from independence of sampling errors and treatment effects. Combining equation (3) and Proposition 1 gives the result.

Proof of Proposition 2.

With fractions γ and $1-\gamma$ of data in the training and validation subexperiments, we have $\sigma_{\varepsilon_1}^2 = \sigma_\varepsilon^2/\gamma$ and $\sigma_{\varepsilon_2}^2 = \sigma_\varepsilon^2/(1-\gamma)$. Define

$$h(\gamma, \sigma_\varepsilon^2, \sigma_\theta^2, N_E) = \frac{(N_E - 2)\sigma_\theta^2\sigma_\varepsilon^2/\gamma}{\sigma_\theta^2 + \sigma_\varepsilon^2/\gamma} + \frac{\sigma_\varepsilon^2}{(1-\gamma)}.$$

By Corollary 1, the problem of minimizing the mean squared error is $\min_{\gamma\in[\underline{\gamma},\overline{\gamma}]} h(\gamma, \sigma_\varepsilon^2, \sigma_\theta^2, N_E)$. Some calculations show

$$\frac{\partial h}{\partial \gamma} = \sigma_\varepsilon^2\left(\frac{1}{(1-\gamma)^2} - \frac{(N_E - 2)\sigma_\theta^4}{(\sigma_\varepsilon^2 + \gamma\sigma_\theta^2)^2}\right), \tag{4}$$

$$\frac{\partial^2 h}{\partial \gamma^2} = 2\sigma_\varepsilon^2\left(\frac{(N_E - 2)\sigma_\theta^6}{\left(\sigma_\varepsilon^2 + \gamma\sigma_\theta^2\right)^3} + \frac{1}{1-\gamma^3}\right). \tag{5}$$

From equation (5), $\frac{\partial^2 h}{\partial \gamma^2} > 0$. By convexity of h in γ, if $\frac{\partial h}{\partial \gamma}\Big|_{\gamma=\overline{\gamma}} < 0$ then $\gamma^* = \overline{\gamma}$, and if $\frac{\partial h}{\partial \gamma}\Big|_{\gamma=\underline{\gamma}} > 0$ then $\gamma^* = \underline{\gamma}$. Otherwise γ^* solves $\frac{\partial h}{\partial \gamma}\Big|_{\gamma=\gamma^*} = 0$. The limiting behavior of γ^* follows from these observations and equation (4). For large N_E, $\frac{\partial h}{\partial \gamma}\Big|_{\gamma=\overline{\gamma}} < 0$ and so $\lim_{N_E\to\infty} \gamma^* = \overline{\gamma}$. Arguments for limits wrt σ_θ^2 and σ_ε^2 are similar.

REFERENCES

[1] Hirotogu Akaike. 1998. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*. Springer, 199–213.

[2] Michael L Anderson and Jeremy Magruder. 2017. *Split-sample strategies for avoiding false discoveries.* Technical Report. National Bureau of Economic Research.

[3] Susan Athey and Guido Imbens. 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* 113, 27 (2016), 7353–7360.

[4] Susan Athey, Julie Tibshirani, and Stefan Wager. 2016. Generalized random forests. *arXiv preprint arXiv:1610.01271* (2016).

[5] Susan Athey and Stefan Wager. 2017. Efficient policy learning. *arXiv preprint arXiv:1702.02896* (2017).

[6] Eduardo M Azevedo, Alex Deng, Jose Luis Montiel Olea, Justin Rao, and E Glen Weyl. 2018. The A/B Testing Problem. In *Proceedings of the 2018 ACM Conference on Economics and Computation*. ACM, 461–462.

[7] Gerard Biau. 2012. Analysis of a random forests model. *Journal of Machine Learning Research* 13, Apr (2012), 1063–1095.

[8] Thomas Blake and Dominic Coey. 2014. Why marketplace experimentation is harder than it seems: The role of test-control interference. In *Proceedings of the fifteenth ACM conference on Economics and computation*. ACM, 567–582.

[9] Lawrence D Brown. 2008. In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies. *The Annals of Applied Statistics* (2008), 113–152.

[10] Lawrence D Brown and Eitan Greenshtein. 2009. Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *The Annals of Statistics* (2009), 1685–1704.

[11] Bradley P Carlin and Thomas A Louis. 2010. *Bayes and empirical Bayes methods for data analysis*. Chapman and Hall/CRC.

[12] George Casella. 1985. An introduction to empirical Bayes data analysis. *The American Statistician* 39, 2 (1985), 83–87.

[13] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21, 1 (2018), C1–C68.

[14] Victor Chernozhukov, Whitney Newey, and James Robins. 2018. Double/debiased machine learning using regularized Riesz representers. *arXiv preprint arXiv:1802.08667* (2018).

[15] Alex Deng. 2015. Objective bayesian two sample hypothesis testing for online controlled experiments. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, 923–928.

[16] Bradley Efron. 2011. Tweedie's formula and selection bias. *J. Amer. Statist. Assoc.* 106, 496 (2011), 1602–1614.

[17] Bradley Efron. 2012. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*. Cambridge University Press.

[18] Bradley Efron and Trevor Hastie. 2016. *Computer age statistical inference*. Cambridge University Press.

[19] Bradley Efron and Carl Morris. 1973. Stein's estimation rule and its competitors–an empirical Bayes approach. *J. Amer. Statist. Assoc.* 68, 341 (1973), 117–130.

[20] Bradley Efron and Carl Morris. 1975. Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.* 70, 350 (1975), 311–319.

[21] Bradley Efron, Carl Morris, et al. 1976. Multivariate empirical Bayes and estimation of covariance matrices. *The Annals of Statistics* 4, 1 (1976), 22–32.

[22] Bradley Efron, Robert Tibshirani, John D Storey, and Virginia Tusher. 2001. Empirical Bayes analysis of a microarray experiment. *Journal of the American statistical association* 96, 456 (2001), 1151–1160.

[23] Marcel Fafchamps and Julien Labonne. 2017. Using Split Samples to Improve Inference on Causal Effects. *Political Analysis* 25, 4 (2017), 465–482.

[24] Jerome Friedman, Trevor Hastie, Holger Höfling, Robert Tibshirani, et al. 2007. Pathwise coordinate optimization. *The Annals of Applied Statistics* 1, 2 (2007), 302–332.

[25] F. Hayashi. 2011. *Econometrics*. Princeton University Press.

[26] William James and Charles Stein. 1961. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, Vol. 1. 361–379.

[27] Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. 2013. Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1168–1176.

[28] Colin L Mallows. 1973. Some comments on Cp. *Technometrics* 15, 4 (1973), 661–675.

[29] Whitney K Newey and James R Robins. 2018. Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138* (2018).

[30] Alexander Peysakhovich and Dean Eckles. 2018. Learning Causal Effects From Many Randomized Experiments Using Regularized Instrumental Variables. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 699–707.

[31] Herbert Robbins. 1956. An Empirical Bayes Approach to Statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California.

[32] Charles M Stein. 1956. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proc. Third Berkeley Symp. Math. Statist. Probab., 1956*, Vol. 1. Univ. California Press, 197–206.

[33] Charles M Stein. 1962. Confidence sets for the mean of a multivariate normal distribution. *Journal of the Royal Statistical Society. Series B (Methodological)* (1962), 265–296.

[34] Charles M Stein. 1981. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics* (1981), 1135–1151.

[35] William E Strawderman. 1971. Proper Bayes minimax estimators of the multivariate normal mean. *The Annals of Mathematical Statistics* 42, 1 (1971), 385–388.

[36] Mark J Van der Laan and Sherri Rose. 2011. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media.

[37] Stefan Wager and Susan Athey. 2017. Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* just-accepted (2017).

[38] Stefan Wager, Wenfei Du, Jonathan Taylor, and Robert J Tibshirani. 2016. High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences* 113, 45 (2016), 12673–12678.