FACULTY OF SOCIAL SCIENCES
Department of Economics
University of Copenhagen

Master's Thesis

Ditlev Kiersgaard Frisch

Variable Selection and Inference in High-Dimensional Sparse Econometric Models

A Bayesian Perspective

Supervisor: Rémi Piatek

ECTS points: 30

Date of submission: 31/12/2018

Keystrokes: 136,196

Abstract

This thesis provides a Bayesian perspective to estimation of high-dimensional sparse (HDS) linear regression models in econometrics. These models have a large number of regressors p relative to the sample size N, but only a small number $p_0 < N$ of these are needed to capture the important characteristics of the regression function, whereby estimation procedures for detecting and removing the zero regression coefficients from the model are needed. In the frequentist literature, the post-Lasso (Belloni et al., 2013) and double-Lasso (Belloni et al., 2014b) estimators have proven successful in, respectively, estimating the regression function and estimating treatment effects in HDS linear regression models, and this paper therefore studies the spike-and-slab model alternatives to these two estimators that provide natural ways to represent sparsity within the Bayesian paradigm. These models are applicable to various settings, and an extension to a HDS finite mixture model of linear regressions is therefore also considered.

To improve variable selection in the HDS linear regression model, a novel-type mixture prior for the spike-and-slab hierarchy is developed, and a two-step spike-and-slab analogue to the double-Lasso is also proposed to improve inference on the treatment parameter that performs better than the double-Lasso in simulation studies. The Lasso and Bayesian approaches to three HDS models - the standard linear regression model; the treatment effects model; and the finite mixture model of linear regressions - are then compared in a horse race study. The post-Lasso performs best in terms of variable selection, bias and fit in the standard linear regression model, whereas in a treatment effects model and in the finite mixture model of linear regressions the spike-and-slab performs better. To illustrate the applicability of the spike-and-slab models in economic research, an empirical example of the effects of abortion on crime is considered.

Contents

1 Introduction

1.1 Motivation

Estimating the regression function or the causal effect of a variable on an outcome in observational data is a common goal in econometrics. For causal inference, for instance, the traditional approach is to select a regression model to estimate the effect, include the confounding variables to the model, and report the estimate and its statistical significance conditional on the model being correct. Inference on the parameter of interest is then valid insofar 1) the model is correctly specified and 2) the true set of confounders is included. While the intuitive approach to validating these two conditions is widely employed in academia and industry - such as ad hoc robustness checks and/or simply telling a convincing story - it has come to be challenged by new, data-driven methods where the nuisance function of confounders or the entire regression function is flexibly modelled via high-dimensional or so-called machine learning methods that are able to select the correct model and, for causal inference, identify the confounding variables. These methods are characterized by performing well on data sets with a large number of variables p relative to the number of observations N where standard methods (such as ordinary least squares, OLS) fail to supply useful insights because the parameters are estimated with a large variance - or may even seize to exist when $p > N$. Prominent and widely used examples from the machine learning literature are the Least Absolute Shrinkage and Selection Operator (Lasso) (Tibshirani, 1996) that leaves variable selection fully data-driven, and the non-parametric random forests (Breiman, 2001) that flexibly estimates the regression function by partitioning the sample into smaller parts.

However, while machine learning provides a powerful, flexible way of making high-quality predictions, it does not in itself produce stable estimates for inference on the structural parameters, which is most often the interest to economists. For instance, the Lasso procedure stains the non-zero coefficients with a negative bias as it pulls all coefficients towards zero, which leaves inference infeasible. Therefore, while these machine learning methods have a long and successful history in the statistical literature, they have only recently caught interest in economics where the models have been recast

in order to utilize the powerful prediction toolbox for inference, see Varian (2014), Mullainathan and Spiess (2017) and Athey (2018) for reviews on this. The overall take-away from these reviews is that in an era of "big data" and increasing computing power, econometrics better get used to incorporating high-dimensional machine learning methods. Thus far, most of the successful applications of machine learning in econometrics have been in the frequentist literature, e.g. the post-Lasso (Belloni et al., 2013) for estimating high-dimensional sparse regression functions, the double-Lasso procedure for selection among high-dimensional controls (Belloni et al., 2014b) and the causal forests for heterogenous treatment effects (Wager and Athey, 2017), where Bayesian applications in high-dimensional econometrics have in large part been overlooked. The objective of this paper is to fill part of this gap and offer a Bayesian perspective to variable selection and (causal) inference in these high-dimensional econometric settings.

1.2 Problem Formulation

Specifically, I consider Bayesian approaches to estimation of high-dimensional sparse (HDS) linear regression models, which are models where the number of regressors p is large relative to the sample size N, but only a small number p_0 of these are needed to capture the important characteristics of the regression function. These types of econometric models were first discussed in Belloni and Chernozhukov (2011), where the assumption of p_0 being sufficiently small enables estimation by searching for the true regressors (with non-zero coefficients) via e.g. ℓ_1 regularization like the Lasso, and it therefore provides a meaningful framework that I adopt throughout the rest of this thesis to analyze high-dimensional methods.

Three cases are considered: Variable selection and inference on the regression function in the normal linear regression model, inference on a treatment parameter in a linear model among high-dimensional controls, and the generalization of the standard linear regression model to a finite mixture model of linear regressions.

For the first case, I consider the normal linear regression model

$$y_i = Z_i\mu + X_i\beta + \epsilon_i \; \epsilon_i \sim \mathrm{N}(0, \sigma^2), \; i = 1, ..., N \tag{1}$$

where $Z_i \in \mathbb{R}^l$ and $X_i \in \mathbb{R}^p$ are two vectors of regressors with p large (and potentially $p \gg N$), and Z represents the set of regressors that the analyst always wants to include in the model, e.g. an intercept term. Because of the sparsity assumption underlying HDS models, the dual goal then becomes to learn the support of β, support$(\beta) \subset (1, ..., p)$, and estimate the non-zero entries precisely, where the first goal naturally affects the second. To this end, I focus attention on the Bayesian spike-and-slab models that were introduced in Mitchell and Beauchamp (1988) for simultaneous variable selection and parameter estimation in (1), and later popularized in George and McCulloch (1993) with a Markov chain Monte Carlo (MCMC) scheme that allowed for feasible sampling from the posterior. These models provide a natural Bayesian way to represent sparsity, and I also consider a modification of the models where I develop a novel mixture prior for the spike-and-slab hierarchy that is intended to improve on the variable selection performance. The performance of the spike-and-slab models are then compared to the post-Lasso estimator (Belloni et al., 2013), which is a powerful frequentist method to estimation and variable selection of regression functions in HDS models like (1).

Building on these methods, I then consider the structural model

$$y_i = \mu + \alpha_0 d_i + x_i \beta_y + \epsilon_i, \ \epsilon_i \sim N(0, \sigma_\epsilon^2) \tag{2}$$

$$d_i = x_i \beta_d + \nu_i, \ \nu_i \sim N(0, \sigma_\nu^2) \tag{3}$$

where d_i is a scalar continuous treatment variable, x_i is a high-dimensional $p \times 1$ vector of potential confounders, and α_0 is the treatment parameter of interest. The double-Lasso method (Belloni et al., 2014b) has proven successful in achieving precise estimates of α_0 even in $p > N$ designs after selecting controls via the Lasso in both equations (hence, the name "double"), and I provide a novel two-step spike-and-slab analogue whose posterior mean estimator actually performs better in simulation studies in terms of bias and variance than the double-Lasso estimator.

While the main bulk of the thesis is concerned with the above two cases, I also briefly consider the generalization to a finite normal mixture model of linear regressions to showcase the power of Bayesian methods over frequentist when the optimization routine of the Lasso becomes troublesome due to loss of convexity in the objective function which happens in mixture models. I consider the

model in Lee et al. (2016), where the outcome is assumed to derive from a heterogenous population of M sub-populations

$$y_i \mid \pi_m, \beta_m, \sigma_m^2 \sim \sum_{m=1}^{M} \pi_m N(X_i\beta_m, \sigma_m^2),\ \pi_m \geq 0,\ \sum_{m=1}^{M} \pi_m = 1, \qquad i = 1, ..., N \tag{4}$$

and the goal is to select the true model for each of the M mixture components (or sub-populations) via variable selection. Here, I compare the performance of the spike-and-slab approach and the Lasso-procedure outlined in Khalili and Chen (2007).

Ultimately, this thesis therefore covers two simultaneous and overlapping goals. The first goal is to compare the performance of the Lasso and its relatives to the Bayesian spike-and-slab approach in HDS models (a horse race), where my simulation studies suggest that the Lasso toolbox outperforms the Bayesian in terms of variable selection, inference and fit in the standard linear regression model (1), whereas the Bayesian methods fare better in the treatment effects model (2),(3) and in the finite mixture model of linear regressions (4). While there are of course many other frequentist methods one could compare the Bayesian methods to, the Lasso and its relatives are probably the most widely used and well-studied frequentist tools to analyze HDS linear models in both the machine learning literature and economics, and it therefore makes for a natural benchmark.

The second goal is to compare the performance of the priors I develop for the spike-and-slab hierarchy for, respectively, the standard linear model and the treatment effects model. The simulation studies suggest that the first of these priors does not improve on variable selection and bias compared to existing priors, whereas the second prior for the treatment effects improves performance dramatically and is the reason that Bayesian inference on α_0 is more precise than the double-Lasso.

1.3 Structure

While analytical focus is on the spike-and-slab models for variable selection and high-dimensional inference, most chapters include an overview of the corresponding frequentist literature with emphasis on the Lasso methods. This is motivated both by the fact that many popular Bayesian ideas in this

field originate from the frequentist literature, e.g. the Bayesian Lasso (Park and Casella, 2008), which is also the case when I develop a spike-and-slab analogue to the double-Lasso procedure; and second because they are used to benchmark the performance of the spike-and-slab models, where I do not expect the reader to already be familiar with the Lasso and its relatives.

The rest of the paper is organized as follows: Section 2 surveys the most important developments in applying machine learning methods to econometrics for inference. Section 3 first briefly surveys the frequentist literature on variable selection and post-variable selection inference in the linear regression model, and then digs into the spike-and-slab models and their MCMC implementations, where the novel mixture prior for the spike-and-slab hierarchy is also developed. Section 4 describes the double-Lasso procedure and develops its two-step spike-and-slab analogue. Section 5 outlines the application of the Lasso and the spike-and-slab models to finite mixture models of linear regressions with MCMC implementation. Section 6 conducts a simulation study to evaluate the performance of the proposed Bayesian methods. Section 7 shows a real data example of the spike-and-slab in action when inference on a treatment effect is the goal among a high-dimensional set of control variables. Section 8 concludes.

1.4 Note on Contribution

Before diving in to the subject, it is important to state a clarifying note on contribution. This project started out with no prior knowledge on how to derive and code up the MCMC algorithms for the spike-and-slab models, and I therefore choose for convenience of the potential reader in the same situation to present the derivations for two generic models outlined in Malsiner-Walli and Wagner (2011) that are used as foundations for all spike-and-slab analyses throughout this paper. These generic MCMC algorithms in Sections 3.2.2 and 3.2.5 are *not* my work, but I take credit for the derivations I present. The mixture prior I develop in Section 3.3 is my work alone, whereas I only modify the Bayesian variable selection procedure of the finite mixture model of linear regressions in Lee et al. (2016) slightly in Section 5. The two-step spike-and-slab prior for treatment effects in Section 4 also builds upon work by Wang et al. (2012) in another Bayesian context, but I take full credit for implementing the ideas within the spike-and-slab framework, which to my knowledge has not been done before. No analytical work has been carried out on the frequentist methods, and they are presented

solely with the purpose of comparison to the Bayesian methods. All code for the Bayesian models are my work alone.

Code

Throughout, the presented algorithms for the Bayesian models include hyperlinks to the code that are of course only accesible in the online version of this paper. All code is written in the object-oriented language R and can be accessed on Github via the following link

https://github.com/DitlevF/MasterThesis2018

An example code for running the models and evaluating them can be accessed *here*.

2 Machine Learning and Econometrics: A Brief Review

It is hard to come up with a precise definition of machine learning, as the term covers many subfields - such as computer science, engineering and statistics - and many applications in e.g. robotics, recommender systems, prediction tasks in industry, and now also causal inference. In this paper, machine learning is used synonymously with high-dimensional methods, that is, as statistical (learning) methods that are capable of analyzing and structuring high-dimensional data with p large relative to N. There are many possible applications of machine learning in social science research, and this thesis narrowly focuses on the use of machine learning to analyze high-dimensional models for inference on parameters, where the main focus on frequentist methods will be on the popular Lasso and its relatives. Other applications, for instance, include using satellite data and neural networks to analyze and measure agriculture and urban land use in development economics, see e.g. Donaldson and Storeygard (2016), but these are not covered here.

While econometrics is predominantly focused on inference, machine learning is predominantly focused on prediction. For instance, reiterating the example in the introduction, the Lasso provides a tool to avoid overfitting HDS models by shrinking coefficients towards zero, which provides strong predictions on new data, but leaves inference on the model parameters inoperative. Thus, to enable

inference in high-dimensional data with frequentist machine learning methods, one needs to "translate" the predictive toolbox from machine learning to inference. This paper is concerned with the use of spike-and-slab models to estimate regression functions and treatment parameters in HDS models, and in the following I briefly cover the state of the art and the most important developments in the econometrics literature that have fruitfully translated and utilized machine learning methods for these two tasks.

For treatment effects and causal inference, one of the earliest examples of the application of machine learning methods is the double-Lasso procedure (Belloni et al., 2014b) where the predictive relationship between the treatment variable and its confounders in (3) is leveraged to the Lasso that helps identify the important confounders. In Chernozhukov et al. (2017), they generalize this approach to incorporate more flexible machine learning methods such as random forests and neural networks to estimate this potentially non-linear nuisance function of confounders. Another prominent example in the literature is the causal trees (Athey and Imbens, 2016) that provides a data-driven means to estimate heterogeneous treatment effects by partitioning the covariate space via a splitting rule that reveals treatment effect heterogeneity using random forest estimation. Essentially, the estimator is a k-nearest-neighbors estimator that matches observations by their relevant confounding characteristics, where "relevant" is determined by the splitting rule at each leaf in the tree. A final interesting development is that of using deep learning for IV-estimation (Hartford et al., 2016), which leverages the powerful prediction tool to the first-stage task of purely predicting the endogenous variable. Belloni et al. (2012) also consider this task using the Lasso in place of deep learning to select the relevant instruments in the first stage, which Belloni and Chernozhukov (2011) demonstrate can be used to select among 180 potential instruments in the famous study of Angrist and Keueger (1991) where quarter-of-birth is originally used as an instrument for education.

Apart from treatment effects estimation, machine learning has also been used to flexibly estimate the regression function in economics. A practical application is in Dubé and Misra (2017) where interest lies in estimating a regression function to enable inference on an elasticity. Here, the authors consider a price experiment done by an online job-search platform firm, Ziprecruiter.com. In the experiment, $7,867$ prices (the monthly subscription fee to be on the platform) were randomly allocated

to prospective customers (the firms paying to be on the platform) with prices ranging from \$19 to \$399. During the experiment they obtained firm characteristics, such as company type or geographical location of the job, and the decision to buy/not buy at the given price which they used to identify the willingness-to-pay as a function of firm characteristics. To capture as much heterogeneity in the firms' willingness-to-pay, they expanded the set of firm characteristics by interactions and quadratic transformations of the original covariate set, and they used this high-dimensional set of characteristics to estimate a structural discrete-choice demand model via a logistic Lasso that identified a sparse set of important characteristics. Having identified the predictive relationship between firm characteristics and the buy/not buy decision at the given prices, they were able to identify the optimal price by using the predicted own-price elasticities, which led them to raise Ziprecruiter's price from \$99 before the experiment to \$249 after. A simulation study on this particular experiment in the seminar paper Frisch and Melchiorsen (2018) confirmed the strong performance of this machine learning approach to identify the elasticities and optimal prices, and they furthermore showed how using a standard logistic regression model instead would lead to very poor estimates with a large variance. This highlights how well inference on some parameters - in this case, the own-price elasticity - can be significantly improved upon by leveraging the strong predictive power of machine learning methods to estimate a high-dimensional sparse regression function.

So, what about Bayes? Although the Bayesian spike-and-slab models provide a natural way to represent sparsity in HDS models that enables simultaneous variable selection and parameter estimation, their applications in econometrics are sparse. To my knowledge, the only application in econometrics is in Scott and Varian (2015), where the spike-and-slab model was used to select predictors of consumer demand in a time-series model. With frequentist machine learning methods successfully flooding into econometric analysis of HDS linear models, it therefore makes sense to ask whether the spike-and-slab models can offer the same, if not better, performance than their frequentist alternatives. I start by looking at the standard linear regression case, where interest lies in estimating a high-dimensional sparse regression function.

3 Variable Selection and Inference in a High-Dimensional Sparse Linear Regression Model

This section considers modelling the outcome variables $y = (y_1, ..., y_N)$ as a linear function of the covariates and a Gaussian iid error term in a HDS linear regression model

$$y_i = Z_i\mu + X_i\beta + \epsilon_i \; \epsilon_i \sim \mathrm{N}(0, \sigma^2), \; i = 1, ..., N \tag{5}$$

where $Z_i \in \mathbb{R}^l$ and $X_i \in \mathbb{R}^p$ are two vectors of regressors with p large (and potentially $p \gg N$), ϵ_i are iid disturbances, Z_i represents the set of regressors that the analyst always wants to include in the model, e.g. an intercept term, and β is the sparse regression coefficient vector with only $p_0 \ll N$ non-zero entries. For ease of derivations later and without loss of generality, the design matrix $X = (X_1, ..., X_N)' \in \mathbb{R}^{N \times p}$ is assumed centered such that $X'\mathbf{1}_n = \mathbf{0}$.

3.1 Frequentist Approach: The Penalized Likelihood Fix with the Lasso

In the frequentist literature, linear regression models of the form in (5) are traditionally fit by OLS. Let (5) be written in the traditional matrix form $Y = X\beta + \epsilon, \epsilon \sim N(\mathbf{0}, I\sigma^2)$ where $X \in \mathbb{R}^{N \times (p+1)}$ with the first column being a vector of ones, then the OLS estimator minimizes the sum of squared residuals

$$\arg\min_{\beta} (Y - X\beta)'(Y - X\beta) \tag{6}$$

When $p \leq N$, $(X'X)^{-1}$ exists, and (6) has a unique solution, $\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$, with variance $\mathrm{Var}(\hat{\beta}_{OLS}) = \sigma^2(X'X)^{-1}$ and zero bias, and when $p \ll N$ the estimator tends to perform quite well with a low variance. However, the performance of the estimator is vulnerable to the problem of multicollinearity that causes $(X'X)^{-1}$ and hence the variance of the estimator to explode, which is naturally an increasingly severe issue as $p \to N$. This, together with the failure of the estimator to exist when $p > N$, has led to the development of alternative fitting procedures for high-dimensional data that aims to reduce the variance and thus also the mean squared error, which oftentimes involves

removing zero coefficient regressors from the model. Commonly used practices in the frequentist literature for removing variables have been forward or backwards subset selection, where in each step, either a variable is included or excluded from the model, and finally a selection among all models at each step is performed by comparing e.g. the AIC, BIC, cross-validated test error or the adjusted R^2 criterion. However, computationally, these methods do not scale very well in p, which has led to the development of other approaches such as the penalized likelihood approach. Perhaps the most successful of these is the Lasso (Tibshirani, 1996), which has sparked an enormous amount of further research within the machine learning literature on penalized likelihoods for variable selection (e.g. Fan and Li (2001), Zou (2006), Ročková and George (2018), Yuan and Lin (2006)), and has now, with some delay, also surfaced within the economics profession in Belloni et al. (2013), Belloni et al. (2014a) and Belloni et al. (2014b) as post-Lasso and double-Lasso estimators. The Lasso is a special case of the more general linear penalized likelihood framework of the form

$$\arg\min_{\beta} \frac{1}{2}(Y - X\beta)'(Y - X\beta) + \sum_{j=1}^{p} p_\lambda(\beta_j) \tag{7}$$

where the last term is introduced to regularize the coefficient vector towards zero, which reduces the variance of the estimator. The Lasso employs $p_\lambda(\beta_j) = \lambda|\beta_j|$, and other popular cost functions include $p_\lambda(\beta_j) = \lambda\beta_j^2$ (ridge, Hoerl and Kennard (1970)) and $p_\lambda(\beta_j) = \lambda_1|\beta_j| + \lambda_2\beta_j^2$ (elastic net, Zou and Hastie (2005)). The parameter λ controls the complexity of the model, where a larger λ is associated with a sparser model. Typically, this parameter is estimated via 10-fold cross-validation, but recently Belloni et al. (2014a) provided a theoretically justified value that can be estimated from the data when inference, rather than prediction, is the goal.

What caused the Lasso to become so popular is first and foremost due to its penalty function's kink at zero that forces some coefficients to be set exactly to zero. That is, at the Lasso solution $\hat{\beta}^{Lasso}$ of (7), one obtains $\hat{\beta}_j^{Lasso} = 0$ for several components $j = 1, ..., p$. This feature, coupled with its convex loss function that allows for very fast optimization enables a straightforward, semi-automatic and feasible variable selection procedure. Indeed, the Lasso is the only procedure with a penalty term of the form $\lambda|\beta|^q$, $q > 0$ that both induces sparsity (by setting some coefficients exactly to zero) and

is convex, which with the development of the least-angle regression algorithm (LARS) for solving the Lasso that scales very well in p has made the model very popular in both industry and academia.

However, the Lasso suffers from two main drawbacks when used in high-dimensional sparse econometric problems where inference is the goal. First, a sufficient amount of sparsity is needed for the Lasso's fit to converge to the Oracle OLS accuracy, see e.g. Bühlmann and Van De Geer (2011) pp. 99-104, where p_0 must be of smaller order than $\sqrt{N/\log(p)}$ for the Lasso to be consistent, where the Oracle OLS estimator is the OLS estimator applied to the true set of active regressors only (which is naturally only implementable in simulation studies). Lack of consistency in turn can cause a severe omitted variable bias if the true set of regressors is not identified. The second drawback is that all estimates are biased towards zero induced by the penalty term in (7), which is an undesirable property when inference is the goal (as opposed to prediction). To tackle the second problem, the post-Lasso estimator has been proposed that applies OLS in the second stage to the set of variables selected by the Lasso in the first stage, which Belloni et al. (2013) show performs better than the Lasso in terms of rates of convergence and bias.

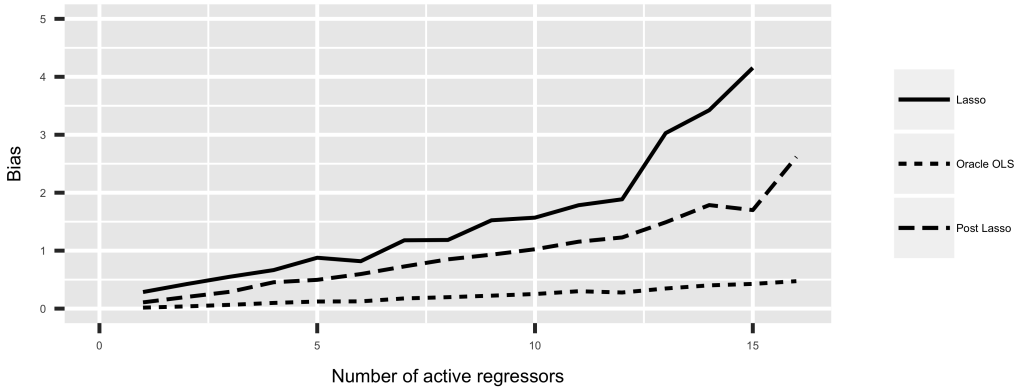
Figure 3.1 illustrates these two caveats with a simple example in a HDS linear regression model

$$Y = X\beta + \epsilon, \text{ with } N = 100,\ p = 100,\ p_0 = 1, ..., 16,\ X \sim N(\mathbf{0}, I),\ \epsilon_i \overset{\text{ind}}{\sim} N(0, 1)$$

where $\beta_j = 5$ if the regressor is active, zero otherwise. The theoretically justified value of λ in Belloni et al. (2014a) is used as tuning parameter. The figure plots the ℓ_2 norm of the bias $\| E(\hat{\beta}) - \beta \|_2$ averaged over 100 data sets for the 16 models with different number of active variables. Clearly, the Lasso performs worse than the Oracle OLS and post-Lasso in all cases because of the bias stained to its estimates. This is the second drawback, which the post-Lasso alleviates. The post-Lasso is competitive to the Oracle OLS for $p_0 = 1, ..., 15$, but then degrades into a much higher bias. This is the first drawback, where the sparsity assumption breaks down.

Albeit these shortcomings, the Lasso, post-Lasso and double-Lasso (the latter to be thoroughly introduced in Section 4) estimators have all turned out very powerful in obtaining both good predictions

Figure 3.1: Lasso, Post Lasso and Oracle OLS



Note - This figure displays the bias as a function of the size of p_0, i.e. the number of non-zero coefficient regressors. Source code can be found *here.*

of the outcome variable and robust inference on the parameters in HDS linear models as long as p_0 is sufficiently small. Therefore, they provide a useful benchmark to the Bayesian spike-and-slab approach to estimating parameters in HDS models that are covered in the coming subsection.

3.2 Bayesian Approach: Spike-and-Slab Priors

It is assumed that the reader is familiar with the essentials of the Bayesian paradigm, where parameters are treated as random quantities, and point estimates of parameters are replaced by distributions that represent the posterior belief about the value of the parameters. Within this paradigm, the spike-and-slab models are probably the most widely used models to analyze high-dimensional data, as they provide simultaneous variable selection and estimation of the parameters averaged over different models, so-called Bayesian model averaging (BMA). In this subsection, the general concept of sparsity in the Bayesian paradigm is first introduced, which is then followed by a thorough description and derivation of MCMC schemes for two popular classes of spike-and-slabs models in Sections 3.2.2 and 3.2.5. Subsequently, a novel prior in the spike-and-slab hierarchy is developed in 3.3 that is intended to improve variable selection.

3.2.1 Bayesian Model Selection and Inference: An Overview

In a Bayesian formulation, the observational model in (5) is interpreted as a conditional distribution

$$y \mid \mu, X, \beta, \sigma^2 \sim N(\mu \mathbf{1}_N + X\beta, \sigma^2 \mathbf{1}_N) \tag{8}$$

where $y = (y_1, ..., y_N)$ is an N-dimensional vector of outcome variables, X is an $N \times p$ matrix of covariates, β is a p-dimensional vector of regression coefficients with $p_0 < N$ non-zero entries, σ^2 is the variance of the Gaussian error term common to all models, and I set $Z = (Z_1, ..., Z_N) = \mathbf{1}_N$ for simplicity. I assume throughout the uninformative Jeffreys prior on the variance and mean terms

$$p(\mu, \sigma^2) \propto \frac{1}{\sigma^2} \tag{9}$$

In contrast to the frequentist penalized likelihood approach, the Bayesian approach to variable selection is conducted by classifying coefficients to be either zero or non-zero rather than shrinking them to zero. Thus, variables are selected based on posterior information of hierarchical models, which starts with indexing the 2^p models $\gamma_m = (\gamma_{1m}, ..., \gamma_{pm})$ with $\gamma_{jm} \in \{0, 1\}$, assign a prior probability of the model $p(\gamma_m)$, a prior on the regression coefficients conditional on the model $p(\beta_m \mid \gamma = \gamma_m)$, where $\beta_m = (\beta_{1m}, ..., \beta_{pm})'$, and ends with specifying the likelihood $p(y \mid \beta_m, \gamma_m)$. For variable selection, an intuitive Bayesian strategy is to calculate the individual posterior model probabilities

$$p(\gamma_m \mid y) = \frac{p(y \mid \gamma_m)p(\gamma_m)}{\sum_m p(y \mid \gamma_m)p(\gamma)}$$

where the marginal likelihood is

$$p(y \mid \gamma_m) = \int p(y \mid \gamma_j, \beta_m)p(\beta_m \mid \gamma_m)d\beta_m$$

and then select the model with highest posterior probability. However, as p increases, it becomes infeasible to explore the entire set of models, which is a problem similar to the frequentist forward and backwards selection. One approach to search for the best model is the MCMC model composition

MC^3 (Madigan et al., 1995), which visits new models $p(\gamma^* \mid y)$ via a Metropolis-Hasting step that causes the algorithm to visit only the most probable models. However, the MC^3 method does not provide parameter estimates of the model which is then left for post-estimation. Post-estimation of the model with the highest posterior probability in turn does not incorporate model selection uncertainty, and if one instead considers estimates averaged over all models weighted by their posterior probability, then all coefficients will be non-zero leading to a non-sparse solution. This means that one cannot obtain both model averaging estimates and variable selection via the MC^3 method.

The spike-and-slab framework, on the other hand, incorporates the updating of the conditional prior $p(\beta_m \mid \gamma = \gamma_m)$ to a posterior throughout the MCMC, and hence offers a simultaneous exploration of the model space, variable selection and parameter estimation averaged over the visited models, which makes it an appealing framework for inference in high dimensions. The priors on the regression coefficients, the spike-and-slab priors, are two-group-mixture-priors that cause weak coefficients to be pulled towards the group whose probability mass is concentrated at zero. Like the general Bayesian framework described above, the generic form of the spike-and-slab prior starts by indexing each of the 2^p possible models by

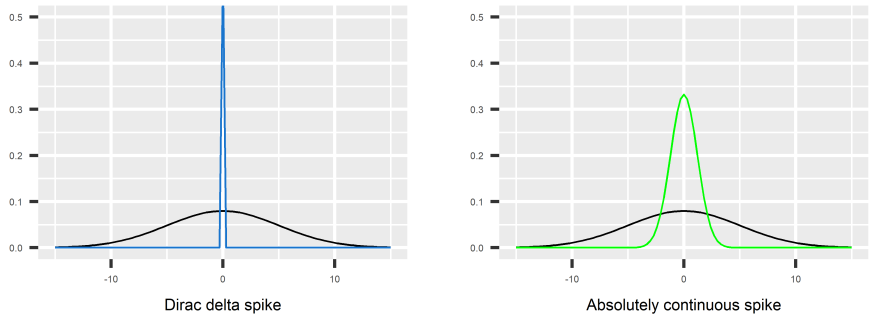
$$\gamma = (\gamma_1, ..., \gamma_p) \tag{10}$$

where $\gamma_j \in \{0, 1\}$ indicates whether the regression coefficient belongs to the spike ($\beta_j = 0$) or slab ($\beta_j \neq 0$) component. Conditional on γ, the prior for the regression coefficients is given hierarchically as

$$\beta_j \mid \gamma_j \sim \gamma_j \psi_{\text{slab}} + (1 - \gamma_j)\psi_{\text{spike}}, \qquad \gamma_j \sim \pi(\gamma_j) \tag{11}$$

where $\pi(\gamma_j)$ is commonly taken to be independent Bernoulli. The initial formulation of (11) in Mitchell and Beauchamp (1988) employed a uniform slab and a Dirac delta function centered at zero as the spike, while the advent of MCMC schemes for posterior exploration has replaced the use of a uniform slab with a normal distribution, and for some applications the spike is also replaced with an absolutely continuous spike that has a (much) narrower variance than the slab, see Figure 3.2 for a comparison between the two. Notably, George and McCulloch (1993) popularized the spike-and-slab framework when they introduced the stochastic search variable selection (SSVS) procedure that allows for fast

Figure 3.2: Dirac Delta Spike and Absolutely Continuous Spike



Note - The figure displays the two types of spike-and-slab models with different spike distributions: One is with a Dirac delta spike (left panel) and the other is with an absolutely continuous spike (right panel), which is a N(0,1) distribution (green line) in this example. The blacklined distribution corresponds to the slab, which for both cases here is a N(0,5) distribution.

Gibbs sampling in the spike-and-slab framework using a scale mixture of two normal distributions

$$\beta_j \mid \gamma_j \sim \gamma_j N(0, c_j^2 \tau_j^2) + (1 - \gamma_j) N(0, \tau_j^2),\ \gamma_j \sim \pi(\gamma_j) \tag{12}$$

with $\pi(\gamma_j)$ being Bernoulli, and where $c_j > 1$ is chosen a priori large enough and $\tau_j > 0$ small enough such that non-zero coefficients have posterior latent variables $\gamma_j = 1$. An example with $\tau_j^2 = 1$ and $c_j^2 = 5$ is given in Figure 3.2, right panel. The SSVS formulation in (12) is insightful, as it reveals the generality of the spike-and-slab framework for other Bayesian sparsity models: for instance, setting $\gamma_j = 1$ and $\tau_j = \tau$ for all j yields the Bayesian Ridge estimator; and letting $c_j^2 \sim \text{Exp}(\frac{\lambda^2}{2})$ yields the popular Bayesian Lasso (Park and Casella, 2008). While β_j is never set to be exactly zero in the formulation in (12) (unlike with the Dirac delta spike in (11)), variable selection can be interpreted from the posterior distribution of γ_j, where $\gamma_j = 0$ suggests that the variable is zero or practically zero.

More recently, Malsiner-Walli and Wagner (2011) compared different MCMC schemes for both absolutely continuous spikes and Dirac delta spikes coupled with different conjugate normal slabs of the form $N(\mathbf{a}_\gamma, \mathbf{A}_\gamma \sigma^2)$, where the subscript γ collects the active columns j for which $\gamma_j = 1$[1]. Their

[1]In Bayesian analysis, the probability density function $f(Z)$ is said to be conjugate of the likelihood $f(X \mid Z)$ if the

simulation studies suggest that both the choice of the slab hyperpriors, $\boldsymbol{a}_\gamma$ and $\boldsymbol{A}_\gamma$, as well as the choice of the form of the spike do not affect the variable selection to a large extent for independent regressors. For correlated regressors, however, the choice has significant influence on whether one is interested in avoiding false positives (including zero coefficients) or false negatives (excluding non-zero coefficients): if one is interested in avoiding false negatives, then slabs like the independence slab $N(\mathbf{0}, cI\sigma^2)$ should be employed, whereas when interest lies in avoiding false positives, the g-slab $N(\mathbf{0}, g(X'X)^{-1}\sigma^2)$ is a proper choice.

To implement the spike-and-slab models in my analyses, I thoroughly describe and derive the rather generic MCMC scheme in Malsiner-Walli and Wagner (2011) with a Dirac delta spike and $N(\mathbf{a}_{0,\gamma}, \mathbf{A}_{0,\gamma}\sigma^2)$ slab as well as the absolutely continuous version in the following. Much of the intuition carries over to other applications than the standard linear regression model, and will thus also be used when investigating a causal setting where the goal is to select the right set of confounders, and the finite mixture model of linear regressions in Lee et al. (2016). Although mentioned in the introduction, it should be stressed again: the models in Sections 3.2.2 and 3.2.5 are the work of Malsiner-Walli and Wagner (2011), but the derivations in Section 3.2.3 and Appendix 9.2 and 9.3, the computational note in Section 3.2.4, the code, and the posterior analysis in Section 3.2.6 are my work.

3.2.2 MCMC with a Dirac Delta Spike

To complete the model specification for the spike-and-slab model with a Dirac delta spike in (8), (9) and (11), priors on the remaining parameters are given as

$$p(\gamma \mid \omega = (\omega_1, ..., \omega_p)) = \prod_{j=1}^{p} \gamma_j^{\omega_j}(1-\gamma_j)^{1-\omega_j} \tag{13}$$

$$\omega_j \sim \pi(\omega_j) \tag{14}$$

$$p_{spike}(\beta_j) = p(\beta_j \mid \gamma_j = 0) = \Delta_0(\beta_j) \tag{15}$$

$$p_{slab}(\beta_\gamma) = N(\mathbf{a}_{0,\gamma}, \mathbf{A}_{0,\gamma}\sigma^2) \tag{16}$$

posterior density $f(Z \mid X)$ has the same form as the prior density function $f(Z)$.

where the prior on γ in (13) is iid Bernoulli, which is a common choice for binary indicator variables, $\Delta_0(\beta_j)$ in (15) is the Dirac delta function centered at zero, and the slab in (16) is the conjugate slab commonly used in the literature. The prior in (14) is left unspecified for now, but a common choice is simply to set $\omega_j = k \in [0, 1]$ for all j, where k is chosen to reflect the prior share of included regressors in the model, or the conjugate choice $\omega_j \sim \text{Beta}(a_\omega, b_\omega)$. The priors in the model are formulated such that one can formulate the full conditional distributions for each component, which makes a Gibbs sampling scheme the most straightforward choice. However, as noted in George and McCulloch (1997), the spike in (15) generates reducible, and thus non-convergent, Markov chains when successively drawing from the full conditionals in a Gibbs sampling scheme, because $\gamma_j = 0 \Leftrightarrow \beta_j = 0$ (unlike the SSVS in (12) where β_j is always drawn from a continuous distribution). Effectively, the Markov chain gets stuck when it generates $\beta_j = 0$, and hence γ_j cannot be updated by conditioning on β_j.

As a fix, Geweke (1996) proposed to jointly draw (β_j, γ_j) conditioning on σ^2 and the non-j elements (indicated by $\setminus j$) of the two coefficient and indicator vectors $(\beta_{\setminus j}, \gamma_{\setminus j})$, which comes at the cost of high serial correlation in the MCMC when the degree of multicollinearity is high. Another approach that avoids this, which will be employed here, is to update via the marginal posterior as proposed in Smith and Kohn (1996)

$$p(\gamma \mid y) \propto p(y \mid \gamma)p(\gamma) \tag{17}$$

where

$$p(y \mid \gamma) = \int_\mu \int_{\sigma^2} \left\{ \int_{\beta_\gamma} p(y \mid \beta_\gamma, \mu, \sigma^2) p(\beta_\gamma \mid \sigma^2) d\beta_\gamma \right\} p(\sigma^2, \mu) d\sigma^2 d\mu \tag{18}$$

As γ can take on 2^p different values, it is computationally infeasible to calculate its posterior unless p is small. Instead, one can explore the parameter space by successively drawing from $p(\gamma_j \mid y, \gamma_{\setminus j})$ in a random permutation order, $j = 1, ..., p$, that will converge to the joint distribution of $p(\gamma \mid y)$ after a sufficient burn-in period, see Gelfand and Smith (1990). This only demands evaluating $2p$ values in each iteration of the MCMC.

Below, I outline the calculation of the conditional posteriors for the MCMC. For ease of notation in the following section, let the following variables and hyperpriors be conditioned on inclusion, i.e. $\theta \equiv \theta_\gamma$ for $\theta = (y, X, \beta, \mathbf{a}_0, \mathbf{A_0})$.

3.2.3 Posterior Sampling Distributions

Marginal likelihood $p(y \mid X)$

The derivations of the likelihood and the marginal likelihood below are moved to Appendix 9.2, and the results are stated here. Using that $X'\mathbf{1}_N = \mathbf{0}$, the likelihood function can be written as (see the Appendix)

$$\mathcal{L}(\sigma^2, \beta, \mu, \mid y, X) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[(y_c - X\beta)'(y_c - X\beta) + N(\bar{y} - \mu)^2\right]\right\} \tag{19}$$

where $y_c = y - \bar{y}$ with $\bar{y}$ being the mean of the vector y. Given the hierarchy described above, the marginal likelihood is then given as

$$p(y \mid X) = \frac{1}{\sqrt{N}(2\pi)^{s_N}} \times \frac{\mid \mathbf{A}_N \mid^{\frac{1}{2}}}{\mid \mathbf{A}_0 \mid^{\frac{1}{2}}} \left(\frac{S_N^{s_N}}{\Gamma(s_N)}\right)^{-1} \tag{20}$$

where

$$\begin{aligned} S_N &= \frac{1}{2}\left[y_c'y_c + \mathbf{a}_0'\mathbf{A}_0^{-1}\mathbf{a}_0 - \mathbf{a}_N'\mathbf{A}_N^{-1}\mathbf{a}_N\right] \\ & s_N = (N-1)/2 \\ & \mathbf{A}_N = \left(X'X + \mathbf{A}_0^{-1}\right)^{-1} \\ & \mathbf{a}_N = \mathbf{A}_N\left(X'y_c + \mathbf{A}_0^{-1}\mathbf{a}_0\right) \end{aligned}$$

Conditional posterior of μ

As I employ the Jeffreys prior $p(\mu) \propto 1$, the conditional posterior is proportional to the likelihood (19)

$$p(\mu \mid y, \sigma^2) \propto \exp\{-\frac{1}{2\sigma^2}N(\mu - \bar{y})^2\} \Rightarrow \mu \mid y, \sigma^2 \sim N(\bar{y}, \sigma^2/N) \tag{21}$$

Conditional posterior of σ^2

We can sample σ^2 in one block with μ and γ as we calculate the marginal likelihood in each MCMC iteration, which is done by drawing σ^2 from its marginal posterior. From (60) in Appendix 9.2, we get

$$p(\sigma^2 \mid y, X) \propto p(y \mid \sigma^2, X)p(\sigma^2) \propto \frac{1}{\sqrt{N}(2\pi\sigma^2)^{\frac{N-1}{2}}} \times \frac{\mid \mathbf{A}_N \mid^{\frac{1}{2}}}{\mid \mathbf{A}_0 \mid^{\frac{1}{2}}} \exp\left(-\frac{S_N}{\sigma^2}\right)(\sigma^2)^{-1} \propto \exp\left(-\frac{S_N}{\sigma^2}\right)(\sigma^2)^{-s_N-1}$$

which is the kernel of an inverse-gamma distribution, thus

$$\sigma^2 \mid y, X \sim \text{IG}(s_N, S_N) \tag{22}$$

Marginal posterior of γ

Using the updating scheme as in Gelfand and Smith (1990), for $j = 1, ..., p$, one obtains

$$\Pr(\gamma_j = 1 \mid \gamma_{\backslash\gamma_j}, y, \omega_j) = \frac{p(y \mid \gamma_j = 1, \gamma_{\backslash\gamma_j})p(\gamma_j = 1, \gamma_{\backslash\gamma_j} \mid \omega_j)}{p(y \mid \gamma_j = 1, \gamma_{\backslash\gamma_j})p(\gamma_j = 1, \gamma_{\backslash\gamma_j} \mid \omega_j) + p(y \mid \gamma_j = 0, \gamma_{\backslash\gamma_j})p(\gamma_j = 0, \gamma_{\backslash\gamma_j} \mid \omega_j)} = \frac{p(y \mid \gamma_j = 1, \gamma_{\backslash\gamma_j})\omega_j}{p(y \mid \gamma_j = 1, \gamma_{\backslash\gamma_j})\omega_j + p(y \mid \gamma_j = 0, \gamma_{\backslash\gamma_j})(1-\omega_j)} = \frac{1}{1 + \frac{1-\omega_j}{\omega_j}R_j}, \; R_j = \frac{p(y \mid \gamma_j = 0, \gamma_{\backslash\gamma_j})}{p(y \mid \gamma_j = 1, \gamma_{\backslash\gamma_j})} \tag{23}$$

which can be drawn in random permutation order. Note that using a common hyperprior inclusion probabilty $\omega_j = \omega$ for all j results in the same posterior with ω_j replaced with ω.

Conditional posterior of β

With prior $p(\beta) = N(\mathbf{a}_0, \mathbf{A}_0\sigma^2)$, the posterior conditional distribution is

$$
\begin{aligned}
p(\beta \mid y, X, \sigma^2) &\propto p(y \mid X, \sigma^2, \beta) p(\beta \mid \sigma^2) \propto \\
&\exp\left(-\frac{1}{2\sigma^2}(y_c - X\beta)'(y_c - X\beta)\right) \times \exp\left(-\frac{1}{2\sigma^2}(\beta - \boldsymbol{a}_0)'(\mathbf{A}_0)^{-1}(\beta - \mathbf{a}_0)\right) \propto \\
&\exp\left(-\frac{1}{2\sigma^2}\left[-\beta' X' y_c - y_c' X'\beta + \beta' X' X\beta + \beta' \mathbf{A}_0^{-1}\beta - \beta'\mathbf{A}_0^{-1}\mathbf{a}_0 - \mathbf{a}_0\mathbf{A}_0^{-1}\beta\right]\right) \propto \\
&\exp\left(-\frac{1}{2\sigma^2}\left[-\beta'\left(X' y_c + \mathbf{A}_0^{-1}\mathbf{a}_0\right) - \left(y_c' X'\beta + \mathbf{a}_0\mathbf{A}_0^{-1}\right)\beta + \beta'\left(X'\mathbf{X} + \mathbf{A}_\mathbf{0}^{-\mathbf{1}}\right)\beta\right]\right) = \\
&\exp\left(-\frac{1}{2\sigma^2}\left[\beta'\underbrace{\left(\mathbf{X}'\mathbf{X} + \mathbf{A}_0^{-1}\right)}_{\mathbf{A}_N^{-1}}\beta - 2\beta'\underbrace{\left(X' y_c + \mathbf{A}_0^{-1}\mathbf{a}_0\right)}_{\mathbf{A}_N^{-1}\mathbf{a}_N}\right]\right)
\end{aligned}
$$

which is the kernel of a normal distribution, thus

$$\beta \mid y, X, \sigma^2 \sim N(\mathbf{a}_N, \mathbf{A}_N\sigma^2) \tag{24}$$

with

$$
\begin{aligned}
\mathbf{A}_N &= \left(X'X + \mathbf{A}_0^{-1}\right)^{-1} \\
\mathbf{a}_N &= \mathbf{A}_N\left(X' y_c + \mathbf{A}_0^{-1}\mathbf{a}_0\right)
\end{aligned}
$$

Conditional posterior of ω

One choice of prior is the common beta prior

$$\omega \sim \text{Beta}(a_\omega, b_\omega)$$

which leads to the following posterior update

$$p(\omega \mid \gamma) \propto p(\gamma \mid \omega)p(\omega) \propto p(\omega \mid \gamma) \propto \omega^{p_\gamma}(1-\omega)^{p-p_\gamma}\omega^{a_\omega-1}(1-\omega)^{b_w-1} = \omega^{p_\gamma+a_\omega-1}(1-\omega)^{p-p_\gamma+b_\omega-1} \Rightarrow$$
$$\omega \mid \gamma \sim \text{Beta}(p_\gamma + a_\omega, p - p_\gamma + b_\omega),\ p_\gamma = \sum_{j=1}^{p} \gamma_j \quad (25)$$

If instead we use an individual prior, $\omega_j \sim \text{Beta}(a_{\omega_j}, b_{\omega_j})$, we obtain by similar calculations

$$\omega_j \mid \gamma_j \sim \text{Beta}(\gamma_j + a_{\omega_j}, 1 + b_{\omega_j} - \gamma_j) \quad (26)$$

This individual prior can especially be useful in econometric settings, where the researcher wants to impose prior information of inclusion/exclusion on some regressors in the model.

3.2.4 The Sampling Scheme and a Computational Note

The above conditional posteriors suggest the sampling scheme presented in Algorithm 1. Note that due to marginalization, we can sample the posterior (γ, σ^2, μ) in one block from $p(\gamma, \sigma^2, \mu \mid y, X) = p(\mu \mid \sigma^2, y)p(\sigma^2 \mid y, \gamma)p(\gamma \mid y)$.

Although the posteriors are analytically straightforward to simulate from, Algorithm 1 is computationally burdensome, as the calculation of R_j in Step 1a requires the calculation of determinants of possibly high-dimensional matrices (see (20)) two times within each permutation, which yields $2p$ calculations of the marginal likelihood within each MCMC iteration. To increase computational efficiency, I start by noting that following this implementation, half of the values in the sequence of marginal likelihoods are recurrent. To fix ideas, let γ_j be the current permutation, and let γ_{j+q} denote the next permuted value that is initially set to $\gamma_{j+q} = l \in \{0, 1\}$, then if conditional on the set $\{X, \gamma_{\setminus\gamma_j,\gamma_{j+q}}, \gamma_{j+q} = l, \omega\}$ we sample $\gamma_j = k \in \{0, 1\}$, it is evident that

$$p(y \mid X, y, \gamma_{\setminus\gamma_j,\gamma_{j+q}}\gamma_j = k, \gamma_{j+q} = l) = p(y \mid X, y, \gamma_{\setminus\gamma_{j+q}}, \gamma_{j+q} = l)$$

wherefore we only need to calculate the new value $p(y \mid X, y, \gamma_{\setminus\gamma_{j+q}}, \gamma_{j+q} = 1 - l)$ to sample γ_{j+q} as

Algorithm 1 Spike-and-slab in a Linear Model with a Dirac Delta Spike. Example function code with $\omega_j \sim \text{Beta}(a_\omega, b_\omega)$ and independence slab *here.*

Initialize $\gamma^{(0)} = (\gamma_1, ..., \gamma_p)^{(0)}$ and $\omega^{(0)} = (\omega_1, ..., \omega_p)^{(0)}$

MCMC. At each iteration $t = 1, ..., T$, run through

1a) In a random permutation order, sample $\gamma_j^{(t)}$ from $p(\gamma_j \mid y, \gamma_{\setminus\gamma_j}^{(t,t-1)}, \omega^{(t-1)})$ for all $j = 1, ..., p$, see Eq. (23)

1b) Sample $(\sigma^2)^{(t)}$ from $p(\sigma^2 \mid y, X, \gamma^{(t)})$, see Eq. (22)

1c) Sample $\mu^{(t)}$ from $p(\mu \mid y, (\sigma^2)^{(t)})$, see Eq. (21)

2) Sample $\beta^{(t)}$ from $p(\beta \mid y, (\sigma^2)^{(t)}, \gamma^{(t)})$, see Eq. (24)

3) Sample $\omega^{(t)}$ from $p(\omega \mid \gamma^{(t)})$, e.g. from Eq. (25) or (26)

Note: Conditioning on γ implies that only the columns j of $X, \boldsymbol{A}_0, \boldsymbol{A}_N, \boldsymbol{a}_0, \boldsymbol{a}_N$ for which $\gamma_j = 1$ are used.

$p(y \mid X, y, \gamma_{\setminus\gamma_{j+q}}, \gamma_{j+q} = l)$ is already calculated. This lowers the $2p$ calculations to p at the minor expense of storing a scalar marginal likelihood value, which greatly increases computational efficiency. In fact, for most problems it reduces computational time on the order of $30 - 40\%$.

3.2.5 MCMC with an Absolutely Continuous Spike

While the traditional understanding of a spike-and-slab prior was a mixture with a Dirac delta spike, the SSVS interpretation in George and McCulloch (1993) and its normal-inverse gamma (NMIG) extension in Ishwaran et al. (2005) with a continuous spike has been increasingly used in recent research aimed at analyzing very high-dimensional data as the computational time scales better in p, see e.g. Ročková and George (2018), because calculation of the marginal likelihood is not required. Coupled

with the linear model in (5), both SSVS and NMIG can be represented as

$$\beta_j \mid \gamma_j \sim \gamma_j N(0, c_j^2\tau_j^2) + (1-\gamma_j)N(0,\tau_j^2),\ \tau_j^2 \sim p(\tau_j^2 \mid -) \tag{27}$$

$$p(\gamma \mid \omega = (\omega_1, ..., \omega_p)) = \prod_{j=1}^{p} \gamma_j^{\omega_j}(1-\gamma_j)^{1-\omega_j} \tag{28}$$

$$\omega_j \sim \pi(\omega_j) \tag{29}$$

where the SSVS prior is with fixed $\tau_j^2 = V$, and the NMIG prior is with $\tau_j^2 \sim \text{IG}(\nu, Q)$. The choice of τ_j^2 and c_j^2 has a large impact on determining whether β_j is interpreted to be (practically) zero, which motivated the NMIG alternative hierarchical formulation to allow the data to affect this decision. The impact of these two parameters on variable selection can be seen by the intersection point between the two Gaussian densities that is naturally increasing (in absolute value) in both τ_j and c_j. Derivations are similar in logic to the MCMC with a Dirac spike, and are hence put in Appendix 9.3. The sampling scheme is summarized in Algorithm 2.

Algorithm 2 Spike-and-slab in a Linear Model with an Absolutely Continuous Spike. Function code with NMIG prior *here.*

Initialize $\gamma^{(0)} = (\gamma_1, ..., \gamma_p)^{(0)}$, $\omega^{(0)} = (\omega_1, ..., \omega_p)^{(0)}$ and $\beta^{(0)} = (\beta_1, ..., \beta_p)'^{(0)}$

MCMC. At each iteration $t = 1, ..., T$, run through

1a) In a random permutation order, sample $\gamma_j^{(t)}$ from $p(\gamma_j \mid \beta_j^{(t-1)}, \gamma_{\setminus\gamma_j}^{(t,t-1)}, \omega^{(t-1)})$ for all $j = 1, ..., p$, see Eq. (62)

1b) Sample $(\tau_j^2)^{(t)}$ from $p(\tau_j^2 \mid \gamma_j^{(t)}, \beta_j^{(t-1)})$, see Eq. (63). If using the SSVS, simply set $\tau_j^2 = V$.

2) Sample $(\sigma^2)^{(t)}$ from $p(\sigma^2 \mid y, \beta^{(t-1)})$, see Eq. (65)

3) Sample $\mu^{(t)}$ from $p(\mu \mid y, (\sigma^2)^{(t)})$, see Eq. (21)

4) Sample $\beta^{(t)}$ from $p(\beta \mid y, (\sigma^2)^{(t)}, \gamma^{(t)}), (\tau^2)^{(t)})$, see Eq. (64)

5) Sample $\omega^{(t)}$ from $p(\omega \mid \gamma^{(t)})$, e.g. from Eq. (25) or (26)

Note: For the NMIG prior, marginally both spike and slab distributions are Student distributions, such that $p_{spike} = t_{2\nu}(0, c^2Q/\nu)$ and $p_{slab} = t_{2\nu}(0, Q/\nu)$.

3.2.6 Posterior Analysis: Variable Selection and Estimation

After having generated a sufficiently large sample from the target posterior distribution, one can evaluate the posteriors simply by the kernel density estimates or summarize the results via the posterior mean or mode. E.g., the posterior mean for the parameters of interest $\Theta = (\beta, \sigma^2, \mu)$ and γ can be evaluated by

$$\hat{\Theta} = \frac{1}{M-B} \sum_{m=M-B}^{M} \Theta^{(m)}, \qquad \hat{\gamma}_j = \frac{1}{M-B} \sum_{m=M-B}^{M} \gamma_j^{(m)},\ j = 1, ..., p \tag{30}$$

where M is the length of the Markov chain and B the burn-in period. The posterior results from this analysis are in fact averages over different models indexed by $\gamma^{(m)} = (\gamma_1^{(m)}, ..., \gamma_p^{(m)})$, whereby $\hat{\Theta}$ is a BMA estimate of the parameters over all 2^p models weighted according to the posterior probability of each model (some of which are likely not visited during the MCMC). The most appealing quality to this kind of estimate is that it incorporates model uncertainty in the estimation, and many prediction tasks are shown to benefit tremendously from averaging over an ensemble of models (Madigan and Raftery, 1994).

On the other hand, one may wish to select a parsimonious model and perform post-regression on the selected variables. Contrary to what one might think, Barbieri et al. (2004) showed that the best single-model approximation to model averaging under orthogonal and nested correlated designs is not the model with highest posterior probability

$$p(\gamma \mid y) = \sum_{m=M-B}^{M} \mathbf{1}(\gamma^{(m)} = \gamma)/M$$

but the median probability model that is defined as the model consisting of variables with marginal inclusion probabilities larger than 0.5. This can readily be gathered from the data by simply selecting the variables for which $\hat{\gamma}_j > 0.5$. The median probability model criterion is widely used in the literature for model selection, and I also base model selection on this criterion in the simulation studies. After selecting the variables to be included, post-estimation of the model parameters is then straightforward, but does not incorporate model uncertainty, which can lead to biased estimates (George and Foster,

2000). Thus, when model selection is the goal, the median probability model criterion is a good choice to employ, whereas when inference on parameters is the goal, there are plenty of good reasons to utilize the posterior BMA output from the spike-and-slab models.

Bias of the posterior mean estimator? A brief comment

As touched upon earlier, in the Bayesian paradigm, parameters are treated as random quantities and thus point estimates are replaced by posterior distribution. However, in the simulation study in Section 6 I compare the performance of the posterior mean (point) estimator and its maximum likelihood alternative to a "true value" using the bias as a metric to do so. The bias concept is inherently frequentist as it relates the mean of the estimator with the true value of the parameter, which philosophically makes little sense from a Bayesian point of view where random quantities obviously do not have true values. However, albeit being on the frequentist playing field, it is a useful metric to compare frequentist and Bayesian estimators, and it is therefore important to understand how the posterior mean estimator behaves in terms of this bias. For the Dirac Delta spike-and-slab prior, for instance, conditional on inclusion, the posterior mean estimator is given from (24)

$$E(\beta \mid y, X, \sigma^2) = (X'X + \mathbf{A}_0^{-1})^{-1}(X'y_c + \mathbf{A}_0^{-1}\mathbf{a}_0)$$

which is the unbiased OLS estimator as $A_0^{-1} \to \mathbf{0}_{p\times p}$ and $\mathbf{a}_0 \to \mathbf{0}$. When using slabs of the form $N(\mathbf{0}, \sigma^2\mathbf{A}_0)$, the estimator will therefore be slightly biased towards zero whenever $\mathbf{A}_0$ is chosen small. If the regression coefficients of interest are furthermore not fixed in the model, a larger bias towards zero occurs as they will be sampled from the spike part of the distribution as well during MCMC, which will especially happen for weak coefficients that are sampled many times from the spike. Although this is simply the prior influencing the posterior, this caveat should be kept in mind when I compare biases of the estimators in the simulation study, where I also choose $\mathbf{A}_0$ fairly large to ameliorate this effect.

3.3 A Mixture Prior on the Individual Inclusion Probabilities

There are many possible extensions to the spike-and-slab models, and most of them are concerned with developing new priors on the regression coefficients, e.g. the Horseshoe prior (Carvalho et al., 2010), and the Lasso prior (Ročková and George, 2018). However, the literature is rather silent on the choice of priors on the inclusion probabilities $\omega_j = p(\gamma_j = 1 \mid \omega_j)$, and it is standard to choose either a fixed prior $\omega_j = k \in [0, 1]$ for all j, a beta prior common to all inclusion indicators $\omega \sim \text{Beta}(a_\omega, b_\omega)$ or an individual beta prior $\omega_j \sim \text{Beta}(a_\omega, b_\omega)$. In this subsection, I take a step back and consider a new prior on ω_j to improve on variable selection. To understand how the prior influences variable selection, notice how γ_j is sampled during MCMC in the spike-and-slab model with a Dirac spike from (23)

$$\Pr(\gamma_j = 1 \mid \gamma_{\setminus\gamma_j}, y, \omega_j) = \frac{1}{1 + \frac{1-\omega_j}{\omega_j} R_j}, \; R_j = \frac{p(y \mid \gamma_j = 0, \gamma_{\setminus\gamma_j})}{p(y \mid \gamma_j = 1, \gamma_{\setminus\gamma_j})}$$

where a large value of ω_j naturally increases the probability of sampling $\gamma_j = 1$. Both the fixed prior and the common beta prior effectively place the conceived share of non-zero regression coefficients as the prior on all indicators, whereas the individual beta prior applies some inertia to the sampling as the probability of sampling $\gamma_j^{(t+1)} = 1$ is larger when $\gamma_j^{(t)} = 1$, independent of $\gamma_{\setminus\gamma_j}^{(t)}$. This makes it more likely to include weak coefficients into the model compared to the other two, which can be of interest if the goal lies in e.g. capturing weak confounders. However, for the individual beta prior, the values for the parameters of the posterior distribution $\omega_j \mid \gamma_j \sim \text{Beta}(\gamma_j + a_{\omega_j}, 1 + b_{\omega_j} - \gamma_j)$ are confined to either $(a_\omega + 1, b_\omega)$ or $(a_\omega, b_\omega + 1)$, which motivates a more flexible estimation of the posterior parameters that enables a steeper learning process from the data, thereby making a beta mixture distribution a natural choice

$$p(\omega_j \mid \pi, a, b) = \sum_{m=1}^{M} \pi_m \text{Beta}(\omega_j : a_m, b_m), \qquad \sum_{m=1}^{M} \pi_m = 1, \; \pi_m \geq 0 \text{ for all } m$$

where $(\pi, a, b) = ((\pi_1, a_1, b_1), ..., (\pi_M, a_M, b_M))$. The steeper learning process comes from the mixture distributions being able to flexibly model data, which allows the prior on ω_j to be updated throughout MCMC, such that, in theory, a strong regressor z will have its prior on ω_z move towards a component

m with much probability mass near 1, and a weak regressor x will have its prior ω_x move towards a component with probability mass near 0. When this happens, the posterior output will be less blurred as strong regressors are less likely to be sampled from the spike distribution, and zero coefficient regressors are less likely to be sampled from the slab distribution.

Before turning to the implementation of this prior, the concept of finite mixture models is briefly introduced. Both the general frequentist and Bayesian approach is introduced, as they are both also used for the case of finite mixture models of linear regressions in Section 5.

3.3.1 The Finite Mixture Model

A finite mixture model provides a way to generate a flexible form of the density of the data. Mathematically, the M-component finite mixture model can be represented as

$$p(x) = \sum_{m=1}^{M} \pi_m p_m(x \mid \theta_m), \qquad \sum_{m=1}^{M} \pi_m = 1,\ \pi_m \geq 0 \text{ for all } m \tag{31}$$

where p is the unknown density of x that we wish to approximate, $\pi = (\pi_1, ..., \pi_M)$ are the mixture weights, the mixture component p_m is the distribution for group m, and $\theta = (\theta_1, ..., \theta_M)$ the model parameters (for instance, $\theta_m = (\mu_m, \sigma_m^2)$ in a finite normal mixture model). For $1 < M < N$, the mixture model can be viewed as a semi-parametric compromise between a fully parametric model represented by a single component $M = 1$ and a non-parametric model represented by $M = N$, i.e. a kernel density estimate (McLachlan and Peel, 2000).

In their seminal work, Dempster et al. (1977) noted that mixture models of the form in (31) can be represented hierarchically by latent data that allows for feasible estimation of the model parameters, so-called data augmentation. Indeed, introducing the unobserved vector $z_i = (z_{i1}, ..., z_{iM})$, where $z_{im} \in \{0, 1\}$ denotes whether observation i belongs to group m, the likelihood of the "complete data" $\{x_i, z_i\}_{i=1}^N$ is

$$\mathcal{L}(\theta, \pi \mid x, z) \prod_{i=1}^{N} \sum_{m=1}^{M} z_{im} \pi_m p_m(x_i \mid \theta_m) \tag{32}$$

The frequentist approach to estimate this data-augmented model that was established in Dempster

et al. (1977) is to employ the expectation-maximization (EM) algorithm that iteratively estimates the expected value of all z_i given θ (the E-step) and maximizes (32) with z_{im} replaced by its E-step value (the M-step).

Following Tanner and Wong (1987), Bayesians, in a similar manner, simulate the latent data directly in an iterative scheme, where $\{z_i\}_{i=1}^N$ is sampled from $p(z_i \mid x, \theta)$ and θ is sampled from $p(\theta \mid z, x)$ until convergence. The most standard approach for estimating mixture models is the Gibbs sampler (Diebolt and Robert, 1994) using Dirichlet-Multinomial conjugate priors for π_m and z_{im} that successively simulates the parameters conditional on one another and the data. This is also the approach that will be applied for the beta mixture prior next.

3.3.2 Implementation of the Beta Mixture Prior

My Gibbs sampling approach following Diebolt and Robert (1994) for the beta mixture prior is based on successive simulation of $\pi = (\pi_1, ..., \pi_M)$, $z = (z_1, ..., z_p)$ and $\theta = \{a_m, b_m\}_{m=1}^M$. After initialization, the sampling scheme is, for $t = 1, ..., T$:

1. Generate $z_j^{(t)}$ from $p(z_j \mid \omega^{(t-1)}, \pi^{(t-1)}, \theta^{(t-1)})$
2. Generate $\pi^{(t)}$ from $p(\pi \mid z^{(t)})$
3. Generate $\theta^{(t)}$ from $p(\theta \mid z^{(t)}, \omega^{(t-1)})$
4. Generate $\omega^{(t)}$ from $p(\omega \mid z^{(t)}, \theta^{(t)}, \gamma)$

where $\omega = (\omega_1, ..., \omega_p)$ are the points to estimate, and as these are latent data, we also need to update them conditional on $\gamma = (\gamma_1, ..., \gamma_p)$ in the fourth step. Because the vector π is defined on the simplex

$$\{(\pi_1, ..., \pi_M) : \sum_{m=1}^{M-1} \pi_m < 1)\}$$

the natural choice for $p(\pi)$ is the Dirichlet distribution with parameters $(\alpha_1, ..., \alpha_M)$. Together with a

discrete prior on z, this gives the following posterior update in Step 2

$$p(\pi \mid z) \propto p(z \mid \pi)p(\pi) \propto \prod_{j=1}^{p} \pi_1^{\mathbf{1}(z_{j1}=1)} ... \pi_M^{\mathbf{1}(z_{jM}=1)} \times \pi_1^{\alpha_1 - 1} ... \pi_M^{\alpha_M - 1} \propto \pi_1^{n_1 + \alpha_1 - 1} ... \pi_M^{n_M + \alpha_M - 1} \Rightarrow$$

$$\pi \mid z \sim \text{Dirichlet}(n_1 + \alpha_1, ..., n_M + \alpha_M), \; n_m = \sum_{j=1}^{p} \mathbf{1}(z_{jm} = 1) \quad (33)$$

where $\alpha = \alpha_1, ..., \alpha_M$ are hyperpriors such that smaller values of α are associated with a high probability on models with only few number of components out of the total M components. It is standard to employ an uninformative prior $\alpha = (1, 1, ..., 1)$, which I also do.

The group membership vector in Step 1 can be sampled discretely via

$$\Pr(z_{mj} = 1 \mid \pi_m, \omega_j, a_m, b_m) \propto p(\omega_j \mid z_{mj} = 1)p(z_{mj} = 1 \mid \pi_m) \propto \pi_m \times \omega_j^{a_m - 1}(1 - \omega_j)^{b_m - 1} \quad (34)$$

and conditional on the membership vector, ω is sampled in Step 4 for $j = 1, ..., p$ via

$$p(\omega_j \mid \gamma, z_{mj} = 1) \propto p(\gamma \mid \omega_j, z_{mj} = 1)p(\omega_j \mid z_{mj} = 1) \propto \omega_j^{\gamma_j + a_m - 1}(1 - \omega_j)^{\gamma_j + b_m} \Rightarrow$$

$$\omega_j \mid \gamma, z_{mj} = 1 \sim \text{Beta}(a_m + \gamma_j, b_m + 1 - \gamma_j) \quad (35)$$

For Step 3, Bouguila et al. (2006) provide a Metropolis-Hastings step to sample θ from its conditional posterior, since the beta distribution does not have an analytically tractable conjugate distribution. I follow this approach as well as their tuning of the hyperparameters, which is described in detail in Appendix 9.4. The sampling steps are summarized in Algorithm 3.

The algorithm requires one to select the number of mixture components, where a straightforward choice is either two or three to allow the mixture components to have differing masses spread out over the $[0, 1]$ line. In simulation studies with $p = 40$, I have found three groups to have optimal performance in terms of variable selection and fit. The posterior of the mixture components for the three-component analysis typically stabilize and take forms like the ones displayed in Figure 3.3 that hence allow strong coefficient priors to move towards the first mixture (left panel) and weak coefficient priors towards the second mixture (middle panel). The right panel component is a stabilizing component that aids

Algorithm 3 Gibbs Sampler for Finite Beta Mixture Density on Prior Inclusion Probabilities

For use within a spike-and-slab model, the following sampling scheme is used to replace Step 3 in Algorithm 1 or Step 5 in Algorithm 2, where all other steps remain the same. Full code and implementation of the sampling can be viewed *here*.

Initialize $\boldsymbol{\pi}^{(0)} = (\pi_1, ..., \pi_K)^{(0)}$, $((a_1, b_1), ..., (a_M, b_M))^{(0)}$ and $\omega^{(0)} = (\omega_1, ..., \omega_p)$.

Gibbs Sampling. At each iteration $t = 1, ..., T$, run through

1) Sample $z_j^{(t)}$ from $p(z_j \mid \pi_m^{(t-1)}, \omega_j^{(t-1)})$ for $j = 1, ..., p$, and compute $n_m = \sum_{j=1}^{p} \mathbf{1}(z_{mj} = 1)$, see Eq. (34)

2) Sample $\pi^{(t)}$ from $p(\pi \mid z^{(t)})$, see Eq. (33)

3) Sample $\{a_m^{(t)}, b_m^{(t)}\}_{m=1}^{M}$ via the Metropolis-Hastings steps described in Appendix 9.4

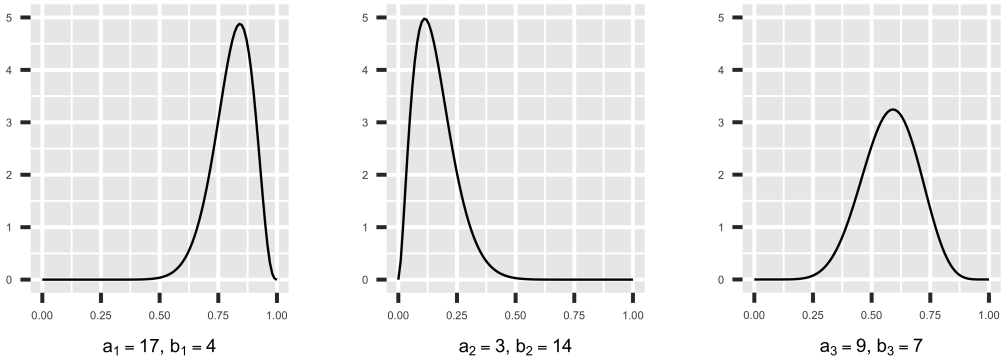
4) Sample $\omega_j^{(t)}$ from $p(\omega_j \mid \gamma, z_{mj}^{(t)} = 1)$ for $j = 1, ..., p$, see Eq. (35)

ω_j's to jump from belonging to one component to another - for instance, from the figure it is evident that if $\omega_j^{(t)} \approx 0.7$ then $\omega_j^{(t)}$ is as likely to be allocated to the left panel component or the right panel component if $\pi_1^{(t)} = \pi_3^{(t)}$. This stabilizing component is typically missing in the two-component case, which can explain why the three-component mixture seems to perform best.

One caveat of this prior is that (a_m, b_m) is unbounded, which means that when (a_m, b_m) grows large, the likelihood effect in Step 4 of Algorithm 3 diminishes. To counter this, it is possible instead to fix $\{a_m, b_m\}_{m=1}^{M}$ a priori and skip Step 3, but in my experience it does not improve either variable selection or inference of the parameters. It is naturally also possible to place priors on (a_m, b_m) that binds them to a narrow interval, but this has not been investigated yet.

Later, in the simulation studies, I investigate how this prior fares relative to the two usual choices - the common and individual beta prior - in a linear regression model, but attention is now turned to the development of another prior for treatment effects studies that takes the relationship between a treatment variable and its confounding variables into account.

Figure 3.3: Example of the Posterior Parameters of the Mixture Components



Note - This figure displays a representative example of the Beta(a_j, b_j) distributions corresponding to the parameters of the mixture components placed on ω_j that are sampled throughout MCMC.

4 High-Dimensional Inference on a Treatment Parameter

The previous section was concerned with the simultaneous estimation and variable selection in the linear regression model, which enables inference on the regression function in HDS models. The natural and well-studied choice for a Bayesian was argued to be the spike-and-slab model that, in addition to variable selection, provided robust BMA estimates of the model parameters, and the post-Lasso estimator (Belloni et al., 2013) was a good frequentist choice for attaining near-Oracle OLS performance in terms of convergence and bias under the assumption of sparsity. In treatment effects studies, where interest lies in capturing confounders, these methods can also be applied to conduct a data-driven search for the true set of confounders. However, when using the Lasso to select confounders in a treatment effects model, model selection mistakes may occur, which can severely bias inference of the object of interest. To tackle this, Belloni et al. (2014b) proposed a two-step procedure, the so-called double-Lasso, that drastically limits the possibility of these mistakes, which will be presented in Section 4.1. Because of the powerful performance of this estimator to estimate treatment effects, I propose a Bayesian interpretation of this method inspired by ideas from Wang et al. (2012) with a two-step spike-and-slab prior in Section 4.2.

4.1 The Double-Lasso

The double-Lasso is applicable to settings where interest is on causal inference on a model parameter with confounding variables. Consider the structural treatment effects HDS model from the introduction

$$y_i = \mu + \alpha_0 d_i + x_i \beta_y + \epsilon_i, \; i = 1, ..., N \tag{36}$$

$$d_i = x_i \beta_d + \nu_i \tag{37}$$

where d_i is a scalar continuous treatment variable, x_i is a $p \times 1$ vector of potential confounders with p large, ϵ_i and ν_i are iid disturbances, and α_0 is the parameter of interest. If $p \ll N$ and x_i includes all proper controls such that

$$E(d_i \epsilon_i \mid x_i) = 0 \tag{38}$$

then one can simply proceed by estimating (36) via e.g. OLS, and the parameter α_0 can be interpreted as the average treatment effect under few other conditions, see e.g. Imbens (2004).

However, as $p \to N$, the estimate of α_0 will have a high variance, and as p grows larger than N, standard methods like OLS are no longer applicable. Under the assumption of sparsity, namely that there exist approximations to the true regression function that require only $p_0 < N$ non-zero coefficients, Belloni et al. (2014b) develop a solution to these cases via the double-Lasso. The double-Lasso effectively provides a robust data-driven method to select the right set of controls and subsequently estimate the treatment effect α_0 given these controls via OLS. The algorithmic scheme is as follows

1. Fit a Lasso predicting the dependent variable y_i on all controls x_i. Retain the active (non-zero) set of predictors $x_i^{\mathrm{A},1}$

2. Fit a Lasso predicting the treatment parameter d_i on all controls x_i. Retain the active set of predictors $x_i^{\mathrm{A},2}$. If the experiment is randomized, this set should naturally be empty.

3. Fit a linear regression (i.e. post-Lasso) of y_i on d_i and the union of variables selected in Steps 1 and 2, $X_i^{\mathrm{A},1} \bigcup X_i^{\mathrm{A},2}$.

Without the second step, the estimator is simply the post-Lasso estimator with d_i fixed in the model.

The intuition for introducing the second step is to avoid excluding important confounders in Step 1 that in turn could lead to a substantial omitted-variable bias. For instance, if some β_{jy} is small, then the Lasso in Step 1 risks excluding this from the model, and if β_{jd} in turn is large, this would lead to a substantial omitted-variable bias, which motivates including Step 2. The same intuition carries over if one considered only using Steps 2 and 3 for estimation, where the Lasso would risk excluding a small β_{jd} that could lead to a large bias if β_{jy} is large.

One crucial parameter to be specified in the above scheme is the penalty parameter λ, which controls the complexity of the model. For optimal predictive performance, 10-fold cross-validation is a common and useful choice to estimate λ, but Belloni et al. (2014b) also develop a theoretically grounded value that can be estimated from the data that takes potential heteroskedasticity and the goal of inference into account. This can be employed using their R-package hdm, but in practice there is negligible difference between the two when the sparsity assumption is fulfilled.

4.2 Two-step Spike-and-Slab

For Bayesians, a natural choice to estimate α_0 in (36) is to place spike-and-slab priors on β_y as described in Section 3.2 with e.g. a mixture prior on ω, fix d_i in the model, and use the posterior distribution of α_0 to draw inference on the effect of d_i on y_i. If the MCMC draws from the posterior explores and identifies the true model, this approach is perfectly valid. However, like the Lasso, spike-and-slab models also face the issue of risking to omit weak regressors in the outcome equation (36) that can lead to a severe omitted-variable bias if these are not weak in the selection equation (37). Thus, for robust inference, it is crucial to incorporate information from the selection equation when estimating α_0, which motivates incorporating information of β_d into the inference for β_y via a joint prior.

4.2.1 A Two-step Prior

Wang et al. (2012) provided an intuitive way to incorporate such information from the selection equation into the outcome model to avoid missing important confounders. Let γ^y denote the indicator variables for inclusion in (36) and γ^d the indicator variables for inclusion in (37), then they propose a

conditional prior distribution of $\gamma^y \mid \gamma^d$ that can be specified as

$$\frac{\Pr(\gamma_j^y = 1 \mid \gamma_j^d = 1)}{\Pr(\gamma_j^y = 0 \mid \gamma_j^d = 1)} = \zeta, \qquad \frac{\Pr(\gamma_j^y = 1 \mid \gamma_j^d = 0)}{\Pr(\gamma_j^y = 0 \mid \gamma_j^d = 0)} = \phi, \; j = 1, ..., p \tag{39}$$

where $\zeta \in [1, \infty]$ is a dependence parameter denoting the odds of including x_j in the outcome model (36) when x_j is included in the selection model (37). As $\zeta \to \infty$, all covariates selected in the selection equation are automatically included in the outcome model. $\phi \in [0, 1]$ denotes the odds of including x_j if it is not included in the selection model. Wang et al. (2012) introduced this prior in another context, where the MC^3 method (Madigan et al., 1995) is used for posterior exploration, and they proposed both a one-step and two-step procedure. In order to incorporate the prior in (39) into the spike-and-slab framework, I adopt ideas from their two-step procedure with two separate MCMCs.

The first stage first calculates $p(\gamma^d \mid d)$ by applying the spike-and-slab Algorithm 1 or 2 to the selection model (37) with e.g. a mixture prior on γ^d, and then updates an integrated prior of $p(\gamma^y \mid d)$ as

$$p(\gamma^y \mid d) = \sum_{\gamma^d} p(\gamma^y, \gamma^d \mid d) = \sum_{\gamma^d} p(\gamma^y \mid \gamma^d, d) p(\gamma^d \mid d) = \sum_{\gamma^d} p(\gamma^y \mid \gamma^d) p(\gamma^d \mid d) \tag{40}$$

where the third equality arises from an assumption of conditional independence of d and γ^y given γ^d, i.e. that the outcome model does not contain any additional information on d given the selection model. As interest in the first stage is only on γ^d, one only needs to run Step 1a in Algorithm 1 if one were to use a Dirac delta spike. Then, the second stage again applies Algorithm 1 or 2 to the outcome model (36) with d fixed and with $p(\gamma^y \mid d)$ used as an updated prior on γ^y.

4.2.2 Implementation

While theoretically appealing, this approach is computationally intensive as it first off requires an additional MCMC run to count the frequencies of each γ^d from $p(\gamma^d \mid d)$, and secondly it subsequently requires the density evaluation of all 2^p permutations of γ^y. To deal with this, I employ an approximation to the prior in (40) where the first stage is replaced with a simple Lasso regression with a 10-fold cross-validated penalty term, which delivers only one occurrence of γ^d with $\gamma_j^d = 0$ if the Lasso estimate is 0. The second stage then maps the vector γ^d into the vector $\gamma^y \mid \gamma^d$ via (39). This procedure can be

viewed as a hybrid empirical Bayes approximation to the fully Bayesian treatment given above that works remarkably well in simulation studies. Thus, the algorithmic scheme is as follows

1. Fit a Lasso predicting the treatment parameter d_i on all controls x_i. Let $\gamma_j^d = 1$ if $\hat{\beta}_j^{Lasso} \neq 0$.

2. Update the prior γ^y

$$\Pr(\gamma_j^y = 1 \mid d) = \begin{cases} \frac{\zeta}{1+\zeta} & \text{if } \gamma_j^d = 1 \\ \frac{\phi}{1+\phi} & \text{if } \gamma_j^d = 0 \end{cases} \qquad j = 1, ..., p$$

3. Run Algorithm 1 or 2 to the outcome model (36) with d fixed and with $p(\gamma^y \mid d)$ from Step 2 used as an updated prior on γ^y, thereby skipping the sampling of ω in, respectively, Steps 3 and 5.

Indeed, the virtue of the spike-and-slab over the Lasso is its ability to simultaneously conduct variable selection and BMA parameter estimation, but as the parameter estimation is not needed in the first stage, the Lasso provides a computationally favorable alternative. It should, however, be noted that this Lasso approximation changes the target distribution of the Markov Chain, and we are thus in theory not drawing from the exact posterior, but the promising results in the simulation studies in Section 6 indicate that this is not a big issue. The function code for the treatment prior model can be found *here*.

5 Variable Selection and Inference in a Finite Mixture Model of Linear Regressions

The two preceding sections demonstrated how the spike-and-slab models are able to readily serve as a Bayesian alternative to the existing frequentist toolbox for variable selection, fit and inference in the linear model with and without a structural link between a parameter of interest and confounders. As a final horse race between the frequentist and Bayesian high-dimensional toolbox, I consider the application to a finite mixture model of linear regressions, where the linear relationship between outcome

and covariates is allowed to flexibly vary between groups. This is a fruitful Bayesian application of the spike-and-slab, as the Lasso utilized for frequentist variable selection loses much of its appeal because the EM algorithm employed for estimation is only guaranteed to converge to a stationary point of the negative log-likelihood. The analysis takes its starting point from Khalili and Chen (2007) who considered frequentist variable selection in a finite mixture model of linear regressions, and then turns to its Bayesian analogue in Lee et al. (2016) who considered variable selection using spike-and-slab priors, which I modify to fit more easily into the framework of Section 3.2.

5.1 Frequentist Approach

Khalili and Chen (2007) consider applications where it is assumed that the N observations derive from a heterogeneous population with M sub-populations or mixture components, and within each sub-population the outcome is modelled as a separate linear regression in the covariates. They therefore consider the extension of (31) to a normal regression mixture, such that

$$y_i \mid \pi_m, \beta_m, \sigma_m^2 \sim \sum_{m=1}^{M} \pi_m N(y; X_i\beta_m, \sigma_m^2),\ \pi_m \geq 0,\ \sum_{m=1}^{M} \pi_m = 1 \qquad i = 1, ..., N \tag{41}$$

where $\pi = (\pi_1, ..., \pi_M)$ are the mixture weights, $\beta_m = (\beta_{m1}, ..., \beta_{mp})'$ is the coefficient vector for sub-population m, and σ_m^2 is the variance for the Gaussian residual errors in group m. Define $\beta = (\beta_1, ..., \beta_M)$ and $\sigma^2 = (\sigma_1^2, ..., \sigma_M^2)$, then the corresponding log-likelihood is

$$l(\beta, \sigma^2, \pi) = \sum_{i=1}^{N} \log \left\{ \sum_{m=1}^{M} \pi_m N(y_i; X_i\beta_m, \sigma_m^2) \right\}$$

In order to conduct variable selection, the authors consider minimizing the penalized negative log-likelihood

$$\tilde{l}(\beta, \sigma^2, \pi) = -l(\beta, \sigma^2, \pi) + p(\beta, \sigma^2, \pi) \tag{42}$$

where $p(\cdot)$ is the penalty function, here the Lasso as described in Section 3.1. Minimizing (42) results in some coefficients being exactly zero, and the optimization routine makes use of both the EM algorithm and Newton-Rhapson updating, where they also provide a numerical choice of the tuning penalty

parameter λ. One caveat of this approach is that unlike the standard Lasso, the negative log-likelihood in the mixture model is no longer globally convex, and as such the EM algorithm is only guaranteed to converge towards a stationary point, which therefore can lead to quite unstable performance. This motivates using Bayesian methods, where it is possible to visit several local minima during MCMC.

5.2 Spike-and-Slab Priors in a Finite Normal Mixture Model with a Dirac Spike

Like the beta mixture prior in Section 3.3, the natural Bayesian way to estimate this type of model in (41) is to introduce the latent vector $z_i = (z_{i1}, ..., z_{iM})$, where $z_{im} \in \{0, 1\}$ denotes whether observation i belongs to group m, which we sample during MCMC.

To allow for different sparsity patterns across groups, instead of introducing a penalty term, Lee et al. (2016) introduce the latent inclusion vector $\gamma_m = (\gamma_{m1}, ..., \gamma_{mp})$ as in Section 3.2.1 with a Dirac spike such that $\gamma_{mj} = 0 \Leftrightarrow \beta_{mj} = 0$. Then, the model in (41) can be expressed as

$$y_i \mid \pi_m, \beta_m, \sigma_m^2, \gamma_m \sim \sum_{m=1}^{M} \pi_m N(X_i^{\gamma_m} \beta_m^{\gamma_m}, \sigma_m^2), \qquad i = 1, ..., N \tag{43}$$

where $X_i^{\gamma_m}$ collects the active columns j of X_i for which $\gamma_{mj} = 1$, and similarly for the other variables. As with the beta mixture prior, priors on z and π are chosen to be conjugate

$$z_i \mid \pi \sim \text{Discrete}(\pi_1, ..., \pi_M) \tag{44}$$

$$\pi \mid \alpha_1, ..., \alpha_M \sim \text{Dirichlet}(\alpha_1, ..., \alpha_M) \tag{45}$$

To complete the model, priors on the remaining parameters γ, β, σ^2 have to be specified. I consider a slightly modified model than Lee et al. (2016) where the prior slab is changed, which I find to perform better on simulated data. While they propose the slab

$$\beta_m^{\gamma_m} \sim N(\hat{\beta}_m^{\gamma, OLS}, n_m \sigma_m^2 (X'X)^{-1})$$

where $\hat{\beta}_m^{\gamma,OLS} = (X'_m X)_m^{-1} X'_m y_m$, I consider the independence slab, and I also consider the limiting inverse-gamma prior on the variance term, which is the Jeffreys prior[2], such that the remaining priors are

$$\beta_m^{\gamma_m} \sim N(\mathbf{0}_{p_{\gamma_m}}, \sqrt{n_m} I_m \sigma^2), \qquad p_{\gamma_m} = \sum_{j=1}^{p} \gamma_{mj}, \qquad n_m = \sum_{i=1}^{N} \mathbf{1}(z_{im} = 1) \tag{46}$$

$$p(\sigma^2) \propto \frac{1}{\sigma^2} \tag{47}$$

$$p(\gamma_m) = \prod_{j=1}^{p} \omega_{mj}^{\gamma_{mj}} (1 - \omega_{mj})^{1-\gamma_{mj}} \tag{48}$$

Conditional on $z = (z_1, ..., z_N)$, the model for each group m is then exactly the same as the standard linear case analyzed in Section 3.2.2 for all groups m, when z_i's are assumed independent of $(\gamma_m, \beta_m, \sigma_m^2)$, and $(\gamma_m, \beta_m, \sigma_m^2)$ are assumed independent of each other. The sampling steps for π and z are similar in logic to before:

1. Sample $\pi^{(t)}$ from $p(\pi \mid z)$, where $\pi \mid z \sim \text{Dirichlet}(n_1 + \alpha_1, ..., n_M + \alpha_M)$, $n_m = \sum_{i=1}^{N} \mathbf{1}(z_{im} = 1)$

2. Sample the group membership vector via

$$\Pr(z_{im} = 1 \mid y, \beta, \sigma^2, \pi, \gamma) \propto \pi_m \times \text{N}(y_i \mid X_i^{\gamma_m} \beta_m^{\gamma_m}, \sigma_{m,\gamma}^2)$$

Thus, for each group m, the sampling is the same as outlined in Algorithm 1, where the above two steps are added in each iteration of the MCMC that allocates the memberships of the observations to each group. The function code for the generic version with $\omega_j \sim \text{Beta}(0.5, 0.5)$ can be accessed *here*, and the default is with an independence slab, but the function also easily allows for using the g-prior slab $\beta_m^{\gamma_m} \sim N(\mathbf{0}, n_m (X'_m X_m)^{-1})$ with a Moore-Penrose generalized inverse.

[2]Specifically, I consider the limit of $\sigma^2 \sim \text{IG}(\frac{a_m}{2}, \frac{b_m}{2})$ and I let $a_m, b_m \to 0$ such that $p(\sigma^2) \propto \frac{1}{\sigma^2}$. Changing the prior on $\beta_m^{\gamma_m}$ is due to the the OLS estimates being quite unstable with few observations, which in turn leads to unstable variable selection. As a matter of fact, placing a non-zero prior mean on the regression coefficients makes little sense when variable selection is the goal. I also consider a slightly modified sampling scheme, where I block-update σ_m^2 with $\beta_m^{\gamma_m}$ being integrated out.

5.2.1 Posterior analysis

Similar to the standard linear case, variable selection is performed by employing the median probability model criterion (Barbieri et al., 2004), and group membership can be determined by assigning the observation to the group from which it has belonged to most times during MCMC.

A common problem with mixture models is the nonidentifiability of the component parameters, which leads to the so-called "label switching" problem (Celeux, 1998) caused by the symmetry in the likelihood of the model parameters. This symmetry can cause group switches of the parameters during MCMC that leads to a blurred posterior output, where posterior output from one group can contain posterior output that belongs to a second group. However, the trace plots from the posterior output do not indicate these switches in the experiments, and there was thus no need to correct for this.

6 Simulation Study

I investigate the performance of the spike-and-slab in the three models presented in Sections 3 (the linear regression model), 4 (the treatment effects model) and 5 (the finite mixture model of linear regressions), and I compare them in a horse race study to their Lasso alternatives. Ultimately, the objective is to evaluate when spike-and-slab models provide a good alternative to the Lasso and its relatives, and whether the proposed new priors, the beta mixture and the two-step treatment, in fact improve performance.

For the linear regression model, I compare the spike-and-slab with the post-Lasso estimator in terms of fit, variable selection and inference on the regression function $f(X_i) = X_i\beta$ in a HDS model. Because I also proposed a new mixture prior on ω_j in Section 3.3, I simultaneously investigate whether this new prior performs better relative to the existing ones in terms of the three performance metrics.

For the treatment effects model, I proposed the two-step spike-and-slab analogue to the double-Lasso method, and I evaluate its ability to obtain precise estimates on the treatment parameter of interest in settings with a high-dimensional sparse set of confounders. I also use a (standard) one-step spike-and-slab as a placebo to evaluate whether the proposed two-step procedure in fact improves on inference.

The study of the finite mixture model of linear regressions is a brief comparison between the Lasso procedure proposed in Khalili and Chen (2007) and the spike-and-slab approach in Lee et al. (2016), which is intended to showcase the benefit of the Bayesian alternative in models where the frequentist optimization routine loses its appeal.

Throughout all three studies, I confine attention to spike-and-slab models with a Dirac delta spike. This is motivated by there being more theoretical evidence in favor of this type of spike rather than a continuous spike, see e.g. Ročková and George (2018) for a discussion on this, where they also refer to this point-mass spike as being "the theoretically ideal". Qualitatively, the results seem to carry over to absolutely continuous spikes, which the interested reader can readily confirm by applying the code available on Github to the data analyzed, but it has been omitted in the coming section for brevity. For the Dirac spike-and-slab, there are many possible slabs one can choose. I use the independence slab

$$\beta_\gamma \sim N(\mathbf{0}_{p_\gamma}, cI_{p_\gamma}\sigma^2), \qquad p_\gamma = \sum_{j=1}^{p} \gamma_j$$

where β_γ denotes the non-zero set of β for which $\gamma_j = 1$ and I_{p_γ} is the p_γ-dimensional identity matrix. This slab is chosen because the simulation study in Malsiner-Walli and Wagner (2011) suggested that it traded off the false positive/false negative ratio with having more false positives (included zero coefficients), where e.g. the g-prior (Zellner, 1986) had relatively more false negatives (excluded non-zero coefficients) when using the median probability model criterion (Barbieri et al., 2004) for variable selection. For causal inference in Section 6.2 it therefore makes sense to select this type of slab to avoid missing important confounders, and I stick to this slab throughout for simplicity.

6.1 Linear Regression Model

I consider variable selection, fit and inference in the model analyzed in Section 3

$$y_i = \mu + x_i\beta + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2),\ i = 1, ..., N \tag{49}$$

with $N = 40$ and $\mu = 1$. x_i is a 40-dimensional vector of covariates where $x_i \sim N(\mathbf{0}, \Sigma)$ with correlation $\Sigma_{ij} = \pi^{|i-j|}$ as in Tibshirani (1996), and I consider two cases with either $\pi = 0$ (independent regressors) or $\pi = 0.5$ (correlated regressors). The coefficient vector in $\mathbb{R}^{40}$ is sparse with five active entries, such that

$$\beta = (1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, 0, ..., 0)'$$

and I consider two types of signals, one with a strong signal $\sigma^2 = 0.1$ and one with a weak signal $\sigma^2 = 1$. Using the definition of the signal-to-noise ratio in the linear model in Dicker (2012)

$$\frac{\tau^2}{\sigma^2} = \frac{\beta'\Sigma\beta}{\text{Var}(\epsilon)}$$

these two signals correspond to signal-to-noise ratios of 1.4 and 14 for the independent regressors, and 2.6 and 26 for the correlated regressors.

When variable selection is the goal, the tuning parameter c plays an important role in trading off false positives with false negatives (like λ in the Lasso), where a larger c enforces stronger sparsity, see e.g. Malsiner-Walli and Wagner (2011), and thus also more false negatives. Therefore, for variable selection, I consider three cases for $c \in \{1, \sqrt{N} \approx 6, 100\}$, where letting c grow with N makes intuitive sense to favor parsimonious models as the number of observations grow and more information becomes available[3]. As starting values for the MCMC, I set $\gamma^{(0)} = (0, 0, ..., 0)$, and $\omega^{(0)}$ is drawn from its prior distribution. Different starting values do not change convergence.

Five estimators are compared with three being Bayesian spike-and-slabs and two being frequentist. The three Bayesian spike-and-slab models differ on the priors put on the inclusion probabilities - one is with the common beta prior $\omega \sim \text{Beta}(1, 1)$, one is with the individual beta prior $\omega_j \sim \text{Beta}(0.5, 0.5)$ and the final one is with the mixture prior described in Section 3.3 with three mixture components[4]. They are compared to the post-Lasso estimator with either the theoretically grounded value of λ (called r-Lasso where "r" stands for rigorous following the notation of the hdm R-package) or the cross-

[3]This is also the rule of thumb for the g-prior slab, which I find to carry nicely over to the independence slab as well.

[4]The choice of hyperpriors do not have much impact on the common beta prior, whereas for the individual beta prior the Beta$(0.5, 0.5)$ (a U-shaped prior) is a choice that I have found to yield the best variable selection performance. As mentioned earlier, the three-component mixture also has a slightly better performance than the two-component in my experience.

validated value of λ (called CV-Lasso) using the "one standard error" rule, in which one chooses the most parsimonious model whose error is no more than one standard error above the error of the best model, which is a rule of thumb choice, cf. Hastie et al. (2009). The models are evaluated first by their variable selection performance, and afterwards by their bias and predictive fit.

6.1.1 Variable Selection

Independent Regressors

To evaluate the performance of the variable selection, I estimate the model in (49) on $n = 100$ data sets where I count the number of times the regressors were included in the model. For Bayesian variable selection, the median probability model criterion (Barbieri et al., 2004) is employed, and only the results for $c = \sqrt{N}$ are displayed. Table 9.1.1 in Appendix 9.1 provides the numbers behind the graphs, which also includes the cases for $c = 1$ and $c = 100$[5].

Figure 6.1 displays the results for the case of independent regressors, which shows the number of times each of the five regressors were selected into the model over $n = 100$ data sets, where the bottom-right graph shows the average occurrence of the remaining 35 redundant regressors (e.g., for the individual beta prior in the weak signal case, each of the 35 zero coefficient regressors entered the model in 15 data sets on average out of the $n = 100$ - note the change of scale on the second axis). The black histogram is the strong signal case, the blue is the weak signal case.

For the strong signal case, all five estimators capture almost all of the three largest coefficients, where the CV-Lasso and the individual and mixture beta prior are better at capturing the weak coefficients. On the other hand, they also by far have the most false positives, i.e. they include too many zero coefficient regressors, as seen by the bottom-right figure. The same picture emerges for the weak signal case, where the five estimators, however, are only able to correctly include the strongest coefficient in almost all data sets. Looking at the figure, the results indicate two classes of estimators: one which includes most of both the non-zero and zero coefficients (the CV-Lasso, individual and beta mixture, grouped together to the left in the figures) and the other group, which is more conservative

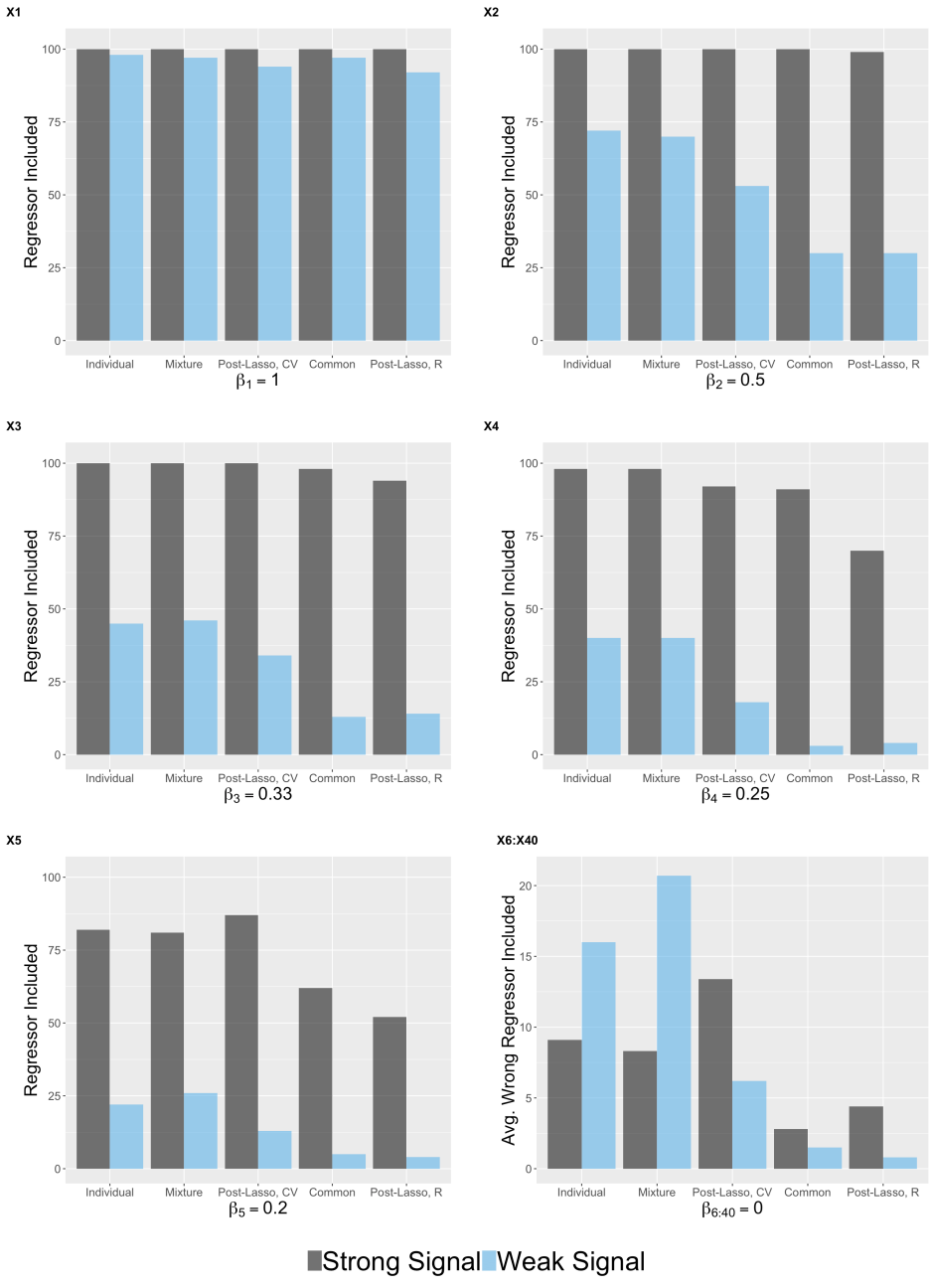
[5] In general $c = 100$ seems to impose too much sparsity, whereas the cases for $c = 1$ or $c = 6$ are quite similar.

and excludes more non-zero and zero coefficients (the r-Lasso and the common beta prior).

Within the first class, the individual and mixture beta prior are comparable and generally detect the non-zero coefficients better than the CV-Lasso, whose ability to exclude zero coefficients depend on the signal of the data: for a weak signal, the CV-Lasso becomes overconfident and includes too many zero coefficients, whereas in the weak signal case it includes fewer. For the mixture prior, especially, the ability to exclude zero coefficient regressors from the model degrades in the weak signal regime, where it is barely able to conduct any meaningful variable selection.

Within the second class, the r-Lasso and the common beta prior are comparable, where the common beta prior is slightly better in the strong signal case with most non-zero coefficients and fewest zero coefficients. This lead vanishes, however, for the weak signal case.

Figure 6.1: Variable Selection for Independent Regressors



Correlated Regressors

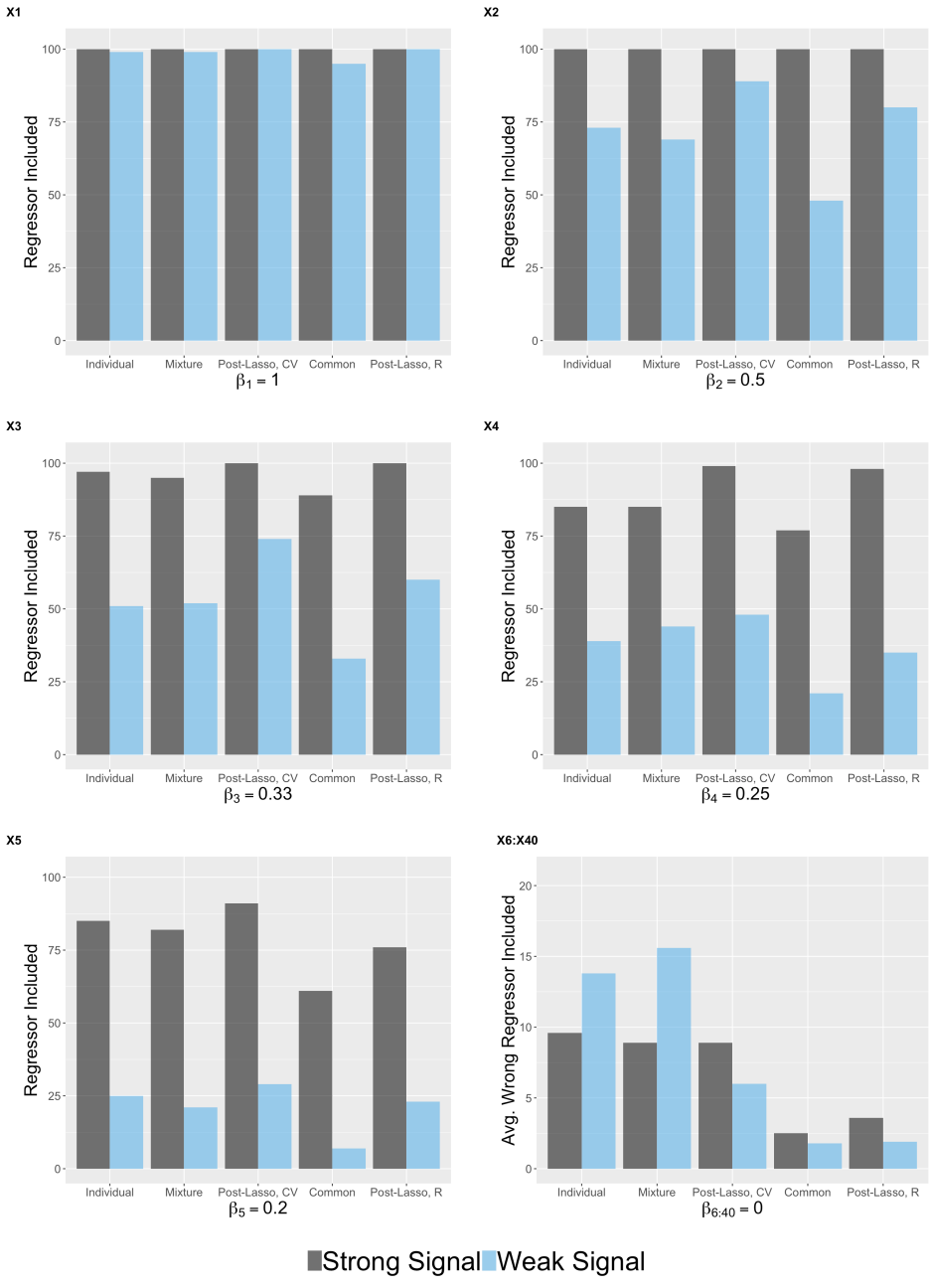
Figure 6.2 displays the same results for the case of correlated regressors, and Table 9.1.2 provides the numbers behind the figure in Appendix 9.1. While the two-class pattern from the independent case carries over, the Lasso estimators - especially the r-Lasso - perform considerably better relative to the Bayesian estimators as well as to the independent regressors regime. These frequentist estimators are superior at correctly classifying both zero and non-zero coefficient regressors.

Take-aways

The variable selection study provides the following insights

- The mixture beta prior and individual beta prior provide more or less the same variable selection performance with a high number of true and false positives. However, for the weak signal case, the mixture prior is worse at excluding zero coefficients.
- The CV-Lasso and r-Lasso differ in the sense that the CV-Lasso has many true and false positives, whereas the r-Lasso selects a sparser set of variables with fewer true positives and less false positives. For correlated regressors, the r-Lasso seems to perform best, where it is very good at excluding the zero coefficients.
- For the independent regressors case, the Bayesian and frequentist methods perform similarly.
- For the correlated regressors case, the r-Lasso and CV-Lasso outperformed the Bayesian methods in both detecting the zero and non-zero coefficient regressors.

Figure 6.2: Variable Selection for Correlated Regressors



6.1.2 Prediction Error and Bias

While variable selection can be a goal in itself, inference is usually the goal in econometrics. I compare the performance of the posterior mean estimators, the post-Lasso estimator with a theoretically grounded penalty term (r-Lasso) and the Oracle OLS estimator. The CV-Lasso is omitted because there is no theoretical justification for its performance, whereas these are provided rigorously for the Post r-Lasso (Belloni et al., 2013). Related to the discussion of the bias of the posterior mean estimator in Section 3.2.6, I only report the results for $c = \sqrt{N}$ to limit the bias without enforcing too much sparsity.

Independent Regressors

The results are summarized in Table 6.1, which displays the mean ℓ_0 norm $E(\mid \hat{\beta} \mid_0)$ (i.e. the average number of non-zero components); the ℓ_2 norm of the bias $\parallel E(\hat{\beta}) - \beta \parallel_2$; and the mean prediction error $E[\frac{1}{N}\sum_{i=1}^{N} \mid x_i'(\hat{\beta} - \beta) \mid]$. For the ℓ_2 norm of the bias, I also report in parenthesis the corresponding bias on the active set only.

For the Bayesian estimators, the common beta prior performs better than the other two in terms of both bias and prediction error, and it identifies the ℓ_0 norm quite precisely under the strong signal regime. However the other two perform better in terms of bias on the true active regression coefficients, which makes intuitive sense taking their ℓ_1 norms and the variable selection performance studied previously into account, as they are less likely to omit regressors during MCMC, which comes at the cost of a bias on the zero-regression coefficients, but less bias on the non-zero coefficients. Compared to the r-Lasso estimator, the common beta prior has a comparable performance, where the r-Lasso seems to perform slightly better under a weak signal regime, whereas the common beta prior performs relatively better under a strong signal.

Table 6.1
Independent Regressors: Inferential and Predictive Performance for $N = p = 40$

	Strong signal			*Weak signal*		
	Mean ℓ_0 norm	ℓ_2 bias	Pred. error	Mean ℓ_0 norm	ℓ_2 bias	Pred. error
Ind. incl. prior	7.30	0.07 (0.04)	0.15	7.02	0.91 (0.37)	0.52
Common prior	4.77	0.05 (0.04)	0.13	2.14	0.54 (0.49)	0.52
Mixture Prior	6.29	0.06 (0.04)	0.13	6.88	1.08 (0.43)	0.52
Post Lasso, r	4.75	0.10 (0.09)	0.18	2.07	0.50 (0.44)	0.50
Oracle OLS	5.00	0.02 (0.02)	0.09	5.00	0.14 (0.14)	0.28

Correlated Regressors

Table 6.2 displays the same results for the correlated data. Like the variable selection performance, the r-Lasso is clearly the best performing estimator with lowest ℓ_2 bias and prediction error under both strong and weak signal regimes. It is also striking how close the r-Lasso comes to the Oracle OLS performance in terms of ℓ_2 bias when the signal is strong, which is also theoretically grounded in Belloni et al. (2013). For the Bayesian methods, both the individual beta prior and mixture beta prior perform relatively worse than the common beta prior in terms of prediction error and bias on all coefficients. However, again, both of the two perform better in terms of bias on the true active regression coefficients.

Table 6.2
Correlated Regressors: Inferential and Predictive Performance for $N = p = 40$

	Strong signal			*Weak signal*		
	Mean ℓ_0 norm	ℓ_2 bias	Pred. error	Mean ℓ_0 norm	ℓ_2 bias	Pred. error
Ind. incl. prior	7.09	0.09 (0.05)	0.16	7.75	1.38 (0.50)	0.55
Common prior	4.52	0.08 (0.07)	0.14	2.31	0.66 (0.59)	0.51
Mixture Prior	5.39	0.09 (0.05)	0.15	8.63	1.53 (0.51)	0.54
Post Lasso, r	5.36	0.04 (0.03)	0.11	3.41	0.38 (0.33)	0.39
Oracle OLS	5.00	0.02 (0.02)	0.09	5.00	0.22 (0.22)	0.28

Take-aways

- In terms of bias and predictive performance, the r-Lasso performs best on especially the correlated

data where it also attains a near-Oracle OLS bias when the signal is strong.

- Among the Bayesian methods, the common beta prior performs best in terms of bias and predictive fit of the regression function $f(X_i) = X_i\beta$, whereas the individual and mixture beta prior perform relatively better in estimating the non-zero coefficients. The fit and bias of the individual and mixture prior were similar.

The conclusion from this study is that the post-Lasso estimator with a theoretically grounded penalty term generally performs best in terms of both variable selection, inference and fit, which is especially evident on correlated data. The Bayesian estimators were, however, competitive when the regressors were independent. In terms of fitting the regression function, the common beta prior seemed preferable to the two other Bayesian alternatives in terms of bias and predictive fit, but it did not provide a better fit for the non-zero coefficients.

Finally, the proposed mixture prior did not improve upon performance relative to the individual beta prior, and since the individual beta prior is simpler and easier to implement, it seems as the preferable choice between the two. For weak signal regimes, especially, the mixture prior makes little sense to use for variable selection as it includes $15 - 20$ out of the 35 zero coefficients, and it has a quite large bias here as well. The poor performance of the mixture prior in the weak signal case is likely due to the many hierarchies in the model and limited information. Indeed, the latent variable ω_j is updated conditional on another latent variable γ_j, and it may simply be too hard of a job to estimate the parameters of the mixture components precisely and allocate ω_j to the right components so far down the hierarchy when $p = 40$ and the signal is scarce.

Apart from generally performing best, the post-r-Lasso is also computationally efficient. It takes less than a second for it to solve the model with dimensions $N = p = 40$, whereas for the Bayesian methods it takes roughly 80 seconds to run $6,000$ MCMC iterations on a Macbook Pro with a 2.3 GHz Intel Core i5 processor. Thus, unless one has some strong prior information that can be incorporated into e.g. the individual prior, the post-Lasso estimator seems as a preferable alternative to the spike-and-slab models if the goal is a good fit of the regression function, variable selection or predictive accuracy.

6.2 Treatment Effects

While the frequentist toolbox fared best in terms of variable selection, bias, fit and computational speed in the linear regression model when estimation of the linear regression function $f(X_i) = X_i\beta$ was the goal, oftentimes interest lies in estimating a single coefficient that can be given a causal interpretation, e.g. the effect of education on wages. This section considers this scenario where the goal is inference of a single treatment parameter in sparse high-dimensional settings. I investigate the performance of the treatment prior described in Section 4.2, which I also benchmark against the individual beta prior $\omega_j \sim \text{Beta}(0.5, 0.5)$ that fared relatively well in the simulation study of the linear regression model with fewest false negatives for $c = \sqrt{N}$ and lowest bias on the non-zero coefficients. I then compare the two to the double-Lasso estimator and an Oracle OLS estimator. As it is relatively more important to avoid false negatives than false positives that could lead to a substantial omitted-variable bias, I again employ the independence slab and fix the hyperprior $c = \sqrt{N}$ (to ameliorate the bias of the estimator as discussed in Section 3.2.6). For this study, I consider the structural model

$$y_i = \mu + \alpha_0 d_i + x_i\beta_y + \epsilon_i \tag{50}$$

$$d_i = x_i\beta_d + \nu_i \tag{51}$$

where d_i is a scalar continuous treatment variable, x_i is a p-dimensional vector of controls where $x_i \sim N(\mathbf{0}, \Sigma)$ with correlation $\Sigma_{ij} = 0.5^{|i-j|}$, $(\epsilon_i, \nu_i)' \sim N(\mathbf{0}, \Omega)$ where $\Omega = \text{diag}(\sigma_\epsilon^2, 1)$, and α_0 is the parameter of interest. For this study, I increase the number of variables and observations such that $p = N = 100$. The parameters are first set in a "sparse" case as

$$\begin{aligned} \beta_y &= (1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, 0, 0, 0, 0, 0, 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, 0, 0...) \\ \beta_d &= (1, \frac{1}{2}, \frac{1}{3}..., \frac{1}{10}, 0, 0, ...) \\ \alpha_0 &= 1, \mu = 1 \end{aligned} \tag{52}$$

and in a second study, I consider a non-sparse case with twice as many active entries in β_y, such that

$$\begin{aligned}
\beta_y &= (1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, 0, 0, 0, 0, 0, 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, 0, 0, 0, 0, 0, \\
&\quad 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, 0, ...) \\
\beta_d &= (1, \frac{1}{2}, \frac{1}{3} ..., \frac{1}{10}, 0, 0, ...) \\
\alpha_0 &= 1, \mu = 1
\end{aligned} \tag{53}$$

Two cases of noise are investigated: One with a low level of noise $\sigma_\epsilon^2 = 1$ ("strong" signal) and one with a high level $\sigma_\epsilon^2 \approx 2$ ("weak" signal) leading to signal-to-noise ratios of, respectively, 7 and 3, see Appendix 9.5 for a calculation of the signal-to-noise ratio in this case. Convergence of the MCMC is usually obtained in less than $1,000$ iterations, and I run $M = 6,000$ iterations and discard the first $1,000$ as a burn-in. I consider $n = 100$ data sets of which the results are averages from.

For the double-Lasso's tuning of the penalty parameter λ, I employ the theoretically justified value in Belloni et al. (2014b), which can be iteratively estimated from the data and allows for e.g. heteroskedasticity (using the function rLassoEffect from the R package hdm). One could also use a cross-validated value as the choice of theoretical vs. cross-validated λ seems to make little difference in practice[6].

For the treatment prior, hyperpriors were selected as $\phi = 0.2$ and $\zeta = 10$ in order to deliver a fairly sparse representation in the second step, as the first-step lasso approximation in general overestimates the true active set of regressors (Bühlmann and Van De Geer (2011), Chapter 2). In general, I find the results to be robust for any $\zeta > 5$ and $\phi < 0.5$, which is likely due to the BMA effect.

6.2.1 Results

The results from the study are summarized in Table 6.3 for both the sparse case (52) and the non-sparse case (53), which records the performance of the estimators of α_0 for the treatment prior, the individual beta prior, the double-Lasso and the Oracle OLS. For the two Bayesian methods, $\hat{\alpha}_0$ is

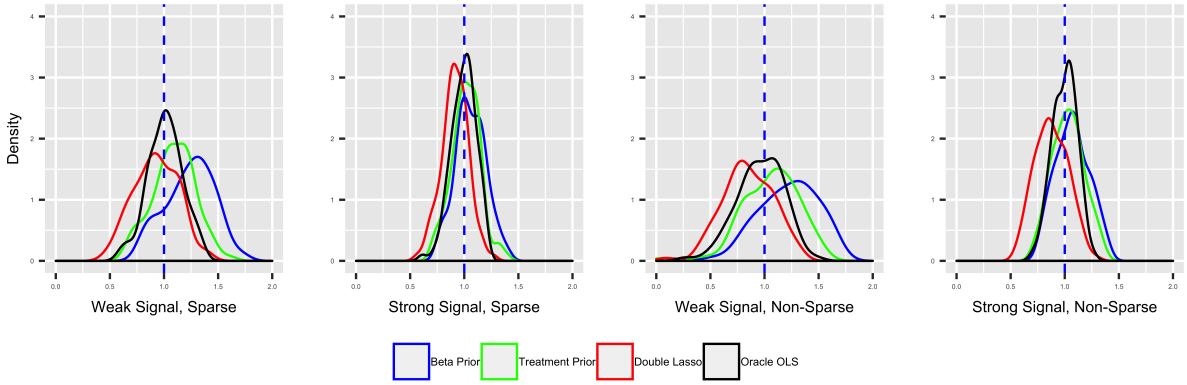
[6] One of the authors of Belloni et al. (2014b), Christian Hansen, has an illuminating NBER talk on this *here*, where his discussion on the practical indifference between the theoretical and cross-validated penalty term can be found around minute 35.

the posterior mean and the median probability model criterion (Barbieri et al., 2004) is used to select variables. The performance is evaluated by the bias - which is the difference between the mean of the estimates and the true value $\alpha_0 = 1$ - the mean squared error, the coverage probability of the 95% credible or confidence interval, the coverage length, the number of falsely classified regressors, and the number of false positives. Density plots for $\hat{\alpha}_0$ are also reported in Figure 6.3.

For the sparse model, the treatment prior performs best with both strong and weak signals by having the lowest bias and mean squared error, highest coverage probability and fewest falsely classified regressors. For a weak signal, however, the difference between the double-Lasso and the treatment prior is small. Considering that it does not take the selection equation into account in its estimation, the individual beta prior $\omega_j \sim \text{Beta}(0.5, 0.5)$ fares surprisingly well in the strong-signal regime, but performs relatively bad in the weak-signal regime, which is due to it missing important confounders. This omitted-variable bias is most clearly seen from Figure 6.3, where the individual beta prior has a positive bias in all weak signal regimes, which is to be expected when the model excludes confounders that are both positively correlated with the outcome and treatment variable.

It is also interesting to see how the double-Lasso and the Bayesian methods differ in their model selection mistakes: where the double-Lasso has a relatively high number of false positives, the two Bayesian methods have a relatively high number of false negatives. As the Bayesian estimates are BMA estimates, this does not bias the estimator as severely as it would for the double-Lasso, and the lower number of false positives relative to the double-Lasso also seems to unfold into more certainty about the estimate with a higher coverage probability and narrower coverage length.

Figure 6.3: Densities of $\hat{\alpha}_0$



Note - The figures report the densities of $\hat{\alpha}_0$ averaged over $n = 100$ data sets with varying signal-to-noise ratios and sparsity patterns where the true value is $\alpha_0 = 1$, which is indicated by the blue dotted line.

Table 6.3

Comparison of estimates of α_0 for three methods with $N = 100$, $p = 100$

Sparse β_y

	Strong signal						*Weak signal*					
	Bias	MSE	Cov95	CL	Wrong	FP	Bias	MSE	Cov95	CL	Wrong	FP
Beta prior	0.12	0.02	0.83	0.42	5.35	1.72	0.27	0.11	0.77	0.80	7.85	1.60
Treatment prior	0.10	0.01	0.90	0.43	4.09	1.35	0.18	0.05	0.93	0.80	6.46	1.46
Double lasso	0.12	0.02	0.84	0.44	7.39	5.96	0.18	0.05	0.91	0.82	8.63	6.26
Oracle OLS	0.09	0.01	0.92	0.39	0.00	0.00	0.12	0.03	0.96	0.74	0.00	0.00

Non-Sparse β_y

	Strong signal						*Weak signal*					
	Bias	MSE	Cov95	CL	Wrong	FP	Bias	MSE	Cov95	CL	Wrong	FP
Beta prior	0.14	0.03	0.83	0.43	11.77	8.00	0.29	0.13	0.76	0.82	14.43	8.28
Treatment prior	0.12	0.02	0.86	0.45	9.40	6.10	0.21	0.07	0.88	0.86	12.93	8.06
Double lasso	0.16	0.04	0.74	0.48	12.46	11.11	0.22	0.08	0.89	0.86	13.32	11.29
Oracle OLS	0.09	0.01	0.95	0.42	0.00	0.00	0.18	0.05	0.93	0.79	0.00	0.00

Note - This table compares estimates of α_0 over $n = 100$ data sets for three methods and the Oracle OLS. For the Bayesian methods, $\hat{\alpha}_0$ is the posterior mean. The bias is the difference between the mean of the estimates and the true value, MSE is the mean squared error, Cov95 is the coverage probability of the 95% credible or confidence interval, CL is the coverage length, Wrong is the number of falsely classified regressors, and FP is the number of false positives.

For the non-sparse case, both the treatment prior and double-Lasso estimators perform worse in terms of all metrics relative to the sparse case, where they have more difficulties in identifying the right variables to include in the model, but the number of false negatives do not increase by much, which ameliorates the effect of this poor variable selection performance on the bias, but increases the coverage length. Interestingly, it is also evident that the individual beta prior now performs better than the double-Lasso in terms of bias when the signal is strong, which again highlights the importance of the sparsity assumption on the performance of Lasso estimators for inference. When the signal is weak, however, the picture turns again. Although all estimators move away from the Oracle OLS performance when sparsity breaks down (which was also illustrated in the simple example in Figure 3.1 for the Lasso and post-Lasso earlier in Section 3.1), the treatment prior still performs best. It is also reassuring that even in the non-sparse case with $p_0 = 20$ and $p = N = 100$, the estimators, especially the two-step treatment prior, still perform relatively well, which can be seen most easily from Figure 6.3, where the distributions of $\hat{\alpha}_0$ from both the double-Lasso and the treatment prior closely align that of the Oracle OLS for both types of signals.

Thus, in terms of inference on the parameter of interest, the treatment prior seems as an attractive alternative to the double-Lasso procedure with a lower bias and higher precision. This is likely due to its robust BMA estimation routine that is not as vulnerable to model selection mistakes as the double-Lasso. Computationally, however, the method is far more demanding than the double-Lasso. On a Macbook Pro with a 2.3 GHz Intel Core i5 processor, running one MCMC with $6,000$ iterations with $N = p = 100$ takes 200 seconds, whereas it takes less than a second for the double-Lasso.

6.3 Finite Mixture Model of Linear Regressions

As the final race, a brief study is conducted on variable selection and inference performance in a setting where the Bayesian approach has good conditions for performing relatively well due to the Lasso's difficulty in optimizing a non-convex objective function. I consider variable selection and inference in a sparse finite mixture model of linear regressions with two components, such that

$$y_i \sim \rho N(x_i'\beta_1, \sigma_1^2) + (1-\rho)N(x_i'\beta_2, \sigma_2^2), \qquad i = 1, ..., N$$

where I set $\sigma_1^2 = \sigma_2^2 = 1$ and $\rho = 0.5$, and vary the signal of the data by considering three cases with $N = 100$, $N = 200$ and $N = 500$. x_i is a 15-dimensional vector of covariates where again $x_i \sim N(\mathbf{0}, \Sigma)$ with correlation $\Sigma_{ij} = \pi^{|i-j|}$, and I consider two cases with either $\pi = 0$ (independent regressors) or $\pi = 0.5$ (correlated regressors). The coefficient vectors are sparse with five active entries, such that

$$\beta_1 = (0, 0, 1, \frac{1}{2}, \frac{1}{3}, 1, 0, ..., 0)$$
$$\beta_2 = (0, 0, 0, 0, 1, -4, \frac{1}{2}, \frac{1}{3}, 0, ..., 0))$$

where the negative coefficient in β_2 is chosen for ease of posterior identification of the components. Both coefficient vectors thus have 11 zero and 4 non-zero entries. For Bayesian estimation of the model I use the spike-and-slab approach from Section 5.2, and with the same three different priors on the individual inclusion probabilities as before - common, individual and mixture - and I compare the approach to the frequentist estimation of Khalili and Chen (2007) by using their R package fmrs with an adaptive penalty for the Lasso. As focus is on variable selection and inference, I set the number of mixture components to the true number, 2, for both algorithms.

6.3.1 Results

The results from the study are summarized in Table 6.4. Averaged over $n = 100$ data sets and averaged over both components, the table shows the average number of correctly classified zero coefficients (out of 11); the average number of incorrectly classified zero coefficients (out of 4); and the ℓ_2 norm of the bias $\| E(\hat{\beta}) - \beta \|_2$.

In all cases, all Bayesian methods are better at identifying the regressors with zero coefficients, whereas the frequentist approach is slightly better at detecting the non-zero coefficients (i.e. fewer incorrectly classified zero coefficients), because it includes almost all regressors in the model. The bias is much higher for the frequentist approach, which is due to quite large instability in the estimates, especially for the $N = 100$ case, the poor variable selection performance, and the notorious negative bias stained to the regression coefficients[7]. This highlights the benefit of the Bayesian approach to

[7]It is of course possible to conduct post-Lasso estimation to get rid of some of this bias, but this would require a post-identification step to determine to which group each observation belongs, which has not been considered here.

variable selection for this application, as the negative log-likelihood is not globally convex, which yields high uncertainty on the frequentist point estimates. As the number of observations grow, the Lasso performs much better, but it still performs strongly worse than the individual and mixture beta prior for both independent and correlated data having a higher bias, fewer correctly classified zero coefficients and more incorrectly classified zero coefficients.

Regarding the choice of priors, the same picture from the standard linear regression model carries over for both the independent and correlated case: relative to the individual and beta mixture prior, the common beta prior is better at correctly classifying the zero-coefficients ('Correct'), but also excludes more of the non-zero coefficients ('Incorrect'). The average bias on the regression functions, however, is higher for the common prior in this case, which was not the case for the standard linear regression model. The individual and mixture beta prior perform similarly, where the mixture beta prior excludes slightly more zero and non-zero coefficients. In general, it is difficult to conclude which of them are preferable, and therefore the proposed mixture prior again does not seem to improve performance.

Computationally, again, the Lasso approach is much faster. For the different N, the Lasso approach takes in the order of 1-10 seconds to estimate the model, whereas it takes 140-180 seconds to run an MCMC with $6,000$ iterations.

Table 6.4
Average Number of Correct and Incorrect Zero Coefficients for $n = 100$

		$N = 100$			$N = 200$			$N = 500$		
		Correct	Incorrect	Bias	Correct	Incorrect	Bias	Correct	Incorrect	Bias
$\pi = 0$	Ind. incl. prior	10.61	0.98	0.30	10.78	0.44	0.11	10.96	0.06	0.03
	Common prior	10.92	1.66	0.40	11.00	0.96	0.16	11.00	0.20	0.04
	Mixture prior	10.88	1.26	0.29	10.97	0.72	0.13	11.00	0.10	0.03
	Lasso	4.75	0.57	5.57	7.97	0.24	0.59	7.48	0.17	0.12
$\pi = 0.5$	Ind. incl. prior	10.66	1.02	0.40	10.88	0.72	0.17	10.93	0.17	0.05
	Common prior	10.91	1.67	0.59	10.99	1.05	0.21	11.00	0.50	0.09
	Mixture prior	10.85	1.29	0.44	10.97	0.90	0.19	10.98	0.35	0.07
	Lasso	4.69	0.57	4.37	7.77	0.38	0.58	8.05	0.37	0.31

Note - The table compares variable selection and inferential performance for the Lasso and spike-and-slab approach to finite mixture models of linear regressions. The "Correct" column gathers the average number of correctly classified zero coefficients (out of 11) averaged over both components; "Incorrect" gather the average number of incorrectly classified zero coefficients (out of 4) averaged over both components; and "Bias" gathers the ℓ_2 bias averaged over both components.

7 Empirical Example: Abortion and Crime

The results from the simulation studies suggest that frequentist methods are preferable in the standard linear regression model when interest lies in estimating the high-dimensional sparse regression function, whereas the spike-and-slab models perform better when interest lies in estimating a treatment parameter or estimating several regression functions for different sub-populations. Because of the promising result on the proposed two-step spike-and-slab for treatment effects estimation - which is also oftentimes the application of interest for economists - I illustrate this estimator in action on a famous study by Donohue III and Levitt (2001). Belloni et al. (2014a) also considered this study as a test for the double-Lasso, and I use their data set which they have made publicly available[8]. If interested in empirical examples of the spike-and-slab models for the standard linear regression case or the finite mixture model of linear regression, the reader is referred to Malsiner-Walli and Wagner (2011) and Lee et al. (2016).

[8]Creating the data set with an expanded covariate set as well as estimation can be found on the Github page linked to in the beginning.

7.1 Basic Specification

I consider the estimation of the effects of abortion on crime analyzed in Donohue III and Levitt (2001). The paper tells quite an interesting story: the increase in abortion rates following the Roe v. Wade decision in 1973 that eased legal restrictions on abortion in many states, is associated with lower rates of crime in the 1990s. The paper sparked quite a debate, as the main channel is hypothesized to be through a reduction in unwanted births, which countered the general belief at the time that the reduction in crime was due to the government being tough on crime.

The authors look at state-level panel data in the period running from 1985 to 1997, where their empirical specification is

$$\ln(\text{CRIME}_{st}) = \alpha_0 \text{ABORT}_{st} + x_{st}\beta + \mu_s + \lambda_t + \epsilon_{st} \tag{54}$$

where $\text{CRIME}_{st} = \{\text{Property}_{st}, \text{Murder}_{st}, \text{Violence}_{st}\}$ is the crime rate for the particular crime in state s at time t; ABORT_{st} is the "effective abortion rate" that is the abortion rate relevant for the crime, which is determined by the age criminals tend to commit crimes (e.g. if violent crimes are typically committed for ages 20-25, then the effective abortion rate in 1995 is determined by weighting the actual abortion rates over the period 1970-75); μ_s is a vector of state-dummies; λ_t a vector of time dummies; and x_{st} is a vector of controls to correct for abortion rates not being randomly assigned. In their original specification, the following eight controls are included

1. log of lagged prisoners per capita
2. log of lagged police per capita
3. state unemployment rate
4. log of state income per capita
5. state poverty rate
6. measures of welfare generosity at time $t - 15$

7. a dummy for having a concealed weapons law

8. beer consumption per capita

If these controls capture all factors that may be associated with both state-level crime and the effective abortion rate, then the estimate of α_0 measures the causal effect. However, missing important trends, interactions or other confounders (such as e.g. drug use) can lead to an omitted-variable bias. Indeed, critics of the paper such as Foote and Goetz (2008) showed how controlling for other crime-associated factors reduced the estimated effects by roughly a half, and Kahane et al. (2008) replicated the study on England and Wales and concluded that they could obtain Donohue III and Levitt (2001)'s findings from the US, but that "this association breaks down under the scrutiny of robustness checks". This motivates expanding the set of possible confounders and perform a data-driven selection of controls by applying high-dimensional methods such as the spike-and-slab to learn about the effect.

7.2 Expanding the Set of Controls

Belloni et al. (2014a) also analyzed this data with the double-Lasso, and I follow their example by expanding the set of covariates to capture all possible trends and interactions. The data set consists of 50 states with 12 time-series observations for a total of 600 observations, and I consider the within-state first-differenced structural model

$$\Delta \ln(\text{CRIME}_{st}) = \alpha_0 \Delta \text{ABORT}_{st} + z_{st}\beta + \lambda_t + \Delta\epsilon_{st} \tag{55}$$

$$\Delta \text{ABORT}_{st} = z_{st}\theta + \kappa_t + \Delta\nu_{st} \tag{56}$$

where $\Delta\text{CRIME}_{st} = \text{CRIME}_{st} - \text{CRIME}_{s,t-1}$ and similarly for all other Δ's; λ_t and κ_t are time effects, and z_{st} is the vector of expanded confounders. This set consists of levels, differences, initial levels, initial differences, and within-state averages of the eight confounders; the initial level and initial differences of the effective abortion rate relevant to the crime, quadratics of each of the aforementioned variables, interactions of all variables with t and t^2 and also t and t^2. Including time dummies, this leaves a total of $p = 322$ potential confounders. With $N = 600$, it is of course possible to estimate the model with

OLS, but few reasonable researchers would consider this approach. Instead, high-dimensional tools are needed that can search for the true, sparse model.

7.3 Estimation and Results

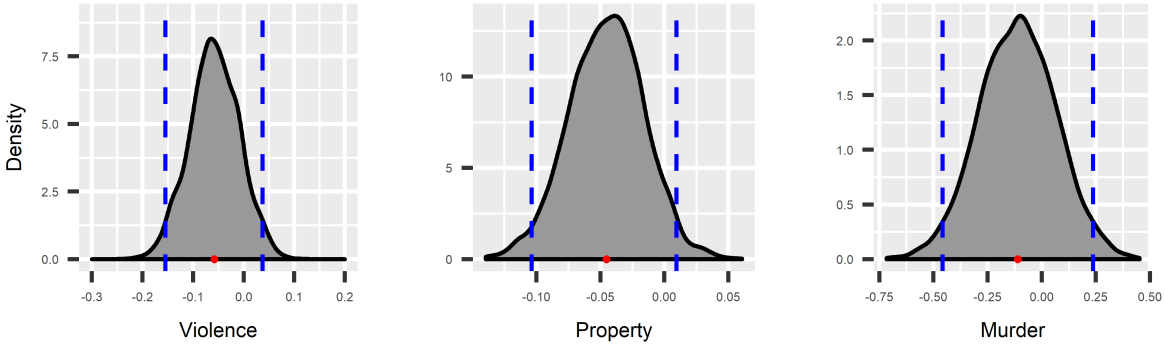
MCMC is performed for the spike-and-slab with independence slab, Dirac spike and treatment prior, and was run for $M = 6,000$ iterations with a burn-in of $1,000$ iterations, after which the trace plots indicate proper convergence. Hyperpriors were selected as $c = \sqrt{N} \approx 25$ for the independence slab and $\phi = 0.2$ and $\zeta = 10$ for the treatment prior. The priors ϕ and ζ were again chosen to deliver a fairly sparse representation in the second step.

The spike-and-slab was run by considering two scenarios: One in which variable selection is fully data-driven, and one in which prior beliefs of Donohue III and Levitt (2001) are incorporated by fixing the eight confounders in the baseline specification (54) in the model. However, as the results show that neither of the eight confounders enter the model significantly conditional on inclusion (i.e. the posterior distributions are centered around 0), the estimates from the two different scenarios are practically identical, and I choose to only present the results from the former.

7.3.1 Results

The posterior densities for the three α_0's are displayed in Figure 7.1. While all three posteriors indicate a negative effect of abortion on crime (consistent with the original story), 0 is included in all 95% credible intervals, thus invalidating a strong causal claim. Similar results occur when using the double-Lasso, which can be seen from Table 7.1. The table shows the posterior inference of the spike-and-slab model compared to the double-Lasso and a first-difference estimation in (55) with the eight original controls of Donohue III and Levitt (2001). As in the original paper, the first-difference shows significant negative effects, where, for example, an increase in the effective abortion rate of 10 per 100 live births is associated with around 9% reduction in property crimes. The double-Lasso and spike-and-slab model both agree on the direction of this first-difference estimate, but not the significance. This is interesting, as one concludes qualitatively differently depending on whether one chooses an intuitive set of controls or leaves it to be flexibly estimated from the data, which is in line with the

Figure 7.1: Posterior densities of α_0 for the three types of crime



Note - The figure displays the density plots of the effect of abortion on crime estimated from the final 5,000 MCMC draws. The blue lines indicate the 95% credible interval. The red dot is the posterior mean.

critique of the robustness of the results mentioned above.

That both the double-Lasso and the spike-and-slab deliver similar results is reassuring, as this was also the picture painted in the simulation study in Section 6. However, in addition to coming from two different strands of statistics, the approaches to estimation are also fundamentally different, which helps explain the different credible/confidence intervals. On the one hand, the double-Lasso estimator relies on it being able to detect the correct set of controls (Steps 1 and 2) in order for the post-OLS inference (Step 3) to be valid. For the violence estimation, for example, the controls include (in addition to the time dummies that are fixed in the OLS regression) lagged prisoners, lagged income, mean state income, lagged police squared interacted with a quadratic time trend, initial beer consumption interacted with a trend, initial violence and initial violence interacted with a quadratic time trend. Because omitting confounders can bias the results a lot, the double-Lasso and its theoretically tuned penalty term is geared towards overestimating, rather than underestimating, the active set of confounders, which comes at the expense of a higher variance on the estimator.

The spike-and-slab estimator, on the other hand, is the outcome of averaging over many different models indexed by the latent inclusion variables, i.e. Bayesian model averaging, which leaves it rather

robust to model uncertainty. Of all variables in the violence equation, only lagged prisoners, lagged police, mean state police, police interacted with a quadratic time trend entered with non-zero estimates in all MCMC iterations, whereas several others entered and exited throughout. As there are fewer confounders entering the sampling of α_0 in each MCMC iteration compared to the double-Lasso, the variance of its estimator seems to be smaller, which was also evident in the simulation study. In this empirical example, it is especially evident for the Property equation, where the credible interval is much narrower.

Table 7.1
Credible/confidence intervals and parameter estimates

	Violence			*Property*			*Murder*		
	2.5%	$\hat{\alpha_0}$	97.5%	2.5%	$\hat{\alpha_0}$	97.5%	2.5%	$\hat{\alpha_0}$	97.5%
Treatment Prior	-0.15	-0.06	0.04	-0.10	-0.05	0.00	-0.46	-0.11	0.24
Double-Lasso	-0.57	-0.21	0.15	-0.28	-0.06	0.16	-0.54	-0.19	0.16
FD OLS	-0.22	-0.12	-0.02	-0.15	-0.09	-0.04	-0.33	-0.19	-0.06

Note - This table reports credible/confidence intervals as well as estimates of the treatment parameter. For the spike-and-slab model, the posterior mean is used as the point estimate.

8 Conclusion

This thesis developed two new priors for the Bayesian spike-and-slab models and conducted a horse race study between the Lasso methods and the spike-and-slab models using these priors for three HDS models: the standard linear regression model; the treatment effects model; and the finite mixture model of linear regressions.

The first prior, the beta mixture prior, did not improve on performance in terms of variable selection and bias relative to the individual beta prior, which is arguably because an inadequate amount of information flows down the Bayesian hierarchy for a proper flexible density estimation of ω_j. On the other hand, the two-step treatment prior had a drastic impact on the ability to estimate the treatment parameter precisely among high-dimensional controls, because it incorporated important information from the selection equation into the variable selection procedure via an updated prior.

The horse race study showed how the post-Lasso performed best in the standard linear regression model in terms of variable selection, bias and predictive fit on especially correlated data, whereas for the treatment effects model the proposed two-step treatment spike-and-slab was better at obtaining more precise and less biased estimates of the treatment parameter than the double-Lasso alternative, which is likely in part due to its robust BMA estimation routine. The extension to the finite mixture model of linear regressions also showed how the spike-and-slab had a better performance than the Lasso alternative, because the Lasso procedure lost much of its appeal due to the loss of convexity in the objective function that led to unstable variable selection and estimation of the model parameters.

Putting the debate between the Lasso methods and the spike-and-slab aside, the example of the effects of abortion on crime showcased how useful both of these high-dimensional methods can be in empirical research, as they can, among other things, enable a data-driven selection of confounders in an expanded covariate space and provide a robust estimation routine for the parameter of interest. While this should not necessarily replace intuitively selecting controls, it nevertheless helps serve as a robustness check to the validity of inference based on the choices made by researchers, which seemed especially important in the Donohue III and Levitt (2001) paper. The relatively strong performance of the spike-and-slab and the post-Lasso estimator for estimating the regression function in the HDS linear regression model also highlights how useful these methods can be in obtaining flexible fits of regression functions that can e.g. be used for inference on own-price elasticities (Dubé and Misra, 2017) that was discussed briefly.

As high-dimensional machine learning methods for inference continue to impact economics, the results from this thesis suggest that it is natural to include the Bayesian paradigm into the movement when one seeks to estimate HDS models. Indeed, the spike-and-slab models were superior in estimating treatment effects; they were superior in settings where the frequentist optimization routine became troublesome; and they are likely also strong candidates in many other high-dimensional settings that this thesis did cover, but hopefully other researchers will.

References

Angrist, J. D. and Keueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4):979–1014.

Athey, S. (2018). The impact of machine learning on economics. In *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

Barbieri, M. M., Berger, J. O., et al. (2004). Optimal predictive model selection. *The annals of statistics*, 32(3):870–897.

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

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

Belloni, A., Chernozhukov, V., et al. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547.

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

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

Bouguila, N., Ziou, D., and Monga, E. (2006). Practical bayesian estimation of a finite beta mixture through gibbs sampling and its applications. *Statistics and Computing*, 16(2):215–225.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

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

Celeux, G. (1998). Bayesian inference for mixture: The label switching problem. In *Compstat*, pages 227–232. Springer.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

Dicker, L. H. (2012). Residual variance and the signal-to-noise ratio in high-dimensional linear models. *arXiv preprint arXiv:1209.0012*.

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 363–375.

Donaldson, D. and Storeygard, A. (2016). The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives*, 30(4):171–98.

Donohue III, J. J. and Levitt, S. D. (2001). The impact of legalized abortion on crime. *The Quarterly Journal of Economics*, 116(2):379–420.

Dubé, J.-P. and Misra, S. (2017). Scalable price targeting.

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

Foote, C. L. and Goetz, C. F. (2008). The impact of legalized abortion on crime: Comment. *The Quarterly Journal of Economics*, 123(1):407–423.

Frisch, D. K. and Melchiorsen, F. (2018). Seminar paper: Scalable price targetting. *Seminar Paper at University of Copenhagen, https://github.com/DitlevF/Scalable-Price-Targetting*.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409.

George, E. and Foster, D. P. (2000). Calibration and empirical bayes variable selection. *Biometrika*, 87(4):731–747.

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

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

Geweke, J. (1996). Variable selection and model comparison in regression. *In Bayesian Statistics 5*.

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. (2016). Counterfactual prediction with deep instrumental variables networks. *arXiv preprint arXiv:1612.09596*.

Hastie, T., Robert, T., and JH, F. (2009). The elements of statistical learning: data mining, inference, and prediction.

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

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29.

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

Kahane, L. H., Paton, D., and Simmons, R. (2008). The abortion–crime link: Evidence from england and wales. *Economica*, 75(297):1–21.

Khalili, A. and Chen, J. (2007). Variable selection in finite mixture of regression models. *Journal of the american Statistical association*, 102(479):1025–1038.

Lee, K.-J., Chen, R.-B., and Wu, Y. N. (2016). Bayesian variable selection for finite mixture model of linear regressions. *Computational Statistics & Data Analysis*, 95:1–16.

Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association*, 89(428):1535–1546.

Madigan, D., York, J., and Allard, D. (1995). Bayesian graphical models for discrete data. *International Statistical Review/Revue Internationale de Statistique*, pages 215–232.

Malsiner-Walli, G. and Wagner, H. (2011). Comparing spike and slab priors for bayesian variable selection. *Austrian Journal of Statistics*, 40(4):241–264.

McLachlan, G. and Peel, D. (2000). Finite mixture models, willey series in probability and statistics.

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

Mullainathan, S. and Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

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

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

Scott, S. L. and Varian, H. R. (2015). Bayesian variable selection for nowcasting economic time series. In *Economic analysis of the digital economy*, pages 119–135. University of Chicago Press.

Smith, M. and Kohn, R. (1996). Nonparametric regression using bayesian variable selection. *Journal of Econometrics*, 75(2):317–343.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398):528–540.

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

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2):3–28.

Wager, S. and Athey, S. (2017). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, (just-accepted).

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

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zellner, A. (1986). On assessing prior distributions and bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques*.

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

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

9 Appendix

9.1 Tables

9.1.1 Variable Selection for Independent Regressors

Table 9.1.1
Independent Data, Model Selection Performance for $N = p = 40$:
Number of Data Sets out of $n = 100$ with Regressors Included

		Strong signal, $\sigma^2 = 0.1$						*Weak signal*, $\sigma^2 = 1$					
		x_1	x_2	x_3	x_4	x_5	$\bar{x}_{6:40}$	x_1	x_2	x_3	x_4	x_5	$\bar{x}_{6:40}$
	Coefficient, β_j	1	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	0	1	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	0
$c = 1$	Ind. incl. prior	100	100	100	97	90	7.7	100	85	58	43	30	20.7
	Common prior	100	100	99	89	72	2.7	97	45	21	12	7	3.0
	Mixture prior	100	100	100	97	84	7.0	100	84	58	41	32	20.1
$c \approx 6$	Ind. incl. prior	100	100	100	98	82	9.1	98	72	45	40	22	16.0
	Common prior	100	100	98	91	62	2.8	97	30	13	3	5	1.5
	Mixture prior	100	100	100	98	81	8.3	97	70	46	40	26	20.7
$c = 100$	Ind. incl. prior	53	21	21	15	10	0.3	48	22	18	8	3	1.7
	Common prior	50	18	12	9	3	0.9	44	16	16	8	2	3.8
	Mixture prior	50	43	38	35	30	12.0	45	41	34	22	17	14.2
	Post-Lasso, r	100	99	94	70	52	4.4	92	30	14	4	4	0.8
	Post-Lasso, CV	100	100	100	92	87	13.4	94	53	34	18	13	6.2

Note - The table reports the number of times the regressors were included in the model over $n = 100$ data sets with correlated regressors, using the median probabilty model criterion Barbieri et al. (2004) to select variables for the Bayesian methods. All prior slabs were the independence slab with tuning parameter $c \in \{1, \sqrt{N} \approx 6, 100\}$, and the priors indicated in the table refer to the prior put on the inclusion probability ω. The individual inclusion probability prior is $\omega_j \sim \text{Beta}(\frac{1}{2}, \frac{1}{2})$, the common prior is $\omega \sim \text{Beta}(1, 1)$ and the mixture prior is the prior outlined in Section **??**. "Post-Lasso, r" refers to using the theoretically justified value of λ as tuning parameter. "Post-Lasso, CV" uses 10-fold cross-validation. $\bar{x}_{6:40}$ is the average number of times each of the 35 redundant regressors entered the model.

9.1.2 Variable Selection for Correlated Regressors

Table 9.1.2
Correlated Data, Model Selection Performance for $N = p = 40$:
Number of Data Sets out of $n = 100$ with Regressors Included

		Strong signal, $\sigma^2 = 0.1$						*Weak signal*, $\sigma^2 = 1$					
		x_1	x_2	x_3	x_4	x_5	$\bar{x}_{6:40}$	x_1	x_2	x_3	x_4	x_5	$\bar{x}_{6:40}$
	Coefficient, β_j	1	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	0	1	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	0
$c = 1$	Ind. incl. prior	100	100	100	90	86	8.0	100	78	63	59	41	19.7
	Common prior	100	100	97	83	62	2.2	98	52	36	31	11	2.9
	Mixture prior	100	100	100	98	89	7.2	100	75	63	54	39	19.6
$c \approx 6$	Ind. incl. prior	100	100	97	85	85	9.6	99	73	51	39	25	13.8
	Common prior	100	100	89	77	61	2.5	95	48	33	21	7	1.8
	Mixture prior	100	100	95	85	82	8.9	99	69	52	44	21	15.6
$c = 100$	Ind. incl. prior	50	22	19	12	7	0.5	45	23	16	7	3	2.2
	Common prior	49	15	11	7	1	1.0	46	21	17	6	2	4.2
	Mixture prior	46	40	36	35	32	9.5	42	40	35	25	24	10.2
	Post-Lasso, r	100	100	100	98	76	3.6	100	80	60	35	23	1.9
	Post-Lasso, CV	100	100	100	99	91	8.8	100	89	74	48	29	6.0

Note - The table reports the number of times the regressors were included in the model over $n = 100$ data sets with correlated regressors, using the median probabilty model criterion Barbieri et al. (2004) to select variables for the Bayesian methods. All prior slabs were the independence slab with tuning parameter $c \in \{1, \sqrt{N} \approx 6, 100\}$, and the priors indicated in the table refer to the prior put on the inclusion probability ω. The individual inclusion probability prior is $\omega_j \sim \text{Beta}(\frac{1}{2}, \frac{1}{2})$, the common prior is $\omega \sim \text{Beta}(1, 1)$ and the mixture prior is the prior outlined in Section **??**. "Post-Lasso, r" refers to using the theoretically justified value of λ as tuning parameter. "Post-Lasso, CV" uses 10-fold cross-validation. $\bar{x}_{6:40}$ is the average number of times each of the 35 redundant regressors entered the model.

9.2 Deriving the Marginal Likelihood

The likelihood can be written as

$$L(\sigma^2, \beta, \mu, \mid y, X) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2}(y - \mathbf{1}\mu - X\beta)'(I\sigma^2)^{-1}(y - \mathbf{1}\mu - X\beta)\right\} =$$
$$(2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2}(y_c + \mathbf{1}\bar{y} - \mathbf{1}\mu - X\beta)'(y_c + \mathbf{1}\bar{y} - \mathbf{1}\mu - X\beta)\right\} =$$
$$(2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[(y_c - X\beta)'(y_c - X\beta) + (\mathbf{1}\bar{y})'(\mathbf{1}\bar{y}) - (\mathbf{1}\bar{y})'(\mathbf{1}\mu) - ....\right]\right\} =$$
$$(2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[(y_c - X\beta)'(y_c - X\beta) + N(\bar{y} - \mu)^2\right]\right\} \quad (57)$$

where I have used that $X'\mathbf{1} = 0$ and defined $y_c = y - \mathbf{1}\bar{y}$ (where $\bar{y}$ is the mean of y), such that $y_c'(\mathbf{1}\bar{y} - \mathbf{1}\mu) = \beta' X'(\mathbf{1}\bar{y} - \mathbf{1}\mu) = 0$.

Step I: Integrating out μ

Using the expression in (57) and the Jeffreys prior assumption $p(\mu) \propto 1$, marginalizing over μ yields

$$p(\mathbf{y} \mid \sigma^2, \beta, X) = \int p(y \mid \sigma^2, \mu, \beta, X)p(\mu)d\mu =$$
$$(2\pi\sigma^2)^{-\frac{N}{2}} \exp\left(-\frac{1}{2\sigma^2}(y_c - X\beta)'(y_c - X\beta)\right) \int \exp\left\{-\frac{N}{2\sigma^2}\left[(\bar{y} - \mu\right]^2\right\} d\mu =$$
$$(2\pi\frac{\sigma^2}{N})^{\frac{1}{2}} \times (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left(-\frac{1}{2\sigma^2}(y_c - X\beta)'(y_c - X\beta)\right) \int (2\pi\frac{\sigma^2}{N})^{-\frac{1}{2}} \exp\left\{-\frac{N}{2\sigma^2}\left[(\mu - \bar{y}\right]^2\right\} d\mu =$$
$$\frac{1}{\sqrt{N}(2\pi\sigma^2)^{\frac{N-1}{2}}} \exp\left(-\frac{1}{2\sigma^2}(y_c - X\beta)'(y_c - X\beta)\right)$$

Step II: Integrating out β

With $\beta \mid \sigma^2 \sim N(\boldsymbol{a_0}, \boldsymbol{A}_0\sigma^2)$, marginalizing over β yields

$$p(y \mid \sigma^2, X) = \int p(y, \beta \mid \sigma^2, X)d\beta = \int p(y \mid \sigma^2, \beta, X)p(\beta \mid \sigma^2)d\beta =$$
$$\underbrace{\frac{1}{\sqrt{N}(2\pi\sigma^2)^{\frac{N-1}{2}}}}_{\Delta} \times \mid 2\pi\mathbf{A_0}\sigma^2 \mid^{-\frac{1}{2}} \int \exp\left(-\frac{1}{2\sigma^2}(y_c - X\beta)'(y_c - X\beta)\right) \times$$
$$\exp\left(\underbrace{-\frac{1}{2}(\beta - \mathbf{a}_0)'(\mathbf{A}_0\sigma^2)^{-1}(\beta - \mathbf{a}_0)}_{\Omega}\right) d\beta =$$
$$\Delta \times \mid 2\pi\mathbf{A_0}\sigma^2 \mid^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}\left[y_c'y_c\right]\right) \int \exp\left(-\frac{1}{2\sigma^2}\left[-\beta'Xy_c - y_c'X'\beta + \beta'X'X\beta\right]\right) \exp(\Omega)\, d\beta \quad (58)$$

Expanding the two exponential terms in the integral, we obtain

$$= \Delta \times \mid 2\pi\mathbf{A_0}\sigma^2 \mid^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}\left[y_c'y_c + \mathbf{a}_0'\mathbf{A}_0^{-1}\mathbf{a}_0\right]\right) \times$$
$$\int \exp\left(-\frac{1}{2\sigma^2}\left[-\beta'X'y_c - y_c'X'\beta + \beta'X'X\beta + \beta'\mathbf{A}_0^{-1}\beta - \beta'\mathbf{A}_0^{-1}\mathbf{a}_0 - \mathbf{a}_0\mathbf{A}_0^{-1}\beta\right]\right) d\beta$$

Now, ordering terms in the exponential yields

$$\int \exp\left(-\frac{1}{2\sigma^2}\left[-\beta'X'y_c - y_c'X'\beta + \beta'X'X\beta + \beta'\mathbf{A}_0^{-1}\beta - \beta'\mathbf{A}_0^{-1}\mathbf{a}_0 - \mathbf{a}_0\mathbf{A}_0^{-1}\beta\right]\right) d\beta =$$
$$\int \exp\left(-\frac{1}{2\sigma^2}\left[-\beta'\left(X'y_c + \mathbf{A}_0^{-1}\mathbf{a}_0\right) - \left(y_c'X'\beta + \mathbf{a}_0\mathbf{A}_0^{-1}\right)\beta + \beta'\left(X'X + \mathbf{A_0^{-1}}\right)\beta\right]\right) d\beta =$$
$$\int \exp\left(-\frac{1}{2\sigma^2}\left[\beta'\underbrace{(X'X + A)}_{\mathbf{A}_N^{-1}}\beta - 2\beta'\underbrace{\left(X'y_c + \mathbf{A}_0^{-1}\mathbf{a}_0\right)}_{\mathbf{A}_N^{-1}\mathbf{a}_N}\right]\right) d\beta$$

which is clearly the kernel of a normal distribution. Thus, completing the distribution, the integral

can can be written as

$$\mid 2\pi \mathbf{A}_N \sigma^2 \mid^{\frac{1}{2}} \exp\left\{\frac{1}{2\sigma^2}\mathbf{a}_N' \mathbf{A}_N^{-1}\mathbf{a}_N\right\} \times$$

$$\int \mid 2\pi \mathbf{A}_N \sigma^2 \mid^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}\left[\beta' \mathbf{A}_N^{-1}\beta - 2\beta' \mathbf{A}_N^{-1} \underbrace{\mathbf{A}_N\left(X' y_c + \mathbf{A}_0^{-1}\mathbf{a}_0\right)}_{\mathbf{a}_N}\right]\right) \times \exp\left\{-\frac{1}{2\sigma^2}\mathbf{a}_N' \mathbf{A}_N^{-1}\mathbf{a}_N\right\} d\beta =$$

$$\exp\left\{-\frac{1}{2\sigma^2}\mathbf{a}_N' \mathbf{A}_N^{-1}\mathbf{a}_N\right\} \mid 2\pi \mathbf{A}_N \sigma^2 \mid^{\frac{1}{2}} \quad (59)$$

Combining the results of (58) and (59) yields

$$p(y \mid \sigma^2, X) = \frac{1}{\sqrt{N}(2\pi\sigma^2)^{\frac{N-1}{2}}} \times \mid 2\pi \mathbf{A_0} \sigma^2 \mid^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}\left[y_c' y_c + \mathbf{a}_0' \mathbf{A}_0^{-1}\mathbf{a}_0 - \mathbf{a}_N' \mathbf{A}_N^{-1}\mathbf{a}_N\right]\right) \mid 2\pi \mathbf{A}_N \sigma^2 \mid^{\frac{1}{2}}$$

which can be simplified to

$$p(y \mid \sigma^2, X) = \frac{1}{\sqrt{N}(2\pi\sigma^2)^{\frac{N-1}{2}}} \times \frac{\mid \mathbf{A}_N \mid^{\frac{1}{2}}}{\mid \mathbf{A}_0 \mid^{\frac{1}{2}}} \exp\left(-\frac{S_N}{\sigma^2}\right) \quad (60)$$

where

$$\begin{aligned} S_N &= \frac{1}{2}\left[y_c' y_c + \mathbf{a}_0' \mathbf{A}_0^{-1}\mathbf{a}_0 - \mathbf{a}_N' \mathbf{A}_N^{-1}\mathbf{a}_N\right] \\ \mathbf{A}_N &= \left(X'X + A\right)^{-1} \\ \mathbf{a}_N &= \mathbf{A}_N\left(X' y_c + \mathbf{A}_0^{-1}\mathbf{a}_0\right) \end{aligned}$$

Step III: Integrating out σ^2

Using the Jeffreys prior $p(\sigma^2) \propto \frac{1}{\sigma^2}$, we obtain

$$p(y \mid X) = \int p(y, \sigma^2 \mid X) d\sigma^2 = \int p(y \mid \sigma^2, X) p(\sigma^2) d\sigma^2 =$$

$$\int \frac{1}{\sqrt{N}(2\pi\sigma^2)^{\frac{N-1}{2}}} \times \frac{\mid \mathbf{A}_N \mid^{\frac{1}{2}}}{\mid \mathbf{A}_0 \mid^{\frac{1}{2}}} \exp\left(-\frac{S_N}{\sigma^2}\right) \frac{1}{\sigma^2} d\sigma^2 = \frac{1}{\sqrt{N}(2\pi)^{\frac{N-1}{2}}} \times \frac{\mid \mathbf{A}_N \mid^{\frac{1}{2}}}{\mid \mathbf{A}_0 \mid^{\frac{1}{2}}} \int (\sigma^2)^{-\frac{N-1}{2}-1} \exp\left(-\frac{S_N}{\sigma^2}\right) d\sigma^2 =$$

$$\frac{1}{\sqrt{N}(2\pi)^{\frac{N-1}{2}}} \times \frac{\mid \mathbf{A}_N \mid^{\frac{1}{2}}}{\mid \mathbf{A}_0 \mid^{\frac{1}{2}}} \left(\frac{S_N^{\frac{N-1}{2}}}{\Gamma(\frac{N-1}{2})}\right)^{-1} \int \frac{S_N^{\frac{N-1}{2}}}{\Gamma(\frac{N-1}{2})} \exp\left(-\frac{S_N}{\sigma^2}\right) (\sigma^2)^{-\frac{N-1}{2}-1} d\sigma^2 =$$

$$\frac{1}{\sqrt{N}(2\pi)^{s_N}} \times \frac{\mid \mathbf{A}_N \mid^{\frac{1}{2}}}{\mid \mathbf{A}_0 \mid^{\frac{1}{2}}} \left(\frac{S_N^{s_N}}{\Gamma(s_N)}\right)^{-1}$$

where $s_N = (N-1)/2$. The log-marginal likelihood then is

$$\log p(y \mid X) = -0.5 \log N - s_N \log(2\pi) + \frac{1}{2} \left(\log \mid \mathbf{A}_N \mid - \log \mid \mathbf{A}_0 \mid\right) - \left(s_N \log(S_N) - \log(\Gamma(s_N))\right)$$

9.3 Derivations for MCMC with an Absolutely Continuous Spike

The conditional posteriors of μ and ω_j are the same as with a Dirac delta spike, and the following derivations outline the conditional posteriors for the remaining parameters. For simpler derivations, note that (27) can be written hierarchically as

$$\beta_j \mid \gamma_j, \tau_j^2 \sim N(0, c_j^2(\gamma_j)\tau_j^2) \tag{61}$$

$$c_j^2(\gamma_j) = \begin{cases} c_j^2, \text{ if } \gamma_j = 1 \\ 1, \text{ if } \gamma_j = 0 \end{cases}$$

Conditional posterior of γ_j

As the Markov chain does not have absorbing states when $\beta_j = 0$ is sampled, we can update γ_j conditional on β_j and ω_j

$$\begin{aligned} p\left(\gamma_j = 1|\beta_j, \omega_j\right) &= \frac{p\left(\beta_j|\gamma_j = 1, \omega_j\right) p\left(\gamma_j = 1, \beta_j|\omega_j\right)}{p\left(\beta_j|\gamma_j = 1, \omega_j\right) p\left(\gamma_j = 1, \beta_j|\omega_j\right) + p\left(\beta_j|\gamma_j = 0, \omega_j\right) p\left(\gamma_j = 0, \beta_j|\omega_j\right)} \\ &= \frac{p_{slab}\left(\beta_j\right)\omega}{p_{slab}\left(\beta_j\right)\omega_j + p_{spike}\left(\beta_j\right)\left(1-\omega_j\right)} \\ &= \frac{1}{1 + L_j \frac{1-\omega_j}{\omega_j}}, \qquad\qquad L_j = \frac{p_{spike}\left(\beta_j\right)}{p_{slab}\left(\beta_j\right)} \end{aligned} \tag{62}$$

Conditional posterior of τ_j^2

Trivially, for SSVS set $\tau_j^2 = V$. For NMIG, update the prior $\tau_j^2 \sim \text{IG}(\nu, Q)$, using the formulation in (61), to

$$\begin{aligned} &p(\tau_j^2 \mid \gamma_j, \beta_j) \propto p(\beta_j \mid \gamma_j, \tau_j^2) p(\gamma_j) p(\tau_j^2) \propto p(\beta_j \mid \gamma_j, \tau_j^2) p(\tau_j^2) \propto \\ &\propto \frac{1}{\sqrt{2\pi c_j^2\left(\gamma_j\right)\tau_j^2}} \exp\left\{\frac{-\beta_j^2}{2c_j^2\left(\gamma_j\right)\tau_j^2}\right\} \times \frac{Q^\nu}{\Gamma\left(\nu\right)} (\tau_j^2)^{-\nu-1} \exp\left\{-\frac{Q}{\tau_j^2}\right\} \propto \\ &\propto \frac{1}{\sqrt{2\pi c_j^2\left(\gamma_j\right)}} (\tau_j^2)^{-\frac{1}{2}-\nu-1} \exp\left\{-\frac{1}{\tau_j^2}\left(\frac{\beta_j^2}{2c_j^2\left(\gamma_j\right)} + Q\right)\right\} \Rightarrow \\ &\tau_j^2|\beta_j, \gamma_j \sim \mathcal{G}\left(\nu + 1, \frac{\beta_j^2}{2c_j^2\left(\gamma_j\right)} + Q\right) \end{aligned} \tag{63}$$

Conditional posterior of β_j

In matrix notation, we can write the prior on $\beta = (\beta_1, ..., \beta_p)'$ as $\beta \mid \gamma, \tau \sim N(\mathbf{0}, \mathbf{D})$, where

$$\boldsymbol{D} = \begin{bmatrix} c_1^2\left(\gamma_1\right)\tau_1^2 & 0 & \dots & 0 \\ 0 & c_2^2\left(\gamma_2\right)\tau_2^2 & & \vdots \\ \vdots & & & 0 \\ 0 & \cdots & 0 & c_p^2\left(\gamma_p\right)\tau_p^2 \end{bmatrix}$$

Then, using the likelihood formulation in (57), we obtain

$$
\begin{aligned}
&p(\beta \mid y, \gamma, \tau, \sigma^2) \propto p(y \mid \beta, \gamma, \sigma^2) p(\beta \mid \gamma, \tau) \propto \\
&\exp\left\{-\frac{1}{2\sigma^2}\left[(y_c - X\beta)'(y_c - X\beta)\right]\right\} \times \exp\left\{-\frac{1}{2}\left[\beta' \boldsymbol{D}^{-1} \beta\right]\right\} \propto \\
&\exp\left\{-\frac{1}{2\sigma^2}\left[(y_c' y_c - y_c X\beta - \beta X' y_c + \beta' X' X \beta)\right] - \frac{1}{2}\beta' (\boldsymbol{D})^{-1} \beta\right\} \propto \\
&\exp\left\{-\frac{1}{2}\left[\beta' \underbrace{\left(\frac{X'X}{\sigma^2} + \boldsymbol{D}^{-1}\right)}_{A_N^{-1}} \beta - 2\beta' \underbrace{\frac{X' y_c}{\sigma^2}}_{A_N^{-1} a_N}\right]\right\}
\end{aligned}
$$

which is the kernel of a normal distribution, such that

$$
\beta \mid y, \gamma, \tau^2, \sigma^2 \sim \mathcal{N}(a_N, A_N) \tag{64}
$$

with

$$
A_N = \left(\frac{X'X}{\sigma^2} + \boldsymbol{D}^{-1}\right)^{-1}
$$

$$
a_N = A_N \frac{X' y_c}{\sigma^2}
$$

Conditional posterior of σ^2

The posterior of σ^2 follows from its Jeffreys prior $p(\sigma^2) \propto \frac{1}{\sigma^2}$ updated via the likelihood with μ integrated out

$$
\begin{aligned}
&p(\sigma^2 \mid y, \beta) \propto p(y \mid \beta, \sigma^2) p(\beta \mid \sigma^2) p(\sigma^2) \propto p(y \mid \beta, \sigma^2) p(\sigma^2) \propto \\
&\quad \propto \frac{1}{\sqrt{N}(2\pi\sigma^2)^{\frac{N-1}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\left[(y_c - X\beta)'(y_c - X\beta)\right]\right\} \frac{1}{\sigma^2} \propto \\
&\quad \exp\left\{-\frac{(y_c - X\beta)'(y_c - X\beta)}{2\sigma^2}\right\} \times \left(\sigma^2\right)^{-\frac{N-1}{2}-1} \Rightarrow \\
&\quad \sigma^2 \mid y, \beta \sim \mathrm{IG}(s_N, SN)
\end{aligned} \tag{65}
$$

with

$$S_N \equiv \frac{(y_c - X\beta)'(y_c - X\beta)}{2}$$

$$s_N \equiv (N-1)/2$$

9.4 Metropolis-Hastings Step for the Beta Mixture Prior

The steps outlined here are paraphrased from Bouguila et al. (2006), and for a more detailed explanation the reader is referred to that paper. To sample $\theta_n = (a_m, b_m)$, the following representation is employed, where

$$s_m = a_m + b_m$$

$$m_m = a_m / s_m$$

such that $p(\omega_j \mid s_m, m_m, z_{jm} = 1) \propto \text{Beta}(s_m m_m, s_m(1 - m_m))$. The priors are then given

$$p(s_m, m_m) \propto \left[1 - \exp\left(-\delta((s_m - 2)^2 + (m_m - 0.5)^2)\right)\right) \exp\left(\frac{-\rho}{s_m^2 m_m (1 - m_m)}\right) - \kappa s_m^2/2\right]$$

where δ, ρ and κ are hyperparameters, and the choice of prior is motivated by the desire to avoid $(s_m, m_m) = (2, 0.5) \Leftrightarrow (a_m, b_m) = (1, 1)$, i.e. a uniform distribution. Because $m_m \in [0, 1]$, the following transformation is used $T(m_m^*) = \exp(m_m^*/(1 + m_m^*)$ where $m_m^* = m_m/(1 - m_m)$ for sampling efficiency, and the proposal densities are

$$\tilde{s}_m \sim LN\left(\log(s_m^{(t-1)}), \sigma^2\right)$$
$$\tilde{m}_m^* \sim LN\left(\log(m_m^{*(t-1)}), \sigma^2\right)$$

where LN is the log-normal distribution. The Metropolis-Hastings sampling is then composed of the following steps:

1. Generate $\tilde{s}_m \sim LN\left(\log(s_m^{(t-1)}), \sigma^2\right)$, $\tilde{m}_m^* \sim LN\left(\log(m_m^{*(t-1)}), \sigma^2\right)$ and $u \sim \text{Unif}(0, 1)$

2. Compute the acceptance probability using the sigmoid transformation $T(m^*_m) = \exp(m^*_m)/(1 + \exp(m^*_m))$

$$r = \frac{p(\tilde{s}_m, \tilde{m^*}_m \mid z, \omega) \times LN(m_m^{*(t-1)} \mid \log(\tilde{m^*}_m), \sigma^2) \times LN(s_m^{(t-1)} \mid \log(\tilde{s}_m), \sigma^2)}{p(s_m^{(t-1)}, m_m^{*(t-1)} \mid z, \omega) \times LN(\tilde{m^*}_m \mid \log(m_m^{*(t-1)}), \sigma^2) \times LN(\tilde{s}_m \mid \log(s_m^{(t-1)}), \sigma^2)}$$

where

$$p(s_m, m^*_m \mid z, \omega) \propto \left[1 - \exp\left(-\delta((s_m - 2)^2 + (m^*_m - 0.5)^2)\right)\right) \exp\left(\frac{-\rho}{s_m^2 m^*_m(1 - m^*_m)}\right) - \kappa s_m^2/2\Big] \times \left(\frac{\Gamma(s_m)}{\Gamma(s_m T(m^*_m)\Gamma(s_m(1 - T(m^*_m)))}\right)^{n_m} \left(\prod_{z_{jm}=1} \omega_j\right)^{s_m T(m^*_m)-1} \times \left(\prod_{z_{jm}=1} (1 - \omega_j)\right)^{s_m(1-T(m^*_m))-1} \frac{\exp(m^*_m)}{(1 + \exp(m^*_m))^2}$$

where the last term is the Jacobian of the transformation needed for the change-of-variables formula to be applied

3. If $r < u$ let $(s_m^{(t)}, m_m^{(t)}) = (\tilde{s}_m, \tilde{m^*}_m)$ else $(s_m^{(t)}, m_m^{(t)}) = (s_m^{(t-1)}, m_m^{(t-1)})$

There are four hyperparameters to be tuned δ, ρ, κ and σ^2. In my experience, the choice of the first three do not impact the variable selection performance and I follow Bouguila et al. (2006) and set $(\delta, \rho, \kappa) = (3, 0.1, 0.00001)$. The choice of σ^2 affects the performance a lot, as it effectively decides the proposal jumps of the MH algorithm. Values in the range $[0.1, 0.3]$ provide reasonable acceptance rates and relatively stable performance, and I therefore employ $\sigma^2 = 0.1$ in the simulation studies.

9.5 Signal-to-Noise Ratio

I follow Dicker (2012) and define the signal-to-noise ratio in the linear model $y = X\beta + \epsilon$, $\epsilon \sim N(0, \sigma^2 I)$ as

$$\frac{\tau^2}{\sigma^2} = \frac{\beta' \Sigma \beta}{\text{Var}(\epsilon)}$$

where Σ is the positive definite variance-covariance matrix of X. Plugging (51) into (50), one obtains

$$y_i = \mu + (\alpha_0\beta_{d1} + \beta_{y1})x_{1i} + ... + (\alpha_0\beta_{dp} + \beta_{yp})x_{pi} + u_i, \equiv \mu + x_i\tilde{\beta} + u_i, \qquad u_i = \alpha_0\nu_i + \epsilon_i$$

leading to the following signal-to-noise ratio for the model

$$\frac{\tilde{\beta}'\Sigma\tilde{\beta}}{\alpha_0^2\sigma_\nu^2 + \sigma_\epsilon^2}$$