Sparse Estimation and Uncertainty with Application to Subgroup Analysis[*]

Marc Ratkovic[†] Dustin Tingley[‡]

First Draft: March 2015
This Draft: October 20, 2016

Abstract

We introduce a Bayesian method, LASSOplus, that unifies recent contributions in the sparse modeling literatures, while substantially extending pre-existing estimators in terms of both performance and flexibility. Unlike existing Bayesian variable selection methods, LASSOplus both selects and estimates effects while returning estimated confidence intervals for discovered effects. Furthermore, we show how LASSOplus easily extends to modeling repeated observations and permits a simple Bonferroni correction to control coverage on confidence intervals among discovered effects. We situate LASSOplus in the literature on how to estimate subgroup effects, a topic that often leads to a proliferation of estimation parameters. We also offer a simple pre-processing step that draws on recent theoretical work to estimate higher-order effects that can be interpreted independently of their lower-order terms. A simulation study illustrates the method's performance relative to several existing variable selection methods. In addition, we apply LASSOplus to an existing study on public support for climate treaties to illustrate the method's ability to discover substantive and relevant effects. Software implementing the method is publicly available in the **R** package `sparsereg`.

Key Words: subgroup analysis, LASSO, Bayesian LASSO, conjoint analysis, heterogeneous treatment effects

[*] We are grateful to Neal Beck, Scott de Marchi, In Song Kim, John Londregan, Luke Miratrix, Michael Peress, Jasjeet Sekhon, Yuki Shiraito, Brandon Stewart, and Susan Athey for helpful comments on an earlier draft. Earlier versions presented at the 2015 Summer Methods Meeting, Harvard IQSS Applied Statistics Workshop, Princeton Political Methodology Colloquium, DARPA/ISAT Conference "What If? Machine Learning for Causal Inference", and EITM 2016. We are also grateful to two anonymous reviewers for insanely detailed feedback on an earlier version. All mistakes are because of the authors.

[†] Assistant Professor, Department of Politics, Princeton University, Princeton NJ 08544. Phone: 608-658-9665, Email: ratkovic@princeton.edu, URL: http://www.princeton.edu/∼ratkovic

[‡] Professor of Government, Harvard University, Email: dtingley@gov.harvard.edu, URL: scholar.harvard.edu/dtingley

1 Introduction

Social scientists regularly confront situations that require some type of model selection. This may include selecting the shape of the relationship between an explanatory variable and an outcome. Or, as is the focus in this paper, it may involve "sparse modeling," an estimation technique that zeroes out all but the most relevant of variables from hundreds or thousands of possible candidates. Earlier approaches to these situations, such as step-wise regression, have in recent years been replaced by new tools, including those provided by the growing machine learning literature.[1] Many of these tools use some form of "regularization" or "sparsity" where estimates can be shrunk or removed from the model according to a threshholding rule. As one applied example of variable selection, a *subgroup analysis* involves estimating which combinations of treatments are most (or least) efficacious, and for which observations (Berry, 1990). As experimental designs grow more complex (e.g., Hainmueller, Hopkins and Yamamoto, 2014), the number of candidate subgroup effects has grown as well. In turn, social scientists have grown increasingly interested in methods for uncovering subgroup effects (e.g. Grimmer, Messing and Westwood, 2014; Imai and Ratkovic, 2013; Green and Kern, 2012), with some of these methods using regularization-based variable selection methods.

In this paper, we introduce LASSOplus,[2] a Bayesian method for variable selection in high dimensional settings. LASSOplus offers two major statistical advances over existing Bayesian shrinkage and selection methods (Park and Casella, 2008; Carvalho, Polson and Scott, 2010; Armagan, Dunson and Lee, 2012; Leng, Tran and Nott, 2014; Griffin and Brown, 2012, 2010; Kang and Guo, 2009). First, LASSOplus is the first *sparse* Bayesian method for variable selection, whereby effects are simultaneously estimated and selected. Second, LASSOplus has been designed to possess desirable theoretical properties. We focus our theoretical attention on cases where the number of possible effects are either growing in or even greater than the sample size. LASSOplus has the Oracle Property described by Fan and Li (2001), which means that it is asymptotically indistinguishable from a model fit only to covariates with an in-truth non-zero effect. It also satisfies an Oracle Inequality (Candes, 2006), so it achieves a prediction error of the same order as least squares fit to the true model. Third, LASSOplus returns approximate confidence intervals. As a Bayesian method, LASSOplus returns credible intervals, but previous methods have found that these credible intervals are more narrow than the confidence intervals (Kyung, Gill, Ghosh, Casella et al., 2010). For the researcher interested in confidence intervals, we return uncertainty intervals that are

[1] For example, kernel regularized least squares (Hainmueller and Hazlett, 2013) and the adaptive LASSO (Kenkel and Signorino, 2012) have been deployed for estimating functional forms while the LASSO (Tibshirani, 1996) has been used for variable selection.

[2] For *p*seudo-*l*ikelihood *u*nbiased *s*elector.

calibrated to achieve nominal coverage (Efron, 2015). Our goal is to produce a method useful to the applied researcher, so our implementation of LASSOplus includes extensions to several commonly encountered data structures and types. The software, which we make publicly available in the **R** programming language, handles binary and truncated outcomes, computes up to three-way random effects, and has both a full Markov Chain Monte Carlo implementation and a faster Expectation Maximization implementation, a useful tool for practical modeling.

To illustrate the usefulness of LASSOplus, we apply the method to subgroup analysis. Existing subgroup analysis methods face several shortcomings. Frequentist tree-based methods identify subgroups, but they do not offer uncertainty estimates nor can they handle experiments with repeated observations (Loh, Heb and Manc, 2015; Foster, Taylor and Ruberg, 2011; Imai and Strauss, 2011; Lipkovich et al., 2011; Su et al., 2009), but see Wager and Athey (2015) for recent work on both fronts. Frequentist variable selection methods also cannot handle repeated observations (Imai and Ratkovic, 2013), and their methods for estimating confidence intervals perform poorly in our simulations (Minnier, Tian and Cai, 2011). Recent work has implemented ensemble or high-dimensional Bayesian methods (Berger, Wang and Shen, 2015; Green and Kern, 2012; Grimmer, Messing and Westwood, 2014). These methods are powerful predictive tools, but they do not point-identify relevant subgroups. They work by fitting either a single black-box model or several different models, and subgroups are identified through an ex post, ad hoc search. The uncertainty estimates (if they are even available) are not guaranteed to have nominal coverage, and implementations of the machine learning methods do not accommodate repeated observations.

We present a simulation study that compares LASSOplus to other cutting edge methods. LASSOplus achieves a false discovery rate lower than that of several existing methods, often dramatically so. The method remains reasonably powerful, and its approximate confidence intervals achieve nominal or near-nominal coverage. We also apply the method to a recent conjoint experiment by Bechtel and Scheve (2013) in which the authors estimate the effects of different features of an international climate change agreement on voter support. The original authors conduct an extensive set of subgroup analyses by running a regression after repeatedly subsetting their data. Our method recovers many of the same effects, avoids the arbitrary nature of subsetting, is implemented in one line of code, and returns uncertainty estimates on each effect that take into account the fact that individuals respond to multiple versions of the experiment.

The structure of the paper reflects our three main goals. First, in Section 2, we introduce readers to core concepts and existing methods for LASSO-based variable selection. Second, Sections 3 and 4 introduce the LASSOplus and state its statistical properties, while Section 5 discusses relevant issues arising during the application and interpretation of the method for subgroup analysis. Third, we illustrate the application of LASSOplus in Section 6 by comparing it to earlier methods using an extensive Monte Carlo simulation study and in Section 7 we apply the method to the case of subgroup analysis by analyzing the data in Bechtel and Scheve (2013). We show how LASSOplus recovers many of the original authors' subgroup results while making minimal modeling decisions. Section 8 concludes with key contributions and discusses future research opportunities.

2 Variable Selection and Shrinkage

Given observed outcome Y_i and vector of K observed covariates X_i on observation $i \in \{1, 2, \ldots, N\}$, researchers will commonly turn to the linear model to connect the two, as

$$Y_i = X_i^\top \beta^o + \epsilon_i \tag{1}$$

where β^o is the population-level vector of parameters associated with each covariate and ϵ_i is the error term, assumed mean-zero and equivariant. We will also assume that Y_i is mean-zero, so $\sum_{i=1}^N Y_i = 0$ and that each element of X_i is scaled to be mean-zero and have a sample standard deviation one, so $\sum_{i=1}^N X_{ik} = 0$ and $\sum_{i=1}^N X_{ik}^2 = N - 1$.

Social scientists are well-trained on how to handle the case where $N >> K$. Valid inference can be conducted using the familiar t- or z-statistics, p-values, and confidence intervals. We work here on a different problem: how to fit this model when there are hundreds or thousands of elements in X_i and return the handful that best explain the outcome. When K is large, least squares estimates are unstable, and when $K > N$, a unique estimate does not even exist.

This setting may at first seem unfamiliar or esoteric, but it is not. Once we consider X_i as consisting of all main effects and two- and three-way interaction terms, even a modest number of variables can produce a large number of covariates (e.g., Gillen et al., 2016). In our application below, we consider data from a conjoint experiment, where main effects plus *treatment* × *covariate* interactions generated 215 possible subgroup effects. Rather than present three or four pages of output from a regression table, we implement LASSOplus, producing 41 non-zero effects. Thus LASSOplus opens the door to allow for models that are saturated with interaction terms, while still returning efficient estimates that can add nuance to the underlying theory.

LASSOplus is an example of a "sparse model." Sparse modeling involves fitting a model that zeroes out all but some small subset of $\widehat{\beta}$. The literature on sparse modeling is large and diverse, so we first introduce several of the the key concepts, contributions, and insights. For ease of exposition, we focus on variable selection in the single-parameter case, where X_i is simply a scalar and the issue at hand is whether or not to set the estimate of β^o to zero. After a brief survey of sparse modeling methods we turn to a description and evaluation of LASSOplus.

2.1 Standard Practice: Variable Selection with One Covariate

For this section, we assume an outcome Y_i and single covariate, X_i, both scaled as described above. We are going to consider variable selection in this simplified setting, with model

$$Y_i = X_i \beta^o + \epsilon_i \tag{2}$$

where the goal is how to decide whether or not to zero out $\widehat{\beta}$. We focus on this simplified setting because it provides analytic results not available in the multivariate setting, and we use these results to convey the basic intuitions of variable selection. Later we return to the multivariate setting.

With a single covariate, variable selection is normally done in two stages: first, the effect is estimated and then some p-value or t-statistic threshold for statistical significance is used to determine whether the effect can be differentiated from zero. A standard estimate for β^o is the least squares estimate,

$$\begin{align} \widehat{\beta}^{LS} &= \underset{\widetilde{\beta}}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{N} (Y_i - X_i \widetilde{\beta})^2 \tag{3} \\ &= \frac{\sum_{i=1}^{N} Y_i X_i}{\sum_{i=1}^{N} X_i^2} = \frac{\sum_{i=1}^{N} Y_i X_i}{N-1} \tag{4} \end{align}$$

which gives a point estimate. The point estimate is then compared to its standard error, $\widehat{\sigma}_{\widehat{\beta}}$:

$$\widehat{\sigma}_{\widehat{\beta}} = \frac{\widehat{\sigma}_\epsilon}{\sqrt{N-1}}; \;\; \widehat{\sigma}_\epsilon = \sqrt{\frac{\sum_{i=1}^{N} (Y_i - X_i \widehat{\beta}^{LS})^2}{N-2}}. \tag{5}$$

If the t-statistic is larger in magnitude than some critical value, normally 1.96, the effect is considered statistically significant. In this framework, estimation and selection occur in two separate steps. Estimation and selection cannot, in fact, be simultaneous: the least squares estimate is never zero, outside of pathological cases.

2.2 LASSO with a Single Covariate

Next we introduce the LASSO of Tibshirani (1996) in the case with a single covariate. LASSO is an acronym for *L*east *A*bsolute *S*hrinkage and *S*election *O*perator. In the one-parameter case, the LASSO estimator is the solution to

$$\widehat{\beta}^L = \underset{\widetilde{\beta}}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{N} (Y_i - X_i \widetilde{\beta})^2 + \lambda |\widetilde{\beta}|. \tag{6}$$

The first term in this objective function is the residual sum of a squares. The second term has two components: the tuning parameter λ, indexed by sample size, and the penalty term $|\widetilde{\beta}|$. The least absolute component references the structure of the penalty, $\lambda|\widetilde{\beta}|$.

Denote the sign of the least squares estimate as $\widehat{s}^L = \text{sgn}(\widehat{\beta}^{LS}) \in \{-1, 1\}$. With one parameter, the LASSO estimate is (Tibshirani, 1996, sec 2.2)

$$\widehat{\beta}^L = \left(\widehat{\beta}^{LS} - \widehat{s}^L \frac{\lambda}{N-1}\right) \mathbf{1}\left(|\widehat{\beta}^{LS}| > \frac{\lambda}{N-1}\right) \tag{7}$$

The first part of Equation 7 illustrates the shrinkage: the estimate is the least squares estimate biased towards zero by $\lambda/(N-1)$. The second part of Equation 7 illustrates the selection component. If $\widehat{\beta}^{LS}$ is less than $\lambda/(N-1)$ in magnitude, then the LASSO estimate is simply zero.

Equation 7 reveals three shortcomings of LASSO estimation. First, λ must be estimated from the data. Researchers commonly turn to an AIC or BIC criterion or to cross-validation to select this tuning parameter. Unfortunately, these three methods may give quite different results, with little theoretical guidance to choose among them. Second, both large effects and small effects are biased towards zero by the same amount, $\lambda/(N-1)$. A more accurate estimator would shrink smaller effects but leave larger effects untouched. The third problem arises due to competing demands on $\lambda/(N-1)$. In the left side of the equation, it is a bias term. As such, we want it to go to zero quickly. In the right side of the equation, though, we see that if $\lambda/(N-1)$ goes to zero *too* quickly, then we will end up not zeroing out any effects. As we show in Appendix C,[3] no LASSO model indexed by a single λ can produce estimates that are sparse, consistent, and asymptotically normal. The LASSOplus method proposed in this paper helps to resolve all three issues.

2.3 Single Covariate Adaptive LASSO

Zou (2006) introduced the adaptive LASSO, which extends the LASSO by including parameter-specific weights that cause differential shrinkage (see also Kenkel and Signorino, 2012). These weights serve to shrink larger effects less than smaller effects. The adaptive LASSO is a two-stage estimator. In the first stage, weights that are inversely related to β are constructed. In the second stage, these weights are used to weight a LASSO problem. The weights are constructed as

$$\widehat{w} = 1/|\widehat{\beta}^1|^\gamma \tag{8}$$

where $\widehat{\beta}^1$ is a first-stage, root-N consistent estimate of β^o and $\gamma > 0$. With these weights, the estimator becomes

$$\widehat{\beta}^L(\lambda) = \underset{\widetilde{\beta}}{\text{argmin}} \frac{1}{2}\sum_{i=1}^{N}(Y_i - X_i\widetilde{\beta})^2 + \lambda\widehat{w}|\widetilde{\beta}|. \tag{9}$$

[3] See also Fan and Li (2001, Remark 1, pp. 1353).

In this setting, the univariate adaptive LASSO estimator is

$$\widehat{\beta}^{aL}(\lambda, w) = \left(\widehat{\beta}^{LS} - s\frac{w\lambda}{N-1}\right)\mathbf{1}\left(\left|\widehat{\beta}^{LS}\right| > \frac{w\lambda}{N-1}\right). \tag{10}$$

For a fixed value of λ, the adaptive LASSO estimator has a lower bias for larger effects than smaller ones. With one effect, the adjustment is trivial, but with hundreds of possible effects, the gains from differential shrinkage can be substantial.

Several problems emerge with adaptive LASSO estimation. The first is that any number of methods can return a root-N consistent estimate of β: least squares, ridge regression, or Bayesian regression models (e.g., Gelman et al., 2008). Second, the decay parameter γ must be either assumed or estimated from the data, which can grow computationally costly. Third, the adaptive LASSO inherits the same uncertainty over tuning parameter selection as the LASSO.

2.4 An Empirical Process Approach with a Single Variable

We turn next to LASSO estimation through an empirical process approach. Early seminal works include Donoho and Johnstone (1994), Candes and Tao (2007), and Bickel, Ritov and Tsybakov (2009); the approach has been recently popularized in economics in work by Victor Chernozhukov and colleagues (Belloni and Chernozhukov, 2013; Chernozhukov, Fernández-Val and Melly, 2013; Belloni et al., 2012; Belloni, Chernozhukov and Hansen, 2011). Harding and Lamarche (2016) have extended this work to estimating indivdual-level heterogeneous effect in quantile regression panel models. We recommend Buhlmann and van de Geer (2013) as a particularly accessible introduction to this line of work.

A central result in this literature is deriving an "Oracle Inequality," a bound showing that for a particular value of λ, the excess risk goes to zero at rate $1/N$, up to a penalty incurred for not knowing the true covariates. In the one-parameter case, this inequality is given below:

Lemma 1 *Oracle Inequality for the LASSO in the Single Parameter Case*

For $\lambda = \sigma \times t \times \sqrt{(N-1)}$*, the single-parameter LASSO estimator satistfies the Oracle Inequality*[4]

$$\frac{1}{N}\left\{\sum_{i=1}\left(X_i(\widehat{\beta}^L - \beta^o)\right)^2 + \lambda\left|\widehat{\beta}^L - \beta^o\right|\right\} \leq \frac{C_{L1}\sigma^2 t^2}{N} \tag{11}$$

with probability at least $1 - 2\exp\{-t^2/2\}$.

[4] Note that the Oracle Inequality is distinct from the Oracle Property, which we discuss below. Across the literature, an estimator that satisfies either is called an "oracle estimator," so we will be clear as to which we are discussing in each section. We will discuss the two different concepts after introducing the Oracle Property below.

We denote as t the variable by which we control the probability of the bound holding, i.e. the variable t enters into both the bound $C_{L1}\sigma^2 t^2/(N-1)$ and the probability of it holding $1 - 2\exp\{-t^2/2\}$. We will use $C_{\cdot}$ to denote constants that do not change in N or K.

Though we state the result with $K = 1$, this approach is most useful in the $K > N$ setting, a point to which we return below.

2.5 The Multivariate LASSO and its Variants

We now turn to the multivariate setting, where $X_i^\top$ is a vector of K observation-level covariates, $i \in \{1, 2, \ldots, N\}$. These covariates may include the values of treatment variables, pre-treatment moderators, and interactions within and between the two. We assume X_i is of length K, where K is a finite but possibly large number, say in the hundreds or thousands. The k^{th} element of X_i and β are X_{ik} and β_k, and we also assume all fourth moments of $[Y_i X_i^\top]$ exist.[5] We assume the data are generated as

$$Y_i = X_i^\top \beta^o + \epsilon_i \tag{12}$$

where we desire a sparse representation of the K-dimensional vector β. We are *not* assuming that elements of β^o are zero, but that some effects are of negligible magnitude.[6] Instead, we are seeking the best representation of the model in which most effects are estimated as zero, so as to allow the researcher to focus on relevant effects. We return to this point more fully below.

In the multivariate setting, the LASSO estimate is the solution to

$$\widehat{\beta}^L(\lambda) = \underset{\widetilde{\beta}}{\operatorname{argmin}} \frac{1}{2}\sum_{i=1}^{N}(Y_i - X_i^\top \widetilde{\beta})^2 + \lambda \sum_{k=1}^{K} |\widetilde{\beta}_k|. \tag{13}$$

The LASSO tends to over-select small effects. Previous research has addressed this problem in two ways. First, first-stage adaptive weights can be incorporated into the estimation. This leads to the adaptive LASSO, given above. Alternatively, small coefficients can be trimmed ex post and OLS fit to the surviving covariates. This approach has been developed in the empirical process framework discussed above and we focus on the popular LASSO+OLS method of Belloni and Chernozhukov (2013) below and in our simulations.

We start with the first. In the multivariate setting, the adaptive LASSO gives each parameter its own weight in the penalty $\widehat{w}_k = 1/|\widehat{\beta}_k^0|^\gamma$, for some $\gamma > 0$ and β_k^0 a root-N consistent estimate.

[5] Notation for sample size and number of covariates varies across literatures, with the number of covariates represented with either p or n. We use K to align with common social science notation.

[6] We are more precise about differentiating "relevant" and "irrelevant" effects in Appendix B.

The model is now

$$\widehat{\beta}^{aL}(\lambda) = \underset{\widetilde{\beta}}{\operatorname{argmin}} \frac{1}{2}\sum_{i=1}^{N}(Y_i - X_i^\top \widetilde{\beta})^2 + \lambda \sum_{k=1}^{K} \widehat{w}_k |\widetilde{\beta}_k|. \tag{14}$$

We move now to the second approach. LASSO+OLS proceeds in two steps. First, a LASSO fit is used to select an initial set of possible effects. Second, OLS is conducted on the all subsets of the initial fit, and the OLS fit with residual variance close to the residual variance of the LASSO fit is selected (Belloni and Chernozhukov, 2013). From our experience, LASSO+OLS serves as a helpful tool for sparse modeling and variable selection, so we include it both in our software package and in our simulations below. We provide a more complete discussion of LASSO+OLS in Appendix H.

These methods, pre-estimation weighting and post-estimation selection, raise several concerns. First, each have tuning parameters that must be selected and, ideally, estimated. As we describe in Appendix H, LASSO+OLS has three tuning parameters that must be selected. The authors provide reasonable defaults and algorithms on how to estimate the tuning parameters, but provide no theoretical guidance as to how to choose them. Similarly, with the adaptive LASSO, the user must choose a method for estimating the first-stage weights and for the exponent in the weights. A second area of concern is in generating confidence intervals. The probabilistic bounds in LASSO+OLS generate confidence intervals (Belloni, Chernozhukov and Hansen, 2011), but require user-tuning of several parameters. For the adaptive LASSO methods, generating confidence intervals is still a field of active research.[7]

2.6 Two Statistical Properties of Sparse Models

We now discuss two statistical properties of sparse models that have been described in the literature. We present these properties and their relationship to existing methods here in order to set up our discussion in Section 4 of how and when LASSOplus has these properties. Though the two properties have similar names, the Oracle Property and Oracle Inequality, they are actually distinct statistical concepts. The Oracle Property requires that, as N grows, the estimator select the correct effects and converge in distribution to OLS using only variables from the true subset. This property was initially cast in the fixed-K, growing-N framework (Fan and Li, 2001; Zou, 2006), but we focus on the Oracle Property in when N and K can both be thought of as growing (see also Fan and

[7] Recent work has proposed resampling methods, through either a bootstrap or repeatedly reweighting observations (Minnier, Tian and Cai, 2011; Chatterjee and Lahiri, 2011). A second strand of research uses normal or truncated-normal approximations to construct confidence sets on selected effects (Berk et al., 2013; Leeb, Potscher and Ewald, 2015). A third has considered variable selection in terms of nested hypothesis testing and generated p-values (Lockhart et al., 2014). Our proposed estimator in Section 3 offers a straightforward way of calculating effect uncertainty and thus offers a clear improvement on existing estimation strategies.

Peng, 2004). The Oracle Inequality is a predictive property, such that the fitted values are close to the fitted values from OLS estimated only on the true subset variables. Satisfying the Oracle Inequality is weaker than satisfying the Oracle Property, as estimating fitted values well is easier than estimating β^o correctly. The key advantage to the Oracle Inequality is that it allows for analysis when $K > N$.

The Oracle Property. The adaptive LASSO satisfies the Oracle Property of Fan and Li (2001, p. 1353, Thm 2). An Oracle estimator must satisfy two conditions. First, it must be consistent for variable selection, so in-truth non-zero effects are selected and in-truth zero effects are zeroed out. Second, the estimate has to be consistent and asymptotically normal, with variance equal to a model fit only to covariates associated with the non-zero effects. In other words, the estimator achieves the same asymptotic performance as an identical estimator told *ex ante* by an "oracle" which effects are zero and which are not. We give a formal definition in Definition 1.

Definition 1 **The Oracle Property**

Assume the model $Y_i = X_i^\top \beta^o + \epsilon_i$ *with* ϵ_i *mean zero with four finite moments. Let the set* S *denote the set of indices of in-truth non-zero elements of* β*, i.e.* $S = \{k : k \in \beta_k^o \neq 0\}$.

An Oracle estimator $\widehat{\beta}^{oracle}$ *has the following two properties (Zou, 2006, p. 1418):*

1. *Consistent Variable Selection:* $\lim_{N\to\infty}\{k : \widehat{\beta}_k^{oracle} \neq 0\} = S$
2. *Optimal Estimation Rate:* $\sqrt{N}\left(\widehat{\beta}_S^{oracle} - \beta_S^o\right) \xrightarrow{\text{d}} \mathcal{N}(\mathbf{0}_{|S|}, \Sigma_S^*)$ *where* Σ_S^* *is the asymptotic variance matrix from the true subset model.*

Satisfying the Oracle Property is desirable because it offers asymptotic efficiency gains over the normal least squares estimator.[8] Denote the asymptotic relative efficiency of two estimators of vector θ, $\widehat{\theta}_1$ and $\widehat{\theta}_2$, as

$$ARE_\theta(\widehat{\theta}_1, \widehat{\theta}_2) = \lim_{N\to\infty} \frac{\mathbb{E}\left\{||\widehat{\theta}_1 - \theta||_2^2\right\}}{\mathbb{E}\left\{||\widehat{\theta}_2 - \theta||_2^2\right\}} \tag{15}$$

The asymptotic relative efficiency of an estimator with the Oracle Property will never perform worse than the least squares estimate, a result we state below:

[8] As with any estimator, disagreement exists over the desirability of Oracle estimators. On the one hand, Oracle estimators reduce the tendency of the LASSO to return a large number of false positives with small coefficient estimates. We find this in our simulations below. On the other hand, a critique offered by Leeb and Potscher (2008), but first acknowledged in Fan and Li (2001, p. 1348, 1353), showed that Oracle estimators that the superefficiency property comes at the cost of losing uniform convergence. We refer the reader to Leeb and Potscher (2008) for more details, but note that LASSOplus has both an Oracle and non-Oracle implementation in Section 3.1.

Proposition 1 *Assume the least squares estimator exists and is unique. Then, an estimator with the Oracle Property is at least as efficient as the non-Oracle least squares estimator, asymptotically:*

$$ARE_{\beta^o}(\widehat{\beta}^{LS}, \widehat{\beta}^{Oracle}) \geq 1 \tag{16}$$

with equality if and only if none of the elements of β^0 are 0.

Proof: *See Appendix A.*

We note that the Oracle Property has been extended by Fan and Peng (2004). The authors consider the general case of a penalized likelihood and give conditions for the estimates to satisfy the Oracle Property. Among these are that the bias induced by shrinkage disappear and that the likelihood eventually swamp the prior. The key assumption, from an applied standpoint, is that $K^5/N \to 0$. Of course this holds when K is fixed. When K is allowed to grow, the result illustrates that the Oracle Property–basically, getting the model right pointwise and in distribution–requires quite a bit of data.[9]

The Oracle Inequality. Again, we take as our benchmark the least squares estimator fit to only the in-truth-nonzero effects, with the goal of producing an estimator that has similar properties. The least squares estimator, fit only to true covariates, achieves a predictive risk of order $\mathbb{E}\left(\frac{1}{N}||X_S(\beta^o_S - \widehat{\beta}^{LS}_S)||_2^2\right) = \sigma^2|S|/N$, where $|S|$ denotes the number of in-truth non-zero effects. An Oracle Inequality bounds the predictive order at a rate going to zero as $1/N$, so it performs comparably to OLS on the true subset. Often in these inequalities, though, the bound will include a penalty that grows in K, as the true model is not known in advance.

The Oracle Inequality in the multivariate case requires two additional constructions. Denote as X_S the subset of X corresponding with in-truth non-zero effects and denote as ϕ_o the smallest eigenvalue of $\frac{1}{N}\sum_{i=1}^N X_S^\top X_S$. The Compatibility Condition holds when $\phi_o > 0$.[10]

Proposition 2 *Oracle Inequality for the LASSO in the Multivariate Case*

For

$$\lambda = C_\epsilon \widehat{\sigma}\sqrt{(t + \log(K)) \times (N-1)} \tag{17}$$

the LASSO estimator satistfies the Oracle Inequality

$$\frac{1}{N}\left\{\left|\left|X_i(\widehat{\beta}^L - \beta^o)\right|\right|_2^2 + \lambda\left|\left|\widehat{\beta}^L - \beta^o\right|\right|_1\right\} \leq \frac{C_L\widehat{\sigma}^2\,(t + \log(K))}{\phi_0^2 N} \tag{18}$$

[9] For example, with 10 covariates, the Oracle Property would require N to be of order $10^5 = 100,000$.

[10] The assumption shows up under several formulations. For example, the Restricted Eigenvalue assumption of Bickel, Ritov and Tsybakov (2009), that all submatrices in X of size $|S|$ are full rank and all components $X_{S^\complement}$ are linearly independent of X_S.

with probability at least $1 - 2\exp\{-t^2/2\} - \Pr(C_\epsilon\widehat{\sigma} \leq \sigma)$.

Corollary *The LASSO risk is consistent for the population risk when* $\log(K)/N \to 0$.

Proof: *See Buhlmann and van de Geer (2013, ch. 6.2)*

We focus on two insights from the Oracle Inequality. First, it is achieved when λ is of order $\sqrt{N\log(K)}$ and actually allows for closed feasible estimation of the tuning parameter. The results also highlights that we are paying a "cost" of $\log(K)$ for not knowing the true model in advance. The second is that the requirements are quite mild for consistency. For K covariates, we only need the sample size N to be of order $\log(K)$. For example, then, going from $K = 100$ to $K = 200$ would only require a 15% increase in sample size ($= \log(200)/\log(100)$) to maintain the Oracle Inequality bound. Relative to the Oracle Property, we see that the Oracle Inequality requires a good bit less data, but it also guarantees less than the Oracle Property.

2.7 The Bayesian LASSO

LASSOplus is a Bayesian estimator and thus with the above review of recent frequentist based LASSO methods in mind, we turn to the existing Bayesian LASSO literature. In a Bayesian framework, the LASSO can be interpreted as the maximum a posteriori (MAP) estimate of a model with a double-exponential prior $\Pr(\beta_j|\lambda) = \frac{1}{2\lambda}\exp(-\lambda|\beta_j|) = DE(\lambda)$. The Bayesian LASSO model of Park and Casella (2008, hereafter PC)[11], can be written as

$$Y_i|X_i, \beta, \sigma^2 \sim \mathcal{N}(X_i^\top\beta, \sigma^2) \tag{19}$$

$$\beta_k|\lambda, \sigma^2 \sim DE(\lambda/\sigma) \tag{20}$$

$$\lambda^2 \sim \Gamma(\delta, \rho) \tag{21}$$

where we denote as $DE(a)$ the double exponential density $f(x; a) = \frac{a}{2}\exp(-a|x|)$. PC show that parameterizing the prior on β_k with λ/σ instead of λ ensures a unimodal posterior. The prior on the tuning parameter is over λ^2 rather than λ in order to maintain conjugacy in the augmented model, given below. Any positive value for the shape (δ) and rate parameters (ρ) will give a proper prior; PC take $(\delta, \rho) = (1, 1.78)$. PC complete the hierarchy by assuming $\Pr(\sigma^2) \propto 1/\sigma^2$, though any gamma prior on $1/\sigma^2$ will maintain conjugacy.

The posterior mode of the Bayesian LASSO is a LASSO estimator. The negative log-posterior

[11] See also Hans (2009); Kyung, Gill, Ghosh, Casella et al. (2010).

of β under this model is, up to an additive constant that does not depend on β,

$$-\log\left(\Pr\left(\beta|\lambda, \sigma^2, \mathcal{D}_N\right)\right) = \frac{1}{\sigma^2}\left\{\frac{1}{2}\sum_{i=1}^{N}(Y_i - X_i^\top\beta)^2 + \lambda\sigma\sum_{k=1}^{K}|\beta_k|\right\} \tag{22}$$

$$\propto \frac{1}{2}\sum_{i=1}^{N}(Y_i - X_i^\top\beta)^2 + \lambda\sigma\sum_{k=1}^{K}|\beta_k| \tag{23}$$

Factoring out $\frac{1}{\sigma^2}$ reveals the posterior mode of the PC model is exactly a LASSO estimate with tuning parameter $\widetilde{\lambda} = \lambda\sigma$.

Shrinkage Priors and Scale Mixtures. The normal likelihood above is not conjugate with the double-exponential prior on β. In order to restore conjugacy, PC augment the parameter space by representing the double-exponential distribution as a scale mixture of normals with exponential mixing density (see also West, 1987):

$$\frac{\widetilde{\lambda}}{2}e^{-\widetilde{\lambda}|\beta_k|} = \int_0^\infty \frac{1}{\sqrt{2\pi\tau_k^2}}e^{-\beta_k^2/(2\tau_k^2)}\frac{\widetilde{\lambda}^2}{2}e^{-\widetilde{\lambda}^2\tau_k^2/2}d\left(\tau_k^2\right). \tag{24}$$

This suggests the following augmented representation of the double exponential prior:

$$\beta_k \sim DE(\lambda/\sigma) \Rightarrow \beta_k|\tau_k^2, \sigma^2 \sim \mathcal{N}(0, \tau_k^2\sigma^2);\;\; \tau_k^2 \sim \exp(\lambda^2/2) \tag{25}$$

Under the augmented parameterization, the likelihood and prior for β are both normal and hence conjugate. Let $D_\tau = \text{diag}(\tau^2)$ and $A = \left(X^\top X + D_\tau^{-1}\right)^{-1}$. The Gibbs updates are:

$$\beta|\cdot \sim \mathcal{N}(AXY, \sigma^2 A) \tag{26}$$

$$\sigma^2|\cdot \sim \text{InvGamma}\left((N-1)/2 + K/2, \frac{1}{2}\left\{\sum_{i=1}^{N}(Y_i - X_i^\top\beta)^2 + \sum_{k=1}^{K}\beta_k^2/\tau_k^2\right\}\right) \tag{27}$$

$$1/\tau_k^2|\cdot \sim \text{InvGaussian}\left(\lambda\sigma/|\beta_k|, \lambda^2\right) \tag{28}$$

$$\lambda^2|\cdot \sim \Gamma(K+\delta, \sum_{k=1}^{K}\tau_k^2/2 + \rho) \tag{29}$$

where $\Gamma(a, b)$ denotes a Gamma distribution with shape parameter a and rate parameter b.

Additional methods have implemented different mixing densities within the scale mixture representation of shrinkage priors (Polson and Scott, 2012). The "horseshoe prior" of Carvalho, Polson and Scott (2010) is

$$\beta_k|\lambda, \lambda_k \sim \mathcal{N}(0, \lambda_k^2\lambda^2) \tag{30}$$

$$\lambda_k \sim C^+(0, s) \tag{31}$$

where s is taken as either 1 or σ^2 and $C^+(a, b)$ denotes the half-Cauchy distribution (Gelman, 2006). The model is so-named because the posterior density places most of its mass at either no shrinkage or full shrinkage, giving the posterior density a horseshoe shape. The horseshoe prior has proven to be an excellent default choice in sparse modeling, so we include it in our simulation study below.[12]

2.8 Shrinkage without Selection

The Bayesian estimators have shown better performance, in terms of mean-squared error and prediction, than their frequentist counterparts; see Kyung, Gill, Ghosh, Casella et al. (e.g. 2010) as well as our simulations below. These estimators, however, are not sparse. By a sparse Bayesian model, we mean one where either the mean, median, or mode of the conditional posterior density of $\Pr(\beta_k|\cdot)$ takes on a value of zero with non-zero probability. By the Bernstein-Von Mises Theorem, we know that the data will swamp the prior and the posterior density will converge to the same density as the maximum likelihood estimate. The maximum likelihood estimate for a continuous parameter is never sparse, outside of pathological cases, and therefore neither are these Bayesian methods. In order to move towards a sparse estimate, the LASSOplus estimator "slows down" the rate at which the prior disappears asymptotically.

Under existing Bayesian methods, variable selection occurs in one of two ways. One, the variable may be selected by examining the posterior density of the estimated effect size; see Figure 8 or Hahn and Carvalho (2015, Section 3). Two, effects may be selected off a summary statistic of the posterior density. Kyung, Gill, Ghosh and Casella (2010) propose fitting a frequentist LASSO such that the sums of absolute values of the frequentist method agree with the posterior mean of the sum of absolute values of the parameters. Carvalho, Polson and Scott (2010) suggest a threshold for selecting off the parameter-specific weight parameters, a process shown to satisfy the Oracle Property (Datta and Ghosh, 2013). Hahn and Carvalho (2015) suggest fitting a non-sparse Bayesian model and then selecting a frequentist model closest to these fitted values.

LASSOplus selects effects off a statistic of the posterior density. It is a sparse approximation to an underlying posterior density, constructed to achieve the Oracle Property.

3 LASSOplus: The Proposed Method

This section progresses in three parts. First, we introduce the LASSOplus model (Section 3.1). Second, we describe how we calculate confidence intervals, including how LASSOplus accommodates

[12] For additional estimators using a scale-mixture normal representation, see Hahn and Carvalho (2015); Bhadra et al. (2015); Bhattacharya et al. (2015); Leng, Tran and Nott (2014); Griffin and Brown (2012); Armagan, Dunson and Lee (2012); Griffin and Brown (2010). For a full discussion of this family of shrinkage estimators, see Polson and Scott (2012).

repeated observations (Section 3.2). Finally, we briefly discuss how LASSOplus easily accommodates parametric flexibility (Section 3.3). In Section 4 we detail the statistical properties of the estimator, including results on the Oracle Property and Oracle Inequality. We then contrast our prior structure for the parameter-specific weights to alternatives in the earlier literature.

3.1 The LASSOplus Model

The LASSOplus model contains two components. The first, which we term the consistent model, returns a consistent estimate for each effect. The second component is a thresholding rule, whereby small effect estimates are trimmed to zero. The LASSOplus estimate consists of the consistent estimates that are not zeroed out by the thresholding rule. We present each component in turn.

The consistent model. We constructed the prior structure for the LASSOplus with two goals in mind. First, the log-posterior takes the same functional form as an adaptive LASSO problem. We show this property below. Second, the posterior mean of λ grows as $N^{1/4}K^{1/2}$. We show in the next section how this growth rate helps the LASSOplus estimator achieve the Oracle Property and satisfy an Oracle Inequality.

The consistent model for LASSOplus can be written as

$$Y_i|X_i, \beta, \sigma^2 \sim \mathcal{N}(X_i^\top \beta, \sigma^2) \tag{32}$$

$$\beta_k|\lambda, w_k, \sigma \sim DE\left(\lambda w_k/\sigma\right) \tag{33}$$

$$\lambda^2|N, K \sim \Gamma\left(K\left(\sqrt{N}-1\right), \rho\right) \tag{34}$$

$$w_k|\gamma \sim \text{generalizedGamma}(1, 1, \gamma) \tag{35}$$

$$\gamma \sim \exp(1) \tag{36}$$

with the generalized Gamma density $f(x; a, d, p) = \frac{p/a^d}{\Gamma(d/p)} x^{d-1} \exp\{-(x/a)^p\}$.

The prior on β is a reweighted version of that in the PC model. The tuning paramter, λ, was constructed to grow in N. This growth is evident in the prior on λ^2: we replace the δ parameter in the PC model with $K(\sqrt{N}-1)$. Any value $\rho > 0$ returns a proper prior; we take $\rho = 1$. The Gamma prior on λ^2 returns the Gibbs update in Formula 39. Lastly, the priors on the weights were derived so that joint posterior of $(\beta, \{w_k\}_{k=1}^K)$ would resemble the adaptive LASSO model. To see this, note that up to an additive constant that does not depend on β or the weights,

$$-\log\left(\Pr\left(\beta, \{w_k\}_{k=1}^K|\lambda, \sigma^2, \gamma\right)\right) = \frac{1}{\sigma^2}\left\{\frac{1}{2}\sum_{i=1}^N (Y_i - X_i^\top\beta)^2 + \lambda\sigma\sum_{k=1}^K w_k|\beta_k|\right\} + \sum_{k=1}^K w_k^\gamma. \tag{37}$$

which combines the elements of equations (8) and (9).

Our model differs from earlier implementations of adaptive weights from Leng, Tran and Nott (2014); Alhamzawi, Yu and Benoit (2012); Griffin and Brown (2012, 2010); Kang and Guo (2009) by placing a prior over w_k^γ rather than w_k and γ separately. Like existing methods, we estimate the decay parameter γ from the data. As we show below in Figure 1, adjusting the parameter allows the model to adapt to the global level of sparsity in the data. Taking $w_k = 1$ for all k returns the Bayesian LASSO model prior structure on β_k;[13] this model is the implementation of LASSOplus that does not have the Oracle Property.

Estimation. Estimation is nearly identical to the augmented PC model. We augment the model as

$$\beta_k|\lambda, w_k, \sigma \sim DE(\lambda w_k/\sigma) \Rightarrow \beta_k|\tau_k^2, \sigma^2, w_k^2 \sim \mathcal{N}(0, \tau_k^2\sigma^2/w_k^2);\;\; \tau_k^2 \sim \exp(\lambda^2/2) \tag{38}$$

There are only two adjustments to the PC Gibbs sampler:

$$\lambda^2|\cdot \sim \Gamma(K\sqrt{N}, \sum_{k=1}^K \tau_k^2/2 + \rho) \tag{39}$$

$$1/\tau_k^2|\cdot \sim \text{InvGaussian}\left(\lambda w_k \sigma/(|\beta_k|), \lambda^2 w_k^2\right) \tag{40}$$

We update w_k and γ using a Griddy Gibbs sampler (Tierney, 1994).

The LASSOplus estimator. As with existing methods, the data generating process above does not return a sparse mode. The LASSOplus estimate is constructed from the estimate β_k and a thresholding function that zeroes out sufficiently small values of $|\beta_k|$. The threshold was constructed such that the final estimate achieves the Oracle Property, a point we return to after defining the estimator itself.

In order to guarantee that we zero out effects in the limit of N, we sample an inflated variance component, σ_{sp}^2

$$\sigma_{sp}^2|\cdot \sim \text{InvGamma}\left(\left(N^{1-2\alpha} - 1\right)/2 + K/2, \frac{1}{2}\left\{\sum_{i=1}^N (Y_i - X_i^\top \beta)^2 + \sum_{k=1}^K \beta_k^2/\tau_k^2\right\}\right). \tag{41}$$

that will enter into the threshold function. The parameter σ_{sp}, which is central to the theoretical properties of LASSOplus, grows approximately as $N^\alpha \sigma$. We implement the model at $\alpha = 1/4$, as this value achieves several desirable theoretical properties as we explain below.

The LASSOplus estimate is constructed from the consistent model and the inflated variance term σ_{sp}. Define as

$$V_i^k = Y_i - X_{i,-k}^\top \beta_{-k} \tag{42}$$

[13] Note that $\beta_k|\lambda, w_k = 1, \sigma \sim DE(\lambda/\sigma)$, which is the PC prior for β_k.

the outcome less the estimated values from all effects except the k^{th}. Next, denote the conditional least squares estimate $\widehat{\beta}_k^{ols}$ as

$$\widehat{\beta}_k^{ols} = \frac{\sum_{i=1}^N X_{ik} V_i^k}{\sum_{i=1}^N X_{ik}^2} \tag{43}$$

Conditional on all other parameters in both models, the LASSOplus estimate for the k^{th} element is then defined as

$$\beta_k^{plus}|\cdot = \beta_k \mathbf{1}\left(\left|\widehat{\beta}_k^{ols}\right| \geq \frac{\lambda \sigma_{sp} w_k}{N-1} \right) \tag{44}$$

As LASSOplus is a Bayesian model, it returns estimates of the full posterior for all parameters. In the examples below, we select and focus on effects for which the median of the posterior density is non-zero, $\text{med}(\beta_k^{plus}|\cdot) \neq 0$. We show that selecting off the median LASSOplus estimates serves as a conservative and powerful rule when trying to identify non-zero effects.

3.2 Approximate Confidence Intervals

A crucial contribution of LASSOplus is uncertainty estimates for model parameters. While the LASSOplus method returns posterior uncertainty estimates, the following discussion shows how to calculate approximate confidence intervals. We focus on confidence intervals because credible intervals are not calibrated to achieve nominal coverage, except in the limit. And as with earlier work, we found coverage to be sub-nominal with credible intervals (Kyung, Gill, Ghosh and Casella, 2010).

In returning approximate confidence intervals, we sample from the approximate sampling distribution of the LASSOplus estimator. To do so, we approximate Equation 44 as

$$\beta_k^{plus}|\cdot \approx \beta_k \mathbf{\Phi}\left(\left|\frac{\widehat{\beta}_k^{ols}}{\widehat{\sigma}_k}\right| \geq \frac{\lambda \sigma_{sp} w_k}{\widehat{\sigma}_k (N-1)} \right) \tag{45}$$

with $\widehat{\sigma}_k$ the variance of β_k^{ols}. As our approximation is differentiable, we apply the delta method to estimate the variance σ_{ci}^2 (see Appendix E for details and Efron (2015) for a more general argument).

The asymptotic normal approximation of the delta method may not hold with small samples. To correct for this, we estimate the error degrees of freedom using Satterthwaite's approximation,

$$df^{ci} = \frac{\left(\sum_{i=1}^N (Y_i - X_i^\top \beta)^2\right)^2}{\sum_{i=1}^N (Y_i - X_i^\top \beta)^4} \tag{46}$$

With an estimate of error degrees of freedom in hand, we exploit the representation of a t-density as a scale mixture of normals with inverse-gamma mixing density. We draw $v_{ci} \sim \text{InvGamma}(df^{ci}/2, df^{ci}/2)$ and use this value to inflate σ_{ci}^2.[14]

[14] In our simulations, this correction made a noticeable difference only at the smallest sample sizes ($N = 50, 100$).

The sampling density of a non-zero univariate LASSO estimate is truncated normal, conditional on the sign of the mode (Potscher and Leeb, 2009). Our approximation of the sampling density is then

$$\beta_k^{ci}|\cdot \sim \begin{cases} \mathcal{N}\left(\beta_k^{plus}, \sigma_{ci}^2 v_{ci}\right);\ \beta_k^{plus} = 0 \\ \mathcal{TN}\left(\beta_k^{plus}, \sigma_{ci}^2 v_{ci}, 0, \infty\right);\ \beta_k^{plus} > 0 \\ \mathcal{TN}\left(\beta_k^{plus}, \sigma_{ci}^2 v_{ci}, -\infty, 0\right);\ \beta_k^{plus} < 0. \end{cases} \tag{47}$$

where $\mathcal{TN}(\mu, \sigma^2, l, u)$ denotes the truncated normal density with mean μ, variance σ^2, and support on (l, u).

The approximate confidence interval is taken from the posterior distribution of β_k^{ci}. For K discovered effects, we take $\widetilde{K} = \max(K, 1)$ and approximate the $1 - \alpha_0\%$ confidence interval as

$$CI_{\alpha_0, \widetilde{K}} = \left(q_{\alpha_0/(2\widetilde{K})}, q_{1-\alpha_0/(2\widetilde{K})}\right) \tag{48}$$

where $q_{\widetilde{\alpha}}$ is the estimated $\widetilde{\alpha}$ quantile of β_k^{ci}, with a Bonferroni correction for the discovered effects. Benjamini and Yekutieli (2005, esp. 74–75) show that implementing a Bonferroni-correction off discovered effects will maintain at least nominal coverage across all discovered effects.

Random effects and uncertainty estimates for designs using repeated unit-level observations. When researchers have repeated units in their sample, ignoring within-unit correlation can produce incorrect uncertainty estimates. Our substantive application, which uses a conjoint experiment by design, features repeated observations at the unit level. Often researchers will utilize some form of clustered standard errors. LASSOplus implements unit-level random effects for the same purpose.[15]

Specifically, assume observation $i \in \{1, 2, \ldots, N\}$ as above. Now, assume each observation was generated by experimental unit j, $j \in \{1, 2, \ldots, J\}$. The function $j[i]$ maps each observation back to one of the experimental units (Gelman and Hill, 2007).

We include random effects u_i as

$$u_i = a_{j[i]} \tag{49}$$

$$a_{j[i]} \sim \mathcal{N}(0, \sigma_a^2). \tag{50}$$

We take the Jeffreys' prior $1/\sigma^2$ on σ^2, though a folded-t density may be used (Gelman, 2006).

[15] For a sweeping discussion and synthesis of related issues, see Stewart (Working Paper).

3.3 Parametric extensions

Most experimental studies implement a linear (mean) model, given its connection to causal estimands. Researchers may prefer alternative models however, such as a probit model for a binary outcome. In this case, LASSOplus models this alternative data generating process in a straightforward way by using the latent variable representation of the probit model (Albert and Chib, 1993). Briefly, the probit regression models the probability of a positive outcome as

$$\Pr(Y_i = 1 | X_i, \beta) = \Phi(X_i^\top \beta) \tag{51}$$

with $\Phi(a)$ representing the cumulative distribution for a standard normal random variable. An observation-specific random variable is introduced, z_i^*, and the problem transforms to

$$z_i^* = X_i^\top \beta + e_i \tag{52}$$

with $e_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. Conditonal on z_i^*, the probit regression is a least squares problem with known variance, and LASSOplus can be applied as above. The latent variable representation also allows straightforward extension to ordinal, categorical, and censored data (Jackman, 2009).

Expectation-Maximization implementation. While we have certainly worked to produce an efficient implementation of LASSOplus, the MCMC method given above may be time-consuming and computationally expensive for large datasets with a large number of possible effects. For practical research and initial model-fitting, we also implement an EM version of the LASSOplus model. We take β, σ^2 as the parameters to be estimated and the remaining parameters, $\left\{\lambda, \{1/\tau_k^2\}_{k=1}^K, \{w_k\}_{k=1}^K, \gamma\right\}$, as "missing."[16] As we have calculated all of the conditional posterior densities or kernels, the EM implementation is straightforward. We defer the details to Appendix F.

4 Statistical Properties of LASSOplus

In this section, we discuss the statistical properties of LASSOplus. First, we present some preliminary results that will be used in later results. Second, we derive condtions under which LASSOplus satisfies the Oracle Property, in both a fixed-K and growing-K setting. Third, we give Oracle Inequality bounds for both the consistent model and LASSOplus. As the Oracle Inequality is a frequentist property, we derive these results in terms of the LASSOplus-EM estimates. Fourth, we compare the LASSOplus prior to several existing Bayesian priors.

[16] See Figueiredo (2004) for a similar setup for a LASSO model.

4.1 Preliminary Results

We present here two preliminary results. First, we consider the role of the weights, $\widehat{w}_k$. Each weight is associated with a parameter in the penalty, where we want to downweight the penalty associated with large effects and upweight the penalty associated with small effects. We first show that the weights and magnitude of the effect estimates are indeed inversely related.

Lemma 2 *The partial derivative of $\widehat{w}_k$ with respect to $|\widehat{\beta}_k|$ is*

$$\frac{\partial \widehat{w}_k}{\partial |\widehat{\beta}_k|} = -\widehat{\lambda}\sqrt{\widehat{\frac{1}{\sigma^2}}}\text{Var}(w_k|\cdot) \tag{53}$$

which is strictly less than zero.

Proof: *See Appendix B.*

This will allow us to associate the largest weights with the smallest effect estimates, and vice versa.

Second, we bound $\widehat{\lambda}$:

Lemma 3 *When N and K grow at some rate, $\widehat{\lambda}$ grows as $N^{1/4}K^{1/2}$.*

Proof: *See Appendix B.*

The bound on $\widehat{\lambda}$ is a crucial component of both proofs below, as the growth rate of $\widehat{\lambda}$ and $\widehat{w}_k$ determine the LASSOplus-EM estimator and, hence, its statistical properties. We turn next to the first result.

4.2 Oracle Property and Oracle Inequality for LASSOplus

We now give conditions on N, K for which LASSOplus achieves the Oracle Property. We then show that both the estimator from the consistent model and LASSOplus each satisfy their own Oracle Inequality.

Oracle Property of the LASSOplus estimator. As described above in Section 2.5, an estimator that satisfies the Oracle Property meets two conditions. First, it is consistent for variable selection, so in-truth non-zero effects are selected and in-truth zero effects are zeroed out. Second, the estimate has to be indistinguishable asymptotically from a consistent, asymptotically efficient, model fit only to the in-truth non-zero effects. In other words, the estimator achieves the same performance as an identical estimator told *ex ante* by an "oracle" which effects are zero and which are not.

The LASSOplus estimator satisfies both properties in the case of K, N growing, and hence is an Oracle estimator. We offer the following:

Proposition 3 **Oracle Property of LASSOplus** *Whenever K is growing in N, but $K^2/N \rightarrow 0$, and $1/4 \leq \alpha < 1$, LASSOplus has the two Oracle Properties:*

1. *Consistent Variable Selection:* $\lim_{N\rightarrow\infty}\{k : \Pr(\widehat{\beta}_k^{plus} \neq 0) = 1\} = S$
2. *Optimal Estimation Rate:* $\sqrt{N}\left(\widehat{\beta}_S^{plus} - \beta_S^o\right) \xrightarrow{\text{d}} \mathcal{N}(\mathbf{0}_{|S|}, \Sigma_S^*)$ *where* Σ_S^* *is the asymptotic variance matrix from the true subset model.*

with S the set of indices of in-truth non-zero elements of β, i.e. $S = \{k : k \in \beta_k^o \neq 0\}$, and. Σ_S^* the asymptotic variance of the least squares estimator fit only to the in-truth non-zero effects.

These properties also hold in the fixed-K setting when $1/4 < \alpha < 1$.

Proof: *See Appendix C.*

Our result of $K^2/N \rightarrow 0$ is much weaker than that of Fan and Peng (2004), who require $K^5/N \rightarrow 0$.[17] This difference appears because Fan and Peng (2004) prove their result in some generality, for general likelihoods and penalties, whereas we are using a normal likelihood and have a particular penalty function.

Oracle Inequality of the LASSOplus estimator. We next present an Oracle Inequality for the LASSOplus-EM model, which we formally prove in Appendix D. Standard Oracle Inequality results generally involve controlling a probabilistic bound on the distance between the estimated and true regression parameters. As LASSOplus-EM recovers the mode of a Bayesian model, the probability in the bound is not controlled by the researcher; it is instead estimated endogenously along with all other parameters. We find that both the consistent model and LASSOplus satisfy an Oracle Inequality.

The Oracle Inequalities below offer two heuristic insights. First, the consistent model performs well with prediction when N is of order $(K/log(K))^2$ or less. Therefore, the consistent model does well when predicting in small-N, large-K settings. Second, LASSOplus satisfies an Oracle Inequality so long as K is growing, and regardless of N. This comes at the cost of a bound that is twice that of the consistent model.

Formal statement of the LASSOplus Oracle Inequality. We state the results here but defer the full set of definitions and derivations to Appendix D. The interested reader will find there all definitions and assumptions, though we note that the practical insights from the theoretical derivations are descibed directly above.

[17] Continuing the example above, for $K = 10$, achieving the Oracle Property with LASSOplus only requires N of order $10^2 = 100$ as opposed to $10^5 = 100,000$.

Denote as $\widehat{W}$ the matrix with the weights along the diagonal; $C.$, i.e. C_λ, C_ϵ, etc., as constants not changing in N or K; $\widehat{\overline{\gamma}}$ the maximal value that can be taken by a weight; $|\widehat{S}|$ is the estimated number of large effects; and $p_\lambda(C_\lambda)$, $p_\epsilon(C_\epsilon)$, and $p_w(C_1, C_2, C_3)$ the probabilites with which necessary bounds are violated. The parameter t is a user-selected term that controls the error, such that the probability that the bound contains a term $\exp(-t^2/2)$.

Proposition 4 **Oracle Inequality for LASSOplus** *The LASSOplus-EM model offers two separate Oracle Inequality results.*

Denote as $\widehat{\delta} = \widehat{\beta} - \beta^o$ *and* $\widehat{\delta}^P$ *the subvector of* $\widehat{\delta}$ *corresponding to effects not zeroed out by LASSOplus. Similarly, let* X_p *and* $\widehat{W}_p$ *denote the submatrices of* X *and* $\widehat{W}$ *associated with elements of* $\widehat{\delta}^p$*. Then, under the assumptions in Appendix D,*

1. *So long as*

$$32 \times \sqrt{(N-1)} \leq \frac{C_\lambda C_\epsilon C_2 \widehat{\overline{\gamma}}^2 K}{\frac{t^2}{2} + \log(K - |\widehat{S}|)} \tag{54}$$

the consistent model will satisfy the Oracle Inequality

$$\frac{1}{N}\left\{||X\widehat{\delta}||_2^2 + \lambda\widehat{\sigma}||\widehat{W}\delta||_1^1\right\} \leq \frac{C_{L1}\widehat{\sigma}^2\widehat{\lambda}^2|S|}{N^2\left\{C_{\phi_1}\frac{\widehat{\lambda}^2\widehat{\sigma}^2\widehat{\beta}^2_{(K)}}{\underline{C}_1\log(|\widehat{S}|)^2} + C_{\phi_2}C_2\widehat{\overline{\gamma}}\right\}^2} \tag{55}$$

with probability at least $1 - \exp(-t^2/2) - p_\lambda(C_\lambda) - p_\epsilon(C_\epsilon) - p_w(C_1, C_2, C_3)$.

2. *So long as*

$$\frac{C_\lambda C_\epsilon C_2 \widehat{\overline{\gamma}}^2 K}{\frac{t^2}{2} + \log(K - |\widehat{S}|)} \geq 32 \tag{56}$$

$\widehat{\beta}^{plus}$ *satisfies an Oracle Inequality.*

$$\frac{1}{N}\left\{||X_p\widehat{\delta}^p||_2^2 + \lambda\widehat{\sigma}|\widehat{W_p}\delta^p|_1^1\right\} \leq 2\frac{C_{L1}\widehat{\sigma}^2\widehat{\lambda}^2|S|}{N^2\left\{C_{\phi_1}\frac{\widehat{\lambda}^2\widehat{\sigma}^2\widehat{\beta}^2_{(K)}}{\underline{C}_1\log(|\widehat{S}|)^2} + C_{\phi_2}C_2\widehat{\overline{\gamma}}\right\}^2} \tag{57}$$

with probability at least $1 - \exp(-t^2/2) - p_\lambda(C_\lambda) - p_\epsilon(C_\epsilon) - p_w(C_1, C_2, C_3)$.

Proof: *See Appendix D.*

The Oracle Inequality offers insight when $K > N$, and LASSOplus performs well in this setting. We next move on to a comparative look at LASSOplus from a Bayesian perspective.

4.3 Comparison to Existing Priors

The Oracle Property and Oracle Inequality are both theoretical results. As we designed LASSOplus for use on real data, we next move on to finite-sample consideration by examining the behavior of the prior structure over the parameter weights w_k. This enables us to compare the prior used in LASSOplus to other priors used in the literature.

There is, of course, no prior structure that performs well in all situations and for all data sets. We have generated a prior structure with four properties. First, the prior is concentrated at zero. This is appropriate for a setting where the researcher confronts hundreds or thousands of effects and wants to winnow these down to a small subset of relevant ones. Second, the prior places a large probability on the existence of large effects. For example, a standard normal prior places a 5.0% prior probability on observing a value larger than 1.96 in magnitude. For a Cauchy prior, this value is 30.0%; for the horseshoe, 20.0%, and for LASSOplus, 31.39%. The more mass in the tails, the less posterior inference on large effects will be impacted by the prior. Third, the decay parameter γ allows the prior to adjust to the level of sparsity implied by the data. Fourth, the prior is less informative than several existing sparse priors. Conditional on assuming a sparse model, we want a prior that drives posterior inference as little as possible.

We illustrate the properties of the LASSOplus prior in Figure 1.[18] In each plot, the y-axis contains the prior probability on a log scale and the x-axis contains the magnitude of the effect, $|\beta|$. The left figure plots the unconditional LASSOplus prior

$$\Pr(\beta) = \int \Pr(\beta|\gamma)\Pr(\gamma)d\gamma \tag{58}$$

against the normal, LASSO (double exponential), Cauchy, and horseshoe priors.

The lefthand plot in Figure 1 illustrates how the different priors will handle large, intermediate, and small effects. The LASSOplus, LASSO, and horseshoe priors all concentrate at zero relative to the normal and Cauchy. The LASSOplus and horseshoe have the most pronounced spike at zero, and therefore will be the most aggressive in shrinking small effects to zero. This suggests that the LASSOplus and horseshoe should make the fewest false positive discoveries. LASSOplus is also relatively aggressive in shrinking intermediate effects to zero. For example, the normal prior places a high mass on effects less than 2 in magnitude, and will therefore shrink those the least. For larger values, say much larger than 4 in magnitude, the normal prior places a vanishingly low probability and will therefore shrink effect estimates quite a bit. The horseshoe and LASSO both place a

[18] The figures were constructed assuming all tuning parameters and the error variance are 1. γ in the unconditional case is calculated by Formula 36. All densities and integrals were calculated empirically at intervals of 0.005 on the range from 0.0001 to 99.9951.

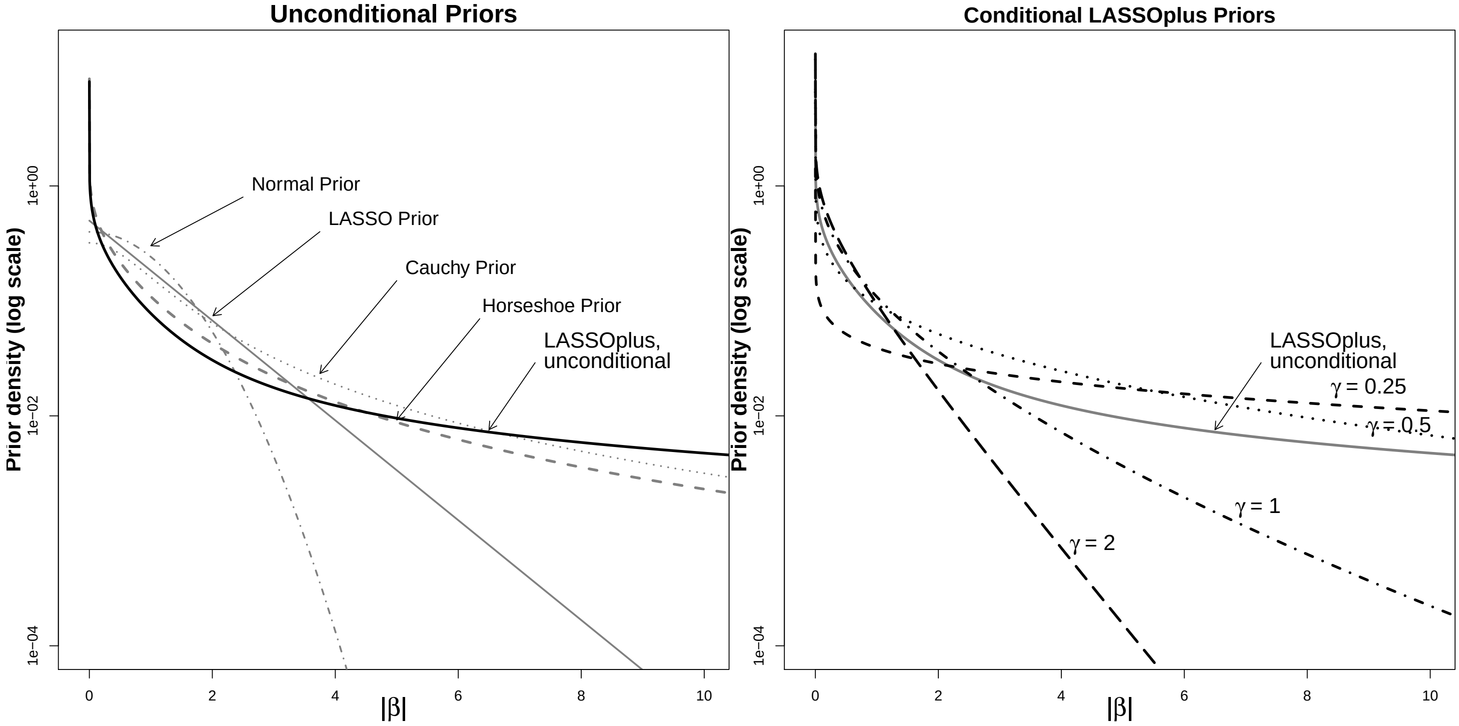

Figure 1: **Comparison of Priors.** This figure compares our unconditional LASSOplus prior to other priors (left) and shows the behavior of our conditional LASSOplus prior under different values of γ (right). The x-axis contains values of β and the y-axis contains the prior density, $\Pr(\beta)$, on a log scale. We also include the two limiting distributions of the t density, the normal and Cauchy, as well as the LASSO and horseshoe priors. The sparse priors (LASSO, horseshoe, LASSOplus) place greater mass at zero than the normal, an effect particularly pronounced for the horseshoe and LASSOplus priors. The normal prior places the least mass, and enacts the most shrinkage, on large effects. LASSOplus has the fattest tails, and thereby will have the least impact on posterior inference for large effects. The righthand plot shows how the decay parameter γ affects the conditional prior, $\Pr(\beta|\gamma)$. The larger γ, the more aggressive the shrinkage. For smaller γ, the more the model is adjusting to large effects. As γ increases, less mass is placed in the tails and more towards zero. The parameter γ is estimated endogenously within the model.

higher mass on intermediate effects less than approximately 4 in magnitude. This suggests that the horseshoe and LASSO are more likely to discover intermediate effects, and that the LASSOplus will be relatively conservative in this range. We find this behavior in the simulations below. The horseshoe, LASSOplus, and Cauchy all have fatter tails than the LASSO and normal. Of these three, the LASSOplus has the fattest tails, and therefore will have the least impact on posterior inference for large effects.

The righthand plot in Figure 1 shows how the decay parameter γ affects the conditional prior, $\Pr(\beta|\gamma)$. For a larger γ, the more aggressive the shrinkage is. For smaller γ, the more the model is adjusting to large effects. As γ increases, less mass is placed in the tails and more towards zero.

	Unconditional Priors				
Prior	*LASSOplus*	*Normal*	*Cauchy*	*LASSO*	*Horseshoe*
Entropy	5.57	2.84	5.06	3.39	3.55
	LASSOplus Priors				
Prior	*Unconditional*	$\gamma = 0.25$	$\gamma = 0.5$	$\gamma = 1$	$\gamma = 2$
Entropy	5.57	9.19	5.4	2.02	1.02

Table 1: **Comparison of Entropy Across Priors.** From an objective Bayesian viewpoint, priors with a higher entropy are favored. Higher-entropy priors, intuitively, add less information to the model and have less impact on posterior inference. We present the entropy of the priors shown in Figure 1. The unconditional LASSO plus has the highest entropy among the unconditional priors. Among the conditional LASSOplus priors, smaller values of γ correspond with a larger entropy. In the absence of information about expected effect size, we prefer the conditional LASSOplus prior with the global decay parameter γ estimated from the data.

The parameter γ is estimated endogenously within the model.

Another way to evaluate different prior distributions is to compare the amount of information they contribute to posterior inference. Practitioners are naturally wary of results driven by prior assumptions. A long strand of research has sought to identify reference priors such that the resultant posterior is dominated by the data (see particularly Bernardo, 1979; Jaynes, 1982; Berger and Bernardo, 1989; Bernardo, 2005; Berger, Bernardo and Sun, 2009; Berger, 2006). A standard measure of the information in a prior p with support $\mathcal{B}$ is its entropy, with higher values indicating less prior influence on the posterior:[19]

$$H(p) = -\int_{\mathcal{B}} p(\beta) \log \left(p(\beta)\right) d\beta. \tag{59}$$

The reference prior is the prior from a class of densities that maximizes the entropy. For a single-parameter, asymptotically-normal posterior, the reference prior is the Jeffreys' prior.

We present the entropy of the priors in Table 1. The unconditional LASSOplus has the highest entropy among sparse priors (LASSO, horseshoe). The unconditional LASSOplus even has higher entropy than the Cauchy, which is proper but has no finite moments. Among the conditional LASSOplus priors, smaller values of γ correspond with a larger entropy. In the absence of information about expected effect size and the underlying level of sparsity, we prefer the conditional LASSOplus prior with the global decay parameter γ estimated from the data.

[19] Uninformativeness is not the only consideration when selecting priors; for example, Jeffreys' prior was originally motived by invariance concerns, while Gelman et al. (2014, p. 129) argue that prior structure should be selected off the sensibility of posterior inference.

5 Application to Subgroup Analysis

Though broadly applicable in a regression framework, we developed LASSOplus for use with subgroup analysis. We discuss next several issues that arise when using the method for subgroup analysis, and how our implementation helps address some of these concerns.

Subgroup analysis through repeatedly split sample analyses. In experimental analyses, subgroup analysis often consists of analysis of repeatedly split samples. For example, authors may conduct a long series of subgroup analyses by splitting the sample into different groups based on one dichotomous or dichotomized pre-treatment covariate at a time. A regression model is fit to each group, and then the marginal effects of the different effects are evaluated within each subsample.

While showing these interactions between treatment conditions and pre-treatment covariates is theoretically interesting, and often requested by audience members and reviewers, the methodology employed has a number of limitations. The decision to split subgroups *one at a time* introduces both conceptual and statistical concerns. Conceptually, the decision to conduct multiple separate subgroup analyses means that we implicitly acknowledge that we have estimated the wrong model. Take for example the simple moderating effects of two separate pre-treatment variables on a treatment effect. The fact that the moderating effect of one of these variables is not included when estimating the moderating influence of another variable implies an awareness of potential mis-specification. Statistically, the decision to conduct multiple separate subgroup analyses by discarding part of the data each time means the estimates are inefficient. This can be remedied by allowing for interactions between the treatment condition and the covariate in the same model. But the proliferation of parameters means that the standard regression framework is ill-equipped to deal with these situations. The basic reason is simple: the number of parameters quickly proliferates and hence some sort of stabilization, such as through sparsity, is necessary. We illustrate the advantages of sparsity-inducing priors below.

5.1 Interpreting Interaction Terms and the Sparsity Assumption

We use LASSOplus to estimate saturated models containing tens, hundreds, or even thousands of interaction terms. Common, and correct, practice proscribes including interaction terms in the absence of lower-order terms. We agree. Our implementation of LASSOplus, at its default, fits a model with all lower-order and interaction terms.

We also want to address the concern that the proposed methodology works only under the assumption that most of the effects are in truth zero. This assumption may arise from the frequentist implementation of the LASSO, whose proponents argue for using the "bet on sparsity principle"

(Hastie, Tibshirani and Friedman, 2010, p. 611–613), arguing that sparse models are to be preferred over dense models. However, we do not assume that the true effects are zero in any sense. Our prior places mass 0 at the point 0. This differs from spike-and-slab priors, where the researcher implicitly places some prior mass on each parameter being zero (Mitchell and Beauchamp, 1988; O'Hara and Silanapaa, 2009; Gill, 2014, ch. 4.6). Instead, we generate a summary of the posterior that takes the value zero with some nonzero probability, allowing the data to tell us *ex post* that some effect is negligible. We are not assuming that the true value, even after the model is fit, is zero–instead, we seek the best sparse representation of the outcome in terms of main and subgroup effects.

Furthermore, researchers interested in characterizing every effect are able to do so. The pseudoposterior density of each effect is not sparse, as illustrated in our applied example. Researchers interested in evaluating non-selected effects are able to do so by evaluating this full posterior density. An important step in any analysis is to look to the data to determine which variables have a non-negligible effect. LASSOplus provides an answer to this question by returning posterior median estimates of zero for negligible variables.

Lastly, the researcher may be nervous in interpreting higher-order interaction terms when the lower-order terms are not selected. The reason is that standard interaction terms cannot be interpreted without referencing their lower-order terms. The problem of interpreting interactions arises because of a correlation between the lower-order terms and the interaction term: the effect of one cannot be considered independently of the other (Esarey and Summer, 2015). To solve this problem, we include interactions terms that are uncorrelated with their lower-order terms. To do so, first we construct the interaction term through elementwise multiplication, regress this term on its lower-order terms, and enter the residuals from this regression into LASSOplus. Under this construction, the effects of interaction terms can be interpreted as the effect above and beyond any lower-order effects. For a proof, see Appendix G.

5.2 Scope of Method for Subgroup Analysis

The development of the LASSOplus method to, in part, facilitate subgroup analysis had two goals. The first is to allow the experimentalist to uncover potentially relevant subgroup effects after implementing an experiment. Second, we designed the method to apply from the simplest of experiments to a conjoint analysis with repeated observations. The analysis of experimental data normally occurs in two steps. An experiment is designed to test a set of pre-specified hypotheses. Upon completing an experiment, these hypotheses are then tested, and the point estimates and p-values for each effect are reported.

Our method is designed for use in the subsequent analysis. After the inferential stage, researchers may be interested in higher-order effects, including *treatment heterogeneity*, when two treatments have an interactive effect; *treatment effect heterogeneity*, when the effect of a treatment varies with a covariate; or *targeted treatment effect*, when the effect of a treatment interaction varies with a covariate (Imai and Ratkovic, 2013). Common practice involves repeatedly subsetting the data on pre-treatment covariates and running separate regressions in each subset. For example, the researcher may find no treatment effect on average, but may find effects of opposite signs for males and females in her experimental data. The problem is the sheer number of effects quickly overwhelms, and researcher-driven repeated subsetting quickly devolves into interaction fishing.

The proposed method provides a means for considering all possible interaction terms and returning a sparse subset estimated as non-zero. The estimation is post-inferential and descriptive, returning the effects that seem pronounced in this experimental data and might be considered for further study in the next experiment.

6 Simulation Study

In this section, we compare LASSOplus to several sparse estimators, assessing each method in terms of discovery of all effects, discovery of small effects, and coverage. We find that LASSOplus performs competitively across each dimension.

6.1 Setup

Our simulation is motivated by conjoint experiments, where the researcher wants to search through a large number of treatment/covariate interactions. The simulation mimics an experiment with 3 treatments, having 2, 3, and 4 levels, respectively. We also assume a set of p pre-treatment covariates, with $p \in \{10, 25, 50, 75, 100\}$. The design matrix consists of the matrix of treatment indicators, T, the pre-treatment covariates, X^0, and their residualized interactions. We designate the first level of each treatment as the baseline, dropping it and all of its interactions from X. After creating all interactions, we are left with a design matrix X with one of $\{76, 181, 356, 531, 706\}$ covariates. We run simulations with sample sizes $N \in \{50, 100, 250, 500, 1000, 2500\}$. For 16/30 of our simulation setups (53%), we have more observations than covariates.

We assume the following model:

$$\begin{aligned}
E(Y_i|X_i, T_{i1}, T_{i2}, T_{i3}) \propto & 3X_{i2} + 3X_{i3} + 3X_{i4} + && (60)\\
& 2 \times \mathbf{1}(T_{i2} = b) + 2 \times \mathbf{1}(T_{i3} = b) + 2 \times \mathbf{1}(T_{i3} = c) + && (61)\\
& X_{i2} \times \mathbf{1}(T_{i2} = b) + X_{i2} \times \mathbf{1}(T_{i3} = c) + X_{i2} \times \mathbf{1}(T_{i2} = a) + && (62)\\
& X_{i4} \times \mathbf{1}(T_{i1} = b) + X_{i3} \times \mathbf{1}(T_{i3} = b) + X_{i4} \times \mathbf{1}(T_{i3} = c) && (63)
\end{aligned}$$

with noise from a t_5 density and the systematic component is scaled to give a true R^2 of 0.5. The variables in X_i are drawn from a $Wishart_p(N)$, and each treatment condition is equiprobable. Each simulation setting was run 1000 times. The simulation design mimics a situation where there a few large main effects, some medium-sized average treatment effects, and the remaining interaction terms are small relative to the other effects.

6.2 Alternative Estimators

We compare LASSOplus to three alternative sparse estimators: the frequentist LASSO, frequentist adaptive LASSO, and the horseshoe estimator. The LASSO and adaptive LASSO are fit using `glmnet` from **R** package `glmnet`. First-stage estimates come from ridge regression with the tuning parameter selected to minimize cross-validated error. The horseshoe prior is implemented in `STAN`.[20] We coded up an implementation of LASSO+OLS with all tuning parameters set at the defaults suggested by the original authors.

For details on the implementation of additional methods, see Appendix H.

6.3 Results

A primary concern is differentiating relevant from irrelevant effects. As a secondary concern, estimators must be responsive to main effects as well as smaller subgroup effects. We first compare methods based on their ability to reliably pick up the former and not mislead on the latter. We consider all effects (Figure 2) and smaller interaction effects (Figure 3).

Consider first the results for all effects, in Figure 2. The plots are arranged in rows by the number of possible effects, with the sample size along the x-axis. Columns contain false positives (left), false negatives (center), and false discovery rates (right). Starting in the rightmost column, the LASSOplus has the lowest false positive rate across all settings except for the largest-N, smallest-K setting, where it is outperformed by the horseshoe. In terms of false negatives, consider first the setting with $K = 76$ and $K = 181$. In this setting, the LASSOplus achieves a false negative rate either lower than or approximately the same as the horseshoe. As K increases, the horseshoe grows more aggressive in identifying effects. The LASSO and adaptive LASSO are both more aggressive than LASSOplus in identifying effects. Across simulation setups, the LASSOplus dominates existing methods in terms of false discovery rate except for two settings: the largest N, smallest K setting where the horseshoe performs best and the smallest N, smallest K setting where the LASSO performs best. LASSO+OLS identifies the fewest correct effects as the sample size grows. As the

[20] The code for the horseshoe was adapted from the code at http://brandonwillard.bitbucket.org/bayes-horseshoe-plus/horseshoe-stan-nvm.html, last accessed September 28, 2015. After acceptance, we found that a simpler implementation can be found in **R** package `rstanarm`.

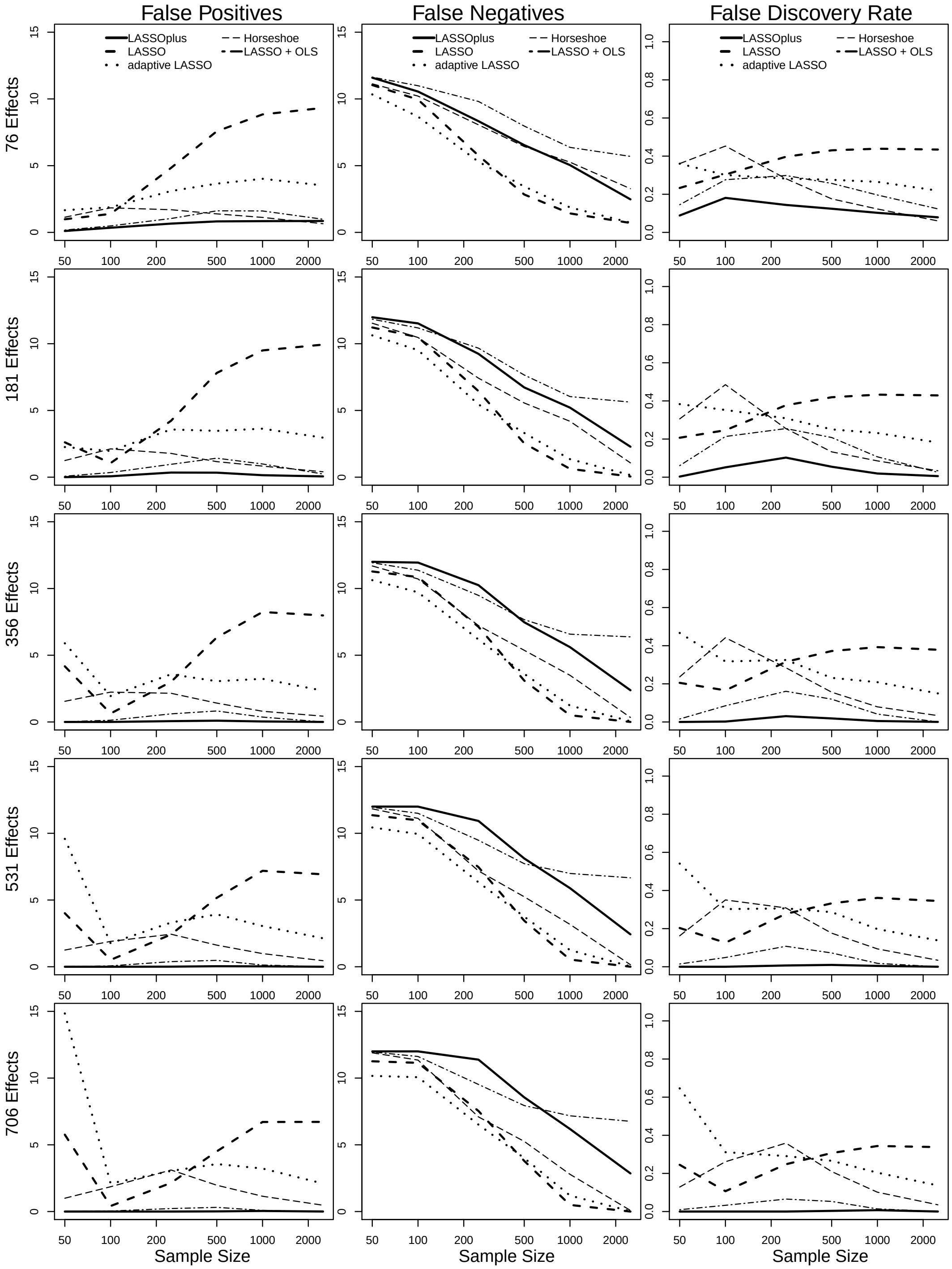

Figure 2: **False Positive, False Negative, and False Discovery Rates, All Effects.** Each row represents the results for a different number of effects. The sample size is along the x-axis. LASSOplus achieves a lower false positive rate and false discovery rate across simulation settings (first and third columns), while still maintaining reasonable power in terms of the false negative rate (middle column).

number of covariates increases, LASSOplus achieves a lower false discovery rate relative to the alternative methods. LASSOplus, the horseshoe, adaptive LASSO, and LASSO + OLS all have false discovery rates that are decreasing in sample size.

Figure 3 contains the same results but considers only the six interactive effects. These effects are 1/3 the size of the non-zero main effects and 1/2 the size of the average non-zero treatment effects. Again, plots are arranged in rows by the number of possible effects, with the sample size along the x-axis. Columns contain false positives (left), false negatives (center), and false discovery rates (right). We do not report values for the false discovery rates (FDR) if there are no discovered effects, which is why LASSOplus is missing entries at the lower sample sizes in the FDR column. We do so in order to differentiate an FDR of zero because there are no discoveries from an FDR of zero because none of the discovered effects are false.

Figure 3 contains patterns similar to those in Figure 2: the false positive rate is lowest for LASSOplus, and LASSOplus or LASSO + OLS tends to make the most false negatives aside from the smallest K setting, where the horseshoe performs worse with large N. Again, the LASSO and adaptive LASSO are the most aggressive in identifying effects. The pattern among false discovery rates is similar but more pronounced than that in Figure 2. Aside from a single setting (N=50, K=76), LASSOplus achieves the lowest false discovery rate among all methods.

These results are consistent with the prior structure illustrated in Figure 1. Both the LASSOplus and horseshoe have a higher prior density at zero, thereby zeroing out effects more aggressively. We see this in the higher number of false positives for the horseshoe and LASSO than the adaptive LASSO. LASSOplus places a lower prior density on intermediate effects and a higher density on large effects than the horseshoe and LASSO.

Finally we turn to coverage in Figure 4. Columns contain coverage for all nonzero effects (left), discovered effects (middle), and discovered interactions (right). The gray line at 0.9 is the nominal rate. We do not return coverage results on in-truth zero effects, as these are close to 1 for all methods as they all shrink effects towards zero.

Confidence intervals are calibrated to be nominal on discovered effects, with coverage shown in the middle column. We find coverage from LASSOplus to be nominal or near-nominal on discovered effects across simulations. The horseshoe estimator consistently returns conservative confidence intervals. The boostrapped LASSO + OLS returns near-nominal coverage, similar to LASOSplus. Across most settings, the perturbation confidence intervals of the adaptive LASSO are too narrow. The lefthand column reports coverage on all nonzero effects. By this measure, LASSOplus generally outperforms alternative methods, growing closer to nominal in sample size. Across settings, the

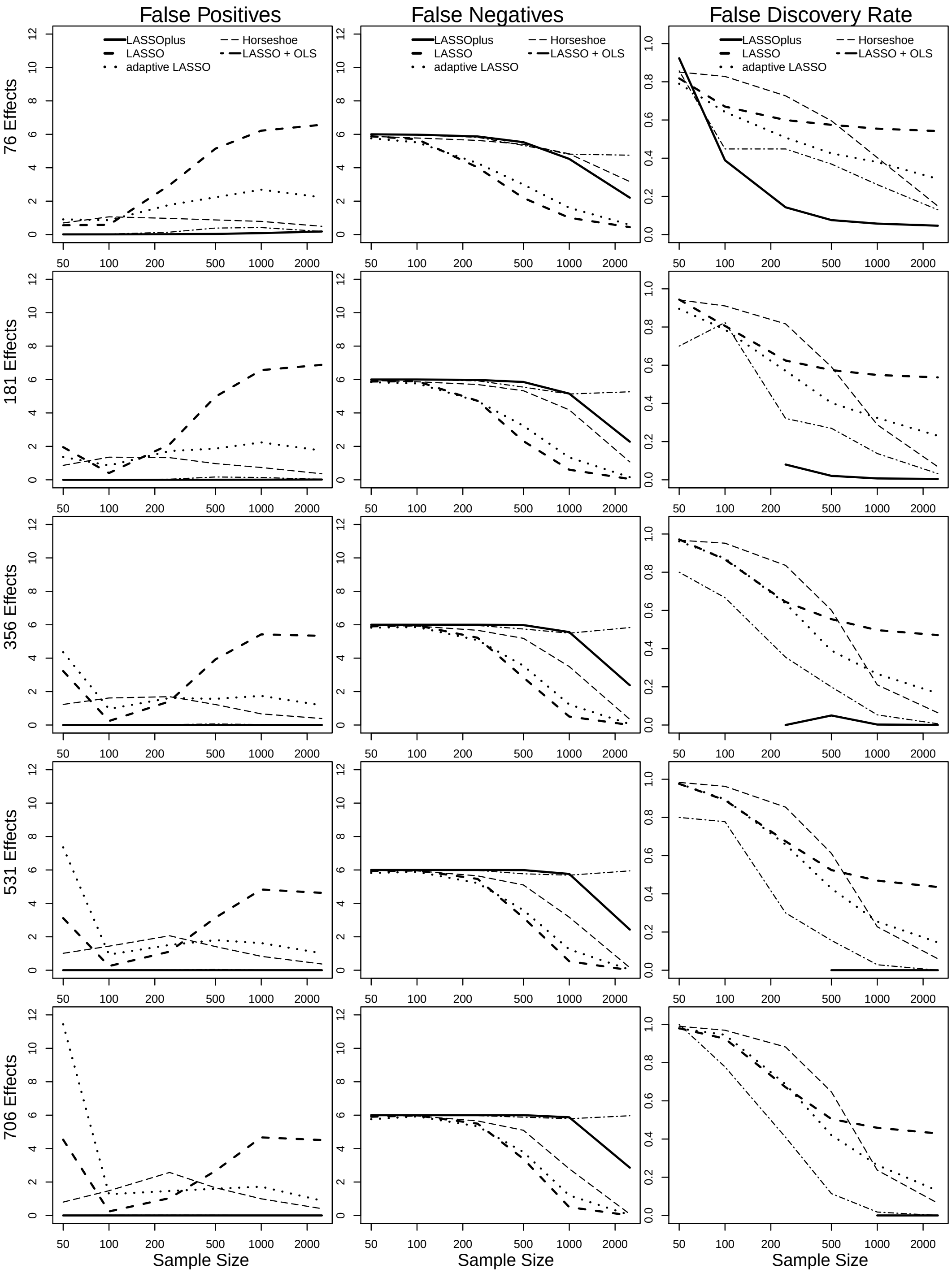

Figure 3: **False Positive, False Negative, and False Discovery Rates, Interaction Terms Only.** This figure considers each method's ability to uncover the six small interaction effects. The columns contain false positives (left), false negatives (middle), and false discovery rate (right).

horseshoe goes from too narrow at a small N to too wide for large N. The rightmost column reports coverage on only the six interaction effects that pick up a subgroup effect. LASSOplus performs

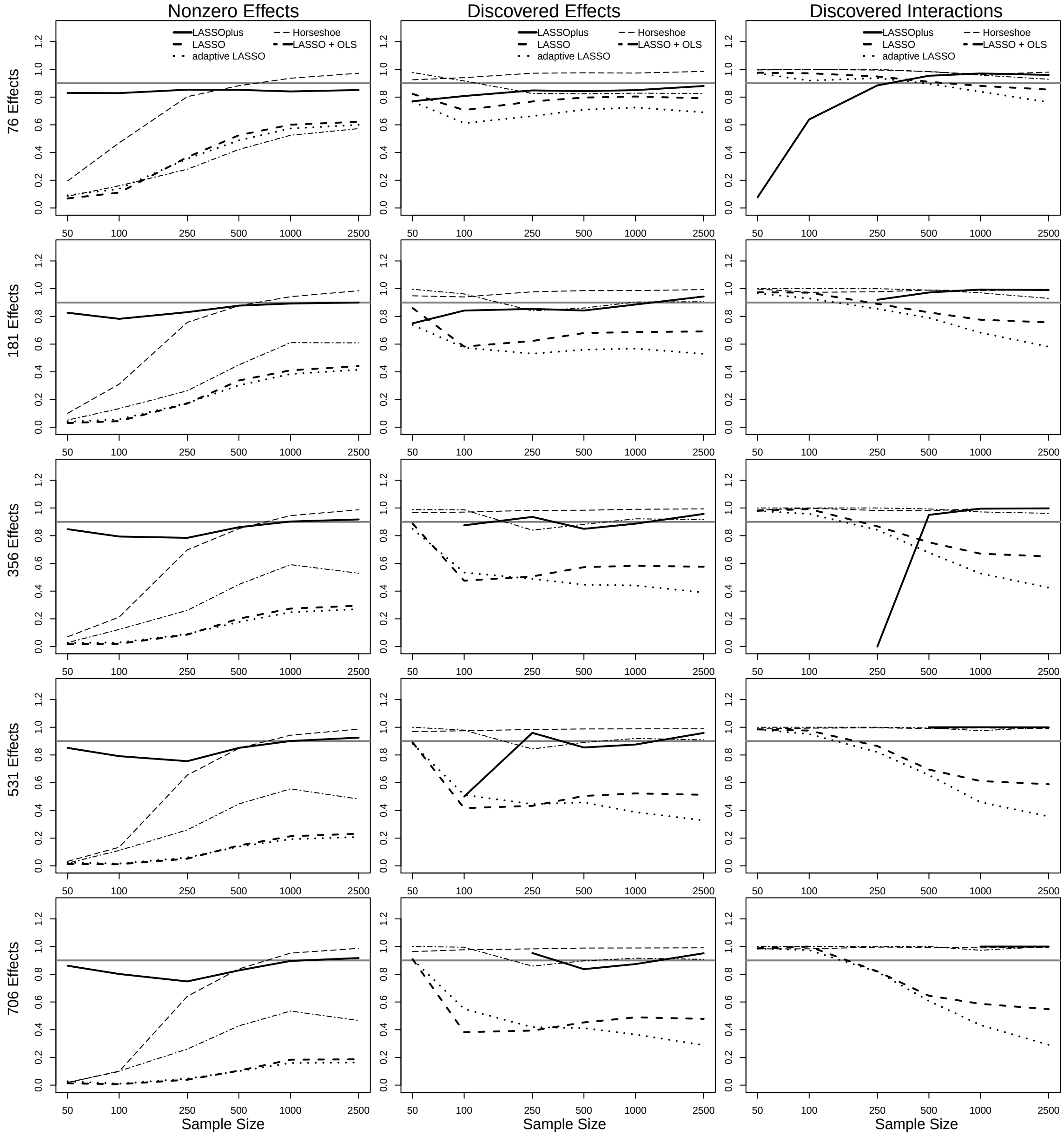

Figure 4: **A Comparison of Coverage Across Methods.** Columns contain coverage for all nonzero effects (left), discovered effects (middle), and discovered interactions (right). The gray line at 0.9 is the nominal rate. Confidence intervals are calibrated to be nominal on discovered effects. We find coverage from LASSOplus to be nominal or near-nominal on discovered effects across simulations. The horseshoe estimator consistently returns conservative confidence intervals. Across most settings, the perturbation confidence intervals are too narrow.

poorly in the small N, small K settings, but as N grows it achieves conservative coverage on these effects. As with the FDR above, areas where LASSOplus does not show coverage, no effects were discovered. The horseshoe maintains conservative coverage on these effects, while coverage for the LASSO and adaptive LASSO grows worse in sample size.

On the whole, the LASSOplus appears successful in identifying effects, identifying small effects, and in generating approximate confidence intervals with nominal coverage. We next illustrate how

LASSOplus performs on data from a recent conjoint analysis.

7 Application

To illustrate the proposed method we analyze a conjoint experiment that examines preferences over different dimensions of international climate agreements (Bechtel and Scheve, 2013). Conjoint experiments expose subjects to multiple different treatment conditions at once. In this study the authors varied the expected costs of the agreement, how costs would be distributed across different groups of countries, the participation rates of countries, the extent to which emissions would be reduced, the severity of sanctions for violations, and the identity of organizations that would monitor compliance. The survey was fielded to nationally representative samples in the United States, United Kingdom, France, and Germany. In the conjoint experiment, respondents considered two agreements, each with various values for the aforementioned dimensions, and then chose the agreement they preferred. The authors then transformed the data to examine the probability that each agreement was chosen as a function of its attributes. To estimate effects the authors implemented a simple dummy regression framework for each of the dimensions and clustered the standard errors at the respondent level to account for the repeated observations at the respondent level.

Figure 5 presents the original results, which we produce using the `cjoint` package (Strezhnev et al., 2014) that implements the methods described in Hainmueller, Hopkins and Yamamoto (2014), which in this case is equivalent to the dummy variable regression used by the original authors. As can be seen, the different dimensions of international agreements have an impact on support. For example, as the cost of the agreement goes up, support goes down. The original paper emphasized other aspects of the design of agreements, such as how increasing the number of participating countries leads to greater support for the agreement, which the authors took as evidence of the important role of reciprocity in international agreements (see also Tingley and Tomz, 2013).

The authors then conducted a long series of subgroup analyses by splitting the sample into different groups based on one dichotomous pre-treatment covariate at a time. Next they estimated the same dummy variable regression model for each group and then examined whether the marginal effects of the different conjoint conditions varied. In the main body of the paper they focused on two effect modifiers: the respondents' general level of environmentalism and the respondents' propensity to engage in reciprocity in a two-player linear public good game that was included in the survey after the conjoint experiment. They found, for example, that the effect of a high cost agreement on opposition to the deal was lower for individuals who are environmentalists compared to individuals who are not environmentalists. The authors also explored a range of other subgroup analyses in

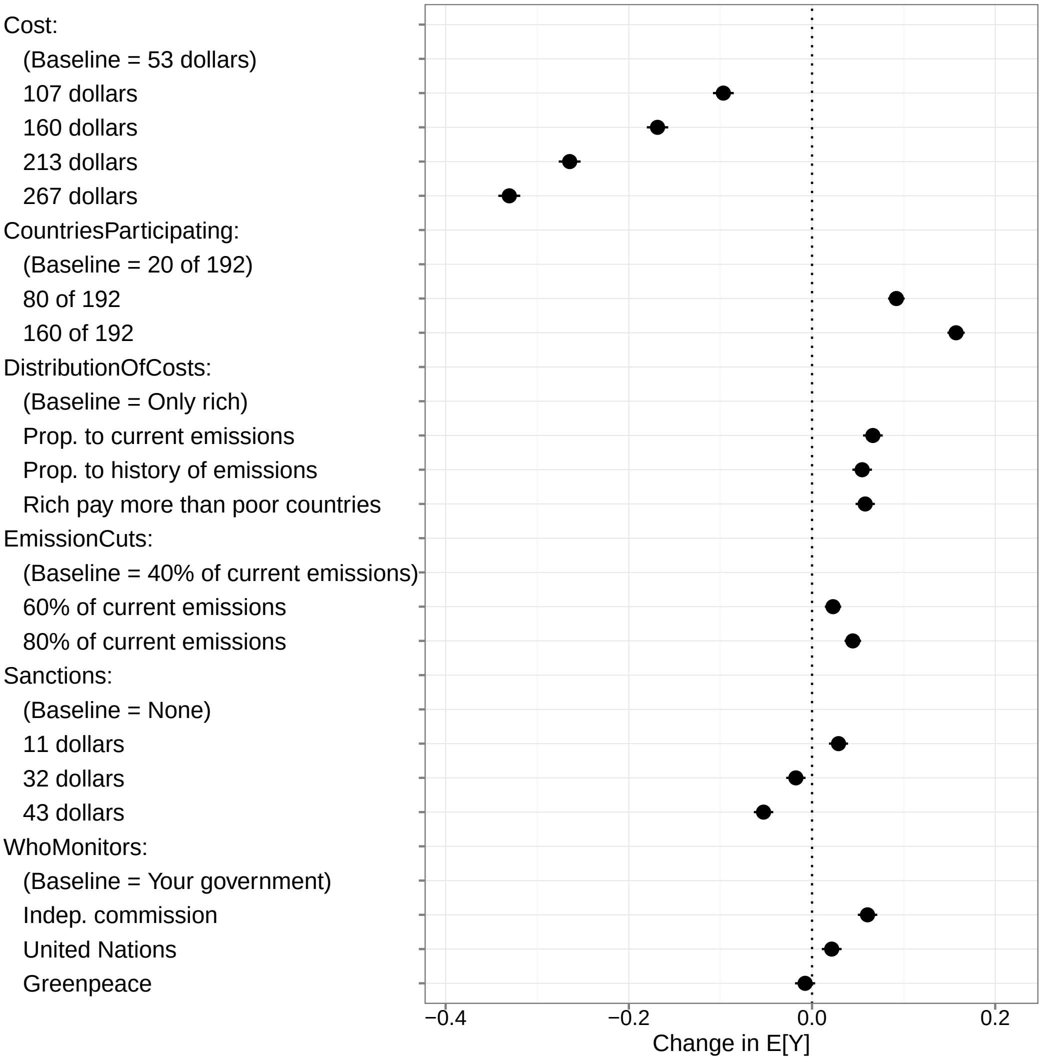

Figure 5: Original Results from Figure 2 in Bechtel and Scheve (2013).

a lengthy supplementary materials section. Importantly, each subgroup analysis was conducted by subsetting the data to one smaller group and calculating the effects of the conjoint conditions. LASSOplus achieves the same goal with one line of code.[21]

To illustrate the application of the LASSOplus algorithm we present the case where a set of pre-treatment covariates can moderate the effect of a set of treatment variables.[22] Conceptually this is analogous to the case analyzed in Bechtel and Scheve (2013). We estimated the LASSOplus model using the Gibbs sampler with 30,000 burnin iterations, 30,000 posterior samples, and thinning

[21] Given available human time and estimation strategies, this was a reasonable approach by the authors. We offer LASSOplus as an alternative and use this example for illustration purposes only.

[22] The pre-treatment variables were gender, and dichotomized values of the the Ideology, Environmentalism, and Reciprocity variables used in the original study, as well as factoral variables for country of the survey, and age (coded as low, middle and high corresponding to the 33rd and 66th quantiles of the age distribution). There were 67,992 observations and 215 potential effects. Each covariate level is interacted with each level of each treatment.

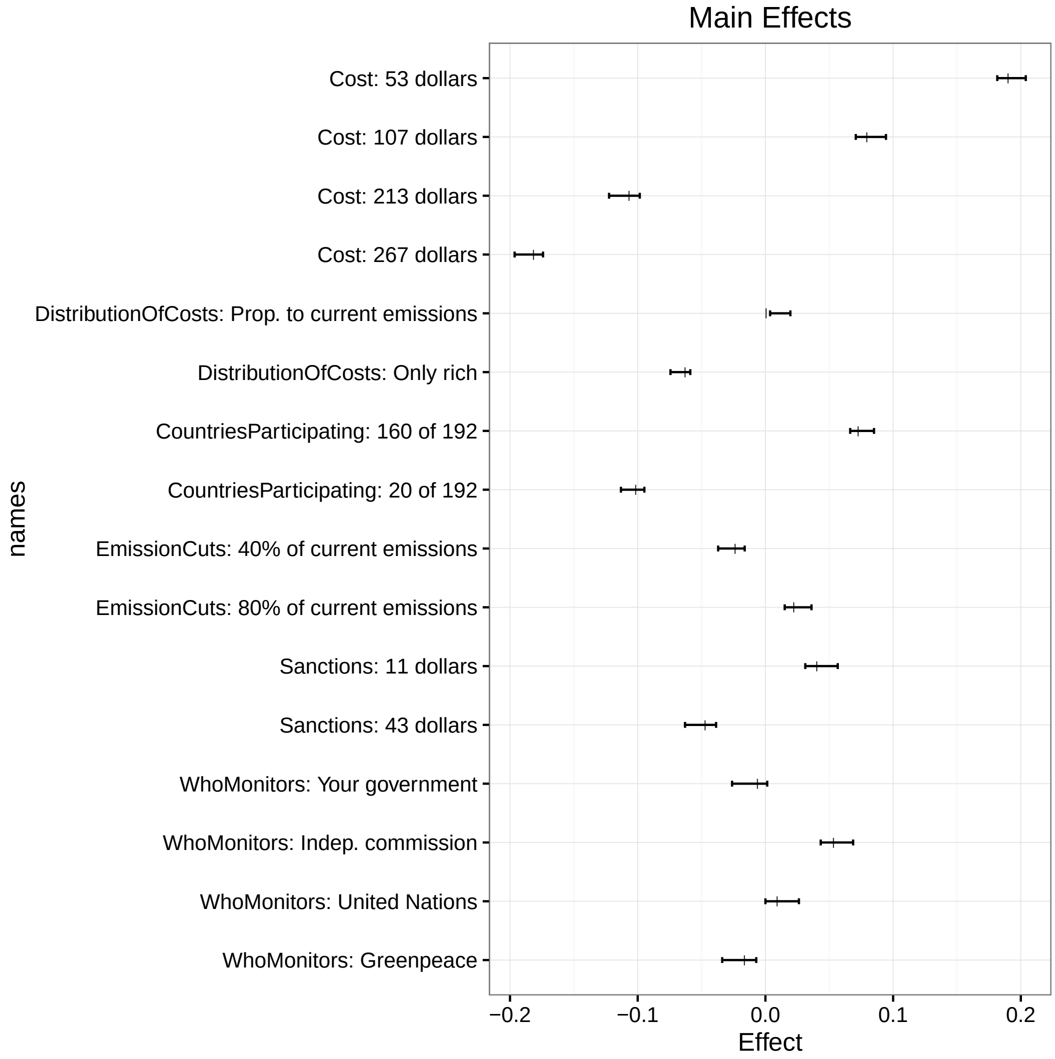

Figure 6: Estimated effects for non-zero coefficients for main effects. Normal linear model for outcome. Each conjoint level represented by the category of treatment (e.g., Cost, WhoMonitors) and the specific level of each treatment (e.g., 53 dollars, United Nations).

every 30 samples, which yielded 1,000 draws from the posterior. LASSOplus selected 41 effects. We present the non-zero effects and their 95% intervals for the main effects (i.e., non-interacted variables) in Figure 6 and the non-zero effects for interaction terms in Figure 7.[23]

We immediately see a number of effects that were strong in the original analysis. Less expensive agreements and those with broad participation were favored, and expensive agreements and those with limited participation were less popular. The fact that these main effects do not disappear within our framework is important, and thus these effects are unlikely to be false positives. More interestingly we see a number of interactions between pre-treatment covariates and treatment con-

[23] Uncertainty estimates were calculated by taking the 5th and 95th quantile of the approximate confidence interval discussed in Section 3.2.

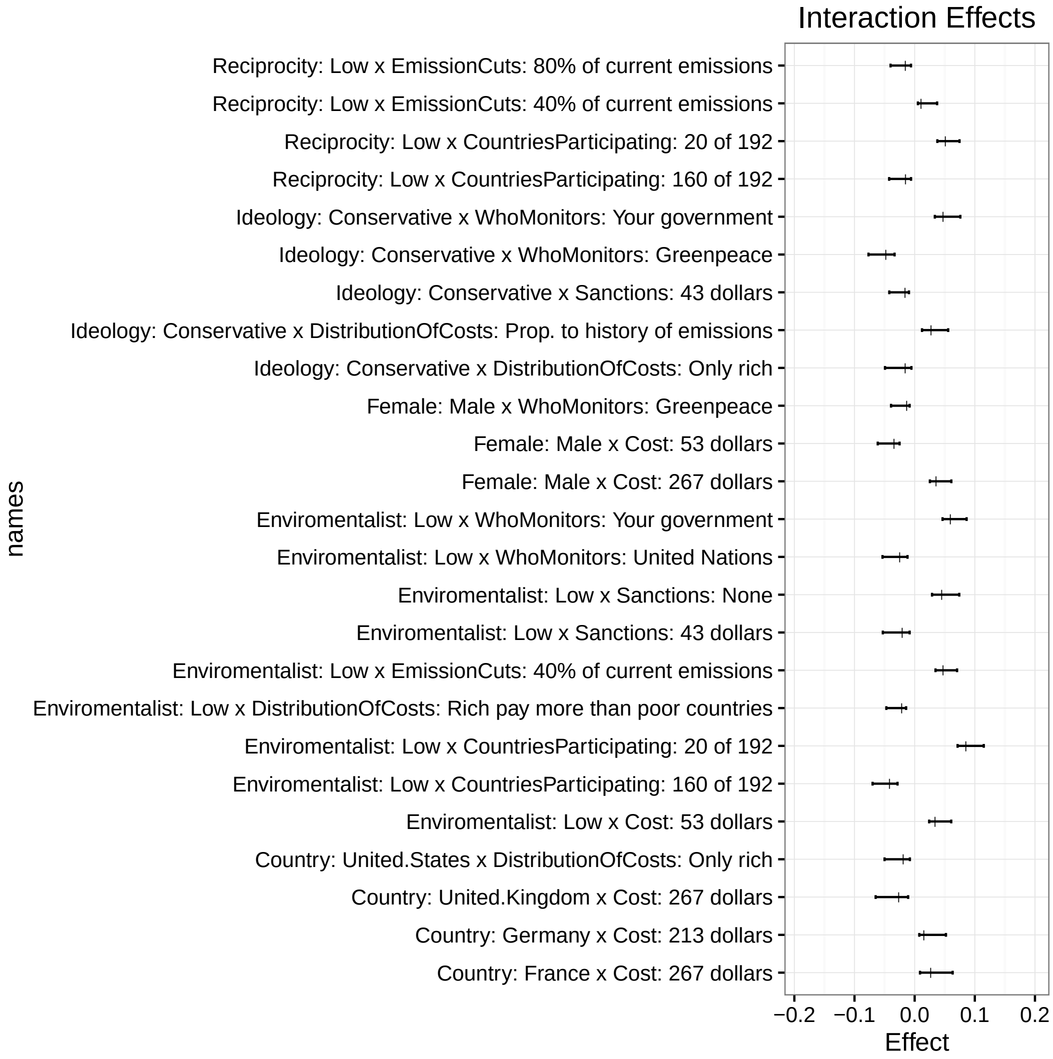

Figure 7: Estimated effects for non-zero coefficients for interaction effects. Normal linear model for outcome. Each interaction term represented as moderating variable and its level x the category of treatment and its level. For example, “Ideology: Conservative x WhoMonitors: Your government” is the interaction between conservative respondents and the conjoint condition of their own country monitoring the agreement.

ditions. For example, individuals with low environmentalism scores had a positive evaluation of the agreement when 20 out of 192 countries joined the deal. This contrasts with how these individuals responded to a treaty with 160 out of 192 countries: in this case they were more opposed to the agreement. This result was present using the original analysis method in Bechtel and Scheve (2013).

Finally, the effect of the ideology variable also has some interesting results. As in the supplementary materials presented in Bechtel and Scheve (2013), conservatives were less enamored than liberals in having the monitoring done by Greenpeace, and more enthusiastic about having their own government conduct the monitoring. Figure 8 plots the posterior distribution of the effects for these interaction terms, as well as the interaction with the United Nations serving as the monitoring agency. In the last case, there is substantial support at 0 (no effect) but some of the mass is negative, indicating an intermediate position between the other two effects, which were cleanly positive or negative.[24]

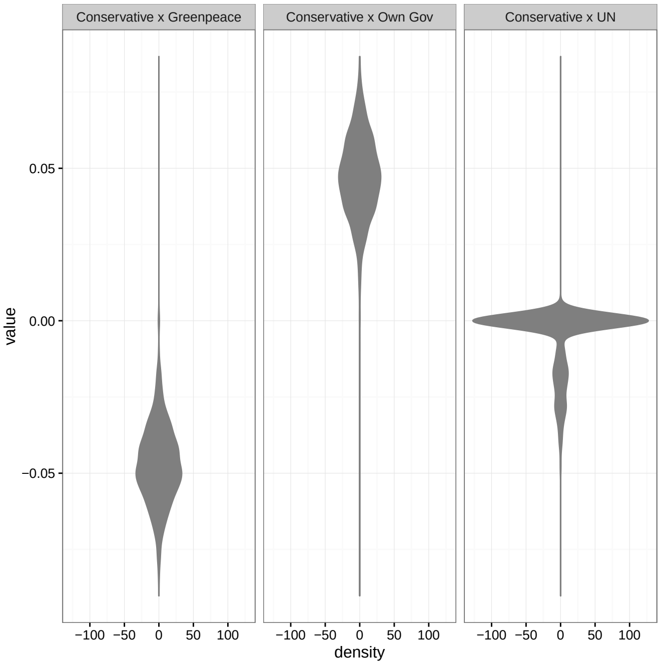

Figure 8: Posterior distribution for interactions between Conservative ideology and monitoring source.

8 Conclusion

The LASSOplus unifies recent contributions in the sparse modeling literatures and provides competitive performance with substantially more flexibility. The method offers several advantages. First, unlike existing Bayesian methods, LASSOplus both selects and estimates effects, returning point estimates and whether the effect is relevant. Second, LASSOplus returns conservative confidence intervals that maintain nominal coverage among discoveries. Third, unlike existing software, LASSOplus models repeated observations, a boon to experimentalists using the same unit (e.g., individuals) several times as is common in conjoint analyses.

We apply the method in the context of analyzing subgroup effects. Given the proliferation of potential effects that subgroup analyses can generate, a sparse model like LASSOplus is useful. However, we go beyond the core estimation strategy to show how scaling covariates properly can

[24] In separate models we allowed for interactions between treatments. We did not see any interactions between treatment conditions on their own. For example, there is no interaction between the cost of the agreement and extent of other country participation. For a similar observation and discussion see Tingley and Tomz (2013).

allow for straightforward identification of causal effects under our estimation framework. We also show that the same scaling can generate interaction effects that can be interpreted independent of their lower-order terms. This proves useful in interpreting models where higher-order effects are selected but lower-order terms are not.

We contrast LASSOplus to other LASSO based approaches in great detail, including unpacking different types of Oracle results. We also conduct one of the most systematic simulation studies to date, comparing LASSOplus to many of these alternatives. The simulation study and application to the data from a recent experiment highlight the method's use and efficacy. We find that the method performs well relative to the frequentist LASSO, adaptive LASSO, and LASSO+OLS models in terms of effect discovery and coverage. But the LASSOplus comes with the additional aforementioned advantages. We show with real world data that the method returns many results uncovered by the authors through their own split-sample regressions, but does so within a coherent statistical framework.

Another advantage of the LASSOplus is substantial flexibility. For example, the method easily extends to other parametric models, such as the probit and type 1 and 2 tobit. Finally, we make a software package available, `sparsereg`, in the **R** programming language that implements the methodology discussed in this paper. Future work will involve extending the method to panel data, censored data, and various modes of causal inference: propensity score methods, instrumental variable methods, mediation methods, and selection models.

References

Albert, James H. and Siddhartha Chib. 1993. "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association* 88:669–679.

Alhamzawi, Rahim, Keming Yu and Dries F Benoit. 2012. "Bayesian adaptive Lasso quantile regression." *Statistical Modelling* 12(3):279–297.

Armagan, Artin, David B. Dunson and Jaeyong Lee. 2012. "Generalized Double Pareto Shrinkage." *Statistica Sinica* .

Bechtel, Michael M and Kenneth F Scheve. 2013. "Mass support for global climate agreements depends on institutional design." *Proceedings of the National Academy of Sciences* 110(34):13763–13768.

Belloni, A., D. Chen, V. Chernozhukov and C. Hansen. 2012. "Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain." *Econometrica* 80(6):2369–2429. **URL:** *http://dx.doi.org/10.3982/ECTA9626*

Belloni, Alexandre and Victor Chernozhukov. 2013. "Least squares after model selection in high-dimensional sparse models." *Bernoulli* 19(2):521–547.

Belloni, Alexandre, Victor Chernozhukov and Christian Hansen. 2011. Inference for high-dimensional sparse econometric models. CeMMAP working papers CWP41/11 Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Benjamini, Yoav and Daniel Yekutieli. 2005. "False Discovery Rate-Adjusted Multiple Confidence Intervals for Selected Parameters." *Journal of the American Statistical Association* 100(469):71–93.

Berger, J. O. and J. M. Bernardo. 1989. "Estimating a product of means: Bayesian analysis with reference priors." *Journal of American Statistical Association* 84:200–207.

Berger, James O. 2006. "The case for objective bayesian analysis." *Bayesian Analysis* 1(3):385–402.

Berger, James O., Jose M. Bernardo and Dongchu Sun. 2009. "The formal definition of reference priors." *The Annals of Statistics* 37(2):905–938.

Berger, James O., Xiaojing Wang and Lei Shen. 2015. "A Bayesian Approach to Subgroup Identification." *Journal of Biopharmaceutical Statistics* 24(1):110–129.

Berk, Richard, Lawrence Brown, Andreas Buja, Kai Zhang and Linda Zhao. 2013. "Valid Post-Selection Inference." *Annals of Statistics* 41(2):802–837.

Bernardo, J. M. 1979. "Reference posterior distributions for Bayesian inference." *Journal of the Royal Statistical Society Series B* 41:113–147.

Bernardo, Jose M. 2005. Reference analysis. In *Handbook of Statistics*, ed. D. K. Dey and C. R. Rao. Elsevier.

Berry, Donald. 1990. "Subgroup Analysis." *Biometrics* 46(4):1227–1230.

Bhadra, Anindya, Jyotishka Datta, Nicholas G. Polson and Brandon Willard. 2015. "The Horseshoe+ Estimator of Ultra-Sparse Signals." Working paper.

Bhattacharya, Anirban, Debdeep Pati, Natesh S. Pillai and David B. Dunson. 2015. "Dirichlet-Laplace priors for optimal shrinkage." *Journal of the Americal Statistical Association* In print.

Bickel, Peter, Ya'acov Ritov and Alexandre Tsybakov. 2009. "Simultaneous Analysis of Lasso and Dantzig Selector." *Annals of Statistics* 37(4):1705–1732.

Buhlmann, Peter and Sara van de Geer. 2013. *Statistics for High-Dimensional Data*. Springer.

Candes, E. and T. Tao. 2007. "The Dantzig selector: statistical estimation when p is much larger than n (with discussion)." *Annals of Statistics* 35:2313–2404.

Candes, Emmanuel J. 2006. "Modern statistical estimation via oracle inequalities." *Acta Numerica* pp. 1–69.

Carvalho, C, N Polson and J Scott. 2010. "The Horseshoe Estimator for Sparse Signals." *Biometrika* 97:465–480.

Chatterjee, A and SN Lahiri. 2011. "Bootstrapping Lasso Estimators." *Journal of the American Statistical Association* 106(494):608–625.

Chatterjee, A and SN Lahiri. 2013. "Rates of convergence of the adaptive LASSO estimators to the oracle distribution and higher order refinements by the bootstrap." *The Annals of Statistics* 41(3):1232–1259.

Chatterjee, Sourav. 2014. Assumptionless Consistency of the LASSO. arxiv:1303.5817v5.

Chernozhukov, Victor, Iván Fernández-Val and Blaise Melly. 2013. "Inference on Counterfactual Distributions." *Econometrica* 81(6):2205–2268.

Datta, Jyotishka and Jayanta K. Ghosh. 2013. "Asymptotic Properties of Bayes Risk for the Horseshoe Prior." *Bayesian Analysis* 8(1):111–132.

Donoho, David L. and Iain M. Johnstone. 1994. "Ideal Spatial Adaptation by Wavelet Shrinkage." *Biometrika* 81(3):425–455.

Efron, Bradley. 2015. "Frequentist accuracy of Bayesian estimates." *Journal of the Royal Statistical Society Series B* 77(3):617–646.

Esarey, Justin and Jane Lawrence Summer. 2015. "Marginal Effects in Interaction Models: Determining and Controlling the False Positive Rate." Working Paper.

Fan, Jianqing and Heng Peng. 2004. "Nonconcave Penalized Likelihood with a Diverging Number of Parameters." *The Annals of Statistics* 32(3):928–961.

Fan, Jianqing and Runze Li. 2001. "Variable selection via nonconcave penalized likelihood and its oracle properties." *Journal of the American statistical Association* 96(456):1348–1360.

Figueiredo, Mario. 2004. Lecture Notes on the EM Algorithm. Lecture notes. Instituto de Telecomunicacoes, Instituto Superior Tecnico.

Foster, J. C., J. M. Taylor and S. J. Ruberg. 2011. "Subgroup identification from randomized clinical trial data." *Statistics in Medicine* 30(2867-2880).

Gelman, Andrew. 2006. "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)." *Bayesian Analysis* 1(3):515–534.

Gelman, Andrew, Aleks Jakulin, Maria Grazia Pittau and Yu-Sung Su. 2008. "A weakly informative default prior distribution for logistic and other regression models." *Annals of Applied Statistics* 2(4):1360–1383.

Gelman, Andrew and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambrdige University Press.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari and Donald B. Rubin. 2014. *Bayesian Data Analysis*. Text in Statistical Science Series CRC Press.

Gill, Jeff. 2014. *Bayesian Methods: A Social and Behavioral Sciences Approach*. 3rd ed. CRC Press.

Gillen, B., S. Montero, H.R. Moon and M. Shum. 2016. "BLP-Lasso for Aggregate Discrete Choice Models Applied to Elections with Rich Demographic Covariates." Working paper.

Green, Donald P. and Holger L. Kern. 2012. "Modeling heterogeneous treatment effects in survey experiments with Bayesian Additive Regression Trees." *Public Opinion Quarterly* 76:491–511.

Griffin, J. E. and P. J. Brown. 2010. "Inference with normal-gamma prior distributions in regression problems." *Bayesian Analysis* 5(1):171–188.

Griffin, J. E. and P. J. Brown. 2012. "Structuring shrinkage: some correlated priors for regression." *Biometrika* 99(2):481–487.

Grimmer, Justin, Solomon Messing and Sean Westwood. 2014. "Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods.".

Hahn, P Richard and Carlos M Carvalho. 2015. "Decoupling shrinkage and selection in Bayesian linear models: a posterior summary perspective." *Journal of the American Statistical Association* 110(509):435–448.

Hainmueller, Jens and Chad Hazlett. 2013. "Kernel Regularized Least Squares: Reducing Misspecification Bias with a Flexible and Interpretable Machine Learning Approach." *Political Analysis* .

Hainmueller, Jens, Daniel J Hopkins and Teppei Yamamoto. 2014. "Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments." *Political Analysis* 22(1):1–30.

Hans, Chris. 2009. "Bayesian lasso regression." *Biometrika* 96(4):835–845.

Harding, Matthew and Carlos Lamarche. 2016. "Penalized Quantile Regression with Semiparametric Correlated Effects: An Application with Heterogeneous Preferences." *Journal of Applied Econometrics* Forthcoming.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2010. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer-Verlag.

Imai, Kosuke and Aaron Strauss. 2011. "Estimation of Heterogeneous Treatment Effects from Randomized Experiments, with Application to the Optimal Planning of the Get- Out-the-Vote Campaign." *Political Analysis* 19(1):1–19.

Imai, Kosuke and Marc Ratkovic. 2013. "Estimating treatment effect heterogeneity in randomized program evaluation." *The Annals of Applied Statistics* 7(1):443–470.

Jackman, Simon. 2009. *Bayesian Analysis for the Social Sciences.* Wiley.

Jaynes, E. T. 1982. "On the rationale of maximum-entropy methods." *Proceedings of the IEEE* 70(939–952).

Kang, Jian and Jian Guo. 2009. "Self-adaptive Lasso and its Bayesian Estimation." Working Paper.

Kenkel, Brenton and Curtis Signorino. 2012. "A Method for Flexible Functional Form Estimation: Bootstrapped Basis Regression with Variable Selection." Working paper.

Kyung, Minjung, Jeff Gill, Malay Ghosh and George Casella. 2010. "Penalized Regression, Standard Errors, and Bayesian Lassos." *Bayesian Analysis* 5(2):369–412.

Kyung, Minjung, Jeff Gill, Malay Ghosh, George Casella et al. 2010. “Penalized regression, standard errors, and Bayesian lassos.” *Bayesian Analysis* 5(2):369–411.

Leeb, Hannes and Benedikt Potscher. 2008. “Sparse Estimators and the Oracle Property, or the Return of Hodges Estimator.” *Journal of Econometrics* 142:201–211.

Leeb, Hannes, Benedikt Potscher and Karl Ewald. 2015. “On Various Confidence Intervals Post-Model-Selection.” *Statistical Science* 30(2):216–227.

Leng, Chenlei, Minh-Ngoc Tran and David Nott. 2014. “Bayesian Adaptive LASSO.” *Annals of the Institute of Statistical Mathematics* 66(2):221–244.

Lipkovich, I., A. Dmitrienko, J. Denne and G. Enas. 2011. “Subgrosup identification based on differential effect search—A recursive partitioning method for establishing response to treatment in patient subpopulations.” *Statistics in Medicine* 30:2601–2621.

Liu, H. and B. Yu. 2013. “Asymptotic properties of Lasso+mLS and Lasso+Ridge in sparse high-dimensional linear regression.” *Electronic Journal of Statistics* 7(3124–3169).

Lockhart, Richard, Jonathan Taylor, Ryan J. Tibshirani and Robert Tibshirani. 2014. “A significance test for the lasso.” *The Annals of Statistics* 42(2):413–468.

Loh, Wei-Yin, Xu Heb and Michael Manc. 2015. “A regression tree approach to identifying subgroups with differential treatment effects.” *Statistics in Medicine* 34:1818–1833.

Minnier, Jessica, Lu Tian and Tianxi Cai. 2011. “A perturbation method for inference on regularized regression estimates.” *Journal of the American Statistical Association* 106(496).

Mitchell, T.J. and J. J. Beauchamp. 1988. “Bayesian Variable Selection in Linear Regression.” *Journal of the Americal Statistical Association* 83(404):1023–1032.

O’Hara, R. B. and M. J. Silanapaa. 2009. “A Review of Bayesian Variable Selection Methods: What, How and Which.” *Bayesian Analysis* 4(1):85–118.

Park, Trevor and George Casella. 2008. “The bayesian lasso.” *Journal of the American Statistical Association* 103(482):681–686.

Polson, Nicholas and James Scott. 2012. “Local shrinkage rules, Levy processes and regularized regression.” *Journal of the Royal Statistical Society, Series B* 74(2):287–311.

Potscher, Benedikt and Hannes Leeb. 2009. "On the Distribution of Penalized Maximum Likelihood Estimators: The LASSO, SCAD, and Thresholding." *Journal of Multivariate Analysis* 100(9):2065–2082.

Ratkovic, Marc and Dustin Tingley. Replication Data for: Sparse Estimation and Uncertainty with Application to Subgroup Analysis. doi:10.7910/DVN/RNMB1Q, Harvard Dataverse, September 6, 2016.

Stewart, Brandon M. Working Paper. "Latent Factor Regressions for the Social Sciences.".

Strezhnev, Anton, Jens Hainmueller, Daniel Hopkins and Teppei Yamamoto. 2014. *cjoint: AMCE Estimator for Conjoint Experiments.* R package version 1.0.3.

Su, Xiaogang, Chih-Ling Tsai, Hansheng Wang, David M. Nickerson and Bogong Li. 2009. "Subgroup Analysis via Recursive Partitioning." *Journal of Machine Learning Research* 10:141–158.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society, Series B.* 58:267–88.

Tierney, Luke. 1994. "Markov Chains for Exploring Posterior Distributions." *The Annals of Statistics* 22(4):1701–1728.

Tingley, Dustin and Michael Tomz. 2013. "Conditional cooperation and climate change." *Comparative Political Studies* p. 0010414013509571.

Tripathi, Gautam. 1999. "A matrix extension of the Cauchy-Schwarz inequality." *Economics Letters* 63:1–3.

Wager, S. and S. Athey. 2015. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests." Working paper.

Wang, Hansheng and Chenlei Leng. 2007. "Unified LASSO Estimation by Least Squares Approximation." *Journal of the American Statistical Association* 102(479):1039–1048.

West, M. 1987. "On Scale Mixtures of Normal Distributions." *Biometrika* 74:646–648.

Zou, Hui. 2006. "The Adaptive Lasso and Its Oracle Properties." *Journal of the American Statistical Association* 101(476):1418–1429.

Zou, Hui, Trevor Hastie and Robert Tibshirani. 2007. "On the degrees of freedom of the lasso." *The Annals of Statistics* 35(5):2173–2192.

A Proof of Relative Efficiency of Oracle Estimator and OLS.

Proof: Denote as X_S the submatrix of X for which $\beta_k \neq 0$ and the Gram matrix for X as

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} X_i^\top X_i = \Sigma_X \tag{64}$$

and in block-partition form

$$\Sigma_X = \begin{pmatrix} \Sigma_{SS} & \Sigma_{SS^\complement} \\ \Sigma_{SS^\complement}^\top & \Sigma_{S^\complement S^\complement} \end{pmatrix} \tag{65}$$

Σ_X is invertible, since the least squares estimate exists and is unique. Since Σ_X is invertible, every square submatrix of Σ_X is also invertible.

The asymptotic relative efficiency of the least squares estimate and Oracle estimate is then

$$\lim_{N\to\infty} \frac{\frac{\sigma^2}{N}\mathrm{Tr}\left\{\Sigma^{-1}\right\}}{\frac{\sigma^2}{N}\mathrm{Tr}\left\{\Sigma_{SS}^{-1}\right\}} = \frac{\mathrm{Tr}\left\{\Sigma^{-1}\right\}}{\mathrm{Tr}\left\{\Sigma_{SS}^{-1}\right\}} \tag{66}$$

By the block inverse partition formula,

$$\mathrm{Tr}\left(\Sigma_X^{-1}\right) = \mathrm{Tr}\left\{ \begin{pmatrix} \Sigma_{SS} & \Sigma_{SS^\complement} \\ \Sigma_{SS^\complement}^\top & \Sigma_{S^\complement S^\complement} \end{pmatrix}^{-1} \right\} \tag{67}$$

$$= \mathrm{Tr}\left\{ \left(\Sigma_{SS} - \Sigma_{SS^\complement}\Sigma_{S^\complement S^\complement}^{-1}\Sigma_{SS^\complement}^\top \right)^{-1} \right\} + \mathrm{Tr}\left\{ \left(\Sigma_{S^\complement S^\complement} - \Sigma_{SS^\complement}^\top\Sigma_{S^\complement S^\complement}^{-1}\Sigma_{SS^\complement} \right)^{-1} \right\} \tag{68}$$

Consider the first summand inside the parentheses on the r.h.s. and apply Morrison-Woodbury-Sherman

$$\left(\Sigma_{SS} - \Sigma_{SS^\complement}\Sigma_{S^\complement S^\complement}^{-1}\Sigma_{SS^\complement}^\top \right)^{-1} = \Sigma_{SS}^{-1} + \Sigma_{SS}^{-1}\Sigma_{SS^\complement}\left(\Sigma_{S^\complement S^\complement} - \Sigma_{SS^\complement}^\top\Sigma_{SS}^{-1}\Sigma_{SS^\complement}\right)^{-1}\Sigma_{SS^\complement}^\top\Sigma_{SS}^{-1} \tag{69}$$

By Cauchy-Schwarz, the term $\Sigma_{S^\complement S^\complement} - \Sigma_{SS^\complement}^\top \Sigma_{SS}^{-1}$ is positive semi-definite, see e.g. Tripathi (1999, esp. the last line of the proof of Theorem 1.1.). By symmetry, we get an analogous result for the second summand in side the trace operator,

This gives

$$\begin{aligned} \mathrm{Tr}\left(\Sigma_X^{-1}\right) = \mathrm{Tr}\left(\Sigma_{SS}^{-1}\right) + \mathrm{Tr}\left(\Sigma_{S^\complement S^\complement}^{-1}\right) + \quad & \quad (70) \\ \mathrm{Tr}\left(\Sigma_{SS}^{-1}\Sigma_{SS^\complement}\left(\Sigma_{S^\complement S^\complement} - \Sigma_{SS^\complement}^\top\Sigma_{SS}^{-1}\Sigma_{SS^\complement}\right)^{-1}\Sigma_{SS^\complement}^\top\Sigma_{SS}^{-1}\right) + & \\ \mathrm{Tr}\left(\Sigma_{S^\complement S^\complement}^{-1}\Sigma_{SS^\complement}^\top\left(\Sigma_{SS} - \Sigma_{SS^\complement}\Sigma_{S\complement S\complement}^{-1}\Sigma_{SS^\complement}^\top\right)^{-1}\Sigma_{SS^\complement}\Sigma_{S\complement S\complement}^{-1}\right) & \\ \geq \mathrm{Tr}\left(\Sigma_{SS}^{-1}\right) \quad & \quad (71) \end{aligned}$$

and therefore an estimator with the Oracle Property is asymptotically more efficient than least squares.

To establish when equality holds, if $X = X_S$, then clearly the asymptotic relative efficiency is 1. For only if, the inequality above is an equality only when $\mathrm{Tr}\left(\Sigma^{-1}_{S^\complement S^\complement}\right) = 0$, which is not possible unless $X = X_S$.

B Preliminaries

We offer three sets of preliminary results. First, we show that the weights, $\widehat{w}_k$, and magnitude of $|\widehat{\beta}_k|$ are inversely related. Second, we formally differentiate between "large" and "small" estimates. This will help us derive bounds on $\widehat{w}_k$. Third, we provide a bound on $\widehat{\lambda}$. Note that we refer to the k^{th} order statistic of vector a as $a_{(k)}$, where $a_{(1)}$ is the smallest element of a.

B.1 Inverse relationship between weights and effect size.

Proposition 5

$$\frac{\partial \widehat{w}_k}{\partial |\widehat{\beta}_k|} = -\widehat{\lambda}\sqrt{\widehat{\frac{1}{\sigma^2}}}\mathrm{Var}(w_k|\cdot) < 0. \tag{72}$$

Derivation: *The weights are calculated as*

$$\widehat{w}_k = \mathbb{E}(w_k|\cdot) = \frac{\int_{w=0}^{\infty} w e^{-w^{\widehat{\gamma}} - \widehat{\lambda} w \sqrt{\widehat{\frac{1}{\sigma^2}}}|\widehat{\beta}_k|} dw}{\int_{w=0}^{\infty} e^{-w^{\widehat{\gamma}} - \widehat{\lambda} w \sqrt{\widehat{\frac{1}{\sigma^2}}}|\widehat{\beta}_k|} dw}. \tag{73}$$

Denote as $A = e^{-w^{\widehat{\gamma}} - \widehat{\lambda} w \sqrt{\widehat{\frac{1}{\sigma^2}}}|\widehat{\beta}_k|}$. *Then,*

$$\frac{\partial \widehat{w}_k}{\partial |\widehat{\beta}_k|} = \frac{-\int_{w=0}^{\infty} A dw \times \int_{w=0}^{\infty} w^2 \widehat{\lambda}\sqrt{\widehat{\frac{1}{\sigma^2}}} A dw + \int_{w=0}^{\infty} w A dw \int_{w=0}^{\infty} w \widehat{\lambda}\sqrt{\widehat{\frac{1}{\sigma^2}}} A dw}{\left(\int_{w=0}^{\infty} A dw\right)^2} \tag{74}$$

$$= -\widehat{\lambda}\sqrt{\widehat{\frac{1}{\sigma^2}}}\left\{\frac{\int_{w=0}^{\infty} w^2 A dw}{\int_{w=0}^{\infty} A dw} - \left(\frac{\int_{w=0}^{\infty} w A dw}{\int_{w=0}^{\infty} A dw}\right)^2\right\} \tag{75}$$

$$= -\widehat{\lambda}\sqrt{\widehat{\frac{1}{\sigma^2}}}\mathrm{Var}(w_k|\cdot) \tag{76}$$

where moving the derivative under the integral in the first line is allowed by the monotone convergence theorem.

This result allows us to associate the largest weight, $\widehat{w}_{(K)}$ with the smallest estimate, $\widehat{\beta}_{(1)}$, the second largest weight with the second smallest estimate, and so on. In general, weight $\widehat{w}_{(k)}$ is associated with $|\widehat{\beta}|_{(K-k+1)}$

B.2 Separating large and small weights and effect estimates.

We next distinguish between weights near zero from weights close to the maximal value $\overline{\widehat{\gamma}}$. This is our equivalent of either assuming the estimates are "well-separated" (Belloni and Chernozhukov, 2013), or separating "relevant" from "irrelevant" effects (Buhlmann and van de Geer, 2013). The key difference is that these authors separate large and small "true" effects, whereas we separate large and small estimated effects. As is common in the literature, our bounds will be more informative the better we can distinguish between zero- and non-zero effect estimates.

We separate the weights into two groups. In the kernel for $\Pr(w_k|\cdot)$, the numerator in Equation 73, is approximately exponential for large $|\beta_k|$, small w_k, and is approximately constant for $|\beta_k| \approx 0$, w_k large. Define as

$$p_k(C_1, C_2) = \max\left\{\Pr\left(\widehat{w}_k > \frac{C_1 \log(\widehat{S})}{\lambda\widehat{\sigma}|\widehat{\beta}_k|}\right), \Pr\left(\widehat{w}_k < C_2\overline{\widehat{\gamma}}\right)\right\};\ C_1 > 0, 0 < C_2 < 1 \tag{77}$$

where the first inequality allows us to bound with some high probability small weights from above and the second, larger weights from below. We use this distinction to differentiate between weights tending to zero (the lefthand set) and those tending to the maximum (the righthand set).

$$\widehat{S} = \left\{k : \Pr\left(\widehat{w}_k > \frac{C_1 \log\left(|\widehat{S}|\right)}{\lambda\widehat{\sigma}|\widehat{\beta}_k|}\right) < \Pr\left(\widehat{w}_k < C_2\overline{\widehat{\gamma}}\right)\right\}. \tag{78}$$

The $\log(|\widehat{S}|)$ term on the left comes from using the union bound applied to $\{p_k\}_{k=1}^K$ and a subexponential (rather than subgaussian) bound applied to each value p_k, as the kernel is approximately exponential in this range. Define

$$\Pr(\max(p_k) > C_3 \log(K)) = p_w(C_1, C_2, C_3). \tag{79}$$

such that, with probablity at least $p_w(C_1, C_2, C_3)$, the weights can be bounded by one of the bounds above, i.e. is either "small" or "large."

Lastly, denote as $\underline{C}_1$ the value that satisfies

$$\Pr\left(\widehat{w}_k > \frac{C_1 \log(|\widehat{S}|)}{\widehat{\lambda}\widehat{\sigma}|\widehat{\beta}_k|}\right) = \Pr\left(\widehat{w}_k \leq \frac{\underline{C}_1}{\widehat{\lambda}\widehat{\sigma}|\widehat{\beta}_k| \log(|\widehat{S}|)}\right) \tag{80}$$

which will give us a lower bound on all $\widehat{w}_k$ with probability at least $p_w(C_1, C_2, C_3)$.

B.3 Bounding the tuning parameter $\widehat{\lambda}$.

Given the results above, we can bound $\widehat{\lambda}$. For the Oracle results below, we need to bound $\widehat{\lambda}$ from below, though we note that a similar upper bound of the same order of N, K can be found using the strategy below.

As $\lambda^2|\cdot \sim \Gamma(\sqrt{N}K, \frac{1}{2}\sum_{k=1}^{K}\widehat{\tau}_k^2 + \rho)$, this gives

$$\widehat{\lambda^2} = \frac{\sqrt{N}K}{\frac{1}{2}\sum_{k=1}^{K}\widehat{\tau}_k^2 + \rho}. \tag{81}$$

Change of variables gives $\lambda|\cdot \sim$ generalizedGamma $\left(2 \times (\frac{1}{2}\sum_{k=1}^{K}\widehat{\tau}_k^2 + \rho)^{-1/2}, 2\sqrt{N}K, 2\right)$, which gives the estimate

$$\widehat{\lambda} = \frac{\widetilde{\Gamma}(\sqrt{N}K + 1/2)/\widetilde{\Gamma}(\sqrt{N}K)}{\sqrt{\frac{1}{2}\sum_{k=1}^{K}\widehat{\tau}_k^2 + \rho}} \tag{82}$$

with $\widetilde{\Gamma}()$ the Gamma function. Note $\widehat{\lambda^2} \geq (\widehat{\lambda})^2$ and if $\sqrt{N}K > 1$, then $\Gamma(3/2)^2(\widehat{\lambda})^2 = \frac{4}{\pi}(\widehat{\lambda})^2 > \widehat{\lambda^2}$. Lastly,

$$1/\tau_k^2|\cdot \sim \text{InvGaussian}\left(\lambda w_k \sigma / |\beta_k|, w_k^2\lambda^2\right) \Rightarrow \tag{83}$$

$$\sum_{k=1}^{K}\widehat{\tau}_k^2 = \sum_{k=1}^{K}\frac{|\widehat{\beta}_k|}{\widehat{\lambda}\widehat{w}_k\widehat{\sigma}} + \frac{1}{\widehat{\lambda^2}\widehat{w}_k^2} \tag{84}$$

and we use the bound

$$\sum_{k=1}^{K}\widehat{\tau}_k^2 \leq \frac{|\widehat{S}| \times |\widehat{\beta}_{(K)}|}{\widehat{\lambda}\widehat{w}_{(1)}\widehat{\sigma}} + \frac{|\widehat{S}|}{(\widehat{\lambda})^2\widehat{w}_{(1)}^2} + \frac{\left(K - |\widehat{S}|\right)|\widehat{\beta}|_{\left(K-|\widehat{S}|-1\right)}}{\widehat{\lambda}\widehat{w}_{\left(K-|\widehat{S}|-1\right)}\widehat{\sigma}} + \frac{\left(K - |\widehat{S}|\right)}{(\widehat{\lambda})^2\widehat{w}^2_{\left(K-|\widehat{S}|-1\right)}} \tag{85}$$

$$\leq \frac{|\widehat{S}|\widehat{\beta}_{(K)}^2}{\underline{C}_1\log(\widehat{S})} + \frac{|\widehat{S}|\widehat{\sigma}^2\widehat{\beta}_k^2}{\underline{C}_1^2\log(\widehat{S})^2} + \frac{\left(K - |\widehat{S}|\right)|\widehat{\beta}|_{\left(K-|\widehat{S}|-1\right)}}{\widehat{\lambda}C_2\widehat{\widehat{\gamma}}\widehat{\sigma}} + \frac{\left(K - |\widehat{S}|\right)}{(\widehat{\lambda})^2C_2^2\widehat{\widehat{\gamma}}^2} \tag{86}$$

$$= \frac{|\widehat{S}|\widehat{\beta}_{(K)}^2\left(\underline{C}_1\log(\widehat{S}) + \widehat{\sigma}^2\right)}{\underline{C}_1^2\log(\widehat{S})^2} + \frac{\left(K - |\widehat{S}|\right)|\widehat{\beta}|_{\left(K-|\widehat{S}|-1\right)}}{\widehat{\lambda}C_2\widehat{\widehat{\gamma}}\widehat{\sigma}} + \frac{\left(K - |\widehat{S}|\right)}{(\widehat{\lambda})^2C_2^2\widehat{\widehat{\gamma}}^2} \tag{87}$$

The first line follows from the inverse relationship between $|\widehat{\beta}_k|$ and $\widehat{w}_k$; the second comes from the lower bounds on $\widehat{w}_k$ in $\widehat{S}$ and $\widehat{S}^{\complement}$. The third line is just simplifying.

Combining inequalities gives

$$\frac{4}{\pi}(\widehat{\lambda})^2 \geq \widehat{\lambda^2} = \frac{\sqrt{N}K}{\frac{1}{2}\sum_{k=1}^{K}\widehat{\tau}_k^2 + \rho} \tag{88}$$

$$\Rightarrow (\widehat{\lambda})^2 \geq \frac{\pi}{4} \times \frac{\sqrt{N}K}{\frac{|\widehat{S}|\widehat{\beta}^2_{(K)}\left(\underline{C}_1 \log(\widehat{S}) + \widehat{\sigma}^2\right)}{2\underline{C}_1^2 \log(\widehat{S})^2} + \frac{\left(K - |\widehat{S}|\right)|\widehat{\beta}|_{\left(K - |\widehat{S}| - 1\right)}}{2\widehat{\lambda}C_2\widehat{\overline{\gamma}}\sigma} + \frac{\left(K - |\widehat{S}|\right)}{2(\widehat{\lambda})^2 C_2^2 \widehat{\overline{\gamma}}^2} + \rho} \tag{89}$$

$$\Rightarrow \widehat{\lambda} \geq \frac{\pi}{4} \times \frac{\sqrt{N}K}{\frac{\widehat{\lambda}|\widehat{S}|\widehat{\beta}^2_{(K)}\left(\underline{C}_1 \log(\widehat{S}) + \widehat{\sigma}^2\right)}{2\underline{C}_1^2 \log(\widehat{S})^2} + \frac{\left(K - |\widehat{S}|\right)|\widehat{\beta}|_{\left(K - |\widehat{S}| - 1\right)}}{2C_2\widehat{\overline{\gamma}}\widehat{\sigma}} + \frac{\left(K - |\widehat{S}|\right)}{2\widehat{\lambda} C_2^2 \widehat{\overline{\gamma}}^2} + \rho\widehat{\lambda}} \tag{90}$$

where the second line comes from substituting from Inequality 87 and the third from multiplying both sides by $1/\widehat{\lambda}$. Cross-multiplying gives a quadratic equation in $\widehat{\lambda}$ of the form $\widetilde{a}(\widehat{\lambda})^2 + \widetilde{b}\widehat{\lambda} + \widetilde{c} > 0$ where[25]

$$\widetilde{a} = \frac{|\widehat{S}|\widehat{\beta}^2_{(K)}\left(\underline{C}_1 \log(\widehat{S}) + \widehat{\sigma}^2\right)}{2\underline{C}_1^2 \log(\widehat{S})^2} + \rho \tag{91}$$

$$\widetilde{b} = \frac{\left(K - |\widehat{S}|\right)|\widehat{\beta}|_{\left(K - |\widehat{S}| - 1\right)}}{2C_2\widehat{\overline{\gamma}}\widehat{\sigma}} \tag{92}$$

$$\widetilde{c} = -\left(\frac{\pi}{4}\sqrt{N}K - \frac{\left(K - |\widehat{S}|\right)}{2C_2^2\widehat{\overline{\gamma}}^2}\right). \tag{93}$$

The quadratic equation gives

$$\widehat{\lambda} \geq \frac{-\frac{\left(K - |\widehat{S}|\right)|\widehat{\beta}|_{\left(K - |\widehat{S}| - 1\right)}}{2C_2\widehat{\overline{\gamma}}\widehat{\sigma}} + \sqrt{\left\{\frac{\left(K - |\widehat{S}|\right)|\widehat{\beta}|_{\left(K - |\widehat{S}| - 1\right)}}{2C_2\widehat{\overline{\gamma}}\widehat{\sigma}}\right\}^2 + 4\left\{\frac{\widehat{\lambda}|\widehat{S}|\widehat{\beta}^2_{(K)}\left(\underline{C}_1 \log(\widehat{S}) + \widehat{\sigma}^2\right)}{2\underline{C}_1^2 \log(\widehat{S})^2} + \rho\right\} \times \left\{\frac{\pi}{4}\sqrt{N}K - \frac{\left(K - |\widehat{S}|\right)}{2C_2^2\widehat{\overline{\gamma}}^2}\right\}}}{2\frac{|\widehat{S}|\widehat{\beta}^2_{(K)}\left(\underline{C}_1 \log(\widehat{S}) + \widehat{\sigma}^2\right)}{2\underline{C}_1^2 \log(\widehat{S})^2} + 2\rho} \tag{94}$$

which, for growing N and K, is of order $N^{1/4}K^{1/2}$ by the bound in 94.

C The Oracle Property for LASSOplus

We derive conditions for when LASSOplus possesses the Oracle Property in the case of K, N growing. The proof progresses in four steps. First, we derive the conditions for which the posterior density β converges to the same distribution as the least squares estimate. Second, we show that β^{plus} is consistent in variable selection. Third, we show how a model without the adaptive weights will not be consistent in variable selection. Fourth, we combine the results using Slutsky's Theorem.

[25] We use the convention $0 \log 0 = 0$

Asymptotic behavior of β. The conditional posterior density of β is multivariate normal and given in Equation (26). The vector β shares a limiting distribution with the least squares estimate when $\lim_{N\to\infty}\left(\frac{1}{N}X^\top X+\frac{1}{N}D_\tau^{-1}\right)^{-1}=\lim_{N\to\infty}\left(\frac{1}{N}X^\top X\right)^{-1}$. Therefore, it suffices to identify the rates of N, K for which $\frac{1}{\tau_k^2}=o_p(N)$ for all k.

Since $1/\tau_k^2$ has positive support, we need only find when its posterior mean

$$\mathbb{E}\left(\frac{1}{\tau_k^2}\bigg|\cdot\right)=\frac{\lambda w_k\sigma}{|\beta_k|} \tag{95}$$

grows at a rate less than N. Consider first the case where $|\beta_k|$ is converging to a number away from zero. Then, we know λw_k is $O_p(1)$, since the kernel for w_k is approximately exponential in the case of $\lambda|\beta_k|$ large. Therefore, $1/\tau_k^2=o_p(N)$. Consider the next case, where $|\beta_k|$ is converging to zero. Then, if β_k is consistent, $w_k\to\overline{\gamma}=O_p(1)$ and β_k will go to zero at $1/\sqrt{N}$. This implies $1/N\times\lambda w_k\sigma/|\beta_k|$ is of order $K^{1/2}N^{-1/4}$ which must go to zero. Therefore, β is consistent and shares the same limiting distribution as the least squares estimator when $K^2/N\to 0$.

Consistency in variable selection for β^{plus} Next, we give the requirements on N, K such that plim $\Pr\left(\beta_k^{plus}=0\right)=\mathbf{1}\left(\beta_k^o=0\right)$. We assume $K^2/N\to 0$. Under this condition, β_{-k} is consistent for β_{-k}^o, and we can write $\widehat{\beta}_k^{sp}=\beta_k^o+u_k/\sqrt{N}$ for a sufficiently large N with $\mathrm{Var}(u_k)<\infty$.

The asymptotic probability of a variable being selected is

$$\text{plim }\Pr(\beta_k^{plus}\neq 0)=\lim_{N\to\infty}\Pr\left(\left|\beta_k^o+\frac{u_k}{\sqrt{N}}\right|>\frac{\lambda w_k\sigma_{sp}}{N-1}\right) \tag{96}$$

Recall $\sigma_{sp}=\mathcal{O}_p\left(N^\alpha\right)$ and $\lambda=O_p(N^{1/4}K^{1/2})$. Consider the case $\beta_k^o=0$, such that $w_k\to\overline{\gamma}=O_p(1)$:

$$\begin{aligned}\text{plim }\Pr(\beta_k^{plus}\neq 0|\beta_k^o=0)&=\lim_{N\to\infty}\Pr\left(|u_k|>\sqrt{N}\frac{\lambda w_k\sigma_{sp}}{N-1}\bigg|\beta_k^o=0\right) && (97)\\ &=\lim_{N\to\infty}\Pr\left(|u_k|>C_{u_0}\right) && (98)\\ &=0\text{ if }K^{1/2}N^{\alpha-1/4}\to\infty && (99)\end{aligned}$$

The value C_{u_0} is $O_p(K^{1/2}N^{\alpha-1/4})$, so propertly zeroing out all in-truth-zero effects occurs when $\alpha>1/4$ for fixed K or when K grows in N at any rate when $\alpha=1/4$.

Next, consider the case $\beta_k^0\neq 0$, so $\lambda w_k=O_p(1)$:

$$\begin{aligned}\text{plim }\Pr(\beta_k^{plus}\neq 0|\beta_k^o\neq 0)&=\lim_{N\to\infty}\Pr\left(|\beta_k^o|>\frac{\lambda w_k\sigma_{sp}}{N-1}\bigg|\beta_k^o\neq 0\right) && (100)\\ &=\lim_{N\to\infty}\Pr\left(|u_k|>C_{u_1}\right) && (101)\\ &=1\text{ if }N^{\alpha-1}\to 0. && (102)\end{aligned}$$

since $C_{u_1} = O(N^{\alpha-1})$.

Therefore, β_k^{plus} is consistent for variable selection so long as $K^{1/2}N^{\alpha-1/4} \to \infty$ and $N^{\alpha-1} \to 0$. This will always be achieved for $1 > \alpha \geq 1/4$, if K grows in N, and $1 > \alpha > 1/4$ for fixed K. Under these conditions, β_k^{plus} satisfies the first Oracle condition.

The model with no weights and no variance inflation. We consider the normal LASSO, where $w_k = 1 \; \forall \; k$ and $\alpha = 0$. In this case, β still shares the same limiting distribution as the least squares estimate so long as $K^2/N \to 0$. Setting $\alpha = 0$ in Equation 100 shows that in-truth-zero effects are zeroed out when $K^{1/2}N^{-1/4} \to \infty$, or equivalenlty $K^2/N \to \infty$. But, K^2/N cannot approach both zero and infiinity, so both conditions cannot be met. Therefore, as has been observed several times (Fan and Li, 2001; Buhlmann and van de Geer, 2013; Zou, 2006), the LASSO estimator, if tuned for consistency, will over-select small effects with some positive probability.

The Oracle Property. Denote as S^{plus} the vector which takes on a value of 1 in element k if the effect is selected by the rule in the previous section and a 0 otherwise. Under the conditions on N, K and α given above, we have shown that plim $S_k^{plus} = \mathbf{1}(\beta_k^0 \neq 0)$; $\forall k$. Let $\otimes$ denote the Hadamard (elementwise) product between two vectors. The LASSOplus estimate is $\beta^{plus} = \beta \otimes S^{plus}$, where we have shown that β and the $\widehat{\beta}^{LS}$ share the same limiting distribution. By Slutsky's Theorem, this converges to a normal random variable with mean $\mathbb{E}(\widehat{\beta}^{LS} \otimes S^{plus})$ and variance $\text{Var}(\widehat{\beta}^{LS} \otimes S^{plus})$, which is simply the least squares variance for all non-zero elements of β^o. Taken together, this gives plim $\mathbf{1}(\beta_k^{plus} = 0) = \mathbf{1}(\beta_k^o = 0)$ and

$$\sqrt{N}\left(\widehat{\beta}^{plus}_{S^o} - \beta^o_{S^o}\right) \xrightarrow{\text{d}} \mathcal{N}(\mathbf{0}_{|S^o|}, \Sigma^0_{S^o}) \tag{103}$$

which are the two Oracle Properties.

D The Oracle Inequality for LASSOplus

We shift now from the Oracle Property to the Oracle Inequality. Denote $\widehat{W}$ the $K \times K$ diagonal matrix with $\widehat{W}_{kk} = \widehat{w}_k > 0$ and $\widehat{\delta} = \widehat{\beta} - \beta^o$. Parts of this section follow the argument in Buhlmann and van de Geer (2013).

D.1 Assumptions

Assumption 1 **Weighted Compatibility Condition** *For all $\widehat{\delta}$ in the set that satisfies $||\widehat{\delta}_{S^\complement}||_1 \leq 3||\widehat{\delta}_S||_1^1$, it holds that*

$$||\widehat{W}_S\widehat{\delta}_S||_1^2 \leq \frac{\widehat{\delta}^\top \Sigma_X \widehat{\delta}|S|}{\phi_0^2}. \tag{104}$$

Denoting $\widehat{\widetilde{\delta}} = \widehat{W}_S \widehat{\delta}_S$*, this condition can also be expressed as*

$$||\widehat{\widetilde{\delta}}_S||_1^2 \leq \frac{\widehat{\widetilde{\delta}}^\top \widehat{W}^{-1} \Sigma_X \widehat{W}^{-1} \widehat{\widetilde{\delta}} |S|}{\phi_0^2}. \tag{105}$$

for all $||\widehat{W}_{S^\complement}^{-1} \widehat{\delta}_{S^\complement}||_1 \leq 3||\widehat{W}_S^{-1} \widehat{\delta}_S||_1^1$

For a variation of this compatibility condition expressed as a restricted eigenvalue condition, see Bickel, Ritov and Tsybakov (2009); Belloni and Chernozhukov (2013). Some version of this assumption on the design is standard in the literature, as it is used to combine the L_2 empirical loss with the L_1 penalty. We illustrate below.

We make the following assumption to simplify the analysis. The assumption has two implications. The first, common in the literature (Liu and Yu, 2013), assumes iid Gaussian errors. The second restricts this analysis to the case of in-truth sparse, linear models. The former can be relaxed using Talagrand style bounds, and the second can be relaxed to include a nonparametric setup. We reserve both extensions to future work though we note that the development builds off what is done here (for examples, see Buhlmann and van de Geer, 2013; Belloni and Chernozhukov, 2013).

Assumption 2 **Data-Generating Process** *The data are generated as*

$$Y_i = X_i^\top \beta^o + \epsilon_i \tag{106}$$

such that $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ *and* $\lim_{K \to \infty} |S|/K = 0$ *where* $|S| = \sum_{i=1}^K \mathbf{1}(\beta_k^0 \neq 0)$*. We further assume* X_i *and* β_o *are finite and bounded.*

Our final assumption is similar to the assumption that the model be in-truth sparse, but instead we need an assumption on the distribution of estimated effects. We require that estimated weights be distributed with a few small values and a large number of large ones, or equivalently that there be a few large estimated effects and a large number of small ones.

Assumption 3 **Sufficient sparsity condition** *The thresholds that determine* $|\widehat{S}|$ *and* $|\widehat{S}^\complement|$ *are selected such that the weights satisfy*

$$\sum_{k \in \widehat{S}} \exp\left(-\frac{1}{32} \frac{\widehat{\lambda}^2 \widehat{w}_k^2 \widehat{\sigma}^2}{(N-1)\sigma^2}\right) \leq \sum_{k' \in \widehat{S}^\complement} \left\{ \exp\left(-\frac{1}{32} \frac{\widehat{\lambda}^2 \widehat{w}_{(K-|\widehat{S}|-1)}^2 \widehat{\sigma}^2}{(N-1)\sigma^2}\right) - \exp\left(-\frac{1}{32} \frac{\widehat{\lambda}^2 \widehat{w}_{k'}^2 \widehat{\sigma}^2}{(N-1)\sigma^2}\right) \right\} \tag{107}$$

The better the non-zero and in-truth-zero effects are separated, the wider the gap above. Trivially, taking $|\widehat{S}| = 0 \rightarrow \widehat{w}_{(1)} = \infty$ satisfies the bound, with $K > 1$. The assumption is most easily satisfied when there are a few small large effects and a large number of small effects, as in the cases we consider.

For clarity, we present the probabilistic bounds derived in Appendix B that we use in the proofs below:

1. **Probability of bounds on large or small estimate**: $1 - p_w(C_1, C_2, C_3)$
2. **Probability of bounding residual variance**: $1 - p_\epsilon(C_\epsilon) = \Pr(C_\epsilon \widehat{\sigma}^2 > \sigma^2)$
3. **Probability of bounding tuning parameter**: $1 - p_\lambda(C_\lambda) = \Pr(\widehat{\lambda}^2 > C_\lambda \sqrt{N} K)$

By the union bound, the probability of all of these conditions holding is at least $1 - p_w(C_1, C_2, C_3) - p_\epsilon(C_\epsilon) - p_\lambda(C_\lambda)$.

D.2 The consistent model

Bounding the excess risk. Start with the excess risk for the consistent model and simplify, as

$$||Y - X\widehat{\beta}||_2^2 + \widehat{\lambda}\widehat{\sigma}||\widehat{W}\widehat{\beta}||_1^1 \leq ||Y - X\widehat{\beta}^o||_2^2 + \widehat{\lambda}\widehat{\sigma}||\widehat{W}\beta^o||_1^1 \tag{108}$$

$$||X\widehat{\delta}||_2^2 + \widehat{\lambda}\widehat{\sigma}||\widehat{W}\widehat{\beta}||_1^1 \leq 2\epsilon^\top X\widehat{W}^{-1}\widehat{W}(\widehat{\beta} - \beta^o) + \widehat{\lambda}\widehat{\sigma}||\widehat{W}\beta^o||_1^1 \tag{109}$$

$$||X\widehat{\delta}||_2^2 + \widehat{\lambda}\widehat{\sigma}||\widehat{W}\widehat{\beta}||_1^1 \leq 2\sum_{k=1}^{K} |\epsilon^\top X_k/(\widehat{w}_k\widehat{\sigma})||\widehat{\sigma}\widehat{w}_k\widehat{\delta}_k| + \widehat{\lambda}\widehat{\sigma}||\widehat{W}\beta^o||_1^1. \tag{110}$$

We follow Buhlmann and van de Geer (2013) and select λ so as to bound $4\sum_{k=1}^{K} |\epsilon^\top X_k/(\widehat{w}_k\widehat{\sigma})|$, so that we dominate the random part of the problem. For arbitrary k, Assumption 2 allows us to use symmetrization and the Chernoff bound to get

$$\Pr(4|\epsilon^\top X_k/(\widehat{w}_k\widehat{\sigma})| > \widehat{\lambda}) \leq 2\exp\left(-\frac{1}{32}\frac{\widehat{\lambda}^2\widehat{w}_k^2\widehat{\sigma}^2}{(N-1)\sigma^2}\right). \tag{111}$$

The union bound, $\widehat{w}_{(1)} \leq \widehat{w}_{(k')}$; $k' \in \widehat{S}$, and $\widehat{w}_{(K-|\widehat{S}|-1)} \leq \widehat{w}_k$; $k \in \widehat{S}^{\complement}$ gives

$$\Pr(4 \max_{1\leq k\leq K} |\epsilon^\top X_k/(\widehat{w}_k\widehat{\sigma})| > \widehat{\lambda}) \leq \sum_{1\leq k\leq K} 2\exp\left(-\frac{1}{32}\frac{\widehat{\lambda}^2\widehat{w}_k^2\widehat{\sigma}^2}{(N-1)\sigma^2}\right) \tag{112}$$

$$= \sum_{k\in\widehat{S}} 2\exp\left(-\frac{1}{32}\frac{\widehat{\lambda}^2\widehat{w}_k^2\widehat{\sigma}^2}{(N-1)\sigma^2}\right) + \sum_{k'\in\widehat{S}^{\complement}} 2\exp\left(-\frac{1}{32}\frac{\widehat{\lambda}^2\widehat{w}_{k'}^2\widehat{\sigma}^2}{(N-1)\sigma^2}\right) \tag{113}$$

$$\begin{aligned}&= 2(K-|\widehat{S}|)\exp\left(-\frac{1}{32}\frac{\widehat{\lambda}^2\widehat{w}_{(K-|\widehat{S}|-1)}^2\widehat{\sigma}^2}{(N-1)\sigma^2}\right) + \\ &\sum_{k\in\widehat{S}} 2\exp\left(-\frac{1}{32}\frac{\widehat{\lambda}^2\widehat{w}_k^2\widehat{\sigma}^2}{(N-1)\sigma^2}\right) + \sum_{k'\in\widehat{S}^{\complement}} \left\{2\exp\left(-\frac{1}{32}\frac{\widehat{\lambda}^2\widehat{w}_{k'}^2\widehat{\sigma}^2}{(N-1)\sigma^2}\right) - 2\exp\left(-\frac{1}{32}\frac{\widehat{\lambda}^2\widehat{w}_{(K-|\widehat{S}|-1)}^2\widehat{\sigma}^2}{(N-1)\sigma^2}\right)\right\}\end{aligned} \tag{114}$$

$$\leq 2(K-|\widehat{S}|)\exp\left(-\frac{1}{32}\frac{\widehat{\lambda}^2\widehat{w}_{(K-|\widehat{S}|-1)}^2\widehat{\sigma}^2}{(N-1)\sigma^2}\right) \tag{115}$$

$$= 2\exp\left(-\frac{1}{32}\frac{\widehat{\lambda}^2\widehat{w}_{(K-|\widehat{S}|-1)}^2\widehat{\sigma}^2}{(N-1)\sigma^2} + \log(K-|\widehat{S}|)\right) \tag{116}$$

where the second to last line comes from the Assumption 3. Bounding the exponent in the last line gives

$$-\frac{1}{32}\frac{\widehat{\lambda}^2\widehat{w}_{(K-|\widehat{S}|-1)}^2\widehat{\sigma}^2}{(N-1)\sigma^2} + \log(K-|\widehat{S}|) \tag{117}$$

$$\leq -\frac{1}{32}\frac{C_\lambda C_\epsilon C_2 \widehat{\overline{\gamma}}^2 K\sqrt{N-1}}{(N-1)} + \log(K-|\widehat{S}|) \tag{118}$$

with probability at least $1 - p_\lambda(C_\lambda) - p_\epsilon(C_\epsilon) - p_w(C_1, C_2, C_3)$. With the same probability, we want to then bound by the exponent by $-t^2/2$:

$$-\frac{1}{32}\frac{C_\lambda C_\epsilon C_2 \widehat{\overline{\gamma}}^2 K}{\sqrt{N-1}} + \log(K-|\widehat{S}|) \leq -t^2/2 \tag{119}$$

$$\Rightarrow N-1 \leq \left\{\frac{1}{32} \times \frac{C_\lambda C_\epsilon C_2 \widehat{\overline{\gamma}}^2 K}{\frac{t^2}{2} + \log(K-|\widehat{S}|)}\right\}^2 \tag{120}$$

We see then that the LASSOplus consistent model controls the random component, with probability at least $1 - p_\lambda(C_\lambda) - p_\epsilon(C_\epsilon) - p_w(C_1, C_2, C_3) - \exp(-t^2/2)$ so long as Inequality 120 is met. In the limit, this requires N is growing at rate less than $(K/\log(K))^2$.

Geometric bounds. We next move towards the Oracle Inequality, noting that the argument in this section is now standard in the literature(see, e.g. Bickel, Ritov and Tsybakov (2009); Chatterjee (2014); Buhlmann and van de Geer (2013, esp. 6.2, which we follow). Statements in this section hold with probability at least $1 - p_\lambda(C_\lambda) - p_\epsilon(C_\epsilon) - p_w(C_1, C_2, C_3) - \exp(-t^2/2)$ and under the restrictions on N, K directly above.

Continuing from the righthand side of Inequality 110 gives

$$2\sum_{k=1}^{K} |\epsilon^\top X_k/(\widehat{w}_k\widehat{\sigma})||\widehat{\sigma}\widehat{w}_k\widehat{\delta}_k| + \widehat{\lambda}\widehat{\sigma}||\widehat{W}\beta^o||_1^1 \le \frac{1}{2}\widehat{\lambda}\widehat{\sigma}||\widehat{W}\widehat{\delta}||_1^1 + \widehat{\lambda}\widehat{\sigma}||\widehat{W}\beta^o||_1^1 \tag{121}$$

by our probabilistic bound on λ. Continuing Inequality 110 on the lefthand side gives

$$||X\widehat{\delta}||_2^2 + \widehat{\lambda}\widehat{\sigma}||\widehat{W}\widehat{\beta}||_1^1 = \tag{122}$$

$$\begin{aligned} &||X\widehat{\delta}||_2^2 + \widehat{\lambda}\widehat{\sigma}||\widehat{W}_S\widehat{\beta}_S||_1^1 + \widehat{\lambda}\widehat{\sigma}||\widehat{W}_{S^\complement}\widehat{\beta}_{S^\complement}||_1^1 \ge \\ &||X\widehat{\delta}||_2^2 + \widehat{\lambda}\widehat{\sigma}|\widehat{W}_S\beta_S^o|_1^1 - \widehat{\lambda}\widehat{\sigma}||\widehat{W}_S\widehat{\delta}_S||_1^1 + \widehat{\lambda}\widehat{\sigma}||\widehat{W}_{S^\complement}\widehat{\beta}_{S^\complement}||_1^1 \end{aligned} \tag{123}$$

by the triangle inequality. Combining Inequalities 121 and 123 gives

$$||X\widehat{\delta}||_2^2 + \widehat{\lambda}\widehat{\sigma}|\widehat{W}_S\beta_S^o|_1^1 - \widehat{\lambda}\widehat{\sigma}||\widehat{W}_S\widehat{\delta}_S||_1^1 + \widehat{\lambda}\widehat{\sigma}||\widehat{W}_{S^\complement}\widehat{\beta}_{S^\complement}||_1^1 \le \frac{1}{2}\widehat{\lambda}\widehat{\sigma}||\widehat{W}\widehat{\delta}||_1^1 + \widehat{\lambda}\widehat{\sigma}||\widehat{W}\beta^o||_1^1 \tag{124}$$

$$\Rightarrow 2||X\widehat{\delta}||_2^2 - 2\widehat{\lambda}\widehat{\sigma}||\widehat{W}_S\widehat{\delta}_S||_1^1 + 2\widehat{\lambda}\widehat{\sigma}||\widehat{W}_{S^\complement}\widehat{\beta}_{S^\complement}||_1^1 \le \widehat{\lambda}\widehat{\sigma}||\widehat{W}\widehat{\delta}||_1^1 \tag{125}$$

$$\Rightarrow 2||X\widehat{\delta}||_2^2 + \widehat{\lambda}\widehat{\sigma}||\widehat{W}_{S^\complement}\widehat{\beta}_{S^\complement}||_1^1 \le 3\widehat{\lambda}\widehat{\sigma}||\widehat{W}_S\widehat{\delta}_S||_1^1 \tag{126}$$

where the lines come from substitution, simplification, $||\widehat{W}\widehat{\delta}||_1^1 = ||\widehat{W}_S\widehat{\delta}_S||_1^1 + ||\widehat{W}_{S^\complement}\widehat{\delta}_{S^\complement}||_1^1$. This result also gives us $||\widehat{W}_{S^\complement}\widehat{\beta}_{S^\complement}||_1^1 \le 3||\widehat{W}_S\widehat{\delta}_S||_1^1$.

Continuing,

$$\begin{aligned} 2||X\widehat{\delta}||_2^2 + \widehat{\lambda}\widehat{\sigma}||\widehat{W}\widehat{\delta}||_1^1 &= 2||X\widehat{\delta}||_2^2 + \widehat{\lambda}\widehat{\sigma}||\widehat{W}_{S^\complement}\widehat{\beta}_{S^\complement}||_1^1 + \widehat{\lambda}\widehat{\sigma}||\widehat{W}_S\widehat{\delta}_S||_1^1 && (127)\\ &\le 4\widehat{\lambda}\widehat{\sigma}||\widehat{W}_S\widehat{\delta}_S||_1^1 && (128)\\ &\le 4\widehat{\lambda}\widehat{\sigma}\sqrt{\frac{\widehat{\delta}^\top\Sigma_X\widehat{\delta}|S|}{\phi_0^2}} && (129)\\ &= 4\widehat{\lambda}\widehat{\sigma}\sqrt{\frac{\widehat{\delta}^\top X^\top X\widehat{\delta}|S|}{N\phi_0^2}} && (130)\\ &\le ||X\widehat{\delta}||_2^2 + 4\frac{\lambda^2|S|}{N\phi_0^2} && (131) \end{aligned}$$

where we use the inequality $(a-2b)^2 \geq 0 \Rightarrow a^2+4b^2 \geq 4ab$. Simplifying gives a preliminary Oracle Inequality:

$$2||X\widehat{\delta}||_2^2 + \widehat{\lambda}\widehat{\sigma}|\widehat{W\widehat{\delta}}| \leq ||X\widehat{\delta}||_2^2 + 4\frac{\lambda^2|S|}{N\phi_0^2} \tag{132}$$

$$\Rightarrow \frac{1}{N}\left\{|||X\widehat{\delta}||_2^2 + \widehat{\lambda}\widehat{\sigma}|\widehat{W\widehat{\delta}}|\right\} \leq \frac{C_L\lambda^2|S|}{N^2\phi_0^2} \tag{133}$$

where we have inserted C_L since the the 4 is an arbitrary constant that can be adjusted by choosing a different bound on the random component.

Bounding the compatibility condition constant. Unlike the standard LASSO setup, our weighted compatibility condition involves the matrix $\widehat{W_S}$ which may grow of N, K. At the one extreme, the diagonal elements of $\widehat{W_S}$ may all be in $\widehat{S}$, so they are growing as $1/(\lambda\widehat{\sigma}|\beta_k|); k \in \widehat{S}$. At the other extreme, all elements of $\widehat{W_S}$ may be be in $\widehat{S}^{\complement}$, so they are approximately $\widehat{\gamma}$. There then exist constants C_{ϕ_1}, C_{ϕ_2} that satisfy

$$\phi_0^2 \geq C_{\phi_1}\frac{\widehat{\lambda}^2\widehat{\sigma}^2\widehat{\beta}^2_{(K)}}{\underline{C}_1\log(|\widehat{S}|)^2} + C_{\phi_2}C_2\widehat{\widehat{\gamma}} \tag{134}$$

with probability at least $1-p_w(C_1, C_2, C_3)$.

Statement of Oracle Inequality for consistent model. We next give our Oracle Inequality for the consistent model.

$$\frac{1}{N}\left\{||X\widehat{\delta}||_2^2 + \lambda\widehat{\sigma}||\widehat{W\widehat{\delta}}||_1^1\right\} \leq \frac{C_{L1}\widehat{\sigma}^2\widehat{\lambda}^2|S|}{N^2\phi_0^2}. \tag{135}$$

Using the bound in Inequality 134 gives

$$\frac{1}{N}\left\{||X\widehat{\delta}||_2^2 + \lambda\widehat{\sigma}||\widehat{W\widehat{\delta}}||_1^1\right\} \leq \frac{C_{L1}\widehat{\sigma}^2\widehat{\lambda}^2|S|}{N^2\left\{C_{\phi_1}\frac{\widehat{\lambda}^2\widehat{\sigma}^2\widehat{\beta}^2_{(K)}}{\underline{C}_1\log(|\widehat{S}|)^2} + C_{\phi_2}C_2\widehat{\widehat{\gamma}}\right\}^2} \tag{136}$$

which will hold with probability at least $1-p_\lambda(C_\lambda)-p_\epsilon(C_\epsilon)-p_w(C_1, C_2, C_3)-\exp(-t^2/2)$ so long as Inequality 120 is met.

D.3 Results for LASSOplus

The results for LASSOplus are similar to those from the consistent model. LASSOplus-EM is essentially an adaptive LASSO with endogenously estimated weights plus a threshold to zero out small effects. As in the theoretical analyses in Buhlmann and van de Geer (2013, ch. 7.8) , we find that the adaptive LASSO and thresholded LASSO achieve similar Oracle bounds, but our LASSOplus bound is twice that of the consistent model.

Probabilistic bound on LASSOplus-EM. The basic approach with LASSOplus is to endogenously estimate parameter-specific weights and generate an inflated the variance component so as to threshold small effects. For our consistent model, our probabilistic bound (see Inequality (117)) took the form

$$-\frac{1}{32}\frac{\widehat{\lambda}^2\widehat{w}^2_{(K-|\widehat{S}|-1)}\widehat{\sigma}^2}{(N-1)\sigma^2}+\log(K-|\widehat{S}|)\leq-\frac{t^2}{2} \tag{137}$$

The problem is that the first term is $O_p(\sqrt{K}/\sqrt{N})$, meaning that as the sample size grows, the probabilistic bound will not be met. LASSOplus-EM inflates the variance to get a bound as follows:

$$-\frac{1}{32}\frac{\widehat{\lambda}^2\widehat{w}^2_{(K-|\widehat{S}|-1)}\widehat{\sigma}^2_{sp}}{(N-1)\sigma^2}+\log(K-|\widehat{S}|)\leq-\frac{t^2}{2} \tag{138}$$

which now makes the first term $O_p(\sqrt{K}N^{2\alpha}/\sqrt{N})$. Setting $\alpha=1/4$, our default, makes the first term $O_p(\sqrt{K})$, which will dominate the $\log(K-|\widehat{S}|)$ term for large enough K, regardless of N. Specifically, it will hold whenever

$$\left\{\frac{C_\lambda C_\epsilon C_2\widehat{\widehat{\gamma}}^2 K}{\frac{t^2}{2}+\log(K-|\widehat{S}|)}\right\}^2\geq 32^2 \tag{139}$$

Therefore, LASSOplus-EM will bound the empirical process for sufficiently large K.

Oracle Inequality for LASSOplus-EM. LASSOplus achieves an Oracle bound regardless of N, but at the cost of increasing the bound. Denote as $\widehat{\delta}_p, X_p, \widehat{W}_p$ the elements of $\widehat{\delta}$ and columns of X and $\widehat{W}$ that are associated with non-zero elements of LASSOplus-EM, and $\widehat{\delta}_{p^\complement}$ and $X_{p^\complement}, \widehat{W}_{p^\complement}$ the complements of these terms. We know then

$$\frac{1}{N}\left\{||X\widehat{\delta}||_2^2+\lambda\widehat{\sigma}||\widehat{W}\widehat{\delta}||_1^1\right\} \tag{140}$$

$$=\frac{1}{N}\left\{||X_p\widehat{\delta}_p||_2^2+||X_{p^\complement}\widehat{\delta}_{p^\complement}||_2^2+2\widehat{\delta}^{p\top}X_p^\top X_{p^\complement}\widehat{\delta}_{p^\complement}+\lambda\widehat{\sigma}||\widehat{W}_p\widehat{\delta}_p||_1^1+\lambda\widehat{\sigma}||\widehat{W}_{p^\complement}\widehat{\delta}_{p^\complement}||_1^1\right\}. \tag{141}$$

Combining with the Oracle Inequality for the consistent model gives

$$\frac{1}{N}\left\{||X_p\widehat{\delta}_p||_2^2+\lambda\widehat{\sigma}||\widehat{W}_p\widehat{\delta}_p||_1^1+\lambda\widehat{\sigma}||\widehat{W}_p\widehat{\delta}_p||_1^1\right\}$$

$$\leq\frac{C_{L1}\widehat{\sigma}^2\widehat{\lambda}^2|S|}{N^2\left\{C_{\phi_1}\frac{\widehat{\lambda}^2\widehat{\sigma}^2\widehat{\beta}^2_{(K)}}{\underline{C}_1\log(|\widehat{S}|)^2}+C_{\phi_2}C_2\widehat{\widehat{\gamma}}\right\}^2}-\frac{1}{N}\left\{||X_{p^\complement}\widehat{\delta}_{p^\complement}||_2^2+2\widehat{\delta}_p^\top X_p^\top X_{p^\complement}\widehat{\delta}_{p^\complement}+\lambda\widehat{\sigma}||\widehat{W}_{p^\complement}\widehat{\delta}_{p^\complement}||_1^1\right\} \tag{142}$$

$$\leq\frac{C_{L1}\widehat{\sigma}^2\widehat{\lambda}^2|S|}{N^2\left\{C_{\phi_1}\frac{\widehat{\lambda}^2\widehat{\sigma}^2\widehat{\beta}^2_{(K)}}{\underline{C}_1\log(|\widehat{S}|)^2}+C_{\phi_2}C_2\widehat{\widehat{\gamma}}\right\}^2}+\frac{1}{N}\left\{||X_p\widehat{\delta}_p||_2^2-\lambda\widehat{\sigma}||\widehat{W}_{p^\complement}\widehat{\delta}_{p^\complement}||_1^1\right\} \tag{143}$$

$$\leq\frac{2C_{L1}\widehat{\sigma}^2\widehat{\lambda}^2|S|}{N^2\left\{C_{\phi_1}\frac{\widehat{\lambda}^2\widehat{\sigma}^2\widehat{\beta}^2_{(K)}}{\underline{C}_1\log(|\widehat{S}|)^2}+C_{\phi_2}C_2\widehat{\widehat{\gamma}}\right\}^2} \tag{144}$$

where the first line is just rearranging, the second uses the inequality $-2ab \leq a^2 + b^2$, and the third come from re-applying the Oracle bound from the consistent model to the second term in the second line.

We see that LASSOplus-EM satisfies an Oracle Bound without the constraints on N and K required by the consistent model, but this comes at the cost of a looser bound.

E Variance Estimation

We sample from the approximate sampling distribution of the the LASSOplus estimator at each Gibbs update:

$$\begin{aligned}
&\beta_k \mathbf{1}\left(|\widehat{\beta}_k^{sp}| \geq \frac{\lambda w_k \sigma_{sp}}{N-1}\right) && (145)\\
&\approx \beta_k \Phi\left\{\left||\widehat{\beta}_k^{sp}/\widehat{\sigma}_{ls}| - \frac{\lambda w_k \sigma_{sp}}{\widehat{\sigma}_{ls} \times (N-1)}\right|\right\} && (146)\\
&= \beta_k \Phi\left\{\sqrt{N-1}\left||\widehat{\beta}_k^{sp}/\sigma| - \frac{\lambda w_k \sigma_{sp}}{\sigma \times (N-1)}\right|\right\} && (147)\\
&= g\left(\beta_k, \widehat{\beta}_k^{sp}, \sigma, \sigma_{sp}, \lambda, w_k\right) && (148)
\end{aligned}$$

where $\Phi(a)$ is the cumulative distribution for a standard normal random variable and we approximate the standard error of the least squares coefficient as $\widehat{\sigma}_{ls} \approx \sigma/\sqrt{N-1}$. Define

$$\begin{aligned}
z_k &= \sqrt{N-1}\left||\widehat{\beta}_k^{sp}/\sigma| - \frac{\lambda w_k \sigma_{sp}}{\sigma \times (N-1)}\right| && (149)\\
p_k &= \Phi\left\{z_k\right\} && (150)
\end{aligned}$$

Define the vector of partial derivatives

$$\nabla g\left(\beta_k, \widehat{\beta}_k^{sp}, \sigma, \sigma_{sp}, \lambda, w_k\right) = \left[\frac{\partial g(\cdot)}{\partial \beta_k}, \frac{\partial g(\cdot)}{\partial \widehat{\beta}_k^{sp}}, \frac{\partial g(\cdot)}{\partial \sigma}, \frac{\partial g(\cdot)}{\partial \sigma_{sp}}, \frac{\partial g(\cdot)}{\partial \lambda}, \frac{\partial g(\cdot)}{\partial w_k}\right]^\top \tag{151}$$

$$= \begin{bmatrix} p_k \\ \beta_k \times \phi(z_k) \times \sqrt{N-1}/\sigma \operatorname{sgn}(\widehat{\beta}_k^{sp}) \\ \beta_k \times \phi(z_k) \times \sqrt{N-1}\left(-\frac{\left|\widehat{\beta}_k^{sp}\right|}{\sigma^2} + \frac{\lambda w_k \sigma_{sp}}{\sigma^2 \times (N-1)}\right) \\ \beta_k \times \phi(z_k) \times \sqrt{N-1} \times \frac{\lambda w_k}{\sigma \times (N-1)} \\ \beta_k \times \phi(z_k) \times \sqrt{N-1} \times \frac{w_k \sigma_{sp}}{\sigma \times (N-1)} \\ \beta_k \times \phi(z_k) \times \sqrt{N-1} \times \frac{\lambda \sigma_{sp}}{\sigma \times (N-1)} \end{bmatrix} \tag{152}$$

and the 6×6 matrix

$$V = \operatorname{diag}\left[\operatorname{Var}(\beta_k), \operatorname{Var}(\widehat{\beta}_k^{sp}), \operatorname{Var}(\sigma), \operatorname{Var}(\sigma_{sp}), \operatorname{Var}(\lambda), \operatorname{Var}(w_k)\right] \tag{153}$$

where we are assuming zero covariance between elements. All elements of V are calculated analytically from the variance of the conditional pseudoposterior densities except for $\mathrm{Var}(w_k)$ which is calculated from the approximate density used in the griddy Gibbs sampler. Our approximate variance is then

$$\widehat{\sigma}_j^2 = \nabla g^\top(\cdot) V \nabla g(\cdot) \tag{154}$$

.

F EM Updates for LASSOplus-EM

For our EM implementation, we treat in $\beta^{plus-EM}$ and σ^2 as parameters and the remaining parameters as "missing," i.e. to be estimated. As we have already calculated the conditional posterior densities for all parameters, the EM updates is straightforward.

Standardize Y and all columns of X to be mean-zero, sample variance one. Initialize $\forall k : \widehat{\beta}_k \leftarrow u_k$ with $u_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 0.01)$; $\widehat{\sigma}^2 \leftarrow ||Y - X\widehat{\beta}||_2^2/N$; $\widehat{\lambda} \leftarrow 1$; $\widehat{w}_k \leftarrow 1$.

At each given step, the most current updates from the previous steps are used. To convergence,

- E-steps

  1. $\forall k :\ \widehat{(1/\tau_k^2)} \leftarrow \widehat{\lambda}\widehat{w}_k\widehat{\sigma}/|\widehat{\beta}_k|$; $\widehat{\tau_k^2} \leftarrow |\widehat{\beta}_k|/(\widehat{\lambda}\widehat{w}_k\widehat{\sigma}) + 1/(\widehat{\lambda}^2\widehat{w}_k^2)$
  2. $\widehat{\lambda} \leftarrow \frac{\widetilde{\Gamma}(\sqrt{N}K+1/2)/\widetilde{\Gamma}(\sqrt{N}K)}{\sqrt{\frac{1}{2}\sum_{k=1}^K \widehat{\tau}_k^2+\rho}}$; $\widehat{\lambda^2} \leftarrow \frac{\sqrt{N}K}{\frac{1}{2}\sum_{k=1}^K \widehat{\tau}_k^2+\rho}$ with $\widetilde{\Gamma}()$ the gamma function.
  3. $\forall k$: update $\widehat{w}_k$ via numerical integration using kernel $\Pr(w_k|\cdot) \propto e^{-w^{\widehat{\gamma}}-\widehat{\lambda}w\sqrt{\widehat{\frac{1}{\sigma^2}}}|\widehat{\beta}_k|}$
  4. Update $\widehat{\gamma}$ via numerical integration using kernel $\Pr(\gamma|\cdot) \propto \gamma e^{-\sum_{k=1}^K \widehat{w}_k^\gamma - \gamma}$

- M-Steps

- $\widehat{\sigma^2} \leftarrow \frac{\sum_{i=1}^N (Y_i - X_i^\top \widehat{\beta})^2 + \sum_{k=1}^K (\widehat{\beta}_k)^2 \times \widehat{\frac{1}{\tau_k^2}}}{N+K}$; $\widehat{\frac{1}{\sigma^2}} \leftarrow \frac{N+K-2}{\sum_{i=1}^N (Y_i - X_i^\top \widehat{\beta})^2 + \sum_{k=1}^K (\widehat{\beta}_k)^2 \times \widehat{\frac{1}{\tau_k^2}}}$

- Conditional M-steps: $\forall k : \widehat{\beta}^k \leftarrow \frac{\sum_{i=1}^N X_{ik}(Y_i - \sum_{k' \neq k} X_{ik'}\widehat{\beta}_{k'})}{(N-1)+\widehat{\frac{1}{\tau_k^2}}}$ where it is understood that at update $\widetilde{k}$, updated estimates of $\left\{\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_{\widetilde{k}-1}\right\}$ are used.

LASSOplus updates:

- $\widehat{\sigma}_{sp}^2 \leftarrow\leftarrow \frac{\sum_{i=1}^N (Y_i - X_i^\top \widehat{\beta})^2 + \sum_{k=1}^K (\widehat{\beta}_k)^2 \times \widehat{\frac{1}{\tau_k^2}}}{\sqrt{N}+K}$
- $\widehat{\beta}_k^{plus} \leftarrow \widehat{\beta}_k \mathbf{1}\left(\big|\sum_{i=1}^N X_{ik}(Y_i - \sum_{k' \neq k} X_{ik'}\widehat{\beta}_{k'})\big| > \widehat{\lambda}\widehat{w}_k\sqrt{\widehat{\sigma}_{sp}^2}\right)$

G Independence between Adjusted Higher-Order Terms and Lower-Order Terms

We prove first that, under the residualized construction, the least squares coefficient on the a higher-order interaction term is uncorrelated with the coefficients on lower-order terms. By this means, the effect of the higher-order term does not vary with its lower-order components, and hence can be interpreted on its own. We then extend the result to the conditional pseudoposterior density of the estimates.

Denote the $N \times 1$ vector of outcomes Y, $N \times L$ matrix of lower-order terms $\mathbf{X}_{lower}$ and vector of mean-zero, equivariant errors ϵ. Define as $X_{inter} = [X_{inter}]_i = \prod_{1 \leq l' \leq L} x_{il'}$, the elementwise product of the lower-order terms. Assume $[\mathbf{X}_{lower}|X_{inter}]$ is full rank. Using parameters $\{\beta_0, \vec{\beta}_l, \beta_{inter}\}$, define the model, with $\vec{\beta}_l$ an $L \times 1$ vector and the others scalars, as

$$Y = \mathbf{X}_{lower}\vec{\beta}_l + X_{inter}\beta_{inter} + \epsilon. \tag{155}$$

Define the matrices

$$\mathbf{X} = [\mathbf{X}_{lower}|X_{inter}] \tag{156}$$

$$\mathbf{M}_{lower} = \mathbf{I}_L - (\mathbf{X}_{lower})(\mathbf{X}_{lower}^\top \mathbf{X}_{lower})^{-1}\mathbf{X}_{lower}^\top \tag{157}$$

$$\widetilde{X}_{inter} = \mathbf{M}_{lower}X_{inter} \tag{158}$$

$$\mathbf{X}_{adjust} = [\mathbf{X}_{lower}|\widetilde{X}_{inter}] \tag{159}$$

The vector $\widetilde{X}_{inter}$ is the residualized interaction term described in the text, giving parameterization

$$Y = \mathbf{X}_{lower}\widetilde{\vec{\beta}}_l + \widetilde{X}_{inter}\widetilde{\beta}_{inter} + \epsilon \tag{160}$$

where the error vector ϵ, stays unchanged since the two parameterizations differ only by a linear transformation of the covariates.

The covariance of the least squares estimates in the first parameterization is proportional to the inverse of the cross product of the design matrix. Using the block-partition matrix formula gives

$$(\mathbf{X}^\top\mathbf{X})^{-1} = \begin{bmatrix} (\mathbf{X}_{lower}^\top\mathbf{X}_{lower}) & \mathbf{X}_{lower}^\top X_{inter} \\ X_{inter}^\top\mathbf{X}_{lower} & X_{inter}^\top X_{inter} \end{bmatrix}^{-1} \tag{161}$$

$$= \begin{bmatrix} \left(\mathbf{X}_{lower}^\top\mathbf{X}_{lower} - \frac{1}{c_0}\mathbf{X}_{lower}^\top X_{inter}X_{inter}^\top\mathbf{X}_{lower}\right)^{-1} & -\frac{1}{c_0}(\mathbf{X}_{lower}^\top\mathbf{X}_{lower})^{-1}X_{lower}^\top X_{inter} \\ -\frac{1}{c_0}X_{inter}^\top\mathbf{X}_{lower}(\mathbf{X}_{lower}^\top\mathbf{X}_{lower})^{-1} & \frac{1}{c_0} \end{bmatrix} \tag{162}$$

with the constant $c_0 = X_{inter}^\top X_{inter} - X_{inter}^\top \mathbf{X}_{lower}(\mathbf{X}_{lower}^\top \mathbf{X}_{lower})^{-1}\mathbf{X}_{lower}^\top X_{inter}$. This implies

$$\text{Cov}(\widehat{\beta}_{inter}, \widehat{\beta}_k) \propto -\left[\frac{1}{c_0}(\mathbf{X}_{lower}^\top \mathbf{X}_{lower})^{-1}\mathbf{X}_{lower}^\top X_{inter}\right]_j \text{ for } j \in \{1, 2, \ldots, L\} \tag{163}$$

In general, this covariance will not be zero, suggesting that under the normal parameterization the effect of the interaction term varies with movements in its lower order terms. Repeating the same exercise with a model parameterized in terms of $\widetilde{X}_{inter}$ gives

$$\text{Cov}\left(\widehat{\widetilde{\beta}}_{inter}, \widehat{\widetilde{\beta}}_k\right) \propto -\left[\frac{1}{c_0}(\mathbf{X}_{lower}^\top \mathbf{X}_{lower})^{-1}\mathbf{X}_{lower}^\top \widetilde{X}_{inter}\right]_j \tag{164}$$

$$= -\left[\frac{1}{c_0}(\mathbf{X}_{lower}^\top \mathbf{X}_{lower})^{-1}\mathbf{X}_{lower}^\top \mathbf{M}_{lower} X_{inter}\right]_j \tag{165}$$

$$= 0 \text{ for } j \in \{1, 2, \ldots, L\} \tag{166}$$

Therefore, under the parameterization with residualized interaction terms, the marginal effect of each interaction term is uncorrelated with that of its lower order terms. To extend to the multivariate case, assume the full design matrix of all effects is full-rank, and all other effects have been partialed out. The case of $K > N$ requires an assumption similar to the restricted eigenvalue assumption (Bickel, Ritov and Tsybakov, 2009), that all submatrices of size $L + 2$ are full rank and all components of the submatrices not in $\mathbf{X}$ are linearly independent of $\mathbf{X}$. Partialing out with respect to the other covariates in either case leaves the results unchanged.

Next, we show the result holds for the conditional pseudoposterior density under a conditional independent normal prior, as with the augmented LASSOplus. Assume $[\vec{\beta}_l^\top, \beta_{inter}]^\top \sim \mathcal{N}(0_{L+1}, D)$ with D an $(L + 1) \times (L + 1)$ diagonal matrix with positive entries along the diagonal. In this case, the conditional posterior of $[\vec{\beta}_l^\top, \beta_{inter}]^\top$ under a normal likelihood takes the form

$$\Pr([\vec{\beta}_l^\top, \beta_{inter}]^\top | \cdot) \sim \mathcal{N}(A^{-1}\mathbf{X}^\top Y, \sigma^2 A^{-1}) \tag{167}$$

with $A = \mathbf{X}^\top \mathbf{X} + D$. Carrying through the same derivation as above gives the posterior covariance between β_{L+1}, the parameter on the interaction term, and β_k, $1 \leq j \leq L$, as

$$A_{j,L+1}^{-1} \propto -\left[\frac{1}{c_0'}(\mathbf{X}_{lower}^\top \mathbf{X}_{lower} + D_{1:L,1:L})^{-1}\mathbf{X}_{lower}^\top X_{inter}\right]_j \text{ for } j \in \{1, 2, \ldots, L\} \tag{168}$$

which will not be 0, in general. In this case, the constant $c_0' = (X_{inter} + D_{L+1,L+1})^\top (X_{inter} + D_{L+1,L+1}) - X_{inter}^\top \mathbf{X}_{lower}(\mathbf{X}_{lower}^\top \mathbf{X}_{lower} + D_{1:L,1:L})^{-1}\mathbf{X}_{lower}^\top X_{inter}$.

Considering the residualized interaction term instead of the standard term gives

$$A_{j,L+1}^{-1} \propto -\left[\frac{1}{c_0'}(\mathbf{X}_{lower}^\top \mathbf{X}_{lower} + D_{1:L,1:L})^{-1}\mathbf{X}_{lower}^\top \widetilde{X}_{inter}\right]_j = 0 \text{ for } j \in \{1, 2, \ldots, L\} \tag{169}$$

H Alternative Estimators

For the LASSO and adaptive LASSO, we found the BIC statistic of Wang and Leng (2007) performed poorly when $K > N$, sometimes including dozens of false positives. We instead use a standard BIC statistic where we take the degrees of freedom as the number of non-zero coefficients (Zou, Hastie and Tibshirani, 2007).

In terms of uncertainty estimates, we implement the approximate confidence intervals for the LASSOplus. We use the posterior intervals for the horseshoe model. For the frequentist LASSO and adaptive LASSO, we implement the perturbation method of Minnier, Tian and Cai (2011). For $p \in \{1, 2, \ldots, P\}$ for some large P, the method requires fitting

$$\widehat{\beta}^{alasso,p}(\lambda|w_{\cdot}, g_{\cdot}) = \underset{\beta}{\text{argmin}} \sum_{i=1}^{N} g_i^p (Y_i - X_i^\top \beta)^2 + \lambda \sum_{k=1}^{K} w_k |\beta_k|; \tag{170}$$

$$w_k = 1/|\beta_k^0| \tag{171}$$

where the weights are $g_i^p \overset{\text{i.i.d.}}{\sim} \exp(1)$. For the LASSO, we simply take $w_k = 1$ for all k. Minnier, Tian and Cai (2011) prove that the set $\{\widehat{\beta}^{alasso,p}(\lambda|w_{\cdot}, g_{\cdot})\}_{p=1}^{P}$ will achieve nominal coverage asymptotically, though the result does not hold for the LASSO. We fit the perturbation method to both for comparison. We found the perturbation method performs better than the parametric bootstrap suggested by Chatterjee and Lahiri (2011, 2013), so we do not present the results.

We next move on to the LASSO+OLS method of Belloni and Chernozhukov (2013), hereafter BC.The empirical process approach selects the tuning parameter in order to bound $2 \max \left(\epsilon^\top X_{\cdot k}\right)$ with some high probability. BC note that, up to a scale parameter σ, the tuning parameter value can be simulated quite easily, and they define $\Lambda(1 - \alpha_{sig}|X)$ as the $1 - \alpha_{sig}$ quantile of $2 \max \left(\epsilon^\top X_{\cdot k}/\sigma_b\right)$ for $\mathbb{E}(\epsilon_i|X_i) = 0$; $\text{Var}(\epsilon_i|X_i) = \sigma_b^2$ as approximated through a simulation.

Second stage variable selection. Tuning λ in order to satisfy the Oracle Inequality will generally over-select effects. The reason is that the LASSO induces bias in the coefficient estimaates, and that bias leaves a gap for irrelevant effects that are correlated with the relevant effects to be drawn in and selected. Several methods in the empirical process framework have used the Oracle Inequality-tuned LASSO to over-select covariates and then, in a second stage, select a subset of these.

One way to do so is simply thresholding the LASSO estimates, so

$$\widehat{\beta}^{thresh} = \widehat{\beta}^{L} \odot \mathbf{1}\left(\left|\widehat{\beta}^{L}\right| > \tau\right) \tag{172}$$

where the inequality and multiplication $\odot$ are taken elementwise. A second option is to take then rerun OLS on variables that survive the threshold. Define X_{thresh} as the submatrix of X corresponding with elements of $\widehat{\beta}^{thresh}(\tau)$ that are not zero. Then,

$$\widehat{\beta}^{thresh+OLS}(\tau) = (X_{thresh}^{\top} X_{thresh})^{-1} X_{thresh}^{\top} Y. \tag{173}$$

In the case X_{thresh} is rank-deficient, either ridge regression or partial least squares can be used (Liu and Yu, 2013). The post LASSO OLS estimator is then $\widehat{\beta}^{thresh+OLS}(0)$, which is simply OLS used on all selected LASSO covariates.

Belloni and Chernozhukov (2013) propose a different means of selecting a subset of relevant effects and eliminating the first-stage LASSO bias. Denote $Q(\theta) = ||Y - X\theta||_2^2$. The select τ such that

$$t_\gamma = \max_{t \geq 0} Q\left(\widehat{\beta}^{thresh+OLS}(\tau)\right) - Q\left(\widehat{\beta}^{L}\right) \leq \gamma \tag{174}$$

for $\gamma \leq 0$. Taking $\gamma = 0$ returns the sparsest OLS-reflated model that generates a lower residual sum of squares than the LASSO estimator. We follow the suggestion of Belloni and Chernozhukov (2013, expr 2.14) and take $\gamma = \left\{Q\left(\widehat{\beta}^{thresh+OLS}(0)\right) - Q\left(\widehat{\beta}^{L}\right)\right\}/2$ in the simulations.