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

Research Report

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

Zhipei (Kim) Wang (5041333)

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

Utrecht University

Word count: 2447

January 18, 2025

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 exceeds the number of observations. This "small n, large p" dilemma introduces several challenges. Notably, 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).

To address this, various regularization methods have been developed within both frequentist and Bayesian frameworks. Frequentist statisticians typically approach it by adding a penalty term to the loss function (e.g. Hoerl & Kennard, 2000; Zou & Hastie, 2005), which shrinks parameter estimates and potentially sets them to zero (Tibshirani, 1996), thereby performing variable selection. Bayesian statisticians use shrinkage priors that aim to shrink small effects to zero, while leaving the substantial effects intact. 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 such as ridge (Hsiang, 1975) and LASSO (Park & Casella, 2008), which are the Bayesian counterparts of frequentist penalties. After fitting the model, summaries of the posterior distribution are used to determine whether a predictor should remain in the model, for example based on thresholds for posterior point summaries or 95% credibility 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 Markov chain Monte Carlo (MCMC) samples of its coefficient were assigned to the slab, with a threshold often involved often 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 (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 giving good prediction accuracy, and then identifies a minimal subgroup of predictors that can approximate this accuracy.

Given that the PPVS method is relatively new, and its performance is not investigated in many scenarios, we aim to implement it under different conditions and evaluate its performance. To benchmark its effectiveness, we compare it against 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 a simulation study to compare their performance and discuss the results. Finally, we examine the limitations of the current simulation study and propose potential future research directions.

2 Methods

In Bayesian analysis, parameters are considered random variables with prior distributions that encapsulate prior beliefs about their values, unlike the frequentist approach that treats parameters as fixed but unknown. These beliefs are updated through the likelihood of the observed data, leading to posterior distributions that reflect updated beliefs post-data.

This section delves into the specific Bayesian variable selection methods employed in this study: Stochastic Search Variable Selection (SSVS) and Projection Predictive Variable Selection (PPVS). Each method leverages distinct approaches to address the challenges of high-dimensional data analysis, particularly the identification and inclusion of significant predictors.

2.1 Stochastic search variable selection

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 i^{th} observation of the j^{th} 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 their marginal distributions. For the SSVS method, this is achieved using the spike-and-slab prior.

The spike-and-slab prior (George & McCulloch, 1993; Mitchell & Beauchamp, 1988) falls into the category of discrete mixture priors. The idea is that it is a combination of a point mass at 0 (the spike), and a diffused distribution (the slab). Such a prior can often be written as a mixture of two normal distributions (Piironen et al., 2020)

$$\beta_j \mid \lambda_j, c, \varepsilon \sim \lambda_j \mathrm{N}(0, c^2) + (1 - \lambda_j)\mathrm{N}(0, \varepsilon^2),$$

$$\lambda_j \mid \pi \sim \text{Bernoulli}(\pi), \quad j = 1, \ldots, p.$$

Here, ε is typically much smaller than c, with $\lambda_j = 1$ indicating that the coefficient $\beta_j$is drawn from the slab, and $\lambda_j = 0$ indicating it is drawn from the spike. The spike can either be formulated to be very peaked at zero (Mitchell & Beauchamp, 1988) by setting ε to zero, or to be highly concentrated around zero (George & McCulloch, 1993) by setting ε to a small positive value. The prior probability π (typically 0.5) of including the predictor or not is included so a binary decision about the inclusion of a predictor can be made. A hyperprior can also be set for π by assigning it for example a standard uniform distribution (Ishwaran & Rao, 2005).

The SSVS method utilizes spike-and-slab priors for variable selection. 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 & Berger, 2004).

2.2 Projection predictive variable selection

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. McLatchie et al. (2024) proposed an efficient projection predictive workflow, which is briefly reviewed as follows.

Initially, a reference model is recommended to be built with a sparsifying prior using all the predictors (Piironen et al., 2020). The regularized horseshoe prior was specifically chosen for use in the simulation study's PPVS component due to its nature as a continuous mixture of normal distributions. This is in contrast to the spike-and-slab prior, which is a discrete mixture. Both of these priors represent the state-of-the-art shrinkage priors, each offering a distinct approach to regularization. Additionally, in high-dimensional settings, preliminary feature screening or supervised principal component analysis is recommended to reduce computational demands, and a useful technique to combine those two is known as supervised principal components (Bair et al., 2006).

The solution path is then established, detailing which predictors are selected at each step as the regularization process unfolds. Two primary search approaches are utilized: KL divergence-based forward search and Lasso regularization, with the latter noted for its computational efficiency but higher variability across different data realizations (McLatchie et al., 2024).

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 it 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, that intermediates between the draw-by-draw (Dupuis & Robert, 2003; Goutis & Robert, 1998) 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 heuristic for determining an upper bound on the submodel size.

Ultimately, the optimal submodel size is determined based on its predictive closeness to the reference model. McLatchie et al. (2024) suggested two primary heuristics for this decision: one based on differential utility intervals, as discussed in several studies (Catalina et al., 2021; Piironen et al., 2020; Piironen et al., 2023; Piironen & Vehtari, 2017; Weber et al., 2024), and another that relies on the mean expected log point-wise predictive density (elpd) difference compared to the reference model (Sivula et al., 2023). The final step involves projecting the reference model onto the chosen submodel and using the projected posterior for inference as in standard MCMC analyses.

3 Simulation Study

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, adopted from Bainter et al. (2023), consisted of 100 observations and 50 predictors, among which 10 had true effects. The predictors X were generated from a multivariate normal distribution with diagonal variances of 1 and block diagonal covariance structures. In the condition with correlated true effects, sets of five predictors exhibited high pairwise correlations of .8. The regression coefficients for this scenario, β was defined as $\beta = (1.5, .9, .9, .3, .3, 1.5, .9, .9, .3, .3, 0, ..., 0)'$, grouping the impactful predictors into two clusters.

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 $\beta = (1.5, 0, 0, 0, 0, .9, 0, ..., 0)'$. This modification ensured that the effects were isolated rather than clustered.

The response vector Y was drawn from $N(\mathbf{X}\beta, \sigma^2\mathbf{I})$, 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.

The simulated datasets used in this study were generously provided by Bainter et al. (2023). Due to time constraints in this preliminary analysis, we randomly sampled 100 replications from the 500 available replications for each condition to evaluate the two methods.

3.1 Computational details

3.1.1 Stochastic search variable selection specifications

Each iteration for SSVS was conducted using the `ssvs()` function from the SSVS package (Bainter et al., 2024). The function was executed with the default prior settings where the spike is a point mass at zero. We ran one MCMC chain for each simulation, consisting of 20,000 iterations, with the initial 5,000 iterations discarded as warm-up.

3.1.2 Projection predictive variable selection specifications

For PPVS, each iteration involved fitting a Bayesian multiple regression model using the function `horseshoe()` in the package `brms` (Bürkner, 2017, 2018, 2021), which is based on Stan (Stan Development Team, 2023).

We configured the true predictor-to-total predictor ratio parameter par_ratio at 0.2 to match the data-generating mechanism. The model was run across 4 chains with 2,000 iterations each, discarding 1,000 iterations per chain for 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, as we knew the number of true predictors to be 10 (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 for `suggest_size()` were explored: the default setting, "type_lower" where the type argument was set to lower, "thres_lower" combining a threshold for the elpd difference of -4 with type set to lower, and "thres_upper" with the same elpd difference threshold but type set to upper. Finally, the variables selected by each heuristic for every replication were saved.

3.2 Measures of performance

Given our primary interest in variable selection accuracy, we concentrated our comparison of the two methods on their performance in this domain. Specifically, we calculated inclusion rates stratified by different effect sizes for both methods (Figure 1 visualizing the overall inclusion rates unstratified can be found in the Appendix). Inclusion rates quantify the proportion of true predictors correctly identified by the methods, differentiated by effect sizes. Therefore, the inclusion rates for small, medium, and large effects represent true inclusion rates, whereas those for null effect size serve as false inclusion rates. For SSVS, predictors were considered included if their marginal inclusion probabilities (MIPs) exceeded a threshold of 0.5, following the approach used in Bainter et al. (2023), which is deemed optimal under specific conditions (Barbieri & Berger, 2004). For PPVS, four different heuristics were employed in the variable selection process, resulting in a comparison across five distinct methods.

Table 1: Inclusion rates for PPVS and SSVS under correlated true predictors

Effect Size	PPVS (%)				SSVS (%)
	Default	Type_lower	Thres_Upper	Thres_Lower	
Null	0.3	**1.7**	0.2	0.2	0.8
Small	12.3	25.5	11.8	12.3	**27.8**
Medium	88.6	97.8	88.6	89.1	**98.3**
Large	100.0	100.0	100.0	100.0	100.0

Table 2: Inclusion rates for PPVS and SSVS under uncorrelated true predictors

Effect Size	PPVS (%)				SSVS (%)
	Default	Type_lower	Thres_Upper	Thres_Lower	
Null	4.6	**7.3**	4.3	4.6	4.7
Small	20.5	**30.7**	17.0	17.1	27.5
Medium	100.0	100.0	100.0	100.0	99.0
Large	100.0	100.0	100.0	100.0	100.0

4 Results

4.1 Variable selection accuracy

We examined the inclusion rates for both methods, PPVS across four heuristics, and SSVS, under conditions of correlated and uncorrelated true predictors. The results are presented separately Table 1 for correlated predictors and Table 2 for uncorrelated predictors. The inclusion rates for small, medium, and large effect sizes are thus the true inclusion rates, while the inclusion rates for the null effect size are false inclusion rates.

In the scenario of correlated true predictors detailed in Table 1, both methods displayed high inclusion rates for large effects. For medium effects, the inclusion rate for SSVS was slightly higher at 98.3%, compared with 97.8% for the best performing PPVS heuristic (type_lower). For small effects, SSVS continued to perform slightly better than PPVS, achieving an inclusion rate of 27.8% compared to 25.5% for the heuristic type_lower. The inclusion rates for null effects were significantly lower under correlated conditions, which is expected as the presence of correlation likely reduces the number of predictors with negligible coefficients being incorrectly included.

With uncorrelated true effects as shown in Table 2, both PPVS and SSVS achieved inclusion rates of 100% for medium and large effects, irrespective of the heuristic used by PPVS. This demonstrates robust performance across both methods when effects are significant. For smaller and null effect sizes, the results varied more significantly between the heuristics. Specifically, the false inclusion rate, or the inclusion rate for null effects, was highest using the type_lower heuristic at 7.3% for PPVS, compared to 4.7% for SSVS. Meanwhile, for small effects, SSVS generally showed superior performance compared to most PPVS heuristics, except the type_lower heuristic which closely matched SSVS at 30.7%.

These findings underscore the effectiveness of SSVS in managing both small and null effects more efficiently than PPVS, particularly when using the type_lower heuristic. Although this heuristic is competitive, it exhibits a higher propensity for false inclusions, suggesting that its use should be carefully considered. The analysis clearly highlights the strengths and limitations of each method and heuristic, providing essential insights for their application in statistical modeling where the true predictor structure is unknown.

5 Discussion

Overall, SSVS demonstrated consistent and robust performance across various conditions compared to PPVS, especially for smaller effect sizes where it generally outperformed PPVS. Although PPVS occasionally exceeded SSVS with heuristic type_lower, it was at the expense of higher false inclusion rates.

Notably, the `suggest_size()` function encountered numerous failures with some PPVS heuristics, such as returning 78 NAs for heuristic type_lower in the uncorrelated condition, potentially impacting the reliability of the mean inclusion rates reported. Future analysis could explore why this function failed and consider adjusting the maximum number of terms to search, which could help streamline the workflow for extensive replications.

The computational cost also plays a significant role; the PPVS workflow typically takes 8 to 10 minutes per run, whereas SSVS usually completes in less than a minute, highlighting the efficiency of SSVS.

Despite operating in a relatively low-dimensional space with 50 predictors and 10 true effects against a sample size of 100, PPVS is known to excel in high-dimensional settings. Moving forward, we plan to extend the simulation to higher dimensions to assess if PPVS's performance merits the increased computational demand.

6 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*, 411–421. https://doi.org/10.1097/01.psy.0000127692.23278.a9

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

Bainter, S. A., McCauley, T., Fahmy, M., & Attali, D. (2024). *SSVS: Functions for stochastic search variable selection (SSVS)*. https://github.com/sabainter/ssvs

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. *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

Catalina, A., Bürkner, P., & Vehtari, A. (2021). *Latent space projection predictive inference* (arXiv:2109.04702). arXiv. 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

Erp, S. van, 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

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. *Technometrics*, *42*(1), 80–86. https://doi.org/10.2307/1271436

Hsiang, T. C. (1975). A bayesian view on ridge regression. *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. *The Annals of Statistics*, *33*(2), 730–773. https://www.jstor.org/stable/3448605

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

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *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

Piironen, J., Paasiniemi, M., Catalina, A., Weber, F., & Vehtari, A. (2023). *projpred: Projection predictive feature selection*. 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. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, *11*(2). https://doi.org/10.1214/17-EJS1337SI

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

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. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288. https://www.jstor.org/stable/2346178

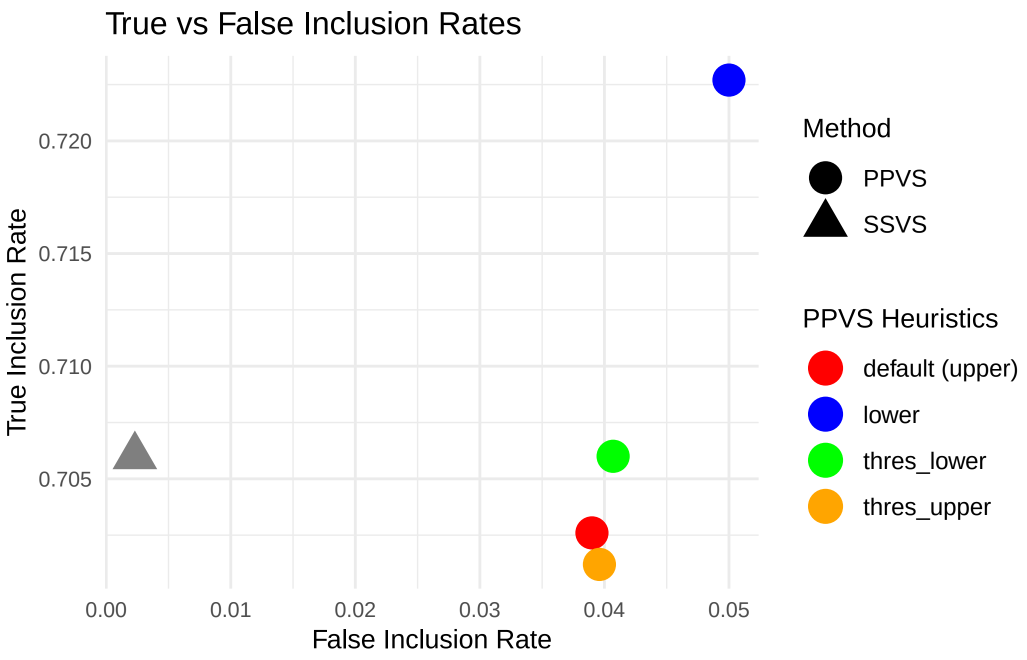
Tran, M.-N., Nott, D. J., & Leng, C. (2012). The predictive lasso. *Statistics and Computing*, *22*(5), 1069–1084. https://doi.org/10.1007/s11222-011-9279-3

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

7 Appendix

Figure 1: Overall true vs false inclusion rates by method


True vs False Inclusion Rates
True Inclusion Rate
False Inclusion Rate
Method
PPVS
SSVS
PPVS Heuristics
default (upper)
lower
thres_lower
thres_upper