Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Bayesian Variable Selection for Understanding Mixtures in Environmental Exposures

Daniel R. Kowal[1] | Mercedes Bravo[2,3] | Henry Leong[3] | Alexander Bui[4] | Robert J. Griffin[4] | Katherine B. Ensor[1] | Marie Lynn Miranda[3,5]

[1]Department of Statistics, Rice University, Texas, U.S.A.
[2]Biostatistics and Epidemiology Division, RTI International, North Carolina, U.S.A.
[3]Children's Environmental Health Initiative, University of Notre Dame, Indiana, U.S.A.
[4]Department of Civil and Environmental Engineering, Rice University, Texas, U.S.A.
[5]Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Indiana, U.S.A.

Correspondence
Daniel R. Kowal. Email: daniel.kowal@rice.edu

Present Address
Department of Statistics, MS 138, Rice University, Houston, TX 77251-1892

Summary

Social and environmental stressors are crucial factors in child development. However, there exists a multitude of measurable social and environmental factors—the effects of which may be cumulative, interactive, or null. Using a comprehensive cohort of children in North Carolina, we study the impact of social and environmental variables on 4th end-of-grade exam scores in reading and mathematics. To identify the essential factors that predict these educational outcomes, we design new tools for Bayesian linear variable selection using decision analysis. We extract a predictive optimal subset of explanatory variables by coupling a loss function with a novel model-based penalization scheme, which leads to coherent Bayesian decision analysis and empirically improves variable selection, estimation, and prediction on simulated data. The Bayesian linear model propagates uncertainty quantification to all predictive evaluations, which is important for interpretable and robust model comparisons. These predictive comparisons are conducted out-of-sample with a customized approximation algorithm that avoids computationally intensive model re-fitting. We apply our variable selection techniques to identify the joint collection of social and environmental stressors—and their interactions—that offer clear and quantifiable improvements in prediction of reading and mathematics exam scores.

KEYWORDS:
Air Quality, Educational Outcomes, Lead, Prediction, Regression

1 | INTRODUCTION

It is widely agreed that health and well-being are determined by multiple forces. Environmental exposures such as air pollution and lead adversely affect health and cognitive development, and are often elevated in disadvantaged communities. These same communities may also experience a multitude of social stressors: inadequate access to health care, poor housing, high unemployment, high crime, high poverty, and racial segregation, among others.[1,2] While social and environmental stressor co-exposures may or may not interact, they almost certainly cumulate. In addition, health and developmental impacts during childhood are strongly associated with poor outcomes in adults. Thus, understanding the effects of co-exposures to social and environmental stressors among children is critical to improving the nation's health. Here, we develop methods to better understand the relationships among environmental mixtures and social context, as well as their impacts on developmental outcomes.

Although the evidence is not yet conclusive, a growing body of epidemiological research suggests that pre- and post-natal exposure to ambient air pollution, including $PM_{2.5}$, adversely impacts neurological development in children.[3] Prenatal $PM_{2.5}$ exposure specifically has been linked with poorer function across memory and attention domains in children.[4] Meanwhile, there is a plethora of evidence that lead exposure in young children, even at low levels, is associated with intellectual deficits,[5] learning and behavioral disorders,[6] and lower scores on intelligence and standardized tests.[7] The adverse effects of childhood lead exposure persist into adulthood, affecting intelligence and socioeconomic status.[8]

Social stressors, often present in the same communities as these adverse environmental exposures, may also affect health and development. Neighborhood measures of poverty, disadvantage, and deprivation are associated with decrements in cognitive development in children,[9] and a substantial body of evidence links segregation with adverse health outcomes.[10] Despite evidence linking neighborhood conditions with health[11] and developmental outcomes in children,[12] the effects of environmental exposures such as air pollution and lead are not always evaluated in conjunction with neighborhood conditions.

Our overarching objective is to identify the environmental exposures (e.g., air pollution exposure, lead exposure) and social stressors (e.g., racial residential segregation, neighborhood deprivation, income status)—as well as their interactions—that predict child educational outcomes. For this task, we develop new decision tools for Bayesian variable selection. Although shrinkage and sparsity priors can encode prior beliefs or preferences for sparsity, the prior alone cannot *select* variables. In particular, the posterior distribution encapsulates information about variable importance, but a *decision* about the active or inactive variables is still required. A natural solution is a two-stage approach: after obtaining the posterior distribution of a Bayesian (linear) model, we extract sparse linear summaries that are optimized for prediction. From a decision analysis perspective, we pair a loss function with a novel model-based penalization scheme that enables simultaneous optimization for predictive accuracy and sparsity. Unlike marginal decision criteria that examine each regression coefficient individually—hard-thresholding,[13] posterior inclusion probabilities,[14] or posterior credible intervals that exclude zero[15]—our approach selects variables that are *jointly* predictive, which is particularly important in the presence of correlated variables. We quantify the predictive contributions of particular variables or sets of variables by comparing the out-of-sample predictive accuracy across distinct sparsity levels. Crucially, we leverage the predictive distribution of the Bayesian linear model to augment these predictive evaluations with posterior uncertainty quantification. The uncertainty quantification is valuable for informed comparisons and provides a mechanism for selecting the smallest subset among *nearly*-optimal linear models. For computational scalability, we design an approximation algorithm that avoids intensive model re-fitting for the out-of-sample evaluations and instead only requires a single model fit.

The statistical analysis is enabled by linking multiple administrative datasets to construct an *analysis dataset* that tracks children in North Carolina from birth through time of lead testing and finally time of 4th grade standardized testing in reading and mathematics. This work is also responsive to calls from leading researchers in support of research agendas that leverage longitudinal population-based data linkages.[12]

The paper is organized as follows: Section 2 introduces the comprehensive database constructed for our study; in Section 3, we develop in full the Bayesian variable selection methodology; results are presented in Section 4, which includes a simulation study; we conclude in Section 5.

2 | DATA

The analysis dataset[16] for this study was created by linking three administrative databases for the State of North Carolina (NC)—detailed birth records, blood lead surveillance data, and end-of-grade (EOG) standardized testing data—with extensive exposure data. The exposure data include ambient air quality and temperature, socioeconomic factors, and indices that measure racial isolation and neighborhood deprivation. In all, 25 variables and features are included. A short summary of the dataset is included here, with a comprehensive list of variables provided in the Appendix. More detailed discussion of the dataset is available in Bravo et al.[17]

Detailed birth records were obtained from the NC State Center for Health Statistics, Vital Statistics Department. The data include information on maternal demographics, maternal and infant health, and maternal obstetrics history for all documented live births in NC. The analysis dataset restricts to students born in 2000, which facilitates linking with the Census-based datasets below.

Blood lead surveillance records were obtained from the state registry maintained by the Childhood Lead Poisoning Prevention Program of the Children's Environmental Health Unit, Department of Health and Human Services in Raleigh, NC. Blood lead levels are stored as integer values. The analysis dataset includes measurements from 2000-2011.

EOG standardized testing data were obtained from the NC Education Research Data Center of Duke University in Durham, NC. Children in NC in grades 3–8 are tested in reading and mathematics at the end of the school year.[18] The database contains records for all EOG test results statewide for tests from the 1995–1996 school year to the present; the analysis dataset restricts to 4th grade EOG exams in reading and mathematics in 2010-2012. The dataset also includes identifying information such as name and birth date and data on demographics and socioeconomics, English proficiency, and school district.

Air pollution exposure metrics are based on daily 24-hour averages of ambient $PM_{2.5}$ concentrations obtained from two different sources: the US Environmental Protection Agency National Air Monitor Stations or State and Local Air Monitoring Stations (NAMS/SLAMS) monitoring network (1999-2000); and publicly available output from the Community Multiscale Air Quality (CMAQ) downscaler (2010-2012). These exposure metrics are converted into *pre-natal exposure*, which is the average exposure by trimester of pregnancy (1-13 weeks, 14-26 weeks, and 27 weeks to birth); *chronic exposure*, which is the average exposure for one year prior to EOG testing; and *acute exposure*, which is the average exposure for 30 days prior to EOG testing. The prenatal metrics use the closest monitor within 30km of the mother's residence, while the chronic and acute metrics use the closest monitor within 30km of the child's residence.

Ambient temperature data were obtained from the State Climate Office of NC. Temperature exposure was computed as the average exposure by trimester of pregnancy (1-13 weeks, 14-26 weeks, and 27 weeks to birth) using the nearest weather station within 50km of the mother's residence.

Racial isolation (RI) and neighborhood deprivation (NDI) are critical indices that measure social stress. RI is based on 2000 and 2010 Census data using a previously derived local, spatial measure of RI,[19] which is in turn derived from the global spatial isolation index developed by Reardon and O'Sullivan.[20] The RI index ranges from 0 to 1, with values close to 1 indicating the neighborhood environment is almost entirely non-Hispanic black (NHB). NDI values were calculated at the census tract level based on 2000 and 2010 Census data using a previously derived measure of NDI.[21] The NDI empirically summarizes eight census variables representing five socio-demographic domains (income/poverty, education, employment, housing, and occupation), including: percent of males in management and professional occupations, percent of crowded housing, percent of households in poverty, percent of female headed households with dependents, percent of households on public assistance and households earning <$30,000 per year estimating poverty, percent earning less than a high school education, and the percent unemployed. Higher values of NDI indicate more severe neighborhood deprivation. To account for neighborhood dynamics, each child was assigned an RI and NDI value at time of birth and at time of EOG testing. RI and NDI at birth were assigned based on the child's tract of residence at time of birth (obtained from the detailed birth records), using RI and NDI calculated from 2000 Census data. RI and NDI at time of EOG testing were assigned based on the child's tract of residence at time of testing (obtained from the EOG testing data), using RI and NDI calculated from 2010 Census data.

Methods for receiving, storing, linking, analyzing, and presenting results related to this study were all governed by a research protocol approved by the Rice University Institutional Review Board.

3 | METHODS

3.1 | The Bayesian linear model

Bayesian linear regression provides a compelling platform from which to address the challenges in our analysis: the linear modeling structure inherits the interpretability of classical linear regression, while the prior distribution conveys advantages for regularization of unimportant variables, (posterior) uncertainty quantification, and improved predictions. Consider the paired observations $\{y_i, x_i\}_{i=1}^n$ where $y_i \in \mathbb{R}$ denotes the response variable and x_i is the p-dimensional predictor, which may include interaction terms. The Bayesian linear model is

$$y_i = x_i'\beta + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2) \tag{1}$$

$$\beta \sim p(\beta) \tag{2}$$

where $p(\beta)$ denotes a prior distribution on the regression coefficients. The prior distribution (2) has long served as a mechanism for introducing regularization in the linear model: Bayesian ridge regression $\beta_j \sim N(0, \sigma_\beta^2)$ and the Bayesian lasso[22] are adaptations of non-Bayesian regularization techniques which seek to combat multicollinearity and mitigate the effects of unimportant predictor variables. More commonly, Bayesian regression utilizes *shrinkage priors* such as the horseshoe prior[23] and the Dirichlet-Laplace prior,[24] which demonstrate improved point estimation and posterior contraction properties,[13,25] especially for sparse regressions in which many coefficients may be negligible. Shrinkage priors are continuous distributions, usually scale

mixtures of Gaussian distributions, which are designed for both regularization and computational scalability. By comparison, sparsity-inducing priors, such as spike-and-slab priors,[26,27,28,29] place a positive probability on $\beta_j = 0$. Although conceptually appealing, sparsity-inducing priors face significant challenges for computation and inference. Nonetheless, sparsity-inducing priors are compatible with the methods introduced below.

Regardless of the choice of prior, Bayesian inference proceeds via the posterior distribution $p(\beta|y)$ from model (1)-(2). The posterior distribution is rarely available analytically, especially for shrinkage priors, and therefore posterior inference most commonly uses Markov Chain Monte Carlo (MCMC) algorithms. For many choices of (2), an efficient elliptical slice sampler[30] is available for obtaining samples from the posterior distribution of the regression coefficients. Using these samples, posterior summaries such as expectations, standard deviations, and credible intervals are easily computable.

Although the posterior distribution $p(\beta|y)$ encapsulates information about variable importance, additional tools are needed to select and evaluate joint subsets of variables. In practice, the vast majority of Bayesian variable selection proceeds marginally for each coefficient β_j, including hard-thresholding under shrinkage priors,[13] marginal inclusion probabilities under sparsity-inducing priors,[14] or selection based on credible intervals that omit zero.[15] While useful summaries, these approaches fail to consider variables jointly, which is particularly important for correlated predictors and interactions. In addition, marginally-selected variables do not necessarily comprise an adequately predictive linear basis relative to the full model, and indeed are not constructed or evaluated for this purpose. The implication is that, for a set of *marginally-selected* variables, it is incorrect to interpret them jointly as a predictive model: for instance, we cannot claim that these variables are sufficient to explain the variability in the data. Marginal selection is not without benefits, for example as a tool to screen unimportant variables in ultrahigh-dimensional data,[31] but must be interpreted carefully. Alternatively, some point estimation techniques produce sparse estimates of β within a Bayesian linear model,[32] but these methods are not accompanied by uncertainty quantification for either the coefficients β or metrics for evaluating predictive accuracy associated with various sparsity levels. Consequently, techniques that provide both selection *and* uncertainty quantification for the selection process are at a premium.

3.2 | Decision analysis for Bayesian variable selection

Bayesian variable selection is a *decision problem*: using the full model posterior, the task is to determine a minimal subset of variables that nearly preserve—or if possible, improve upon—the fitness of the full model. Fitness is evaluated using predictive performance and measured by a loss function, such as squared error loss. Importantly, the loss function may be augmented with a *sparsity-inducing penalty*, which admits a formal expression of the tradeoff between *predictive accuracy* and *model simplicity*. Decision analysis proceeds to minimize the posterior predictive expected loss; intuitively, this procedure averages the loss function over future or unobserved values of the response variable and then minimizes the resulting expression. Importantly, we obtain computationally convenient representations of this objective, with model selection uncertainty quantification (Section 3.4) and rapidly-computable out-of-sample comparisons (Section 3.5).

For any covariate value $\tilde{x}$, the linear model assigns a *posterior predictive* distribution to the associated response variable $\tilde{y}$:

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta, \quad \theta = (\beta, \sigma^2) \tag{3}$$

where $\tilde{y}|\theta \sim N(\tilde{x}'\beta, \sigma^2)$ is conditionally independent from the observed data $y = (y_1, \dots, y_n)'$. Notably, (3) defines a distribution over the possible realizations of the response variable at a particular covariate value and conditional on all observed data. The posterior predictive distribution is fundamental for point prediction, but also provides uncertainty quantification for predictive functionals, such as $\tilde{y}$ itself and measures of predictive loss (see Section 3.4).

More generally, consider the task of prediction at covariates $\{\tilde{x}_i\}_{i=1}^{\tilde{n}}$, which may be distinct from the observed covariates $\{x_i\}_{i=1}^{n}$. The choice of $\{\tilde{x}_i\}_{i=1}^{\tilde{n}}$ can target predictions for specific designs, subject characteristics, or subpopulations of interest. The corresponding predictive variables are denoted $\{\tilde{y}_i\}_{i=1}^{\tilde{n}}$, which are the unobserved or future values of the response variable at each predictor $\tilde{x}_i$. The posterior predictive distribution $p(\tilde{y}_i|y)$ is given by (3) with $\tilde{y}_i|\theta \sim N(\tilde{x}_i'\beta, \sigma^2)$, and is not required to be known analytically: samples $\{\tilde{y}_i^s\}_{s=1}^{S} \sim p(\tilde{y}_i|y)$ are sufficient, which usually are obtained by sampling $\theta^s \sim p(\theta|y)$ from the posterior and $\tilde{y}_i^s \sim p(\tilde{y}_i|\theta^s)$ from the aforementioned Gaussian distribution. Regardless of the choice of covariate values $\{\tilde{x}_i\}_{i=1}^{\tilde{n}}$, the posterior predictive distribution conditions on *all* available data.

For simultaneous evaluation of both predictive accuracy and model complexity, we consider the following loss function:

$$\mathcal{L}_\lambda(\{\tilde{y}_i\}_{i=1}^{\tilde{n}}, \delta) = \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} ||\tilde{y}_i - \tilde{x}_i'\delta||^2 + \lambda \mathcal{P}(\delta) \tag{4}$$

where δ is a p-dimensional vector of linear coefficients and $\mathcal{P}$ is a complexity penalty. The complexity parameter $\lambda \geq 0$ controls the tradeoff between accuracy and complexity, with larger values of λ encouraging simpler models. The squared error loss measures fitness of the linear point prediction $\tilde{x}_i'\delta$ for some value of δ, which is to be determined. Many choices of complexity penalization are available, but for the purpose of variable selection, we consider sparsity-inducing penalties such as ℓ_1-penalization and its variants (see Section 3.3). In this context, a simpler model is one which includes fewer variables, i.e., $\delta_j = 0$ for some or many $j \in \{1, \dots, p\}$, and the goal is to maximize predictive performance in the presence of sparsity.

To obtain optimal coefficients δ for a given complexity level λ, we minimize the *posterior predictive expectation* of (4), which averages over the uncertainty implicit in each predictive variable $\tilde{y}_i$ conditional on the observed data y:

$$\hat{\delta}_\lambda = \arg\min_\delta \mathbb{E}_{[\tilde{y}|y]} \mathcal{L}_\lambda(\{\tilde{y}_i\}_{i=1}^{\tilde{n}}, \delta) \tag{5}$$

where the expectation is taken with respect to (3) for each $p(\tilde{y}_i|y)$. By varying λ, the solution in (5) generates a path of optimal coefficients at varying complexity levels. From (3), it follows that $\mathbb{E}_{[\tilde{y}|y]}\mathcal{L}_\lambda(\{\tilde{y}_i\}_{i=1}^{\tilde{n}}, \delta) = \mathbb{E}_{[\theta|y]}\mathbb{E}_{[\tilde{y}|\theta]}\mathcal{L}_\lambda(\{\tilde{y}_i\}_{i=1}^{\tilde{n}}, \delta)$. We evaluate these expectations sequentially. The penalty term is fixed and remains constant throughout, so for notational simplicity we consider the unpenalized case, $\lambda = 0$. Observing $\mathbb{E}_{\tilde{y}|\theta}\tilde{y}_i = \tilde{x}_i'\beta$, we simplify as follows:

$$\begin{aligned}\mathbb{E}_{[\tilde{y}|\theta]}\mathcal{L}_0(\{\tilde{y}_i\}_{i=1}^{\tilde{n}}, \delta) &= \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} \mathbb{E}_{[\tilde{y}|\theta]}||\tilde{y}_i - \tilde{x}_i'\delta||^2 = \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} \mathbb{E}_{[\tilde{y}|\theta]}||(\tilde{y}_i - \tilde{x}_i'\beta) + (\tilde{x}_i'\beta - \tilde{x}_i'\delta)||^2 \\ &= \sigma^2 + \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} ||\tilde{x}_i'\beta - \tilde{x}_i'\delta||^2\end{aligned}$$

since $\sigma^2 = \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} \mathbb{E}_{[\tilde{y}|\theta]}||\tilde{y}_i - \tilde{x}_i'\beta||^2$. Letting $\hat{\beta} = \mathbb{E}_{\theta|y}(\beta)$ and $\widehat{\sigma^2} = \mathbb{E}_{\theta|y}(\sigma^2)$ be the posterior expectations, the outer expectation of the loss function simplifies similarly:

$$\begin{aligned}\mathbb{E}_{[\theta|y]}\left\{\sigma^2 + \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} ||\tilde{x}_i'\beta - \tilde{x}_i'\delta||^2\right\} &= \widehat{\sigma^2} + \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} \mathbb{E}_{[\theta|y]}||(\tilde{x}_i'\beta - \tilde{x}_i'\hat{\beta}) + (\tilde{x}_i'\hat{\beta} - \tilde{x}_i'\delta)||^2 \\ &= \widehat{\sigma^2} + \frac{1}{\tilde{n}}\text{trace}\left(\tilde{X}'\tilde{X}\widehat{\Sigma}_\beta\right) + \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} ||\tilde{x}_i'\hat{\beta} - \tilde{x}_i'\delta||^2\end{aligned}$$

where $\tilde{X} = (\tilde{x}_1, \dots, \tilde{x}_{\tilde{n}})'$ is the specified design matrix and $\widehat{\Sigma}_\beta = \text{Cov}_{[\theta|y]}(\beta)$ is the posterior covariance of β. Conclusively, we omit additive constants that do not depend on δ to identify an equivalent expression for the optimal parameters from (5):

$$\hat{\delta}_\lambda = \arg\min_\delta \left\{\frac{1}{\tilde{n}}||\hat{y} - \tilde{X}\delta||^2 + \lambda\mathcal{P}(\delta)\right\} \tag{6}$$

where $\hat{y} = \mathbb{E}_{[\tilde{y}|y]}\tilde{y} = \tilde{X}\hat{\beta}$ is the point prediction at $\tilde{X}$ under the linear model (1)-(2). This expectation is easily and quickly computable using posterior draws of β, e.g., $\hat{y} \approx \tilde{X}\bar{\beta}$ for $\bar{\beta} = S^{-1}\sum_{s=1}^S \beta^s$ with $\beta^s \sim p(\beta|y)$.

The representation in (6) is valuable: the decision analysis posited in (5) is solved as a penalized least squares problem. For many choices of penalties $\mathcal{P}$, efficient algorithms exist for solving (6); all that is required is the pseudo-response variable $\hat{y} = \tilde{X}\hat{\beta}$ and the accompanying design matrix $\tilde{X}$. By varying λ, we obtain solutions to (5) that vary in complexity. When $\mathcal{P}$ is a sparsity-inducing penalty, we obtain subsets of variables that are optimized for the point predictions under the full model. For example, the ℓ_1-penalty, $\mathcal{P}(\delta) = \sum_{j=1}^p |\delta_j|$, induces a lasso regression problem[33] with $\hat{y}$ as data, so the parameters $\hat{\delta}_\lambda$ are computable using standard software such as the `glmnet`[34] package in `R`.

There is a fundamental distinction between the linear coefficients β and δ. In particular, β is a parameter of the Bayesian model and therefore is endowed with a prior distribution. By comparison, δ is any p-dimensional vector, and in that sense is entirely disconnected from the Bayesian model. The operation in (5) optimizes δ based on the model and the loss function, and the solution (6) is an expected functional under the posterior distribution. Hence, it is not coherent to place a prior on δ or λ: these define the actions and loss functions, respectively, in the decision analysis but they are not parameters of the model. While the prior on β may record our preference for sparse linear coefficients, this does not guarantee sparsity of any particular posterior functional or summary. For illustration, consider the unpenalized case with $\lambda = 0$. Using any design matrix $\tilde{X}$, the (possibly nonunique) solution to (6) is simply $\hat{\delta}_{\lambda=0} = \hat{\beta}$, which is the posterior expectation of the regression coefficients in (1). However, this quantity is not sparse even in the case of sparsity-inducing priors and therefore cannot provide variable selection. By allowing $\lambda > 0$, we admit linear coefficients that jointly optimize for prediction and sparsity.

The decision analysis approach to variable selection has historical roots[35] with modern development.[36,37] Related approaches, often termed *posterior summarization*, have been developed for multiple linear regression and logistic regression,[36] nonlinear regressions,[38] time-varying parameter models,[39] functional regression,[40] graphical models,[41] and seeming unrelated regressions.[42] However, existing methods face several important limitations. First, they all rely on in-sample quantities for evaluation. Out-of-sample metrics are more suitable for predictive evaluation, which is essential for model comparison and selection. In-sample metrics have a greater risk of overfitting, and therefore may favor models that are unnecessarily complex. Second, these methods provide limited guidelines for model determination, and often use in-sample graphical summaries. It is important to identify both (i) the *best* model for predictive accuracy and (ii) the set of models that are *close enough* to the best model in predictive performance. Leaving precise definitions to the next section, we emphasize that these notions fundamentally rely on evaluations of predictive performance—which is most appropriately determined out-of-sample. Lastly, these approaches commonly adopt an adaptive penalty akin to the adaptive lasso[43] in place of the ℓ_1-penalty. However, these existing adaptive penalization schemes are not coherent within the decision analysis framework of (4) and (5). This issue is addressed and resolved in the next section.

3.3 | Generalized penalties and the Bayesian adaptive lasso

The task of variable selection requires a sparsity-inducing penalty for $\mathcal{P}$. Equally important, the optimal coefficients are only available upon solving the penalized least squares problem (6). Mutual consideration of sparsity and computational scalability suggests ℓ_1-penalization, for which the penalized least square problem is equivalent to the lasso[33] using $\hat{y}$ as data. However, the ℓ_1-penalty is known to introduce substantial bias for large coefficients, and cannot simultaneously provide correct variable selection and optimal estimation in a frequentist context.[43] Adaptive penalization methods[44,43] attempt to correct the suboptimality of the ℓ_1-penalty and have demonstrated wide success in practice. These approaches are distinct from the decision analysis framework of (4)-(5): the response variable in the penalized least squares problem (6) is $\hat{y}$ and not the original data y. More generally, classical adaptive penalization methods do not operate in the same Bayesian framework of (1)-(2), and therefore require additional tools to provide uncertainty quantification.

In the context of decision analysis for Bayesian variable selection, several authors have considered adaptive penalties for $\mathcal{P}$.[36,37,38,39,40] These approaches adopt a common form which mimics the adaptive lasso[43]:

$$\mathcal{P}^0(\delta) = \sum_{j=1}^{p} \hat{\omega}_j^0 |\delta_j|, \quad \hat{\omega}_j^0 = |\hat{\beta}_j|^{-\gamma} \tag{7}$$

where $\hat{\beta} = \mathbb{E}_{\theta|y}(\beta)$ and $\gamma > 0$, typically with $\gamma = 1$[36,38,40] or $\gamma = 2$.[37,39] The adaptive penalty in (7) does not introduce any additional complexity for computation relative to the ℓ_1-penalty. However, this approach is not coherent within the decision analysis framework of (4)-(5): the penalty function $\mathcal{P}$ is treated as known in (4), while (7) requires an estimate of a posterior quantity. Implicitly, this approach ignores the uncertainty in unknown model parameters.

We propose fully Bayesian model-based penalization, which allows the regularization behavior and notion of complexity in the predictive optimization problem to be informed by the model within a coherent decision analysis framework. Specifically, we generalize the loss function (4) as follows:

$$\mathcal{L}_\lambda(\{\tilde{y}_i\}_{i=1}^{\tilde{n}}, \delta, \theta) = \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} ||\tilde{y}_i - \tilde{x}_i'\delta||^2 + \lambda \mathcal{P}(\delta, \theta) \tag{8}$$

where the penalty is now permitted to depend on model parameters θ. Within this framework, we introduce a *Bayesian adaptive lasso penalty*:

$$\mathcal{P}(\delta, \theta) = \sum_{j=1}^{p} \omega_j |\delta_j|, \quad \omega_j = |\beta_j|^{-\gamma} \tag{9}$$

where $\gamma > 0$. Notably, (9) depends on the regression coefficients β, which are unknown and random quantities.

To extract optimal coefficients from the generalized loss function, we now integrate over two sources of uncertainty: (i) the predictive variables $\tilde{y}_i$ and (ii) the model parameters θ, in both cases conditional on the observed data y. The optimal coefficients are identified by minimizing the iterated expected loss:

$$\hat{\delta}_\lambda = \arg\min_{\delta} \mathbb{E}_{[\theta|y]} \left\{ \mathbb{E}_{[\tilde{y}|y,\theta]} \mathcal{L}_\lambda(\{\tilde{y}_i\}_{i=1}^{\tilde{n}}, \delta, \theta) \right\}. \tag{10}$$

Under the mild condition that $\mathbb{E}_{[\theta|y]}|\mathcal{P}(\delta,\theta)| < \infty$ for some δ, the arguments in Section 3.2 imply that (10) is equivalent to

$$\hat{\delta}_\lambda = \arg\min_\delta \left\{ \frac{1}{\tilde{n}}||\hat{y} - \tilde{X}\delta||^2 + \lambda\overline{\mathcal{P}(\delta)} \right\} \tag{11}$$

where $\overline{\mathcal{P}(\delta)} = \mathbb{E}_{[\theta|y]}\mathcal{P}(\delta,\theta)$ is the posterior expectation under (1)-(2). For the Bayesian adaptive lasso in (9), the penalty required to solve (11) is

$$\overline{\mathcal{P}(\delta)} = \sum_{j=1}^{p} \hat{\omega}_j|\delta_j|, \quad \hat{\omega}_j = \mathbb{E}_{[\theta|y]}\left(|\beta_j|^{-\gamma}\right). \tag{12}$$

The necessary weights are easily estimable using posterior simulations: $\hat{\omega}_j \approx S^{-1}\sum_{s=1}^{S}|\beta_j^s|^{-\gamma}$ with $\beta^s \sim p(\beta|y)$. Therefore, the optimal coefficients in (10) are readily available—again using standard software such as the `glmnet`[34] package in `R`—with minimal additional costs relative to (non-adaptive) ℓ_1-penalization.

The adaptive weights in (12) are distinct from both the adaptive lasso[43] and previous approaches for sparse decision analysis.[36,37,38,39,40] In the adaptive lasso, the estimator for the weights is $|\hat{\beta}_j^{ols}|^{-\gamma}$ where $\hat{\beta}_j^{ols}$ is the ordinary least squares estimate of β_j in (1). Notably, $\hat{\beta}_j^{ols}$ is only well-defined for $p < n$; by comparison, (12) is valid for $p > n$ for many choices of shrinkage priors in (2). More specifically, the adaptive lasso requires a consistent estimate of β_j in (1) in order to obtain both selection consistency and an optimal estimation rate. From the continuous mapping theorem, this is equivalent to consistent estimation of the weight $\omega_j = |\beta_j|^{-\gamma}$, so in that context there is no need to distinguish between estimation of β_j and ω_j. However, from a Bayesian perspective, the order of estimation matters due to the presence of (posterior) expectations. Jensen's inequality implies that the proposed weights are larger than those in $\mathcal{P}^0$: $\hat{\omega}_j \geq \hat{\omega}_j^0$. The discrepancy between the weights is greatest when $|\hat{\beta}_j|$ is small; when $|\hat{\beta}_j|$ is large, the weights are effectively indistinguishable. Therefore, the proposed penalty is more aggressive in weighting small coefficient estimates toward zero without introducing additional bias for the large coefficient estimates. Nonetheless, the effects of small sampled values of $|\beta_j^s|$ on $\hat{\omega}_j$ can be mitigated by decreasing γ. More broadly, the proposed approach is coherent within a decision analysis framework: the loss function (8) depends on predictive and posterior variables $\tilde{y}_i$ and θ, respectively, which are integrated out using the posterior distribution under model (1)-(2).

3.4 | Model determination from predictive performance

The decision analysis elicits optimal coefficients $\hat{\delta}_\lambda$ in (10) for each value of the complexity parameter, λ. It is informative to evaluate and compare these coefficients at various levels of complexity. The goal is determine (i) which λ delivers the *best* predictive performance and (ii) which complexity levels λ are *close enough* in predictive accuracy relative to the best model. Criterion (ii) encapsulates the core tradeoff between accuracy and simplicity: although the best model might be complex, there may exist one or more simpler models that achieve nearly the same performance. In the context of variable selection, λ controls the tradeoff between accuracy and sparsity: for an ℓ_1-penalty or the Bayesian adaptive lasso (9), $\lambda = 0$ produces a solution with maximal complexity, so all variables are included, while $\lambda \to \infty$ returns the null model. Alternative specifications of $\mathcal{P}$ are readily available, such as stepwise regression with forward selection or backward elimination, for which λ can be defined to index the model size. However, the resulting solutions $\hat{\delta}_\lambda$ are not continuous in λ, which suggests a lack of stability and robustness.[45] Nonetheless, the subsequent methodology remains valid for any choice of penalty as long as (6) is numerically solvable.

Our approach for evaluating predictive performance follows two core principles, which are similarly emphasized by Kowal[46]: (i) predictions should be evaluated *out-of-sample* and (ii) evaluations should be accompanied by *full uncertainty quantification*. Naturally, out-of-sample predictive evaluations are strongly preferred as more reliable assessments of the predictive capability of a model. In-sample metrics tend to favor more complex models, which simultaneously encourages overfitting and undermines the goal of producing simple or sparse solutions. Yet fundamentally, there is nonignorable uncertainty in the evaluation process: predictive performance is inherently random and will vary across different testing datasets, even for the same values of the predictors. This uncertainty is extremely valuable: it provides strength of evidence for a model and introduces a notion of equivalence among models with similar predictive performance. Importantly, the uncertainty quantification is provided automatically from model (1)-(2) and does not require additional assumptions.

Among frequentist methods, it is common to select the complexity parameter using cross-validation. It has been shown that accounting for the uncertainty in model evaluation can improve selection and reduce bias in prediction.[47,48] The one-standard-error rule[48] is a popular choice for model selection: instead of selecting the best model, this rule selects the simplest model within one standard error of the best model. Similarly, Lei[47] proposes a K-fold cross-validation procedure coupled with a multiplier bootstrap to construct a confidence set of acceptable models. These techniques are successful because they account

for nonignorable uncertainty in the evaluation process, and then leverage that information for model selection. Our approach is conceptually adjacent to these methods, yet fundamentally different due to the presence of the Bayesian model (1)-(2). Bayesian analysis attributes uncertainty to a probability distribution. In the context of predictive evaluation, that distribution is the posterior predictive distribution. We use this distribution to construct interpretable metrics of predictive performance accompanied by full posterior uncertainty quantification.

Note that one approach for decision analytic Bayesian variable selection, called signal adaptive variable selection (SAVS),[37] claims to avoid the issue of selecting λ. However, SAVS uses the optimization criterion (6) with the penalty (7), but simply sets $\lambda = 1$. This does not obviate the need to select λ: it is just one possible choice for the sparsity level. Furthermore, this approach fails to provide any (out-of-sample) comparative evaluation for models of differing sparsity levels, and consequently cannot provide the necessary uncertainty quantification for predictive performance among models of varying complexity.

In conjunction with the concurrent work by Kowal,[46] we introduce two complementary notions of proximity in predictive performance, which are made precise below. First, let $\eta \geq 0$ define a *margin* for acceptable performance: any model with predictive performance within η of the best model is considered acceptable. The margin η is selected by the scientific investigator and varies according to the need to achieve a certain level of accuracy, including $\eta = 0$, and may be expressed relative to the error variance σ^2 in (1). Second, note that predictive performance is inherently random: model accuracy will vary across different testing datasets, even for the same values of the predictors. In particular, a simplified (or sparse) model may not perform within η of the best model for a particular testing dataset, but nonetheless may achieve this margin with some (possibly small) probability. We define $\varepsilon \in [0, 1]$ to be a *probability level* such that, according to the predictive distribution (3) under the linear model, there is *at least* ε probability that the simpler model matches or exceeds the predictive ability of the best model (within a margin of η). *Acceptable models* must satisfy these criteria, which depend on both the margin η and the probability level ε.

For concreteness, we proceed to measure accuracy using mean squared error (MSE), although the subsequent results are generalizable for other performance metrics. The *in-sample* MSE is defined by

$$\text{MSE}_\lambda^{in} = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\hat{\delta}_\lambda)^2 \tag{13}$$

which evaluates the point prediction parameters $\hat{\delta}_\lambda$ at the observed data $\{y_i, x_i\}_{i=1}^n$. For an analogous *predictive* quantity, consider the MSE evaluated at $\tilde{y}_i \sim p(\tilde{y}_i|y)$ in (3):

$$\widetilde{\text{MSE}}_\lambda^{in} = \frac{1}{n}\sum_{i=1}^{n}(\tilde{y}_i - x_i'\hat{\delta}_\lambda)^2 \tag{14}$$

which is the *predictive* (in-sample) MSE. Notably, (14) incorporates the distribution of future or unobserved values $\tilde{y}_i$ at covariate value x_i and conditional on all observed data. Since $\widetilde{\text{MSE}}_\lambda^{in}$ is a predictive quantity, it is accompanied by posterior uncertainty quantification under model (1)-(2). However, $\widetilde{\text{MSE}}_\lambda^{in}$ is an in-sample quantity: the distribution is conditional on all observed data $\{y_i\}_{i=1}^n$, and therefore incorporates the information in the *observed* value of y_i at each x_i. By comparison, out-of-sample prediction does not have the benefit of observing each y_i, so there is inherently greater uncertainty in the corresponding predictive distribution of $\tilde{y}_i$. It is the latter setting that is more representative of practical prediction problems.

To construct out-of-sample MSEs and predictive MSEs, we adjust the key components of (13) and (14)—namely, the optimal coefficients $\hat{\delta}_\lambda$ and the predictive variables $\tilde{y}_i$—to condition only on a subset of the data, the *training data*, and evaluate on the remaining (now out-of-sample) *testing data*. Repeating this procedure for K distinct and randomly-selected training/testing splits reduces the sensitivity to any particular split, and produces a Bayesian K-fold cross-validation procedure. Let $\mathcal{I}_k$ denote the indices of the kth holdout (test) dataset, $k = 1, \ldots, K$, with $\cup_{k=1}^K \mathcal{I}_k = \{1, \ldots, n\}$. The holdout sets $\mathcal{I}_k$ are typically equally-sized, mutually exclusive, and selected randomly from $\{1, \ldots, n\}$. Let $y = \{y^{-\mathcal{I}_k}, y^{\mathcal{I}_k}\}$, so $y^{-\mathcal{I}_k}$ and $y^{\mathcal{I}_k}$ refer to the training and testing datasets, respectively. For each fold $k = 1, \ldots, K$, we define the out-of-sample MSE and predictive MSE as follows:

$$\text{MSE}_\lambda^{out}(k) = \frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k}\left(y_i - x_i'\hat{\delta}_\lambda^{-\mathcal{I}_k}\right)^2, \quad \widetilde{\text{MSE}}_\lambda^{out}(k) = \frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k}\left(\tilde{y}_i^{-\mathcal{I}_k} - x_i'\hat{\delta}_\lambda^{-\mathcal{I}_k}\right)^2 \tag{15}$$

where $\hat{\delta}_\lambda^{-\mathcal{I}_k}$ minimizes the expected loss (10), but restricted to the training data $y^{-\mathcal{I}_k}$,

$$\hat{\delta}_\lambda^{-\mathcal{I}_k} = \arg\min_\delta \mathbb{E}_{[\theta|y^{-\mathcal{I}_k}]}\left\{\mathbb{E}_{[\tilde{y}|y^{-\mathcal{I}_k},\theta]}\mathcal{L}_\lambda(\tilde{y}^{-\mathcal{I}_k}, \delta, \theta)\right\} \tag{16}$$

and similarly $\tilde{y}_i^{-\mathcal{I}_k} \sim p(\tilde{y}_i|y^{-\mathcal{I}_k})$ is the predictive variable at x_i but conditional only on the training data $y^{-\mathcal{I}_k}$. This *out-of-sample* predictive distribution is fundamentally distinct from the *in-sample* predictive distribution, $p(\tilde{y}_i|y) = p(\tilde{y}_i|y^{-\mathcal{I}_k}, y^{\mathcal{I}_k})$, which inappropriately conditions on the testing data $y^{\mathcal{I}_k}$. The out-of-sample MSEs in (15) better capture the essence of the prediction

task, where the predictand is not available for model-fitting of (1)-(2) or solving (16). Since these quantities may be sensitive to the specific testing set $\mathcal{I}_k$, we average (15) over all K testing sets:

$$\text{MSE}_\lambda^{out} = \frac{1}{K}\sum_{k=1}^{K}\text{MSE}_\lambda^{out}(k), \quad \widetilde{\text{MSE}}_\lambda^{out} = \frac{1}{K}\sum_{k=1}^{K}\widetilde{\text{MSE}}_\lambda^{out}(k) \tag{17}$$

which is advocated in the literature.[47,46] When the holdout sets are mutually exclusive, each observation $i = 1, \ldots, n$ appears exactly once in (17). Computation of (17) requires solving (16) and obtaining the out-of-sample predictive distribution $\tilde{y}_i^{-\mathcal{I}_k} \sim p(\tilde{y}_i | y^{-\mathcal{I}_k})$. Although these quantities may be calculated using brute-force model re-fitting for each $k = 1, \ldots, K$, we provide a substantially simpler and faster approximation in Section 3.5 which completely avoids the need for Bayesian model re-fitting.

Given the out-of-sample MSEs, we are now able to define formally the evaluation and selection procedure. First, we define the *best* model to be the minimizer of out-of-sample MSE:

$$\lambda_{min} = \arg\min_\lambda \text{MSE}_\lambda^{out} \tag{18}$$

which equivalently identifies the complexity parameter λ_{min} corresponding to the smallest K-fold cross-validated MSE. This choice of λ serves as a reference point: we are interested in identifying models that perform nearly as well as λ_{min}, and among those, models with minimal complexity. Next, consider the out-of-sample predictive variable

$$\widetilde{\text{gapMSE}}_\lambda^{out} = \widetilde{\text{MSE}}_\lambda^{out} - \widetilde{\text{MSE}}_{\lambda_{min}}^{out} \tag{19}$$

which is the gap between model λ and the best model λ_{min}. The set of acceptable models[46] is

$$\Lambda_{\eta,\varepsilon} = \left\{\lambda : \mathbb{P}\left(\widetilde{\text{gapMSE}}_\lambda^{out} \leq \eta\sigma^2\right) \geq \varepsilon\right\}. \tag{20}$$

According to (20), any model with at least $\varepsilon \in [0, 1]$ probability of matching the accuracy of the MSE_λ^{out}-optimal model λ_{min} within $\eta\sigma^2$ is considered acceptable. The probability $\mathbb{P}$ is computed with respect to the out-of-sample predictive distributions in (17) under model (1)-(2) and estimated using posterior predictive simulations. The set of acceptable models is nonempty, since $\lambda_{min} \in \Lambda_{\eta,\varepsilon}$ for all $\eta \geq 0$ and $\varepsilon \in [0, 1]$. More useful, (20) gathers the set of models which demonstrate near-optimal performance—some of which may be more simple or sparse than λ_{min}.

The acceptable set of models in (20) is equivalently derived using posterior predictive intervals[46]: $\lambda \in \Lambda_{\eta,\varepsilon}$ if and only if there exists a $(1 - \varepsilon)$ lower posterior predictive credible interval for $\widetilde{\text{gapMSE}}_\lambda^{out}/\sigma^2$ that includes η. Naturally, if the predictive interval for the increase in out-of-sample MSE *excludes* η, such as $\eta = 0$, then this choice of λ is deemed unacceptable and omitted from $\Lambda_{\eta,\varepsilon}$. From this interpretation, (20) is a Bayesian analog of "cross-validating with confidence" (CVC).[47] CVC computes a confidence set of acceptable models by testing the null hypothesis that each λ produces the best predictive risk; if the null is rejected, then λ is excluded from the confidence set. By comparison, the proposed Bayesian approach leverages the posterior predictive distribution—automatically available from the linear model (1)-(2) with no additional assumptions—which precludes the need for computationally-intensive bootstrapping procedures, and incorporates a margin $\eta \geq 0$ to relax the accuracy requirements and permit simpler models.

From the set of acceptable models, we identify the simplest, or most penalized model:

$$\lambda_{\eta,\varepsilon} = \max \Lambda_{\eta,\varepsilon}. \tag{21}$$

Both $\lambda_{\eta,\varepsilon}$ and $\Lambda_{\eta,\varepsilon}$ depend on the complexity penalty $\mathcal{P}$ and the solution path. For a sparsity-inducing penalty $\mathcal{P}$, $\lambda_{\eta,\varepsilon}$ is the smallest model along the λ path that belongs to the set of acceptable models $\Lambda_{\eta,\varepsilon}$. Selection via (21) is a Bayesian analog of the one-standard-error rule[48] which, instead of selecting the *best* model, selects the simplest model that is *close enough* to the best model. The intuition behind (21) is the same, but instead incorporates uncertainty quantification through the posterior predictive distribution.

3.5 | Approximations for fast out-of-sample validation

The construction of the set of acceptable models $\Lambda_{\eta,\varepsilon}$ hinges on the ability to estimate out-of-sample quantities, namely, (16) and (17). Re-fitting model (1)-(2) for each training set $k = 1, \ldots, K$ is impractical and may be computationally infeasible for moderate to large sample sizes n. We design an alternative approximation based on importance sampling that only requires a single model fit of (1)-(2) using the full dataset—which is necessary for standard posterior inference.

To proceed, we begin by simplifying the out-of-sample predictive expected loss in (16). Following the arguments in Section 3.2 - 3.3, it is straightforward to show

$$\hat{\delta}_\lambda^{-\mathcal{I}_k} = \arg\min_\delta \left\{ (n - |\mathcal{I}_k|)^{-1} \sum_{j \notin \mathcal{I}_k} (\hat{y}_j^{-\mathcal{I}_k} - x_j'\delta)^2 + \lambda \overline{\mathcal{P}(\delta)^{-\mathcal{I}_k}} \right\} \tag{22}$$

where $\hat{y}_j^{-\mathcal{I}_k} = \mathbb{E}_{[\tilde{y}|y^{-\mathcal{I}_k}]}\tilde{y}_j = \tilde{x}_j'\hat{\beta}^{-\mathcal{I}_k}$ is the out-of-sample point prediction at $\tilde{x}_j$, $\hat{\beta}^{-\mathcal{I}_k} = \mathbb{E}_{[\theta|y^{-\mathcal{I}_k}]}\beta$ is the posterior expectation of β conditional *only* on the training data $y^{-\mathcal{I}_k}$, and similarly $\overline{\mathcal{P}(\delta)^{-\mathcal{I}_k}} = \mathbb{E}_{[\theta|y^{-\mathcal{I}_k}]}\mathcal{P}(\delta, \theta)$ is the posterior expectation of (9) conditional *only* on $y^{-\mathcal{I}_k}$. Given estimators of $\hat{\beta}^{-\mathcal{I}_k}$ and $\overline{\mathcal{P}(\delta)^{-\mathcal{I}_k}}$, the out-of-sample predictive expected loss (16) is equivalent to a penalized least squares problem (22), which has the same form as the in-sample version (6) or (11). Importantly, (22) is efficiently solved for many choices of complexity penalty $\mathcal{P}$, including the Bayesian adaptive lasso penalty (9).

Building upon the representation in (22), we proceed to approximate the core quantities needed for out-of-sample evaluation: $\hat{\beta}^{-\mathcal{I}_k} = \mathbb{E}_{[\theta|y^{-\mathcal{I}_k}]}\beta$, $\overline{\mathcal{P}(\delta)^{-\mathcal{I}_k}} = \mathbb{E}_{[\theta|y^{-\mathcal{I}_k}]}\mathcal{P}(\delta, \theta)$, and $\tilde{y}_i^{-\mathcal{I}_k} \sim p(\tilde{y}_i|y^{-\mathcal{I}_k})$ for $i \in \mathcal{I}_k$. All approximations are constructed using the same importance sampling framework. For the regression coefficients, we approximate

$$\hat{\beta}^{-\mathcal{I}_k} = \int \beta p(\beta|y^{-\mathcal{I}_k})\, d\beta = \int \beta p(\theta|y^{-\mathcal{I}_k})\, d\theta \approx \sum_{s=1}^S \beta^s w_k^s \tag{23}$$

where $\theta = (\beta, \sigma^2)$ are model parameters, $\{\beta^s\}_{s=1}^S$ are drawn from a proposal distribution, and $\{w_k^s\}_{s=1}^S$ are the importance weights. For the proposal distribution, we select the full posterior, $p(\theta|y) = p(\theta|y^{-\mathcal{I}_k}, y^{\mathcal{I}_k})$, which will produce an accurate approximation for $p(\theta|y^{-\mathcal{I}_k})$ unless the testing points $y^{\mathcal{I}_k}$ are highly influential for the posterior of θ. For this proposal, the importance weights simplify to

$$w_k^s \propto 1/p(y_{\mathcal{I}_k}|\theta^s) = \prod_{i \in \mathcal{I}_k} 1/p(y_i|\theta^s), \quad \theta^s \sim p(\theta|y) \tag{24}$$

under the conditional independence assumption accompanying model (1). The importance weights in (24) constitute an out-of-sample adjustment: relative to (23), the in-sample posterior expectation is typically estimated by $\bar{\beta} = S^{-1}\sum_{s=1}^S \beta^s$.

The approximation for the penalty proceeds similarly: $\overline{\mathcal{P}(\delta)^{-\mathcal{I}_k}} \approx \sum_{s=1}^S \mathcal{P}(\delta, \theta^s) w_k^s$, which further simplifies for the Bayesian adaptive lasso to $\overline{\mathcal{P}(\delta)^{-\mathcal{I}_k}} \approx \sum_{j=1}^p \bar{\omega}_j |\delta_j|$ with $\bar{\omega}_j = \sum_{s=1}^S w_k^s |\beta_j^s|^{-\gamma}$. Importantly, the out-of-sample adjustment from (24) requires only (i) the posterior draws $\theta^s \sim p(\theta|y)$ and (ii) computation of the Gaussian likelihood (1), and therefore incurs minimal additional computational costs. The use of the full posterior distribution as the proposal for an out-of-sample predictive distribution has been employed successfully for Bayesian model selection[49] and evaluating predictive distributions.[50,46]

To obtain draws from the out-of-sample predictive distribution $\tilde{y}_i^{-\mathcal{I}_k} \sim p(\tilde{y}_i|y^{-\mathcal{I}_k})$, we directly build upon the importance weights constructed in (24). The out-of-sample predictive distribution has the same form as the (in-sample) predictive distribution (3), $p(\tilde{y}_i|y^{-\mathcal{I}_k}) = \int p(\tilde{y}_i|\theta)p(\theta|y^{-\mathcal{I}_k})\, d\theta$, so draws from this distribution can be generated by iteratively sampling $\theta^s \sim p(\theta|y^{-\mathcal{I}_k})$ and $\tilde{y}_i^s \sim p(\tilde{y}_i|\theta^s)$. Again using the full posterior $p(\theta|y)$ as the proposal, we generate draws from the out-of-sample predictive distribution $p(\tilde{y}_i|y^{-\mathcal{I}_k})$ using *sampling importance resampling* (SIR) based on the importance weights in (24). Specifically, let $\tilde{S} \ll S$ denote the number of SIR simulations. Given samples $\{\theta^s\}_{s=1}^S$ from the full posterior, we iterate the following algorithm: (i) sample $s^* \in \{1, \ldots, S\}$ with associated probabilities $\{w_k^1, \ldots, w_k^S\}$; (ii) sample $\tilde{y}_i^{s^*} \sim p(\tilde{y}_i|\theta^{s^*})$ from the likelihood (1); (iii) discard s^* and $w_k^{s^*}$ from the set of indices and weights; (iv) repeat for $\tilde{S}$ iterations. The final sample $\{\tilde{y}_i^{s^*}\}_{s^*=1}^{\tilde{S}}$ is distributed $p(\tilde{y}_i|y^{-\mathcal{I}_k})$. Sampling *without replacement* from step (iii) helps mitigate the effects of excessively large weights: when $w_k^{s^*}$ is large, the same value $\tilde{y}_i^{s^*}$ may be sampled many times and can produce a degenerate sample of $\{\tilde{y}_i^{s^*}\}_{s^*=1}^{\tilde{S}}$.

Using importance sampling and SIR, the out-of-sample predictive evaluations in (17)—including uncertainty quantification via the predictive MSE—are rapidly computable using only (i) posterior (predictive) simulations under the linear model (1)-(2), (ii) pointwise likelihood evaluations for the Gaussian distribution (1), and (iii) the solution to the penalized least square problem (22). For completeness, the algorithm is presented in Algorithm 1.

Note that steps 2(d) and 3 are the only components that depend on the complexity level λ. Therefore, the remaining steps represent a one-time cost for use with all approximating models or penalties. Similarly, only step 3 depends on the choice of metric: alternatives to MSE, such as mean absolute error, may be included at this stage without affecting the broader algorithm. For the Bayesian adaptive lasso penalty (9), the optimization in step 2(d) is done using the `glmnet`[34] package in `R`. In the event of large importance weights, the out-of-sample posterior expectations $\hat{\beta}^{-\mathcal{I}_k}$ and $\bar{\omega}_j$ instead may be estimated using the sample means from the SIR draws, i.e., $\hat{\beta}^{-\mathcal{I}_k} \approx \tilde{S}^{-1}\sum_{s^*=1}^{\tilde{S}} \beta^{s^*}$ and $\bar{\omega}_j \approx \tilde{S}^{-1}\sum_{s^*=1}^{\tilde{S}} |\beta_j^{s^*}|^{-\gamma}$. In particular, sampling without replacement eliminates the effect of excessively large importance weights. As a last resort, highly influential test data $y^{\mathcal{I}_k}$ can be evaluated out-of-sample without approximation, e.g., by re-fitting the linear model (1)-(2) on the training data $y^{-\mathcal{I}_k}$.

Algorithm 1 Out-of-sample MSE and predictive MSE.

1. Obtain posterior samples $\{\theta^s\}_{s=1}^S \sim p(\theta|y)$;
2. For each holdout set $k = 1, \ldots, K$:
   (a) Compute the (log) weights, $\log w_k^s \stackrel{c}{=} -\log p(y_{\mathcal{I}_k}|\theta^s) = -\sum_{i\in\mathcal{I}_k} \log p(y_i|\theta^s)$ using the (Gaussian) log-likelihood in (1), where $\stackrel{c}{=}$ denotes equality up to a constant;
   (b) Estimate $\hat{\beta}^{-\mathcal{I}_k} \approx \sum_{s=1}^S \beta^s w_k^s$ and $\bar{\omega}_j = \sum_{s=1}^S w_k^s |\beta_j^s|^{-\gamma}$;
   (c) Set $\hat{y}_j^{-\mathcal{I}_k} = \tilde{x}_j' \hat{\beta}^{-\mathcal{I}_k}$ for $j \notin \mathcal{I}_k$;
   (d) Compute $\hat{\delta}_\lambda^{-\mathcal{I}_k}$ by solving (22) for each λ;
   (e) Sample $\{\tilde{y}_i^{s^*}\}_{s^*=1}^{\tilde{S}}$ from $p(\tilde{y}_i|y^{-\mathcal{I}_k})$ using the SIR algorithm for $i \in \mathcal{I}_k$;
3. Compute MSE_λ^{out} using $\hat{\delta}_\lambda^{-\mathcal{I}_k}$ and $\widetilde{\text{MSE}}_\lambda^{out}$ using $\hat{\delta}_\lambda^{-\mathcal{I}_k}$ and $\{\tilde{y}_i^{s^*}\}_{s^*=1}^{\tilde{S}}$ in (17).

Perhaps most interesting, the proposed analysis is conducted entirely with out-of-sample objectives using in-sample computing. In particular, the evaluation metrics (17) and the set of acceptable models $\Lambda_{\eta,\varepsilon}$—as well as the best model λ_{min} and the simplest acceptable model $\lambda_{\eta,\varepsilon}$—are all constructed from out-of-sample quantities. At the same time, all of the necessary computations are performed using in-sample quantities, namely, the posterior samples under the linear model (1)-(2) coupled with the importance sampling approximations described above. Consequently, the proposed approach reaps the benefits of out-of-sample predictive analysis with minimal additional computational burden relative to the original posterior sampler.

4 | RESULTS

4.1 | Simulation Study

The proposed techniques are evaluated using simulated data. We focus on three key properties: (i) variable and model selection capabilities, (ii) point prediction accuracy of the response variable, and (iii) point estimation accuracy of the regression coefficients. First, we investigate the ability of the acceptable model sets $\Lambda_{\eta,\varepsilon}$ to identify the correct subset of active (nonzero) predictors. Naturally, this performance depends on the choice of margin η and the probability level ε, as well as other features of the simulation design such as the sample size and the strength of the signal. Next, we study the point prediction accuracy of the proposed decision analytic approach. The decision analysis optimizes for point prediction under a sparsity or complexity constraint. Thus, it is important to assess whether there are empirical gains in accuracy relative to the full Bayesian model, among other competitors. The estimation accuracy results are similar to the prediction results and are provided in the Appendix.

The simulated data are constructed to match the general features of the real dataset in Section 4.2, including the dimensionality of the data, the signal-to-noise ratio, and the correlations, discreteness, and interactions among the predictor variables. We construct $p - 5$ main effects and 5 interactions. First, we simulate continuous predictors $x_{i,j}$ from marginal standard normal distributions with persistent correlation among predictors, $\text{Cor}(x_{i,j}, x_{i,j'}) = (0.75)^{|j-j'|}$, and randomly permute the columns. Next, discrete covariates are included by selecting 25% of these $p - 5$ covariates at random and binarizing at zero, i.e., among the selected predictors j we update $x_{i,j} \leftarrow \mathbf{1}\{x_{i,j} \geq 0\}$ for all $i = 1, \ldots, n$. Matching the pre-processing steps in Section 4.2, the continuous predictors are centered and scaled to have sample mean zero and sample standard deviation 0.5, which helps align the continuous and binary variables on similar scales.[51] The 5 interaction effects are then generated via multiplication. In particular, we include 7 signals among the $p-5$ main effects and 2 signals among the 5 interactions. The main effect signals are $\beta_j^* = \pm 2$, while the interactions are smaller, $\beta_j^* = \pm 1.5$, with the signs distributed evenly. Each active interaction corresponds to at least one active main effect (weak hierarchy), while inactive interactions arise from both active and inactive main effects. The intercept $\beta_0^* = 1$ is included and fixed throughout. Letting $y_i^* = x_i'\beta^*$, we generate response variables $y_i = y_i^* + \epsilon_i$ with independent errors $\epsilon_i \sim N(0, \sigma_{true}^2)$, where $\sigma_{true}^2 = \text{var}(y_i^*)/\text{SNR}$ and SNR is the signal-to-noise ratio. We consider the designs $(n, p) \in \{(10{,}000, 100), (1{,}000, 100), (200, 500)\}$ with weak (SNR = 0.5) and strong (SNR = 3) signals. The randomization steps are repeated for each of 500 simulations.

For each simulated dataset, we implement the Bayesian linear regression model (1)-(2) with a horseshoe prior.[23] Samples from the posterior and posterior predictive distributions are obtained using the elliptical slice sampler in the R package `bayeslm`.[30] We save $S = 10{,}000$ MCMC samples $\{\theta^s\}_{s=1}^S \sim p(\theta|y)$ after discarding 5,000 initial iterations as a burnin. Using these posterior samples, we solve (10) for $\hat{\delta}_\lambda$ across a path of λ values using the `glmnet`[34] package in R. The out-of-sample evaluations from Section 3.4 are applied to this path of λ values, specifically using $K = 10$-fold cross-validation and $\tilde{S} = S/2 = 5{,}000$ SIR samples. For any choice of η and ε, we subsequently construct the set of acceptable models $\Lambda_{\eta,\varepsilon}$.

Among competing methods, we include both Bayesian and frequentist alternatives. To represent the full Bayesian linear model, we compute the posterior means of $\hat{\beta}$, which are used in both estimation and prediction, and select variables based on whether the 95% highest posterior density interval for each β_j excludes zero. For frequentist methods, we include the adaptive lasso[43] with ordinary least squares weights; for $p > n$, we use the (non-adaptive) lasso.[33] The tuning parameter is selected using the one-standard-error rule[48] from 10-fold cross-validation and implemented using the `glmnet`[34] package in R. In addition, we include classical or exhaustive subset selection (after pre-screening to the first 30 predictors that enter the adaptive lasso model) using the `leaps`[52] package in R with the final subset determined by Mallow's C_p. Lastly, we include minimax convex penalization (MCP)[53] implemented using `ncvreg`[54] with the tuning parameter selected using 10-fold cross-validation.

Ideally, the set of acceptable models should include the *true* model, i.e., the subset of active variables $\{j : \beta_j^* \neq 0\}$. This task necessitates joint selection across all predictors, and is highly demanding: incorrect labeling of any one variable as active or inactive leads to an incorrect model. Since the set of acceptable models is nested by (ε, η)—increasing $\varepsilon \in [0, 1]$ or decreasing $\eta \geq 0$ shrinks the acceptable set—it is of interest to determine the largest values of ε and smallest values of η for which the true model is acceptable. For any model indexed by λ, consider the following constituent of (20):

$$\varepsilon_{max}(\lambda) = \mathbb{P}\big(\widetilde{\text{gapMSE}}_\lambda^{out} \leq \eta\sigma^2\big). \tag{25}$$

Equivalently, $\varepsilon_{max}(\lambda)$ is the maximum probability level ε for which λ belongs to the set of acceptable models with margin η. By design, for any smaller probability level $\varepsilon' \leq \varepsilon_{max}(\lambda)$—which provides more leniency for inaccuracy—λ remains acceptable, $\lambda \in \Lambda_{\eta,\varepsilon'}$. We are interested in $\varepsilon_{max}(\lambda^*)$, where λ^* denotes the true model along the λ path. If the true model is not along the λ path, we set $\varepsilon_{max}(\lambda^*) = 0$. This definition is important: it interrogates the ability of the entire procedure—including the elicitation of the λ path—to capture the true model, rather than merely checking whether the true model belongs to the acceptable set. Consequently, the true model must be both *discovered* and *evaluated correctly* to achieve $\varepsilon_{max}(\lambda^*) > 0$.

For each simulated dataset, we compute $\varepsilon_{max}(\lambda^*)$ for a grid of η values. The results are presented in Figure 1 for SNR $= 3$. For the larger sample sizes $(n, p) = (10{,}000, 100)$ and $(n, p) = (1{,}000, 100)$, there is rapid convergence to $\varepsilon_{max}(\lambda^*) = 1$, while $\varepsilon = 0.05$ is capable of capturing the true model often even without a margin $\eta = 0$. The high-dimensional case $(n, p) = (200, 500)$ is more challenging: convergence to $\varepsilon_{max}(\lambda^*) = 1$ does not occur for a majority of simulations. The reason is that the λ path in this $p > n$ setting often does not include the true model, and for those simulations $\varepsilon_{max}(\lambda^*) = 0$. As a result, increasing the margin η cannot admit the true model in the set of acceptable models. Importantly, these results suggest that the set of acceptable models produced by a small probability level ε, such as 0.05, and a small margin, including $\eta = 0$, can adequately capture the true model. However, the quality of the λ path of models is a limiting factor.

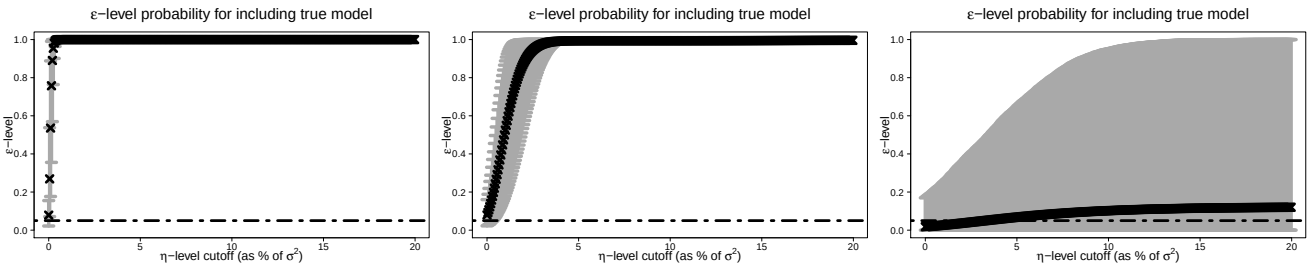

FIGURE 1 The maximum ε-level at which the true model is included for each η-margin (as percent of σ^2). **Left to right:** $(n, p) \in \{(10{,}000, 100), (1{,}000, 100), (200, 500)\}$ for SNR $= 3$. The horizontal line indicates $\varepsilon = 0.05$. The points (x-marks) are sample means and the gray bars are 90% intervals of $\varepsilon_{max}(\lambda^*)$ computed across 500 simulations.

By design, the proposed decision analysis produces coefficient estimates that optimize for point prediction accuracy. Therefore, it is relevant and interesting to evaluate predictions of y_i^*. Informed by Figure 1, we consider the smallest acceptable model $\lambda_{\eta,\varepsilon}$ with $\eta = 0$ and $\varepsilon = 0.05$, and compare the predictions and estimates to the aforementioned competitors. In addition to root mean squared errors (RMSEs) for y_i^*, we compute the median size of the selected subsets across simulations for each method. The results are displayed in Figure 2 across various simulation designs. Most notably, the proposed out-of-sample approach is highly competitive and offers significant improvements relative to the full Bayesian model (i.e., the posterior mean), classical subset selection, and the adaptive lasso in nearly all cases. The distinction between the in-sample and out-of-sample version is most noticeable when the SNR is small, which favors the out-of-sample approach. The main competitor is MCP, which offers predictive accuracy yet tends to select too many variables. The most challenging setting is $p > n$ with low SNR: since $\lambda_{\eta,\varepsilon}$ selects as few variables as needed to satisfy (20), the weak signal results in overly sparse models. Decreasing λ includes additional variables and, when $\lambda = 0$, returns the posterior mean $\hat{\beta}$—which is the best predictor in this setting. Our analysis in Section 4.2 considers the λ path from this broader perspective (see Figure 3).

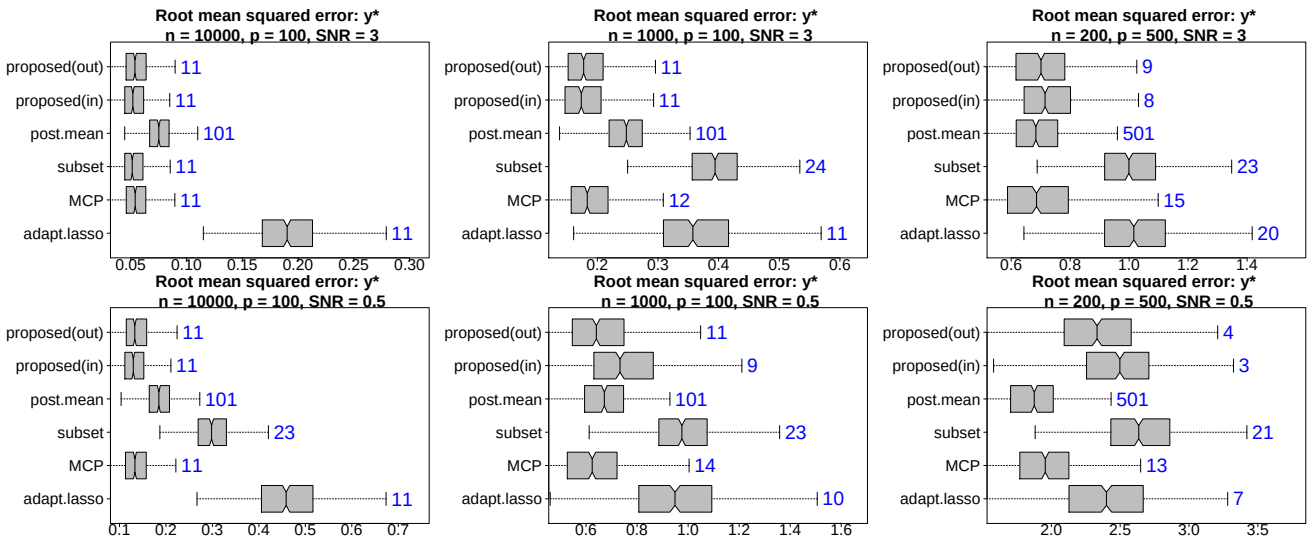

FIGURE 2 Root mean squared errors (boxplots) for $y_i^* = x_i'\beta^*$ with median subset sizes (annotations) for high (**top**) and low (**bottom**) SNR with varying n, p. The proposed out-of-sample decision analytic approach provides highly competitive point predictions with fewer covariates than competing methods, including large improvements over the full Bayesian model and frequentist alternatives. Including the intercept, the true model size is 11.

For completeness, we also assess *marginal* selection properties using true positive rates and true negative rates averaged across all predictors and simulations. The results are in Table 1. In general, the out-of-sample approach offers competitive performance that is similar to the adaptive lasso and MCP, but far superior to classical subset selection which tends to select too many variables. Most notably, the primary Bayesian competitor—selection based on HPD intervals—is excessively conservative and fails to identify many of the active variables. From a Bayesian perspective, joint variable selection using decision analysis offers clear improvements relative to this popular marginal selection technique.

The Appendix includes RMSEs for β^* (Figure B.1) and a sensitivity analysis on the choice of prior for β. Specifically, Figure B.2 replicates the results in Figure 2 under the Bayesian lasso prior[22] instead of the horseshoe prior. The well-known deficiencies of the Bayesian lasso[55] propagate to the decision analysis, yet the proposed approach maintains large advantages relative to the full Bayesian model and some frequentist competitors.

4.2 | Analysis of end-of-grade testing data

To identify joint collections of variables with predictive capabilities for educational outcomes, we apply the proposed methodology to end-of-grade (EOG) reading and mathematics scores. The EOG reading and mathematics scores were analyzed separately

using the same sample of $n = 16{,}839$ students and the same set of $p = 37$ covariates including 12 interactions. The continuous predictors are centered and scaled to have sample mean zero and sample standard deviation 0.5—prior to computing interactions—which better aligns the continuous and binary variables on comparable scales.[51] Posterior inference from the Bayesian linear model (1)-(2) with a horseshoe prior[23] was obtained using the `bayeslm`[30] package in `R`. We save $S = 10{,}000$ MCMC samples $\{\theta^s\}_{s=1}^S \sim p(\theta|y)$ after discarding 5,000 initial iterations as a burnin and thinning to every 10th iteration for more satisfactory effective sample sizes. Trace plots confirm convergence and good mixing of the MCMC samples. As in Section 4.1, we solve (10) for $\hat{\delta}_\lambda$ across a path of λ values using the `R` package `glmnet`,[34] and apply the out-of-sample evaluations from Section 3.4 to this path of λ values using $K = 10$-fold cross-validation and $\tilde{S} = S/2 = 5{,}000$ SIR samples.

To quantify how particular variables or sets of variables contribute to the predictive ability of the model, we present the predictive MSEs for both EOG reading and EOG math scores in Figure 3 with accompanying variable selection details in Table 2. Specifically, we display posterior predictive means and 95% intervals for $\widetilde{\text{gapMSE}}_\lambda^{out}$ from (19), which is the gap in predictive accuracy between each model along the λ path and the best model according to out-of-sample MSE. Figure 3 serves a different purpose from Figure 1: Figure 3 illustrates the predictive ability across sparsity levels λ, while Figure 1 explores the ability to select the "true" model as (η, ε) vary. Although λ and (η, ε) are linked, the selection capabilities (Figure 1) are distinct from relative predictive performance (Figure 3).

For both EOG reading and math scores, λ_{min} is the full model, which is unhelpful for selecting variables. However, the predictive uncertainty quantification provides an avenue for selection: by considering models with a nonnegligble probability of improving upon the best model λ_{min}, we obtain a set *acceptable* models with far fewer variables. The smallest acceptable model $\lambda_{\eta,\varepsilon}$ for $\eta = 0$ and $\varepsilon = 0.05$ has 21 and 22 covariates for EOG reading and EOG math scores, respectively, and in each case ensures a probability of at least 5% that this sparse model is superior to the full model for out-of-sample prediction. Naturally, if we admit a margin $\eta > 0$, we obtain even smaller acceptable models: for η less than 0.4% of the error variance σ^2, the smallest acceptable models for both EOG reading and math scores have fewer than 15 variables. Figure 3 clearly illustrates that small models with fewer than 10 variables are noncompetitive, while the gains beyond around 25 variables are marginal.

To extract interpretable conclusions from Figure 3, we report the selected variables corresponding to these sparse models in Table 2. Variables are reported up to the smallest acceptable model for zero margin $\eta = 0$ and probability level $\varepsilon = 0.05$. We characterize the incremental gains offered by additional sets of predictors—expressed visually by the vertical jumps in Figure 3—by considering the effect of varying the margin η. For any fixed probability level, say $\varepsilon = 0.05$, each model along the λ path will be acceptable at some choice of margin η, which we denote by $\eta_{min}(\lambda) = \min\{\eta : \lambda \in \Lambda_{\eta,0.05}\}$. For any larger $\eta' > \eta_{min}(\lambda)$, this model $\lambda \in \Lambda_{\eta',0.05}$ remains acceptable. These values $\eta_{min}(\lambda)$ are interpretable: they indicate a best case scenario in out-of-sample predictive performance (at the $\varepsilon = 0.05$ probability level) for each model λ relative to the best model λ_{min}. In particular, we are interested in large decreases in $\eta_{min}(\lambda)$ that occur via addition of a small number of variables. These occurrences are indicated by vertical black lines in Figure 3 and demarcated in Table 2.

The results from Table 2 indicate substantial overlap in selected variables between EOG reading and EOG math scores. In addition, many of the basic demographic, social, and environmental variables are confirmed by adaptive lasso selection and (marginal) selection via the 95% posterior credible intervals that exclude zero. These variables include maternal characteristics (education level, ethnicity, age, and marital status), environmental exposures (temperature, $PM_{2.5}$, blood lead level), and social conditions (economically disadvantaged, neighborhood deprivation). The selected environmental exposures and social conditions cover a broad temporal spectrum, including the gestational period, at the time of birth, chronic exposures, and at the time of the exam.

The proposed approach selects additional variables that offer clear and quantifiable improvements in prediction accuracy. Perhaps most notably, we identify evidence for *social and environmental interactions* that predict EOG reading and math scores—even after adjusting for a multitude of other effects. For both EOG reading and math scores, we find that chronic $PM_{2.5}$ interacts with racial isolation at the time of the exam. EOG math scores are additionally predicted by interactions between blood lead level and neighborhood deprivation. Given the conservative nature of selection via credible intervals (see Table 1), it is unsurprising that these variables are undetected by that approach. Yet the evidence for interactions is nontrivial: these effects are present for $\eta_{min}(\lambda)$ values as large as 0.307% of the error variance σ^2 for EOG math scores, and of course persist for $\eta = 0$. While $\eta = 0$ is a strict criterion, it is necessary to provide the assurance that the smallest acceptable model $\lambda_{0,0.05}$ is not definitively suboptimal relative to larger models. Cumulatively, these results suggest that a full *predictive* picture of EOG test scores requires simultaneous consideration of the exam topic; demographic, socioeconomic, and environmental variables; and interactions between environmental exposures and social conditions.

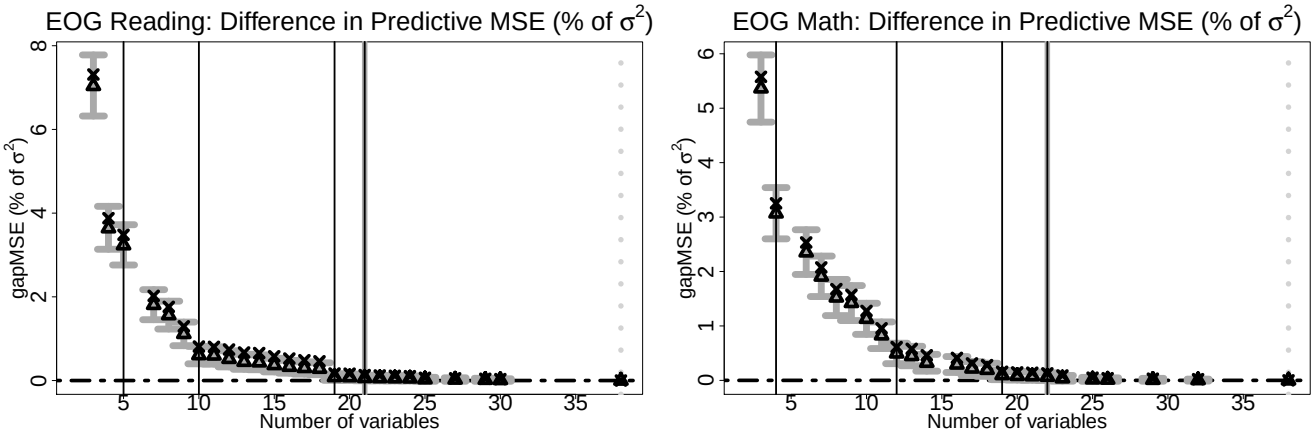

FIGURE 3 The difference in out-of-sample predictive MSEs between each model along the λ path and the best λ_{min} model for EOG reading scores (**left**) and EOG math scores (**right**) as a percent of σ^2. Expectations (triangles) and 95% intervals (gray bars) from the predictive distribution are included along with the out-of-sample MSEs (x-marks). The vertical black lines denote the $\eta_{min}(\lambda)$ values summarized in Table 2. The solid gray line denotes $\lambda_{0,0.05}$ and the dotted gray line denotes λ_{min}.

5 | DISCUSSION

We have introduced a comprehensive and decision-analytic approach for Bayesian linear variable selection, which is designed to advance the state-of-the-art in exposure science. By representing variable selection as a decision problem, we are able to identify *joint* and *optimal* subsets of explanatory variables. The decision problem balances predictive accuracy with model complexity through a penalized loss function, and uniquely incorporates model parameters in the complexity penalty to improve selection and estimation capabilities. Out-of-sample predictive evaluations are conducted for each linear subset of variables, which inherit a posterior predictive distribution from the underlying Bayesian regression model. Using this out-of-sample predictive uncertainty quantification, we identify not only the *best* subset for prediction, but also those subsets that are *nearly* optimal with some nonnegligible probability. The collection of these *acceptable models* offers insights into the covariates needed—which and how many—for sufficiently accurate prediction. Comparisons on simulated datasets demonstrate the excellent variable selection, estimation, and prediction capabilities of the proposed approach.

Using population data from North Carolina, our analysis brings new understanding to the cumulative and interactive impact of social and environmental stressors on the educational growth of children as measured by their 4th grade end-of-grade exam scores in reading and mathematics. The selected variables span a broad temporal spectrum, including pre- and post-natal and throughout the life of the child. This more complete predictive picture of child educational outcomes highlights the critical roles of demographics and other information at birth, social stressors, environmental exposures, and key social-environmental interactions.

The next stage in our methodological development is to extend the decision analysis selection approach for (i) all pairwise interactions and (ii) nonlinear models. While our focus is restricted to interactions between social stressors and environmental exposures, other analyses may require consideration of the complete collection of pairwise interactions.[56,57] The proposed model-based penalization strategy (8) offers a path forward and welcomes hierarchical or interaction-sparsity penalties[58]—as well as the accompanying estimation algorithms—into our decision analysis framework. For the nonlinear case, both the Bayesian linear model (1)-(2) and the formulation of the decision problem (4) will require modifications to incorporate nonlinearity. In particular, the abundance of nonlinear models demands careful consideration of the appropriate model specification, as well as the ability to evaluate the out-of-sample predictive capability of each nonlinear model efficiently. The decision analysis framework—in conjunction with the proposed fast out-of-sample approximation algorithms and accompanying mechanisms to elicit acceptable models—builds upon both Bayesian models and non-Bayesian penalized regression techniques, and consequently provides a promising path to achieve these goals.

DATA ACCESSIBILITY

The dataset [16] (see Appendix) cannot be released due to privacy protections. However, access to the dataset can occur through establishing affiliation with the Children's Environmental Health Initiative (contact cehi@nd.edu).

ACKNOWLEDGEMENTS

We thank the reviewers and editors for comments that greatly improved this manuscript. Research reported in this publication was supported by the National Institute of Environmental Health Sciences of the National Institutes of Health under award number R01ES028819, the Army Research Office under award number W911NF-20-1-0184 (Kowal), and the National Institute on Minority Health and Health Disparities under award R00MD011304 (Bravo). The content, views, and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the National Institutes of Health, the North Carolina Department of Health and Human Services, Division of Public Health, the Army Research Office, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

1. Williams DR, Collins C. Racial residential segregation: A fundamental cause of racial disparities in health. 2001

2. Morello-Frosch R, Shenassa ED. The environmental "Riskscape" and social inequality: Implications for explaining maternal and child health disparities. 2006

3. Suades-Gonzalez E, Gascon M, Guxens M, Sunyer J. Air Pollution and Neuropsychological Development: A Review of the Latest Evidence. *Endocrinology* 2015; 156(10): 3473–3482.

4. Chiu YHM, Hsu HHL, Coull BA, et al. Prenatal Particulate Air Pollution and Neurodevelopment in Urban Children: Examining Sensitive Windows and Sex-specific Associations. *Environment International* 2017; 87: 56–65.

5. Canfield RL, Henderson C.R. J, Cory-Slechta DA, Cox C, Jusko TA, Lanphear BP. Intellectual impairment in children with blood lead concentrations below 10 microg per deciliter. *N.Engl.J.Med.* 2003; 348(16): 1517–1526. doi: 10.1056/NEJMoa022848 [doi];348/16/1517 [pii]

6. Liu J, Liu X, Wang W, et al. Blood lead concentrations and children's behavioral and emotional problems: A cohort study. *JAMA Pediatrics* 2014. doi: 10.1001/jamapediatrics.2014.332

7. Miranda ML, Kim D, Galeano MAO, Paul CJ, Hull AP, Morgan SP. The relationship between early childhood blood lead levels and performance on end-of-grade tests. *Environmental Health Perspectives* 2007. doi: 10.1289/ehp.9994

8. Reuben A, Caspi A, Belsky DW, et al. Association of Childhood Blood Lead Levels With Cognitive Function and Socioeconomic Status at Age 38 Years and With IQ Change and Socioeconomic Mobility Between Childhood and Adulthood. *Journal of the American Medical Association* 2017; 317(12): 1244–1251.

9. Sampson RJ, Sharkey P, Raudenbush SW. Durable effects of concentrated disadvantage on verbal ability among African-American children. *Proceedings of the National Academy of Sciences of the United States of America* 2008. doi: 10.1073/pnas.0710189104

10. Kramer MR, Hogue CR. Is segregation bad for your health?. 2009

11. Brooks-Gunn J. *Neighborhood poverty: Context and consequences for children*. 1. Russell Sage Foundation . 1997.

12. Minh A, Muhajarine N, Janus M, Brownell M, Guhn M. A review of neighborhood effects and early child development: How, where, and for whom, do neighborhoods matter?. *Health and Place* 2017; 46: 155–174.

13. Datta J, Ghosh JK. Asymptotic properties of Bayes risk for the horseshoe prior. *Bayesian Analysis* 2013; 8(1): 111–132.

14. Barbieri MM, Berger JO. Optimal predictive model selection. *The Annals of Statistics* 2004; 32(3): 870–897.

15. Meyer MJ, Coull BA, Versace F, Cinciripini P, Morris JS. Bayesian function-on-function regression for multilevel functional data. *Biometrics* 2015; 71(3): 563–574.

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

17. Bravo MA, Kowal DR, Leong H, et al. Understanding Mixtures in Environmental Exposures through Generalized Additive Models. *working manuscript*.

18. Schools NCP. Understanding North Carolina end-of-grade testing. report, 2004.

19. Anthopolos R, Kaufman JS, Messer LC, Miranda ML. Racial residential segregation and preterm birth: built environment as a mediator.. *Epidemiology* 2014; 25(3): 397–405.

20. Measures of spatial segregation. *Sociological Methodology* 2004. doi: 10.1111/j.0081-1750.2004.00150.x

21. Messer LC, Laraia BA, Kaufman JS, et al. The Development of a Standardized Neighborhood Deprivation Index. *Journal of Urban Health* 2006; 83(6): 1041–1062.

22. Park T, Casella G. The Bayesian Lasso. *Journal of the American Statistical Association* 2008; 103(482): 681–686.

23. Carvalho CM, Polson NG, Scott JG. The horseshoe estimator for sparse signals. *Biometrika* 2010: 465–480.

24. Bhattacharya A, Pati D, Pillai NS, Dunson DB. Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association* 2015; 110(512): 1479–1490.

25. Pas v. dSL, Kleijn BJK, Vaart v. dAW. The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics* 2014; 8(2): 2585–2618.

26. Mitchell TJ, Beauchamp JJ. Bayesian variable selection in linear regression. *Journal of the American Statistical Association* 1988; 83(404): 1023–1032.

27. George EI, McCulloch RE. Approaches for Bayesian variable selection. *Statistica Sinica* 1997: 339–373.

28. Ishwaran H, Rao JS. Spike and slab variable selection: frequentist and Bayesian strategies. *The Annals of Statistics* 2005; 33(2): 730–773.

29. O'Hara RB, Sillanpää MJ. A review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis* 2009; 4(1): 85–117.

30. Hahn PR, He J, Lopes HF. Efficient sampling for Gaussian linear regression with arbitrary priors. *Journal of Computational and Graphical Statistics* 2019; 28(1): 142–154.

31. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2008; 70(5): 849–911.

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

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

34. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* 2010; 33(1): 1–22.

35. Lindley DV, Smith AFM. Bayes estimates for the linear model. *Journal of the Royal Statistical Society. Series B (Methodological)* 1972: 1–41.

36. Hahn PR, Carvalho CM. Decoupling shrinkage and selection in Bayesian linear models: a posterior summary perspective. *Journal of the American Statistical Association* 2015; 110(509): 435–448.

37. Ray P, Bhattacharya A. Signal Adaptive Variable Selector for the Horseshoe Prior. *arXiv preprint arXiv:1810.09004* 2018.

38. Woody S, Carvalho CM, Murray JS. Model interpretation through lower-dimensional posterior summarization. *Journal of Computational and Graphical Statistics* 2020: 1–9.

39. Huber F, Koop G, Onorante L. Inducing sparsity and shrinkage in time-varying parameter models. *Journal of Business & Economic Statistics* 2020(just-accepted): 1–48.

40. Kowal DR, Bourgeois DC. Bayesian Function-on-Scalars Regression for High-Dimensional Data. *Journal of Computational and Graphical Statistics* 2020(just-accepted): 1–26.

41. Bashir A, Carvalho CM, Hahn PR, Jones MB. Post-Processing Posteriors Over Precision Matrices to Produce Sparse Graph Estimates. *Bayesian Analysis* 2019; 14(4): 1075–1090.

42. Puelz D, Hahn PR, Carvalho CM. Variable selection in seemingly unrelated regressions with random predictors. *Bayesian Analysis* 2017; 12(4): 969–989.

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

44. Breiman L. Better subset regression using the nonnegative garrote. *Technometrics* 1995; 37(4): 373–384.

45. Fan J, Lv J. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* 2010; 20(1): 101.

46. Kowal DR. Fast, Optimal, and Targeted Predictions using Parametrized Decision Analysis. *Journal of the American Statistical Association* 2021. doi: 10.1080/01621459.2021.1891926

47. Lei J. Cross-validation with confidence. *Journal of the American Statistical Association* 2019: 1–20.

48. Tibshirani RJ, Tibshirani R. A bias correction for the minimum error rate in cross-validation. *The Annals of Applied Statistics* 2009: 822–829.

49. Gelfand AE, Dey DK, Chang H. Model determination using predictive distributions with implementation via sampling-based methods. *Bayesian Statistics* 1992; 4: 147–167.

50. Vehtari A, Ojanen J, Others . A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys* 2012; 6: 142–228.

51. Gelman A. Scaling regression inputs by dividing by two standard deviations. *Statistics in medicine* 2008; 27(15): 2865–2873.

52. Alan Miller b. o. F. c. bTL. *leaps: Regression Subset Selection*. 2020. R package version 3.1.

53. Zhang CH, others . Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics* 2010; 38(2): 894–942.

54. Breheny P, Huang J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics* 2011; 5(1): 232–253.

55. Polson NG, Scott JG. Shrink globally, act locally: sparse Bayesian regularization and prediction. *Bayesian Statistics* 2010; 9: 501–538.

56. Zhou F, Ren J, Lu X, Ma S, Wu C. Gene-Environment Interaction: A Variable Selection Perspective. *arXiv preprint arXiv:2003.02930* 2020.

57. Ren J, Zhou F, Li X, et al. Semiparametric Bayesian variable selection for gene-environment interactions. *Statistics in medicine* 2020; 39(5): 617–638.

58. Bien J, Taylor J, Tibshirani R. A lasso for hierarchical interactions. *Annals of Statistics* 2013; 41(3): 1111.

$n = 10{,}000, p = 100, \text{SNR} = 0.5$

	adaptive lasso	MCP	subset selection	HPD interval	proposed(in)	proposed(out)
TPR	1.00	1.00	1.00	1.00	1.00	1.00
TNR	1.00	0.99	0.86	1.00	1.00	0.99

$n = 10{,}000, p = 100, \text{SNR} = 3$

	adaptive lasso	MCP	subset selection	HPD interval	proposed(in)	proposed(out)
TPR	1.00	1.00	1.00	1.00	1.00	1.00
TNR	1.00	0.99	1.00	1.00	1.00	0.99

$n = 1{,}000, p = 100, \text{SNR} = 0.5$

	adaptive lasso	MCP	subset selection	HPD interval	proposed(in)	proposed(out)
TPR	0.90	0.92	0.95	0.77	0.81	0.89
TNR	0.99	0.96	0.86	1.00	1.00	0.99

$n = 1{,}000, p = 100, \text{SNR} = 3$

	adaptive lasso	MCP	subset selection	HPD interval	proposed(in)	proposed(out)
TPR	1.00	1.00	1.00	1.00	1.00	1.00
TNR	1.00	0.98	0.86	1.00	1.00	0.99

$n = 200, p = 500, \text{SNR} = 0.5$

	adaptive lasso	MCP	subset selection	HPD interval	proposed(in)	proposed(out)
TPR	0.42	0.50	0.54	0.06	0.24	0.30
TNR	0.99	0.98	0.97	1.00	1.00	1.00

$n = 200, p = 500, \text{SNR} = 3$

	adaptive lasso	MCP	subset selection	HPD interval	proposed(in)	proposed(out)
TPR	0.93	0.85	0.93	0.67	0.77	0.79
TNR	0.98	0.99	0.97	1.00	1.00	1.00

TABLE 1 True positive rates (TPR) and true negative rates (TNR) for synthetic data.

EOG Reading Scores	
Margin $\eta_{min}(\lambda)$	Variable Name
2.76% of σ^2	(Intercept)$^{*+}$ Temp_T2^{*+} mEdu:College^{*+} mRace:NH Black^{*+} EconDisadv^{*+}
0.4% of σ^2	PM25_T1^{*+} mAge^{*+} mEdu:HS^{*+} Male^{*+} NDI_test^{*+}
0.005% of σ^2	Temp_T3^{*+} PM25_chronic BWTpct^{+} WeeksGest^{+} NotMarried^{+} PTB^{+} Smoker Blood_lead^{+} PM25:chronic x RI_test
$\eta = 0$	mRace:Hispanic RI_test

EOG Math Scores	
Margin $\eta_{min}(\lambda)$	Variable Name
2.6% of σ^2	(Intercept)$^{*+}$ mEdu:College^{*+} mRace:NH Black^{*+} EconDisadv^{*+}
0.307% of σ^2	Temp_T2^{*+} Temp_T3^{*+} PM25_T1^{*+} BWTpct^{*+} mEdu:HS^{*+} mRace:Hispanic^{*+} NotMarried^{+} PM25:chronic x RI_test^{*+}
0.013% of σ^2	PM25_T3^{+} mAge^{+} PM25:chronic WeeksGest^{+} Smoker^{+} Blood_lead^{+} NDI_test
$\eta = 0$	PTB RI_birth Blood_lead x NDI_birth

TABLE 2 Variables selected (in order and incrementally) for EOG Reading (**left**) and Math (**right**) scores. The margin $\eta_{min}(\lambda)$ expresses the difference in predictive MSE between the best model λ_{min} and the noted subset of variables (in addition to the preceding variables). As the margin decreases to zero, more variables are required to achieve optimal predictive performance. Variables selected by adaptive lasso ($*$) or credible intervals that exclude zero ($+$) are also denoted. The indicated groupings of $\eta_{min}(\lambda)$ are marked in Figure 3.

□

APPENDIX

A ADDITIONAL DETAILS ON THE DATA

The common restrictions are the following: singletons only (no plural births); no congenital anomalies; maternal age is 15-44; birth weight is at least 400g; gestational age is 24-42 weeks; mother's race/ethnicity is non-Hispanic White, non-Hispanic Black, or Hispanic; birth order is 1-4. One acute $PM_{2.5}$ outlier (50% larger than all other values) was removed.

We estimated $PM_{2.5}$ exposures during pregnancy using ambient monitoring data from the nearest $PM_{2.5}$ monitor within 30km from a mother's residence at time of the child's birth. Air pollution exposure during pregnancy was calculated based on residence, birth date, and length of gestation. Specifically, the time period of exposure for each birth was calculated using the date of birth and the weeks of gestation at delivery, as recorded in the NCDBR. Using 24-hour average ambient $PM_{2.5}$ concentrations, we calculated average exposure for the first (1-13 weeks), second (14-26 weeks weeks), and third (27 weeks-birth) trimester.

We estimated chronic pre-test $PM_{2.5}$ exposure using downscaler-reported $PM_{2.5}$ concentrations at the census tract level for the census tract in which the child resided at time of EOG testing. Chronic pre-test $PM_{2.5}$ exposure was estimated based on the mean 24-hour average $PM_{2.5}$ concentration during the 12 months prior to the test date. The exact test date is not recorded in the

EOG dataset, but the test month (May or June) is available. Thus, a mean of 24-hour average $PM_{2.5}$ concentration was generated for the 12 months prior to the first of May or June for each year of EOG data.

We estimated acute pre-test $PM_{2.5}$ exposure using downscaler-reported $PM_{2.5}$ concentrations at the census tract level for the census tract in which the child resided at time of testing. Acute pre-test $PM_{2.5}$ exposure was estimated based on the daily 24-hour average $PM_{2.5}$ concentration during the 30 days prior to the test date. Because only the test month is recorded in the EOG dataset, the acute exposure estimate is based on the mean 24-hour average $PM_{2.5}$ concentration in the 30 days prior to the first of May or first of June for each year of EOG data.

Since the EOG testing dataset spans multiple years (2010-2012), the reading and math scores were standardized within each year prior to analysis.

Economically disadvantaged students are indicated by participation in the free/reduced price lunch program (binary variable, 1 = participation in the program).

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
Temp_T1	Temperature – Average of gestation weeks 1-13 (trimester 1) – Temperature
Temp_T2	Temperature – Average of gestation weeks 14-26 (trimester 2) – Temperature
Temp_T3	Temperature – Average of gestation weeks 27-delivery (trimester 3) – Temperature
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)
WeeksGest	Weeks Gestation, Clinical estimate
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)
PTB	Pre-Term Birth (<37 wks, Clinical)? (1 = Yes)
Education / EOG test information	
EOG_Reading	Reading scale score for chronologically first EOG test taken (4^{th} grade); centered and scaled by year (2010, 2011, 2012)
EOG_Math	Math scale score for chronologically first EOG test taken (4^{th} grade); centered and scaled by year (2010, 2011, 2012)
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)

B ADDITIONAL SIMULATION RESULTS

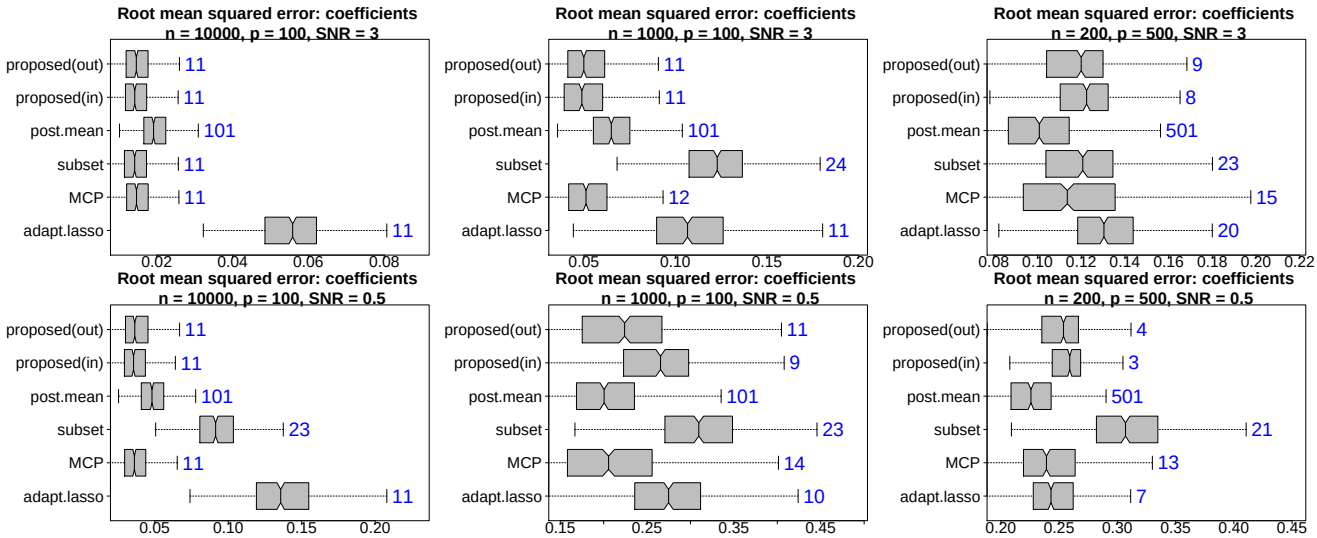

FIGURE B.1 Root mean squared errors (boxplots) for β^* with median subset sizes (annotations) for high (**top**) and low (**bottom**) SNR with varying n, p. The proposed out-of-sample decision analytic approach provides highly competitive point estimates with fewer covariates than competing methods, including large improvements over the full Bayesian model and frequentist alternatives. Including the intercept, the true model size is 11.

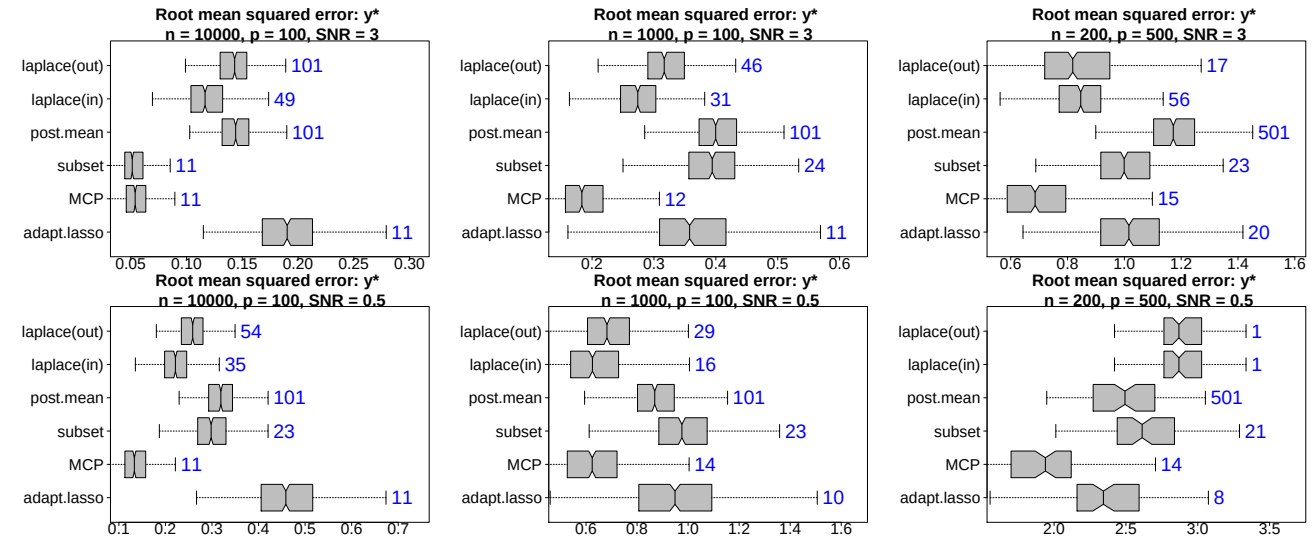

FIGURE B.2 Using a Laplace prior for β: root mean squared errors (boxplots) for $y_i^* = x_i'\beta^*$ with median subset sizes (annotations) for high (**top**) and low (**bottom**) SNR with varying n, p. The decision analytic approach cannot overcome the shortcomings of the Bayesian lasso prior, but nonetheless maintains large advantages relative to the posterior mean in most cases. Results (not shown) for estimation of β^* are similar.