Bayesian subset selection and variable importance for interpretable prediction and classification

Daniel R. Kowal   DANIEL.KOWAL@RICE.EDU
Department of Statistics
Rice University
Houston, TX 77005, USA

Editor: Sayan Mukherjee

Abstract

Subset selection is a valuable tool for interpretable learning, scientific discovery, and data compression. However, classical subset selection is often avoided due to selection instability, lack of regularization, and difficulties with post-selection inference. We address these challenges from a Bayesian perspective. Given any Bayesian predictive model $\mathcal{M}$, we extract a *family* of near-optimal subsets of variables for linear prediction or classification. This strategy deemphasizes the role of a single "best" subset and instead advances the broader perspective that often many subsets are highly competitive. The *acceptable family* of subsets offers a new pathway for model interpretation and is neatly summarized by key members such as the smallest acceptable subset, along with new (co-) variable importance metrics based on whether variables (co-) appear in all, some, or no acceptable subsets. More broadly, we apply Bayesian decision analysis to derive the optimal linear coefficients for *any* subset of variables. These coefficients inherit both regularization and predictive uncertainty quantification via $\mathcal{M}$. For both simulated and real data, the proposed approach exhibits better prediction, interval estimation, and variable selection than competing Bayesian and frequentist selection methods. These tools are applied to a large education dataset with highly correlated covariates. Our analysis provides unique insights into the combination of environmental, socioeconomic, and demographic factors that predict educational outcomes, and identifies over 200 distinct subsets of variables that offer near-optimal out-of-sample predictive accuracy.

Keywords: education; linear regression; logistic regression; model selection; penalized regression

1. Introduction

Subset and variable selection are essential components of regression analysis, prediction, and classification. By identifying subsets of important covariates, the analyst can acquire simpler and more interpretable summaries of the data, improved prediction or classification, reduced estimation variability among the selected covariates, lower storage requirements, and insights into the factors that determine predictive accuracy (Miller, 1984). Classical subset selection is expressed as the solution to the constrained least squares problem

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \quad \text{subject to} \quad \|\boldsymbol{\beta}\|_0 \leq k \tag{1}$$

where $\boldsymbol{y}$ is an n-dimensional response, $\boldsymbol{X}$ is an $n \times p$ matrix of covariates, and $\boldsymbol{\beta}$ is the p-dimensional vector of unknown coefficients. Traditionally, the goal is to determine (i) the best subset of each size k, (ii) an estimate of the accompanying nonzero linear coefficients, and (iii) the best subset of any size. Subset selection has garnered additional attention recently, in part due to the algorithmic advancements from Bertsimas et al. (2016) and the detailed comparisons of Hastie et al. (2020). More broadly, variable selection has been deployed as a tool for interpretable machine learning, including for highly complex and nonlinear models (e.g., Ribeiro et al., 2016; Afrabandpey et al., 2020).

Although often considered the "gold standard" of variable selection, subset selection remains underutilized due to several critical limitations. First, subset selection is inherently unstable: it is common to obtain entirely distinct subsets under perturbations or resampling of the data. This instability undermines the interpretability of a single "best" subset, and is a significant motivating factor for the proposed methods. Second, the solutions to (1) are unregularized. While it is advantageous to avoid *overshrinkage*, Hastie et al. (2020) showed that the lack of any regularization in (1) leads to deteriorating performance relative to penalized regression in low-signal settings. Third, inference about $\boldsymbol{\beta}$ requires careful adjustment for selection bias, which limits the available options for uncertainty quantification. Finally, solving (1) is computationally demanding even for moderate p, which has spawned many algorithmic advancements spanning multiple decades (e.g., Furnival and Wilson, 2000; Gatu and Kontoghiorghes, 2006; Bertsimas et al., 2016). For these reasons, penalized regression techniques that replace the ℓ_0-penalty with convex or nonconvex yet computationally feasible alternatives (Fan and Lv, 2010) are often preferred.

From a Bayesian perspective, (1) is usually translated into a predictive model via a Gaussian (log-) likelihood and a sparsity (log-) prior. Indeed, substantial research efforts have been devoted to both sparsity (e.g., Ishwaran and Rao, 2005) and shrinkage (e.g., Polson and Scott, 2010) priors for $\boldsymbol{\beta}$. Yet the prior alone cannot *select* subsets: the prior is the component of the data-generating process that incorporates prior beliefs, information, or regularization, while selection is ultimately a decision problem (Lindley, 1968). For instance, Barbieri and Berger (2004) and Liang et al. (2013) applied sparsity priors to obtain posterior inclusion probabilities, which were then used for *marginal* selection and screening, respectively. Jin and Goh (2020) selected subsets using marginal likelihoods, but required conjugate priors for Gaussian linear models. Hahn and Carvalho (2015) more fully embraced the decision analysis approach for variable selection, and augmented a squared error loss with an ℓ_1-penalty for linear variable selection. Alternative loss functions have been proposed for seemingly unrelated regressions (Puelz et al., 2017), graphical models (Bashir et al., 2019), nonlinear regressions (Woody et al., 2020), functional regression (Kowal and Bourgeois, 2020), time-varying parameter models (Huber et al., 2020), and a variety of Kullback-Leibler approximations to the likelihood (Goutis and Robert, 1998; Nott and Leng, 2010; Tran et al., 2012; Piironen et al., 2020).

Despite the appropriate deployment of decision analysis for selection, each of these Bayesian methods relies on ℓ_1-penalization or forward search. As such, they are restricted to limited search paths that cannot fully solve the (exhaustive) subset selection problem. More critically, these Bayesian approaches—as well as most classical ones—are unified in their emphasis on selecting a *single* "best" subset. However, in practice it is common for many subsets (or models) to achieve near-optimal predictive performance, known as the

Rashomon effect (Breiman, 2001). This effect is particularly pronounced for correlated covariates, weak signals, or small sample sizes. Under these conditions, it is empirically and theoretically possible for the "true" covariates to be predictively inferior to a proper subset (Wu et al., 2007). As a result, the "best" subset is not only less valuable but also less interpretable. Reporting a single subset—or a small number of subsets along a highly restricted search path—obscures the likely presence of many distinct yet equally-predictive subsets of variables.

We advance an agenda that instead curates and summarizes a *family* of optimal or near-optimal subsets. This broader analysis alleviates the instability issues of a single "best" subset and provides a more complete predictive picture. The proposed approach operates within a decision analysis framework and is compatible with *any* Bayesian model $\mathcal{M}$ for prediction or classification. Naturally, $\mathcal{M}$ should represent the modeler's beliefs about the data-generating process and describe the salient features in the data. Several key developments are required:

1. We derive *optimal* (in a decision analysis sense) linear coefficients for any subset of variables.

Crucially, these coefficients inherit regularization *and* uncertainty quantification via $\mathcal{M}$, but avoid the overshrinkage induced by ℓ_1-penalization. As such, these point estimators resolve multiple limitations of classical subset selection and (ℓ_1-penalized) Bayesian decision-analytic variable selection, and further are adapted to both regression and classification problems. Next,

2. We design a modified branch-and-bound algorithm for *efficient exploration* over the space of candidate subsets.

The search process is a vital component of subset selection, and our modular framework is broadly compatible with other state-of-the-art search algorithms (e.g., Bertsimas et al., 2016). Until now, these algorithms have not been fully deployed or adapted for Bayesian subset selection. Additionally,

3. We leverage the predictive distribution under $\mathcal{M}$ to collect the *acceptable family* of near-optimal subsets.

A core feature of the acceptable family is that it is defined using out-of-sample metrics and predictive uncertainty quantification, yet is computed using *in-sample posterior functionals* from a single model fit of $\mathcal{M}$. Hence, we maintain computational scalability and coherent uncertainty quantification that avoids data reuse. Lastly,

4. We *summarize* the acceptable family with key member subsets, such as the "best" (in terms of cross-validation error) predictor and the smallest acceptable subset, along with new (co-) variable importance metrics that measure the frequency with which variables (co-) appear in all, some, or no acceptable subsets.

Unlike variable importances based on effect size, this inclusion-based metric effectively measures how many "predictively plausible" explanations (i.e., near-optimal subsets) contain each (pair of) covariate(s) as a member. Notably, each of these developments is presented for both prediction and classification.

The importance of curating and exploring a collection of subsets has been acknowledged previously. Existing approaches are predominantly frequentist, including fence methods (Jiang et al., 2008), Rashomon sets (Semenova and Rudin, 2019), bootstrapped confidence sets (Lei, 2019), and subsampling-based forward selection (Kissel and Mentch, 2021). Although the acceptable family has appeared previously for Bayesian decision analysis (Kowal, 2021; Kowal et al., 2021), it was applied only along the ℓ_1-path which does not enumerate a sufficiently rich collection of competitive subsets. Further, previous applications of the acceptable family did not address points 1., 2., and 4. above.

The paper is outlined as follows. Section 2 contains the Bayesian subset search procedures, the construction of acceptable families, the (co-) variable importance metrics, and the predictive uncertainty quantification. Section 3 details the simulation study. The methodology is applied to a large education dataset in Section 4. We conclude in Section 5. The Appendix provides additional algorithmic details, simulation studies, and results from the application. An `R` package is available online at `https://github.com/drkowal/BayesSubsets`. Although the education dataset (Children's Environmental Health Initiative, 2020) cannot be released due to privacy protections, access to the dataset can occur through establishing affiliation with the Children's Environmental Health Initiative (contact cehi@nd.edu).

2. Methods

2.1 Predictive decision analysis

Decision analysis establishes the framework for extracting actions, estimators, and predictions from a Bayesian model (e.g., Bernardo and Smith, 2009). These tools translate probabilistic models into practical decision-making and can be deployed to summarize or interpret complex models. However, additional methodology is needed to convert subset selection into a decision problem, and further to evaluate and collect many near-optimal subsets.

Let $\mathcal{M}$ denote any Bayesian model with a proper posterior distribution $p_{\mathcal{M}}(\boldsymbol{\theta}|\boldsymbol{y})$ and posterior predictive distribution

$$p_{\mathcal{M}}(\tilde{\boldsymbol{y}}|\boldsymbol{y}) \coloneqq \int p_{\mathcal{M}}(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}, \boldsymbol{y}) p_{\mathcal{M}}(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta},$$

where $\boldsymbol{\theta}$ denotes the parameters of the model $\mathcal{M}$ and $\boldsymbol{y}$ is the observed data. Informally, $p_{\mathcal{M}}(\tilde{\boldsymbol{y}}|\boldsymbol{y})$ defines the distribution of future or unobserved data $\tilde{\boldsymbol{y}}$ conditional on the observed data $\boldsymbol{y}$ and according to the model $\mathcal{M}$. Decision analysis evaluates each *action* $\boldsymbol{\delta}$ based on a loss function $\mathcal{L}(\tilde{\boldsymbol{y}}, \boldsymbol{\delta})$ that enumerates the cost of each action when $\tilde{\boldsymbol{y}}$ is realized. Examples include point prediction (e.g., squared error loss) or classification (e.g., cross-entropy loss), interval estimation (e.g., minimum length subject to $1 - \alpha$ coverage), and selection among a set of hypotheses (e.g., 0-1 loss). Since $\tilde{\boldsymbol{y}}$ is unknown yet modeled probabilistically under $\mathcal{M}$, an optimal action minimizes the posterior predictive expected loss

$$\hat{\boldsymbol{\delta}} \coloneqq \arg\min_{\boldsymbol{\delta}} \mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{y}} \mathcal{L}(\tilde{\boldsymbol{y}}, \boldsymbol{\delta}) \tag{2}$$

with the expectation taken under the Bayesian model $\mathcal{M}$. The operation in (2) averages the predictive loss over the possible realizations of $\tilde{\boldsymbol{y}}$ according to the posterior probability under $\mathcal{M}$ and then minimizes the resulting quantity over the action space.

Yet without careful specification of the loss function $\mathcal{L}$, (2) does not provide a clear pathway for subset selection. To see this, let $\tilde{\boldsymbol{y}}(\boldsymbol{x})$ denote the predictive variable at covariate $\boldsymbol{x}$. For point prediction of $\tilde{\boldsymbol{y}}(\boldsymbol{x})$ under squared error loss, the optimal action is

$$\hat{\boldsymbol{\delta}}(\boldsymbol{x}) \coloneqq \arg\min_{\boldsymbol{\delta}} \mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{y}} \|\tilde{\boldsymbol{y}}(\boldsymbol{x}) - \boldsymbol{\delta}(\boldsymbol{x})\|_2^2$$
$$= \mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{y}}\{\tilde{\boldsymbol{y}}(\boldsymbol{x})\},$$

i.e., the posterior predictive expectation at $\boldsymbol{x}$. Similarly, for classification of $\tilde{\boldsymbol{y}}(\boldsymbol{x}) \in \{0,1\}$ under cross-entropy loss (see (10)), the optimal action is the posterior predictive probability $\hat{\boldsymbol{\delta}}(\boldsymbol{x}) = p_{\mathcal{M}}\{\tilde{\boldsymbol{y}}(\boldsymbol{x}) = 1|\boldsymbol{y}\}$. For a generic model $\mathcal{M}$, there is not necessarily a closed form for $\hat{\boldsymbol{\delta}}(\boldsymbol{x})$: these actions are computed separately for each $\boldsymbol{x}$ with no clear mechanism for inducing sparsity or specifying distinct subsets. Hence, additional techniques are needed to supply actions that are not only optimal but also selective and interpretable.

Note that we use observation-driven rather than parameter-driven loss functions. Unlike the parameters $\boldsymbol{\theta}$, which are unobservable and model-specific, the predictive variables $\tilde{\boldsymbol{y}}$ are observable and directly comparable across distinct Bayesian models. A decision analysis based on $\tilde{\boldsymbol{y}}$ operates on the same scale and in the same units as the data that have been—and will be—observed, which improves interpretability. Perhaps most important, an observation-driven decision analysis enables empirical evaluation of the selected actions.

2.2 Subset search for linear prediction

The subset search procedure is built within the decision framework of (2). For any Bayesian model $\mathcal{M}$, we consider *linear* actions $\boldsymbol{\delta}(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\delta}$ with $\boldsymbol{\delta} \in \mathbb{R}^p$, which offer both interpretability and the capacity for selection. Let $\boldsymbol{\delta}_{\mathcal{S}}$ denote the linear action with zero coefficients for all $j \notin \mathcal{S}$, where $\mathcal{S} \subseteq \{1, \ldots, p\}$ is a subset of active variables. For prediction, we assemble the aggregate and weighted squared error loss

$$\mathcal{L}(\{\tilde{y}_i\}_{i=1}^{\tilde{n}}, \boldsymbol{\delta}_{\mathcal{S}}) = \sum_{i=1}^{\tilde{n}} \omega(\tilde{\boldsymbol{x}}_i) \|\tilde{y}_i - \tilde{\boldsymbol{x}}_i'\boldsymbol{\delta}_{\mathcal{S}}\|_2^2 \tag{3}$$

where $\tilde{y}_i \coloneqq \tilde{\boldsymbol{y}}(\tilde{\boldsymbol{x}}_i)$ is the predictive variable at $\tilde{\boldsymbol{x}}_i$ for each $i = 1, \ldots, \tilde{n}$ and $\omega(\tilde{\boldsymbol{x}}_i) > 0$ is a weighting function. The covariate values $\{\tilde{\boldsymbol{x}}_i\}_{i=1}^{\tilde{n}}$ can be distinct from the observed covariate values $\{\boldsymbol{x}_i\}_{i=1}^{n}$, for example to evaluate the action for a different population.

The loss in (3) evaluates linear coefficients $\boldsymbol{\delta}_{\mathcal{S}}$ for any given subset $\mathcal{S}$ by accumulating the squared error loss over the covariate values $\{\tilde{\boldsymbol{x}}_i\}_{i=1}^{\tilde{n}}$. Since (3) depends on $\{\tilde{y}_i\}_{i=1}^{\tilde{n}}$, the loss inherits a joint predictive distribution $p_{\mathcal{M}}(\tilde{y}_1, \ldots, \tilde{y}_{\tilde{n}}|\boldsymbol{y})$ from $\mathcal{M}$. The loss is decoupled from the Bayesian model $\mathcal{M}$: the linear action does not require a linearity assumption for $\mathcal{M}$. The weights $\omega(\tilde{\boldsymbol{x}}_i)$ can be used to target actions locally, which provides a sparse and local linear approximation to $\mathcal{M}$. For example, we might parametrize the weighting function as $\omega(\tilde{\boldsymbol{x}}_i) \propto \exp(-\|\tilde{\boldsymbol{x}}_i - \boldsymbol{x}^*\|_2^2/\ell)$ with range parameter ℓ in order to weight based on proximity to some particular $\boldsymbol{x}^*$ of interest (Ribeiro et al., 2016). Alternatively, ω can be specified via a probability model for the likelihood of observing each covariate value,

including $\omega(\tilde{\boldsymbol{x}}_i) = \tilde{n}^{-1}$ as a simple yet useful example, especially when using the observed covariate values $\{\tilde{\boldsymbol{x}}_i\}_{i=1}^{\tilde{n}} = \{\boldsymbol{x}_i\}_{i=1}^{n}$; this is our default choice.

Crucially, the optimal action can be solved directly for any subset $\mathcal{S}$:

Lemma 1 *Suppose* $\mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{y}}\|\tilde{\boldsymbol{y}}(\tilde{\boldsymbol{x}}_i)\|_2^2 < \infty$ *for* $i = 1, \ldots, \tilde{n}$. *The optimal action* (2) *for the loss* (3) *is given by the nonzero entries*

$$\hat{\boldsymbol{\delta}}_{\mathcal{S}} = \arg\min_{\boldsymbol{\delta}_{\mathcal{S}}} \sum_{i=1}^{\tilde{n}} \omega(\tilde{\boldsymbol{x}}_i)\|\hat{y}_i - \tilde{\boldsymbol{x}}_i'\boldsymbol{\delta}_{\mathcal{S}}\|_2^2 \tag{4}$$

$$= (\tilde{\boldsymbol{X}}_{\mathcal{S}}'\boldsymbol{\Omega}\tilde{\boldsymbol{X}}_{\mathcal{S}})^{-1}\tilde{\boldsymbol{X}}_{\mathcal{S}}'\boldsymbol{\Omega}\hat{\boldsymbol{y}}, \tag{5}$$

with zeros for indices $j \notin \mathcal{S}$, *where* $\hat{y}_i := \mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{y}}\{\tilde{\boldsymbol{y}}(\tilde{\boldsymbol{x}}_i)\}$, $\hat{\boldsymbol{y}} := (\hat{y}_1, \ldots, \hat{y}_{\tilde{n}})'$, $\boldsymbol{\Omega} := diag\{\omega(\tilde{\boldsymbol{x}}_i)\}_{i=1}^{\tilde{n}}$, *and* $\tilde{\boldsymbol{X}}_{\mathcal{S}}$ *is the* $\tilde{n} \times |\mathcal{S}|$ *matrix of the active covariates in* $\{\tilde{\boldsymbol{x}}_i\}_{i=1}^{\tilde{n}}$ *for subset* $\mathcal{S}$.

Proof It is sufficient to observe that $\mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{y}}\|\tilde{\boldsymbol{y}}(\tilde{\boldsymbol{x}}_i) - \tilde{\boldsymbol{x}}_i'\boldsymbol{\delta}_{\mathcal{S}}\|_2^2 = \mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{y}}\|\{\tilde{\boldsymbol{y}}(\tilde{\boldsymbol{x}}_i) - \hat{y}_i\} + (\hat{y}_i - \tilde{\boldsymbol{x}}_i'\boldsymbol{\delta}_{\mathcal{S}})\|_2^2 = \mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{y}}\|\tilde{\boldsymbol{y}}(\tilde{\boldsymbol{x}}_i) - \hat{y}_i\|_2^2 + \|\hat{y}_i - \tilde{\boldsymbol{x}}_i'\boldsymbol{\delta}_{\mathcal{S}}\|_2^2$, where the first term is a (finite) constant that does not depend on $\boldsymbol{\delta}_{\mathcal{S}}$. The remaining steps constitute a weighted least squares solution. ∎

The consequence of Lemma 1 is that the optimal action for each subset $\mathcal{S}$ is simply the (weighted) least squares solution based on pseudo-data $(\tilde{\boldsymbol{X}}_{\mathcal{S}}, \hat{\boldsymbol{y}})$—i.e., a "fit to the fit" from $\mathcal{M}$. The advantages of $\mathcal{M}$ can be substantial: the Bayesian model propagates regularization (e.g., shrinkage, sparsity, or smoothness) to the point predictions $\hat{\boldsymbol{y}}$, which typically offers sizable improvements in estimation and prediction relative ordinary (weighted) least squares. This effect is especially pronounced in the presence of high-dimensional (large p) or correlated covariates. The optimal action may be non-unique if $\tilde{\boldsymbol{X}}_{\mathcal{S}}'\boldsymbol{\Omega}\tilde{\boldsymbol{X}}_{\mathcal{S}}$ is noninvertible, in which case the inverse in (5) can be replaced by a generalized inverse.

At this stage, the Bayesian model $\mathcal{M}$ is only needed to supply the pseudo-response variable $\hat{y}_i$; different choices of $\mathcal{M}$ will result in distinct values of $\hat{y}_i$ and therefore distinct actions $\hat{\boldsymbol{\delta}}_{\mathcal{S}}$. An illuminating special case occurs for linear regression:

Corollary 2 *For the linear regression model* $\mathcal{M}$ *with* $\mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{\theta}}\{\tilde{\boldsymbol{y}}(\boldsymbol{x})\} = \boldsymbol{x}'\boldsymbol{\beta}$ *and* any *set of covariate values* $\{\tilde{\boldsymbol{x}}_i\}_{i=1}^{\tilde{n}}$ *and weights* $\omega(\tilde{\boldsymbol{x}}_i) > 0$, *the optimal action* (2) *under* (3) *for the full set of covariates is* $\hat{\boldsymbol{\delta}}_{\{1,\ldots,p\}} = \hat{\boldsymbol{\beta}}$, *where* $\hat{\boldsymbol{\beta}} := \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y}}\boldsymbol{\beta}$.

Depending on the choice of $\{\tilde{\boldsymbol{x}}_i\}_{i=1}^{\tilde{n}}$ and ω, $\hat{\boldsymbol{\delta}}_{\{1,\ldots,p\}}$ may be non-unique. Corollary 2 links the optimal action to the model parameters: the posterior expectation $\hat{\boldsymbol{\beta}}$ is also the optimal action under the parameter-driven squared error loss $\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\delta}) = \|\boldsymbol{\beta} - \boldsymbol{\delta}\|_2^2$. Similarly, a linear model for $\mathcal{M}$ implies that $\hat{y}_i = \tilde{\boldsymbol{x}}_i'\hat{\boldsymbol{\beta}}$, so the optimal action (5) for any subset $\mathcal{S}$ is intrinsically connected to the (regression) model parameters. This persists for other regression models as well. By contrast, these restrictions also illustrate the generality of (3)-(5): the optimal linear actions are derived explicitly under any model $\mathcal{M}$ (with $\mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{y}}\|\tilde{\boldsymbol{y}}(\tilde{\boldsymbol{x}}_i)\|_2^2 < \infty$) and using any set of covariate values $\{\tilde{\boldsymbol{x}}_i\}_{i=1}^{\tilde{n}}$, active covariates $\mathcal{S} \subseteq \{1, \ldots, p\}$, and weighting functions $\omega(\tilde{\boldsymbol{x}}) > 0$.

The critical remaining challenge is optimization—or at least evaluation and comparison—among the possible subsets $\mathcal{S}$. Our strategy emerges from the observation that there may

be many subsets that achieve near-optimal predictive performance, often referred to as the *Rashomon effect* (Breiman, 2001). The goal is to collect, characterize, and compare these near-optimal subsets of linear predictors. Hence, there are two core tasks: (i) identify candidate subsets and (ii) filter to include only those subsets that achieve near-optimal predictive performance. These tasks must overcome both computational and methodological challenges—similar to classical (non-Bayesian) subset selection—which we resolve in the subsequent sections.

An exhaustive enumeration of all possible subsets presents an enormous computational burden, even for moderate p. Although tempting, it is misguided to consider direct optimization over all possible subsets of $\{1, \ldots, p\}$,

$$\hat{\boldsymbol{\delta}}_{\widehat{\mathcal{S}}} := \arg \min_{\mathcal{S}, \boldsymbol{\delta}_{\mathcal{S}}} \mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{y}} \mathcal{L}(\{\tilde{y}_i\}_{i=1}^{\tilde{n}}, \boldsymbol{\delta}_{\mathcal{S}}), \tag{6}$$

for the aggregate squared error loss (3). To see this—and find suitable alternatives—consider the following result:

Lemma 3 *Let $RSS(\boldsymbol{y}, \boldsymbol{\mu}) := \|\boldsymbol{y} - \boldsymbol{\mu}\|_2^2$ and $\hat{\boldsymbol{y}} := \mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{y}} \tilde{\boldsymbol{y}}$, and suppose $\mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{y}} \|\tilde{\boldsymbol{y}}\|_2^2 < \infty$. For any point predictors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, we have*

$$\mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{y}} \{RSS(\tilde{\boldsymbol{y}}, \boldsymbol{\mu}_1)\} \leq \mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{y}} \{RSS(\tilde{\boldsymbol{y}}, \boldsymbol{\mu}_2)\} \iff RSS(\hat{\boldsymbol{y}}, \boldsymbol{\mu}_1) \leq RSS(\hat{\boldsymbol{y}}, \boldsymbol{\mu}_2). \tag{7}$$

Proof Since $\mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{y}} \{\mathrm{RSS}(\tilde{\boldsymbol{y}}, \boldsymbol{\mu})\} - \mathrm{RSS}(\hat{\boldsymbol{y}}, \boldsymbol{\mu}) = \mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{y}} \|\tilde{\boldsymbol{y}}\|_2^2 - \|\hat{\boldsymbol{y}}\|_2^2$ is finite and does not depend on $\boldsymbol{\mu}$, the ordering of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ will be identical whether using $\mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{y}} \{\mathrm{RSS}(\tilde{\boldsymbol{y}}, \boldsymbol{\mu})\}$ or $\mathrm{RSS}(\hat{\boldsymbol{y}}, \boldsymbol{\mu})$. ∎

Notably, $\mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{y}} \{\mathrm{RSS}(\tilde{\boldsymbol{y}}, \boldsymbol{\mu})\}$ is the key constituent in optimizing the predictive squared error loss (3), while $\mathrm{RSS}(\hat{\boldsymbol{y}}, \boldsymbol{\mu})$ is simply the usual residual sum-of-squares (RSS) with $\hat{\boldsymbol{y}}$ replacing $\boldsymbol{y}$.

Now recall the optimization of (6). For any subset $\mathcal{S}$, the optimal action is the least squares solution (5) with pseudo-data $\hat{\boldsymbol{y}}$. However, RSS in linear regression is ordered by nested subsets: $\mathrm{RSS}(\hat{\boldsymbol{y}}, \tilde{\boldsymbol{X}} \hat{\boldsymbol{\delta}}_{\mathcal{S}_1}) \leq \mathrm{RSS}(\hat{\boldsymbol{y}}, \tilde{\boldsymbol{X}} \hat{\boldsymbol{\delta}}_{\mathcal{S}_2})$ whenever $\mathcal{S}_2 \subseteq \mathcal{S}_1$. By Lemma 3, it follows that the solution of (6) is simply

$$\hat{\boldsymbol{\delta}}_{\widehat{\mathcal{S}}} = (\tilde{\boldsymbol{X}}' \boldsymbol{\Omega} \tilde{\boldsymbol{X}})^{-1} \tilde{\boldsymbol{X}}' \boldsymbol{\Omega} \hat{\boldsymbol{y}}, \quad \widehat{\mathcal{S}} = \{1, \ldots, p\}$$

for $\tilde{\boldsymbol{X}} = (\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_{\tilde{n}})'$. As with (5), a generalized inverse can be substituted if necessary. The main consequence is that the optimal actions in (4) and (6) alone cannot select variables or subsets: (4) provides the optimal action for a given subset $\mathcal{S}$, while (6) trivially returns the full set of covariates. Despite the posterior predictive expectation in (6), this optimality is only valid in-sample and is unlikely to persist for out-of-sample prediction. Hence, this optimality is unsatisfying.

Yet Lemma 3 provides a path forward. Rather than fixing a subset $\mathcal{S}$ in (4) or optimizing over all subsets in (6), suppose we compare among all subsets up to size $k \leq p$. Equivalently, this constraint can be representation as an ℓ_0-penalty augmentation to the loss function (3), i.e.,

$$\hat{\boldsymbol{\delta}}_k := \arg \min_{\mathcal{S}, \boldsymbol{\delta}_{\mathcal{S}}} \mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{y}} \mathcal{L}(\{\tilde{y}_i\}_{i=1}^{\tilde{n}}, \boldsymbol{\delta}_{\mathcal{S}}) \quad \text{subject to} \quad \|\boldsymbol{\delta}_{\mathcal{S}}\|_0 \leq k. \tag{8}$$

In direct contrast with previous approaches for Bayesian variable selection via decision analysis (e.g., Hahn and Carvalho, 2015; Woody et al., 2020; Kowal et al., 2021) we do *not* use convex relaxations to ℓ_1-penalties, which create unnecessarily restrictive search paths and introduce additional bias in the coefficient estimates.

Under the loss (3), the solution to (8) reduces to

$$\hat{\boldsymbol{\delta}}_k = \arg\min_{\mathcal{S}, \boldsymbol{\delta}_{\mathcal{S}}} \sum_{i=1}^{\tilde{n}} \omega(\tilde{\boldsymbol{x}}_i) \|\hat{y}_i - \tilde{\boldsymbol{x}}_i' \boldsymbol{\delta}_{\mathcal{S}}\|_2^2 \quad \text{subject to} \quad \|\boldsymbol{\delta}_{\mathcal{S}}\|_0 = k \tag{9}$$

using the "fit-to-the-fit" arguments from Lemma 1 and the RSS ordering results from Lemma 3. In particular, the solution $\hat{\boldsymbol{\delta}}_k$ resembles classical subset selection (1), but uses the fitted values $\hat{y}_i$ from $\mathcal{M}$ instead of the data $\boldsymbol{y}$ and further generalizes to include possibly distinct covariates $\{\tilde{\boldsymbol{x}}_i\}$ and weights $\{\omega(\tilde{\boldsymbol{x}}_i)\}$.

Because of the representation in (9), we can effectively solve (8) by leveraging existing algorithms for subset selection. However, our broader interest in the collection of near-optimal subsets places greater emphasis on the search and filtering process. For any two subsets $\mathcal{S}_1$ and $\mathcal{S}_2$ of equal size $k = |\mathcal{S}_1| = |\mathcal{S}_2|$, Lemma 3 implies that $\mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{y}}\{\text{RSS}(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{X}}\boldsymbol{\delta}_{\mathcal{S}_1})\} \leq \mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{y}}\{\text{RSS}(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{X}}\boldsymbol{\delta}_{\mathcal{S}_2})\}$ if and only if $\text{RSS}(\hat{\boldsymbol{y}}, \tilde{\boldsymbol{X}}\boldsymbol{\delta}_{\mathcal{S}_1}) \leq \text{RSS}(\hat{\boldsymbol{y}}, \tilde{\boldsymbol{X}}\boldsymbol{\delta}_{\mathcal{S}_2})$. Therefore, we can order the linear actions from (4) among all equally-sized subsets simply by ordering the values of $\text{RSS}(\hat{\boldsymbol{y}}, \tilde{\boldsymbol{X}}\boldsymbol{\delta}_{\mathcal{S}})$. This result resembles the analogous scenario in classical linear regression on $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$: subsets of fixed size maintain the same ordering whether using RSS or information criteria such as AIC, BIC, or Mallow's C_p. Here, the criterion of interest is the posterior predictive expected RSS, $\mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{y}}\{\text{RSS}(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{X}}\boldsymbol{\delta}_{\mathcal{S}})\}$, and the RSS reduction occurs with the model $\mathcal{M}$ fitted values $\hat{y}_i = \mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{y}}\tilde{\boldsymbol{y}}(\tilde{\boldsymbol{x}}_i)$ serving as pseudo-data.

Because of this RSS-based ordering among equally-sized linear actions, we can leverage the computational advantages of the *branch-and-bound algorithm* (BBA) for efficient subset exploration (Furnival and Wilson, 2000). Using a tree-based enumeration of all possible subsets, BBA avoids an exhaustive subset search by carefully eliminating non-competitive subsets (or branches) according to RSS. BBA is particularly advantageous for (i) selecting $s_{max} \leq p$ covariates, (ii) filtering to $m_k \leq \binom{p}{k}$ subsets for each size k, and (iii) exploiting the presence of covariates that are almost always present for low-RSS models (Miller, 1984). In the present setting, the key inputs to the algorithm are the covariates $\{\tilde{\boldsymbol{x}}_i\}$, the pseudo-data $\{\hat{y}_i\}$, and the weights $\{\omega(\tilde{\boldsymbol{x}}_i)\}$. In addition, we specify the maximum number of subsets $m_k \leq \binom{p}{k}$ to return for each size k. As m_k increases, BBA returns more subsets—with higher RSS—yet computational efficiency deteriorates. We consider default values of $m_k = 100$ and $m_k = 15$ and compare the results in Section 4. An efficient implementation of BBA is available in the `leaps` package in `R`. Note that our framework is also compatible with many other subset search algorithms (e.g., Bertsimas et al., 2016).

In the case of moderate to large p, we screen to $s_{max} \leq p$ covariates using the original model $\mathcal{M}$. Specifically, we select the s_{max} covariates with the largest absolute (standardized) linear regression coefficients. When $\mathcal{M}$ is a nonlinear model, we use the optimal linear coefficients $\hat{\boldsymbol{\delta}}_{\mathcal{S}}$ on the full subset $\mathcal{S} = \{1, \ldots, p\}$ of (standardized) covariates. The use of marginal screening is common in both frequentist (Fan and Lv, 2008) and Bayesian (Bondell and Reich, 2012) high-dimensional linear regression models, with accompanying consistency results in each case. Here, sceening is motivated by computational scalability

and interpretability: BBA is quite efficient for $p \leq 35$ and $m_k \leq 100$, while the interpretation of a subset of covariates—acting jointly to predict accurately—is muddled as the subset size increases. By default, we fix $s_{max} = 35$. We emphasize that although this screening procedure relies on marginal criteria, it is based on a joint model $\mathcal{M}$ that incorporates all p covariates. In that sense, our *screening* procedure resembles the most popular Bayesian variable *selection* strategies based on posterior inclusion probabilities (Barbieri and Berger, 2004) or hard-thresholding (Datta and Ghosh, 2013).

2.3 Subset search for logistic classifiers

Classification and binary regression operate on $\{0, 1\}$, rendering the squared error loss (3) unsuitable. Consider a binary predictive functional $h(\tilde{y}) \in \{0, 1\}$ where $\tilde{y} \sim p_{\mathcal{M}}(\tilde{y}|\boldsymbol{y})$. In this framework, binarization can come from one of two sources. Most common, the data are binary $y_i \in \{0, 1\}$, paired with the identity functional $h(\tilde{y}) = \tilde{y}$, and $\mathcal{M}$ is a Bayesian classification (e.g., probit or logistic regression) model. Less common, non-binary data can be modeled via $\mathcal{M}$ and paired with a functional h that maps to $\{0, 1\}$. For example, we may be interested in selecting variables to predict exceedance of a threshold, $h(\tilde{y}) = \mathbb{I}\{\tilde{y} \geq \tau\}$, for some τ based on real-valued data $\boldsymbol{y}$. The latter case is an example of *targeted prediction* (Kowal, 2021), which customizes posterior summaries or decisions for any functional h. This approach is distinct from fitting separate models to each empirical functional $\{h(y_i)\}_{i=1}^{n}$— which is still compatible with the first setting above—and instead requires only a single Bayesian reference model $\mathcal{M}$ for all target functionals h.

For classification or binary prediction of $h(\tilde{y}) \in \{0, 1\}$, we replace the squared error loss (3) with the aggregate and weighted cross-entropy loss,

$$\mathcal{L}[\{h(\tilde{y}_i)\}_{i=1}^{\tilde{n}}, \boldsymbol{\delta}_{\mathcal{S}}] = \sum_{i=1}^{\tilde{n}} \omega(\tilde{\boldsymbol{x}}_i)\big[h(\tilde{y}_i)\log\{\pi_{\mathcal{S}}(\tilde{\boldsymbol{x}}_i)\} + \{1 - h(\tilde{y}_i)\}\log\{1 - \pi_{\mathcal{S}}(\tilde{\boldsymbol{x}}_i)\}\big], \quad (10)$$

where $h(\tilde{y}_i) \in \{0, 1\}$ is the predictive variable at $\tilde{\boldsymbol{x}}_i$ under $\mathcal{M}$ and $\pi_{\mathcal{S}}(\tilde{\boldsymbol{x}}_i) := \{1 + \exp(-\tilde{\boldsymbol{x}}_i'\boldsymbol{\delta}_{\mathcal{S}})\}^{-1}$. The cross-entropy is also the *deviance* or negative log-likelihood of a series of independent Bernoulli random variables $h(\tilde{y}_i)$ each with success probability $\pi_{\mathcal{S}}(\tilde{\boldsymbol{x}}_i)$ for $i = 1, \ldots, \tilde{n}$. However, (10) does not imply a distributional assumption for the decision analysis; all distributional assumptions are encapsulated within $\mathcal{M}$, including the posterior predictive distribution of $h(\tilde{y}_i)$. As before, $\boldsymbol{\delta}_{\mathcal{S}}$ is the linear action with zero coefficients for all $j \notin \mathcal{S}$, where $\mathcal{S} \subseteq \{1, \ldots, p\}$ is a subset of active variables.

The optimal action (2) is obtained for each subset $\mathcal{S}$ by computing expectations with respect to the posterior predictive distribution under $\mathcal{M}$ and minimizing the ensuing quantity. As in the case of squared error loss, key simplifications are available:

$$\hat{\boldsymbol{\delta}}_{\mathcal{S}} = \arg\min_{\boldsymbol{\delta}_{\mathcal{S}}} \sum_{i=1}^{\tilde{n}} \omega(\tilde{\boldsymbol{x}}_i)\big[\hat{h}_i\log\{\pi_{\mathcal{S}}(\tilde{\boldsymbol{x}}_i)\} + (1 - \hat{h}_i)\log\{1 - \pi_{\mathcal{S}}(\tilde{\boldsymbol{x}}_i)\}\big] \quad (11)$$

where $\hat{h}_i := \mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{y}}h(\tilde{y}_i)$ is the posterior predictive expectation of $h(\tilde{y}_i)$ under $\mathcal{M}$. The representation in (11) is quite useful: it is the negative log-likelihood for a logistic regression model with pseudo-data $\{(\tilde{\boldsymbol{x}}_i, \hat{h}_i)\}_{i=1}^{\tilde{n}}$. Standard algorithms and software, such as iteratively

reweighted least squares (IRLS) in the `glm` package in `R`, can be applied to solve (11) for any subset $\mathcal{S}$.

Instead of fitting a logistic regression to the observed binary variables $h(y_i) \in \{0, 1\}$, the optimal action under cross-entropy (10) fits to the posterior predictive probabilities $\hat{h}_i = p_{\mathcal{M}}\{h(\tilde{y}_i) = 1 | \boldsymbol{y}\} \in [0, 1]$ under $\mathcal{M}$. For a well-specified model $\mathcal{M}$, these posterior probabilities $\hat{h}_i$ can be more informative than the binary empirical functionals $h(y_i)$: the former lie on a continuum between the endpoints zero and one. Furthermore, for non-degenerate models $\mathcal{M}$ with $\hat{h}_i \in (0, 1)$, the optimal action (11) resolves the issue of separability, which is a persistent challenge in classical logistic regression.

Despite the efficiency of IRLS for a fixed subset $\mathcal{S}$, the computational savings of BBA for subset search rely on the RSS from a linear model. As such, solving (11) for all or many subsets incurs a much greater computational cost. Yet IRLS is intrinsically linked with RSS. At convergence, IRLS obtains a weighted least squares solution

$$\hat{\boldsymbol{\delta}}_{\mathcal{S}} = (\tilde{\boldsymbol{X}}_{\mathcal{S}}' \boldsymbol{W} \tilde{\boldsymbol{X}}_{\mathcal{S}})^{-1} \tilde{\boldsymbol{X}}_{\mathcal{S}}' \boldsymbol{W} \boldsymbol{z} \tag{12}$$

where $\boldsymbol{W} := \mathrm{diag}\{w_i\}_{i=1}^{\tilde{n}}$ for weights $w_i := \omega(\tilde{\boldsymbol{x}}_i) \hat{\pi}_{\mathcal{S}}(\tilde{\boldsymbol{x}}_i)\{1 - \hat{\pi}_{\mathcal{S}}(\tilde{\boldsymbol{x}}_i)\}$, fitted probabilities $\hat{\pi}_{\mathcal{S}}(\tilde{\boldsymbol{x}}_i) := \{1 + \exp(-\tilde{\boldsymbol{x}}_i' \hat{\boldsymbol{\delta}}_{\mathcal{S}})\}^{-1}$, and pseudo-data

$$z_i := \log \frac{\hat{\pi}_{\mathcal{S}}(\tilde{\boldsymbol{x}}_i)}{1 - \hat{\pi}_{\mathcal{S}}(\tilde{\boldsymbol{x}}_i)} + \frac{\hat{h}_i - \hat{\pi}_{\mathcal{S}}(\tilde{\boldsymbol{x}}_i)}{\hat{\pi}_{\mathcal{S}}(\tilde{\boldsymbol{x}}_i)\{1 - \hat{\pi}_{\mathcal{S}}(\tilde{\boldsymbol{x}}_i)\}} \tag{13}$$

with $\boldsymbol{z} := (z_1, \ldots, z_{\tilde{n}})'$. By design, the weighted least squares objective associated with (12) is a second-order Taylor approximation to the predictive expected cross-entropy loss:

$$\mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{y}} \mathcal{L}[\{h(\tilde{y}_i)\}_{i=1}^{\tilde{n}}, \hat{\boldsymbol{\delta}}_{\mathcal{S}}] \approx \sum_{i=1}^{\tilde{n}} \omega(\tilde{\boldsymbol{x}}_i) \frac{\{\hat{h}_i - \hat{\pi}_{\mathcal{S}}(\tilde{\boldsymbol{x}}_i)\}^2}{\hat{\pi}_{\mathcal{S}}(\tilde{\boldsymbol{x}}_i)\{1 - \hat{\pi}_{\mathcal{S}}(\tilde{\boldsymbol{x}}_i)\}} = \sum_{i=1}^{\tilde{n}} w_i(z_i - \tilde{\boldsymbol{x}}_i' \hat{\boldsymbol{\delta}}_{\mathcal{S}})^2. \tag{14}$$

The weighted least squares approximation in (14) summons BBA for subset search. Hosmer et al. (1989) adopted this strategy for subset selection in classical logistic regression on $y_i \in \{0, 1\}$. Here, both the goals and the optimization criterion are distinct: we are interested in curating a collection of near-optimal subsets—rather than selecting a single "best" subset—and the weighted least squares objective (14) inherits the fitted probabilities $\hat{h}_i$ from the Bayesian model $\mathcal{M}$ along with the weights $\omega(\tilde{\boldsymbol{x}}_i)$.

Ideally, we might apply BBA directly based on the covariates $\{\tilde{\boldsymbol{x}}_i\}$, the pseudo-data $\{z_i\}$, and the weights $\{w_i\}$. However, both z_i and w_i depend on $\hat{\pi}_{\mathcal{S}}(\tilde{\boldsymbol{x}}_i)$ and therefore are subset-specific. As a result, BBA cannot be applied without significant modifications. A suitable alternative is to construct subset-invariant psuedo-data and weights by replacing $\hat{\pi}_{\mathcal{S}}(\tilde{\boldsymbol{x}}_i)$ with the corresponding estimate from the full model, $\hat{h}_i$. Specifically, let

$$\hat{z}_i := \log\{\hat{h}_i/(1 - \hat{h}_i)\}, \quad \hat{w}_i := \omega(\tilde{\boldsymbol{x}}_i) \hat{h}_i(1 - \hat{h}_i), \tag{15}$$

both of which depend on $\mathcal{M}$ rather than the individual subsets $\mathcal{S}$. The pseudo-data $\hat{z}_i$ is defined similarly to z_i in (13), where the second term now cancels. Finally, BBA subset search can be applied using the covariates $\{\tilde{\boldsymbol{x}}_i\}$, the pseudo-data $\{\hat{z}_i\}$, and the weights $\{\hat{w}_i\}$. As for squared error, we restrict each subset size to $m_k = 100$ or $m_k = 15$ for all subsets of size $k = 1, \ldots, p$. Despite the critical role of the weighted least squares approximation in (14) for subset search, all further evaluations and comparisons rely on the exact cross-entropy loss (10).

2.4 Predictive evaluations for identifying near-optimal subsets

The BBA subset search filters the 2^p possible subsets to the m_k best subsets for each size $k = 1, \ldots, s_{max}$ according to posterior predictive expected loss using the weighted squared error loss for prediction (Section 2.2) or the cross-entropy loss for classification (Section 2.3). However, as noted below Lemma 3, further comparisons based on these expected losses are trivial: the full set $\mathcal{S} = \{1, \ldots, p\}$ always obtains the minimum, which precludes variable selection. Selection of a single "best" subset of each size k invites additional difficulties: if multiple subsets perform similarly—which is common for correlated covariates—then selecting $m_k = 1$ subset will not be robust or stable against perturbations of the data. Equally important, restricting to $m_k = 1$ subset is blind to competing subsets that offer similar predictive performance yet may differ substantially in the composition of covariates; see Sections 3-4. These challenges persist for both classical and Bayesian approaches.

We instead curate and explore a collection of near-optimal subsets. The notion of "near-optimal" derives from the *acceptable family* of Kowal (2021). Informally, the acceptable family uses out-of-sample predictive metrics to gather those predictors that match or nearly match the performance of the best out-of-sample predictor with nonnegligible posterior predictive probability under $\mathcal{M}$. The out-of-sample evaluation uses a modified K-fold cross-validation procedure. Let $\mathcal{I}_k \subset \{1, \ldots, n\}$ denote the kth validation set, where each data point appears in (at least) one validation set, $\cup_{k=1}^{K} \mathcal{I}_k = \{1, \ldots, n\}$. By default, we use $K = 10$ validation sets that are equally-sized, mutually exclusive, and selected randomly from $\{1, \ldots, n\}$. Define an evaluative loss function $L(y, \tilde{\boldsymbol{x}}' \hat{\boldsymbol{\delta}}_{\mathcal{S}})$ for the optimal linear coefficients of subset $\mathcal{S}$, and let $\mathbb{S}$ denote the collection of subsets obtained from the BBA filtering process. Typically, the evaluative loss L is identical to the loss $\mathcal{L}$ used for optimization, but this restriction is not required. For each data split k and each subset $\mathcal{S} \in \mathbb{S}$, the out-of-sample *empirical* and *predictive* losses are

$$L_{\mathcal{S}}^{out}(k) := \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} L(y_i, \boldsymbol{x}_i' \hat{\boldsymbol{\delta}}_{\mathcal{S}}^{-\mathcal{I}_k}), \quad \widetilde{L}_{\mathcal{S}}^{out}(k) := \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} L(\tilde{y}_i^{-\mathcal{I}_k}, \boldsymbol{x}_i' \hat{\boldsymbol{\delta}}_{\mathcal{S}}^{-\mathcal{I}_k}) \tag{16}$$

respectively, where $\hat{\boldsymbol{\delta}}_{\mathcal{S}}^{-\mathcal{I}_k} := \arg\min_{\boldsymbol{\delta}_{\mathcal{S}}} \mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{y}^{-\mathcal{I}_k}} \mathcal{L}(\{\tilde{y}_i\}_{i \notin \mathcal{I}_k}, \boldsymbol{\delta}_{\mathcal{S}})$ is estimated using only the training data $\boldsymbol{y}^{-\mathcal{I}_k} := \{y_i\}_{i \notin \mathcal{I}_k}$ and $\tilde{y}_i^{-\mathcal{I}_k} \sim p_{\mathcal{M}}(\tilde{y}_i | \boldsymbol{y}^{-\mathcal{I}_k})$ is the posterior predictive distribution at $\boldsymbol{x}_i$ conditional only on the training data. Averaging across all data splits, we obtain $L_{\mathcal{S}}^{out} := K^{-1} \sum_{k=1}^{K} L_{\mathcal{S}}^{out}(k)$ and $\widetilde{L}_{\mathcal{S}}^{out} := K^{-1} \sum_{k=1}^{K} \widetilde{L}_{\mathcal{S}}^{out}(k)$.

The distinction between the empirical loss $L_{\mathcal{S}}^{out}$ and the predictive loss $\widetilde{L}_{\mathcal{S}}^{out}$ is important. The empirical loss is a point estimate of the risk under predictions from $\hat{\boldsymbol{\delta}}_{\mathcal{S}}$ based on the data $\boldsymbol{y}$. By comparison, the predictive loss provides a distribution of the out-of-sample loss based on the model $\mathcal{M}$. Both are valuable: $L_{\mathcal{S}}^{out}$ is entirely empirical and captures the classical notion of K-fold cross-validation, while $\widetilde{L}_{\mathcal{S}}^{out}$ leverages the Bayesian model to propagate the uncertainty from the model-based data-generating process.

For any two subsets $\mathcal{S}_1$ and $\mathcal{S}_2$, consider the percent increase in out-of-sample predictive loss from $\hat{\boldsymbol{\delta}}_{\mathcal{S}_1}$ to $\hat{\boldsymbol{\delta}}_{\mathcal{S}_2}$:

$$\widetilde{D}_{\mathcal{S}_1, \mathcal{S}_2}^{out} := 100 \times (\widetilde{L}_{\mathcal{S}_2}^{out} - \widetilde{L}_{\mathcal{S}_1}^{out}) / \widetilde{L}_{\mathcal{S}_1}^{out}. \tag{17}$$

Since $\widetilde{D}_{\mathcal{S}_1, \mathcal{S}_2}^{out}$ inherits a predictive distribution under $\mathcal{M}$, we can leverage the accompanying uncertainty quantification to determine whether the predictive performances of $\hat{\boldsymbol{\delta}}_{\mathcal{S}_1}$ and $\hat{\boldsymbol{\delta}}_{\mathcal{S}_2}$

are sufficiently distinguishable. In particular, we are interested in comparisons to the best subset for out-of-sample prediction,

$$\mathcal{S}_{min} \coloneqq \arg\min_{\mathcal{S}\in\mathbb{S}} L_{\mathcal{S}}^{out}, \tag{18}$$

so that $\hat{\boldsymbol{\delta}}_{\mathcal{S}_{min}}$ is the optimal linear action associated with the subset $\mathcal{S}_{min}$ that minimizes the empirical K-fold cross-validated loss. Unlike the RSS-based in-sample evaluations from Section 2.2, the subset $\mathcal{S}_{min}$ can—and usually will—differ from the full set $\{1,\ldots,p\}$, which enables variable selection driven by out-of-sample predictive performance.

Using $\mathcal{S}_{min}$ as an anchor, the acceptable family broadens to include near-optimal subsets:

$$\mathbb{A}_{\eta,\varepsilon} \coloneqq \big\{\mathcal{S}\in\mathbb{S} : \mathbb{P}_{\mathcal{M}}\big(\widetilde{D}_{\mathcal{S}_{min},\mathcal{S}}^{out} < \eta\big) \geq \varepsilon\big\}, \quad \eta \geq 0, \varepsilon \in [0,1]. \tag{19}$$

With (19), we collect all subsets $\mathcal{S}$ that perform within a margin $\eta \geq 0\%$ of the best subset, $\widetilde{D}_{\mathcal{S}_{min},\mathcal{S}}^{out} < \eta$, with probability at least $\varepsilon \in [0,1]$. Equivalently, a subset $\mathcal{S}$ is acceptable if and only if there exists a lower $(1-\varepsilon)$ posterior prediction interval for $\widetilde{D}_{\mathcal{S}_{min},\mathcal{S}}^{out}$ that includes η. Hence, *unacceptable* subsets are those for which there is insufficient predictive probability (under $\mathcal{M}$) that the out-of-sample accuracy of $\mathcal{S}$ is within a predetermined margin of the best subset. The acceptable family is nonempty, since $\mathcal{S}_{min} \in \mathbb{A}_{\eta,\varepsilon}$, and is expanded by increasing η or decreasing ε. By default, we select $\eta = 0\%$ and $\varepsilon = 0.10$ and assess robustness in the simulation study (see also Kowal, 2021; Kowal et al., 2021).

Within the acceptable family, we isolate two subsets of particular interest: the best subset $\mathcal{S}_{min}$ from (18) and the smallest acceptable subset,

$$\mathcal{S}_{small} \coloneqq \arg\min_{\mathcal{S}\in\mathbb{A}_{\eta,\varepsilon}} |\mathcal{S}|. \tag{20}$$

When $\mathcal{S}_{small}$ is nonunique, so that multiple subsets of the same minimal size are acceptable, we select from those subsets the one with the smallest empirical loss $L_{\mathcal{S}}^{out}$. By definition, $\mathcal{S}_{small}$ is the smallest set of covariates that satisfies the near-optimality condition in (19). As noted by Hastie et al. (2009) and others, selection based on minimum cross-validated error, such as $\mathcal{S}_{min}$, often produces models or subsets that are more complex than needed for adequate predictive accuracy. The acceptable family $\mathbb{A}_{\eta,\varepsilon}$, and in particular $\mathcal{S}_{small}$, exploits this observation to provide alternative—and often much smaller—subsets of variables.

To ease the computational burden, we adapt the importance sampling algorithm from Kowal (2021) to compute (19) and all constituent quantities (see Appendix A). Crucially, this algorithm is based entirely on the in-sample posterior distribution under model $\mathcal{M}$, which avoids both (i) the intensive process of re-fitting $\mathcal{M}$ for each data split k and (ii) data reuse that adversely affects downstream uncertainty quantification. Briefly, the algorithm uses the complete data posterior $p_{\mathcal{M}}(\boldsymbol{\theta}|\boldsymbol{y})$ as a proposal for the training data posterior $p_{\mathcal{M}}(\boldsymbol{\theta}|\boldsymbol{y}^{-\mathcal{I}_k})$. The importance weights are then computed using the likelihood $p_{\mathcal{M}}(\boldsymbol{y}^{\mathcal{I}_k}|\boldsymbol{\theta})$ under model $\mathcal{M}$. Similar algorithms have been deployed for Bayesian model selection (Gelfand et al., 1992), evaluating prediction distributions (Vehtari and Ojanen, 2012), and ℓ_1-based Bayesian variable selection (Kowal et al., 2021).

In the case of new (out-of-sample) covariate values $\{\tilde{\boldsymbol{x}}_i\}_{i=1}^{\tilde{n}} \neq \{\boldsymbol{x}_i\}_{i=1}^{n}$, the predictive loss may be defined without cross-validation: $\widetilde{L}_{\mathcal{S}}^{out} \coloneqq \tilde{n}^{-1} \sum_{i=1}^{\tilde{n}} L(\tilde{y}_i, \tilde{\boldsymbol{x}}_i'\hat{\boldsymbol{\delta}}_{\mathcal{S}})$, where $\tilde{y}_i \coloneqq \tilde{\boldsymbol{y}}(\tilde{\boldsymbol{x}}_i)$ is the predictive variable at $\tilde{\boldsymbol{x}}_i$ for each $i = 1,\ldots,\tilde{n}$. The empirical loss is undefined

without a corresponding observation of the response variable for each $\tilde{\boldsymbol{x}}_i$. Hence, we modify the acceptable family (19) to instead reference the full subset $\mathcal{S} = \{1, \ldots, p\}$ in place of $\mathcal{S}_{min}$, which is no longer defined. When $\mathcal{M}$ is a linear model, Corollary 2 implies that the corresponding reference is simply the posterior predictive expectation under $\mathcal{M}$.

2.5 Co-variable importance

Although it is common to focus on a single subset for selection, the acceptable family $\mathbb{A}_{\eta,\varepsilon}$ provides a broad assortment of competing explanations: linear actions with distinct sets of covariates that all provide near-optimal predictive accuracy. Hence, we seek to further summarize $\mathbb{A}_{\eta,\varepsilon}$ beyond $\mathcal{S}_{min}$ and $\mathcal{S}_{small}$ to identify (i) which covariates appear in any acceptable subset, (ii) which covariates appear in all acceptable subsets, and (iii) which covariates appear together in the same acceptable subsets.

For covariates j and ℓ, the sample proportion of joint inclusion in $\mathbb{A}_{\eta,\varepsilon}$ achieves each of these goals:

$$\mathrm{VI}_{\mathrm{incl}}(j, \ell) := |\mathbb{A}_{\eta,\varepsilon}|^{-1} \sum_{\mathcal{S} \in \mathbb{A}_{\eta,\varepsilon}} \mathbb{I}\{j, \ell \in \mathcal{S}\} \tag{21}$$

and measures *(co-) variable importance*. Naturally, (21) is generalizable to more than two covariates, but is particularly interesting for a single covariate: $\mathrm{VI}_{\mathrm{incl}}(j) := \mathrm{VI}_{\mathrm{incl}}(j, j)$ is the proportion of acceptable subsets to which covariate j belongs. When $\mathrm{VI}_{\mathrm{incl}}(j) > 0$, covariate j belongs to at least one acceptable subset. Such a result does not imply that covariate j is necessary for accurate prediction, but rather that covariate j is a component of at least one near-optimal linear subset. When $\mathrm{VI}_{\mathrm{incl}}(j) = 1$, we refer to covariate j as a *keystone covariate*: it belongs to all acceptable subsets and therefore is deemed essential. For $j \neq \ell$, $\mathrm{VI}_{\mathrm{incl}}(j, \ell)$ highlights not only the covariates that co-occur, but also the covariates that rarely appear together. It is particularly informative to identify covariates j and ℓ such that $\mathrm{VI}_{\mathrm{incl}}(j)$ and $\mathrm{VI}_{\mathrm{incl}}(\ell)$ are both large yet $\mathrm{VI}_{\mathrm{incl}}(j, \ell)$ is small. In that case, covariates j and ℓ are both important yet also redundant, as might be expected for highly correlated variables.

Although $\mathrm{VI}_{\mathrm{incl}}$ is defined based on linear predictors for each subset $\mathcal{S}$, the variable importance metric applies broadly to (possibly nonlinear) Bayesian models $\mathcal{M}$ and a variety of evaluative loss functions L, and can be targeted locally via the weights $\omega(\tilde{\boldsymbol{x}}_i)$. The inclusion-based metric $\mathrm{VI}_{\mathrm{incl}}$ can be extended to incorporate effect size, which is a more common strategy for variable importance. Related, Dong and Rudin (2019) aggregated model-specific variable importances across many "good models". Alternative approaches use leave-one-covariate-out predictive metrics (e.g., Lei et al., 2018), but are less appealing in the presence of correlated covariates.

2.6 Posterior predictive uncertainty quantification

A persistent challenge in classical subset selection is the lack of accompanying uncertainty quantification. Given a subset $\hat{\mathcal{S}}$ selected using the data $(\boldsymbol{X}, \boldsymbol{y})$, familiar frequentist and Bayesian inferential procedures applied to $(\boldsymbol{X}_{\hat{\mathcal{S}}}, \boldsymbol{y})$ are in general no longer valid. In particular, we cannot simply proceed as if only the selected covariates $\hat{\mathcal{S}}$ were supplied from the onset. Such analyses are subject to selection bias (Miller, 1984).

A crucial feature of our subset filtering (Section 2.2 and Section 2.3) and predictive evaluation (Section 2.4) techniques is that, despite the broad searching and the out-of-sample targets, these quantities all remain *in-sample posterior functionals* under $\mathcal{M}$. There is no data re-use or model re-fitting: every requisite term is a functional of the complete data posterior $p_{\mathcal{M}}(\boldsymbol{\theta}|\boldsymbol{y})$ or $p_{\mathcal{M}}(\tilde{\boldsymbol{y}}|\boldsymbol{y})$ from a single Bayesian model. Hence, the posterior distribution under $\mathcal{M}$ remains a valid facilitator of uncertainty quantification.

We elicit a posterior predictive distribution for the action by removing the expectation in (2):

$$\tilde{\boldsymbol{\delta}} := \arg\min_{\boldsymbol{\delta}} \mathcal{L}(\tilde{\boldsymbol{y}}, \boldsymbol{\delta}) \tag{22}$$

which no longer integrates over $p_{\mathcal{M}}(\tilde{\boldsymbol{y}}|\boldsymbol{y})$ and hence propagates the posterior predictive uncertainty. For the squared error loss (3), the *predictive action* is

$$\tilde{\boldsymbol{\delta}}_{\mathcal{S}} = (\tilde{\boldsymbol{X}}_{\mathcal{S}}'\boldsymbol{\Omega}\tilde{\boldsymbol{X}}_{\mathcal{S}})^{-1}\tilde{\boldsymbol{X}}_{\mathcal{S}}'\boldsymbol{\Omega}\tilde{\boldsymbol{y}} \tag{23}$$

akin to (5), where $\tilde{\boldsymbol{y}} = (\tilde{y}_1, \ldots, \tilde{y}_{\tilde{n}})' \sim p_{\mathcal{M}}(\tilde{\boldsymbol{y}}|\boldsymbol{y})$. The linear coefficients $\tilde{\boldsymbol{\delta}}_{\mathcal{S}}$ inherit a posterior predictive distribution from $\tilde{\boldsymbol{y}}$ and can be computed for any subset $\mathcal{S}$. Similar computations are available for the cross-entropy loss (10). In both cases, draws $\{\tilde{\boldsymbol{y}}^s\}_{s=1}^S \sim p_{\mathcal{M}}(\tilde{\boldsymbol{y}}|\boldsymbol{y})$ from the posterior predictive distribution are sufficient for uncertainty quantification of $\tilde{\boldsymbol{\delta}}_{\mathcal{S}}$. Under the usual assumption that $p_{\mathcal{M}}(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}, \boldsymbol{y}) = p_{\mathcal{M}}(\tilde{\boldsymbol{y}}|\boldsymbol{\theta})$, these draws are easily obtained by repeatedly sampling $\boldsymbol{\theta}^s \sim p_{\mathcal{M}}(\boldsymbol{\theta}|\boldsymbol{y})$ from the posterior and $\tilde{\boldsymbol{y}}^s \sim p_{\mathcal{M}}(\tilde{\boldsymbol{y}}|\boldsymbol{\theta} = \boldsymbol{\theta}^s)$ from the likelihood.

The predictive actions are computable for any subset $\mathcal{S}$, include those selected based on the predictive evaluations in Section 2.4. Let $\hat{\mathcal{S}}$ denote a subset identified based on the posterior (predictive) distribution under $\mathcal{M}$, such as $\mathcal{S}_{min}$ or $\mathcal{S}_{small}$. The predictive action $\tilde{\boldsymbol{\delta}}_{\hat{\mathcal{S}}}$ using this subset in (23) is a posterior predictive functional. Unlike for a generic subset $\mathcal{S}$, the predictive action $\tilde{\boldsymbol{\delta}}_{\hat{\mathcal{S}}}$ is a functional of both $\tilde{\boldsymbol{y}}$ and $\hat{\mathcal{S}}$, which factors into the interpretation.

A similar projection-based approach is developed by Woody et al. (2020) for Gaussian regression. In place of $\mathcal{L}(\tilde{\boldsymbol{y}}, \boldsymbol{\delta})$ in (22), Woody et al. (2020) suggest $\mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{\theta}}\mathcal{L}(\tilde{\boldsymbol{y}}, \boldsymbol{\delta})$, which is a middle ground between (2) and (22): it integrates over the uncertainty of $\tilde{\boldsymbol{y}}$ given model parameters $\boldsymbol{\theta}$, but preserves uncertainty due to $\boldsymbol{\theta}$. For example, under the linear regression model $\mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{\theta}}\{\tilde{\boldsymbol{y}}(\tilde{\boldsymbol{x}})\} = \tilde{\boldsymbol{x}}'\boldsymbol{\beta}$, the analogous result in (23) is $(\tilde{\boldsymbol{X}}_{\mathcal{S}}'\boldsymbol{\Omega}\tilde{\boldsymbol{X}}_{\mathcal{S}})^{-1}\tilde{\boldsymbol{X}}_{\mathcal{S}}'\boldsymbol{\Omega}\tilde{\boldsymbol{X}}\boldsymbol{\beta}$; when $\mathcal{S} = \{1, \ldots, p\}$, this simplifies to the regression coefficient $\boldsymbol{\beta}$. Both approaches have their merits, but we prefer (22) because of the connection to the observable random variables $\tilde{\boldsymbol{y}}$ rather than the model-specific parameters $\boldsymbol{\theta}$.

3. Simulation study

We conduct an extensive simulation study to evaluate competing subset selection techniques for prediction accuracy, uncertainty quantification, and variable selection. Prediction for real-valued data is discussed here; classification for binary data is presented in Appendix B. Although we do evaluate individual subsets from the proposed framework—namely, $\mathcal{S}_{min}$ and $\mathcal{S}_{small}$—we are more broadly interested in the performance of the *acceptable family* of subsets. In particular, the acceptable family is designed to collect *many subsets* that offer *near-optimal* prediction; both cardinality and aggregate predictive accuracy are critical.

The simulation designs feature varying signal-to-noise ratios (SNRs) and dimensions (n, p), including $p \gg n$, with correlated covariates and sparse linear signals. Covariates $x_{i,j}$ are generated from marginal standard normal distributions with $\mathrm{Cor}(x_{i,j}, x_{i,j'}) = (0.75)^{|j-j'|}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. The p columns are randomly permuted and augmented with an intercept. The true linear coefficients $\boldsymbol{\beta}^*$ are constructed by setting $\beta_0^* = -1$ and fixing $p_* = 5$ nonzero coefficients, with $\lceil p_*/2 \rceil$ equal to 1 and $\lfloor p_*/2 \rfloor$ equal to -1, and the rest at zero. The data are generated independently as $y_i \sim N(y_i^*, \sigma_*^2)$ with $y_i^* := \boldsymbol{x}_i' \boldsymbol{\beta}^*$ and $\sigma_* := \mathrm{sd}(\boldsymbol{y}^*)/\sqrt{\mathrm{SNR}}$. We consider $(n, p) \in \{(50, 50), (200, 400), (500, 50)\}$ and $\mathrm{SNR} \in \{0.25, 1\}$ for low and high SNR. Evaluations are conducted over 100 simulated datasets.

First, we evaluate point prediction accuracy across competing collections of near-optimal subsets. Each collection is built from the same Gaussian linear regression model with horseshoe priors (Carvalho et al., 2010) $\mathcal{M}$ and estimated using `bayeslm` (Hahn et al., 2019). The collections are generated from different candidate sets $\mathbb{S}$ based on distinct search methods: the proposed BBA method (`bbound(bayes)`), the adaptive lasso search (`lasso(bayes)`) proposed by Hahn and Carvalho (2015) and Kowal et al. (2021), and both forward (`forward(bayes)`) and backward (`backward(bayes)`) search. For each collection of candidate subsets, we compute the acceptable families for $\eta = 0\%$, $\varepsilon = 0.10$, $m_k = 15$, and the observed covariate values $\{\tilde{\boldsymbol{x}}_i\}_{i=1}^{\tilde{n}} = \{\boldsymbol{x}_i\}_{i=1}^n$; variations of these specifications are presented in Appendix C. These methods differ *only* in the search process: if a particular subset $\mathcal{S}$ is identified by more than one of these competing methods, the corresponding point predictions—based on the optimal action $\hat{\boldsymbol{\delta}}_{\mathcal{S}}$ in (5)—will be identical.

The competing collections of near-optimal subsets are evaluated in aggregate. At each simulation, we compute the qth quantile of the root mean squared errors (RMSEs) for the true mean $\{y_i^*\}_{i=1}^n$ across all subsets within that acceptable family, and then average that quantity across all simulations. For example, $q = 1$ describes the predictive performance if we were to use the worst acceptable subset at each simulation, as determined by an oracle. These quantities summarize the distribution of RMSEs *within* each acceptable family; we report the values for $q \in \{0, 0.25, 0.5, 0.75, 1\}$ and present the results as boxplots in Figure 1. Most notably, the proposed `bbound(bayes)` strategy produces up to 10 times the number of acceptable subsets as the other methods, yet crucially maintains equivalent predictive performance. Naturally, this expanded collection of subsets also produces a minimum RMSE $(q = 0)$ that outperforms the remaining methods. Clearly, the proposed approach is far superior at collecting *more* subsets that provide *near-optimal* predictive accuracy.

Figure 1 also summarizes the point prediction accuracy for several competing methods. First, we include the usual point estimate under $\mathcal{M}$ given by the posterior expectation of the regression coefficients $\boldsymbol{\beta}$ (`post.mean`), which does not include any variable or subset selection. Next, we compute point predictions for the key acceptable subsets $\mathcal{S}_{min}$ and $\mathcal{S}_{small}$ using the proposed `bbound(bayes)` approach; the competing search strategies produced similar results and are omitted. Among frequentist estimators, we use the adaptive lasso (`lasso`; Zou, 2006) with λ chosen by 10-fold cross-validation and the one-standard-error rule (Hastie et al., 2009). In addition, we include classical subset selection (`subset`) implemented using the `leaps` package in R with the final subset determined by AIC. When $p > 35$, we screen to the first 35 predictors that enter the model in the aforementioned adaptive lasso. Forward (`forward`) and backward (`backward`) selection were also considered, but were either noncompetitive or performed similarly to the adaptive lasso and are omitted from this plot.
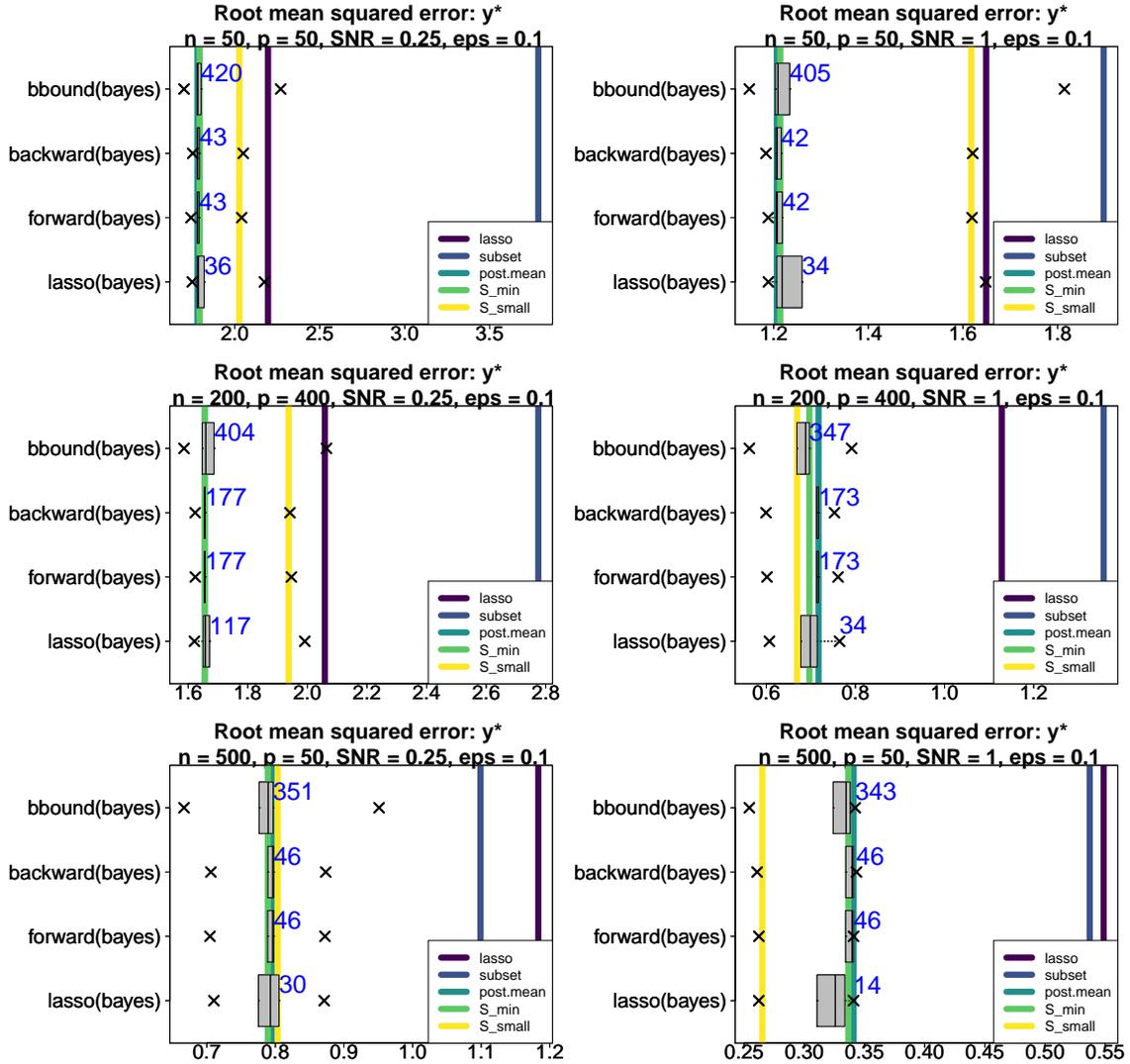
Figure 1: Root mean squared errors (RMSEs) for predicting y^* across (n, p, SNR) configurations. The boxplots summarize the RMSE quantiles for the subsets within each acceptable family, while the vertical lines denote RMSEs of competing methods. The average size of each acceptable family is annotated. The proposed BBA search returns vastly more subsets that remain highly competitive, while $\mathcal{S}_{small}$ performs very well and substantially outperforms classical subset selection.

Most notably, the predictive performance of $\mathcal{S}_{small}$ is excellent, especially for high SNRs, and substantially outperforms the frequentist methods in all settings. Because $\mathcal{S}_{min}$ is overly conservative—i.e., it selects too many variables (see Table 1)—it performs nearly identically to post.mean. Although $\mathcal{S}_{min}$ and post.mean outperform $\mathcal{S}_{small}$ when the SNR is low and the sample size is small, note that $\mathcal{S}_{small}$ is not targeted exclusively for optimal predictive performance; rather, it represents the smallest subset that obtains *near-optimal*

performance. Lastly, the regularization induced by $\mathcal{M}$ is greatly beneficial: *every* member of each acceptable family decisively outperforms classical subset selection across all scenarios.

Next, we compare uncertainty quantification for the regression coefficients $\boldsymbol{\beta}^*$. We compute 90% intervals using the predictive action $\tilde{\boldsymbol{\delta}}_{\mathcal{S}}$ from (23) for $\mathcal{S}_{small}$ and $\mathcal{S}_{min}$ based on the proposed `bbound(bayes)` procedure. In addition, we compute 90% highest posterior density (HPD) credible intervals for $\boldsymbol{\beta}$ under $\mathcal{M}$ (`post.HPD`) and 90% frequentist confidence intervals using Zhao et al. (2021), which computes confidence intervals from a linear regression model that includes only the variables selected by the (adaptive) lasso (`lasso+lm`). The 90% interval estimates are evaluated and compared in Figure 2 using interval widths and empirical coverage. The intervals from $\mathcal{S}_{small}$ using (22) are uniformly better (i.e., narrower) than the usual posterior HPD intervals under $\mathcal{M}$. In addition, the intervals from Zhao et al. (2021) are highly competitive, despite ignoring selection bias. In all cases, the intervals are overly conservative and achieve the nominal empirical coverage.
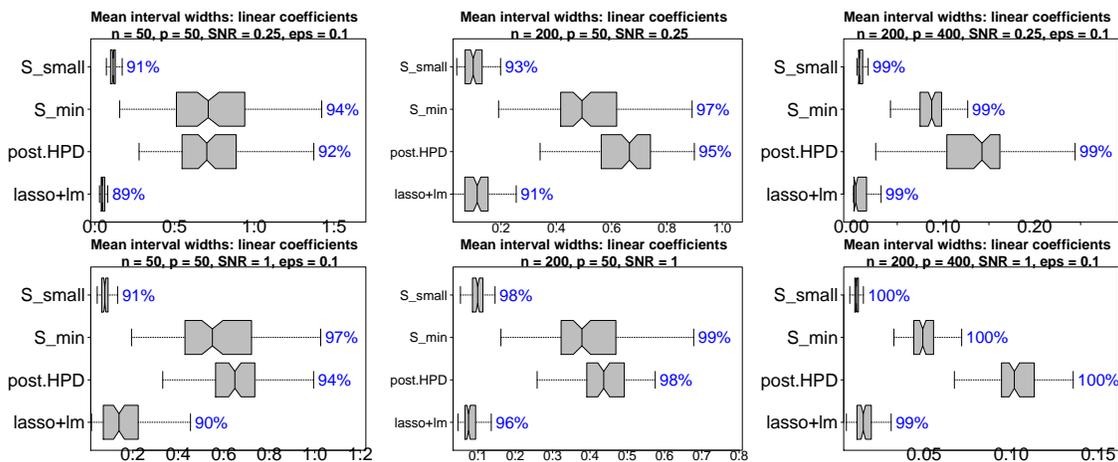

Figure 2: Mean interval widths (boxplots) with empirical coverage (annotations) for $\boldsymbol{\beta}^*$. Non-overlapping notches indicate significant differences between medians. The proposed intervals based on $\mathcal{S}_{small}$ are significantly narrower than the usual HPD intervals under $\mathcal{M}$ yet maintain the empirical nominal 90% coverage.

Lastly, marginal variable selection is evaluated using true positive rates and true negative rates in Table 1. In addition to $\mathcal{S}_{min}$ and $\mathcal{S}_{small}$, we include the common Bayesian selection technique that includes a variable j when the 95% HPD interval for β_j excludes zero (`posterior HPD`). The selection performance of S_{small} is excellent and similar to the the adaptive lasso. Classical subset, forward, and backward selection are too aggressive, while the posterior interval-based selection is too conservative. Hence, despite the primary focus on the *collection* of near-optimal subsets, the smallest acceptable subset $\mathcal{S}_{small}$ is a key member with excellent prediction, uncertainty quantification, and marginal selection performance.

The appendix contains simulation studies for classification (Appendix B) and variations for $\varepsilon \in \{0.01, 0.10, 0.20\}$, $\{\tilde{\boldsymbol{x}}_i\}_{i=1}^{\tilde{n}}$, and the distributions of $\{\boldsymbol{x}_i\}_{i=1}^{n}$ and $\{\tilde{\boldsymbol{x}}_i\}_{i=1}^{\tilde{n}}$ (Appendix C). All results are qualitatively similar to those presented here.

$$n = 50, p = 50, \text{SNR} = 0.25$$

	lasso	forward	backward	subset	posterior HPD	S_{min}	S_{small}
TPR	0.22	0.96	0.93	0.53	0.06	0.51	0.22
TNR	0.98	0.06	0.10	0.63	1.00	0.81	0.98

$$n = 50, p = 50, \text{SNR} = 1$$

	lasso	forward	backward	subset	posterior HPD	S_{min}	S_{small}
TPR	0.54	0.95	0.89	0.72	0.16	0.77	0.34
TNR	0.92	0.06	0.11	0.64	1.00	0.78	0.99

$$n = 200, p = 400, \text{SNR} = 0.25$$

	lasso	forward	backward	subset	posterior HPD	S_{min}	S_{small}
TPR	0.36	0.79	0.78	0.67	0.17	0.75	0.34
TNR	1.00	0.52	0.54	0.96	1.00	0.95	1.00

$$n = 200, p = 400, \text{SNR} = 1$$

	lasso	forward	backward	subset	posterior HPD	S_{min}	S_{small}
TPR	0.97	0.96	0.95	0.95	0.75	0.98	0.93
TNR	0.99	0.57	0.60	0.96	1.00	0.95	1.00

$$n = 500, p = 50, \text{SNR} = 0.25$$

	lasso	forward	backward	subset	posterior HPD	S_{min}	S_{small}
TPR	0.84	0.96	0.95	0.94	0.59	0.99	0.86
TNR	0.98	0.83	0.83	0.83	1.00	0.70	0.98

$$n = 500, p = 50, \text{SNR} = 1$$

	lasso	forward	backward	subset	posterior HPD	S_{min}	S_{small}
TPR	1.00	1.00	1.00	1.00	1.00	1.00	1.00
TNR	1.00	0.84	0.83	0.83	1.00	0.67	0.99

Table 1: True positive rates (TPR) and true negative rates (TNR) for Gaussian synthetic data with $p^* + 1 = 6$ active covariates including the intercept. The selection performance of S_{small} is excellent and similar to the the adaptive lasso. Classical subset, forward, and backward selection are too aggressive, while the posterior interval-based selection is too conservative.

4. Subset selection for predicting educational outcomes

Childhood educational outcomes are affected by adverse environmental exposures, such as poor air quality and lead, as well as social stressors, such as poverty, racial residential isolation, and neighborhood deviation. We study childhood educational development using end-of-grade reading test scores for a large cohort of fourth grade children in North Carolina (Children's Environmental Health Initiative, 2020). The reading scores are accompanied by a collection of student-level information detailed in Figure 3, which includes air quality exposures, birth information, blood lead measurements, and social and economic factors (see

also Kowal et al., 2021). The goal is to identify subsets of these variables and interactions that offer near-optimal prediction of reading scores as well as accurate classification of at-risk students.

Air quality during gestation, life of the child, and at time of testing	
PM25_T1	PM$_{2.5}$ 24 hour - Average of gestation weeks 1-13 (trimester 1) -- ugm3
PM25_T2	PM$_{2.5}$ 24 hour - Average of gestation weeks 14-26 (trimester 2) -- ugm3
PM25_T3	PM$_{2.5}$ 24 hour - Average of gestation weeks 27-delivery (trimester 3) -- ugm3
PM25_chronic	PM$_{2.5}$ Mean of 1 year prior to June 1 of the year of the EOG test
PM25_acute	PM$_{2.5}$ Mean of 30 days prior to May 1 of the year of the EOG test
Birth information	
BWTpct	Birthweight Percentile (based on clinical estimate of gestation)
mEdu	Mother's Education Group at time of birth (NoHS = No high school diploma, HS = High school diploma, College = College diploma)
mRace	Mother's Race/Ethnicity Group (White = White, NH Black = Non-Hispanic Black, Hispanic = Hispanic)
mAge	Mother's Age at time of birth
Male	Male Infant? (1 = Yes)
NotMarried	Not Married at time of birth? (1 = Yes)
Smoker	Mother Smoked? (1 = Yes)
Education / EOG test information	
EOG_Reading	Reading scale score for chronologically first EOG test taken (4th grade)
Blood lead level information	
Blood_lead	Blood Lead Level (micrograms per deciliter); maximum value if there are multiple tests
Social / Economic factors	
NDI_birth	Neighborhood Deprivation Index, at Birth
NDI_test	Neighborhood Deprivation Index, at time of EOG test
RI_birth	Residential Isolation for non-Hispanic Black, at Birth
RI_test	Residential Isolation for non-Hispanic Black, at time of EOG Test
EconDisadv	Participation in the free/reduced price lunch program (1 = Yes)

Figure 3: Variables in the NC education dataset. Data are restricted to individuals with 37-42 weeks gestation, mother's age 15-44, Blood_lead $\leq$ 80, birth order $\leq$ 4, no current limited English proficiency, and residence in NC at time of birth and time of 4th grade test.

A prominent feature of the data is the correlation among the covariates. After centering and scaling the continuous covariates to mean zero and standard deviation 0.5, we augment the variables in Figure 3 (excluding the test scores) with interactions between race and the social and economic factors. The resulting dimensions are $n = 16806$ and $p = 32$. Figure E.1 displays the pairwise correlations among the covariates and the response variable. There are strong associations among the air quality exposures as well as among race and the social and economic factors. Due to the dependences among variables, it is likely that distinct subsets of similar predictive ability can be obtained by interchanging among these correlated covariates. Hence, it is advantageous to collect and study the near-optimal subsets.

We compute acceptable families for (i) the reading scores y_i under squared error loss and (ii) the indicator $h(\tilde{y}_i) = \mathbb{I}\{\tilde{y}_i \geq \tau_{0.1}\}$ under cross-entropy loss, where $\tau_{0.1}$ is the 0.1-quantile

of the reading scores (see Appendix D). While task (i) broadly considers the spectrum of educational outcomes via reading scores, task (ii) targets at-risk students in the bottom 10% of reading ability. Acceptable families and accompanying quantities for both tasks can be computed using the same Bayesian model $\mathcal{M}$: we focus on Gaussian linear regression with horseshoe priors. The acceptable families are computed using the proposed BBA search with $\eta = 0\%$, $\varepsilon = 0.10$, $m_k = 100$ and $\{\tilde{\boldsymbol{x}}_i\}_{i=1}^{\tilde{n}} = \{\boldsymbol{x}_i\}_{i=1}^{n}$; results for other η values, $\varepsilon = 0.20$, and $m_k = 15$ are noted, while alternative target covariates $\{\tilde{\boldsymbol{x}}_i\}_{i=1}^{\tilde{n}}$ are in Section 4.2.

4.1 Subset selection for predicting reading scores

First, we predict reading scores using a linear model for $\mathcal{M}$ and squared error loss for L. Since acceptable family is defined based on $\widetilde{D}_{\mathcal{S}_{min},\mathcal{S}}^{out}$ in (19), we summarize its distribution in Figure 4. For each $\mathcal{S} \in \mathbb{S}$, we display 80% intervals, expectations, and the empirical analog $D_{\mathcal{S},\mathcal{S}_{min}}^{out} := 100 \times (L_{\mathcal{S}}^{out} - L_{\mathcal{S}_{min}}^{out})/L_{\mathcal{S}_{min}}^{out}$. The smaller subsets of sizes four to six demonstrate clear separation for certain subsets. Along with the intercept and the race indicators, these subsets include mEdu (college diploma), EconDisadv, and mEdu (completed high school) in sequence. However, larger subsets are needed to procure near-optimal predictions for smaller margins, such as $\eta < 2\%$. While the best subset $|\mathcal{S}_{min}| = 29$ includes nearly all of the covariates, many subsets with $|\mathcal{S}| > 10$ achieve within $\eta = 1\%$ of the accuracy of $\mathcal{S}_{min}$.
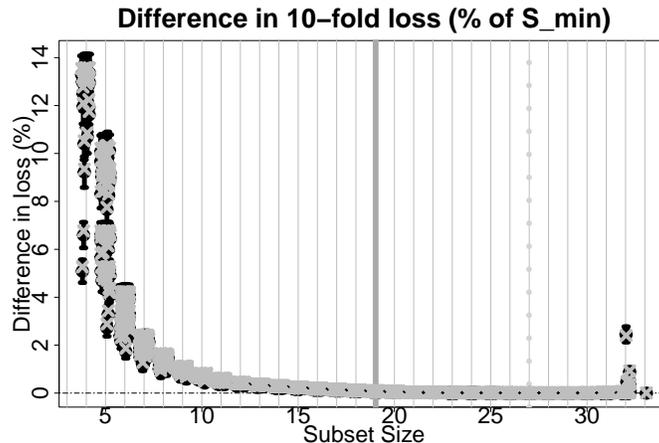

Figure 4: The 80% intervals (bars) and expected values (circles) for $\widetilde{D}_{\mathcal{S}_{min},\mathcal{S}}^{out}$ with $D_{\mathcal{S},\mathcal{S}_{min}}^{out}$ (x-marks) under squared error loss for each subset size $|\mathcal{S}|$ with $\mathcal{S} \in \mathbb{S}$. We annotate $\mathcal{S}_{min}$ (dashed gray line) and $\mathcal{S}_{small}$ (solid gray line) for $\varepsilon = 0.10$ and $\eta = 0$ and jitter the subset sizes for clarity of presentation.

Among the $|\mathbb{S}| = 2761$ candidate subsets identified from the BBA search, there are $|\mathbb{A}_{0,0.1}| = 1183$ acceptable subsets. We summarize $\mathbb{A}_{\eta,\varepsilon}$ via the co-variable importance metrics $\text{VI}_{\text{incl}}(j)$ and $\text{VI}_{\text{incl}}(j,\ell)$ in Figures 5 and E.2, respectively. Unlike many variable importance metrics that measure effect sizes, $\text{VI}_{\text{incl}}(j)$ instead quantifies whether each covariate j is a component of all, some, or no competitive subsets. There are many keystone covariates that appear in (nearly) all acceptable subsets, including environmental exposures (prenatal air quality and blood lead levels), economic and social factors (EconDisadv, mother's edu-

cation level, neighborhood deprivation at time of test), and demographic information (race, gender), among others.

Figure 5: Variable importance $\text{VI}_{\text{incl}}(j)$ for prediction. There are several tiers: variables appear in (nearly) all, many ($> 70\%$), some ($> 30\%$), or no acceptable subsets.

Interestingly, chronic and acute $PM_{2.5}$ exposure each belong to nearly 50% of acceptable subsets (Figure 5), yet rarely appear in the same acceptable subset (Figure E.2). The pairwise correlations (Figure E.1) offer a reasonable explanation: these variables are weakly correlated with reading scores but highly correlated with one another. Similar results persist for neighborhood deprivation and racial residential isolation both at birth and time of test. Moreover, this analysis was conducted after removing one acute $PM_{2.5}$ outlier (50% larger than all other values). When that outlier is kept in the data, acute $PM_{2.5}$ no longer belongs to *any* acceptable subset, while $\text{VI}_{\text{incl}}(j)$ for chronic $PM_{2.5}$ increases. These results are encouraging: the acceptable family identifies redundant yet distinct predictive explanations, but prefers the more stable covariate in the presence of outliers.

Next, we analyze the smallest acceptable subset $\mathcal{S}_{small}$ and incorporate uncertainty quantification for the accompanying linear coefficients. The $|\mathcal{S}_{small}| = 19$ selected covariates are displayed in Figure 6 alongside the point and 90% intervals based on the proposed approach, the Bayesian model $\mathcal{M}$, and the adaptive lasso. The estimates and intervals for covariates excluded from $\mathcal{S}_{small}$ are identically zero for the proposed approach (and, in this case, the adaptive lasso as well), while the estimates and HPD intervals from $\mathcal{M}$ are dense for all covariates. Despite the encouraging simulation results for the Zhao et al. (2021) frequentist intervals, these intervals often exclude the adaptive lasso point estimates, which undermines interpretability.

The estimates from $\mathcal{S}_{small}$ and $\mathcal{M}$ are similar with anticipated directionality: higher mother's education levels, lower blood lead levels in the child, less neighborhood deprivation, and absence of economic disadvantages predict higher reading scores. Prenatal air quality

exposure ($PM_{2.5}$) is significant: due to seasonal effects, the 1st and 3rd trimester exposures have negative coefficients, while the 2nd trimester has a positive effect. Naturally, these effects can only be interpreted jointly. Among interaction terms, the negative effect of NH Black × RI_birth suggests the lower reading scores among non-Hispanic Black students are accentuated by racial residential isolation in the neighborhood of the child's birth. Since we do not force all main effects into each subset, $\mathcal{S}_{small}$ does not contain an estimated effect of RI_birth for other race groups. Other interactions, such as the positive effects of Hisp and NH Black by EconDisadv and Hisp × Blood_lead, must also be interpreted carefully: the vast majority of Hispanic and non-Hispanic Black students belong to the EconDisadv group and have much higher blood lead levels on average, while each of NH Black, EconDisadv, and Blood_lead has a strong negative main effect.
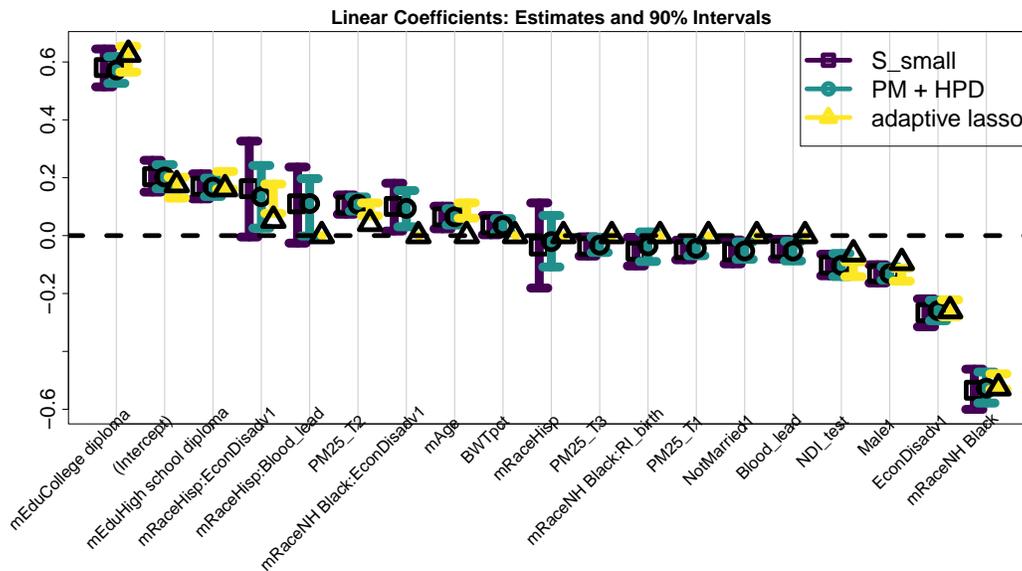
Figure 6: Estimated linear effects and 90% intervals for the variables in $\mathcal{S}_{small}$ based on the proposed approach, model $\mathcal{M}$, and the adaptive lasso.

The results are not particularly sensitive to m_k or ε. When $m_k = 15$, there are $|\mathbb{S}| = 436$ candidate subsets and $|\mathbb{A}_{0,0.1}| = 197$ acceptable subsets. The variable importance metrics broadly agree with Figures 5 and E.2, while $\mathcal{S}_{small}$—and therefore Figure 6—is unchanged. When $\varepsilon = 0.20$ (and $m_k = 100$ as before), the acceptable family reduces slightly to $|\mathbb{A}_{0,0.1}| = 977$ members and $\mathcal{S}_{small}$ differs only in the addition of Smoker and Hisp × NDI_test.

4.2 Out-of-sample prediction

We evaluate the predictive capabilities of the proposed approach for 20 training/testing splits of the NC education data. The same competing methods are adopted from Section 3, including the distinct search strategies for collecting near-optimal subsets. Since $\mathcal{S}_{small}$ and $\mathcal{S}_{min}$ are reasonably robust to m_k, we select $m_k = 15$ for computational efficiency. In addition, we include the acceptable family defined by setting $\{\tilde{\boldsymbol{x}}_i\}_{i=1}^{\tilde{n}}$ to be the testing data

covariate values (`bbound(Xtilde)`), which is otherwise identical to `bbound(bayes)`. Root mean squared prediction errors (RMSPEs) are used for evaluation.

The results are presented in Figure 7. Among single subset methods, $\mathcal{S}_{small}$ outperforms all competitors—including the classical `forward` and `backward` estimators and the smallest acceptable subsets from `lasso(bayes)`, `forward(bayes)`, and `backward(bayes)` discussed in Section 3 (not shown). The adaptive lasso selects fewer variables and is not competitive. Among search methods, Figure 7 confirms the results from the simulation study: the proposed BBA strategy (`bbound(bayes)`) identifies 10-25 times the number of subsets as the other search strategies, yet does not sacrifice any predictive accuracy in this expanded collection. Clearly, `bbound(bayes)` provides a more complete predictive picture, as the competing search strategies omit a massive number of subsets that *do* offer near-optimal prediction. The acceptable family based on the out-of-sample covariates `bbound(Xtilde)` is much larger, which is reasonable: the subsets are computed and evaluated on covariates for which the accompanying response variables are not available. This collection of subsets sacrifices some predictive accuracy relative to the other search strategies, yet still outperforms the adaptive lasso—yet another testament to the importance of the regularization induced by $\mathcal{M}$.
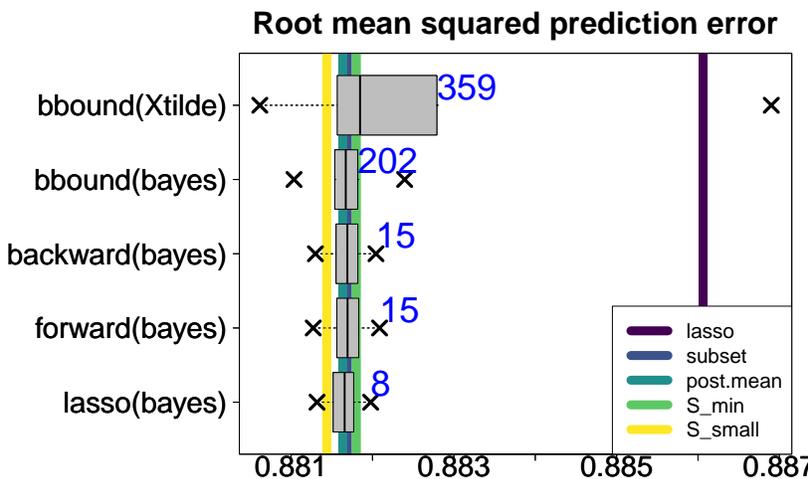
Figure 7: Root mean squared prediction errors (RMSPEs) across 20 training/testing splits of the NC education data. The boxplots summarize the RMSPE quantiles for the subsets within each acceptable family, while the vertical lines denote RMSPEs of competing methods. The average size of each acceptable family is annotated. The proposed BBA search returns vastly more subsets that remain highly competitive, while the accompanying $\mathcal{S}_{small}$ subset performs best overall.

Although the RMSPE differences appear to be small, minor improvements are practically relevant: even a single point on a standardized test score can be the difference between progression to the next grade level (on the low end) or eligibility for intellectually gifted programs (on the high end). More generally, education data are prone to weak signals and small effect sizes, which are conditions under which *many* methods may offer similar predictive performance. Nonetheless, our primary goal is not to substantially *improve* prediction, but rather to identify and analyze a large collection of near-optimal subsets.

The importance of the broader BBA search strategy is further highlighted in comparison `lasso(bayes)`, which uses a lasso search path for decision analytic Bayesian variable selection (Hahn and Carvalho, 2015; Kowal et al., 2021). In particular, `lasso(bayes)` generates only $|\mathbb{S}| = 25$ candidate subsets and $|\mathbb{A}_{0,0.1}| = 9$ acceptable subsets. By comparison, `bbound(bayes)` returns more than 100 times the number of candidate subsets *and* acceptable subsets. Figure 7 shows that the subsets omitted by `lasso(bayes)` yet discovered by `bbound(bayes)` are indeed near-optimal. Further, the restrictive search path of `lasso(bayes)` does not guarantee greater stability: the smallest acceptable subsets for `bbound(bayes)` and `lasso(bayes)` are nearly identical, yet the interquartile range of $|\mathcal{S}_{small}|$ across the 20 training/testing splits under `bbound(bayes)` is 0, while the same quantity under `lasso(bayes)` is 2. Indeed, by this metric, `bbound(bayes)` actually provides the *most stable* smallest acceptable subset across all search strategies considered.

5. Discussion

We developed decision analysis tools for Bayesian subset search, selection, and (co-) variable importance. The proposed strategy is outlined in Algorithm 1. Building from a Bayesian predictive model $\mathcal{M}$, we derived optimal linear actions for any subset of covariates. We explored the space of subsets using an adaptation of the branch-and-bound algorithm. After filtering to a manageable collection of promising subsets, we identified the *acceptable family* of near-optimal subsets for linear prediction or classification. The acceptable family was summarized by a new (co-) variable importance metric—the frequency with which variables (co-) appear in all, some, or no acceptable subsets—and individual member subsets, including the "best" and smallest subsets. Using the posterior predictive distribution from $\mathcal{M}$, we developed point and interval estimates for the linear coefficients of any subset. Simulation studies demonstrated better prediction, interval estimation, and variable selection for the proposed approach compared to existing Bayesian and frequentist selection methods—even for high-dimensional data with $p > n$.

We applied these tools to a large education dataset to study the factors that predict educational outcomes. The analysis identified several keystone covariates that appeared in (almost) every near-optimal subset, including environmental exposures, economic and social factors, and demographic information. The co-variable importance metrics highlighted an interesting phenomenon where certain pairs of covariates belonged to many acceptable subsets, yet rarely appeared in the same acceptable subset. Hence, these variables are effectively interchangeable for prediction, which provides valuable context for interpreting their respective effects. We showed that the smallest acceptable subset offers excellent prediction of end-of-grade reading scores and classification of at-risk students using substantially fewer covariates. The corresponding linear coefficients described new and important effects, for example that greater racial residential isolation among non-Hispanic Black students is predictive of lower reading scores. However, our results also caution against overreliance on any particular subset: we identified over 200 distinct subsets of variables that offer near-optimal out-of-sample predictive accuracy.

Future work will attempt to generalize these tools via the loss functions and the actions. Alternatives to squared error and cross-entropy loss can be incorporated with an IRLS approximation strategy similar to Section 2.3, which would maintain the methodology and

---

Algorithm 1: Bayesian subset selection for interpretable prediction and classification

---

1. Fit a *Bayesian predictive model* $\mathcal{M}$;

2. Specify a *loss function* for prediction or classification, including design points (covariate values) $\{\tilde{\boldsymbol{x}}_i\}_{i=1}^{\tilde{n}}$ and local weights $\{\omega(\tilde{\boldsymbol{x}}_i)\}_{i=1}^{\tilde{n}}$;

3. Filter to a family of *candidate subsets* of covariates:

    (a) Screen to the best $s_{max} \le p$ covariates based on $\mathcal{M}$;

    (b) Apply the branch-and-bound algorithm to select the best $m_k \le \binom{p}{k}$ subsets of each size $k = 1, \ldots, p$;

4. Collect the *acceptable family* of subsets that offer near-optimal prediction or classification;

5. Summarize the acceptable family via *(co-) variable importance*, the *best* subset, and the *smallest* subset.

6. Obtain posterior predictive samples of the coefficients $\tilde{\boldsymbol{\delta}}_{\mathcal{S}}$ for any subset $\mathcal{S}$ of interest.

---

algorithmic infrastructure from the proposed approach. Similarly, the class of parametrized actions can be expanded to include nonlinear predictors, such as trees or additive models, with acceptable families constructed in the same way.

Acknowledgments

Appendix A. Approximations for out-of-sample quantities

The acceptable family $\mathbb{A}_{\eta,\varepsilon}$ in (19) is based on the out-of-sample predictive discrepancy metric $\widetilde{D}_{\mathcal{S}_1,\mathcal{S}_2}^{out}$ and the best subset $\hat{\boldsymbol{\delta}}_{\mathcal{S}_{min}}$ in (18). These quantities both depend on the out-of-sample empirical and predictive losses $L_{\mathcal{S}}^{out}(k)$ and $\widetilde{L}_{\mathcal{S}}^{out}(k)$, respectively, in (16). Hence, it is required to compute (i) the optimal action on the training data, $\hat{\boldsymbol{\delta}}_{\mathcal{S}}^{-\mathcal{I}_k} :=$ $\arg\min_{\boldsymbol{\delta}_{\mathcal{S}}} \mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{y}^{-\mathcal{I}_k}} \mathcal{L}(\{\tilde{y}_i\}_{i\in\mathcal{I}_k}, \boldsymbol{\delta}_{\mathcal{S}})$, and (ii) samples from the out-of-sample predictive distri-

bution $\tilde{y}_i^{-\mathcal{I}_k} \sim p_{\mathcal{M}}(\tilde{y}_i|\boldsymbol{y}^{-\mathcal{I}_k})$. Because of the simplifications afforded by (4)-(5) and (11), computing $\hat{\boldsymbol{\delta}}_{\mathcal{S}}^{-\mathcal{I}_k}$ only requires the out-of-sample expectations $\hat{y}_i^{-\mathcal{I}_k} := \mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{y}^{-\mathcal{I}_k}}\{\tilde{\boldsymbol{y}}(\tilde{\boldsymbol{x}}_i)\}$ or $\hat{h}_i^{-\mathcal{I}_k} := \mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{y}^{-\mathcal{I}_k}} h(\tilde{y}_i)$ for the classification setting. For many models $\mathcal{M}$, there is a further simplification that $\mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{y}^{-\mathcal{I}_k}}\{\tilde{\boldsymbol{y}}(\tilde{\boldsymbol{x}}_i)\} = \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y}^{-\mathcal{I}_k}}[\mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{\theta}}\{\tilde{\boldsymbol{y}}(\tilde{\boldsymbol{x}}_i)\}]$, where often $\mathbb{E}_{\tilde{\boldsymbol{y}}|\boldsymbol{\theta}}\{\tilde{\boldsymbol{y}}(\tilde{\boldsymbol{x}}_i)\}$ has an explicit form (such as in regression models). Absent such simplifications, the expectations can be computed by averaging the draws of $\tilde{y}_i^{-\mathcal{I}_k} \sim p_{\mathcal{M}}(\tilde{y}_i|\boldsymbol{y}^{-\mathcal{I}_k})$.

Although these terms can be computed by repeatedly re-fitting the Bayesian model $\mathcal{M}$ for each training/validation split $k = 1, \ldots, K$, such an approach is computationally demanding. Instead, we apply a sampling-importance resampling (SIR) algorithm to approximate these out-of-sample quantities. Notably, the SIR algorithm requires only a single fit of model $\mathcal{M}$ to the complete data—which is already needed for posterior inference—and therefore contributes minimally to the computational cost of the aggregate analysis.

The details are provided in Algorithm 2. By construction, the samples $\{\tilde{y}_i^{\tilde{s}}\}_{\tilde{s}=\tilde{s}_1}^{\tilde{S}}$ are from the out-of-sample predictive distribution $p_{\mathcal{M}}(\tilde{y}_i|\boldsymbol{y}^{-\mathcal{I}_k})$. Based on Algorithm 2, it is straightforward to compute draws of $\widetilde{D}_{\mathcal{S}_1,\mathcal{S}_2}^{out}$ for any $\mathcal{S}_1, \mathcal{S}_2 \in \mathbb{S}$, as well as the best subset $\hat{\boldsymbol{\delta}}_{\mathcal{S}_{min}}$ in (18). Therefore, the acceptable family $\mathbb{A}_{\eta,\varepsilon}$ is readily computable for any η, ε. By default, we use $\tilde{S} = \lfloor S/2 \rfloor$ SIR samples.

Algorithm 2: Out-of-sample predictive evaluations.

1. Obtain posterior samples $\{\boldsymbol{\theta}^s\}_{s=1}^S \sim p_{\mathcal{M}}(\boldsymbol{\theta}|\boldsymbol{y})$;

2. For each training set $k = 1, \ldots, K$:

   (a) Compute $\log w_k^s \overset{c}{=} -\log p_{\mathcal{M}}(\boldsymbol{y}^{\mathcal{I}_k}|\boldsymbol{\theta}^s) = -\sum_{i \in \mathcal{I}_k} \log p_{\mathcal{M}}(\boldsymbol{y}_i|\boldsymbol{\theta}^s)$ (up to a constant);

   (b) Sample $\{\tilde{s}_1, \ldots, \tilde{s}_{\tilde{S}}\}$ without replacement from $\{1, \ldots, S\}$ with probability weights $\{w_k^1, \ldots, w_k^S\}$;

   (c) Sample $\tilde{y}_i^{\tilde{s}} \sim p_{\mathcal{M}}(\tilde{y}_i|\boldsymbol{\theta}^{\tilde{s}})$ for $\tilde{s} = \tilde{s}_1, \ldots, \tilde{s}_{\tilde{S}}$ and $i = 1, \ldots, n$;

   (d) Compute $\hat{y}_j^{-\mathcal{I}_k} \approx \tilde{S}^{-1} \sum_{\tilde{s}=\tilde{s}_1}^{\tilde{S}} \tilde{y}_j^{\tilde{s}}$ for $j \notin \mathcal{I}_k$;

   (e) Compute $\hat{\boldsymbol{\delta}}_{\mathcal{S}}^{-\mathcal{I}_k}$ for each $\mathcal{S} \in \mathbb{S}$ by solving (5) using the training data covariates $\{\boldsymbol{x}_i\}_{i \notin \mathcal{I}_k}$, the weights $\{\omega(\boldsymbol{x}_i)\}_{i \notin \mathcal{I}_k}$, and the pseudo-data $\{\hat{y}_j^{-\mathcal{I}_k}\}_{j \notin \mathcal{I}_k}$;

   (f) Compute $L_{\mathcal{S}}^{out}(k)$ and $\{\widetilde{L}_{\mathcal{S}}^{out,\tilde{s}}(k)\}_{\tilde{s}=\tilde{s}_1}^{\tilde{S}}$ in (16) using $\hat{\boldsymbol{\delta}}_{\mathcal{S}}^{-\mathcal{I}_k}$ and $\{\tilde{y}_i^{\tilde{s}}\}_{\tilde{s}=\tilde{s}_1}^{\tilde{S}}$;

3. Compute $L_{\mathcal{S}}^{out} = K^{-1} \sum_{k=1}^K L_{\mathcal{S}}^{out}(k)$ and $\widetilde{L}_{\mathcal{S}}^{out,\tilde{s}} = K^{-1} \sum_{k=1}^K \widetilde{L}_{\mathcal{S}}^{out,\tilde{s}}(k)$ for $\tilde{s} = \tilde{s}_1, \ldots, \tilde{S}$.

Algorithm 2 recycles computations: steps 1-2(d) are shared among all subsets $\mathcal{S}$ and loss functions L. Hence, the algorithm is efficient even when the number of candidate subsets $|\mathbb{S}|$ is large—and notably does not require re-fitting the Bayesian model $\mathcal{M}$. Modifications for classification (Section 2.3) are straightforward: steps (c), (d), and (e) are replaced by (c') $\tilde{\boldsymbol{y}}_i^{\tilde{s}} \sim p_{\mathcal{M}}(\tilde{y}_i|\boldsymbol{\theta}^{\tilde{s}})$ for $\tilde{s} = \tilde{s}_1, \ldots, \tilde{s}_{\tilde{S}}$ and $i = 1, \ldots, n$; (d') $\hat{h}_j^{-\mathcal{I}_k} \approx \tilde{S}^{-1} \sum_{\tilde{s}=\tilde{s}_1}^{\tilde{S}} h(\tilde{\boldsymbol{y}}_j^{\tilde{s}})$ for

$j \notin \mathcal{I}_k$; and (e') compute $\hat{\boldsymbol{\delta}}_{\mathcal{S}}^{-\mathcal{I}_k}$ by solving (11) using the training data covariates $\{\boldsymbol{x}_i\}_{i\notin\mathcal{I}_k}$, the weights $\{\omega(\boldsymbol{x}_i)\}_{i\notin\mathcal{I}_k}$, and the pseudo-data $\{\hat{h}_j^{-\mathcal{I}_k}\}_{j\notin\mathcal{I}_k}$.

Appendix B. Simulation study for classification

The synthetic data-generating process for classification mimics that for prediction: the only difference is that the data are generated as $y_i \overset{indep}{\sim} \text{Bernoulli}(\pi_i^*)$ with $\pi_i^* := \{1+\exp(-y_i^*)\}^{-1}$ and $y_i^* := \boldsymbol{x}_i'\boldsymbol{\beta}^*$ as before. For the Bayesian model $\mathcal{M}$, we use a logistic regression model with horseshoe priors and estimated using `rstanarm` (Goodrich et al., 2018). The competing estimators are constructed similarly as before, now using cross-entropy loss (10) for the proposed approach and the logistic likelihood for the adaptive lasso.

The classification performance is evaluated using cross-entropy loss for π_i^* in Figure B.1 (top), using the same competing search methods as in Figure 1. As in the regression case, the proposed `bbound(bayes)` search procedure returns vastly more subsets in the acceptable family, yet maintains excellent classification performance within this collection. In addition, classification based on $\mathcal{S}_{min}$ and $\mathcal{S}_{small}$ substantially outperform the adaptive lasso. We also include the 90% interval comparisons in Figure B.1 (bottom), which confirm the results from the prediction scenario: $\mathcal{S}_{small}$ and Zhao et al. (2021) (modified for the logistic case) achieve the nominal coverage with the narrowest intervals.
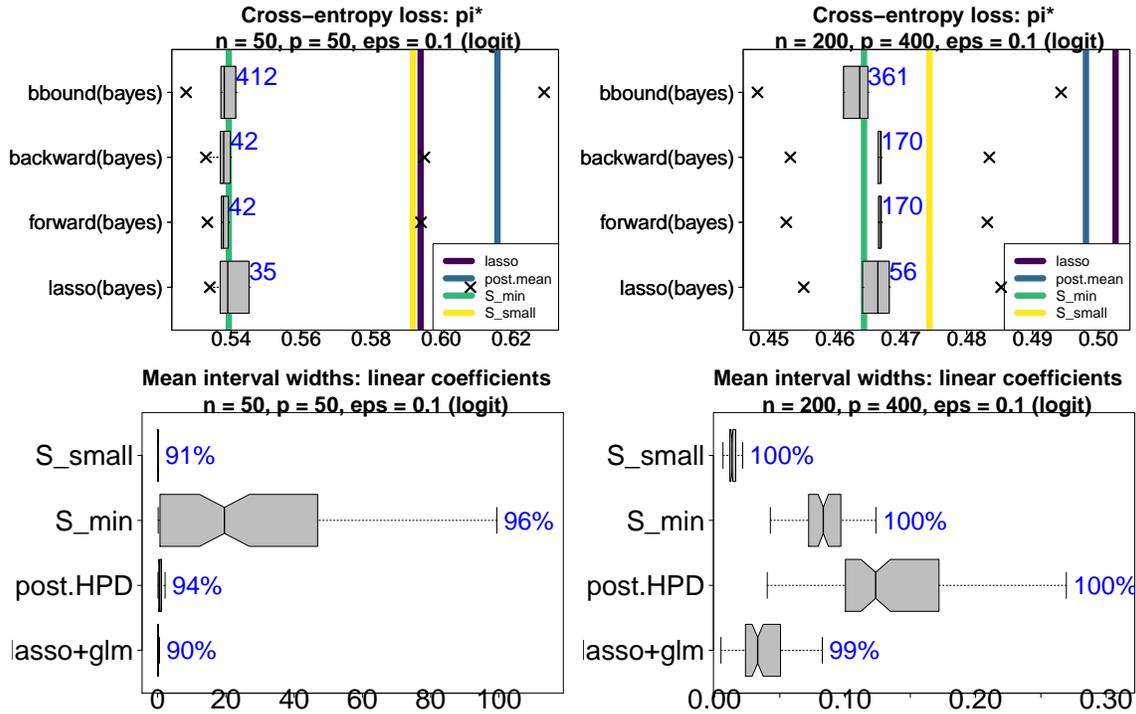
Figure B.1: **Top:** cross-entropy loss for π_i^*. The boxplots summarize the cross-entropy quantiles for the subsets within each acceptable family, while the vertical lines denote cross-entropy of competing methods. The average size of each acceptable family is annotated. **Bottom:** Mean interval widths (boxplots) with empirical coverage (annotations) for $\boldsymbol{\beta}^*$.

Appendix C. Simulation study for sensitivity analysis

We revisit the comparisons in Figure 1 to consider sensitivities to ε and $\{\tilde{\boldsymbol{x}}_i\}_{i=1}^{\tilde{n}}$. In Figure C.1, we report results for the acceptable families with $\varepsilon = 0.01$ and $\varepsilon = 0.20$; the frequentist estimators and `post.mean` are unchanged from Figure 1. Compared also to the analogous case in Figure 1 with $\varepsilon = 0.10$, we see the expected ordering in the cardinalities of the acceptable families: larger values of ε produce a more stringent criterion and therefore fewer acceptable subsets. Notably, the main results are not sensitive to the choice of $\varepsilon \in \{0.01, 0.10, 0.20\}$, and $\mathcal{S}_{small}$ performs exceptionally well in all cases.
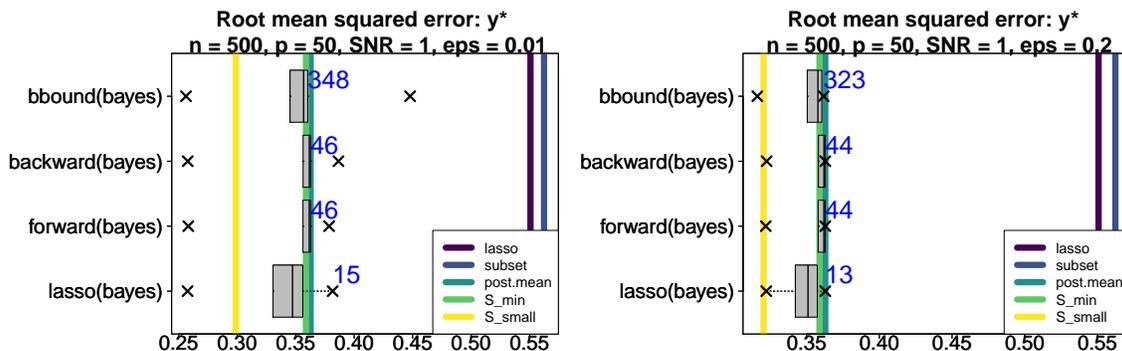
Figure C.1: Root mean squared errors (RMSEs) for predicting y^*. The boxplots summarize the RMSE quantiles for the subsets within each acceptable family; here, the acceptable families use $\varepsilon = 0.01$ (left) and $\varepsilon = 0.20$ (right). The vertical lines denote RMSEs of competing methods and the average size of each acceptable family is annotated. The results suggest only minor sensitivity to the choice of ε.

Next, we again modify the comparisons in Figure 1 to evaluate prediction at a *new* collection of covariates $\{\tilde{\boldsymbol{x}}_i\}_{i=1}^{\tilde{n}}$. We allow variations in the covariate data-generating process—either correlated standard normals from Section 3 or iid Uniform(0,1)—and whether the observed covariates $\{\boldsymbol{x}_i\}_{i=1}^{n}$ follow the same data-generating process as the target covariates $\{\tilde{\boldsymbol{x}}_i\}_{i=1}^{\tilde{n}}$. These configurations generate four scenarios, all with $\{\tilde{\boldsymbol{x}}_i\}_{i=1}^{\tilde{n}} \neq \{\boldsymbol{x}_i\}_{i=1}^{n}$. The competing acceptable families are each generated using this choice of $\{\tilde{\boldsymbol{x}}_i\}_{i=1}^{\tilde{n}}$ and the point predictions are evaluated for $y^*(\tilde{\boldsymbol{x}}_i) = \tilde{\boldsymbol{x}}_i'\boldsymbol{\beta}^*$ for $i = 1, \ldots, \tilde{n} = 1000$. The results are in Figure C.2. Naturally, performance worsens across the board when observed and target covariates differ in either their values ($\{\boldsymbol{x}_i\}_{i=1}^{n} \neq \{\tilde{\boldsymbol{x}}_i\}_{i=1}^{\tilde{n}}$) or their distributions. When the target covariates are iid uniform (top right, bottom left), all methods actually perform better—likely due to the light-tailed (bounded) and independent covariate values. Yet perhaps most remarkably, the relative performance among the methods remains consistent, while $\mathcal{S}_{small}$ outperforms all competitors in all but one scenario—and decisively outperforms the frequentist methods in all cases.

Appendix D. Subset selection for classifying at-risk students

We now focus the analysis on at-risk students via the functional $h(\tilde{y}_i) = \mathbb{I}\{\tilde{y}_i \geq \tau_{0.1}\}$ under cross-entropy loss, where $\tau_{0.1}$ is the 0.1-quantile of the reading scores. The posterior
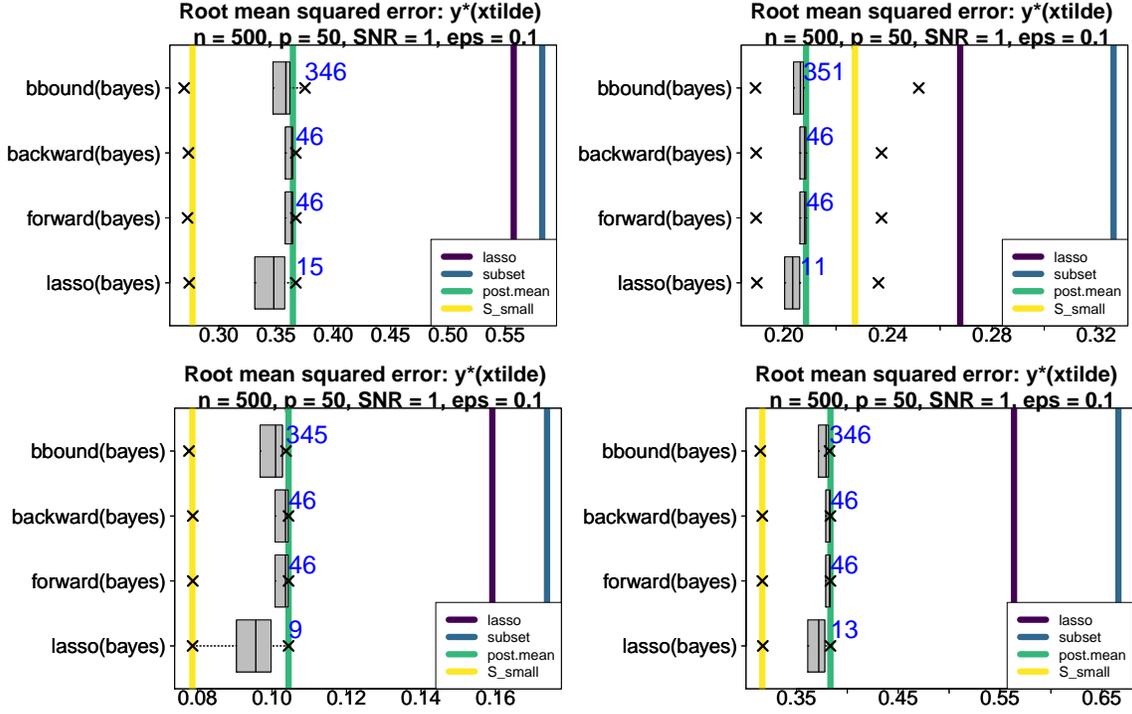
Figure C.2: Root mean squared errors (RMSEs) for predicting $y^*(\tilde{\boldsymbol{x}}_i) = \tilde{\boldsymbol{x}}_i'\boldsymbol{\beta}^*$ for covariates $\{\tilde{\boldsymbol{x}}_i\}_{i=1}^{\tilde{n}=1000}$ when the observed covariates $\{\boldsymbol{x}_i\}_{i=1}^n$ are correlated standard normals (top) or iid uniforms (bottom) and when the target covariates $\{\tilde{\boldsymbol{x}}_i\}_{i=1}^{\tilde{n}=1000}$ have the same distribution (left) or a different distribution (right) from $\{\boldsymbol{x}_i\}_{i=1}^n$. The boxplots summarize the RMSE quantiles for the subsets within each acceptable family. The vertical lines denote RMSEs of competing methods and the average size of each acceptable family is annotated.

predictive variables $\tilde{y}_i$ derive from the same model $\mathcal{M}$ as in Section 4, which eliminates the need for additional model specification, fitting, and diagnostics. However, it is also possible to fit a separate logistic regression model to the empirical functionals $h(y_i)$, which we do for the logistic adaptive lasso competitor.

It is most informative to study how the acceptable families $\mathbb{A}_{0,0.1}$ differ for targeted classification relative to prediction. The comparative distributions of $\widetilde{D}_{\mathcal{S}_{min},\mathcal{S}}^{out}$ (Figures 4 and D.1) clearly show that classification admits smaller subsets capable of matching the performance of $\mathcal{S}_{min}$ within 1%.

The variable importance metrics (Figures D.2 and E.3) identify the same keystone covariates as in the prediction setting. However, although there are many more acceptable subsets for classification ($|\mathbb{A}_{0,0.1}| = 1547$) than for prediction ($|\mathbb{A}_{0,0.1}| = 1183$), each $\mathrm{VI}_{\mathrm{incl}}(j)$ is generally much smaller for classification. Indeed, the smallest acceptable subset for classification is smaller, $|\mathcal{S}_{small}| = 16$, and is accompanied by 21 acceptable subsets of this size. In aggregate, these results suggest that near-optimal linear classification of at-risk students is achievable for a broader variety of smaller subsets of covariates.

Lastly, the selected covariates from $\mathcal{S}_{small}$ are plotted with 90% intervals in Figure D.3. Compared to the prediction version, $\mathcal{S}_{small}$ for classification replaces 1st trimester $\mathrm{PM}_{2.5}$
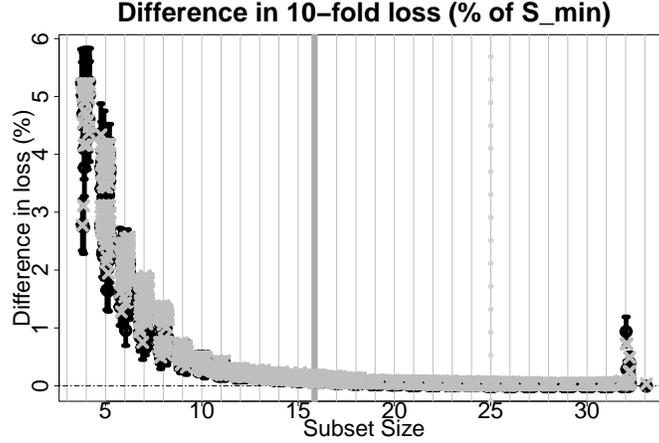
Figure D.1: The 80% intervals (bars) and expected values (circles) for $\widetilde{D}^{out}_{\mathcal{S}_{min},\mathcal{S}}$ with $D^{out}_{\mathcal{S},\mathcal{S}_{min}}$ (x-marks) under cross-entropy for each subset size $|\mathcal{S}|$ with $\mathcal{S} \in \mathbb{S}$. We annotate $\mathcal{S}_{min}$ (dashed gray line) and $\mathcal{S}_{small}$ (solid gray line) and jitter the subset sizes for clarity of presentation.
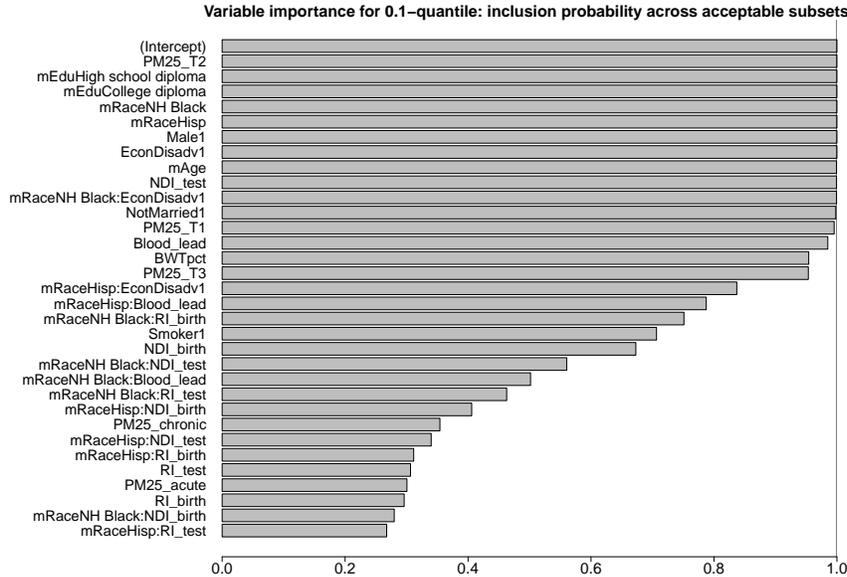

Figure D.2: Variable importance $\mathrm{VI}_{\mathrm{incl}}(j)$ for classification of the 0.1-quantile ($m_k = 100$).

exposure with acute PM$_{2.5}$ exposure and omits the interactions `Hisp` $\times$ `Blood_lead` and `Hisp` $\times$ `EconDisadv`, but is otherwise the same. The posterior expectations under $\mathcal{M}$ are excluded because they are not targeted for prediction of $h(\tilde{y}_i)$ and therefore are not directly comparable. However, there is some disagreement between $\mathcal{S}_{small}$ and the adaptive lasso, the latter of which excludes several keystone covariates and again suffers from inconsistency between the point and interval estimates.
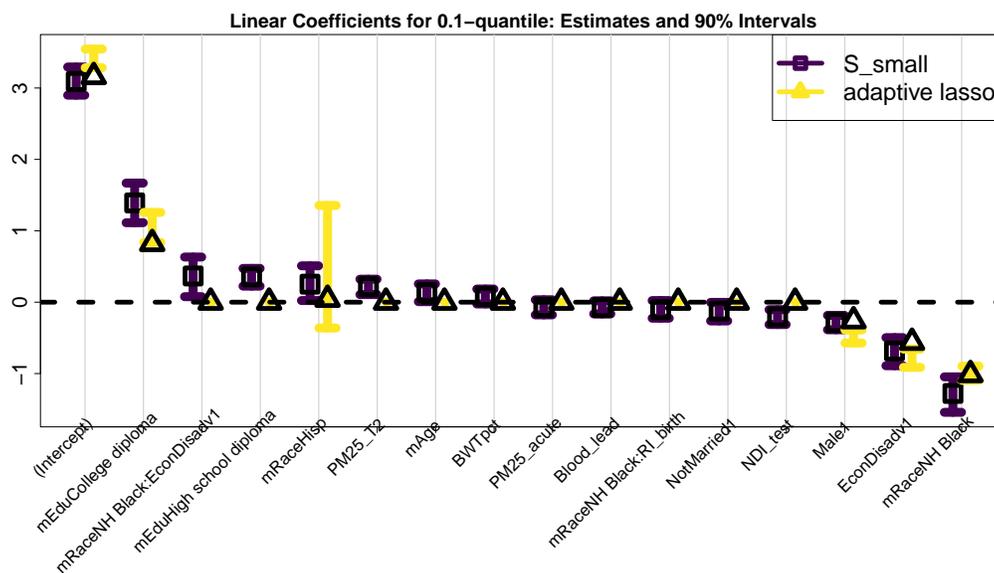
Figure D.3: Estimated linear effects and 90% intervals for the variables in $\mathcal{S}_{small}$ based on the proposed approach and the adaptive lasso.

Appendix E. Supporting figures

These figures include: pairwise correlations among the covariate (Figure E.1) and co-variable importance for prediction (Figure E.2) and classification (Figure E.3).
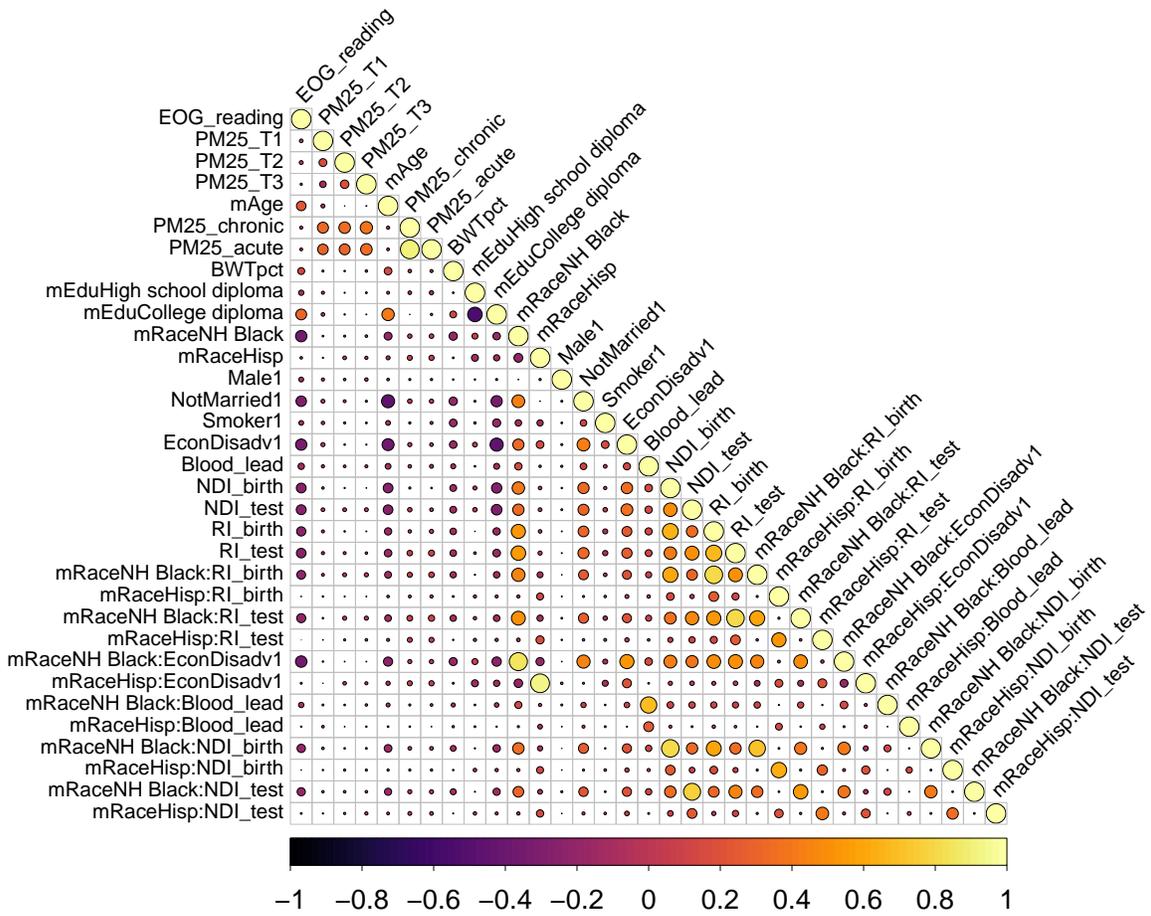
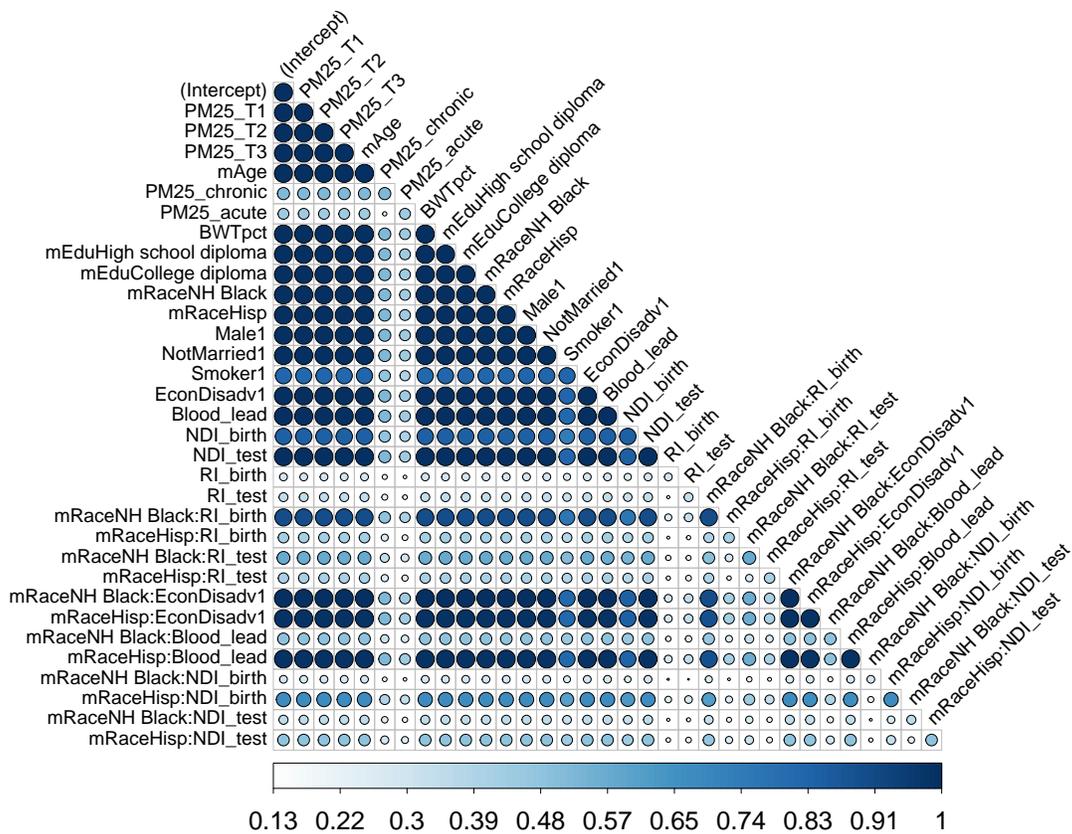Figure E.1: Pairwise correlations among the covariates in the NC education data.

Figure E.2: Co-variable importance $\mathrm{VI}_{\mathrm{incl}}(j, \ell)$ for prediction. For certain pairs of variables (chronic and acute $\mathrm{PM}_{2.5}$ exposure; neighborhood deprivation and racial residential isolation both at birth and time of test), it is common for one—but not both—to appear in an acceptable subset.
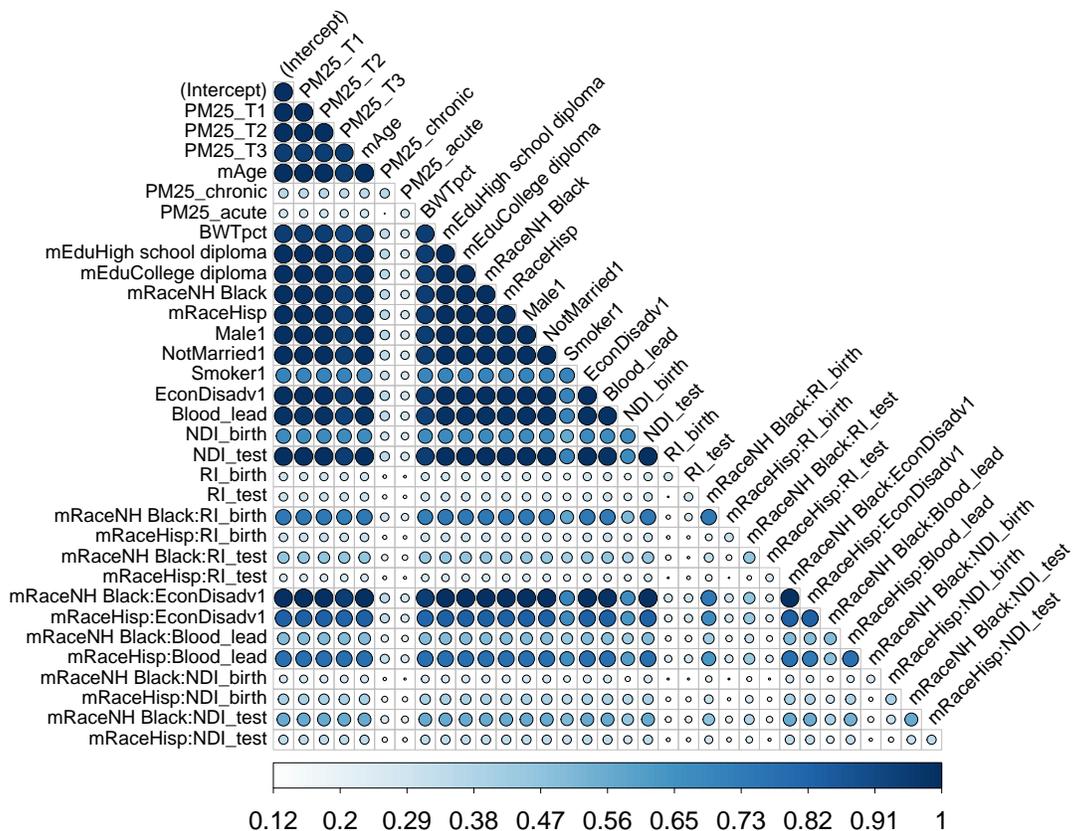
Figure E.3: Co-variable importance $\mathrm{VI}_{\mathrm{incl}}(j, \ell)$ for classification of the 0.1-quantile.

References

Homayun Afrabandpey, Tomi Peltola, Juho Piironen, Aki Vehtari, and Samuel Kaski. A decision-theoretic approach for model interpretability in Bayesian framework. *Machine Learning*, 109(9):1855–1876, 2020. ISSN 1573-0565.

Maria Maddalena Barbieri and James O. Berger. Optimal predictive model selection. *Annals of Statistics*, 32(3):870–897, 2004. ISSN 00905364. doi: 10.1214/009053604000000238.

Amir Bashir, Carlos M. Carvalho, P. Richard Hahn, and M. Beatrix Jones. Post-processing posteriors over precision matrices to produce sparse graph estimates. *Bayesian Analysis*, 14(4):1075–1090, 2019. ISSN 19316690. doi: 10.1214/18-BA1139.

José M Bernardo and Adrian F M Smith. *Bayesian theory*, volume 405. John Wiley and Sons, Inc., 2009. ISBN 047031771X.

Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. *Annals of statistics*, 44(2):813–852, 2016. ISSN 0090-5364.

Howard D. Bondell and Brian J. Reich. Consistent high-dimensional Bayesian variable selection via penalized credible regions. *Journal of the American Statistical Association*, 107(500):1610–1624, 2012. ISSN 01621459. doi: 10.1080/01621459.2012.716344.

Leo Breiman. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001. ISSN 0883-4237. doi: 10.1214/ss/1009213726.

Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010. ISSN 00063444. doi: 10.1093/biomet/asq017.

Children's Environmental Health Initiative. Linked Births, Lead Surveillance, Grade 4 End-Of-Grade (EoG) Scores [Data set], 2020. URL `https://doi.org/10.25614/COHORT_2000`.

Jyotishka Datta and Jayanta K Ghosh. Asymptotic properties of Bayes risk for the horseshoe prior. *Bayesian Analysis*, 8(1):111–132, 2013.

Jiayun Dong and Cynthia Rudin. Variable importance clouds: A way to explore variable importance for the set of good models. *arXiv preprint arXiv:1901.03209*, 2019.

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 70(5):849–911, 2008. ISSN 13697412. doi: 10.1111/j.1467-9868.2008.00674.x.

Jianqing Fan and Jinchi Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101–148, 2010. ISSN 10170405.

George M Furnival and Robert W Wilson. Regressions by leaps and bounds. *Technometrics*, 42(1):69–79, 2000. ISSN 0040-1706.

Cristian Gatu and Erricos John Kontoghiorghes. Branch-and-bound algorithms for computing the best-subset regression models. *Journal of Computational and Graphical Statistics*, 15(1):139–156, 2006. ISSN 1061-8600.

A E Gelfand, D K Dey, and H Chang. Model determination using predictive distributions, with implementation via sampling-based methods (with discussion). *Bayesian Statistics 4*, 4:147–167, 1992. ISSN 01621459.

Ben Goodrich, Jonah Gabry, Imad Ali, and Sam Brilleman. rstanarm: Bayesian applied regression modeling via Stan., 2018. URL `http://mc-stan.org/`.

Constantinos Goutis and Christian P. Robert. Model choice in generalised linear models: A Bayesian approach via Kullback-Leibler projections. *Biometrika*, 85(1):29–37, 1998. ISSN 00063444. doi: 10.1093/biomet/85.1.29.

P. Richard Hahn and Carlos M. Carvalho. Decoupling shrinkage and selection in bayesian linear models: A posterior summary perspective. *Journal of the American Statistical Association*, 110(509):435–448, 2015. ISSN 1537274X. doi: 10.1080/01621459.2014.993077.

P. Richard Hahn, Jingyu He, and Hedibert F. Lopes. Efficient Sampling for Gaussian Linear Regression With Arbitrary Priors. *Journal of Computational and Graphical Statistics*, 28 (1):142–154, 2019. ISSN 15372715. doi: 10.1080/10618600.2018.1482762.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*, volume 2. Springer, 2009. ISBN 0387848584.

Trevor Hastie, Robert Tibshirani, and Ryan Tibshirani. Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons. *Statistical Science*, 35(4):579–592, 2020. ISSN 0883-4237.

David W Hosmer, Borko Jovanovic, and Stanley Lemeshow. Best subsets logistic regression. *Biometrics*, pages 1265–1270, 1989. ISSN 0006-341X.

Florian Huber, Gary Koop, and Luca Onorante. Inducing Sparsity and Shrinkage in Time-Varying Parameter Models. *Journal of Business and Economic Statistics*, pages 1–48, 2020. ISSN 15372707. doi: 10.1080/07350015.2020.1713796.

Hemant Ishwaran and J. Sunil Rao. Spike and slab variable selection: Frequentist and Bayesian strategies. *Annals of Statistics*, 33(2):730–773, 2005. ISSN 00905364. doi: 10.1214/009053604000001147.

Jiming Jiang, J Sunil Rao, Zhonghua Gu, and Thuan Nguyen. Fence methods for mixed model selection. *The Annals of Statistics*, 36(4):1669–1692, 2008. ISSN 0090-5364.

Shiqiang Jin and Gyuhyeong Goh. Bayesian selection of best subsets via hybrid search. *Computational Statistics*, pages 1–17, 2020. ISSN 0943-4062.

Nicholas Kissel and Lucas Mentch. Forward Stability and Model Path Selection. *arXiv preprint arXiv:2103.03462*, 2021.

Daniel R. Kowal. Fast, Optimal, and Targeted Predictions using Parametrized Decision Analysis. *Journal of the American Statistical Association*, 2021. doi: 10.1080/01621459. 2021.1891926.

Daniel R Kowal and Daniel C Bourgeois. Bayesian Function-on-Scalars Regression for High-Dimensional Data. *Journal of Computational and Graphical Statistics*, 29(3):1–10, 2020. ISSN 1061-8600. doi: 10.1080/10618600.2019.1710837.

Daniel R. Kowal, Mercedes Bravo, Henry Leong, Robert J. Griffin, Katherine B. Ensor, and Marie Lynn Miranda. Bayesian Variable Selection for Understanding Mixtures in Environmental Exposures. *Statistics in Medicine*, 40(22):4850–4871, 2021.

Jing Lei. Cross-Validation With Confidence. *Journal of the American Statistical Association*, 0(0):1–53, 2019. ISSN 0162-1459. doi: 10.1080/01621459.2019.1672556. URL https://doi.org/10.1080/01621459.2019.1672556.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018. ISSN 0162-1459.

Faming Liang, Qifan Song, and Kai Yu. Bayesian subset modeling for high-dimensional generalized linear models. *Journal of the American Statistical Association*, 108(502): 589–606, 2013. ISSN 0162-1459.

D. V. Lindley. The Choice of Variables in Multiple Regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(1):31–53, 1968. doi: 10.1111/j.2517-6161. 1968.tb01505.x.

Alan J Miller. Selection of subsets of regression variables. *Journal of the Royal Statistical Society: Series A (General)*, 147(3):389–410, 1984. ISSN 0035-9238.

David J. Nott and Chenlei Leng. Bayesian projection approaches to variable selection in generalized linear models. *Computational Statistics and Data Analysis*, 54(12):3227–3241, 2010. ISSN 01679473. doi: 10.1016/j.csda.2010.01.036.

Juho Piironen, Markus Paasiniemi, and Aki Vehtari. Projective inference in high-dimensional problems: Prediction and feature selection. *Electronic Journal of Statistics*, 14(1):2155–2197, 2020. ISSN 1935-7524. URL http://arxiv.org/abs/1810.02406.

Nicholas G. Polson and James G. Scott. Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction. *Bayesian Statistics*, 9780199694:501–538, 2010. doi: 10. 1093/acprof:oso/9780199694587.003.0017.

David Puelz, P. Richard Hahn, and Carlos M. Carvalho. Variable selection in seemingly unrelated regressions with random predictors. *Bayesian Analysis*, 12(4):969–989, 2017. ISSN 19316690. doi: 10.1214/17-BA1053.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Lesia Semenova and Cynthia Rudin. A study in Rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning. *arXiv preprint arXiv:1908.01755*, 2019. URL `http://arxiv.org/abs/1908.01755`.

Minh Ngoc Tran, David J. Nott, and Chenlei Leng. The predictive Lasso. *Statistics and Computing*, 22(5):1069–1084, 2012. ISSN 09603174. doi: 10.1007/s11222-011-9279-3.

Aki Vehtari and Janne Ojanen. A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6(1):142–228, 2012. ISSN 19357516. doi: 10.1214/12-ss102.

Spencer Woody, Carlos M. Carvalho, and Jared S. Murray. Model interpretation through lower-dimensional posterior summarization. *Journal of Computational and Graphical Statistics*, pages 1–9, 2020. ISSN 1061-8600. URL `http://arxiv.org/abs/1905.07103`.

Shaohua Wu, T J Harris, and K B McAuley. The use of simplified or misspecified models: Linear case. *The Canadian Journal of Chemical Engineering*, 85(4):386–398, 2007. ISSN 0008-4034.

Sen Zhao, Daniela Witten, and Ali Shojaie. In defense of the indefensible: A very naive approach to high-dimensional inference. *Statistical Science*, 36(4):562–577, 2021. ISSN 0883-4237.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006. ISSN 01621459. doi: 10.1198/016214506000000735.