Utrecht University

Department of Methodology and Statistics

Master's degree in Methodology and Statistics for the Behavioural, Biomedical and Social Sciences

Comparing Bayesian Post-estimation Variable Selection Methods: Projection Predictive Variable Selection Versus Stochastic Search Variable Selection

Candidate:
Zhipei (Kim) Wang (5041333)

Supervisors:
Dr. Sara van Erp
Florian van Leeuwen, MSc

FETC approval: 24-2042, 25-1375

May 12, 2025

Abstract

This thesis investigates the comparative performance of two Bayesian variable selection methods: the well-established stochastic search variable selection (SSVS) and the newer projection predictive variable selection (PPVS). Through two simulation studies and empirical applications, we examine their performance in terms of variable selection accuracy, predictive performance, and computational cost. We find that SSVS generally yields better variable selection accuracy, particularly under low correlation and low dimensionality. In contrast, PPVS shows modest gains in predictive performance and interval coverage in higher-dimensional settings, but with consistently lower variable selection accuracy and substantially greater computational burden. Overall, SSVS performs competitively across both variable selection and prediction tasks, particularly given its efficiency and ease of use. Practical limitations of PPVS, such as convergence issues and reliance on heuristic tuning, are also discussed, along with suggestions for future work directions.

Key words: Bayesian, regression, regularization, projection, variable selection, prediction, shrinkage priors, stochastic search variable selection

1 Introduction

Regression models are widely used to investigate the influence of predictors on a response variable and identify which predictors have substantial effects. However, in the era of big data, the number of predictors has grown, and researchers often face high-dimensional problems where the number of predictors is close to or exceeds the number of observations. This "small n, large p" dilemma introduces several challenges. For example, fitting a model can become infeasible when the number of features exceeds the number of observations. Moreover, in such cases, models are prone to overfitting—they rely too heavily on the specific dataset at hand and fail to generalize to new data or to accurately represent the true underlying population characteristics (see Babyak, 2004 for an overview of the problem of overfitting).

Regularization is a widely used technique to address the challenges posed by the "small n, large p" dilemma. Various regularization methods have been developed within both frequentist and Bayesian frameworks. Frequentist approaches typically involve adding a penalty term to the loss function (e.g. Hoerl and Kennard, 2000; Zou and Hastie, 2005), which shrinks parameter estimates and potentially sets them to zero (Tibshirani, 1996), thereby performing variable selection.

Bayesian approaches, by contrast, typically employ shrinkage priors that aim to shrink small effects to zero, while leaving the substantial effects intact. One key advantage of the Bayesian framework is its flexibility: unlike frequentist methods such as the LASSO (Tibshirani, 1996), which can select at most n predictors when $p > n$ due to its formulation as a convex optimization problem (Zou and Hastie, 2005), Bayesian

methods rely on Markov chain Monte Carlo (MCMC) sampling rather than optimization, which allows greater freedom in exploring complex parameter spaces. For a more thorough discussion of the advantages of the Bayesian perspective, see van Erp et al. (2019). Nevertheless, Bayesian approaches face the challenge that shrinkage priors do not always reduce coefficients exactly to zero, making post-estimation variable selection a non-trivial task (Piironen et al., 2020).

In the Bayesian framework, post-estimation variable selection can be broadly divided into two main approaches. The first involves continuous shrinkage priors, which include Bayesian counterparts of frequentist penalization methods such as the ridge (Hsiang, 1975) and LASSO (Park and Casella, 2008). After fitting the model, summaries of the posterior distribution are often used to determine whether a predictor should remain in the model, for example based on thresholds for posterior point summaries or 95% credible intervals. Alternatively, spike-and-slab priors (George & McCulloch, 1993; Mitchell & Beauchamp, 1988) incorporate an indicator variable, assigning each predictor to either the "spike" (negligible effect) or the "slab" (substantial effect). Marginal inclusion probabilities (MIPs) are then computed for each predictor by calculating how many of the MCMC samples of its coefficient were assigned to the slab, with a threshold often involved to decide whether to keep predictors.

Despite the range of existing methods, limitations remain. For example, there is no universal standard for using credibility intervals in real-world analysis, as optimal credibility intervals vary for different data-generating conditions (van Erp et al., 2019). Additionally, as dimensionality grows, the marginal posteriors of relevant features tend to overlap more with zero compared to lower-dimension situations, complicating inference (Piironen et al., 2020). To overcome these difficulties, Piironen et al. (2020) proposed the projection predictive variable selection (PPVS) method, a two-stage approach that first builds a reference model optimized for prediction accuracy while ignoring sparsity, and then identifies a minimal subgroup of predictors that retains comparable predictive performance.

Given that the projection predictive variable selection (PPVS) method is relatively new, its performance in multiple linear regression has not yet been thoroughly examined across a wide range of conditions. While previous work has focused on illustrating the method in small-scale or illustrative scenarios, comprehensive evaluations under varying predictor correlations, high-dimensional settings ($p \geq n$), and different sparsity levels and effect structures remain limited. In this study, we contribute to addressing this gap by evaluating PPVS under a subset of these conditions. To benchmark its performance, we compare it to stochastic search variable selection (SSVS), a well-established and widely used Bayesian variable selection method.

The remainder of this thesis is organized as follows: In the next section, we review the two Bayesian variable selection methods discussed above. Following this, we conduct

simulation studies to compare their performance and discuss the results. In addition, we illustrate the application of these methods using two real-world datasets. Finally, we examine the limitations of the current study and propose potential future research directions.

2 Methods

This section briefly reviews the two Bayesian variable selection methods evaluated in this thesis: stochastic search variable selection (SSVS) and projection predictive variable selection (PPVS). SSVS refers specifically to Bayesian variable selection procedures using a spike-and-slab prior and MCMC sampling, originally introduced under this name by George and McCulloch (1993). In contrast, PPVS is a post-estimation variable selection framework that does not rely on any particular prior. It can be applied to models fitted with a wide range of Bayesian priors. In our implementation, we use the regularized horseshoe prior due to its desirable shrinkage properties and common usage with PPVS.

Consider the standard model for multiple linear regression applied to a response variable Y across n observations and p predictors:

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2) \tag{1}$$

where β_0 is the intercept, β_j represents the regression coefficients, x_{ij} denotes the ith observation of the jth predictor, and ϵ_i is the random error with mean zero and variance σ^2.

In Bayesian regression, a common approach to regularization involves applying sparsifying priors to the regression coefficients β_j to encourage shrinkage towards zero. The relevance of each β_j is assessed by examining its marginal distributions. For SSVS, this is achieved using the spike-and-slab prior.

2.1 Stochastic search variable selection (SSVS)

SSVS is implemented via the spike-and-slab prior, a type of discrete mixture prior (Mitchell and Beauchamp, 1988; George and McCulloch, 1993). This prior combines either a point mass at zero (Mitchell & Beauchamp, 1988) or a normal component very peaked at zero (George & McCulloch, 1993) (the spike), with a more diffuse distribution (the slab). The formulation using a point mass at zero is the default in Bainter et al. (2024), which we used for both our simulation studies and empirical application. This

prior can be written as follows:

$$\begin{aligned} \beta_j \mid \lambda_j, \tau_j &\sim \lambda_j \, \mathrm{N}(0, \tau_j^2) + (1 - \lambda_j) \, \delta_0, \\ \lambda_j \mid \pi &\sim \mathrm{Bernoulli}(\pi), \quad j = 1, \ldots, p. \end{aligned} \tag{2}$$

Here, δ_0 denotes a point mass at zero, and τ_j is set large to represent the slab component. If $\lambda_j = 1$, the coefficient β_j is drawn from the slab; if $\lambda_j = 0$, it is drawn from the spike. The prior probability π (typically 0.5) governs the likelihood that any given predictor is included in the model, which enables binary decisions about inclusion. A hyperprior can also be set for π by assigning it for example a standard uniform distribution (Ishwaran and Rao, 2005).

After estimation, the marginal inclusion probability (MIP) for each predictor is determined from the posterior summaries. The MIP represents the proportion of MCMC samples where $\lambda_j = 1$. Typically, a variable is included in the model if its MIP exceeds 0.5, a threshold shown to be optimal under certain conditions (Barbieri and Berger, 2004).

2.2 Projection predictive variable selection (PPVS)

PPVS involves a two-stage approach starting with fitting a comprehensive reference model with all predictors. This reference model is then projected onto a smaller submodel that maintains comparable predictive performance. The process, though conceptually straightforward, involves multiple complex steps. We follow the workflow proposed by McLatchie et al. (2024), which can be summarized as follows:

1. Fit a comprehensive reference model, and check its performance.
2. Perform a search through the model space to construct the solution path (i.e., a sequence of submodels of increasing size).
3. Use cross-validation to assess predictive performance and avoid overfitting.
4. Choose the submodel size using a utility-based heuristic.

The steps above are discussed in more detail below:

In step 1, a reference model is built with a sparsifying prior using all the predictors (Piironen et al., 2020). A comprehensive review of such priors can be found in van Erp et al. (2019). Once the reference model is fitted, it is important to verify its quality, as the performance of the projected submodel can only be as good as that of the reference model. Since the workflow later involves repeatedly refitting the reference model, strategies can be employed to reduce the computational burden of refitting the reference

model. One approach is supervised principal components (SPCs), a technique that combines preliminary feature screening with dimension reduction (Bair et al., 2006), which was further developed by Piironen and Vehtari (2018) and employed by Pavone et al. (2023) in their study exploring the advantages of incorporating reference models.

In step 2, the solution path is established, detailing which predictors are selected at each step as the process unfolds. The recommended search approach is the Kullback–Leibler (KL) divergence-based forward search, where predictors are added sequentially to minimize the divergence between the predictive distributions of the reference model and the submodel. This process starts from the intercept-only model and adds predictors one at a time based on their impact on projection loss. An alternative is the Lasso regularization that is noted for its computational efficiency but not recommended due to higher variability across different data realizations (McLatchie et al., 2024).

In step 3, to ensure robustness and avoid overfitting, cross-validation is crucial both in evaluating model performance on the solution path and during the model search process. However, given the significant computational demands of cross-validation, employing methods to accelerate the procedure becomes necessary. For example, different projection approaches can be applied when we project the reference model to the final submodel we chose. Piironen et al. (2020) proposed the clustered projection approach, where the posterior draws of the reference model are clustered into many clusters, and a single-point projection is performed for each cluster, which intermediates between the draw-by-draw (Goutis and Robert, 1998; Dupuis and Robert, 2003) technique and the single-point (Tran et al., 2012) technique. Additionally, executing a preliminary run without cross-validating the search process can provide a useful method for determining an upper bound on the submodel size.

Ultimately in step 4, the optimal submodel size is determined based on its predictive closeness to the reference model. While KL divergence can guide this choice by identifying the model with minimal projection loss, Piironen et al. (2020) point out that determining the model complexity using KL divergence is not always easy. They therefore recommend selecting the final submodel based on predictive utility using metrics that are easier to interpret. McLatchie et al. (2024) suggested two primary heuristics for this decision. The first is the standard error (SE) heuristic as discussed in several studies (Catalina et al., 2021; Piironen et al., 2020; Piironen et al., 2023; Piironen and Vehtari, 2017b; Weber et al., 2024), which selects the smallest submodel whose predictive utility is within one standard error of the reference model's utility. The second is the utility difference (Δutility) heuristic, which selects the smallest submodel whose expected log pointwise predictive density (ELPD) is no more than four units lower than that of the reference model (Sivula et al., 2023). After selecting the final submodel size, the projected posterior can be used for inference as in standard MCMC analyses.

While PPVS is compatible with a range of priors, we use the regularized horse-

shoe prior to fit the reference model due to its many desirable properties. Unlike the spike-and-slab, the regularized horseshoe is a continuous shrinkage prior, specifically a continuous analogue of the spike-and-slab with an additional regularization parameter. This enhancement brings it closer in behavior to the spike-and-slab than the original horseshoe prior proposed by Carvalho et al. (2010). For a detailed discussion of the regularized horseshoe and its theoretical motivation, we refer the reader to Piironen and Vehtari (2017b). These two priors represent leading approaches from the families of continuous (one-component) and discrete mixture (two-component) priors, respectively, making their comparison particularly informative.

3 Simulation study setup

To compare the performance of the two Bayesian variable selection methods, two simulation studies were set up. We first give an overview of existing simulation settings where either SSVS or projection methods were used, followed by computational specifications used to run the two methods. Measures of performance used to evaluate and compare the methods were described in detail. Lastly, the simulation conditions and data-generating mechanism used in the two simulation studies are presented. The code to run the simulations is available on GitHub[1].

3.1 Overview of Existing Simulation Settings

We reviewed several existing simulation studies to inform the design of our own, focusing primarily on those that employed projection predictive techniques. These include works by Piironen et al. (2020), Piironen and Vehtari (2017a), Pavone et al. (2023), and McLatchie et al. (2024). In addition, we included the simulation study from Bainter et al. (2023), which compared SSVS to LASSO-type regularization and provided a clear, practically relevant design that informed our Study 1. While these studies vary in terms of dimensionality, sample size, correlation strength, sparsity levels, and effect sizes, they share common goals of evaluating Bayesian variable selection and predictive performance under realistic data-generating scenarios.

Except for Bainter et al. (2023), the reviewed articles used projection techniques, though with varying goals and degrees of emphasis. Piironen et al. (2020) primarily provided illustrative examples to explain the projection framework, including comparisons between shrinkage priors and projection strategies. Their single simulation study focused on evaluating the benefits of reference models, but did not aim to compare PPVS with SSVS. Piironen and Vehtari (2017a) conducted a more systematic evaluation of Bayesian model selection methods, including projection from BMA. Pavone et al. (2023)

[1] https://github.com/zhipeiwang/projpred_vs_spikeslab

Reference	p	n	Sparsity Ratio	Correlation ρ	Effect Size
Bainter et al., 2023	50	{100, 400}	0.2	{0, 0.4, 0.8}	Standardized regression coefficients: $\beta = 0.1$ (4 small), $\beta = 0.3$ (4 medium), $\beta = 0.5$ (2 large)
Piironen et al., 2020	{500}	{50}	0.1	{0.3, 0.5, 0.8}	Latent function $f \sim N(0, 1)$, $y \mid f \sim N(f, 1)$ $x_j \mid f \sim N(\sqrt{\rho} f, 1 - \rho)$, $j = 1, \ldots, 2/p$
Piironen and Vehtari, 2017a	100	{100, 200, 400}	0.15	{0, 0.5, 0.9}	First 15 variables: $(\beta^{1:5}, \beta^{6:10}, \beta^{11:15}) = (\xi, 0.5\xi, 0.25\xi)$
Pavone et al., 2023	70	{80, 100, 150}	0.29	{0.3, 0.5}	Same as Piironen et al., 2020
McLatchie et al., 2024	100	{500}	0.15	{0.9}	Same as Piironen and Vehtari, 2017a

Table 1: Summary of simulation settings in selected studies. For papers with multiple simulation setups, only one representative configuration is shown. p refers to the number of predictors and n to the sample size. The sparsity ratio is defined as p/p_{true} where p_{true} is the number of true predictors.

explored projection-based and reference model strategies under both minimal and complete variable selection frameworks, and considered several additional techniques. Finally, McLatchie et al. (2024) demonstrated the PPVS workflow on several simulated datasets, though their focus was largely methodological and did not involve extensive simulation replications.

Table 1 summarizes the representative features of these designs. Most studies considered moderate-to-high-dimensional settings, with p-to-n ratios between 0.5 and 10, and sparsity ratios between 0.1 and 0.29. Correlation structures ranged from independent to block-diagonal, and some incorporated multiple effect sizes. Notably, Bainter et al. (2023) explicitly varied the spatial distribution of true predictors (clustered vs. mixed).

Taken together, these simulation settings helped shape our study design. We began in low dimensions using the structure from Bainter et al. (2023), and extended the range of conditions by varying both correlation levels and true effect structures under $p = n$ and $p > n$ regimes in Study 2. Our work adds to the literature by directly comparing SSVS and PPVS under these conditions, a comparison that, to our knowledge, has not been conducted before, and is relevant to social sciences.

3.2 Computational details

This section specifies the computational details used in the two studies. For each study, we simulated 100 replications per condition to evaluate the two methods.

3.2.1 Stochastic search variable selection (SSVS)

Each iteration for SSVS was conducted using the `ssvs()` function from the SSVS package (Bainter et al., 2024). The function was executed with default settings, where the spike is a point mass at zero and the slab is a normal distribution with variance 10 (corresponding to τ_j^2 in formulation 2). The prior inclusion probability π in the same formulation was set to 0.5. A noninformative prior $\mathrm{U}(-\infty, \infty)$ was used for the intercept β_0, and a noninformative Gamma(0.01, 0.01) prior was placed on the precision, that is, the inverse of the residual variance σ^2, where σ^2 corresponds to the variance term in Equation 1. By default, we ran one MCMC chain per replication, consisting of 20,000 iterations, discarding the initial 5,000 as warm-up.

3.2.2 Projection predictive variable selection (PPVS)

For PPVS, the regularized horseshoe prior (Piironen & Vehtari, 2017b) was used to fit the reference model using the `horseshoe()` function in the `brms` package (Bürkner, 2017; Bürkner, 2018; Bürkner, 2021), which is based on Stan (Stan Development Team, 2023). We used the default prior settings in `brms`, which specify a student-t prior for the local shrinkage parameters with 1 degree of freedom (`df = 1`), a global shrinkage prior

with `scale_global = 1` and `df_global = 1`, and slab parameters set to `scale_slab = 2` and `df_slab = 4`. The `par_ratio` argument, which indicates the sparsity ratio, was left unspecified to ensure comparability with the SSVS implementation. For a full explanation of these parameters and the theoretical background, see Piironen and Vehtari (2017b). The model was run with four MCMC chains, each with 2,000 iterations, discarding the first 1,000 iterations per chain as warm-up.

Using the function `cv_varsel()` in the package `projpred` (Piironen et al., 2023), and following the suggestions of McLatchie et al. (2024), we performed forward-search cross-validation using the clustered projection technique with 20 clusters and evaluated performance using 400 posterior draws. The search continued until it reached the maximum submodel size of 11 for Study 1 and 16 for Study 2, as we knew the number of true predictors (10 and 15, respectively), to speed up simulations.

We used 5-fold cross-validation because trial runs showed no clear difference between 5-fold and 10-fold results. To automate the process, the `suggest_size()` function from `projpred` was applied to the cross-validation results.

Four heuristic settings were explored for the `suggest_size()` function. The default setting uses the SE heuristic with the confidence interval type set to"upper", while the"lower" heuristic (type set to"lower") uses the lower confidence interval bound. In addition, we considered two alternatives using the Δutility heuristic:"lower threshold" (confidence interval type set to"lower" with an elpd difference threshold of –4 based on McLatchie et al. (2024)), and"upper threshold" (type set to"upper" with the same threshold). For each replication, the variables selected by each heuristic were saved for analysis.

3.3 Measures of performance

We considered the following outcome measures: computational time, convergence, availability of results, variable selection, and predictive performance.

3.3.1 Computational time

For each replication, computational time was recorded and reported in minutes. We computed and reported the mean, standard deviation, and minimum time across replications for each condition. This was particularly relevant because the time required to run a single replication varied considerably for the PPVS method compared to SSVS.

3.3.2 Convergence

Convergence was primarily assessed for the PPVS method, focusing on the reference model fit using the regularized horseshoe prior. We used three diagnostics implemented in Stan (Stan Development Team, 2023) to evaluate convergence: the potential scale

reduction statistic (split $\hat{R}$), introduced by Gelman and Rubin (1992) and extended by Vehtari et al. (2021), the effective sample size for the bulk of the posterior (bulk-ESS), and the number of divergent transitions. We found that the regularized horseshoe prior frequently produces divergent transitions, so we used the combination of divergent transitions and $\hat{R}$ diagnostics to identify replications with clear convergence issues. Specifically, replications with $\hat{R} > 1.1$ on at least one parameter were classified as non-converged and excluded from analysis. These were often accompanied by high rates of divergent transitions. In addition, replications with divergent transition rates exceeding 10%, regardless of $\hat{R}$ values, were also excluded. To further monitor convergence quality, we applied stricter thresholds, flagging replications with $\hat{R} \geq 1.01$ or bulk-ESS ≤ 400 as potentially problematic. These were not removed from the analysis but should be interpreted with caution, particularly for Study 2, where such issues were more frequent. For the SSVS method, we did not assess convergence in detail, as the `ssvs()` function from the SSVS package (Bainter et al., 2024) supports only a single MCMC chain by default.

3.3.3 Availability of results

The success or failure of the `suggest_size()` function is not directly indicative of convergence issues but presents a major practical limitation and is worth discussing separately. As described in the computational details, `suggest_size()` was used to automate submodel size selection. Because we constrained the maximum number of predictors to 11 for Study 1 and 16 for Study 2 (based on the known true model sizes), selected submodels occasionally could not reach predictive performance close enough to the reference model. In such cases, `suggest_size()` fails and returns a missing result (NA). This behavior is expected under strict heuristics and was necessary to allow automated processing across all replications.

In real-world applications, users could manually inspect the utility profiles or ranking proportions to select a model size when `suggest_size()` fails. In our simulations, we observed particularly high failure rates with the SE heuristic using the lower confidence bound. This heuristic is highly conservative, requiring strong evidence that the submodel performs at least as well as the baseline model. As a result, it often favors larger submodel sizes and is more likely to return an error when those submodels with the largest sizes also cannot reach the required utility threshold.

To reduce redundancy, we excluded several heuristics that consistently produced similar results. The upper threshold and lower threshold heuristics typically result in the same submodel sizes, so we retain only the upper threshold for simplicity. Likewise, the default and upper threshold heuristics typically result in the same submodel size, with negligible differences in distributional characteristics, so we retain only the default for comparison. It is also worth noting that the threshold-based heuristic only overrides the SE heuristic when the SE heuristic fails. If both heuristics are successful, the smaller

submodel size is chosen. Thus, even if a threshold is specified, it is not always used to determine the final size, so the comparison can be tricky.

Based on these considerations, we retained only one heuristic for the main comparison: the default (i.e., SE upper bound) for PPVS. Together with SSVS, this yielded two methods for direct comparison. Although the SE lower heuristic produced many missing values, we retained those replications in the dataset for transparency and potential further investigation.

3.3.4 Variable Selection Accuracy

To assess variable selection accuracy, we computed the true inclusion and false inclusion rates for each method. The true inclusion rate is calculated as the number of true predictors selected over the total number of predictors selected, and the false inclusion computed as the number of irrelevant predictors selected over the total number of predictors selected. Additionally, we calculated inclusion rates stratified by effect size groups (large, medium, small, and null) to evaluate how well each method identifies predictors of varying importance.

For each heuristic or method, and for each available replication (as the `suggest_size()` function may fail for some heuristics, the number of available replications may differ), we computed these rates either overall or stratified by effect size. We then summarized performance by examining the means and distributions of the rates across replications for each condition. Accordingly, the inclusion rates for large, medium, and small effect groups correspond to true inclusion rates, reflecting the proportion of true predictors correctly selected by the methods for different effect sizes, while the inclusion rate for null effect predictors is interpreted as the false inclusion rate.

For SSVS, variables were considered selected if their marginal inclusion probabilities (MIPs) exceeded 0.5, consistent with the approach in Bainter et al. (2023), which is considered optimal under certain conditions (Barbieri & Berger, 2004). For PPVS, variable selection was determined based on four distinct heuristics used to define the submodel size through the `suggest_size()` function, allowing for a comparison across five selection strategies (four PPVS heuristics and SSVS).

3.3.5 Predictive performance

To assess predictive performance, we evaluated the prediction mean squared error (PMSE) and prediction interval coverage.

PMSE PMSE was computed using predictions generated by the `proj_predict()` function in the `projpred` package (for PPVS) and manually computed for SSVS. For each replication, a test dataset was simulated from the same distribution as the training data

and scaled using training-set means and standard deviations. Predictions were based on 400 posterior draws per observation in PPVS (default in `projpred`) and 15,000 draws in SSVS (default in the `SSVS` package).

This yielded a matrix of predictive samples with dimensions (number of draws) $\times$ (number of test observations)—i.e., 400 $\times$ 1000 for PPVS. For each test observation, predictive means were obtained by averaging across the posterior draws. PMSE was then computed as: $\text{PMSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, where y_i is the true response for observation i, and $\hat{y}_i$ is the prediction for that observation i. We summarized PMSE across replications by reporting means and visualizing the distribution.

Prediction interval coverage We also looked at the coverage of the prediction interval. For each of the 1000 test observations, we computed a 95% credible interval based on the 400 predictive draws. An observation was considered "covered" if the true value lay within this interval. The coverage percentage for a replication was the proportion of test observations for which coverage was achieved. We then assessed performance by examining the distribution and mean of coverage percentages across replications.

3.4 Conditions simulation study 1

To compare the performance of SSVS and PPVS, we employed simulated datasets with a challenging configuration characterized by highly correlated predictors and a small sample size. The simulation setup was adopted from Bainter et al. (2023), where SSVS was previously compared against LASSO-type regularization, but not directly compared with PPVS—a context particularly relevant to social science applications. Table 2 summarizes the simulation conditions. The simulation setup consisted of 100 observations and 50 predictors, among which 10 had true effects (Table 2). The predictors $\boldsymbol{X}$ were generated from a multivariate normal distribution with mean zero, unit variances, and block-diagonal correlation structures. In conditions with correlated predictors, sets of five predictors were grouped and assigned either moderate $\rho = 0.4$ or high $\rho = 0.8$ within-block correlations. In the condition with clustered true effects, the 10 non-zero coefficients were assigned to two correlated clusters. The regression coefficients for this scenario, β were defined as $\boldsymbol{\beta} = (1.5, .9, .9, .3, .3, 1.5, .9, .9, .3, .3, 0, ..., 0)'$.

For the uncorrelated true effects condition, the magnitudes of the coefficients remained unchanged; however, their distribution across predictors was altered: every fifth predictor was impactful, resulting in a coefficient vector $\boldsymbol{\beta} = (1.5, 0, 0, 0, 0, .9, 0, ..., 0)'$. This modification ensured that the effects were isolated rather than clustered.

The response vector $\boldsymbol{Y}$ was drawn from $N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, where $\sigma^2 = 1$, leading to four small, four medium, and two large effects with standardized regression coefficients of approximately .1, .3, and .5, respectively.

Table 2: Summary of simulation conditions for Study 1 and Study 2. Both studies vary the spatial distribution of true effects (clustered vs. mixed), while Study 2 also introduces variation in sample size and uses a scalable effect size parameter ξ. The sparsity ratio is defined as the number of true predictors over the total number of predictors. The effect sizes are for the true predictors and the ordering is for the clustered scenario.

Study 1		
Variable	Number of Levels	Values
Sample Size	1	100
Number of Predictors	1	50
Sparsity	1	10/50
Correlation	3	{0, 0.4, 0.8}
Effect Size	1	$(\beta^{1,6}, \beta^{2,3,7,8}, \beta^{4,5,9,10}) = (1.5, 0.9, 0.3)$
Covariance Structure	1	Block-diagonal
Pattern of True Effects	2	Clustered vs. Mixed
Study 2		
Variable	Number of Levels	Values
Sample Size	2	{50, 75}
Number of Predictors	1	75
Sparsity	1	15/75
Correlation	2	{0.2, 0.8}
Effect Size	1	$(\beta^{1:5}, \beta^{6:10}, \beta^{11:15}) = (\xi, 0.5\xi, 0.25\xi)$
Covariance Structure	1	Block-diagonal
Pattern of True Effects	2	Clustered vs. Mixed

3.5 Conditions simulation study 2

Table 2 summarizes the simulation conditions. We fixed the number of predictors at 75, while varying the sample size between 50 and 75 to represent two scenarios: $p > n$ (when $n = 50$) and $p = n$ (when $n = 75$). In addition to sample size, two other factors were varied: the correlation strength among predictors and the spatial pattern of true effects.

Predictors were divided into 15 groups of five, with within-block correlations set to either $\rho = 0.2$ (low) or $\rho = 0.8$ (high), using block-diagonal covariance matrices. Following the setup from simulation study 1 and Bainter et al. (2023), the pattern of true effects was either clustered, where the first 15 predictors (three blocks) had non-zero coefficients, and mixed, where every fifth predictor had a non-zero coefficient. In both cases, 15 out of the 75 predictors were assigned non-zero coefficients, resulting in a sparsity ratio of 0.2. The 15 true predictors were grouped into three effect size categories (large, medium, small), with coefficient values defined as $(\beta^{1:5}, \beta^{6:10}, \beta^{11:15}) = (\xi, 0.5\xi, 0.25\xi)$, following Piironen and Vehtari (2017a).

Following McLatchie et al. (2024) and Piironen and Vehtari (2017a), the value of the scaling constant ξ was chosen to ensure that the predictors collectively explained 70% of the variance in the response variable, regardless of the correlation level or the pattern of the true effects. When the true effects were mixed (i.e., uncorrelated), $\xi = 0.59$ was used for both low and high correlation settings. When the true effects were clustered, ξ was set to 0.44 for $\rho = 0.2$ and 0.29 for $\rho = 0.8$. As in Study 1, the response vector $\boldsymbol{Y}$ was generated from $N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, with $\sigma^2 = 1$.

4 Results

This section presents the results from Study 1 and Study 2. For each study, we report outcomes based on the performance measures described earlier, followed by a subset analysis evaluating the SE heuristic using the lower confidence interval bound.

4.1 Simulation study 1

4.1.1 Computational Time

The computational time for each replication was recorded during the simulation, and summary statistics for each condition are reported in Table 3. The simulations were run on a personal laptop (32 GB RAM, 2.20 GHz processor). Times are reported in minutes.

As shown in the table, compared to PPVS, SSVS is extremely fast and stable: across all conditions, it required approximately half a minute per replication, with minimal variability across the 100 replications. In contrast, PPVS was substantially more computationally intensive. Although the number of predictors and the sample size remained

Table 3: Summary of computational time (in minutes) with a 2.20 GHz processor for PPVS and SSVS for Study 1 (under $N = 100$ with 50 predictors). TE refers to the spatial pattern of true effects, and SD refers to standard deviation.

Method	Correlation	TE	Mean	Min	SD
PPVS	0.0	clustered	11.01	4.65	18.10
	0.4	clustered	16.12	4.90	31.05
		mixed	15.57	5.12	34.96
	0.8	clustered	20.38	5.88	47.58
		mixed	19.79	4.81	48.50
SSVS	0.0	clustered	0.52	0.46	0.06
	0.4	clustered	0.52	0.46	0.06
		mixed	0.52	0.46	0.06
	0.8	clustered	0.52	0.46	0.06
		mixed	0.52	0.47	0.06

constant across conditions, the PPVS runtime increased with higher correlation among predictors: the mean runtime rose from approximately 11 minutes ($\rho = 0$), to 16 minutes ($\rho = 0.4$), and to around 20 minutes ($\rho = 0.8$). The structure of the true effects, however, did not appear to influence computational time. Standard deviations were also large—sometimes even exceeding the mean—which likely reflects the memory-intensive nature of the PPVS method. As the simulation progressed and more replications were run in parallel, memory usage increased considerably, which likely contributed to the slowdown in performance over time. These findings confirm that PPVS is markedly more computationally demanding than SSVS. This is expected, as PPVS involves repeated reference model fitting across multiple submodel sizes as part of its cross-validation procedure.

4.1.2 Convergence

Based on the criteria outlined in Section 3.3, one PPVS replication was excluded due to non-convergence, and an additional 6 replications were flagged as potentially problematic but retained, given the small number and mild violations. All SSVS replications were included due to the single-chain default in the `SSVS` package (Bainter et al., 2024).

4.1.3 Availability of results

The SE heuristic with the lower confidence bound produced over 40 failures (i.e., NA values) in every condition, with one case reaching 70. In contrast, the default heuristic (SE upper bound) resulted in fewer than 10 failures in all conditions except for the high-correlation (0.8) and mixed true effects scenario, which had 23. Based on the reasoning in Section 3.3, we retained only the PPVS-default SE heuristic upper bound and SSVS methods for the main analysis. The number of available replications in each condition was considered sufficient.

4.1.4 Variable selection accuracy

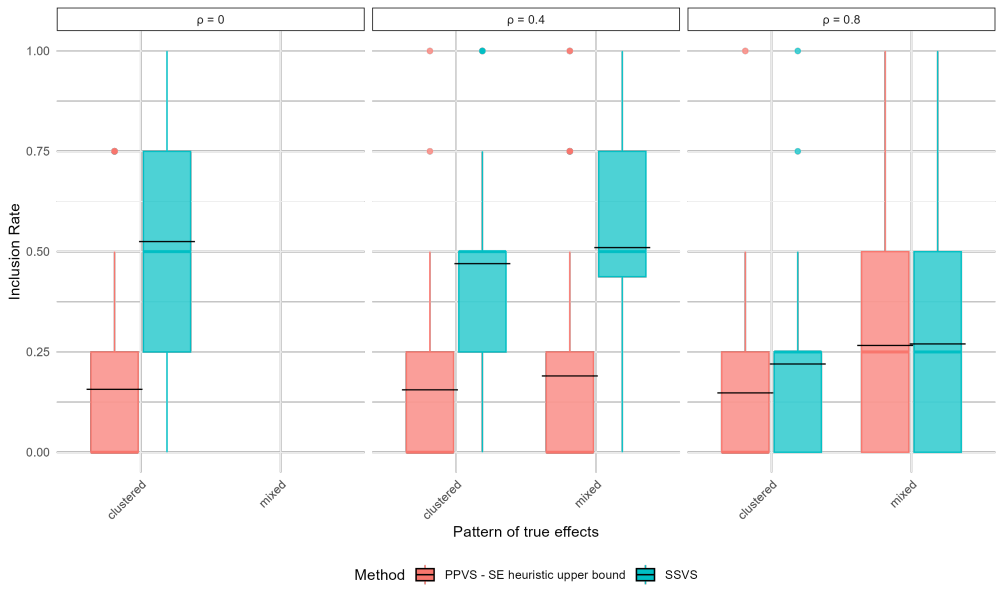

Figure 1: True inclusion rates for predictors for small effect sizes across simulation conditions in Study 1. The black horizontal bars denote the means. The two methods compared are projection predictive variable selection (PPVS), using the SE heuristic (McLatchie et al., 2024) with the upper confidence interval bound, and stochastic search variable selection (SSVS). Inclusion rate is computed as the proportion of true predictors (with small effects) that were selected. The panel for the mixed true effect pattern at $\rho = 0$ is empty because the spatial distribution of true effects is irrelevant in the absence of correlation; only the clustered pattern was included for this setting.

To investigate whether different methods tend to select predictors with different effect sizes, we stratified inclusion rates by effect size group (large, medium, small, null) and summarized results across all replications. The general pattern was the same across effect sizes, although differences were most pronounced for small effects as shown in

Figure 1.

As shown in Figure 1, for small effects, SSVS consistently exhibited higher inclusion rates than PPVS, especially under lower correlation levels. As correlation strength increased, the gap narrowed. However, PPVS inclusion rates remained around 25% across all conditions in this group—suggesting it typically captured only one out of the five small-effect predictors.

For large effects (see appendix Figure 1), both methods achieved near-perfect inclusion rates (close to 1) across all conditions. For medium effects (see appendix Figure 1), inclusion rates remained high under $\rho = 0$ and $\rho = 0.4$, but dropped for $\rho = 0.8$, particularly for PPVS under the clustered effect pattern. For null effects, which represent false inclusions, both methods maintained low rates (close to 0) across conditions (see appendix Figure 2). Yet in the high-correlation, mixed-effects condition, PPVS showed a slightly higher mean inclusion rate. This suggests its improved detection of small effects may come at the cost of slightly increased false inclusions.

4.1.5 Predictive performance

As shown in Figure 2, SSVS generally achieved lower PMSE values than PPVS across all conditions, with an average difference of around 0.1. The only exception occurred in the condition with mixed true effects and high correlation ($\rho = 0.8$), where PPVS showed a slight advantage.

Regarding coverage, Figure 3 reveals that although PPVS tended to have higher PMSE, it consistently achieved higher mean coverage percentages than SSVS, except under conditions with $\rho = 0.8$. These findings suggest that PPVS, while potentially less precise in producing a single prediction value, tends to produce credible intervals that more reliably contain the true value across replications. Nevertheless, it is worth noting that neither of the methods reached the desired 95% coverage in any condition, indicating overconfidence in the estimated 95% credible intervals.

Given that PPVS does not consistently outperform SSVS in terms of variable selection or predictive accuracy—and only demonstrates slight or comparable advantages in a few scenarios—its significantly higher computational cost raises important questions about its practical value in moderate-dimensional settings. This motivated our decision to extend the study to higher-dimensional settings, where PPVS may demonstrate clearer advantages.

4.2 Subset Analysis

To evaluate the performance of the SE heuristic (McLatchie et al., 2024) lower confidence interval bound in PPVS, we focus on the subset of replications where it successfully returned a submodel size. These cases are relatively limited in number per condition

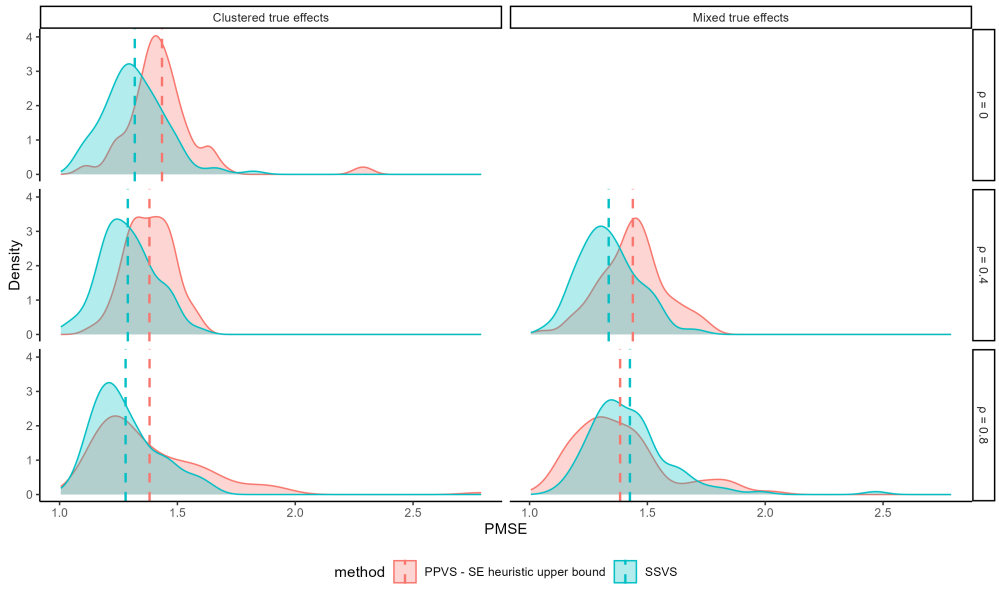

Figure 2: Distributions of the prediction mean squared error (PMSE) across simulation conditions in Study 1. Each panel shows the distribution of PMSE values across replications for one combination of correlation level (ρ) and spatial pattern of true effects (clustered or mixed). The two methods compared are projection predictive variable selection (PPVS), using the SE heuristic with the upper confidence interval bound (McLatchie et al., 2024), and stochastic search variable selection (SSVS). Dashed vertical lines indicate the mean PMSE for each method within a condition. The panel for the mixed true effects at $\rho = 0$ is empty because, in the absence of correlation, the spatial distribution of effects is irrelevant; thus, only the clustered pattern was included for that setting.

(around 30 to 60), which limits the generalizability of the findings. Nevertheless, consistent patterns emerge across conditions.

The lower heuristic yields higher or comparable true inclusion rates than the default (upper bound) heuristic in all conditions (see Appendix Figure 4). It also performs comparably to SSVS in terms of average inclusion rates. The higher inclusion rates stem from its selection of larger submodel sizes—as expected from the heuristic's conservative criterion, which requires stronger evidence that the submodel performs no worse than the reference model. While this improves true inclusion, it also results in higher false inclusion rates (i.e., inclusion rates for null effects).

In terms of predictive performance, the lower heuristic consistently outperforms the default heuristic and performs similarly to SSVS in terms of PMSE (see Appendix Figure 5). It also shows stronger or comparable coverage performance relative to both the default heuristic and SSVS in all conditions (see Appendix Figure 6).

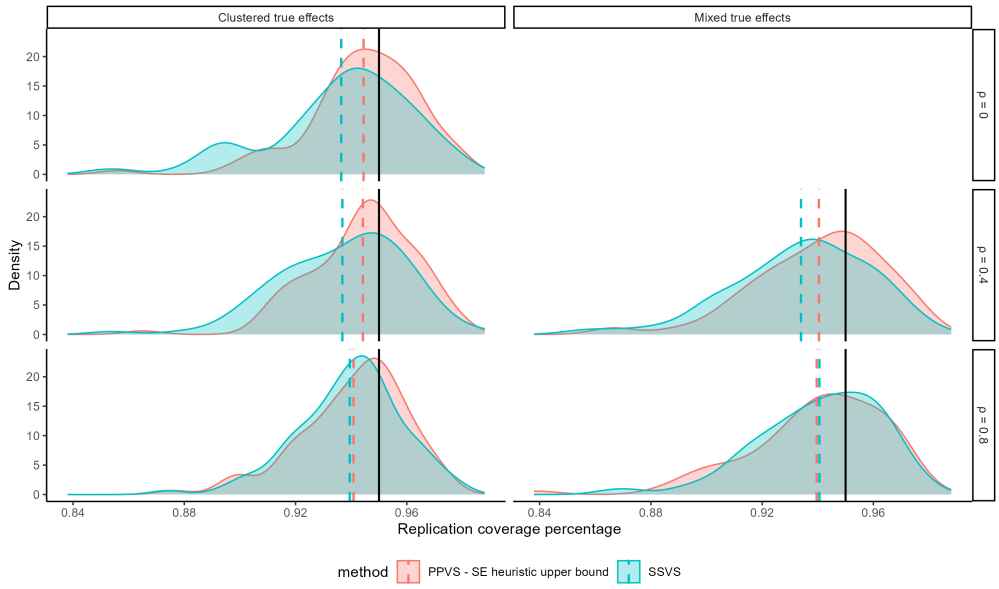

Figure 3: Distributions of coverage percentages across conditions in Study 1. Each curve shows the distribution across replications of the proportion of test observations whose 95% credible interval contained the true value. Dashed vertical lines indicate the mean coverage for each method, while the solid black line marks the nominal 95% target. The two methods compared are projection predictive variable selection (PPVS), using the SE heuristic with the upper confidence interval bound (McLatchie et al., 2024), and stochastic search variable selection (SSVS). The panel for mixed true effects at $\rho = 0$ is empty because the spatial pattern is irrelevant when predictors are uncorrelated.

4.3 Simulation study 2

4.3.1 Computational time

The simulations were run on a departmental server equipped with 64 GB RAM and two Intel Xeon E5-2650 v4 processors (2.20 GHz, 48 threads). Times are reported in minutes. As expected, as shown in Table 4 both methods required more time to complete each replication in Study 2 compared to Study 1, due to the increase in dimensionality. For SSVS, the mean runtime ranged from just over 1 minute to nearly 5 minutes across conditions, with larger variability than observed in Study 1. For PPVS, the computational demands were considerably higher: the average runtime exceeded 30 minutes per condition, with standard deviations around 5 minutes. These results reaffirm that PPVS is substantially more computationally intensive than SSVS. This is consistent with expectations, given the repeated reference model fitting during cross-validation inherent in the PPVS workflow. While the increased dimensionality of Study 2 explains the overall rise in runtime for both methods, the performance gap in computational efficiency

between PPVS and SSVS remains pronounced.

Table 4: Summary of computational time (in minutes) with a 2.20 GHz processor for PPVS and SSVS in Study 2 (75 predictors). N is the sample size, TE indicates the spatial pattern of true effects, and SD refers to standard deviation.

Method	N	Correlation	TE	Mean	Min	SD
PPVS	50	0.2	clustered	33.31	20.02	7.30
			mixed	33.26	20.17	6.47
		0.8	clustered	30.90	21.88	4.05
			mixed	30.49	21.48	3.97
	75	0.2	clustered	31.22	20.66	6.07
			mixed	31.60	21.87	6.45
		0.8	clustered	32.24	21.35	4.45
			mixed	31.95	21.83	4.68
SSVS	50	0.2	clustered	4.75	0.63	34.89
			mixed	2.03	0.69	7.71
		0.8	clustered	1.26	0.66	0.55
			mixed	3.05	0.69	17.91
	75	0.2	clustered	3.85	0.73	24.64
			mixed	3.18	0.78	17.91
		0.8	clustered	2.15	0.77	7.72
			mixed	4.87	0.58	34.89

4.3.2 Convergence

Based on the criteria outlined in Section 3.3, five PPVS replications were excluded due to non-convergence. Another 99 out of 800 replications were flagged as potentially problematic, primarily in settings with small sample sizes (N = 50) and low correlation, and were retained for completeness, though their results should be interpreted with caution. As in Study 1, all replications for SSVS were included due to its one-chain implementation in the `SSVS` package (Bainter et al., 2024).

4.3.3 Availability of results

In Study 2, failures of the `suggest_size()` function were more frequent overall. The SE heuristic with the lower confidence bound failed in over 50 replications in every condition, and in several cases exceeded 75. Even the default heuristic showed increased failure rates, especially under high correlation and mixed true effects, where around 40 failures occurred. As in Study 1, we retained only PPVS-default and SSVS for the main comparison. Despite the higher rate of missing values, 50 to 90 replications per condition remained available for analysis, which we considered sufficient.

4.3.4 Variable selection accuracy

As in Study 1, we focus on two methods in the main comparison PPVS-default SE heuristic upper bound and SSVS. The analyses are structured into a full analysis using all available replications and a subset analysis focusing only on the replications where the lower heuristic was successful.

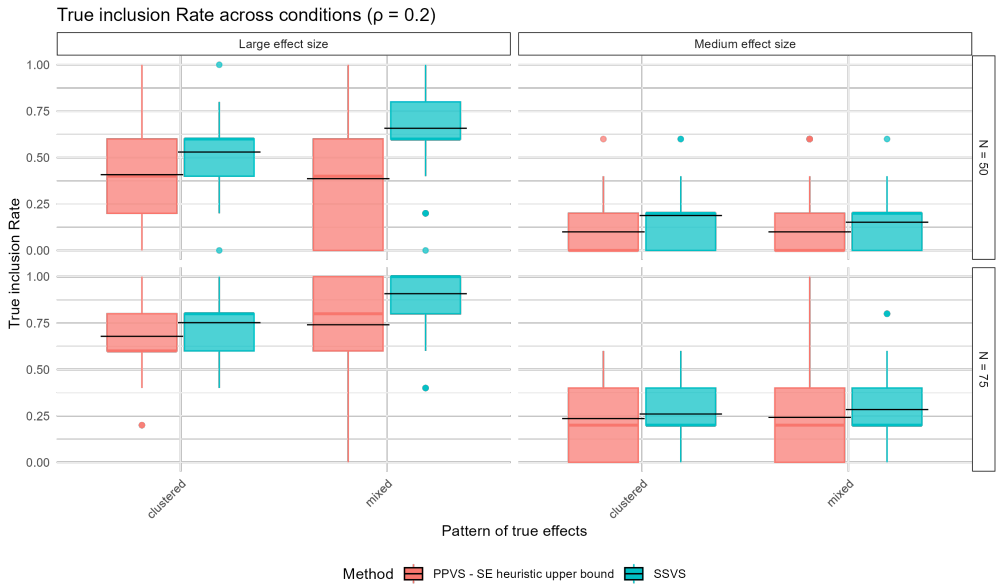

Figure 4: True inclusion rates for predictors for large and medium effect sizes across simulation conditions in Study 2. The black horizontal bars denote the means. The two methods compared are projection predictive variable selection (PPVS), using the SE heuristic (McLatchie et al., 2024) with the upper confidence interval bound, and stochastic search variable selection (SSVS). True inclusion rate is computed as the proportion of true predictors (with large and medium effects) that were selected.

We stratified inclusion rates by effect size group and summarized results across all replications. Figure 4 presents the distributions of inclusion rates for predictors with

large and medium effects under low correlation ($\rho = 0.2$), where differences between methods are most pronounced. Across conditions, SSVS generally achieves higher true inclusion rates than PPVS, and this advantage is most evident for predictors with large effects. The inclusion rate distributions for SSVS also tend to be narrower, suggesting more consistent and stable performance across replications. For medium effects, the advantage of SSVS under low correlation remains, though differences are smaller than for large effects.

Overall, SSVS offers more robust and consistent inclusion behavior under low correlation, while PPVS may offer advantages under high correlation when detecting moderate or clustered effects—though potentially at the cost of higher false inclusion rates (see Appendix Figure 3 for all conditions). To better understand these patterns, we also examined the number of predictors selected. PPVS tends to select fewer predictors than SSVS when correlation is low. When correlation is high, PPVS selects submodels of similar or slightly larger size—particularly when the sample size is large and true effects are mixed, contributing to its higher true and false inclusion rates.

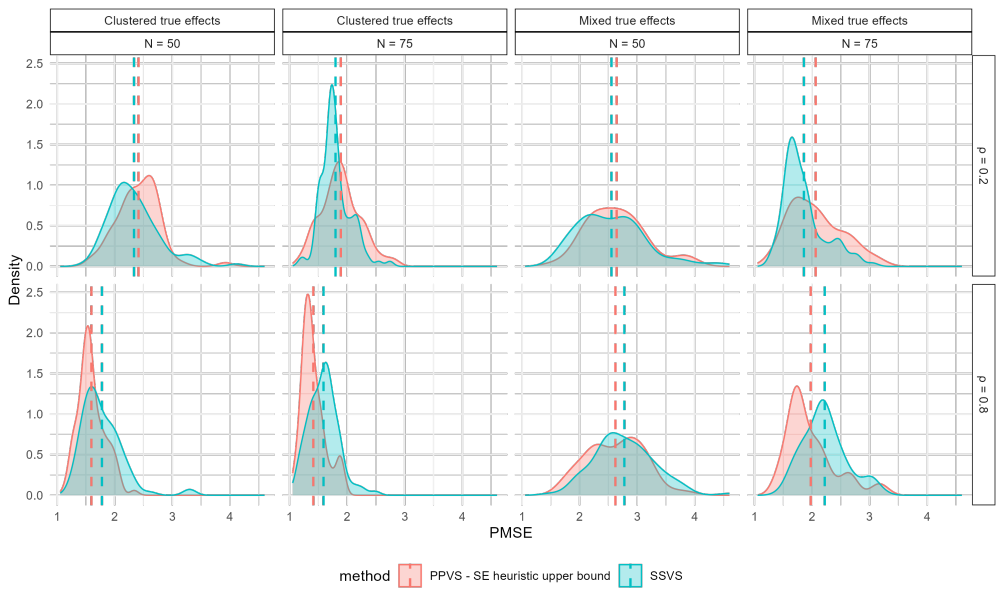

Figure 5: Distributions of the prediction mean squared error (PMSE) across simulation conditions in Study 2. Each panel shows the distribution of PMSE values across replications for one combination of correlation level (ρ), sample size (N), and spatial pattern of true effects (clustered or mixed). The two methods compared are projection predictive variable selection (PPVS), using the SE heuristic with the upper confidence interval bound (McLatchie et al., 2024), and stochastic search variable selection (SSVS). Dashed vertical lines indicate the mean PMSE for each method within a condition.

4.3.5 Predictive performance

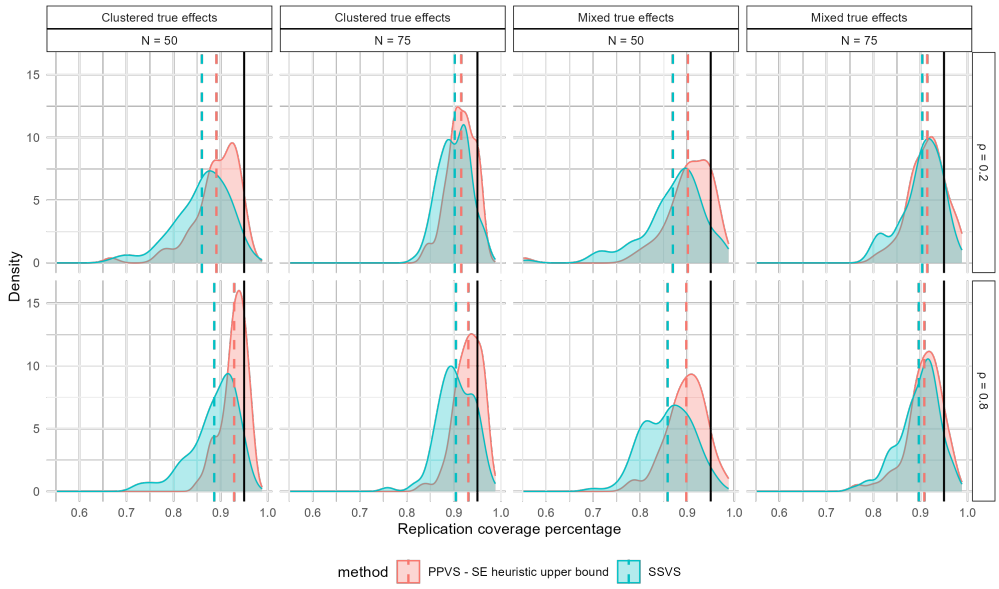

Figure 6: Distributions of coverage percentages across conditions in Study 2. Each curve shows the distribution across replications of the proportion of test observations whose 95% credible interval contained the true value. Dashed vertical lines indicate the mean coverage for each method, while the solid black line marks the nominal 95% target. The two methods compared are projection predictive variable selection (PPVS), using the SE heuristic with the upper confidence interval bound (McLatchie et al., 2024), and stochastic search variable selection (SSVS).

Clear differences in PMSE patterns emerge across conditions, as shown in Figure 5. As expected, smaller sample sizes lead to weaker predictive performance for both methods, regardless of correlation or effect structure. When the correlation is low, SSVS consistently yields lower mean PMSE values than PPVS, regardless of sample size or the structure of the true effects. In contrast, when the correlation is high, PPVS tends to outperform SSVS, with lower mean PMSE across most conditions. Under small sample size and mixed true effects, the PMSE distributions become wider for both methods.

Examining the coverage percentages across conditions, as in Figure 6, PPVS consistently outperforms SSVS, yielding higher mean coverage and comparable or narrower spread in every condition. In general, both methods demonstrate relatively stable predictive performance on this metric. As in Study 1, neither of the methods reached the desired 95% coverage in any condition, demonstrating overconfidence in the estimated 95% credible intervals.

4.4 Subset Analysis

As in Study 1, we examined the subset of replications in which the SE heuristic lower confidence interval bound successfully returned a submodel size. In Study 2, the number of successful cases was smaller—typically between 10 and 40 per condition, further limiting generalizability.

The key patterns observed in Study 1 largely held in Study 2. The lower heuristic again produced higher true inclusion rates than the default (upper bound) heuristic in all conditions (see Appendix Figure 7) and generally outperformed SSVS. These gains were due to larger submodel sizes, which also led to increased false inclusion rates.

Predictive performance remained consistent with earlier findings: the lower heuristic outperformed the default heuristic in all conditions (Appendix Figure 8). Compared to SSVS, it showed a more substantial advantage when correlation was high, and similar performance when correlation was low. In terms of coverage, the lower heuristic again achieved comparable or better results than both the default heuristic and SSVS (Appendix Figure 9).

5 Empirical applications

We applied the PPVS and SSVS methods to two real-world datasets, and their predictive performance and model parsimony were examined.

5.1 Body fat

The percentage body fat dataset [2] is from Johnson (1996), in which body fat was originally measured using underwater weighing techniques—an accurate but complex and costly procedure. Alongside these measurements, the dataset includes 13 simple body measurements such as weight, height, and neck circumference. Fitting a predictive model on these easily obtainable predictors allows body fat percentage to be estimated in practice without relying on an expensive procedure.

In his note, Johnson (1996) flagged several problematic entries: cases 48, 76, and 96 were suspected to contain one-digit errors in the response variable; case 42 had an implausible height of only 29.5 inches despite a weight of over 200 pounds; case 182 reported a body fat percentage of zero; and case 39 had a highly unlikely weight of 363.15 pounds. We excluded these six cases, resulting in a final dataset of 246 observations.

We split the dataset evenly into training and test sets, each containing 123 observations. The primary goal was predictive accuracy—that is, the ability to make reliable predictions for new individuals. Model parsimony was also an important consideration, as we aimed to select a minimal number of predictors without sacrificing accuracy.

[2] The data is openly available from https://jse.amstat.org/v4n1/datasets.johnson.html.

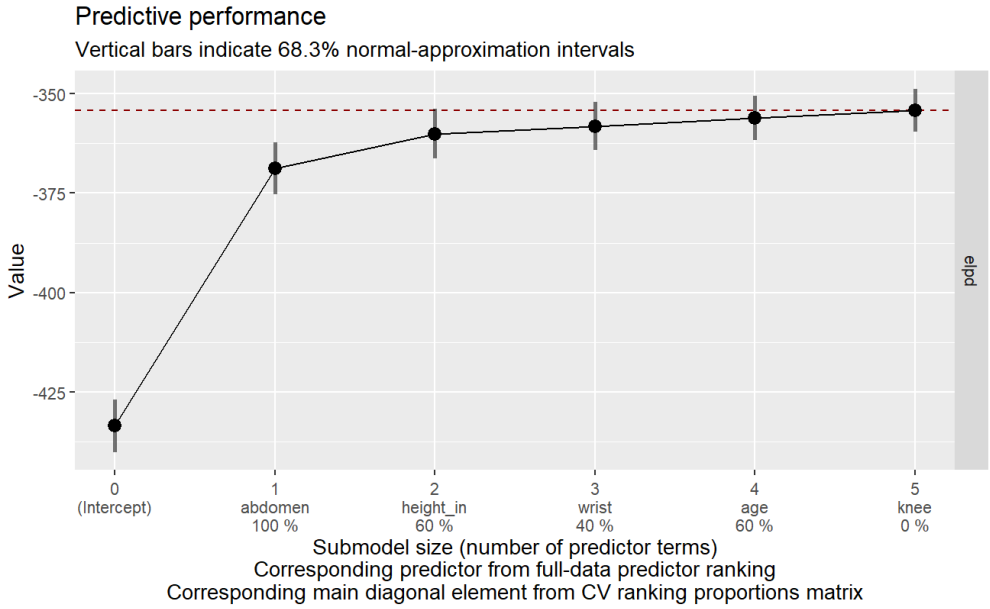

Figure 7: Solution path for the body fat dataset. The y-axis shows the expected log pointwise predictive density (ELPD), and the dashed red horizontal line represents the ELPD of the reference model.

Unlike the simulation studies, where the PPVS workflow was fully automated, we implemented the projection workflow step by step. We first fit the reference model using the regularized horseshoe prior with the same default settings as in the simulation study and assessed whether it was a reasonable model fit, since the quality of the projected submodel depends on the reference model. To assess fit, we examined predictive performance by generating posterior predictive samples and comparing them to observed data (Gabry et al., 2019).

Next, we ran a preliminary search without cross-validation or refitting, followed by another quick search with refitting but still without cross-validation. These preliminary searches helped us decide on the appropriate upper limit for submodel size, since the true number of relevant predictors is unknown in real data. We then ran the full projection search with 5-fold cross-validation.

After the search, we manually inspected the solution path (Figure 7) along with predictor rankings and ranking proportions to determine submodel size. The solution path suggests that the ELPD stabilizes after submodel size 2, and the fifth predictor from the full-data ranking had zero proportion of being added in the cross-validated rankings. To remain conservative and retain proximity to the reference model, we selected a submodel of size 4. The `suggest_size()` function also returned 4, agreeing with our manual selection. We then projected the reference model onto this submodel and generated test

set predictions, from which the Predictive Mean Squared Error (PMSE) was computed.

For SSVS, we followed the same automated workflow as in the simulation studies, using the default priors and the `ssvs()` function from the SSVS package (Bainter et al., 2024). Variables were selected based on the same marginal inclusion probability (MIP) threshold of 0.5. The results showed that PPVS achieved a test PMSE of 19.40, slightly lower than the 19.86 obtained by SSVS, indicating negligible difference in predictive performance. In terms of parsimony, SSVS selected three predictors: abdomen circumference, height, and age. PPVS selected these three plus wrist circumference, resulting in a slightly larger model.

The inclusion of one additional predictor in the PPVS model may not be disadvantageous. Figure 7 and the associated utility values show that submodels of size three and four yield very similar ELPD. In practice, if collecting a particular measurement (e.g., wrist circumference) is costly or impractical, a smaller submodel may be chosen without meaningful loss in predictive utility. Ultimately, the decision depends on the context—whether a researcher prioritizes parsimony and interpretability or is more concerned with complete variable selection and minimizing missed predictors.

Additionally, Figure 8 shows that the posterior predictive distributions obtained from both the PPVS and SSVS after selecting the variables are quite close to the distribution of the true outcome values in the test set.

As in the simulations, PPVS also required more computational time, although this was less burdensome here due to the moderate dimensionality of the dataset.

5.2 Prostate cancer

The prostate cancer data [3] is from Stamey et al. (1989), which contains observations from 97 men who underwent prostatectomy. The outcome of interest is the logarithm of the prostate-specific antigen level, with eight clinical predictors such as log(cancer volume), log(prostate weight), and age. The dataset was split into approximately equal training (n = 48) and test (n = 49) sets.

As with the body fat example, our primary goals were predictive accuracy and model parsimony, and we applied the same modeling workflow. PPVS selected two predictors—log(cancer volume) and log(benign prostatic hyperplasia), and achieved a PMSE of 2.0793. In comparison, SSVS selected only log(cancer volume), yielding a PMSE of 2.0805. The difference in predictive accuracy is negligible.

However, as illustrated in Figure 8, the posterior predictive distributions from both methods deviate noticeably from the distribution of the true data. This suggests that both models struggle to fully capture the variability in the outcome, possibly due to the increase in dimensionality and data characteristics.

[3] Available in the R package `genridge` (Friendly, 2024).

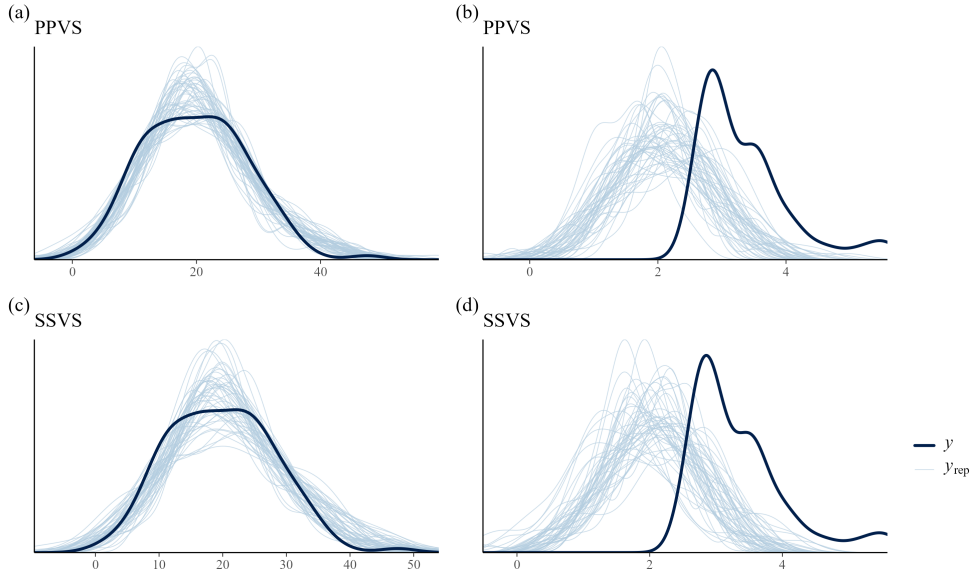

Figure 8: Posterior predictive checks for both empirical datasets after variable selection (projection predictive distribution in the case of PPVS). The left two panels (a) and (c) show results for the body fat dataset, and the right two panels (b) and (d) are for the prostate cancer dataset. The dark curves are the empirical distribution of the true response variable y, while the light blue lines are 50 randomly drawn posterior predictive distributions.

6 Discussion

In conclusion, we reviewed the two methods PPVS and SSVS briefly, and reviewed relevant simulation studies involving projection techniques. We conducted two simulation studies informed by the settings we reviewed, with the first still in low-dimension settings and the second study moved to high-dimensional settings. We also applied the methods to two empirical datasets to demonstrate their practical relevance and real-world utility.

Overall, PPVS did not outperform SSVS in most conditions in terms of variable selection accuracy, largely due to its tendency to select too few predictors. The more lenient SE heuristic lower confidence interval bound helped address the conservative tendency of the default heuristic, but it frequently failed to suggest a size because even the largest submodel did not come close enough to the predictive performance of the reference model. In such cases, the `suggest_size()` function throws an error suggesting the user increase the maximum number of allowed predictors.

Our simulations showed that results depend heavily on the heuristic used. In practice, users can inspect plots such as the solution path in the body fat dataset to guide submodel size decisions. However, this approach introduces arbitrariness and may result

in less parsimonious models. This highlights a practical limitation of the PPVS workflow: while it offers flexibility and strong theoretical justification, it may require manual tuning and expert judgment to be effective.

Interestingly, while PPVS tended to select fewer predictors in the simulation studies, it selected more in the empirical datasets, which may be due to the greater complexity and noise in real-world datasets compared to simulated ones. Additionally, we don't know the true number of relevant predictors in the empirical examples, which means we have to manually inspect the pilot cross-validation results to decide on a maximum model size. That adds another layer of uncertainty to the final submodel size.

The PPVS method is fundamentally designed to prioritize predictive performance rather than identification of the true data-generating process. In Study 1, where dimensionality was still moderate, this advantage did not clearly emerge—PPVS matched or exceeded SSVS in predictive accuracy in only one condition. However, in Study 2, when we entered the higher-dimensional setting, PPVS consistently produced predictive accuracy that was comparable to or better than SSVS across conditions, even though its variable selection accuracy remained lower. Additionally, PPVS produced more reliable prediction intervals in these scenarios, suggesting that its strength lies in stable, generalizable prediction rather than exact variable recovery.

This highlights the practical value of PPVS. Despite its higher computational cost, it may be preferable in situations where the goal is to build a simpler model with minimal loss of predictive accuracy. For example, in applied research where some measurements are expensive or invasive, PPVS can help identify a reduced set of predictors that still yield strong predictions. A common use case might be a pilot study that initially collects a large set of variables; by running PPVS on this full dataset, one can determine which variables are most informative and focus on collecting only those in future studies. By contrast, if the primary goal is to recover as many relevant variables as possible—such as in exploratory or explanatory research contexts—SSVS may be the more suitable choice.

While previous work such as Pavone et al. (2023) explored projection methods in very high-dimensional scenarios where $p \gg n$, they did so by using supervised principal components (SPC; Bair et al., 2006, Piironen and Vehtari, 2018) to reduce the dimensionality of the reference model. Implementing such techniques was beyond the scope of this study. Moreover, even in our moderately high-dimensional settings, we already observed convergence issues and increased failure rates in the automated submodel selection, suggesting that further improvements to the robustness and automation of PPVS are necessary before scaling up.

In addition to computational cost, another practical limitation of PPVS is its complexity. Unlike SSVS, which can be implemented with relatively few decisions, PPVS requires navigating a multi-step workflow. In this thesis, although we followed default settings based on previous recommendations, many of these decisions are non-trivial

and may impact results. This added complexity may deter practitioners who lack the resources or experience to navigate the full workflow.

While SSVS was consistently much faster than PPVS across all conditions, this low computational cost should be interpreted with some caution. In our implementation, the SSVS package runs only a single MCMC chain by default. For more robust posterior inference, it is typically recommended to run multiple chains and assess convergence using diagnostics such as $\hat{R}$. Doing so would increase the runtime somewhat, although likely still far below that of PPVS.

Future work could investigate how PPVS performs in truly high-dimensional regimes (e.g., $p \gg n$), especially when computational shortcuts like SPCs are employed. Comparisons with other Bayesian selection methods under these conditions would also provide a more comprehensive understanding of PPVS's strengths and limitations. Additionally, more research and tutorial-style work is needed to refine and guide the automation of key workflow steps—from fitting the reference model to selecting submodel sizes—to improve accessibility for practitioners.

Acknowledgments

I would like to express my sincere gratitude to my supervisors Dr. Sara van Erp and PhD researcher Florian van Leeuwen for their invaluable supervision, guidance, patience, and support throughout this project. I also want to thank my fellow students from the program for their encouragement and for fostering a supportive atmosphere. In addition, I am grateful to the university for organizing regular check-ins, which provided structure and motivation during the process. I cannot imagine completing this work without all the support I received.

7 References

Babyak, M. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic medicine*, *66*, 411–21. https://doi.org/10.1097/01.psy.0000127692.23278.a9

Bainter, S. A., McCauley, T., Fahmy, M., & Attali, D. (2024). *Ssvs: Functions for stochastic search variable selection (ssvs)* [R package version 2.0.0]. https://github.com/sabainter/ssvs

Bainter, S. A., McCauley, T. G., Fahmy, M. M., Goodman, Z. T., Kupis, L. B., & Rao, J. S. (2023). Comparing bayesian variable selection to lasso approaches for applications in psychology. *Psychometrika*, *88*(3), 1032–1055. https://doi.org/10.1007/s11336-023-09914-9

Bair, E., Hastie, T., Paul, D., & Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association*, *101*, 119–137. https://doi.org/10.1198/016214505000000628

Barbieri, M. M., & Berger, J. O. (2004). Optimal predictive model selection [Publisher: Institute of Mathematical Statistics]. *The Annals of Statistics*, *32*(3), 870–897. https://doi.org/10.1214/009053604000000238

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017

Bürkner, P.-C. (2021). Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software*, *100*(5), 1–54. https://doi.org/10.18637/jss.v100.i05

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals [Publisher: [Oxford University Press, Biometrika Trust]]. *Biometrika*, *97*(2), 465–480. Retrieved October 6, 2024, from https://www.jstor.org/stable/25734098

Catalina, A., Bürkner, P., & Vehtari, A. (2021, September 10). Latent space projection predictive inference. https://doi.org/10.48550/arXiv.2109.04702

Dupuis, J., & Robert, C. (2003). Variable selection in qualitative models via an entropic explanatory power. *Journal of Statistical Planning and Inference*, *111*, 77–94. https://doi.org/10.1016/S0378-3758(02)00286-0

Friendly, M. (2024). *Genridge: Generalized ridge trace plots for ridge regression* [R package version 0.8.0]. https://github.com/friendly/genridge

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization in bayesian workflow. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *182*(2), 389–402. https://doi.org/10.1111/rssa.12378

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences [Publisher: Institute of Mathematical Statistics]. *Statistical Science*, *7*(4), 457–472. Retrieved March 27, 2025, from https://www.jstor.org/stable/2246093

George, E., & McCulloch, R. (1993). Variable selection via gibbs sampling. *Journal of The American Statistical Association - J AMER STATIST ASSN*, *88*, 881–889. https://doi.org/10.1080/01621459.1993.10476353

Goutis, C., & Robert, C. P. (1998). Model choice in generalised linear models: A bayesian approach via kullback-leibler projections. *Biometrika*, *85*(1), 29–37. https://doi.org/10.1093/biomet/85.1.29

Hoerl, A. E., & Kennard, R. W. (2000). Ridge regression: Biased estimation for nonorthogonal problems [Publisher: [Taylor & Francis, Ltd., American Statistical Association, American Society for Quality]]. *Technometrics*, *42*(1), 80–86. https://doi.org/10.2307/1271436

Hsiang, T. C. (1975). A bayesian view on ridge regression [Publisher: [Royal Statistical Society, Wiley]]. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *24*(4), 267–268. https://doi.org/10.2307/2987923

Ishwaran, H., & Rao, J. S. (2005). Spike and slab variable selection: Frequentist and bayesian strategies [Publisher: Institute of Mathematical Statistics]. *The Annals of Statistics*, *33*(2), 730–773. Retrieved October 1, 2024, from https://www.jstor.org/stable/3448605

Johnson, R. W. (1996). Fitting percentage of body fat to simple body measurements [Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10691898.1996.11910505]. *Journal of Statistics Education*, *4*(1). https://doi.org/10.1080/10691898.1996.11910505

McLatchie, Y., Rögnvaldsson, S., Weber, F., & Vehtari, A. (2024, August 6). Advances in projection predictive inference. https://doi.org/10.48550/arXiv.2306.15581

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression [Publisher: [American Statistical Association, Taylor & Francis, Ltd.]]. *Journal of the American Statistical Association*, *83*(404), 1023–1032. https://doi.org/10.2307/2290129

Park, T., & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, *103*, 681–686. https://doi.org/10.2307/27640090

Pavone, F., Piironen, J., Bürkner, P.-C., & Vehtari, A. (2023). Using reference models in variable selection. *Computational Statistics*, *38*(1), 349–371. https://doi.org/10.1007/s00180-022-01231-6

Piironen, J., Paasiniemi, M., Catalina, A., Weber, F., & Vehtari, A. (2023). projpred: Projection predictive feature selection [R package version 2.8.0]. https://mc-stan.org/projpred/

Piironen, J., Paasiniemi, M., & Vehtari, A. (2020). Projective inference in high-dimensional problems: Prediction and feature selection. *Electronic Journal of Statistics*, *14*(1). https://doi.org/10.1214/20-EJS1711

Piironen, J., & Vehtari, A. (2017a). Comparison of bayesian predictive methods for model selection. *Statistics and Computing*, *27*(3), 711–735. https://doi.org/10.1007/s11222-016-9649-y

Piironen, J., & Vehtari, A. (2017b). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, *11*(2). https://doi.org/10.1214/17-EJS1337SI

Piironen, J., & Vehtari, A. (2018, September). Iterative supervised principal components. In A. Storkey & F. Perez-Cruz (Eds.), *Proceedings of the twenty-first international conference on artificial intelligence and statistics* (pp. 106–114, Vol. 84). PMLR. https://proceedings.mlr.press/v84/piironen18a.html

Sivula, T., Magnusson, M., Matamoros, A. A., & Vehtari, A. (2023, October 21). Uncertainty in bayesian leave-one-out cross-validation based model comparison. https://doi.org/10.48550/arXiv.2008.10296

Stamey, T. A., Kabalin, J. N., McNeal, J. E., Johnstone, I. M., Freiha, F., Redwine, E. A., & Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. radical prostatectomy treated patients. *The Journal of Urology*, *141*(5), 1076–1083. https://doi.org/10.1016/s0022-5347(17)41175-x

Stan Development Team. (2023). Stan modeling language users guide and reference manual, version 2.33. https://mc-stan.org

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso [Publisher: [Royal Statistical Society, Oxford University Press]]. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288. Retrieved October 1, 2024, from https://www.jstor.org/stable/2346178

Tran, M.-N., Nott, D. J., & Leng, C. (2012). The predictive lasso. *Statistics and Computing*, *22*(5), 1069–1084. https://doi.org/10.1007/s11222-011-9279-3

van Erp, S., Oberski, D. L., & Mulder, J. (2019). Shrinkage priors for bayesian penalized regression. *Journal of Mathematical Psychology*, *89*, 31–50. https://doi.org/10.1016/j.jmp.2018.12.004

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\widehat{r}$ for assessing convergence of MCMC. *Bayesian Analysis*, *16*(2). https://doi.org/10.1214/20-BA1221

Weber, F., Glass, Ä., & Vehtari, A. (2024). Projection predictive variable selection for discrete response families with finite support. *Computational Statistics*. https://doi.org/10.1007/s00180-024-01506-0

Zou, H., & Hastie, T. (2005). Zou h, hastie t. regularization and variable selection via the elastic net. j r statist soc b. 2005;67(2):301-20. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

A Appendix

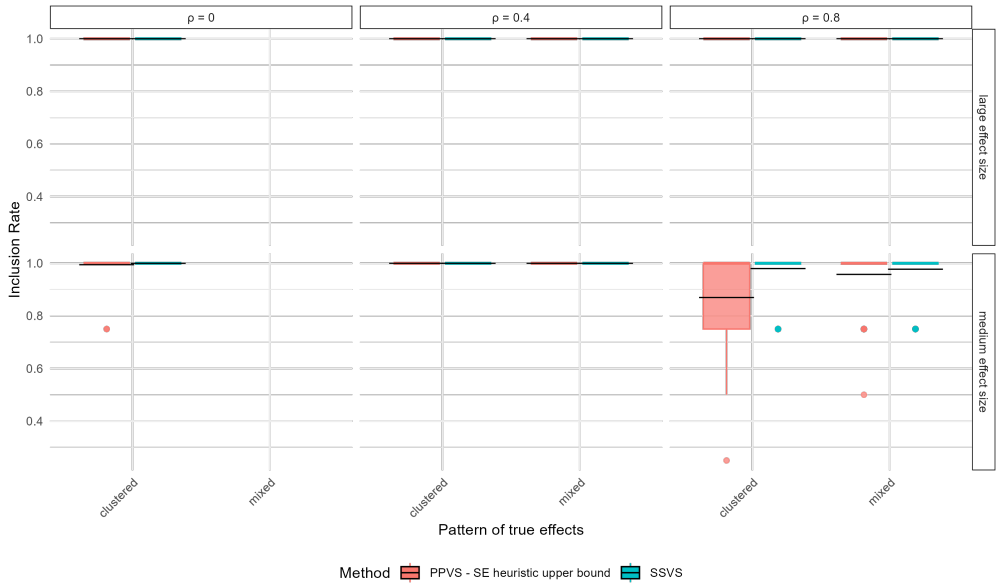

Figure 1: True inclusion rates for predictors for medium and large effect sizes across simulation conditions. The black horizontal bars denote the means. The two methods compared are projection predictive variable selection (PPVS), using the SE heuristic (McLatchie et al., 2024) with the upper confidence interval bound, and stochastic search variable selection (SSVS). Inclusion rate is computed as the proportion of true predictors (with medium and large effects) that were selected. The panels for the mixed true effect pattern at $\rho = 0$ are empty because the spatial distribution of true effects is irrelevant in the absence of correlation; only the clustered pattern was included for this setting.

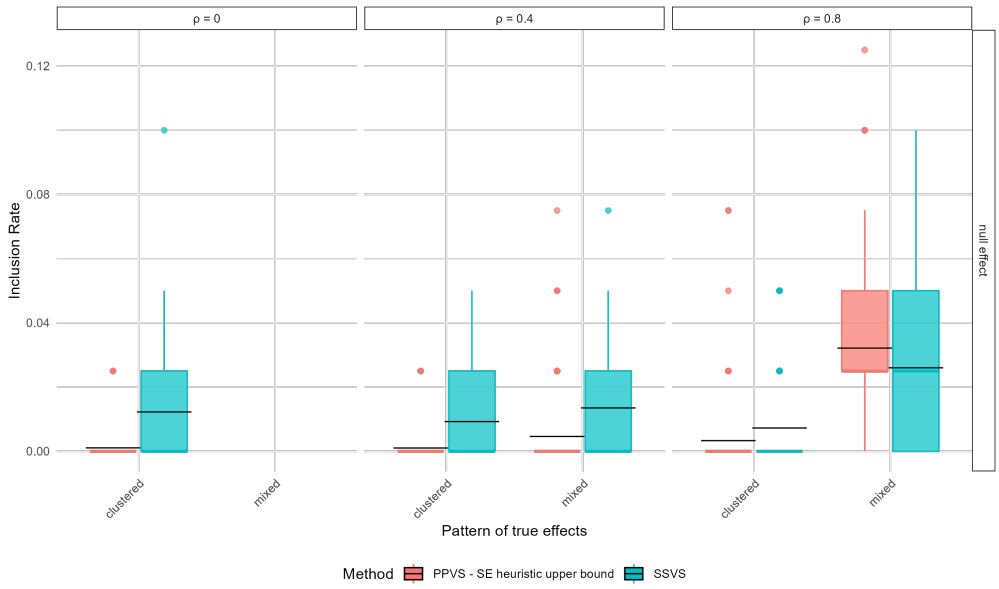

Figure 2: True inclusion rates for predictors for medium and large effect sizes across simulation conditions. The black horizontal bars denote the means. The two methods compared are projection predictive variable selection (PPVS), using the SE heuristic (McLatchie et al., 2024) with the upper confidence interval bound, and stochastic search variable selection (SSVS). Inclusion rate is computed as the proportion of true predictors (with null effects) that were selected. The panels for the mixed true effect pattern at $\rho = 0$ are empty because the spatial distribution of true effects is irrelevant in the absence of correlation; only the clustered pattern was included for this setting.

Figure 3: Inclusion rates across conditions in Study 2. The black horizontal bars denote the means. The two methods compared are projection predictive variable selection (PPVS), using the SE heuristic (McLatchie et al., 2024) with the upper confidence interval bound, and stochastic search variable selection (SSVS). Inclusion rate is computed as the proportion of true predictors (with medium and large effects) that were selected.

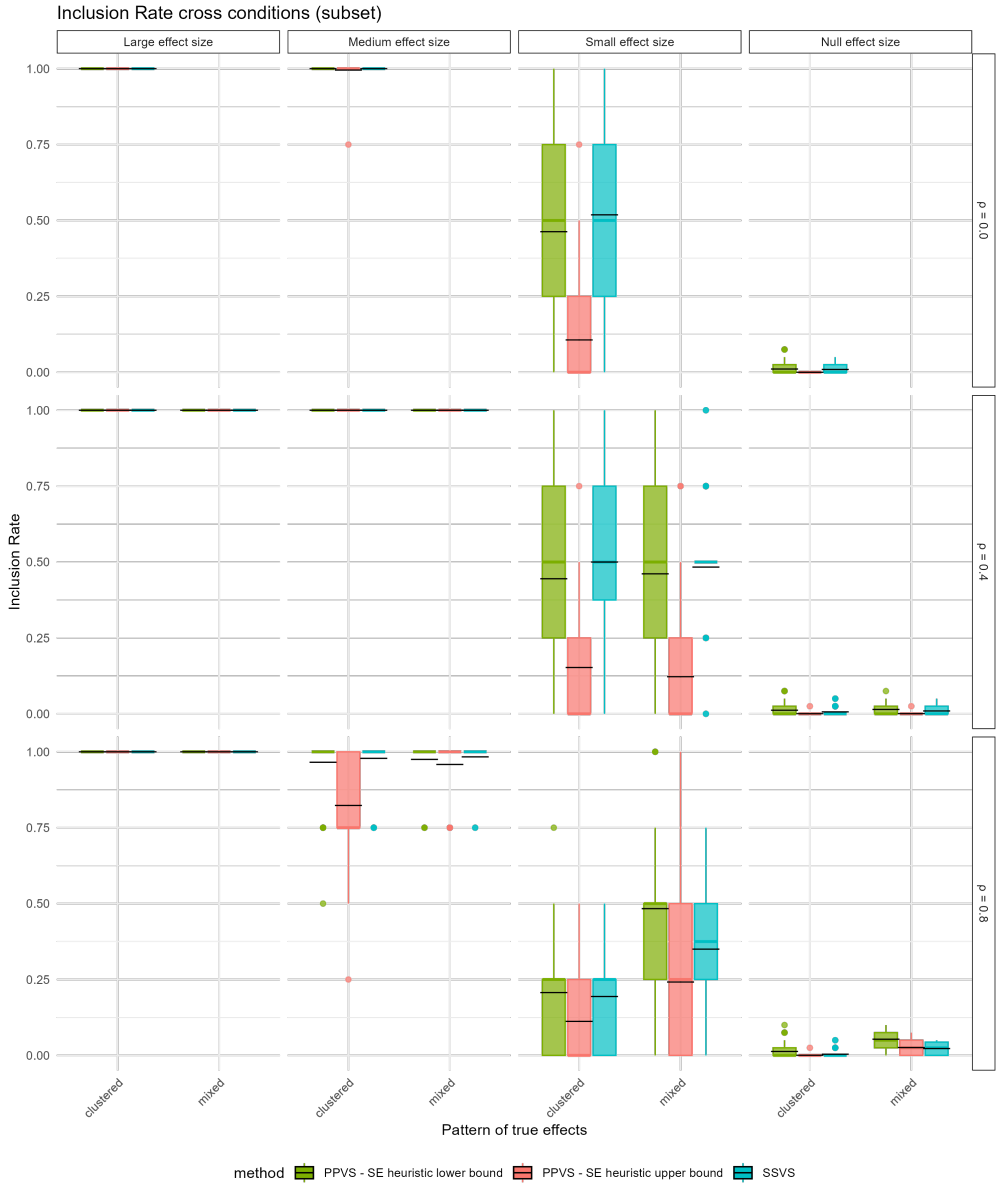

Figure 4: Inclusion rates across conditions in Study 1. The black horizontal bars denote the means. The three methods compared are projection predictive variable selection (PPVS), using the SE heuristic (McLatchie et al., 2024) with both the upper and lower confidence interval bounds, and stochastic search variable selection (SSVS). Only replications where the SE heuristic lower bound returned a value successfully were included. Inclusion rate is computed as the proportion of true predictors (with medium and large effects) that were selected. The panels for the mixed true effect pattern at $\rho = 0$ are empty because the spatial distribution of true effects is irrelevant in the absence of correlation; only the clustered pattern was included for this setting.

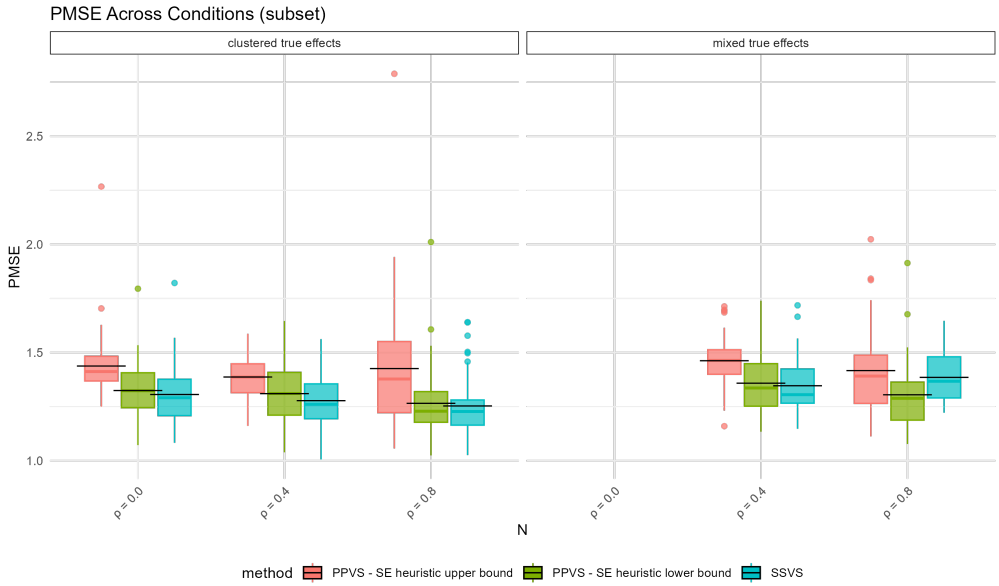

Figure 5: Distributions of the prediction mean squared error (PMSE) across simulation conditions in Study 1. Each panel shows the distribution of PMSE values across replications for one combination of correlation level (ρ), sample size (N), and spatial pattern of true effects (clustered or mixed). The three methods compared are projection predictive variable selection (PPVS), using the SE heuristic (McLatchie et al., 2024) with both the upper and lower confidence interval bounds, and stochastic search variable selection (SSVS). Only replications where the SE heuristic lower bound returned a value successfully were included. Horizontal lines indicate the mean PMSE for each method within a condition. The panel for the mixed true effect pattern at $\rho = 0$ is empty because the spatial distribution of true effects is irrelevant in the absence of correlation; only the clustered pattern was included for this setting.

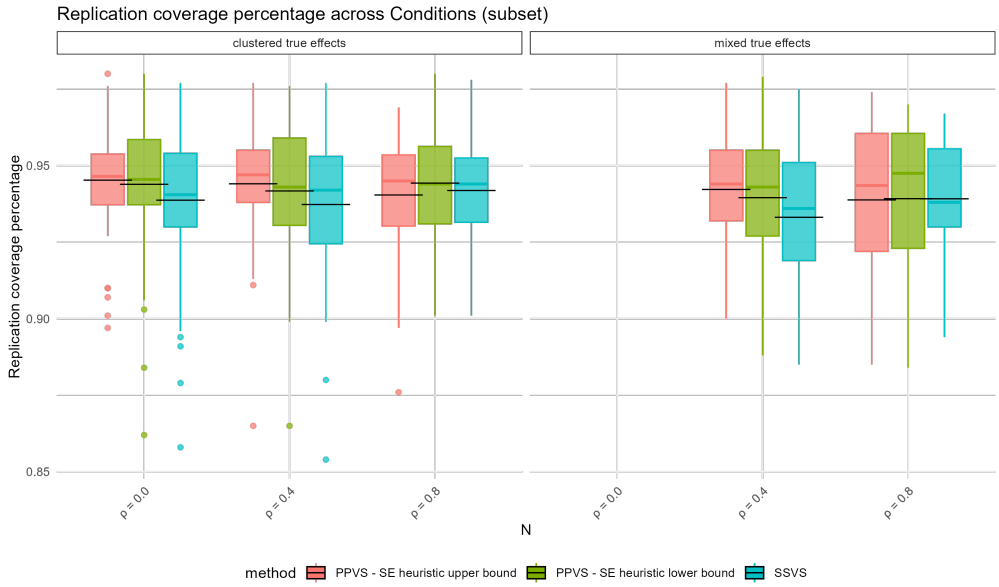

Figure 6: Distributions of coverage percentages across conditions in Study 1. Each curve shows the distribution across replications of the proportion of test observations whose 95% credible interval contained the true value. Dashed vertical lines indicate the mean coverage for each method, while the solid black line marks the nominal 95% target. The three methods compared are projection predictive variable selection (PPVS), using the SE heuristic (McLatchie et al., 2024) with both the upper and lower confidence interval bounds, and stochastic search variable selection (SSVS). Only replications where the SE heuristic lower bound returned a value successfully were included. The panel for the mixed true effect pattern at $\rho = 0$ is empty because the spatial distribution of true effects is irrelevant in the absence of correlation; only the clustered pattern was included for this setting.

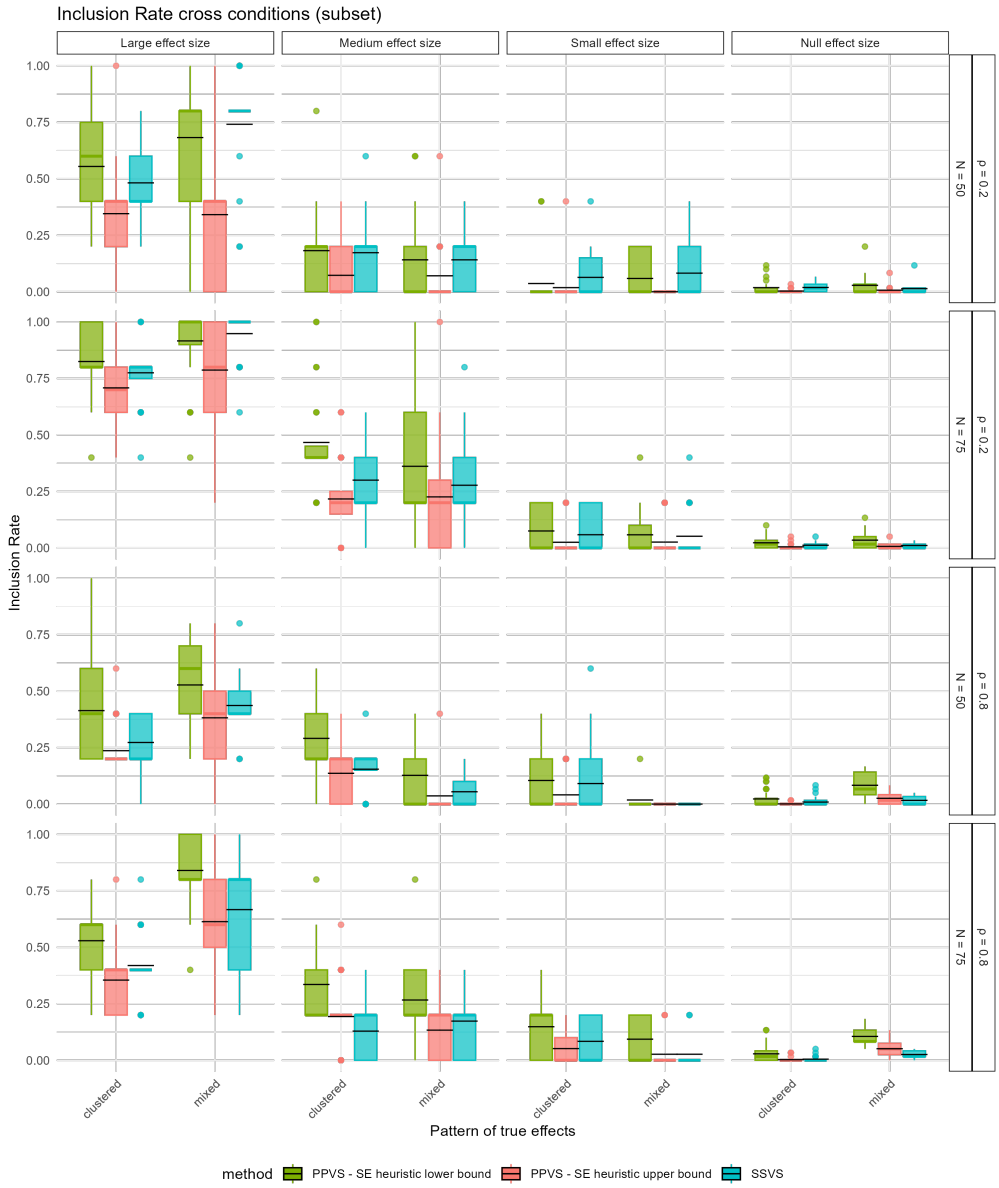

Figure 7: Inclusion rates across conditions in Study 2. The black horizontal bars denote the means. The three methods compared are projection predictive variable selection (PPVS), using the SE heuristic (McLatchie et al., 2024) with both the upper and lower confidence interval bounds, and stochastic search variable selection (SSVS). Only replications where the SE heuristic lower bound returned a value successfully were included. Inclusion rate is computed as the proportion of true predictors (with medium and large effects) that were selected.

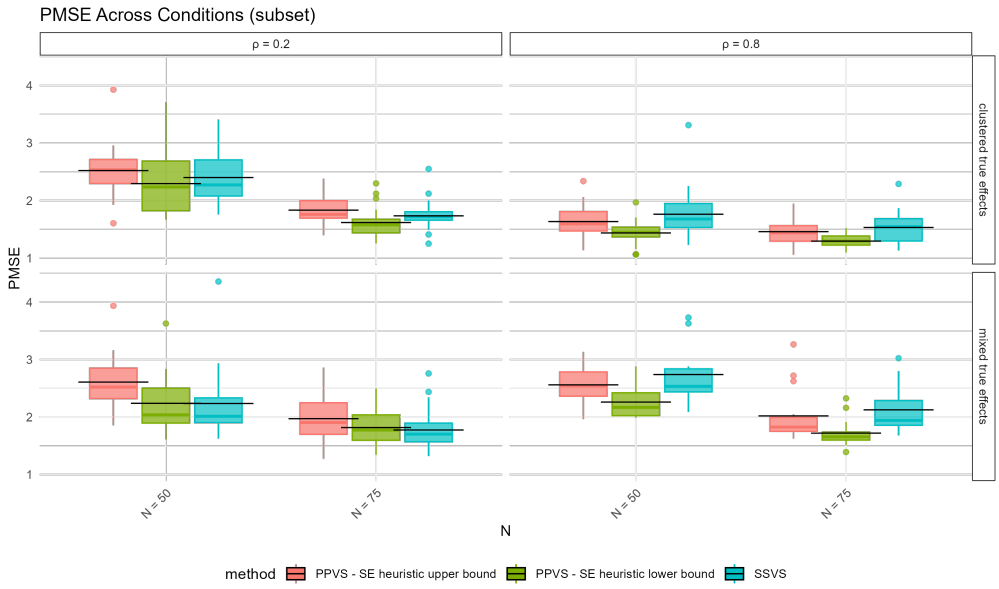

Figure 8: Distributions of the prediction mean squared error (PMSE) across simulation conditions in Study 2. Each panel shows the distribution of PMSE values across replications for one combination of correlation level (ρ), sample size (N), and spatial pattern of true effects (clustered or mixed). The three methods compared are projection predictive variable selection (PPVS), using the SE heuristic (McLatchie et al., 2024) with both the upper and lower confidence interval bounds, and stochastic search variable selection (SSVS). Only replications where the SE heuristic lower bound returned a value successfully were included. Horizontal lines indicate the mean PMSE for each method within a condition.

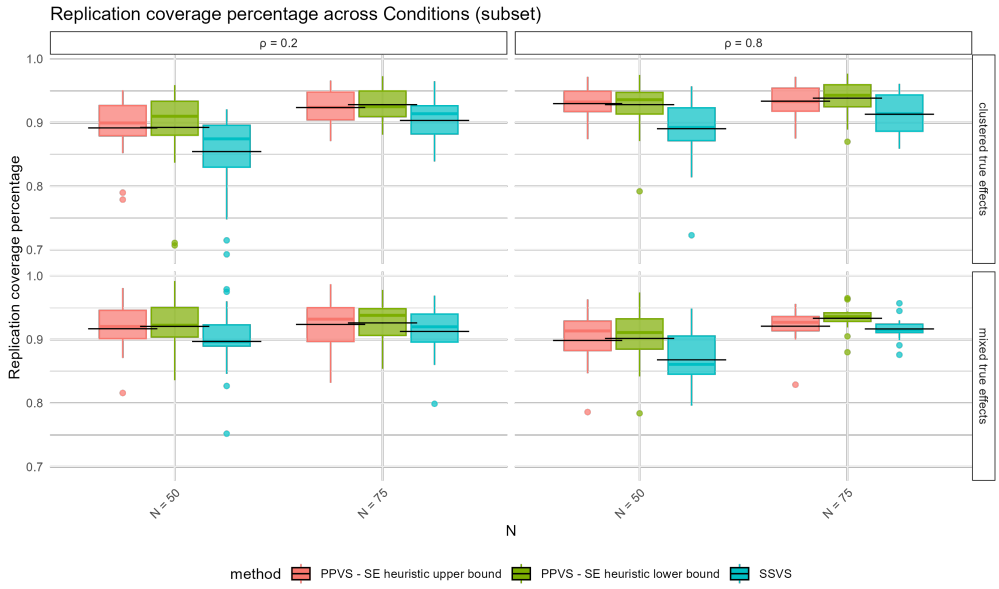

Figure 9: Distributions of coverage percentages across conditions in Study 2. Each curve shows the distribution across replications of the proportion of test observations whose 95% credible interval contained the true value. Dashed vertical lines indicate the mean coverage for each method, while the solid black line marks the nominal 95% target. The three methods compared are projection predictive variable selection (PPVS), using the SE heuristic (McLatchie et al., 2024) with both the upper and lower confidence interval bounds, and stochastic search variable selection (SSVS). Only replications where the SE heuristic lower bound returned a value successfully were included.