Adam Smith Business School
University of Glasgow

INFERENCE ON TREATMENT EFFECTS WITH HIGH-DIMENSIONAL CONTROLS: FREQUENTIST AND BAYESIAN APPROACHES

Duong Trinh - 2494479T

ECON5088P MRES DISSERTATION

Supervisors: Prof. Dimitris Korobilis
Dr. Kenichi Shimizu

Word Count: 14141

Year of submission: 2020-2021

To my family,

CONTENTS

CHAPTER

ONE

INTRODUCTION

Causal inference has long been a fundamental topic in economics, especially when it plays an essential role as the basis for policy or decision making. The gold standard to understand the causal effect of a treatment is a randomized controlled experiment in which the treatment is randomly assigned. In many cases, however, randomized controlled experiments are complicated or impossible to implement due to financial, political, or ethical reasons. This challenge has led to the adoption of a variety of quasi-experimental approaches based on observational data to estimate treatment effects. Nevertheless, inferring causal relationships from observational data remains difficult because treatments are observed rather than assigned randomly, and the Unconfoundedness assumption is often required. This assumption imposes that the treatment is as good as randomly assigned once we condition on observables.

A problem empirical researchers face when relying on the conditional-on-observables identification strategy is to determine which control variables should be included. Even when one may get some intuitions from economic theories or prior knowledge, the corresponding function forms of relevant variables are still uncertain. This lack of clear guidance leaves researchers with a vast set of potential control variables, including raw factors available in the data themselves, as well as many of their possible transformations as regressors. More often than not, studies rely on ad hoc sensitivity analyses in which a researcher reports results for several different sets of controls to convey that their substantive result for the treatment effect is insensitive to changes in the set of control variables [Angrist and Pischke, 2008].

Meanwhile, a growing number of innovative statistical methods (known as machine learning) are available for constructing prediction models in the presence of high-dimensional data

[Hastie et al., 2009]. Although these regularization-based methods tend to perform well at prediction, which they are designed for, they may often suffer a poor assessment of uncertainty around the estimates of regression coefficients [Leeb and Pötscher, 2008]. In other words, they hardly produce valid confidence intervals for parameters of interest.

A novel approach that successfully adopts and modifies machine learning methods to control confounding variables properly and provide high-quality inference on treatment effects is the Post-Double-Selection Lasso (PDS Lasso) method, proposed by Belloni et al. [2014b]. They consider a partially linear model with high-dimensional potential controls. Rather than using the usual Post-Single-Selection procedures that rely on a single selection step with Lasso, they use two different variable selection steps with Lasso followed by a final estimation step with ordinary least squares regression. Their theoretical results demonstrate the uniformly valid inference over a rich class of data-generating processes as the merit of this method.

Another approach that alleviates the issues discussed above is using Bayesian techniques to obtain valid inference from posterior samples. This is exemplified in the work undertaken by Antonelli et al. [2019], the High-dimensional Confounding Adjustment (HDCA) method. In a nutshell, the authors introduce a general formulation of the spike-and-slab Lasso prior to allow the prior probability that the regression coefficient of a given potential control is included in the slab component to depend on the association between this covariate and the treatment. Remarkably, results from their well-designed simulation study indicate a comparable bias yet a better interval coverage rate achieved by this Bayesian method compared to Post-Double-Selection Lasso. One explanation for the superiority of HDCA lies in the way this method could drastically reduce the shrinkage of important confounders, whereas still shrinking to zero the coefficients of instrumental and noise variables in the structural equation. Furthermore, this effect seems to be attributed to the spike-and-slab Lasso prior, a notable member of a broader class of the *Bayesian shrinkage priors*. In fact, the major goal of shrinkage priors is to shrink small coefficients to zero while maintaining true large coefficients, especially in high-dimensional settings. The possible variation in shrinkage amounts among those priors depends on their specific shapes/designs/creations. In particular, the sharper the peak is around zero, the stronger shrinkage for small coefficients. Also, the heavier the tail, the lighter the shrinkage for large coefficients. Naturally, one may think of extending the idea to borrow information from the treatment equation to guide the amount of shrinkage in the structural equation to other high-dimensional shrinkage priors beyond the spike-and-slab Lasso.

Indeed, there are numerous reasons why this extension is potential and worthwhile.

Firstly, it has been known that each Bayesian shrinkage prior could be presented by a function similar to a penalty term in Frequentist regularized regression; thus, penalization could be incorporated naturally within a Bayesian framework through the choice of priors. As a result, this extension could offer a more flexible way to apply regularization-based methods to tackle causal inference problems. Secondly, Bayesian regularization-based methods offer additional advantages such as readily available uncertainty estimates and automatic estimation of the penalty parameter compared to Frequentist regularization-based methods. These merits are also observed in the cases of HDCA and PDS Lasso. In addition, for the same ultimate goal of inference on treatment effects, PDS Lasso allows a moderate variable selection mistakes[1], while HDCA makes use of a more adaptive of variable selection techniques. Therefore, it is attractive to examine if the good quality of the HDCA framework is robust for various choices of shrinkage priors, and if we could enhance its performance to be more effective - better variable selection for better causal inference. At this point, the extension could play a role as a straightforward implementation. Last but not least, little is known about the finite-sample behaviour of different Bayesian shrinkage priors in a specialized setting, such as causal inference. Broadly speaking, there is a lack of a large-scale comparison of Bayesian shrinkage priors, except for the recent surveys of Van Erp et al. [2019] and Polson and Sokolov [2019]. These early attempts, however, devote mostly to evaluating performances in terms of variable selection and prediction. Therefore, the extension could serve as a missing piece of the whole picture.

Taken together, this thesis aims to investigate the application of Frequentist and Bayesian regularization-based methods to inference on treatment effects with high-dimensional potential controls. The contribution is twofold:

Firstly, we generalize the High-dimensional Confounding Adjustment approach developed by Antonelli et al. [2019] to a Bayesian framework for inference on treatment effects. Specifically, generic Stochastic Search Variable Selection (SSVS) priors [George and McCulloch, 1993] and generic Spike and Slab [Kuo and Mallick, 1998] are adopted. The generic framework then allows us to incorporate various Bayesian shrinkage priors include Normal prior, Student-t prior, Laplace prior and Horseshoe prior, hence create eight Bayesian regularization-based methods for causal inference in total. Combined with Post-Double-Selection Lasso [Belloni et al., 2014b], regarded as Frequentist regularization-based method, we evaluate the finite-sample performance of all methods within a dedicated Monte-Carlo study that covers a large class of scenarios to ensure both overall and in-depth analyses. An empirical illustration is also taken into consideration. To the best of our knowledge, there is

[1] which do not affect the asymptotic properties of the estimators

no large-scale evaluation like that before, especially in terms of applying different Bayesian techniques to estimation and inference about treatments effects in high-dimensional settings. This extension, therefore, enriches current understandings of performances of regularization-based methods for causal inference as well as properties of Bayesian shrinkage priors in non-conventional designs, as explained in the previous part.

Secondly, these modern methods discussed in this study could be considered as a data-driven complement to traditional econometric methods, developed to address causality issues from observational data in the context of high-dimensional settings. These developments would help enhance the credibility of empirical economic analysis. Through examining the effect of media on voting outcomes following Enikolopov et al. [2011], our empirical example shows how this approach could be applied to support the causal conclusion in linear regression models.

The general outline of this thesis is as follows: Chapter 2 is a brief literature review that provides three pivots for our analysis throughout this study. In the first part, we formulate a causal framework for inference on treatment effects under the Unconfoundedness assumption. This allows us to specify the puzzle of regression with high-dimensional possible controls, which motivates a data-driven and systematic variable selection approach. The second part of this chapter is a short introduction to regularized regression - a potential solution to cover the gap. A duality between Frequentist and Bayesian approaches is described as well. Examining a range of strategies from the naive to the state-of-the-art, the third part of Chapter 2 illustrates how regularization has been adopted by both Frequentist and Bayesian paradigms when the goal is causal inference. Built upon these foundations, in Chapter 3, we design a set of Frequentist and Bayesian regularization-based methods for inference on treatment effects in high-dimensional settings. Next, the performances of these methods are evaluated within a Monte-Carlo study in section 4 and an empirical illustration in section 5. Finally, Chapter 6 summarizes the main findings of this thesis and discusses several implications for current and future research.

CHAPTER

TWO

LITERATURE REVIEW

2.1 Inference on Treatment Effects under Unconfoundedness

2.1.1 Problem Statement, Notation and Identifying Assumptions

Throughout the manuscript, we consider an observational study which yields i.i.d sample $\mathcal{P}_i = (y_i, T_i, \boldsymbol{X_i})$ for $i = 1, \ldots, n$; where y_i is the outcome, T_i is the scalar treatment variable (we do not need to impose restriction on T_i as a binary or continuous variable), and $\boldsymbol{X_i}$ is a p-dimensional row vector of potential control variables for subject i[1]. We focus on *high-dimensional setting* where p is close to or even larger than the number of observations n.

In a canonical econometric analysis, we are mainly interested in several causal estimands[2] relevant to the effect of treatment T on outcome y, i.e. *Causal treatment effects*. With this goal in mind, we start utilizing the Neyman-Rubin potential outcome framework. Denote $y_i(t)$ the potential outcome the subject i could receive under the treatment level $T_i = t$. Hence,

The (hypothetical) *individual treatment effect* (ITE) of T_i on y_i is defined as

$$\zeta_i = \nabla_t y_i(t) = \begin{cases} y_i(1) - y_i(0), & \text{if } T_i \text{ is binary} \\ \frac{\partial y_i(t)}{\partial t}, & \text{if } T_i \text{ is continuous} \end{cases}$$

[1] potential to select among, i.e. we exclude all bad controls (post-treatment, etc.) which we are sure about. For recent discussion, Cinelli et al. [2020] is a good reference.

[2] the parameters of interest that summarize the causal effect of treatment variable, later denoted by Δ.

where ζ_i measures the impact of a change in treatment variable T_i while holding other factors fixed. However, this quantity is never identified since only one potential outcome for each individual is directly observed. This is widely acknowledged as *the fundamental problem of causal inference*.

Hence, we focus our attention on the *average treatment effect* (ATE), which is defined as

$$\Delta = \mathbb{E}\left[\zeta_i\right] = \mathbb{E}\left[\nabla_t y_i(t)\right] = \begin{cases} \mathbb{E}\left[y_i(1)\right] - \mathbb{E}\left[y_i(0)\right], & \text{if } T_i \text{ is binary} \\ \mathbb{E}\left[\frac{\partial y_i(t)}{\partial t}\right], & \text{if } T_i \text{ is continuous} \end{cases}$$

Another causal estimand, the *conditional average treatment effects* (CATE), is also considered

$$\begin{aligned} \mathbb{E}\left[\zeta_i \mid \boldsymbol{X}_i = \boldsymbol{x}\right] &= \mathbb{E}\left[\nabla_t y_i(t) \mid \boldsymbol{X}_i = \boldsymbol{x}\right] \\ &= \begin{cases} \mathbb{E}\left[y_i(1) \mid \boldsymbol{X}_i = \boldsymbol{x}\right] - \mathbb{E}\left[y_i(0) \mid \boldsymbol{X}_i = \boldsymbol{x}\right], & \text{if } T_i \text{ is binary} \\ \mathbb{E}\left[\frac{\partial y_i(t)}{\partial t} \mid \boldsymbol{X}_i = \boldsymbol{x}\right], & \text{if } T_i \text{ is continuous} \end{cases} \end{aligned}$$

ATE and CATE are only identified under plausible assumptions under which these quantities can be expressed in terms of observed data. In particular, the following assumptions are most commonly imposed and must hold for any t and x for the sake of identification:

- Assumption 1. *Stable Unit Treatment Value Assumption* (SUTVA)

  $$T_i = t \text{ implies } y_i^{obs} = y_i(t) \quad \forall t \in supp(T_i)$$

  This assumption ensures that for each subject j, the same treatment level cannot lead to different observed outcome [Little and Rubin, 2000]. We observe the potential outcome correspond to the realized treatment level. To avoid clutter, denote y_i^{obs} simply by y_i from now on.

- Assumption 2. *Overlap*

  $$0 < \mathbb{P}(T_i = t \mid \boldsymbol{X}_i = \boldsymbol{x}) < 1 \quad \forall t \in supp(T_i), \boldsymbol{x} \in supp(\boldsymbol{X}_i)$$

  This assumption states that all subjects has a positive probability of receiving any treatment level. This is a necessary condition to estimate treatment effects everywhere in covariate space.

- Assumption 3. *Unconfoundedness*

  $$y_i(t) \perp T_i \mid \boldsymbol{X}_i = \boldsymbol{x}$$

The assumption stipulates that the treatment is independent of the potential outcome (as good as randomly assigned), condition on observables $\boldsymbol{X}$. This is the most widely cited assumption in empirical studies and it comes in various names with the same intuition: *No unmeasured confounders* implies there is no unmeasured confounders and that the set of measured variables $\boldsymbol{X}$ contains all common causes of treatment and outcome. *Conditional Independence*, *Selection on Observables*, *Conditional on Observables* are more popular in Econometrics [Barnow et al., 1981]. That means once we condition on observables $\boldsymbol{X}$, the treatment assignment is independent of how each subject would respond to the treatment (potential outcomes). In other words, the rule that decides the level subject i is treated is determined completely by their observable characteristics. *Ignorability* states that the treatment variable may be ignorable/ taken as exogenous once we control for enough observed factors $\boldsymbol{X}$. More details are discussed in Angrist and Pischke [2008], Wooldridge [2010], Morgan and Winship [2015] and Imbens and Rubin [2015]. Despite this jargon, for many purposes, it suffices to assume conditional mean independence:

$$\mathbb{E}\left[y_i(t) \mid T_i, \boldsymbol{X}_i\right] = \mathbb{E}\left[y_i(t) \mid \boldsymbol{X}_i\right]$$

2.1.2 Estimations of Causal Treatment Effects

By definition, there is a clear link between ATE and CATE:

$$\mathbb{E}\left[\zeta_i\right] = \mathbb{E}\{\mathbb{E}\left[\zeta_i \mid \boldsymbol{X}_i = \boldsymbol{x}\right]\}$$

When three assumptions above are satisfied, CATE can be transformed as follows:

$$\begin{aligned}
\mathbb{E}\left[\zeta_i \mid \boldsymbol{X}_i = \boldsymbol{x}\right] &= \mathbb{E}\left[\nabla_t y_i(t) \mid \boldsymbol{X}_i = \boldsymbol{x}\right] \\
&= \mathbb{E}\left[\nabla_t y_i(t) \mid T_i = t, \boldsymbol{X}_i = \boldsymbol{x}\right] \text{ (due to Unconfoundedness)} \\
&= \mathbb{E}\left[\nabla_t y_i \mid T_i = t, \boldsymbol{X}_i = \boldsymbol{x}\right] \quad \text{(due to SUTVA)}
\end{aligned}$$

Now, we proceed by assuming a linear regression model for $\mathbb{E}\left[y_i \mid T_i, \boldsymbol{X}_i\right]$ (Assumption 4.)

$$y_i \mid T_i, \boldsymbol{X_i}, \boldsymbol{\gamma}, \sigma^2 \sim \text{ Normal } \left(\mu(T_i, \boldsymbol{X_i}, \boldsymbol{\gamma}), \sigma^2\right), \quad \forall i = 1, \ldots, n \tag{2.1}$$

where the *conditional expected function* (CEF) is

$$\mu(T_i, \boldsymbol{X_i}; \boldsymbol{\gamma}) = \beta_0 + \alpha T_i + \boldsymbol{X}_i \boldsymbol{\beta}, \quad \boldsymbol{\gamma} = (\beta_0, \alpha, \boldsymbol{\beta}^T)^T$$

thus this extra assumption entails CATE is homogeneous treatment effect

$$\mathbb{E}\left[\zeta_i \mid \boldsymbol{X}_i = \boldsymbol{x}\right] = \nabla_t \mathbb{E}\left[y_i \mid T_i = t, \boldsymbol{X}_i = \boldsymbol{x}\right] = \nabla_t \mu(t, x; \boldsymbol{\gamma}) = \alpha$$

hence the ATE is also straightforward: $\Delta = \mathbb{E}[\zeta_i] = \mathbb{E}[\zeta_i \mid \boldsymbol{X}_i = \boldsymbol{x}] = \alpha$.

Now, we rewrite (2.1) to a more popular form in econometrics [3]:

$$y_i = \beta_0 + \alpha T_i + \boldsymbol{X}_i\boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \mid T_i, \boldsymbol{X_i} \sim \text{ Normal } (0, \sigma^2) \quad \forall i = 1, \ldots, n \tag{2.2}$$

At this point, an natural estimator for the ATE (Δ) we may think of is $\hat{\alpha}$, where $\hat{\alpha}$ is a good[4] estimator of α, obtained from model (2.2).

- Assumption 5. *Approximate sparsity*
  β is non-zero for only a subset $\mathbf{S}$ of the set of potential controls $\mathbf{X}$, where $\mid \mathbf{S} \mid = s < n$. This assumption implies that, the number of potential controls $\mathbf{X}$ (p) may exceed the sample size, but conditional expectation function (CEF) is well-approximated by $s < n$ elements.

So far, if we can identify exactly $\mathbf{S}$ as set of specific controls needed, then the OLS estimator $\hat{\alpha}_{OLS}$ of model: $y_i = \beta_0 + \alpha T_i + \boldsymbol{S}_i\boldsymbol{\beta} + \epsilon_i$ would be an unbiased and consistent estimator for the ATE. Intuitively, we can make a causal interpretation of $\hat{\alpha}$: "the amount Y *would change* if T *were changed* by one unit". Otherwise, a lack of proper control in $\mathbf{S}$ could lead to *Omitted variable bias*.

To sum up, given a vast set of potential control variable $\mathbf{X}$, the basic strategy behind regression analysis to estimate a causal effect is to include a *sufficient* set of *proper* control variables $\boldsymbol{S}$ into the model in addition to the treatment variable T. In other words, the set of controls $\boldsymbol{S}$ plays a role as regression adjustment/confounding adjustment to enable the causal interpretation of the estimated regression coefficient $\hat{\alpha}$. If we cannot control for $\boldsymbol{S}$ properly, we simply cannot obtain causal effect.

2.1.3 The Puzzle of Regression with High-dimensional Controls

There are some guidelines for deciding a candidate variable should be controlled or not. First, we should control for *confounders*, which are pre-treatment variables that determine both the treatment and the outcome. Exclusion of confounders in the regression model results in biased estimation of the treatment effect. Second, we should not control for *"noise" covariates* since inclusion of them may lead to low precision estimates of the treatment effect. However, it seems still vague to verify a set of controls $\boldsymbol{S}$ is *proper* and *sufficient*.

[3] An alternative condition to standard normal distributed error, $\mathbb{E}[\epsilon_i \mid T_i, \boldsymbol{X_i}] = 0$

[4] *good* is used in the sense of desirable properties such as unbiased, consistent, etc.

Empirical researchers often face two issues in realistic implementation: First, which factors are important confounders are never known exactly, thus it would be difficult to categorize initially. Even when one may get some intuitions from economic theories or prior knowledge, the corresponding function forms of relevant variables are still uncertain. Second, the set of potential control variables is often quite large relative to the available sample size. This candidate set could be vast because of the number of baseline factors themselves, or as a result of incorporating many of their possible transformations (polynomials or interactions) as regressors.

There are several ways to response to this dilemma. In early days of empiricism, big set of choices facilitates researcher degrees of freedom[5] [Simmons et al., 2011]. As pessimistically pointed out by Leamer [1983], it is easy for economists to proclaim a seemingly significant finding by cherry-picking a small subset of the potential controls and proceeding with their analysis. Under the influence of "a credibility revolution" in modern empirical work in economics [Angrist and Pischke, 2010], sensitivity analysis has been regarded as a remedy for reducing such false positives. More often than not, researchers often present in the final paper estimates of various sets of controls in order to convey that their substantive result for the treatment effect is insensitive to changes in the set of control variables. Nonetheless, such procedure of robustness checking is still ad hoc. Moreover, in the context of linear regression model armed with a tradition estimation approach like Ordinary Least Squares (OLS), examining the specification with high-dimensional data is impossible in many cases. The term *high-dimensional data* is used here in the sense that the number of regressors (p) being comparable or even larger than the number of observations (n). Particularly, if p is greater than n, the OLS estimator for the parameter of interest is simply non-existent. Even when we restrict our attention to full rank setting (n remains slightly larger than p), ill-posed problems/overfitting problems may occur and make the estimates far less precise.

These problems emerged from common practice continue demanding a more systematic and data-driven way of searching for a small set of influential confounders among the initial broad set of covariates, thereby supporting to a more reliable treatment effect estimation. This consideration suggests that *statistical regularization* method, which is well-known as one of the most effective solution for some primary goals (such as prediction and variable selection) in high-dimensional setting [Hastie et al., 2009], signals it own potential to cover the gap.

[5]Researchers may consciously or unconsciously choose controls to generate results they want

2.2 Regularization and the Duality

By definition, regularization is a statistical technique widely used to solve an *ill-posed problem* for the purpose of *stability* or to guard against *overfitting* for the purpose of *generalization*. These goals can be achieved by introducing additional prior information about the desired solution to the underlying model. In particular, from a classical viewpoint, this information is usually in the form of a penalty for some spectral components of the solution. Interestingly, many regularization techniques correspond to imposing certain prior distributions on model parameters under the Bayesian framework.

Within the scope of this brief review, we focus on regularized regression methods in a popular high-dimensional setting. There are several reasons why this approach has enormous potential to solve the puzzle of regression with high-dimensional controls, which we have discussed in the section 2.1. Generally speaking, regularization methods exert a shrinkage effect on regression coefficients, as an optimal trade-off between model complexity (bias) and out-of-sample performance (variance). Particularly, when the number of predictors p is larger than the sample size n, these methods are able to select variables out of a large set of variables that are relevant for predicting the outcome. Furthermore, even when the number of predictors p is smaller than the sample size n (yet still relatively large), these techniques are beneficial in terms of avoiding overfitting and achieving model parsimony compared to traditional variable selection methods. The advantages regarding prediction and variable selection of regularized regression methods are shown in many previous studies, e.g., Wang et al. [2020] provides a large-scale comparison of different frequentist regularization methods in terms of three related goals - prediction, variable selection and variable ranking in high-dimensional data; Van Erp et al. [2019] and Polson and Sokolov [2019] provide a comprehensive overview of Bayesian shrinkage priors and illustrate their various behaviour in terms of variable selection.

In terms of implementation, a frequentist incorporates a variety of regularizers into a measure of fit, while a Bayesian combines various hierarchical priors with the likelihood. The last several years have recorded tremendous interest in the equivalence between two underlying mechanisms for the same goal of acquiring attractive properties. In parallel with understanding the duality, we will revisit some well-known regularizers and shrinkage priors to set the foundation for our further applications.

2.2.1 Frequentist Regularization

Originally, regularization can be viewed as a constraint on the model space. Consider a classic linear regression model with normally distributed errors

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \sigma_\epsilon^2\, \boldsymbol{I}\right) \tag{2.3}$$

The corresponding regularized maximum likelihood optimization problem is defined as

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \quad \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \quad \text{subject to} \quad \phi(\boldsymbol{\beta}) \leq s \tag{2.4}$$

of which the solution could be written in the alternative form

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \quad \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\phi(\boldsymbol{\beta}) \tag{2.5}$$

where:

- $\boldsymbol{y}$ is the vector of observed outcomes, $\boldsymbol{X}$ is a design matrix, $\boldsymbol{\beta}$ are the model parameters.
- s in (2.4) or λ in (2.5) is a tuning parameter (hyper-parameter) controlling the strength of the penalty.
- $\phi(\boldsymbol{\beta})$ is a regularization term (penalty).

In general, a separable penalty is of the form: $\phi(\boldsymbol{\beta}) = \sum_{j=1}^{p} \phi(\beta_j)$, where $\phi(\beta_j)$ is a penalty function applied for each component β_j. In fact, each appropriate choice of the regularization term is associated with a desirable estimator in (2.5). Mathematically, a regularized solution can be defined by constraining the topology of a search space to a ball.

Ridge

The Ridge estimator [Hoerl and Kennard, 1970] takes the form in (2.5) with an ℓ_2-norm penalty

$$\lambda\phi(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} \phi(\beta_j) = \lambda \sum_{j=1}^{p} \beta_j^2 = \lambda \|\boldsymbol{\beta}\|_2^2 \tag{2.6}$$

The solution is given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} \tag{2.7}$$

As a special case of Tikhonov regularization, Ridge regression aim to address the issue of numerical instability when $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$ is ill-conditioned, which is always a case whenever p is large. Intuitively, the larger tuning parameter λ is, the stronger shrinkage toward zero put on the coefficients. However, the ℓ_2-norm penalty leads to non-sparse solutions because it is not singular at the origin.

Lasso

The Lasso estimator [Tibshirani, 1996] employs an ℓ_1-norm penalty in (2.5):

$$\lambda\phi(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} \phi(\beta_j) = \lambda \sum_{j=1}^{p} |\beta_j| = \lambda \|\boldsymbol{\beta}\|_1^1 \tag{2.8}$$

While being able to perform shrinkage like Ridge, Lasso performs explicit variable selection by making some of the coefficients exactly 0 and producing a true sparse solution. This feature helps distinguish Lasso as a selection-based method from shrinkage-based methods Ridge represents. The underlying explanation is "$\ell_1 - polytope$, unlike $\ell_2 - polytope$, can touch the contours of the least-squares objective function on one or more of the axes leading to estimates of zero for the associated regression coefficients." The tuning parameter λ still plays a role in controlling for the amount of shrinkage and degree of sparsity.

While this superiority gives rise to the popularity of the Lasso method, there are several disadvantages of classical Lasso recognized. "Specifically, (i) it cannot select more predictors than observations, which is problematic when $p > n$; (ii) when a group of predictors is correlated, the lasso generally selects only one predictor of that group; (iii) the prediction error is higher for the lasso compared to the ridge when $n > p$ and the predictors are highly correlated; (iv) it can lead to over-shrinkage of large coefficients [Polson and Scott, 2010]; and (v) it does not always have the oracle property, which implies it does not always perform as well in terms of variable selection as if the true underlying model has been given [Fan and Li, 2001]. The lasso only enjoys the oracle property under specific and stringent conditions [Fan and Li, 2001, Zou, 2006]."

Elastic Net

The Elastic Net estimator [Zou and Hastie, 2005] is (2.5) using a penalty:

$$\lambda\phi(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} \phi(\beta_j) = \lambda \sum_{j=1}^{p} \left[\alpha|\beta_j| + (1-\alpha)\beta_j^2\right] \tag{2.9}$$

That can be seen as a hybrid between ℓ_1 and ℓ_2-norm penalties with a mix parameter $\alpha \in [0, 1]$ for the sake of mitigating drawbacks of each component. Clearly, when $\alpha = 1$ and $\alpha = 0$, the Elastic Net estimator corresponds to the Lasso estimator and Ridge estimator, respectively. While still enjoying some of the benefits of Ridge, Elastic Net can give sparse solutions. In comparison with Lasso, Elastic Net overcomes the limitation of selecting at most n variables in the high-dimensional setting ($p > n$).

However, "a disadvantage of the Elastic Net is that the sequential cross-validation procedure used to determine the tuning parameter results in over-shrinkage of coefficients".

2.2.2 Bayesian Regularization

From a Bayesian perspective, regularization is instead performed by assigning a prior distribution over the model parameters. Consider a Bayesian linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \sigma_\epsilon^2\, \boldsymbol{I}\right), \quad \boldsymbol{\beta} \sim p(\boldsymbol{\beta} \mid \lambda) \tag{2.10}$$

The logarithm of the posterior distribution is then given by

$$-\log p(\boldsymbol{\beta} \mid \boldsymbol{X}, \boldsymbol{y}) = \frac{1}{2\sigma_\epsilon^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \text{const} + \log p(\boldsymbol{\beta} \mid \lambda) \tag{2.11}$$

A regularized maximum a posteriori (MAP) estimate can be found by minimizing the negative log-posterior:

$$\hat{\boldsymbol{\beta}}_{\text{MAP}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \quad \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \phi_\lambda(\beta) \tag{2.12}$$

where:

- $\boldsymbol{y}$ is the vector of observed outcomes, $\boldsymbol{X}$ is a design matrix, $\boldsymbol{\beta}$ are the model parameters.
- The regularization term (penalty) $\phi_\lambda(\beta) \propto \log p(\boldsymbol{\beta} \mid \lambda)$ is interpreted as the log of the prior distribution, and is parametrized by the hyper-parameters λ.

Therefore, under an appropriate regularization prior which induces the corresponding penalty, the resulting maximum a posteriori (MAP) estimate[6] in (2.12) is equivalent to the point estimate from a frequentist regularization in (4). That is a key duality between two standpoints. Some typical examples are an ℓ_2 penalty (or Ridge) [Tikhonov, 1963, Hoerl and Kennard, 1970] corresponding to a Gaussian prior under the same observation distribution, and an ℓ_1 penalty (or Lasso) corresponding to a double-exponential prior [Tibshirani, 1996]. Figure 2.1 compares the geometry of a unit ball used as a constraint in the frequentist approach and the corresponding prior distribution used in the Bayesian approach.

Combining a thorough review and well-designed simulations, Van Erp et al. [2019] and Polson and Sokolov [2019] advocate the following merits of Bayesian regularization methods in comparison to frequentist counterparts. The first two advantages are often cited, while the rest are more specific to regularization context:

[6] can also be interpreted as Bayesian mode of the posterior distribution (assumed to be unimodal)/ a posterior mode.

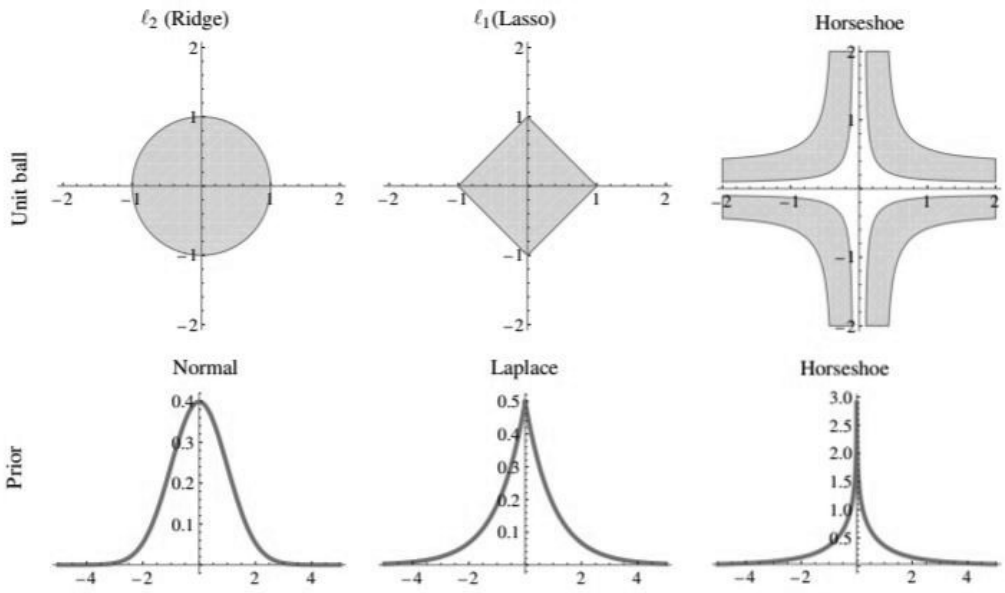

Figure 2.1. Comparison of the geometry of a unit ball induced by Normal, Laplace and Horseshoe priors [Polson and Sokolov, 2019]

- *Automatic uncertainty estimates*
  Bayesian estimation procedures result in posterior distributions over parameters and enable analyses of uncertainty in estimates and predictions. In contrast, frequentist regularized regression procedures can result in estimated standard errors that suffer from multiple problems, such as unstable or poorly performing variance estimates as shown by Casella et al. [2010].

- *Intuitive interpretations*
  Bayesian estimates have intuitive interpretations. For instance, a 95% Bayesian credibility interval can simply be interpreted as the interval which contains the true value with 95% probability.

- *Natural penalization through the prior distribution*
  Penalization can be incorporated naturally within a Bayesian framework through the prior distribution. Specifically, we can choose the prior distribution in such a way that it will shrink small effects towards zero while keeping substantial effects large. Thus, the prior performs similarly to the penalty term in frequentist regularized regression. Moreover, these Bayesian analogues of frequentist regularization methods have been shown to perform at least as good as and in some cases better than the classical

penalization methods (e.g., Casella et al. [2010], Li and Lin [2010]).

- *Simultaneous estimation of the penalty parameter*
  In Bayesian regularization, the penalty parameter can be given its own prior distribution, thereby being estimated with other model parameters in a single step. Additionally, compared to cross-validated selection in the frequentist approach, averaging over the penalty parameter in the Bayesian paradigm has been empirically observed to produce better prediction performance (e.g., Hans [2009]).

- *Flexibility in types of penalties*
  Bayesian regularization offers flexibility in terms of choosing the type of penalties. Frequentist regularization methods rely on optimization techniques to find the minimum of the regularized regression function and gravitate towards convex penalty functions, which results in one minimum. By contrast, Bayesian regularized regression utilizes MCMC sampling, which allows more straightforward implementation of penalties that are not convex. As a result, it enables a more flexible set of models that closely match the data generating process.

Nonetheless, these advantages come at the cost of computation and sparsity. The two main appeals of the frequentist regularization methods, efficient computation and sparse solution vectors of estimated coefficients, were lost in the migration to a Bayesian approach [Hahn and Carvalho, 2015].

Recalling with a suitable regularization prior corresponding to the particular penalty, the resulting posterior mode in (2.12) is equivalent to the point estimate of a classic problem in (2.5), we are now in the position of reviewing some of such priors which are popular in the Bayesian literature. Following the overview of Van Erp et al. [2019], we present in turn different priors in a common framework by setting a standard half-Cauchy priors for the hyper-parameter (penalty parameter) λ as a robust default choice of prior. The conditional densities (given the penalty parameter) for the surveyed shrinkage priors are presented in Table 2.1.

Table 2.1. Conditional prior densities for the regression coefficients $\boldsymbol{\beta}$ implied by the various shrinkage priors and references for each shrinkage prior.

Shrinkage prior	Conditional prior density $p(\beta_j \mid \lambda, \ldots)$	Reference
Ridge	$p\left(\beta_j \mid \sigma^2, \lambda\right) = \sqrt{\frac{\lambda}{2\pi\sigma^2}} \exp\left\{-\frac{\lambda\beta_j^2}{2\sigma^2}\right\}$	Hsiang [1975]
Lasso	$p\left(\beta_j \mid \sigma^2, \lambda\right) = \frac{\lambda}{2\sqrt{\sigma^2}} \exp\left\{\frac{-\lambda\lvert\beta_j\rvert}{\sqrt{\sigma^2}}\right\}$	Park and Casella [2008]
Elastic net	$p\left(\beta_j \mid \sigma^2, \lambda_1, \lambda_2\right) = C \exp\left\{-\frac{1}{2\sigma^2}\left(\lambda_1 \lvert\beta_j\rvert + \lambda_2\beta_j^2\right)\right\}$	Li and Lin [2010]
Student-t	$p\left(\beta_j \mid \sigma^2, \lambda\right) = \frac{\sigma^2}{\pi\lambda}\left(1 + \left(\frac{\sigma^2}{\lambda\beta_j}\right)^2\right)$	Meuwissen et al. (2001)
Horseshoe	Not analytically tractable	Carvalho et al. [2010]
Spike-and-slab	$p\left(\beta_j \mid \gamma_j, \phi_j^2\right) = (1-\gamma_j)\left(\frac{1}{\sqrt{2\pi\phi_j}^2} \exp\left\{-\frac{\beta_j^2}{2\phi_j^2}\right\}\right) + \gamma_j\left(\frac{1}{\pi\left(1+\beta_j^2\right)}\right)$	Mitchell and Beauchamp [1988]

Note. C denotes a normalization constant.

(Bayesian) Ridge

The Ridge penalty in (2.6) corresponds to Gaussian prior centered around 0 on the regression coefficients [Hsiang, 1975], which has a simple structure as follows:

$$\begin{aligned} \beta_j \mid \lambda, \sigma^2 &\sim \text{Normal}\left(0, \frac{\sigma^2}{\lambda}\right), \text{ for } j = 1, \ldots, p \\ \lambda &\sim \text{ half-Cauchy } (0, 1) \\ \sigma^2 &\sim \frac{1}{\sigma^2} \end{aligned} \tag{2.13}$$

The penalty parameter λ determines the amount of shrinkage, with larger values resulting in smaller prior variance and thus more shrinkage of the coefficients towards zero.

(Bayesian) Lasso

The Bayesian counterpart of the Lasso penalty in (2.8) is Laplace prior, which was first proposed by Park and Casella [2008]. The Bayesian Lasso can be obtained as a scale mixture of a Normal density with an Exponential density as below:

$$\begin{aligned} \beta_j \mid \tau_j^2, \sigma^2 &\sim \text{ Normal } \left(0, \sigma^2\tau_j^2\right) \\ \tau_j^2 \mid \lambda^2 &\sim \text{ Exponential } \left(\frac{\lambda^2}{2}\right), \text{ for } j = 1, \ldots, p \\ \lambda &\sim \text{ half-Cauchy } (0, 1) \\ \sigma^2 &\sim \frac{1}{\sigma^2} \end{aligned} \tag{2.14}$$

Integrating τ_j^2 out leads to Double-exponential[7] or Laplace priors on the regression coefficients, i.e.,

$$\beta_j \mid \lambda, \sigma \sim \text{ Double-exponential } \left(0, \frac{\sigma}{\lambda}\right), \text{ for } j = 1, \ldots, p \tag{2.15}$$

Although this version of Bayesian Lasso is the most popular form in literature so far; there are also some alternative formulations suggested by Hans [2009], Mallick and Yi [2014] and Alhamzawi and Taha Mohammad Ali [2020].

In addition to the overall shrinkage parameter λ, the lasso prior has an additional predictor-specific shrinkage parameter τ_j. Therefore, the Lasso prior is more flexible than the Ridge prior, which only relies on the overall shrinkage parameter in (2.13). Figure 2.1 clearly shows that the Lasso prior has a sharper peak around zero compared to the ridge prior.

(Bayesian) Elastic Net

The Bayesian dual prior of Elastic Net penalty in (2.9) can be obtained as the following scale mixture of normal densities [Li and Lin, 2010]:

$$\begin{aligned}
\beta_j \mid \lambda_2, \tau_j, \sigma^2 &\sim \text{ Normal } \left(0, \left(\frac{\lambda_2}{\sigma^2}\frac{\tau_j}{\tau_j - 1}\right)^{-1}\right) \\
\tau_j \mid \lambda_2, \lambda_1, \sigma^2 &\sim \text{ truncated-Gamma } \left(\frac{1}{2}, \frac{8\lambda_2\sigma^2}{\lambda_1^2}\right), \text{ for } j = 1, \ldots, p \\
\lambda_1 &\sim \text{ half-Cauchy } (0, 1) \\
\lambda_2 &\sim \text{ half-Cauchy } (0, 1)
\end{aligned} \tag{2.16}$$

where the truncated Gamma density has support $(1, \infty)$. This implies the following conditional prior distributions for the regression coefficients:

$$\begin{aligned}
&p\left(\beta_j \mid \sigma^2, \lambda_1, \lambda_2\right) = C\left(\lambda_1, \lambda_2, \sigma^2\right) \exp\left\{-\frac{1}{2\sigma^2}\left(\lambda_1 |\beta_j| + \lambda_2 \beta_j^2\right)\right\} \\
&\text{for } j = 1, \ldots, p
\end{aligned} \tag{2.17}$$

where $C\left(\lambda_1, \lambda_2, \sigma^2\right)$ denotes the normalizing constant. This expression illustrates how the Elastic Net prior offers a combination of the double-exponential prior, i.e., the Lasso penalty

[7] Mathematical representation:

$$\int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2 s_j}} e^{\left(-\frac{\beta_j^2}{2\sigma^2 s_j}\right)} \frac{\lambda^2}{2} e^{-\frac{\lambda}{2s_j}} ds_j = \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda|\beta_j|/\sqrt{\sigma^2}}$$

$\lambda |\beta_j|$, and the normal prior, i.e., the Ridge penalty $\lambda \beta_j^2$. Specifically, the two hyper-parameters λ_1 and λ_2 determine the relative influence of the Lasso and Ridge penalty, respectively. By estimating these two hyper-parameters simultaneously, the Bayesian version of Elastic Net overcomes the over-shrinkage problem of the classical version.

Student-t prior

Student-t prior (also named as Normal-iGamma prior) can be seen as an extension of Gaussian prior for Ridge penalty in (2.13) by making the prior variances predictor-specific, thereby allowing for more variation, i.e.,

$$\begin{aligned} \beta_j \mid \tau_j^2 \sim & \text{ Normal } \left(0, \sigma^2 \tau_j^2\right) \\ \tau_j^2 \mid \nu, \lambda \sim & \text{ inv-Gamma } \left(\frac{\nu}{2}, \frac{\nu}{2\lambda}\right), \text{ for } j = 1, \ldots, p \\ \lambda \sim & \text{ half-Cauchy } (0, 1) \end{aligned} \tag{2.18}$$

When integrating τ_j^2 out, the following conditional prior distribution for the regression coefficients is obtained:

$$\beta_j \mid \nu, \lambda, \sigma^2 \sim \text{Student} \left(\nu, 0, \frac{\sigma^2}{\lambda}\right) \tag{2.19}$$

where Student $\left(\nu, 0, \frac{\sigma^2}{\lambda}\right)$ denotes a non-standardized Student's t distribution centered around 0 with ν degrees of freedom and scale parameter $\frac{\sigma^2}{\lambda}$. A smaller value for ν results in a distribution with heavier tails, with $\hat{\nu} = 1$ implying a Cauchy prior for β_j. Larger (smaller) values for λ result in more (less) shrinkage towards m. This prior is considered by Griffin and Brown [2005]. Compared to the ridge prior in in (2.13), the local Student's t prior has heavier tails.

Horseshoe prior

A novel shrinkage prior in the Bayesian literature is the horseshoe prior [Carvalho et al., 2010][8]. This prior is particularly attractive for sparse signal recovery.

$$\begin{aligned} \beta_j \mid \tau_j^2 \sim & \text{ Normal } \left(0, \tau_j^2\right) \\ \tau_j \mid \lambda \sim & \text{ half-Cauchy } (0, \lambda), \text{ for } j = 1, \ldots, p \\ \lambda \mid \sigma \sim & \text{ half-Cauchy } (0, \sigma) \end{aligned} \tag{2.20}$$

[8]Note that [Carvalho et al., 2010] explicitly include the half-Cauchy prior for λ in their specification, thereby implying a full Bayes approach. This formulation results in a horseshoe prior that is automatically scaled by the error standard deviation σ.

The half-Cauchy prior can be written as a mixture of inverse Gamma densities[9] [Makalic and Schmidt, 2015], so that the horseshoe prior in (2.20) can be equivalently specified as:

$$\begin{aligned}
\beta_j \mid \tau_j^2 &\sim \text{ Normal } \left(0, \tau_j^2\right) \\
\tau_j^2 \mid \omega &\sim \text{ inv-Gamma } \left(\frac{1}{2}, \frac{1}{\omega}\right) \\
\omega \mid \lambda^2 &\sim \text{ inv-Gamma } \left(\frac{1}{2}, \frac{1}{\lambda^2}\right) \\
\lambda^2 \mid \gamma &\sim \text{ inv-Gamma } \left(\frac{1}{2}, \frac{1}{\gamma}\right) \\
\gamma \mid \sigma^2 &\sim \text{ inv-Gamma } \left(\frac{1}{2}, \frac{1}{\sigma^2}\right)
\end{aligned} \tag{2.21}$$

An expression for the marginal prior of the regression coefficients β_j is not analytically tractable, but a tight lower bound [Carvalho et al., 2010] can be used instead

$$-\log p\left(\beta_i \mid \lambda\right) \geq -\log\log\left(1 + \frac{2\lambda^2}{\beta_j^2}\right) \tag{2.22}$$

The key features for the appealing performance of horseshoe prior are its Cauchy-like tails and an asymptote at origin (unique advantage), which make horseshoe adaptive to sparsity and robust to large signals so outperform other shrinkage priors we have discussed.

In the search for intuitive reasons, we consider a common framework of shrinkage rules. Define $\kappa_j = 1/(1+\tau_j^2)$, then κ_j is a random shrinkage coefficient in $[0, 1]$. Under a multivariate normal scale mixture prior (i.e. the general form of all shrinkage priors we are discussing), the posterior mean can be written as a linear function of the observation:

$$\mathbb{E}\left[\beta_j \mid y_j\right] = \left\{1 - \mathbb{E}\left[\kappa_i \mid y_j\right]\right\} y_j \tag{2.23}$$

Hence, $\mathbb{E}\left[\kappa_i \mid y_j\right]$ implies the amount of weight that the posterior mean for β_j places on 0 once the data have been observed. A shrinkage coefficient κ_j that is close to zero leads to virtually no shrinkage, thus describes signals. A shrinkage coefficient κ_j that is close to one leads to nearly-total shrinkage, thus describes noises. Intuitively speaking, the behavior of a priori $p(\kappa_j)$ near $\kappa_j = 1$ will control the robustness of signal at tail, while near $\kappa_j = 0$ will control the shrinkage of noise toward 0. Because of difference choice of $p(\tau_j)$, each type of shrinkage prior has distinct $p(\kappa_j)$ reflecting its attempt to separate signal and noise. For horseshoe prior, the attempt is even implied in its name, which arises from the fact that for fixed values $\lambda = \sigma = 1$, $p(\kappa_j)$ is similar to a horseshoe-shaped Beta $(1/2, 1/2)$. This prior is symmetric and unbounded at both 0 and 1; thereby, small coefficients (noises) are heavily

[9] If $x^2 \mid z \sim$ inv-Gamma $(1/2, 1/z)$ and $z \sim$ inv-Gamma $\left(1/2, 1/\alpha^2\right)$ then $x \sim C^+(0, \alpha)$

shrunken towards zero while substantial coefficients (signals) remain large. None of these common shrinkage priors above shares this characteristic. For instance, the Laplace prior, where $p(\kappa_j)$ is bounded at both 0 and 1, tends to over-shrink strong signals yet under-shrink noises. Carvalho et al. [2009, 2010] provide more explanation for other priors.

In fact, unlike local shrinkage priors above, the horseshoe prior is a member of a wider class of *global-local shrinkage priors* [Polson and Scott, 2010, Bhadra et al., 2019] because it enables a clear separation between global and local shrinkage effects. Put another way, this class of priors adapt to sparsity by a global shrinkage parameter and recover signals by a local shrinkage parameter.

Spike-and-slab priors

Apart from the Bayesian priors considered so far, which are all continuous mixtures of normal densities, a *spike-and-slab prior* is a discrete mixture of a peaked prior around zero (the spike) and a vague proper/heavy-tailed prior (the slab). The spike-and-slab prior is first proposed by Mitchell and Beauchamp [1988] and has the popular form as below

$$p\left(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, \sigma^2\right) = \prod_{j=1}^{p} \left[(1-\gamma_j)\delta_0(\beta_j) + \gamma_j p\left(\beta_j \mid \sigma^2\right)\right] \tag{2.24}$$

where δ_0 is a point mass at zero - the concentrated "spike distribution" to model the negligible (small) effect and $p\left(\beta_j \mid \sigma^2\right)$ is a diffuse "slab distribution" to model the non-negligible (large) effects, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, ..., \gamma_n)', \gamma_i \in \{0, 1\}$ is a binary vector that indexes the 2^p possible subset model. Then, based on the data, regression coefficients close to zero will be assigned to the spike, resulting in shrinkage towards 0, while coefficients that deviate substantially from zero will be assigned to the slab, resulting in (almost) no shrinkage.

Although the above point-mass spike and slab prior is often regarded as "theoretically ideal", or a "gold standard" for sparse Bayesian problems, deriving the corresponding full posterior over the entire model space can be a computational burden as a result of the complexity of updating the discrete indicators γ. Alternatively, some variant spike-and-slab models have been developed to replace the point-mass δ_0 by a continuous density that is heavily concentrated about zero. One of such continuous relaxation was made by George and McCulloch [1993][10]. Specifically, they used a normal density with very small variance for the spike and a normal density with very large variance for the slab and propose the

[10]See also George and McCulloch [1997], who describes and compares various hierarchical mixture prior formulations of variable selection uncertainty in normal linear regression models.

following prior for $\boldsymbol{\beta}$

$$p\left(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, \sigma^2\right) = \prod_{j=1}^{p} \left[(1-\gamma_j)\ \text{Normal}\ \left(0, \sigma^2\tau_0^2\right) + \gamma_j\ \text{Normal}\ \left(0, \sigma^2\tau_1^2\right)\right] \tag{2.25}$$

where $0 < \tau_0^2 \ll \tau_1^2$. Ishwaran and Rao [2005] and Narisetty and He [2014] further extended the model by rescaling the variances τ_0^2 and τ_1^2 with sample size n in order to better control the amount of shrinkage for each individual coefficient (particularly suitable for regression settings with very large numbers of covariates). From the regularization viewpoint, prior (2.25) is associated with the spike-and-slab penalty using a mixture of two normal distribution

$$\phi(\beta_j) = -\log\left[\left(\frac{\gamma_j}{\sqrt{2\pi\sigma^2\tau_1^2}}\right) e^{-\frac{\beta_j^2}{\sigma^2\tau_1^2}} + \left(\frac{1-\gamma_j}{\sqrt{2\pi\sigma^2\tau_0^2}}\right) e^{-\frac{\beta_j^2}{\sigma^2\tau_0^2}}\right] \tag{2.26}$$

In terms of variable selection, it could be useful to take into account two auxiliary variables, the indicator γ_j (above) and the latent effect size α_j. One reasonable interpretation for (2.25) is that the regression coefficient and the effect size are the same, $\beta_j \equiv \alpha_j$, thus obviously share the same distribution. Since the indicator affect the prior distribution of α_j (i.e. β_j), so that the prior $\mathbb{P}(\alpha_j \mid \gamma_j = 0)$ would influence posterior. Given some difficulties when tuning hyper-parameters, George and McCulloch [1993] developed a *Stochastic Search Variable Selection* (SSVS) procedure based on posterior sampling with MCMC and thresholding the posterior inclusion probabilities, $\mathbb{P}\left(\gamma_j = 1 \mid \boldsymbol{y}\right), \forall j = \overline{1,p}$. Thereby, the prior in (2.25) is also known as a *SSVS prior*.

Another way to perform variable selection is by re-parametrizing $\beta_j \equiv \gamma_j\alpha_j$, where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_j)$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_j)$ are two independent random variables satisfying:

$$\begin{aligned} \alpha_j \mid \sigma^2 &\sim\ \text{Normal}\ (0, \sigma^2) \\ \gamma_j \mid \theta &\sim\ \text{Bernoulli}\ (\theta) \end{aligned} \tag{2.27}$$

Unlike SSVS prior, this probabilistic structure implies that the prior $\mathbb{P}(\alpha_j \mid \gamma_j = 0)$ has no impact on posterior; therefore, no tuning is required while coefficients can be made exactly zero at positive probability. This is a variable selection approach proposed by Kuo and Mallick [1998]. The recent paper by Polson and Sun [2019] draws an interesting connection between the prior in (2.27) and ℓ_0-regularization.

2.3 Regularization when the Goal is Causal Inference

In section 2.1, we anticipate that regularization could come in to play a role in treatment effect estimation using observational data where the number of control variables is relatively

large compared to the available sample size. Indeed, there are a host of regularization methods from both Frequentist and Bayesian literature, each of which has its own competitive advantages in specific contexts, as illustrated in section 2.2. Nevertheless, we should keep in mind that these tools are originally designed for the purpose of prediction (and sometimes variable selection), while our ultimate goal here is causal inference. Several authors have pointed out that Frequentist and Bayesian regularization procedures that focus on predicting y perform poorly when the inferential goal is estimation of the treatment effect of T on y [Candes and Tao, 2007, Belloni et al., 2014b, 2017, Wang et al., 2012, Hahn et al., 2020]. Even though a well-performed selection of control variables is our main focus and could be partly tackled by a proper regularization approach, this should be considered as a means to an end rather than an end itself. This rationale leads us to another essential question: What is an appropriate procedure for valid inference of treatment effect in parallel with seeking for necessary control variables?

Interestingly, this conundrum has been again investigated by both Frequentists and Bayesians. They both admonish naive regularization approaches while looking for novel solutions from their viewpoint. Several prominent strategies are summarized below, although we would start from some naive methods as benchmarks.

2.3.1 Frequentist Approach

We turn back to the starting point - treatment effect estimation problem in an observational study, focus on a linear regression model where the treatment variable T_i is taken as exogenous after conditioning on control variables

$$y_i = \underbrace{\alpha T_i}_{aim} + \underbrace{\boldsymbol{X}_i\boldsymbol{\beta}}_{nuisance} + \epsilon_i, \quad \forall i = 1, \ldots, n \tag{2.28}$$

with components are defined in section 2.1.

Naive Procedures

The first naive approach would be to apply Lasso to the equation (2.28) while excluding α from the ℓ_1 penalty to impose that T_i always remains in the model, and then interpret that the estimate of α directly obtained from this model reflects treatment effect. Even if the final parsimonious model contains exactly the right control variables, this is a mistake since Lasso estimators lack valid confidence intervals.

The second naive approach is Post-Single-Selection Lasso, a common practice criticized by Belloni et al. [2014a], which could be implemented in two steps:

- Step 1: Use Lasso to estimate equation (2.28) while excluding α from the ℓ_1 penalty (the same as the first naive approach).
- Step 2: Refit the model by least squares after selection (i.e., regress y_i on T_i and only those covariates selected by Lasso as controls), use standard confidence interval and make inference on treatment effect.

The idea of the Post-Single-Selection procedure for inference relies on perfect model selection. Unfortunately, this condition is almost unobtainable, so that the procedure can fail miserably. In fact, Lasso targets prediction, not learning about specific model parameters. Thus, any control variable that is highly correlated with T_i but weakly with y_i tends to drop out of the selection because adding such a control variable does not increase predictive power for y_i so much. As a consequence, this approach fails to describe the relationship between T_i and $\boldsymbol{X_i}$, which is a key to understand the omitted variable bias as well as confounding selection. Moreover, the structural model (2.28) is not representing a prediction rule for y_i given T_i and x_i. Hence, it is not adequate to apply regularization methods such as Lasso directly.

Belloni et al. [2014b], therefore, propose to work with the following system of two reduced-form equations:

$$\text{Treatment Eq.: } T_i = \boldsymbol{X_i}\boldsymbol{\beta_t} + \nu_i, \tag{2.29}$$

$$\text{Outcome Eq.: } y_i = \boldsymbol{X_i}\boldsymbol{\beta_y} + \eta_i \tag{2.30}$$

where $\boldsymbol{\beta_y} = \alpha\boldsymbol{\beta_t} + \boldsymbol{\beta}$ and $\eta_i = \alpha\nu_i + \epsilon_i$ ($\boldsymbol{\beta}$ and ϵ_i are used in (2.28)).

Hence, the nuisance component in structural model (2.28) is modelled as a prediction problem. Since both equations above represent predictive relationships, one can apply regularization methods directly.

These authors also carefully advise against the third naive procedure that uses only one of two reduced form equations for selection because of the similar omitted-variable-bias reason as shown in Post-Single-Selection above. For instance, if we only consider the regression of T_i on the controls, we might miss the controls with a strong predictive power for y_i, but only a moderately sized effect on T_i.

Post-Double-Selection Lasso - Belloni et al. [2014b]

Post-Double-Selection Lasso recommended by Belloni et al. [2014b] includes three following steps:

- Step 1: Use Lasso to estimate Treatment equation (2.29), i.e., we aim to select a set of control variables that are useful for predicting the treatment T_i. Denote the set of Lasso-selected controls by S_1.
- Step 2: Use Lasso to estimate Outcome equation (2.30), i.e., we aim to select a set of control variables that are useful for predicting the outcome y_i. Denote the set of Lasso-selected controls by S_2.
- Step 3: Estimate structural model (2.28) by least squares using the union of selected controls in two above steps, i.e. $\boldsymbol{S_i} = S_1 \cup S_2$. Finally, we can do inference on the treatment effect α of interest.

In a nutshell, the Post Double Selection (PDS) procedure involves selection among the controls that predict *either* T_i and y_i to create robustness. Compared to the Post-Single-Selection procedure, PDS allows moderate selection mistakes and helps guard against "non-negligible" omitted variable bias caused by the omission of some important controls. In essence, this procedure is a model selection version of the Frisch-Waugh-Lovell partialling-out procedure for estimating linear regression with all selected controls from both y_i and T_i [Chernozhukov, 2015].

Once we get the intuition behind the PDS procedure, we may question the role of Lasso as a selection device in this approach[11]. Although all strategies from the naive to the state-of-the-art we have discussed here are related to Lasso, it is worth noting that other selection-based regularization methods might also share the same story. Belloni et al. [2014b] actually extend their Post-Double-Selection procedure to allow for a generic selection method in Frequentist literature (apart from their feasible Lasso, i.e. iterated/square-root Lasso) with one extra assumption [12]. They suggest some of the alternative satisfied methods include: thresholded Lasso [Belloni and Chernozhukov, 2011], Bridge estimator [Huang et al., 2008], Dantzig selector [Candes and Tao, 2007], feasible Dantzig selector [Gautier and Tsybakov,

[11] In fact, Belloni et al. [2014b] (page 20) note that: "Lasso methods generally will not recover $support(\beta_0)$ perfectly... We do not require this condition to hold in our results. All that we need is that the selected model can approximate the regression function well..."

[12] HLMS - with high probability the selected models are sparse and generate a good approximation for the functions g and m corresponding to treatment equation and outcome equation.

2013] and SCAD penalized least squares [Fan and Li, 2001], just to name a few. The key lies on Immunization property of PDS, which states that (possible) moderate selection mistakes of the selection method do not affect the asymptotic distribution of the estimator of the low-dimensional parameter of interest [Ahrens et al., 2020].

Another vital question is how to choose the tuning parameters λ, which can influence the set of selected variables after each selection. In the literature, there are two following popular approaches for selecting the penalty level:

- Theory-driven ("rigorous") approach (*Iterated Lasso*[13]): originally used by Belloni et al. [2014a] and series of their related papers. This approach is supported by theoretically justified and feasible penalty levels and loadings. The penalization is chosen to dominate the noise of the data-generating process (represented by the score vector), which allows the derivation of theoretical results[14] with regard to consistent prediction and parameter estimation. Rigorous penalization is of special interest because it provides the basis for methods to facilitate causal inference.

- Data-driven approach (*Cross-Validated Lasso*): the most popularly used in literature. While this approach is a powerful method for predictive purposes, it is often said to lack theoretical justification. The aim of cross-validation (CV) is to assess the performance of a model on unseen data directly. In the context of regularized regression, CV can be used to select the tuning parameters that yield the best performance; for example, the best out-of-sample mean squared prediction error. Compared to the theory-driven approach, CV Lasso can be computationally expensive and tend to yield more predictors (as a result of smaller penalties) than in the true model and when theory-driven regularization is used.

2.3.2 Bayesian Approach

For clarification before we proceed, the Bayesian approach in this section involves both the use of Bayesian regularization priors as well as the Bayesian inference procedure for treatment

[13]The term used by Belloni et al. [2014b]. More discussion can be found in Ahrens et al. [2020].

[14]The theory of the 'rigorous' LASSO has two main ingredients: *Restricted eigenvalue condition* (REC): REC is much weaker compared to the requirement of OLS - full rank condition, which is too strong in the high-dimensional context. *Penalization level:* We need λ to be large enough to 'control' the noise in the data. At the same time, we want the penalty to be as small as possible (due to shrinkage bias). This allows deriving theoretical results for the LASSO: consistent prediction and parameter estimation. The theory of Belloni et al. [2012] allows for non-Gaussian and heteroskedastic errors and has been extended to panel data [Belloni et al., 2016]

effect estimation.

Similar to pitfalls of the first naive approach with classic Lasso from the Frequentist standpoint, naive regularization which apply a shrinkage prior over β_y using only the structural model (2.28) and directly make inference on the treatment effect from an estimate of α can be problematic. As proved by Hahn et al. [2018], this approach performs poorly in coverage and produces severely biased results. They explain it as a consequence of "regularization-induced confounding", the phenomenon of regularization priors to adversely bias treatment effect estimates by over-shrinking control variable regression coefficients. They also emphasize this is an independent issue arising even when the unconfoundedness assumption is satisfied.

Furthermore, following the caveats of the Post-Single-Selection procedure in the Frequentist approach, a good procedure necessitates exploiting information from the treatment equation (2.29) to avoid omitted variable bias caused by ignoring controls that have a strong predictive power for T_i, but only a moderate effect on y_i. Based on this principle, most of the Bayesian approaches to confounding adjustment for causal inference rely on the specification of the treatment equation (2.29) and the structural-form equation (2.28), or of two reduced-form equations (2.29) and (2.30). For example, Wang et al. [2012] reparameterize the likelihood into a hierarchical model given by:

$$\begin{aligned} \text{Treatment Eq.:} \quad & (T_i \mid \boldsymbol{X}_i) = \boldsymbol{X}_i \boldsymbol{\beta_t} + \nu_i, \quad \nu_i \sim \text{ Normal } \left(0, \sigma_\nu^2\right) \\ \text{Structural Eq.:} \quad & (y_i \mid T_i, \boldsymbol{X}_i) = \alpha T_i + \boldsymbol{X}_i \boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \sim \text{ Normal } \left(0, \sigma_\epsilon^2\right). \end{aligned} \tag{2.31}$$

Their main idea is Bayesian adjustment for confounding (BAC), a variant of Bayesian model averaging to estimate the effect of T on y across various models to account for uncertainty in the set of confounders. The formulation (2.31) enables them to specify informative priors that favour the inclusion of covariates in the structural model if they appear in the treatment model. This approach sets a foundation for some extended works on confounders selection for heterogeneous treatment effect estimation [Talbot et al., 2015, Wang et al., 2015, Antonelli et al., 2017]. However, BAC requires calculating the Bayesian Information Criterion at each posterior draw, which cannot be identified when the number of potential controls surpasses the available sample size. Therefore, these methods are inapplicable in high-dimensional settings. Motivated by the idea of incorporating information from treatment equation into regularized regression on structural model, Antonelli et al. [2019] impose general spike-and-slab lasso priors over the coefficients of all potential control variables on $(y_i \mid T_i, \boldsymbol{X}_i)$ but reduce shrinkage for Lasso-selected variables in $(T_i \mid \boldsymbol{X}_i)$ (which are potentially important confounders). The appealing property of this approach lies in its good performance in high-dimensional confounding adjustment.

As an alternative approach to Wang et al. [2012], Hahn et al. [2018] use two reduced-form equations by applying the transformation $\boldsymbol{\beta_y} = \boldsymbol{\beta} + \alpha\boldsymbol{\beta_t}$ to (2.31), thus obtaining:

$$\begin{aligned} \text{Treatment Eq.:} \quad & (T_i \mid \boldsymbol{X}_i) = \boldsymbol{X}_i\boldsymbol{\beta_t} + \nu_i, \quad \nu_i \sim \text{ Normal } \left(0, \sigma_\nu^2\right) \\ \text{Outcome Eq.:} \quad & (y_i \mid T_i, \boldsymbol{X}_i) = \boldsymbol{X}_i\boldsymbol{\beta_y} + \alpha(T_i - \boldsymbol{X}_i\boldsymbol{\beta_t}) + \epsilon_i, \quad \epsilon_i \sim \text{ Normal } \left(0, \sigma_\epsilon^2\right). \end{aligned} \tag{2.32}$$

and then specify independent shrinkage priors for $\boldsymbol{\beta_t}$ and $\boldsymbol{\beta_y}$ simultaneously. This procedure aims to rectify "regularization-induced confounding," i.e., the bias in estimating α resulting from the naive regularization mentioned above. In comparison with Antonelli et al. [2019]'s, this approach has a computational advantage when using continuous shrinkage priors such as horseshoe rather than spike-and-slab type priors, thereby entailing the ease of posterior sampling. However, Woody et al. [2020] argue that even with careful regularization as a major target of Hahn et al. [2018], the inclusion of a large number of controls could still drown out any signal in the treatment effect given an insufficient number of observations, as shown in their empirical illustrations.

In addition to develop a new Bayesian approach for inference on treatment effects with high-dimensional controls, Antonelli et al. [2019] also provide a simulation study that takes account of many advanced methods from Frequentist approach such as Belloni et al. [2014b], Farrell [2015] and Athey et al. [2016]. Their simulation results are quite appealing: The authors' Bayesian framework for estimating causal effects of binary and continuous treatments in high-dimensional settings produces posterior credible intervals with higher finite-sample coverage compared to Frequentist counterparts. One possible explanation could be that Frequentist measures of uncertainty rely on asymptotic properties, while the Bayesian approach captures the uncertainty in the data (and provides statistically valid inference). The following is a brief description of Antonelli et al. [2019]'s idea, which is a benchmark for our general approach in the next chapter.

High-dimensional Confounding Adjustment - Antonelli et al. [2019]

Antonelli et al. [2019] consider the idea to borrow information from the treatment model to guide the amount of shrinkage in the outcome model. They then propose a spike-and-slab Lasso prior approach to the problem that the naive approach of just ignoring treatment equation and using regularization methods only on the structural equation has: the coefficient on a control variable that is highly correlated with $\boldsymbol{T}_i$ but weakly with $\boldsymbol{y}_i$ tends to be

shrunk to zero. For $j = \overline{1,p}$, their proposed hierarchical formulation is:

$$\begin{aligned}
\boldsymbol{y}_i \mid \boldsymbol{T}_i, \boldsymbol{X}_i, \beta_0, \alpha, \boldsymbol{\beta}, \sigma^2 \sim & \text{ Normal } \left(\beta_0 + \alpha \boldsymbol{T}_i + \boldsymbol{X}_i \boldsymbol{\beta}, \sigma^2\right) \quad \forall i = 1, \ldots, n \\
\beta_0, \alpha \sim & \text{ Normal } (0, K) \\
\beta_j \mid \gamma_j, \sigma^2 \sim & \gamma_j \psi_1 \left(\beta_j; \lambda_1, \sigma^2\right) + (1 - \gamma_j) \psi_0 \left(\beta_j; \lambda_0, \sigma^2\right) \\
\gamma_j \mid \theta, \omega_j \sim & \text{ Bernoulli } (\theta^{\omega_j}) \\
\theta \mid a, b \sim & \text{ Beta } (a, b) \\
\sigma^2 \mid c, d \sim & \text{ Inv-Gamma } (c, d)
\end{aligned} \tag{2.33}$$

where $\psi_0 (\beta_j; \lambda_0, \sigma^2) = \frac{\lambda_0}{2\sigma} e^{-\lambda_0 |\beta_j| / \sigma}$ and $\psi_1 (\beta_j; \lambda_1, \sigma^2) = \frac{\lambda_1}{2\sigma} e^{-\lambda_1 |\beta_j| / \sigma}$. Each of them is a Laplace distribution (corresponding to Lasso penalty - see (2.15)). The hyper-parameter λ_1 is fixed to be a small value, e.g. 0.1, so that the prior variance in the slab component $\psi_1(.)$ is high enough to be uninformative. Meanwhile, the hyper-parameter λ_0 for the spike component $\psi_0(.)$ is chosen via Empirical Bayes.

A new feature that they introduce is the weights ω_j which are tuning parameters that they use to prioritize variables to have $\gamma_j = 1$ if they are correlated with the treatment. Specifically, they firstly fit the standard Lasso in the treatment equation for predicting $\boldsymbol{T}$ given $\boldsymbol{X}$. For $\boldsymbol{x_j}$ with non-zero coefficient from the Lasso estimation, they then set $\omega_j = \delta$ for some $\delta \in (0, 1)$. For other variables, $\omega_j = 1$. On the one hand, a smaller value of δ leads to higher inclusion probability, hence, provides more protection against the omitted variable bias. On the other hand, one needs to ensure a small enough inclusion probability for an unimportant variable in the outcome model (that is $\boldsymbol{x_j}$ with $\beta_j = 0$). The conditional probability that $\boldsymbol{x_j}$ belongs to the slab component is

$$p^*_{\omega_j} \left(\beta_j \mid \theta, \lambda_0, \sigma^2\right) = P \left(\gamma_j \mid \beta_j, \lambda_0, \theta, \sigma^2, \omega_j\right) = \frac{\psi_1 (\beta_j; \lambda_1, \sigma^2) \theta^{\omega_j}}{\psi_1 (\beta_j; \lambda_1, \sigma^2) \theta^{\omega_j} + \psi_0 (\beta_j; \lambda_0, \sigma^2) (1 - \theta^{\omega_j})} \tag{2.34}$$

They first run a Gibbs sampler with $\omega_j = 1$ for all j and then plug in posterior means for the unknown coefficients in the above inclusion probability. The authors then choose $\delta \in (0, 1)$ as the smallest value of ω_j such that $p^*_{\omega_j} (0 \mid \theta, \lambda_0, \sigma^2)$ is less than 0.1.

CHAPTER

THREE

METHODOLOGY

3.1 Overview

In this study, we use both Frequentist and Bayesian regularization-based methods for inference on treatment effects from an observational study using a linear regression model under the *Unconfoundedness* assumption. With respect to Frequentist approach, we employ Post-Double-Selection Lasso (`PDSLasso`, hereafter) proposed by Belloni et al. [2014b]. About Bayesian approach, we generalize the High-dimensional Confounding Adjustment (`HDCA`, hereafter) framework of Antonelli et al. [2019] by utilizing various Bayesian shrinkage priors.

Since both `PDSLasso` and `HDCA` involve Lasso for variable selection to a certain extent, implementation details should be taken into consideration. Particularly, Belloni et al. [2014b] use Lasso in both selection steps corresponding to the treatment equation and the outcome equation. They specifically develop *Iterated Lasso*, which is supported by a theory-driven penalty level. Meanwhile, Antonelli et al. [2019] use Lasso in the first step corresponding to the treatment equation to decide an active covariate set. They apply *CV Lasso* as a common approach but does not explicitly provide theoretical background for their choice. Regarding empirical evidence in causal inference, Angrist and Frandsen [2019] show that CV Lasso retains more controls, but yields similar estimates for treatment effects; while Wuthrich and Zhu [2019] find a poorer performance of CV Lasso when applying to `PDSLasso`. Thus, it is tempting to employ both versions to `PDSLasso` and `HDCA` for a comparison purpose.

Moving on now to consider only Bayesian approach, we focus on the second step (Bayesian part) of `HDCA`. Antonelli et al. [2019] originally use the spike-and-slab Lasso prior for the slope

parameters of the treatment variable while introducing the information from the previous step about the active set through their new feature ω_j. Generalizing their idea, we adopt generic Stochastic Search Variable Selection (SSVS) priors [George and McCulloch, 1993] and generic Spike and Slab [Kuo and Mallick, 1998]. This framework allows us to incorporate various Bayesian shrinkage priors, reviewed in section 2.2. Detailed algorithms are described below. To sum up, we design eight Bayesian methods in total (collectively referred as `HDCA` methods hereafter), which are associated with eight following priors: SSVS with Normal prior (`SSVSNormal`), SSVS with Student-t prior (`SSVSStudent`), SSVS with Laplace priors (`SSVSLasso1`, `SSVSLasso2`, `SSVSLasso3`), SSVS with Horseshoe prior (`SSVSHorseshoe`), Spike and Slab with Normal prior (`SnSNormal`) and Spike and Slab with Laplace prior (`SnSLasso`).

3.2 Generic Stochastic Search Variable Selection (SSVS) priors - George and McCulloch [1993]

For $j = \overline{1,p}$, this hierarchical model can be summarized as:

$$\boldsymbol{y}_i \mid \boldsymbol{T}_i, \boldsymbol{X}_i, \beta_0, \alpha, \boldsymbol{\beta}, \sigma^2 \sim \text{Normal}\left(\beta_0 + \alpha \boldsymbol{T}_i + \boldsymbol{X}_i \boldsymbol{\beta}, \sigma^2\right) \quad \forall i = 1, \ldots, n \tag{3.1}$$

$$\beta_0, \alpha \mid \sigma^2 \sim \text{Normal}\left(0, \sigma^2 K\right) \tag{3.2}$$

$$\beta_j \mid \gamma_j, \sigma^2, \tau_{0j}^2, \tau_{1j}^2 \sim \gamma_j \text{ Normal}\left(0, \sigma^2 \tau_{1j}^2\right) + (1 - \gamma_j) \text{ Normal}\left(0, \sigma^2 \tau_{0j}^2\right) \tag{3.3}$$

$$\tau_{0j}^2, \tau_{1j}^2 \sim \pi\left(\tau_{0j}^2, \tau_{1j}^2\right) \forall j = 1, \ldots, p \tag{3.4}$$

$$\sigma^2 \mid c, d \sim \text{Inv-Gamma}\ (c, d) \tag{3.5}$$

$$\gamma_j \mid \theta, \omega_j \sim \text{Bernoulli}\ (\theta^{\omega_j}) \tag{3.6}$$

$$\theta \mid a, b \sim \text{Beta}\ (a, b) \tag{3.7}$$

where $\pi\left(\tau_{0j}^2, \tau_{1j}^2\right)$ depends on the specified prior.

Follow Antonelli et al. [2019], we fix $a = 1$ and $b = 0.1p$.

Next, we will set $\omega_j = \delta \in (0, 1)$ if $\boldsymbol{X}_j$ belongs to the active set (in Stage 1 above) and $\omega_j = 1$ otherwise. The probability that $\boldsymbol{X}_j$ belongs to the slab component given $\theta, \tau_{0j}^2, \tau_{1j}^2$ is:

$$p_{\omega_j}^*\left(\beta_j \mid \theta, \tau_{0j}^2, \tau_{1j}^2, \sigma^2\right) = \frac{N\left(\beta_j; 0, \sigma^2 \tau_{1j}^2\right) \theta^{\omega_j}}{N\left(\beta_j; 0, \sigma^2 \tau_{1j}^2\right) \theta^{\omega_j} + N\left(\beta_j; 0, \sigma^2 \tau_{0j}^2\right) (1 - \theta^{\omega_j})} \tag{3.8}$$

We first run the Gibbs sampler with $\omega_j = 1$ for all $j = 1, \ldots, p$. Then, the value of $\delta \in (0, 1)$[1]

[1] δ_j could be allowed to vary for each covariate j

is chosen as the smallest value of ω_j such that $p^*_{\omega_j}\left(0 \mid \theta, \tau^2_{0j}, \tau^2_{1j}, \sigma^2\right)$ is below 10%, where for the unknown parameter values we use posterior means based on the initial Gibbs sampling. A Gibbs sampler is summarized below:

1. Sample $\beta^* = \left(1, \alpha, \{\beta_j\}_{j=\overline{1,p}}\right)$ from the full conditional:

$$\beta^* \mid \bullet \sim \text{ Normal } \left(A^{-1}X^{*'}y^*, A^{-1}\right) \tag{3.9}$$

   where $A = X^{*'}X^* + D^{-1}$ with D is a diagonal matrix with diagonal $\left(K, K, \left\{\gamma_j\tau^2_{1j} + (1-\gamma_j)\,\tau^2_{0j}\right\}_{j=\overline{1,}}\right.$

2. Sample σ^2 from the full conditional:

$$\sigma^2 \mid \bullet \sim \text{ Inv-Gamma } \left(c + \frac{n}{2} + \frac{p}{2}, d + \frac{1}{2}\left\|y - \beta_0 - \alpha T - X\beta\right\| + \frac{1}{2}\beta^{*\prime}D^{-1}\beta^*\right) \tag{3.10}$$

3. Sample γ_j from Bernoulli distribution with the mean parameter:

$$\frac{N\left(\beta_j; 0, \sigma^2\tau^2_{1j}\right)\theta^{\omega_j}}{N\left(\beta_j; 0, \sigma^2\tau^2_{1j}\right)\theta^{\omega_j} + N\left(\beta_j; 0, \sigma^2\tau^2_{0j}\right)\left(1-\theta^{\omega_j}\right)} \tag{3.11}$$

4. Sample θ based on a Metropolis-Hastings algorithm:

$$p(\theta \mid \bullet) \propto \theta^{a+\sum_{j=1}^{p}\omega_j\gamma_j}(1-\theta)^b \prod_{j=1}^{p}\left(1-\theta^{\omega_j}\right)^{(1-\gamma_j)} \tag{3.12}$$

5. Update τ^2_{0j}, τ^2_{1j}

Denote $\boldsymbol{Q}$ the diagonal matrix with diagonal elements $\left\{(1-\gamma_j)\,\tau^2_{0j} + \gamma_j\tau^2_{1j}\right\}_{j=1}^{p}$, (3.3) can be concisely written as

$$\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{Q} \sim N_p\left(\mathbf{0}, \sigma^2\boldsymbol{Q}\right) \tag{3.13}$$

Also note that

$$\beta_j \mid \tau^2_{0j}, \sigma^2, \gamma_j = 0 \sim \text{ Normal } \left(0, \sigma^2\tau^2_{0j}\right)$$
$$\beta_j \mid \tau^2_{1j}, \sigma^2, \gamma_j = 1 \sim \text{ Normal } \left(0, \sigma^2\tau^2_{1j}\right)$$

thus by assigning appropriate priors (specific values/ hierarchical priors) $\pi\left(\tau^2_{0j}, \tau^2_{1j}\right)$ for τ^2_{0j} and τ^2_{1j} corresponding to spike and slab components, we could obtain particular regularizers with desirable properties.

QUESTION: Should we decide separable or non-separable priors for τ^2_{0j} and τ^2_{1j}? Which parameter should be fixed - very small τ^2_{0j}, very large τ^2_{1j}, their ratio or their hyper-parameters (e.g.)? Could we mix two different distribution for spike and slab components? Does advice from Chipman et al. [2001] matter, i.e. control $\tau^2_{1j}/\tau^2_{0j} \leq 10000$ regarding thresholds for selection?

3.2.1 SSVS with Normal prior

We consider Normal priors proposed by Narisetty and He [2014] on the spike and slab components. Specifically, the authors fix the value of the prior variance parameters as:

$$\tau_0^2 = \frac{\hat{\sigma}^2}{10n}$$
$$\tau_1^2 = \hat{\sigma}^2 \max\left(\frac{p^{2.1}}{100n}, \log(n)\right)$$

where $\hat{\sigma}^2$ is the sample variance of y_i. The prior inclusion probability θ is chosen so that $\Pr\left(\sum_{j=1}^{p} \gamma_j > K\right) = 0.1$ for $K = \max(10, \log(n))$. The values of τ_0^2 and τ_1^2 are constant across j, so Q is a diagonal matrix with diagonal elements $\{(1-\gamma_j)\,\tau_0^2 + \gamma_j \tau_1^2\}_{j=1}^{p}$.

3.2.2 SSVS with Student-t prior

We assume Student-t priors introduced by Armagan and Zaretzki [2010] on the spike and slab components, where

$$\tau_{0j}^2 = 0.001 \times \tau_{1j}^2$$
$$\tau_{1j}^2 \sim \text{Inv} - \text{Gamma}(\eta, \mu)$$

The conditional posterior of the hyperparameter is:

$$\tau_{1j}^2 \mid \bullet \sim \text{Inv} - \text{Gamma}\left(\eta + 1/2, \beta_j^2/2 + \mu\right)$$

Note: Armagan and Zaretzki [2010] suggest that the inverse scale parameter μ should be fix to be very small, while their experiments recommend values for the shape parameter (larger value encourages further shrinkage) $\eta \in \{3, 4, 5\}$. We fix $\mu = 0.01$, $\eta = 1$.

3.2.3 SSVS with Laplace prior

We consider Laplace prior proposed by Park and Casella [2008]) in three different ways:

- **SSVS Lasso 1**: Apply Laplace priors on both spike and slab components interdependently with a fixed ratio

$$\tau_{0j}^2 = 0.001 \times \tau_{1j}^2$$
$$\tau_{1j}^2 \sim \frac{\lambda_1^2}{2} e^{-\lambda_1^2 \tau_{1j}^2/2} d\tau_{1j}^2$$
$$\lambda_1^2 \sim \left(\lambda_1^2\right)^{r-1} e^{-\delta\lambda_1^2} d\lambda_1^2$$

The conditional posteriors of the hyperparameters are of the form

$$\begin{aligned}\frac{1}{\tau_{1j}^2} \mid \bullet &\sim \text{ Gaussian } \left(\sqrt{\lambda_1^2\sigma^2/\beta_j^2}, \lambda_1^2\right) \\ \lambda_1^2 \mid \bullet &\sim \text{ Gamma } \left(\sum_{j=1}^{p}\gamma_j + r, \sum_{j=1}^{p}\tau_{1j}^2\gamma_j/2 + \delta\right)\end{aligned}$$

- **SSVS Lasso 2**: Apply Laplace priors on only slab component, fix small τ_{0j}^2 to obtain normal/point-mass at 0

$$\begin{aligned}\tau_{0j}^2 &= 0.001 \\ \tau_{1j}^2 &\sim \frac{\lambda_1^2}{2}e^{-\lambda_1^2\tau_{1j}^2/2}d\tau_{1j}^2 \\ \lambda_1^2 &\sim \left(\lambda_1^2\right)^{r-1}e^{-\delta\lambda_1^2}d\lambda_1^2\end{aligned}$$

The conditional posteriors of the hyperparameters are of the form

$$\begin{aligned}\frac{1}{\tau_{1j}^2} \mid \bullet &\sim \text{ Gaussian } \left(\sqrt{\lambda_1^2\sigma^2/\beta_j^2}, \lambda_1^2\right) \\ \lambda_1^2 \mid \bullet &\sim \text{ Gamma } \left(\sum_{j=1}^{p}\gamma_j + r, \sum_{j=1}^{p}\tau_{1j}^2\gamma_j/2 + \delta\right)\end{aligned}$$

- **SSVS Lasso 3**: Apply Laplace priors on spike and slab components separately (similar to Antonelli et al. [2019])

$$\begin{aligned}\tau_{0j}^2 &\sim \frac{\lambda_0^2}{2}e^{-\lambda_0^2\tau_{0j}^2/2}d\tau_{0j}^2 \\ \lambda_0^2 &\sim \left(\lambda_0^2\right)^{r-1}e^{-\delta\lambda_0^2}d\lambda_0^2 \\ \tau_{1j}^2 &\sim \frac{\lambda_1^2}{2}e^{-\lambda_1^2\tau_{1j}^2/2}d\tau_{1j}^2 \\ \lambda_1^2 &= 0.1\end{aligned}$$

The conditional posteriors of the hyperparameters are of the form

$$\begin{aligned}\frac{1}{\tau_{1j}^2} \mid \bullet &\sim \text{ Gaussian } \left(\sqrt{\lambda_1^2\sigma^2/\beta_j^2}, \lambda_1^2\right) \\ \lambda_1^2 \mid \bullet &\sim \text{ Gamma } \left(\sum_{j=1}^{p}\gamma_j + r, \sum_{j=1}^{p}\tau_{1j}^2\gamma_j/2 + \delta\right)\end{aligned}$$

Note: In all cases we fix $r = 1$ and $\delta = 3$.

3.2.4 SSVS with Horseshoe prior

We assume Horseshoe prior according to Makalic and Schmidt [2015]) on the slab component, fix small τ_{0j}^2 to obtain normal/point-mass at 0

$$\tau_{0j}^2 = 0.001 \tag{3.14}$$
$$\tau_{1j}^2 = \kappa^2 \lambda_j^2 \tag{3.15}$$
$$\lambda_j^2 \mid \nu_j \sim \text{Inv} - \text{Gamma}\left(1/2, 1/\nu_j\right) \tag{3.16}$$
$$\kappa^2 \mid \xi \sim \text{Inv} - \text{Gamma}(1/2, 1/\xi) \tag{3.17}$$
$$\nu_1, \ldots, \nu_p, \xi \sim \text{Inv} - \text{Gamma}(1/2, 1) \tag{3.18}$$

The conditional posteriors of the hyper-parameters are:

$$\lambda_j^2 \mid \bullet \sim \text{In}\, v - \text{Gamma}\left(1, 1/v_j + \beta_j^2/\left(2\kappa^2\sigma^2\right)\right) \tag{3.19}$$
$$\kappa^2 \mid \bullet \sim \text{Inv} - \text{Gamma}\left(\sum_{j=1}^{p} \gamma_j/2 + 1/2, 1/\xi + \sum_{j=1}^{p} \beta_j^2/\left(2\sigma^2\lambda_j^2\right)\right) \tag{3.20}$$
$$\nu_j \mid \bullet \sim \text{Inv} - \text{Gamma}\left(1, 1 + 1/\lambda_j^2\right) \tag{3.21}$$
$$\xi \mid \bullet \sim \text{Inv} - \text{Gamma}\left(1, 1 + 1/\kappa^2\right) \tag{3.22}$$

Note: We fix $\kappa = \xi = \nu_j = 1$

3.3 Generic Spike and Slab - Kuo and Mallick [1998]

We reparametrize the regression coefficients $\boldsymbol{\beta}$ by using two independent random vectors $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_o)$ and $\boldsymbol{b} = (b_1, \ldots, b_p)$ such that $\beta_j = \gamma_j b_j$. Here, γ_j is an indicator that only takes value 0 or 1. For $j = \overline{1, p}$, this hierarchical model can be summarized as:

$$\boldsymbol{y}_i \mid \boldsymbol{T}_i, \boldsymbol{X}_i, \beta_0, \alpha, \boldsymbol{b}, \sigma^2 \sim \text{ Normal } \left(\beta_0 + \alpha \boldsymbol{T}_i + \boldsymbol{X}_i \boldsymbol{b}, \sigma^2\right) \quad \forall i = 1, \ldots, n \tag{3.23}$$
$$\beta_0, \alpha \mid \sigma^2 \sim \text{ Normal } \left(0, \sigma^2 K\right) \tag{3.24}$$
$$b_j \mid \sigma^2, \tau_j^2 \overset{iid}{\sim} \gamma_j \text{ Normal } \left(0, \sigma^2 \tau_j^2\right) \tag{3.25}$$
$$\tau_j^2 \sim \pi\left(\tau_j^2\right) \forall j = 1, \ldots, p \tag{3.26}$$
$$\sigma^2 \mid c, d \sim \text{ Inv-Gamma } (c, d) \tag{3.27}$$
$$\gamma_j \mid \theta, \omega_j \overset{iid}{\sim} \text{ Bernoulli } (\theta^{\omega_j}) \tag{3.28}$$
$$\theta \mid a, b \sim \text{ Beta } (a, b) \tag{3.29}$$

where $\pi(\tau_j^2)$ depends on the specified prior.

Next, we will set $\omega_j = \delta \in (0,1)$ if $\boldsymbol{X}_j$ belongs to the active set (in Stage 1 above) and $\omega_j = 1$ otherwise. The probability that $\boldsymbol{X}_j$ belongs to the slab component given θ, τ_j^2 is:

$$p^*_{\omega_j}\left(\beta_j \mid \theta, \tau_j^2, \sigma^2\right) = \frac{l_{1j}}{l_{1j} + l_{0j}} \tag{3.30}$$

where $l_{1j} = f(\boldsymbol{y} \mid \gamma_(j), \gamma_j = 1, \boldsymbol{b}, \sigma^2)\theta^{\omega_j}$ and $l_{0j} = f(\boldsymbol{y} \mid \gamma_(j), \gamma_j = 0, \boldsymbol{b}, \sigma^2)(1-\theta^{\omega_j})$.

We first run the Gibbs sampler with $\omega_j = 1$ for all $j = 1, \ldots, p$. Then, the value of $\delta \in (0,1)$ is chosen as the smallest value of ω_j such that $p^*_{\omega_j}\left(0 \mid \theta, \tau_j^2, \sigma^2\right)$ is below 10%, where for the unknown parameter values we use posterior means based on the initial Gibbs sampling. A Gibbs sampler is summarized below:

1. Sample $\boldsymbol{b}^* = \left(1, \alpha, \{b_j\}_{j=\overline{1,p}}\right)$ from the full conditional:

$$\boldsymbol{b}^* \mid \bullet \sim \text{ Normal } \left(A^{-1}X^{*'}y^*, A^{-1}\right) \tag{3.31}$$

   where $A = X^{*'}X^* + D^{-1}$ with D is a diagonal matrix with diagonal $\left(K, K, \left\{\tau_j^2\right\}_{j=\overline{1;p}}\right)$ and $X^* = \left(\mathbf{1}, \boldsymbol{T}, \{\gamma_j \boldsymbol{X}_j\}_{j=\overline{1;p}}\right)$.

2. Sample σ^2 from the full conditional:

$$\sigma^2 \mid \bullet \sim \text{ Inv-Gamma } \left(c + \frac{n}{2} + \frac{p}{2}, d + \frac{1}{2}\|y - \beta_0 - \alpha T - X\beta\| + \frac{1}{2}\beta^{*\prime}D^{-1}\beta^*\right) \tag{3.32}$$

3. Sample γ_j from Bernoulli distribution with the mean parameter:

$$\frac{l_{1j}}{l_{1j} + l_{0j}} \tag{3.33}$$

   where l_{1j} and l_{0j} are defined as in (3.30).

4. Sample θ based on a Metropolis-Hastings algorithm:

$$p(\theta \mid \bullet) \propto \theta^{a + \sum_{j=1}^p \omega_j \gamma_j}(1-\theta)^b \prod_{j=1}^p (1-\theta^{\omega_j})^{(1-\gamma_j)} \tag{3.34}$$

5. Update τ_j^2

Denote $\boldsymbol{Q}$ the diagonal matrix with diagonal elements $\{\tau_j^2\}_{j=1}^p$, (3.25) is equivalent to

$$\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{Q} \sim \text{ Normal } \left(\mathbf{0}, \sigma^2\boldsymbol{Q}\right) \tag{3.35}$$

Similar to the case of SSVS priors, we now consider some appropriate priors for τ_j.

3.3.1 Spike and Slab with Normal prior

We simply fix $\tau_j = 9 \quad \forall j = \overline{1, p}$, thus leading to Normal priors on each b_j.

3.3.2 Spike and Slab with Laplace prior

To obtain Laplace prior for each b_j, we set the following hierarchical prior for τ_j

$$\tau_j^2 \sim \frac{\lambda^2}{2} e^{-\lambda^2 \tau_j^2/2} d\tau_j^2$$
$$\lambda^2 \sim \left(\lambda^2\right)^{r-1} e^{-\delta\lambda^2} d\lambda^2$$

The conditional posteriors of the hyper-parameters are of the form

$$\frac{1}{\tau_j^2} \mid \bullet \sim \text{ Gaussian } \left(\sqrt{\lambda^2\sigma^2/\beta_j^2}, \lambda^2\right)$$
$$\lambda^2 \mid \bullet \sim \text{ Gamma } \left(\sum_{j=1}^{p} \gamma_j + r, \sum_{j=1}^{p} \tau_j^2 \gamma_j/2 + \delta\right)$$

Note: We fix $r = 1$ and $\delta = 3$.

CHAPTER

FOUR

SIMULATION STUDY

4.1 Overview

Aim

This Monte Carlo study aims to assess the finite-sample performance of different regularization-based methods (Frequentist and Bayesian) for inference on the treatment effect in an observational study with high-dimensional controls. We focus on a linear regression model where the treatment variable T_i is taken as exogenous after conditioning control variables (.i.e. Unconfoundenedness holds).

$$y_i = \alpha T_i + \boldsymbol{X}_i \boldsymbol{\beta} + \epsilon_i, \quad \forall i = 1, \ldots, n \tag{4.1}$$

where y_i is the outcome, T_i is the scalar treatment variable, and $\boldsymbol{X_i}$ is a p-dimensional row vector of potential control variables of subject i (p is comparable or even larger than n).

Estimand

Our estimand α is the regression coefficient of the treatment variable T_i, which would represent the average treatment effect in an observational study (under satisfactions of assumptions specified in section 2.1).

Method implementations

We use nine methods designed in section 3 to analyze each simulated dataset. These methods are consist of `PDSLasso`, `SSVSNormal`, `SSVSStudent`, `SSVSLasso1`, `SSVSLasso2`, `SSVSLasso3`, `SSVSHorseshoe`, `SnSNormal` and `SnSLasso`. The first one represents for Frequentist approach, while the others are Bayesian counterparts (we refer as `HDCA` methods).

Data are simulated, and corresponding performance metrics are calculated using `Matlab` [MATLAB, 2020] with `ParallelComputationalToolbox` for random number generation. The simulation results are analyzed using `R` [R Core Team, 2020]. We describe details in the rest of this chapter.

4.2 Data Generating Processes

We first set $(n, p) = (300, 400)$. For simplicity, we focus on continuous treatments. The model used to generate data is as follows:

$$y_i = \beta_0 + \alpha T_i + x_i'\beta + \epsilon_i \tag{4.2}$$

$$T_i = x_i'\psi + \nu_i \tag{4.3}$$

where $\beta_0 = 0$ and $\alpha = 1$. Set $\beta_j = c_y\bar{\beta}_j$ and $\psi_j = c_t\bar{\psi}_j$ and choose the constants c_y and c_t in order to achieve desired level of signal-to-noise ratios.

Relevant variables x_{ij} are categorized into four types:

(a) *strong confounders* that are strongly correlated with both T_i and y_i:
set $\bar{\psi}_j = 1$ if j is odd (-1 if even) and $\bar{\beta}_j = 1$ if j is odd (-1 if even),

(b) *weak confounders* that are strongly correlated with T_i but weakly with y_i:
set $\bar{\psi}_j = 1$ if j is odd (-1 if even) and $\bar{\beta}_j = 0.3$ if j is odd (-0.3 if even),

(c) *instruments* that are strongly correlated with T_i but uncorrelated with y_i:
set $\bar{\psi}_j = 1$ if j is odd (-1 if even) and $\bar{\beta}_j = 0$, and

(d) *strong predictors* that are strongly correlated with y_i but uncorrelated with T_i:
set $\bar{\psi}_j = 0$ and $\bar{\beta}_j = 1$ if j is odd (-1 if even).

For irrelevant variables (noises) that do not belong to any of the four groups above, we set $\bar{\psi}_j = 0$ and $\bar{\beta}_j \overset{iid}{\sim} N(0, 0.1^2)$.

We vary the following factors in the simulation study:

4.2.1 Correlation among Covariates

The covariates are either uncorrelated or correlated:

(a) Uncorrelated covariates: $x_{ij} \overset{iid}{\sim} N(0, \sigma_x^2)$ for all j and i.

(b) Correlated covariates: $x_i \overset{iid}{\sim} N(0, \Sigma)$ for all i with $\Sigma_{kj} = \rho^{|j-k|}$, where ρ determines how strongly the covariates are correlated.

We fix $\sigma_x^2 = 1$ and $\rho = 0.9$.

4.2.2 Sparsity

Given $q \in (0, 1)$, let $[100 \times q \times p]$ be the percentage of the relevant covariates which fall into the four types defined above.

4.2.3 Error Variance

The error variances are either homoskedastic or heteroskedastic:

(a) For homoskedastic errors, let $v_i \sim$ i.i.d. $N(0, \sigma_t^2)$ and $\epsilon_i \sim$ i.i.d. $N\left(0, \sigma_y^2\right)$ with $\sigma_y^2 = \sigma_t^2 = 1$.

(b) For heteroskedastic errors, let $v_i \sim$ i.i.d. $N(0, \sigma_t^2(x_i))$ and $\epsilon_i \sim N\left(0, \sigma_y^2(T_i, x_i)\right)$ where $\sigma_t(x_i) = \sqrt{\frac{\left(1+x_i'\theta\right)^2}{\mathbb{E}_n\left(1+x_i'\theta\right)^2}}$ and $\sigma_y(T_i, x_i) = \sqrt{\frac{\left(1+\beta_0+\alpha T_i+x_i'\theta\right)^2}{\mathbb{E}_n\left(1+\beta_0+\alpha T_i+x_i'\theta\right)^2}}$ ($\mathbb{E}_n$ denotes the empirical expectation).

4.2.4 Signal-to-noise Ratio

In a general linear regression $y_i = x_i'\beta + \epsilon_i$, the signal-to-noise ratio (SNR) is defined as:

$$SNR = \frac{\left\|\Sigma_X^{1/2}\beta\right\|^2}{\sigma^2} \tag{4.4}$$

where σ^2 is the error variance and Σ_X is a $p \times p$ covariance matrix of x_i. Thus, $\left\|\Sigma_X^{1/2}\beta\right\|^2 = \beta'\Sigma_X\beta$ measures the overall signal strength, where $\|\cdot\|$ is the ℓ^2 -norm.

A related quantity is R^2_{pop} , the population value of R^2, defined as $\frac{SNR}{1+SNR}$. If $\beta_j = c\tilde{\beta}_j$ for $j = 1, \ldots, p$ for some scalar c, then $c = \sqrt{\frac{\sigma^2}{\beta'\Sigma_X\beta}\frac{R^2_{\text{eop}}}{1-R^2_{\text{pop}}}}$. Hence, c could be chosen to achieve a desired value of R^2_{pop} or SNR.

Now, let R_t^2 and R_y^2 be some pre-specified population values of R^2 for the treatment and the outcome equations, respectively. We follow Belloni et al. [2014b] to compute the constants c_t and c_y. For homeskedastic case, these constants are computed according to the following steps:

1. For given (n, p, Σ_X), generate X and for a given q, define $\bar{\psi}_j$ and $\bar{\beta}_j$ for $j = 1, \ldots, p$.
2. Given R_t^2, compute $c_t = \sqrt{\frac{\sigma_t^2}{\psi' \Sigma_X} \frac{R_t^2}{1-R_t^2}}$.
3. Given R_y^2, compute $c_y = \sqrt{\frac{\sigma_y^2}{\beta' \Sigma_X \beta} \frac{R_y^2}{1-R_y^2}}$

For heteroskedastic case, we also use the above formulas for computing the constants as if v_i and ϵ_i are homoskedastic following Belloni et al. [2014b].

4.2.5 Summary

Values of $\left(q, R_t^2, R_y^2\right)$ are chosen to define various data-generating processes (DGPs). Consider the following four scenarios:

(a) uncorrelated predictors AND homoskedasticity

(b) correlated predictors AND homoskedasticity

(c) uncorrelated predictors AND heteroskedasticity

(d) correlated predictors AND heteroskedasticity

For each scenario, consider $q \in \{0.04, 0.4\}$ and $\left(R_t^2, R_y^2\right) \in \{(0.2, 0.2), (0.2, 0.8), (0.8, 0.2), (0.8, 0.8)\}$ which gives us 8 combinations. Hence, there are 32 different DGPs in total. For each design we run $N_s = 48$ Monte Carlo simulations.

4.3 Performance Metrics

Let $\hat{\alpha}$ be a point estimator of a given method. For Bayesian methods, the post median are considered. Let N_s be the number of repeated experiments. We consider five major performance measures as below:

1. Mean-absolute-error (MAE) : $MAE = \frac{1}{N_s} \sum_{s=1}^{N_s} |\hat{\alpha} - \alpha|$ as a measure of bias
2. Root-mean-squared-error (RMSE) : $RMSE = \sqrt{\frac{1}{N_s} \sum_{s=1}^{N_s} (\hat{\alpha} - \alpha)^2}$ as a measure of efficiency

3. Empirical coverage: coverage rates of 95% confidence intervals (Frequentist approach) or posterior credible intervals (Bayesian approach), i.e. the percentage of the time that the interval covers the true parameter. Also, the average interval lengths are reported.

4. Inclusion probability for 5 types of variable x_j for $j = 1, \ldots, p$ in the final model

   (a) For `PDSLasso`, we report $\frac{1}{N_s} \sum_{s=1}^{N_s} 1\left(x_j \in \hat{I}_1 \cup \hat{I}_2\right)$

   (b) For `HDCA` methods, we use (mean and median) posterior probability that $\gamma_j = 1$ and average it over N_s experiments

   (c) 5 types of variable x_j: Strong Confounders, Weak Confounders, Instrumental Variables, Strong Predictors, Irrelevant Variables (Noises - $j = [qp] + 1, \ldots, p$).

4.4 Results and Discussion

4.4.1 Initial Results

The results associated with 32 different DGPs are presented in tables in `Supplementary materials`. We also visualize the calculated values in 32 ordered scenarios for each performance criteria above to identify general patterns. Finally, we summarize initial observations as below:

First of all, we take into account the effect of using Iterated Lasso or using Cross-Validated Lasso in `PDSLasso` and `HDCA` methods. As it can be seen from all tables, `PDSLasso` with Iterated Lasso dominates the counterpart with Cross-Validated Lasso in terms of all metrics: smaller absolute bias, smaller RMSE, higher coverage rate and shorter average interval length. This differential is highly significant, thus, suggests the importance of a theory-driven approach for selecting the penalty level proposed by Belloni et al. [2014b]. However, the case of `HDCAs` is not always clear like that. Two approaches achieve quite similar bias and RMSE. The problem is that several illogical results related to coverage rate and average interval length exist in the case of Cross-Validated Lasso. Additionally, employing Cross-Validated Lasso in the first step of `HDCAs` is much slower than using Iterated Lasso, although the former is originally used in Antonelli et al. [2019]. Therefore, we will focus on `PDSLasso` and `HDCA` methods with Iterated Lasso hereafter.

Secondly, the overall graphs for 32 DGPs show that heteroskedasticity almost does not affect the criteria of interest and the relative performance amongst methods.

Thirdly, a comparison of different versions of spike-and-slab Lasso (i.e. `SSVSLasso1`, `SSVSLasso2`, `SSVSLasso3` and `SnSLasso` can be made straightforward from figure 4.1. We

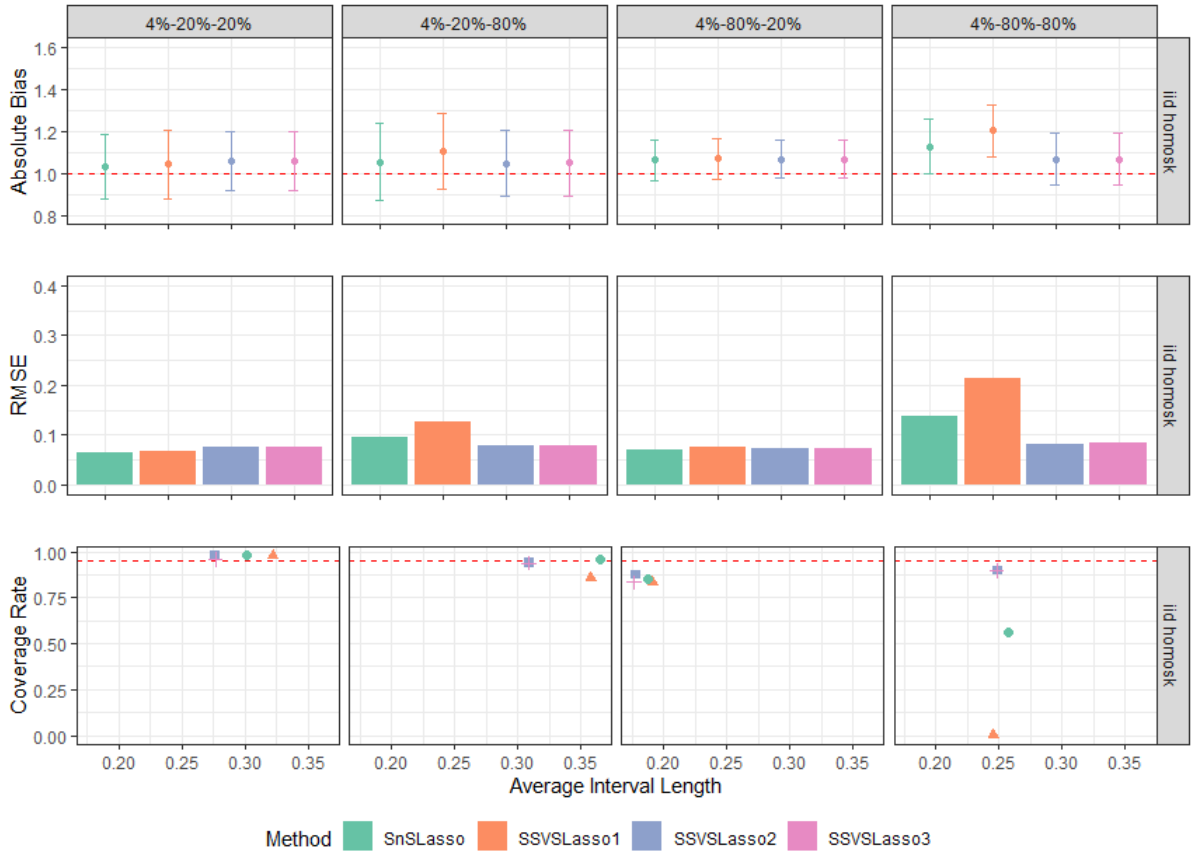

Figure 4.1: Performance of spike-and-slab Lasso priors

consider 4 scenarios corresponding to $(R_t^2, R_y^2) \in \{(20\%, 20\%), (20\%, 80\%), (80\%, 20\%), (80\%, 80\%)\}$ while holding a independent and homoskedastic design with the sparsity level $q = 4\%$. It is evident that `SSVSLasso2` and `SSVSLasso3` act almost identically and outperform `SSVSLasso1` and `SnSLasso` in all considered aspects (bias, RMSE and coverage rate). `SSVSLasso1` performs poorly, especially in the design with high SNR. A further inspection on tables of results suggests that it tends to include all of the variables into the structural-form equation in the second step, thus being closed to a naive Bayesian approach mentioned in section 2.3 and more likely to cause severe bias. In fact, `SSVSLasso1` prior is a mixture of Laplace distribution in both spike and slab components with interdependent penalty parameters, which seems less theoretically supported than the others. `SSVSLasso2` involves a Laplace prior applied on only slab component and a fixed small value τ_{0j}^2 to obtain approximately point-mass at 0, while `SSVSLasso3` entails Laplace priors applied on spike and slab components separately with a fixed λ_1 (similar to Ročková and George [2018]'s approach). These two versions may be linked in the sense that: "Increasing λ_0, while λ_1 is held fixed, corresponds to the deployment of a sequence of SSL priors where the spike concentrates increasingly more mass around zero, approximating the point mass spike $\phi_0(\beta) = \mathrm{I}(\beta = 0)$. Thus, the

Spike-and-Slab LASSO can be seen as a fast computable approximation to mode detection under the spike-and-slab mixture of a point mass at 0 and a diffuse heavy-tailed slab, which is often considered as the Bayesian ideal [Castillo and van der Vaart, 2012]"; hence, more suitable to achieve sparsity. `SnSLasso` is a special case of [Xu and Ghosh, 2015](?). This phenomenon is stable across 32 different scenarios so that we can select `SSVSLasso3` as the "winner" of this group for the sake of simplicity.

Taken together, it is sufficient for us to restrict our attention to 16 DGPs (homoskedastic designs) and a smaller set of methods without losing some key observations.

4.4.2 Key Observations

We proceed analyzing results by performance metrics for six methods: `PD Lasso` with Iterated Lasso for selection in each steps, and `HDCA` methods associated to five spike-and-slab priors comprising `SSVSNormal`, `SSVSStudent`, `SSVSLasso3`, `SSVSHorseshoe` and `SnSNormal` (with Iterated Lasso in the first step). In each of following graphs, the horizontal strips indicate a triple sparsity level - SNR for the treatment equation - SNR for the structural equation, e.g. `4%-20%-20%` means $q = 4\% - R_t^2 = 20\% - R_y^2 = 20\%$. The vertical strips indicate whether the scenario is independent (`iid`) or correlated (`corr`) design (given homoskedastic (`homosk`) design as our concern). Both high-sparsity setting ($q = 4\%$) and low-sparsity setting ($q = 40\%$) are illustrated. For each performance criterion, an axis limit is keep unchanged throughout different plots to facilitate overall comparisons.

High-sparsity designs

Figure 4.2 describes the bias of estimation results provided by six methods. Each dot with the error bar reflects an estimate accompanied by its 95% confident (credible) interval, while the red dashed line presents the true value of treatment effect $\tau = 1$. `PDSLasso` provides smallest bias in cases of independent design with high SNR in the first stage ($R_t^2 = 80\%$) and correlated design with low SNR in the first stage ($R_t^2 = 20\%$), as well as outperforms other methods across different scenarios. Only `SSVSHorseshoe` and `SnSNormal` can produces the comparable bias with `PDSLasso` when SNR are low in both equations ($R_t^2 = R_t^2 = 20\%$). Higher SNR in the second stage ($R_y^2 = 80\%$) entails larger standard errors of all methods. In terms of standard errors, `SSVSHorseshoe`, `SSVSNormal` and `SSVSLasso3` are almost better than `PDSLasso` in the considered DGPs.

The above evaluations are also confirmed in figure 4.3, which illustrates the RMSE of

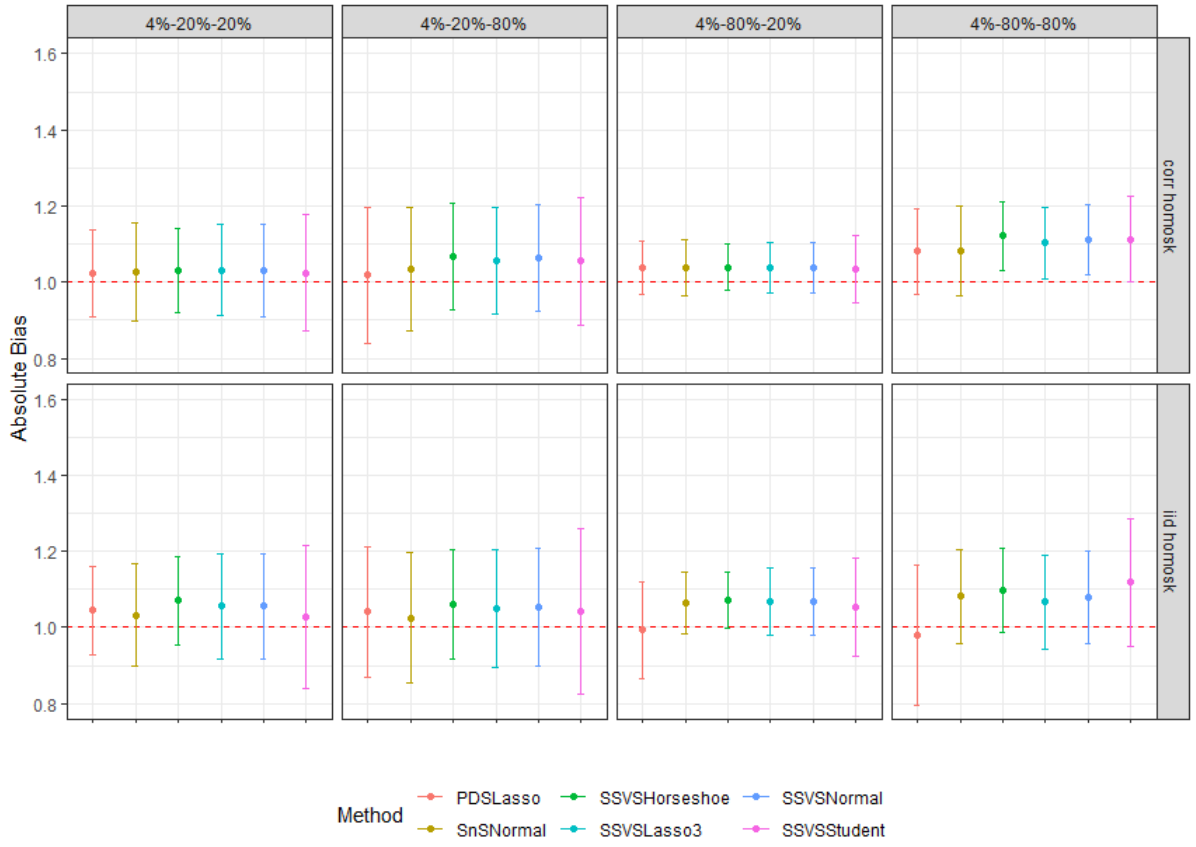

Figure 4.2: The absolute bias of six methods in homoskedastic and high-sparsity designs

estimation results provided by six methods. Higher SNR in the structural model ($R_y^2 = 80\%$) tends to enlarge the RMSE of all methods. While there is no winner, `SSVSHorseshoe` once again performs most poorly compared to others.

Figure 4.4 illustrates the trade-off between the coverage rate and the average interval length. `SSVSStudent` has the highest coverage rate in most of scenarios, except for the case of high SRN ($R_t^2 = R_y^2 = 80\%$) where it is overcame by `PDSLasso`. Nonetheless, `SSVSStudent` also involves a larger average interval length length compared to other methods. By contrast, `SSVSNormal` and `SSVSLasso3` perform similarly (`SSVSLasso3` is slightly better) when producing smaller average interval length at the cost of lower coverage rate. `SSVSHorseshoe` is a dismal one which never reachs the nominal coverage rate. PDSLasso has the most stable performance. In its worst situation, i.e. the correlated design with high SNR in the first stage ($R_t^2 = 80\%$), its coverage rate is still above 65%.

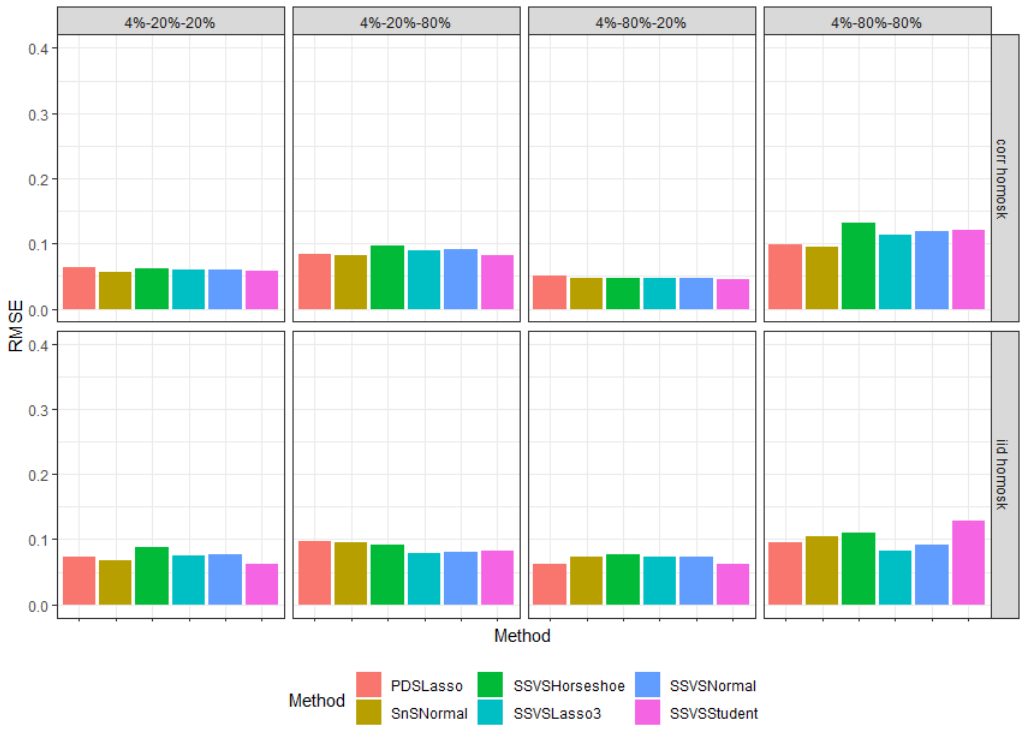

Figure 4.3: RMSE of six methods in homoskedastic and high-sparsity designs

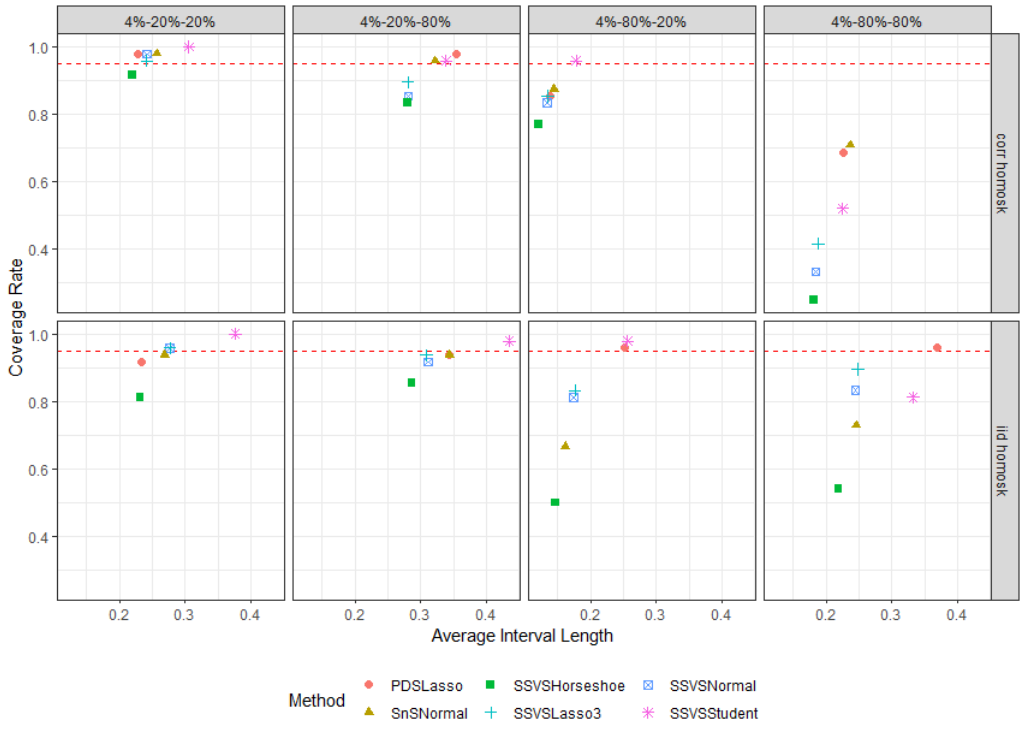

Figure 4.4: Coverage rate and average confidence interval length of six methods in homoskedastic and high-sparsity designs

Low-sparsity designs

When the design deviates from high-sparsity, the performance of all methods deteriorates. We now consider low-sparsity scenarios when $q = 40\%$; the absolute bias, the coverage rate, and the RMSE are summarized in figures 4.5, 4.6 and 4.7, respectively.

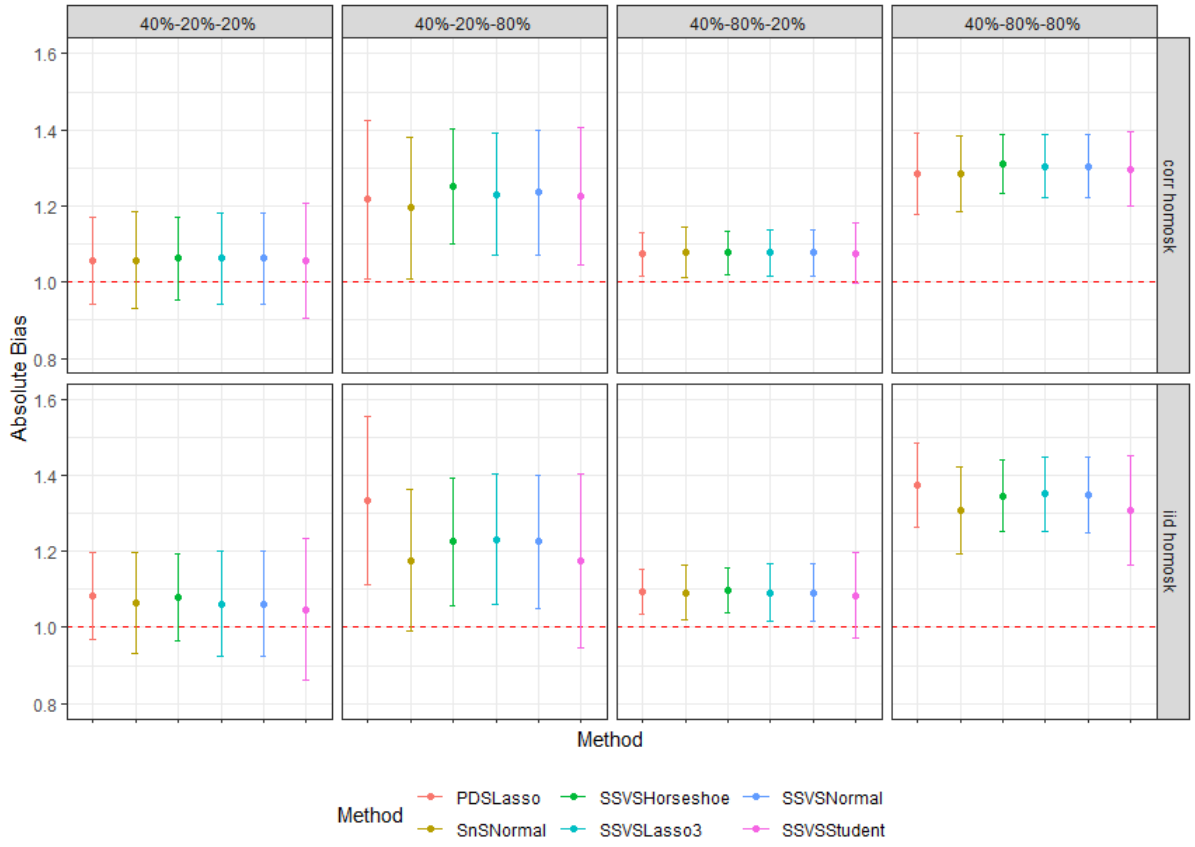

Figure 4.5: The absolute bias of six methods in homoskedastic and low-sparsity designs

Specifically, all methods suffer from the lack of sparsity, but this effect is still moderate in the case of low SRNs in both treatment equation and structural equation ($R_t^2 = R_y^2 = 20\%$). Let high-sparsity settings be the benchmarks; the most challenging situation comes to all methods when $R_t^2 = R_y^2 = 80\%$: absolute bias amplifies, coverage rate drops to 0, and RMSE rises significantly. This phenomenon is expected by Belloni et al. [2014b] since `PDSLasso` requires a certain level of sparsity to become uniformly valid. In fact, it can be seen from figure 6 that `PDSLasso` sometimes is dominated by all `HDCA` methods in terms of coverage rate or average confidence interval length. However, when low sparsity is coupled with high SRN, none of the methods could achieve the nominal coverage rate. We will discuss this in more detail below.

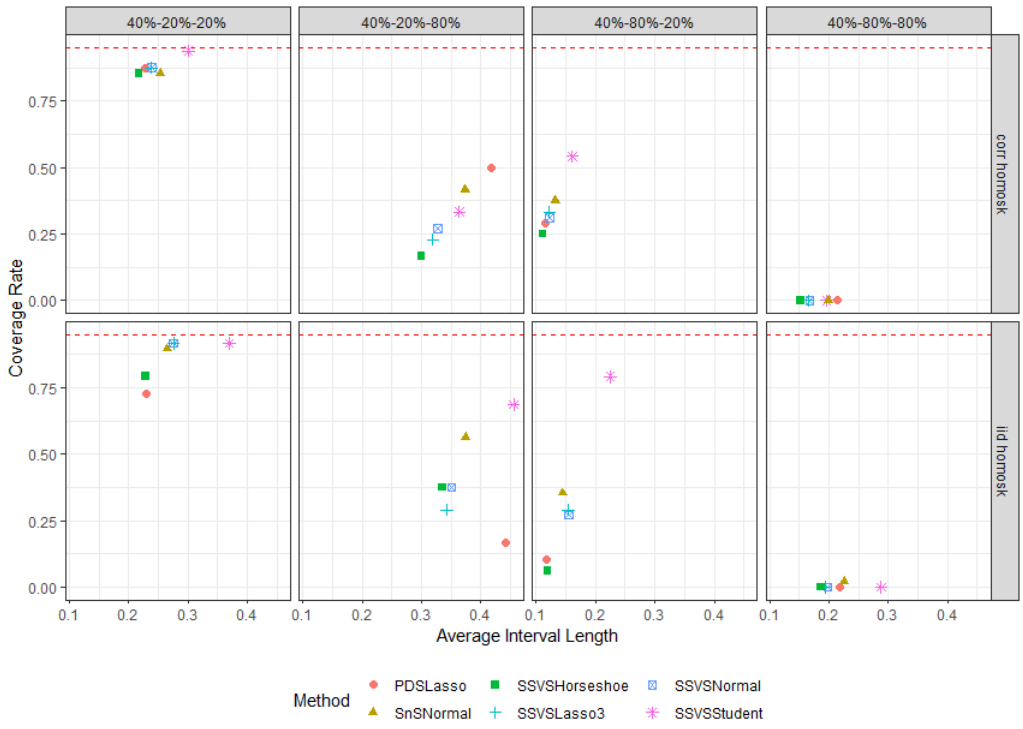

Figure 4.6: Coverage rate and average confidence interval length of six methods in homoskedastic and low-sparsity designs

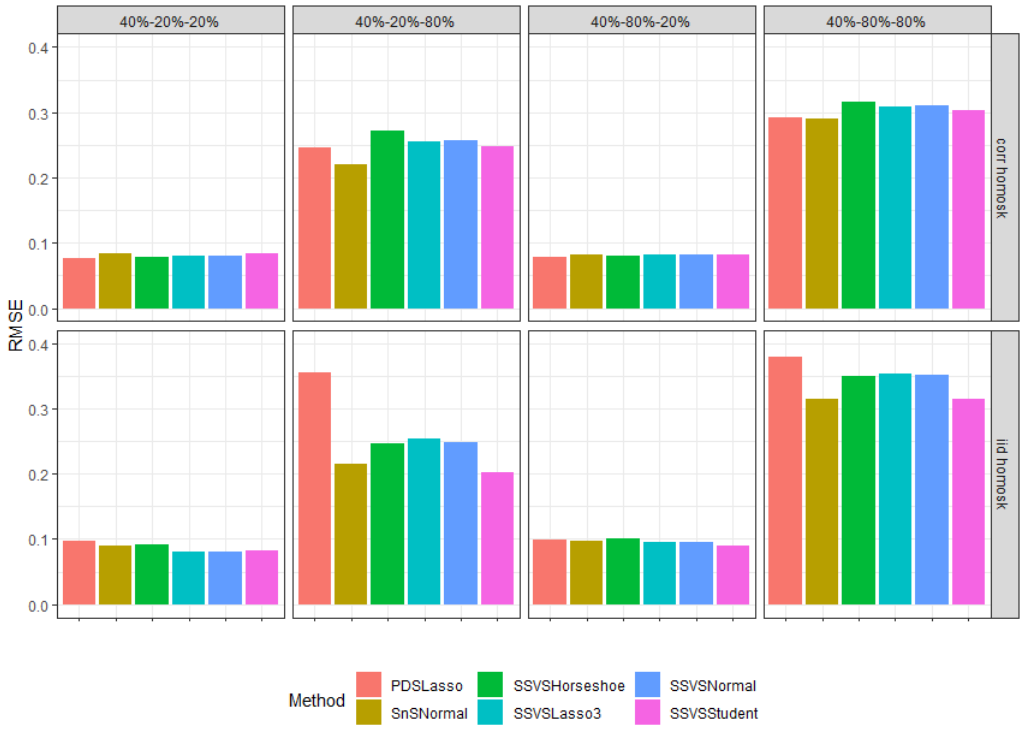

Figure 4.7: RMSE of six methods in homoskedastic and low-sparsity designs

Inclusion Probability

Looking at the inclusion probability can give us some insights into the cause of different performance among the above methods. For `PDSLasso`, the inclusion probability of each variable is the percentage of the time that this variable is included in the final structural model, i.e. it is selected from either treatment equation or outcome equation. For `HDCA` methods, we consider the posterior inclusion probability of each variable in the final model, i.e. the probability this variable enters the slab.

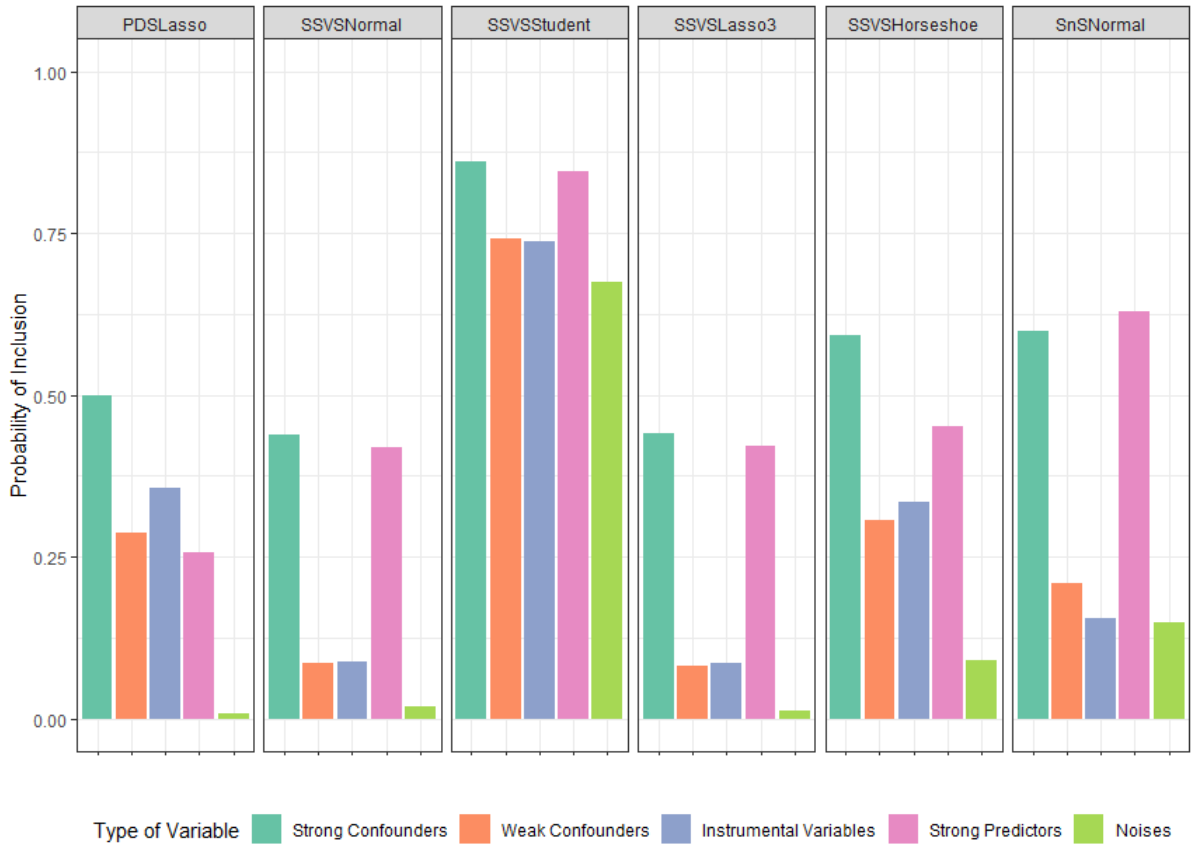

Figure 4.8: Average inclusion probability of different types of variables in high-sparsity designs

Figure 4.8 indicates the average inclusion probability of different types of variables across 16 low-sparsity designs. A confounder is associated with both the treatment variable of interest and the outcome; then, the treatment coefficient can change considerably when this type of variable is added to the model. As shown in figure 4.8, strong confounders have the highest chance of being selected compared to others, whilst weak confounders are chosen with less frequency. A strong predictor is unrelated to the treatment but is highly associated with the outcome variable. Adding this type of variable to the model does not

systematically change the regression coefficient of the treatment variable, but sometimes adding this variable will absorb some of the "noise" (i.e. residual variance) in the outcome, resulting in more precise (i.e. lower standard error) estimation of the treatment coefficient. From figure 4.8, strong predictors often rank second in terms of being selected. It can be explained by the fact that the final step of every `HDCA` method is estimating the structural equation. An instrumental variable is related to the treatment variable but unrelated to the outcome. Adding this type of variable is usually a bad idea: it doesn't improve anything about the model, but it may result in less precise (i.e. higher standard error) estimation of the treatment coefficient as it steals variance from the treatment variable itself. As shown in figure 4.8, instrumental variables have a similar inclusion probability to weak confounders in most methods. These variables are initially selected along with confounders in the first step with Lasso in all methods, and some of them remain after estimation in the final structural equation.

`SSVSStudent` performs very well in terms of selecting confounders, but at the same time it incorporates many non-confounding variables into slab.[1]. This explains why `SSVSStudent` could obtain good coverage rate yet entailing a high variance/average interval length.[2] By contrast, `SSVSLasso3`, By contrast, `SSVSLasso3` and `SSVSNormal` (perform similarly) are good at excluding non-confounding variables, but simultaneously including less confounders. This implies their small average interval length as a trade-off of low coverage rate. `SSVSHorseshoe` and `SnSNormal` are somewhere in between. The stability (and superiority, sometimes) of `PDSLasso` could be also justified by the inclusion probability. Post-double-selection procedure helps `PDSLasso` to outperform `SSVSLasso3` and `SSVSNormal` with respects to choosing both strong and weak confounders, while dominating `SSVSStudent` in terms of precluding non-confounding variables.

Deviating from the high-sparsity context, figure 4.9 demonstrates the average inclusion probability of different variable types across 16 low-sparsity designs. Unsurprisingly, the inclusion probability of all methods drops considerably (especially in the case of `PDSLasso`). Also, strong confounders are no longer the dominants among different types of variables in the model. As a consequence, the performance of all methods deteriorates.

[1]This phenomenon is similar to `SSVSLasso1`. It seems to be a result of putting interdependent hyperparameters, $\tau_{0j}^2 = 0.001 \times \tau_{1j}^2$.

[2]The Student-t prior is unbounded near $\kappa = 0$, reflecting its heavy tails, i.e. it allows strong signals to remain large. But it is bounded near $\kappa = 1$, limiting it ability to shrink noise components back to zero.

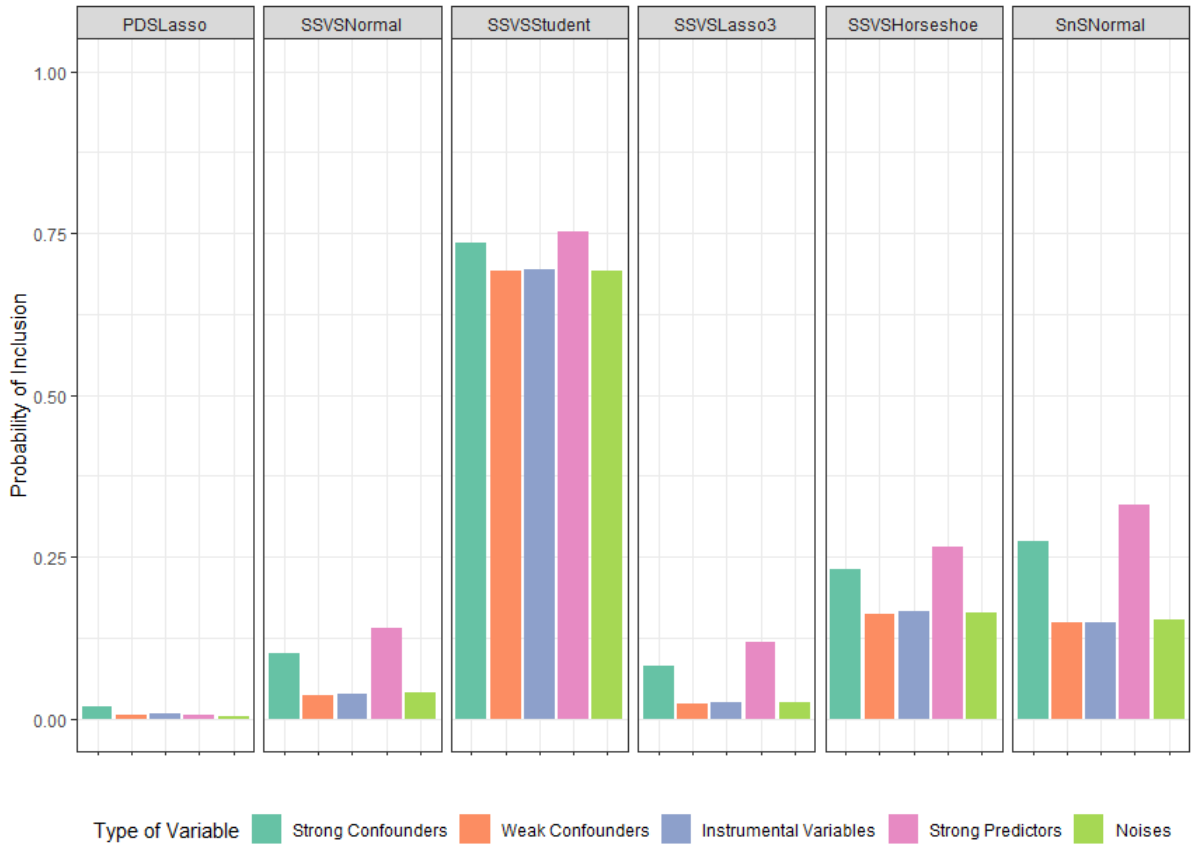

Figure 4.9: Average inclusion probability of different types of variables in low-sparsity designs

CHAPTER

FIVE

EMPIRICAL ILLUSTRATION

5.1 Overview

This section aims to illustrate an economic empirical example when we are interested in inference on treatment effects from an observational study using a linear regression model under the Unconfoundedness assumption. In the previous section, we have conducted a simulation study to evaluate the regularization-based methods for causal inference from both Frequentist and Bayesian perspectives. These methods are now considered in parallel with a traditional ad hoc approach in the revisited observational study: *Media and Political Persuasion: Evidence from Russia* [Enikolopov et al., 2011]. This paper is published in American Economic Review, and supplementary materials (data and `STATA` codes) for replication purposes could be found online. In the following, we first review Enikolopov et al. [2011]'s study of the impact of mass media on political outcomes briefly and then illustrate the use of the considered methods. The new analysis is implemented using `Stata` [StataCorp, 2021] and `Matlab` [MATLAB, 2020].

5.2 Description of Original Analysis

Summary

Enikolopov et al. [2011] investigate the causal effect of the only independent national TV channel, NTV, on voting behaviour during the Russian 1999 parliamentary elections. Despite the overall success of the newly created pro-government Unity Party (with Putin V.) in

the 1999 election, the success was far from uniform across the country. These authors conjecture that massive support from state-owned TV channels caused the rise of Unity. To a large extent, these differences might be explained by the variation in voters' access to an independent media outlet in different parts of the country. [1] Albeit they take account of both aggregate-level effects and individual-level effects, we only focus on the former: the impact of *NTV availability* on *the official electoral results* throughout subregions.

Hypotheses

The 1999 election campaign was the only time in Russia's political history when different TV channels had different political orientations: the pro-government Unity Party was opposed by NTV ("Independent TV") relative to the two main state-controlled TV channels, while the centrist opposition OVR party and liberal opposition SPS and Yabloko parties were supported by NTV relative to the two state-controlled TV channels. At the same time, approximately three-fourths of Russia's population had access to NTV and can watch two-sided political news, while one-fourth of voters located in parts of the country where NTV was not accessible were exposed to only one-sided media coverage of the election campaign. Therefore, three main hypotheses are proposed as below:

- There is a *significant positive effect* of the availability of NTV on voting for all parties supported by NTV (centrist opposition OVR and liberal opposition Yabloko and SPS).
- There is a *significant negative effect* of NTV availability on the vote for pro-government Unity, which was criticized by NTV and praised by the other national TV channels.
- The prediction about the effect of NTV on voter turnout is ambiguous.

Identification strategy and Data

There is a fundamental problem in estimating the causal impact of NTV availability on aggregate voting outcomes: NTV availability during this period were not randomly assigned among subregions. In fact, certain factors may be associated with both state-level NTV availability and state-level voting outcomes. Failing to control for these factors will then lead to omitted variables bias. To address these potential confounding factors, Enikolopov

[1] Indeed, if the governing party controls all major media sources, access to an alternative source of information can be important in helping people to make informed choices.

et al. [2011] introduce the following baseline cross-sectional specification (a linear constant-effect causal model):

$$\text{vote}^j_{s,1999} = \beta_0 + \beta_1 \text{NTV}_{s,1999} + \eta^j_s \tag{5.1}$$

$$= \beta_0 + \beta_1 \text{NTV}_{s,1999} + \beta_2' \boldsymbol{X}_{s,1995} + \beta_3' \boldsymbol{E}_{s,1998} + \delta_r + \varepsilon^j_s \tag{5.2}$$

where:

- $\text{vote}^j_{s,1999}$ is the percent of votes for party j in subregion s at the 1999 Duma elections
- $\text{NTV}_{s,1999}$ is the predicted NTV availability in subregion s in 1999 - based on data on the location and power of NTV transmitters in the respective subregion with a probit regression:

$$\mathbf{Pr}\left\{\text{NTV\_available}_i = 1\right\} = \Phi\left(\underset{[0.00069]}{0.008} \times \text{Signal\_strength}_i + \underset{[0.039]}{0.654}\right)$$

- $\boldsymbol{X}_{s,1995}$ is a vector of electoral outcomes of subregion s in 1995 elections
- $\boldsymbol{E}_{s,1998}$ is a set of socioeconomic characteristics of subregion s measured in 1998
- δ_r are region fixed effects
- Standard errors ε^j_s are adjusted to allow for clusters within each region.

It bears emphasizing that Unconfoundedness is the key assumption here so that β_1 has a causal interpretation: the observable characteristics controlled in (5.2) are the only reason why η^j_s and $\text{NTV}_{s,1999}$ are correlated. Intuitively, "voters in the locations with and without access to NTV are similar in all unobserved characteristics that may drive their voting behaviour once we control for observable differences between these locations."[2] Typically, we cannot verify this assumption exactly. Enikolopov et al. [2011] employ a range of sensitivity analysis techniques to support this argument, such as a placebo experiment [3] in which they estimate the effect of the main explanatory variable, i.e., NTV availability in 1999, on the voting behaviour in 1995 (instead of 1999 as in their baseline specification). Despite these

[2] the availability of NTV was idiosyncratic conditional on observables, i.e., there are no unobserved characteristics of subregions correlated with NTV availability that could drive the observed differences in voting behaviour.

[3] There are two potential reasons why this assumption may not hold. First, there might be reverse causality, as subregions with certain political preferences could be more likely to receive NTV. Second, there might have been some omitted characteristics of subregions that correlated both with the presence of the NTV signal and the political preferences of the population.

noteworthy attempts, one drawback of the approach is that neither Unity nor OVR - two major parties expected to be influenced by NTV the most - were present at the time of the 1995 elections; hence the validity of *Unconfoundedness* is not fully checked. As a result, the plausibility of Enikolopov et al. [2011]'s identification strategy relies strongly on their specification of a control-variable set. This fact directs our attention towards Enikolopov et al. [2011]'s rationale when choosing controls in (5.2):

- The vector of socioeconomic controls $\boldsymbol{E}_{s,1998}$ in the baseline regressions includes *a dummy for cities*, *the fifth-order polynomial of population*, *the fifth-order polynomial of average wage* (as the direct determinants of NTV availability), and *the number of doctors and nurses per capita* (as a proxy for the quality of public goods provision, which can be an important determinant of voting for the pro-government party). In addition, they verify that the results are robust to including a larger set of socioeconomic controls (includes *migration rate*, *average pension*, *the fraction of retired people*, *the fraction of unemployed*, *the number of people employed in farms*, and *crime rate*).
- They present results for each voting outcome: without and with controls for the election results from 1995 ($\boldsymbol{X}_{s,1995}$).

For aggregate analysis, the sample includes a set of from 1686 to 2005 sub-regions (each sub-region is an observation) in 79-81 regions, depending on the specification.

This *ad hoc* selection procedure is not quite rare in the practice of empirical economics, as we have mentioned in section 2.1.3. There are unclear variable choices such as *the fifth-order polynomial of population* and *the fifth-order polynomial of average wage* rather than other transformations. Moreover, the authors argue that *the number of doctors and nurses per capita* is a proxy for the quality of public goods provision, thereby determining vote outcome for the pro-government party. Even if that were the case, this factor might not actually help predict voting for other parties. Imposing such fixed specifications to fit all cases could be inflexible and sometimes problematic. We will relax these constraints in a new analysis.

Results

The results derived by Enikolopov et al. [2011] are consistent with the above hypotheses[4]:

[4]Please see Table 2 — Effect of NTV Availability on Voting Behavior in 1999, Aggregate Data, Cross Section

- The vote for Unity was significantly smaller in sub-regions with higher NTV availability, and the magnitude of the effect is the largest.
- The effect of NTV availability on the combined vote for the three opposition parties, supported by this channel, is significantly positive.
- An increase in NTV availability leads to a decline in turnout.

5.3 New Analysis

For the new analysis, we take the argument that the NTV availability may be taken as exogenous relative to voting outcomes once observables have been conditioned on from Enikolopov et al. [2011] as given. We use the same state-level data as in `NTV_Aggregate_Data.dta` but only keep observations with nonzero value for the largest set of controls stated above. Therefore, the remained sample consists of 1586 sub-regions (observations) in 79 regions.

Departing from the original paper, this new analysis allows for a much richer set of controls than allowed for in $\{\boldsymbol{X}_{s,1995}; \boldsymbol{E}_{s,1998}; \delta_r\}$. Specifically, *population* and *average wage* are used instead of *the fifth-order polynomial of population* and *the fifth-order polynomial of average wage*. Furthermore, all second-order polynomials of the continuous covariates (except for *logarithm of population*, *logarithm of wage* and *the population* change), 78 dummies for regions and all possible first-order interactions of non-region variables are included. After removing collinear columns, we obtain a set of 325 control variables (maximum) as the baseline specification to flexibly select among. With the sample of 1586 observations, we are dealing with the data set in which p is large relative to n. Even though $p < n$, applying the OLS-based methods to a specification with all 325 controls is still not applicable in this context because of the ill-conditioned matrix issue[5].

This phenomenon motivates us to apply Frequentist and Bayesian regularization-based methods designed in section 3. Specifically, we consider Post-Double-Seclection Lasso (`PDSLasso`) and High-dimensional Confounding Adjustment (`HDCA`) methods, which includes `SSVSNormal`, `SSVSStudent`, `SSVSLasso1`, `SSVSLasso2`, `SSVSLasso3` and `SSVSHorseshoe`. In all methods, Iterated Lasso is used for Lasso-based selection steps. Table 5.1 presents the estimation results corresponding to different dependent variables (voter turnout and voting outcomes for major parties). For each dependent variable, we report results using regularization-based methods on on the full set of potential controls ($p = 325$), regression result using OLS for a linear regression with no control (labelled as `No control` in Table 5.1). We also provide

[5]The conditional number of $X'X$ is extremely large, approximate 7.46×10^{22}

original results estimated by Enikolopov et al. [2011] using OLS with two sets of controls with and without the election outcome from 1995, as discussed in section 2.1 (labelled as `Old Large` and `Old Small` in Table 5.1).

Panel A demonstrates the estimated effect of NTV availability on voter turnout. Panel B presents the result for Unity, the main party opposed by NTV. Panel C indicates results for parties supported by NTV, where OVR is the main centrist opposition party. And panel D features the effect on parties who got similar coverage by NTV and the two governmental television channels.

First of all, we take a brief look at relative performances of the Frequentist and Bayesian regularization-based methods in all panels. The findings confirm our analysis of behaviours of these methods in the earlier simulation section. Only `SSVSStudent` suffers from the ill-conditioned matrix issue (which hinders the use of OLS) so that nothing is estimated. It can be explained by the pattern in the Monte Carlo study, where `SSVSStudent` is apt to include the largest number of possible controls into the slab in all scenarios. Among SSVS Lasso priors, `SSVSLasso1` provides a larger credible interval length, as well as having a higher inclusion probability of possible controls. `SSVSLasso2` and `SSVSLasso3` perform nearly identical and produce estimates for treatment coefficients closed to `SSVSNormal` along with smaller standard errors thanks to higher shrinkage effects. Relative performance of `SSVSHorseshoe` is unstable in comparison with other `HDCA` methods. In terms of the Frequentist approach, `PDSLasso` is inclined to retain a more parsimonious final set of controls than the *ad hoc* approach (Column 6). The size of this set varies from party to party, especially significant to the case of Yabloko (24 controls), which suggests the flexibility of the systematic-search approach. Bayesian methods do not produce one specific final set; however, the parsimony level is partly reflected in their average inclusion frequency of a potential control into the slab (Column 7). It is worth noting that even if one variable is included in the slab, this variable can still be shrunk; the effect depends upon the type of Bayesian shrinkage priors. In addition, the CI length of `SSVSLasso3` is always smaller than `PDSLasso`. This finding is consistent with the empirical illustration of Antonelli et al. [2019]. These authors also highlight that while the goal of Post-Double-Selection Lasso is to obtain valid inference in high-dimensional setup, High-dimensional Confounding Adjustment with spike-and-slab Lasso provides the more efficient estimate of the treatment effect; the reason lies in how they address instrumental variables.

Table 5.1. Estimation of the effect of NTV availability on voting outcomes using Regularization-based methods

Panel A. Voter turnout in 1999

	EstTE	LbTE	UbTE	Length	Potential_Set	Final_Set	Percentage
No control	-22.924	-28.0775	-17.771	10.307	n.a.	0	n.a.
Old Small	-6.540	-10.3424	-2.738	7.605	n.a.	91	n.a.
Old Large	-6.670	-9.4532	-3.887	5.566	n.a.	99	n.a.
PDS Lasso	-7.693	-11.4352	-3.952	7.483	325	59	18.154
SSVS Normal	-5.448	-9.08	-1.533	7.547	325	n.a.	60.546
SSVS Student-t	n.a.	n.a.	n.a.	n.a.	325	n.a.	n.a.
SSVS Lasso1	-5.692	-10.0455	-1.161	8.884	325	n.a.	87.287
SSVS Lasso2	-5.787	-9.2316	-2.358	6.874	325	n.a.	53.692
SSVS Lasso3	-5.805	-9.2189	-2.416	6.803	325	n.a.	53.750
SSVS Horseshoe	-4.680	-8.3093	-1.027	7.282	325	n.a.	42.036

Panel B. Opposed by NTV in 1999

Vote for Unity in 1999 (centrist, progovement)

	EstTE	LbTE	UbTE	Length	Potential_Set	Final_Set	Percentage
No control	-34.397	-40.0348	-28.760	11.275	n.a.	0	n.a.
Old Small	-17.720	-22.6396	-12.800	9.839	n.a.	91	n.a.
Old Large	-15.480	-20.9092	-10.051	10.858	n.a.	99	n.a.
PDS Lasso	-12.451	-19.1204	-5.781	13.340	325	54	16.615
SSVS Normal	-15.719	-21.3826	-10.097	11.285	325	n.a.	66.248
SSVS Student-t	n.a.	n.a.	n.a.	n.a.	325	n.a.	n.a.
SSVS Lasso1	-18.597	-24.7771	-12.176	12.601	325	n.a.	90.098
SSVS Lasso2	-13.355	-18.7278	-8.028	10.700	325	n.a.	59.783
SSVS Lasso3	-13.185	-18.276	-7.855	10.421	325	n.a.	60.032
SSVS Horseshoe	-17.058	-22.5285	-11.803	10.725	325	n.a.	44.077

Panel C. Supported by NTV in 1999

Vote for OVR in 1999 (centrist, opposition)

	EstTE	LbTE	UbTE	Length	Potential_Set	Final_Set	Percentage
No control	13.209	6.6589	19.760	13.101	n.a.	0	n.a.
Old Small	5.720	1.898	9.542	7.644	n.a.	91	n.a.
Old Large	3.620	0.2488	6.991	6.742	n.a.	99	n.a.
PDS Lasso	3.171	-1.8406	8.183	10.023	325	57	17.538
SSVS Normal	3.740	-1.039	8.128	9.167	325	n.a.	62.939
SSVS Student-t	n.a.	n.a.	n.a.	n.a.	325	n.a.	n.a.
SSVS Lasso1	5.403	0.3589	10.749	10.390	325	n.a.	89.590
SSVS Lasso2	3.003	-1.1865	7.293	8.479	325	n.a.	56.673
SSVS Lasso3	3.131	-0.8867	7.323	8.210	325	n.a.	57.219
SSVS Horseshoe	4.019	0.25247	7.773	7.521	325	n.a.	37.932

Vote for SPS in 1999 (liberal)

	EstTE	LbTE	UbTE	Length	Potential_Set	Final_Set	Percentage
No control	10.556	8.1479	12.964	4.816	n.a.	0	n.a.
Old Small	4.470	2.3728	6.567	4.194	n.a.	91	n.a.
Old Large	3.520	1.266	5.774	4.508	n.a.	99	n.a.
PDS Lasso	2.538	0.72141	4.354	3.633	325	57	17.538
SSVS Normal	1.943	-0.12751	4.050	4.178	325	n.a.	45.423
SSVS Student-t	n.a.	n.a.	n.a.	n.a.	325	n.a.	n.a.
SSVS Lasso1	3.147	0.90162	5.416	4.514	325	n.a.	75.375
SSVS Lasso2	1.873	0.19983	3.437	3.237	325	n.a.	32.294
SSVS Lasso3	1.869	0.17055	3.460	3.289	325	n.a.	32.108
SSVS Horseshoe	2.852	1.3537	4.345	2.992	325	n.a.	66.570

Vote for Yabloko in 1999 (liberal)

	EstTE	LbTE	UbTE	Length	Potential_Set	Final_Set	Percentage
No control	11.742	10.2813	13.202	2.921	n.a.	0	n.a.
Old Small	4.580	2.9336	6.226	3.293	n.a.	91	n.a.
Old Large	3.850	2.5368	5.163	2.626	n.a.	99	n.a.
PDS Lasso	4.723	3.0329	6.414	3.381	325	24	7.385
SSVS Normal	3.563	2.5207	4.597	2.076	325	n.a.	16.702
SSVS Student-t	n.a.	n.a.	n.a.	n.a.	325	n.a.	n.a.
SSVS Lasso1	3.987	2.147	5.810	3.663	325	n.a.	65.482
SSVS Lasso2	2.944	1.9526	3.907	1.954	325	n.a.	12.896
SSVS Lasso3	2.975	1.9262	3.925	1.999	325	n.a.	12.388
SSVS Horseshoe	3.539	2.5708	4.611	2.040	325	n.a.	93.538

Panel D. Similar coverage by NTV and state TV in 1999

Vote for KPRF in 1999 (communist)

	EstTE	LbTE	UbTE	Length	Potential_Set	Final_Set	Percentage
No control	0.016	-5.9546	5.987	11.941	n.a.	0	n.a.
Old Small	1.680	-2.2988	5.659	7.958	n.a.	91	n.a.
Old Large	3.920	0.2744	7.566	7.291	n.a.	99	n.a.
PDS Lasso	5.946	1.5355	10.357	8.822	325	59	18.154
SSVS Normal	6.808	2.4998	10.994	8.494	325	n.a.	61.144
SSVS Student-t	n.a.	n.a.	n.a.	n.a.	325	n.a.	n.a.
SSVS Lasso1	5.988	0.91651	11.082	10.166	325	n.a.	88.089
SSVS Lasso2	6.177	2.31	10.328	8.018	325	n.a.	55.734
SSVS Lasso3	6.083	2.356	10.503	8.147	325	n.a.	56.074
SSVS Horseshoe	7.086	3.2396	11.055	7.815	325	n.a.	40.667

Vote for LDPR in 1999 (nationalist)

	EstTE	LbTE	UbTE	Length	Potential_Set	Final_Set	Percentage
No control	-5.826	-7.3941	-4.258	3.136	n.a.	0	n.a.
Old Small	-1.720	-3.0136	-0.426	2.587	n.a.	91	n.a.
Old Large	-1.390	-2.5856	-0.194	2.391	n.a.	99	n.a.
PDS Lasso	-1.967	-3.2954	-0.638	2.657	325	63	19.385
SSVS Normal	-1.820	-3.0915	-0.645	2.446	325	n.a.	26.378
SSVS Student-t	n.a.	n.a.	n.a.	n.a.	325	n.a.	n.a.
SSVS Lasso1	-1.777	-3.8401	0.296	4.136	325	n.a.	69.909
SSVS Lasso2	-1.440	-2.6102	-0.271	2.340	325	n.a.	21.165
SSVS Lasso3	-1.488	-2.6108	-0.364	2.246	325	n.a.	20.123
SSVS Horseshoe	-1.740	-3.0105	-0.507	2.504	325	n.a.	87.339

†Column 1 represents the estimates of treatment effect.

††Column 2 and 3 represent the lower bounds and upper bounds of 95% confidence (credible) intervals, respectively.

††Column 4 represents lengths of 95% confidence (credible) intervals.

We are turning now to the comparison with OLS-based approaches. `No control` specification yields estimation results that are vastly different from others, and CI lengths are the largest. It reflects the consequence of omitted variable bias when we ignore important confounders. Compared to original results from intuitive specifications `Old Small` and `Old Large` of Enikolopov et al. [2011], new methods provide results that are similar in magnitude and direction, except for the case of OVR. `PDSLasso`, `SSVSStudent`, `SSVSLasso2` and `SSVSLasso3` (which often dominate `SSVSLasso1` and `Horseshoe`) produce credible intervals that encompass both positive and negative values, thereby suggesting an insignificant effect of NVT availability on vote outcome for OVR. Whereas this result deviates from the positive effect as suggested by Enikolopov et al. [2011], there are several possible explanations. First, the original estimates seem quite sensitive because of large standard errors. Additionally, there is no sensitivity analysis to verify its robustness since OVR did not exist in the preceding periods. Notwithstanding the inconsistency, this ambiguous effect on the major party supported by NTV does not contradict theoretical and empirical evidence that negative political advertising is often more effective than positive.

Fairly speaking, there is no reason to argue that new estimates from regularization-based methods are more reliable than the initial results because performances of these methods hinge on particular designs. For example, as shown in the simulation study, all examined methods might perform poorly in low-sparsity scenarios. Nevertheless, regularization-based methods complement the usual careful specification analysis by providing a researcher with an efficient, data-driven way to search for a small set of influential confounders from a sensibly chosen vast set of potential control variables.

To make the problem further appealing, we generate a high-dimensional setting by drawing random sub-samples of size 300 observations with replacement from the full data set (1568 observations) and implementing estimation throughout 100 simulation repetitions. Unfortunately, this design exaggerates the ill-posed problem so that `PDSLasso` and all `HDCA` methods become not applicable.

CHAPTER

SIX

CONCLUSION

The goal of this thesis was to examine the merits of Frequentist and Bayesian regularization-based methods to inference on treatment effects with high-dimensional controls. We have considered the Post-Double-Selection Lasso [Belloni et al., 2014b] and several Bayesian methods corresponding to a variety of Bayesian shrinkage choices within a common framework - the generalized High-dimensional Confounding Adjustment approach [Antonelli et al., 2019].

The main findings from the simulation study and the empirical illustration implemented in this thesis are as follows:

At the beginning, we have considered the high dimensional econometric settings where the number of potential control variables (p) is very close to or even larger than the number of observations (n). Within this context, a traditional approach such as the OLS method fails to provide reliable results. In specific, in our Monte-Carlo study when $p = 400$ is larger than $n = 300$, the OLS estimator for the treatment effect in a full-control specification is not identified. In our empirical example with a real data-set where $p = 325$ is comparable yet still smaller than $n = 1568$ (i.e. a full rank setting), the OLS estimator becomes very imprecise because of the ill-posed problems. In contrast, estimators associated with regularization-based methods can be computed and perform well in many setups in terms of achieving the nominal coverage rate. Therefore, the advantages of modern methods in comparison with traditional counterparts in high-dimensional scenarios are undeniable.

Regarding the relative performances among regularization-based methods, the conclusion depends on the context. As shown in the simulation study, `PDSLasso` dominates `HDCA` methods in high-sparsity designs thanks to its stable performance (lower bias and safer coverage

rate). However, in low-sparsity designs, `PDSLasso` fails to select confounders and performs the worst. Regarding the empirical illustration, `PDSLasso` provides the estimates qualitatively consistent with `SSVSLasso2`, `SSVSLasso3` and `SSVSNormal`; while the length of 95% confidence (credible) interval of `PDSLasso` is always larger than that of `SSVSLasso3`. In terms of implementation, `HDCA` methods require MCMC samplings, thereby being more computational intensive compared to `PDSLasso`.

Among `HDCA` methods, the performance depends upon the choices of Bayesian shrinkage priors as well as the penalty parameters. There are some general patterns realized in this analysis. `SSVSStudent` and `SSVSLasso1` tend to produce relatively large standard errors (and large average interval length as a result).While it can be seen as a trade-off for the highest coverage rates of `SSVSStudent` in the Monte-Carlo study, this method ranks lowest in the empirical replication because it is the only one that suffers from the ill-posed problem and estimates nothing. `SSVSLasso2` and `SSVSLasso3` share almost identical behaviours and perform quite similar to `SSVSNormal`. Although they cannot achieve the high coverage rates as `SSVSStudent`, their small standard errors are desirable in both simulation and empirical examples. `SSVSHorseshoe` often produces the highest bias and RMSE in simulation designs and does not show any clear behaviour in a real dataset. In terms of variants of spike-and-slab [Kuo and Mallick, 1998], `SnSNormal` and `SnSLasso` present a medium performance so they are not considered in the empirical part. Insights from inclusion probability could serve as quite convincing explanations. `SSVSStudent` and `SSVSLasso1` are good at selecting confounders, but simultaneously include highest number of non-confounding factors. By contrast, `SSVSLasso2`, `SSVSLasso3` and `SSVSHorseshoe` perform well in excluding non-confounding factors, but at the same time, they select less confounders into slab. Notably, although some DGP parameters can influence the isolate performance of each method (e.g. higher SNR in the second stage deteriorates the performance of each method), the relative pattern described above is quite robust across different scenarios.

Finally, regularization-based methods should not be regarded as panaceas, according to the results of this study. The Monte-Carlo study reveals that there are designs in which even well-known theory-based methods like `PDSLasso` could show poor finite-sample performance, e.g. low-sparsity designs. Although outperforming `PDSLasso` in such cases, new Bayesian methods are still far from the desired target. Evidence from the empirical example even poses a more challenging situation. `SSVSStudent` cannot overcome the problem that drowns traditional approaches (ill-posed problem/ multicollinearity even when $p < n$). Moreover, when multicollinearity is compounded with $n < p$, every method considered in this study

fail to estimate treatment effects.[1]

These findings, while preliminary, have some implications for further studies in developing the Bayesian regularization-based methods for inference on treatment effects with high-dimensional controls:

With respect to methodology, this study confirms that there is a link between the finite-sample performance in variable selection and in causal inference of Bayesian methods. Since the choice of Bayesian shrinkage prior could affect variable-selection performance[Van Erp et al., 2019, Polson and Sokolov, 2019], this may influence the causal-inference performance of a method as well. The Bayesian approach offers a huge choice of shrinkage priors, thus being potential for new methods. However, a thoughtful implementation is required. For example, Horseshoe prior, which is novel in the Bayesian literature, does not ensure that `SSVSHorshoe` could perform well in our settings. In fact, the most stable performance in our case belongs to priors which have good established theoretical results - `SSVSNormal` [Narisetty and He, 2014] and `SSVSLasso3` [Ročková and George, 2018]. Even if that were the case, the values of hyper-parameters cannot be neglected. The ability to account for parameter uncertainty of Bayesian methods are often sensitive to those factors, which remains a challenging aspect.

With regards to the simulation study, future research could refine our design by elaborating some parameters which seem to be important in this thesis: sparsity levels, signal-to-noise ratio in both equations. A range of other parameters such as the number of observations, specific hyper-parameters, etc. should be considered as well. Especially, a more sophisticated correlation design is necessary since the issue in our empirical illustration is not well captured by the current Monte-Carlo study. Furthermore, semi-synthetic designs are encouraged.

Regarding implications for empirical works in economics, the findings suggest that Frequentist and Bayesian regularization-based methods offer a coherent data-driven complement to ad hoc robustness checks, thus, support causal analysis in linear regression models[2]. However, the scope of this thesis is limited since it is only a special case of a special case of inference on treatment effects:

First, there are some implicit restrictions on the set of potential controls. Although this study discusses about selecting among controls, the set of potential controls is assumed to be "not bad", e.g. it at least does not contain pre-treatment variables, to enable the equivalent between the causal estimand ATE and the regression coefficent α in section 2.1.

[1] For further research: Yue et al. [2019], Celeux et al. [2012],etc.

[2] This is in light with Angrist and Frandsen [2019], Belloni et al. [2014b], etc. See Wuthrich and Zhu [2019] for caveats.

Distinguishing "good" from "bad" controls remains ambiguous, therefore, a future research in modern methods should pay more systematic attention to moderate the idea of Directed Acyclic Graphs (DAGs) [Cinelli et al., 2020, VanderWeele, 2019] .

Second, this study is based on the Unconfoundedness assumption to focus on a linear regression model with conditional-on-observables identification strategy to carry out estimation and inference on treatment effects. That appears a simplest form in the literature of causal inference, thus, a future study could examine the merits of Frequentist and Bayesian regularization-based methods in settings which allow for unmeasured confounding factors[3].

Finally, machine labor cannot replace brain power. It is evident that the methods introduced in this study are potential yet far from the cure-all. New methods should be applied with inference in mind, to quantify our degree of confidence, also importantly, our degree of uncertainty./

[3]For further research: Abadie and Cattaneo [2018], Athey and Imbens [2017], Belloni et al. [2017], etc.

BIBLIOGRAPHY

A. Abadie and M. D. Cattaneo. Econometric methods for program evaluation. *Annual Review of Economics*, 10:465–503, 2018.

A. Ahrens, C. B. Hansen, and M. E. Schaffer. lassopack: Model selection and prediction with regularized regression in stata. *The Stata Journal*, 20(1):176–235, 2020.

R. Alhamzawi and H. Taha Mohammad Ali. A new gibbs sampler for bayesian lasso. *Communications in Statistics-Simulation and Computation*, 49(7):1855–1871, 2020.

J. Angrist and B. Frandsen. Machine labor. Technical report, National Bureau of Economic Research, 2019.

J. D. Angrist and J.-S. Pischke. *Mostly harmless econometrics*. Princeton university press, 2008.

J. D. Angrist and J.-S. Pischke. The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of economic perspectives*, 24(2):3–30, 2010.

J. Antonelli, C. Zigler, and F. Dominici. Guided bayesian imputation to adjust for confounding when combining heterogeneous data sources in comparative effectiveness research. *Biostatistics*, 18(3):553–568, 2017.

J. Antonelli, G. Parmigiani, and F. Dominici. High-dimensional confounding adjustment using continuous spike and slab priors. *Bayesian analysis*, 14(3):805, 2019.

A. Armagan and R. L. Zaretzki. Model selection via adaptive shrinkage with t priors. *Computational Statistics*, 25(3):441–461, 2010.

S. Athey and G. W. Imbens. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32, 2017.

S. Athey, G. W. Imbens, and S. Wager. Approximate residual balancing: De-biased inference of average treatment effects in high dimensions. *arXiv preprint arXiv:1604.07125*, 2016.

B. Barnow, G. Cain, and A. Goldberg. Selection on observables. *Evaluation Studies*, 1981.

A. Belloni and V. Chernozhukov. High dimensional sparse econometric models: An introduction. In *Inverse Problems and High-Dimensional Estimation*, pages 121–156. Springer, 2011.

A. Belloni, D. Chen, V. Chernozhukov, and C. Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429, 2012.

A. Belloni, V. Chernozhukov, and C. Hansen. High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2):29–50, 2014a.

A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014b.

A. Belloni, V. Chernozhukov, C. Hansen, and D. Kozbur. Inference in high-dimensional panel models with an application to gun control. *Journal of Business & Economic Statistics*, 34 (4):590–605, 2016.

A. Belloni, V. Chernozhukov, I. Fernández-Val, and C. Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017.

A. Bhadra, J. Datta, N. G. Polson, and B. Willard. Lasso meets horseshoe: A survey. *Statistical Science*, 34(3):405–427, 2019.

E. Candes and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The annals of Statistics*, 35(6):2313–2351, 2007.

C. M. Carvalho, N. G. Polson, and J. G. Scott. Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics*, pages 73–80. PMLR, 2009.

C. M. Carvalho, N. G. Polson, and J. G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

G. Casella, M. Ghosh, J. Gill, and M. Kyung. Penalized regression, standard errors, and bayesian lassos. *Bayesian analysis*, 5(2):369–411, 2010.

I. Castillo and A. van der Vaart. Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics*, 40(4):2069–2101, 2012.

G. Celeux, M. El Anbari, J.-M. Marin, and C. P. Robert. Regularization in regression: comparing bayesian and frequentist methods in a poorly informative situation. *Bayesian Analysis*, 7(2):477–502, 2012.

V. Chernozhukov. Mostly dangerous econometrics: How to do model selection with inference in mind. 2015.

H. Chipman, E. I. George, R. E. McCulloch, M. Clyde, D. P. Foster, and R. A. Stine. The practical implementation of bayesian model selection. *Lecture Notes-Monograph Series*, pages 65–134, 2001.

C. Cinelli, A. Forney, and J. Pearl. A crash course in good and bad controls. *Available at SSRN*, 3689437, 2020.

R. Enikolopov, M. Petrova, and E. Zhuravskaya. Media and political persuasion: Evidence from russia. *American Economic Review*, 101(7):3253–85, 2011.

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

M. H. Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23, 2015.

E. Gautier and A. B. Tsybakov. Pivotal estimation in high-dimensional regression via linear programming. In *Empirical inference*, pages 195–204. Springer, 2013.

E. I. George and R. E. McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

E. I. George and R. E. McCulloch. Approaches for bayesian variable selection. *Statistica sinica*, pages 339–373, 1997.

J. Griffin and P. Brown. Alternative prior distributions for variable selection with very many more variables than observations. *University of Kent Technical Report*, 2005.

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

P. R. Hahn, C. M. Carvalho, D. Puelz, and J. He. Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Analysis*, 13(1):163–182, 2018.

P. R. Hahn, J. S. Murray, and C. M. Carvalho. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056, 2020.

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

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer: New York, NY, 2009.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

T. Hsiang. A bayesian view on ridge regression. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 24(4):267–268, 1975.

J. Huang, J. L. Horowitz, and S. Ma. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *The Annals of Statistics*, 36(2):587–613, 2008.

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press, 2015.

H. Ishwaran and J. S. Rao. Spike and slab variable selection: frequentist and bayesian strategies. *The Annals of Statistics*, 33(2):730–773, 2005.

L. Kuo and B. Mallick. Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 65–81, 1998.

E. E. Leamer. Let's take the con out of econometrics. *The American Economic Review*, 73 (1):31–43, 1983.

H. Leeb and B. M. Pötscher. Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, 24(2):338–376, 2008.

Q. Li and N. Lin. The bayesian elastic net. *Bayesian analysis*, 5(1):151–170, 2010.

R. J. Little and D. B. Rubin. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual review of public health*, 21(1):121–145, 2000.

E. Makalic and D. F. Schmidt. A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23(1):179–182, 2015.

H. Mallick and N. Yi. A new bayesian lasso. *Statistics and its interface*, 7(4):571–582, 2014.

MATLAB. *R2020a*. The MathWorks Inc., Natick, Massachusetts, 2020.

T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the american statistical association*, 83(404):1023–1032, 1988.

S. L. Morgan and C. Winship. *Counterfactuals and causal inference*. Cambridge University Press, 2015.

N. N. Narisetty and X. He. Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, 42(2):789–817, 2014.

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

N. G. Polson and J. G. Scott. Shrink globally, act locally: Sparse bayesian regularization and prediction. *Bayesian statistics*, 9(501-538):105, 2010.

N. G. Polson and V. Sokolov. Bayesian regularization: From tikhonov to horseshoe. *Wiley Interdisciplinary Reviews: Computational Statistics*, 11(4):e1463, 2019.

N. G. Polson and L. Sun. Bayesian l 0-regularized least squares. *Applied Stochastic Models in Business and Industry*, 35(3):717–731, 2019.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`.

V. Ročková and E. I. George. The spike-and-slab lasso. *Journal of the American Statistical Association*, 113(521):431–444, 2018.

J. P. Simmons, L. D. Nelson, and U. Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11):1359–1366, 2011.

StataCorp. *Stata Statistical Software: Release 17*. College Station, TX: StataCorp LLC, 2021.

D. Talbot, G. Lefebvre, and J. Atherton. The bayesian causal effect estimation algorithm. *Journal of Causal Inference*, 3(2):207–236, 2015.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

A. N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Mathematics Doklady*, 4:1035–1038, 1963.

S. Van Erp, D. L. Oberski, and J. Mulder. Shrinkage priors for bayesian penalized regression. *Journal of Mathematical Psychology*, 89:31–50, 2019.

T. J. VanderWeele. Principles of confounder selection. *European journal of epidemiology*, 34 (3):211–219, 2019.

C. Wang, G. Parmigiani, and F. Dominici. Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics*, 68(3):661–671, 2012.

C. Wang, F. Dominici, G. Parmigiani, and C. M. Zigler. Accounting for uncertainty in confounder and effect modifier selection when estimating average causal effects in generalized linear models. *Biometrics*, 71(3):654–665, 2015.

F. Wang, S. Mukherjee, S. Richardson, and S. M. Hill. High-dimensional regression in practice: an empirical study of finite-sample prediction, variable selection and ranking. *Statistics and computing*, 30(3):697–719, 2020.

S. Woody, C. M. Carvalho, and J. S. Murray. Bayesian inference for treatment effects under nested subsets of controls. *arXiv preprint arXiv:2001.07256*, 2020.

J. M. Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.

K. Wuthrich and Y. Zhu. Omitted variable bias of lasso-based inference methods: A finite sample analysis. *Available at SSRN 3379123*, 2019.

X. Xu and M. Ghosh. Bayesian variable selection and estimation for group lasso. *Bayesian Analysis*, 10(4):909–936, 2015.

L. Yue, G. Li, H. Lian, and X. Wan. Regression adjustment for treatment effect with multicollinearity in high dimensions. *Computational Statistics & Data Analysis*, 134:17–35, 2019.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.