ORIGINAL PAPER

Using reference models in variable selection

Federico Pavone[1,2] · **Juho Piironen**[2] · **Paul-Christian Bürkner**[2] · **Aki Vehtari**[2]

Abstract

Variable selection, or more generally, model reduction is an important aspect of the statistical workflow aiming to provide insights from data. In this paper, we discuss and demonstrate the benefits of using a reference model in variable selection. A reference model acts as a noise-filter on the target variable by modeling its data generating mechanism. As a result, using the reference model predictions in the model selection procedure reduces the variability and improves stability, leading to improved model selection performance. Assuming that a Bayesian reference model describes the true distribution of future data well, the theoretically preferred usage of the reference model is to project its predictive distribution to a reduced model, leading to projection predictive variable selection approach. We analyse how much the great performance of the projection predictive variable is due to the use of reference model and show that other variable selection methods can also be greatly improved by using the reference model as target instead of the original data. In several numerical experiments, we investigate the performance of the projective prediction approach as well as alternative variable selection methods with and without reference models. Our results indicate that the use of reference models generally translates into better and more stable variable selection.

Keywords Model reduction · Projection predictive approach · Bayesian statistics

1 Introduction

In statistical applications, one of the main steps in the modelling workflow is variable selection, which is a special case of model reduction. Variable selection (also known as feature or covariate selection) may have multiple goals. First, if the variables themselves are of interest, we can use variable selection to infer which variables contain predictive information about the target. Second, as simpler models come

✉ Federico Pavone
  federico.pavone@phd.unibocconi.it

1 Department of Decision Sciences, Bocconi University, Milan, Italy

2 Department of Computer Science, Aalto University, Helsinki, Finland

 🖄 Springer

with the advantages of reduced measurement costs and improved interpretability, we may be interested in finding the minimal subset of variables which still provides good predictive performance (or good balance between simplicity and predictive performance). When the predictive capability is guiding the selection, the true data generation mechanism of future data can be approximated either by using the observed data directly or alternatively by using predictions from a *reference model* (Vehtari and Ojanen 2012).

In data-based approaches, such as Lasso selection (Tibshirani 1996) or stepwise backward/forward regression (Venables and Ripley 2013; Harrell 2015), the observed empirical data distribution is utilised as a proxy of future data, usually in combination with cross-validation or information criteria to provide estimates of out-of-sample predictive performance. In contrast, reference model based methods approximate the future data generation mechanism using the predictive distribution of a reference model, which can be, for example, a full-encompassing model including all variables.

We assume all models are wrong, but we assume we have constructed a model which reflects our beliefs about the future data in the best possible way and which has passed model checking and criticism (Gelman et al. 2020; Gabry et al. 2019). Using the usual best practices for constructing the reference model is important, as using a bad reference model can only lead to selecting a similarly bad smaller model. If the reference model is considered to have useful predictions, then the smaller models selected will also have similar useful predictions. The reference model approach has been used in Bayesian statistics in some form at least since the seminal work of Lindley (1968). For more historical references, see Vehtari and Ojanen (2012) and Piironen and Vehtari (2017a), and for most recent methodological developments see Piironen et al. (2020). Examples of useful reference models can be found for example for small-*n*-large-*p* regression and logistic regression by Piironen and Vehtari (2017a) (with spike-and-slab prior), Piironen and Vehtari (2015) (with horseshoe prior), and Piironen et al. (2020) (with iterative supervised principal components), for generalized linear and additive multilevel models by Catalina et al. (2020), for regression models with non-exponential family observation models by Catalina et al. (2021), and for generic multivariate non-linear regression with higher order interactions by Piironen and Vehtari (2016).

Reference models have been also used in non-Bayesian context. Harrell (2015) describes them as full models that can be thought of as a "gold standard" (for a given application). Faraggi et al. (2001) deal with the necessity of identifying interpretable risk groups in the context of survival data using neural networks, which typically perform very well in terms of prediction, but whose variables are difficult to be understood in terms of relevance. Paul et al. (2008), using the term preconditioning, explore approximating models fitting Lasso or stepwise regression against consistent estimates $\hat{y}$ of a reference model instead of the observed responses y.

All these methods can be framed into the family of reference model approaches. The common denominator is the use of the reference model predictive information instead of simply observed data to guide the selection. Whatever the terminology or applied statistical framework, reference models offer a powerful approach to improving variable selection, as we will demonstrate in the present paper.

The goal of the present paper is to study the impact of reference model approaches by disentangling the benefit of using reference models per se from the benefit of specific variable selection algorithms. In particular, we:

- propose a simple and intuitive approach to combine any variable selection method with the reference model approach. This allows us to investigate the benefit of using reference models independent of the specific variable selection method;
- perform extensive numerical experiments to compare variable selection methods with or without using a reference model, both for complete and minimal subset variable selection and assessing the quality of the selection;
- provide evidence supporting, in particular, the projection predictive approach as a principled way to use reference models in minimal subset variable selection.

The paper is structured as follows. In Sect. 2, we review the concept of the reference model, its benefits with examples and how it can be used as a filter on data in a simple way. In Sects. 3 and 4, we show the benefits of a reference model approaches for minimal and complete variable selection, respectively, before we end with a conclusion in Sect. 5. The code to run all the experiments is available on GitHub.[1]

2 Reference models in variable selection

In this section, we will provide an initial motivation and intuition for the use of reference models to improve variable selection methods. We will start with a case study that is repeatedly used throughout the paper to illustrate the benefits of reference models before we dive deeper into the theoretical reasons why reference models help in variable selection.

2.1 Body fat example: part 1

To motivate the further discussion and experiments, we start by a simple variable selection example using body fat data by Johnson (1996). The same data was used illustrate the variable selection with classic stepwise backward regression by Heinze et al. (2018). We compare the projective prediction approach (*projpred*, Piironen et al. 2020) which uses a reference model, and classic stepwise backward regression (*steplm*). The experiments are implemented in R (R Core Team 2018).

The target variable of interest is the amount of body fat, which is obtained by a complex and expensive procedure consisting in immersing a person in a water tank and carrying out different measurements and computations. Additionally, we have information about 13 variables which are anthropometric measurements (e.g., height, weight and circumference of different body parts). The variables are highly

---

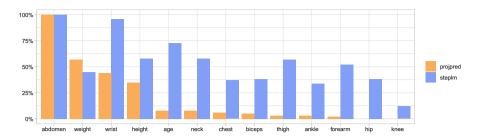[1] https://github.com/fpavone/ref-approach-paper.

Fig. 1 Body fat example: bootstrap inclusion frequencies calculated from 100 bootstrap samples. The projpred approach has less variability on which variables are selected

correlated, which causes additional challenge in the variable selection. In total, we have 251 observations. The goal is to find the model which is able to predict the amount of body fat well while requiring the least amount of measurements for a new person.

Heinze et al. (2018) report results using steplm with a significance level of 0.157 with AIC selection (Akaike 1974), fixing abdomen and height to be always included in the model. For better comparison, we do not fix any of the variables. The steplm approach is carried out combining the `step` and `lm` functions in `R`.

For the selection via projpred, the Bayesian reference model includes all the variables using a regularised horseshoe prior (Piironen and Vehtari 2017b) on the variable coefficients. Submodels are explored using forward search (the results are not sensitive to whether forward or backward search is used), and the predictive utility is the expected log-predictive density (elpd) estimated using approximate leave-one-out cross-validation via Pareto-smoothed importance-sampling (PSIS-LOO-CV; Vehtari et al. 2017). We select the smallest submodel with an elpd score similar to the reference model when taking into account the uncertainty in estimating the predictive model performance. See Appendix A for a brief review of projection predictive approach, and papers by Piironen and Vehtari (2017a) and Piironen et al. (2020) for more details. The complete projpred approach is implemented in the `projpred` R package (Piironen et al. 2019).

The inclusion frequencies of each variable in the final model given 100 bootstrap samples are shown in Fig. 1. In case of projpred there are two variables, 'abdomen' and 'weight', which have inclusion frequencies above 50% ('abdomen' is the only one included always), the third most frequently included is 'wrist' at 44%, and the fourth one is 'height' at 35%. The steplm approach has seven variables with inclusion frequencies above 50%. Such a higher variability and lower stability of steplm can be observed also in the bootstrap model selection frequencies reported in Table 1. For example, the first five selected models have a cumulative frequency of 76% with projpred, but only of 14% with steplm. In addition, the sizes of the selected models with projpred are much smaller than the ones selected with steplm.

The first two rows of Table 2 show the predictive performances, in terms of cross-validated root mean square error (RMSE), of the full model and the selected models using projpred or steplm. There is no significant difference in predictive performance of the selected models by different approaches, even if there is clear difference in the

Table 1 Body fat example: bootstrap model selection frequencies from 100 bootstrap samples

M	Projpred	Freq %	Steplm	Freq %
1	Abdom., weight	39	Abdom., age, forearm, height, hip, neck, thigh, wrist	4
2	Abdom., wrist	10	Abdom., age, chest, forearm, height, neck, thigh, wrist	4
3	Abdom., height	10	Abdom., forearm, height, neck, wrist	2
4	Abdom., height, wrist	9	Abdom., forearm, neck, weight, wrist	2
5	Abdom., weight, wrist	8	Abdom., age, height, hip, thigh, wrist	2
6	Abdom., chest, height, wrist	2	Abdom., age, height, hip, neck, thigh, wrist	2
7	Abdom., biceps, weight, wrist	2	Abdom., age, ankle, forearm, height, hip, neck, thigh, wrist	2
8	Abdom., height, weight, wrist	2	Abdom., age, biceps, chest, height, neck, wrist	2
9	Abdom., age, wrist	2	Abdom., age, biceps, chest, forearm, height, neck, thigh, wrist	2
10	Abdom., age, height, neck, thigh, wrist	2	Abdom., age, ankle, biceps, weight, wrist	2

The projpred approach has less variability on which variable combinations are selected

Table 2 Body fat example: predictive performances with original data (first two rows) and with extra noisy variables (last two rows) estimated with tenfold cross-validation

Data	Method	RMSE Full	RMSE sel	# Sel (10-cv avg, sd)	# Sel noisy (10-cv avg, sd)
Body fat	Projpred	4.4	4.5	2 (2.3, 0.5)	
	Steplm	4.4	4.5	7 (6.0, 0.9)	
+ Noisy variables	Projpred	4.5	4.5	2 (2.0, 0.0)	0 (0, 0)
	Steplm	5.7	5.1	23 (26, 4.5)	15 (19, 4)

RMSE root mean squared error, *Full* full model, *Sel* selected submodel, *# Sel* total number of selected variables, *# Sel noisy* number of selected noisy variables. 10-cv avg = average over the tenfolds in cross-validation. 10-cv sd = standard deviation over the tenfolds in cross-validation

number of selected variables. This can be explained by high correlation between the variablesm and different combinations can provide similar predictive accuracy.

We repeat the experiment with a modified data set by adding 84 unrelated noisy variables, resulting in 100 variables in total. The last two rows of Table 2 show the cross-validated RMSE, the size of the selected model and the number of selected noisy variables using projpred or steplm. The results show that projpred has similar predictive performance and the same number of selected variables as with the original data, whereas the stepwise regression has worse predictive performance and the number of selected variables is much higher and include a large number of irrelevant variables.

Both projpred and steplm compare a large number of models using either forward or backward search, which can lead to selection induced overfitting, but even with 100 variables, projpred is able to select a submodel with similar performance as the full model. In this example, the two compared methods also differ in other aspects than the usage of a reference model, such as that projpred uses Bayesian inference and steplm uses maximum likelihood estimation. To separate the effect of using a reference model, we show that performance of other variable selection methods, including steplm, can also be improved by using reference models.

2.2 Benefits and costs of using a reference model

A properly designed reference model is able to filter parts of the noise present in the data, and hence to provide an improved and more stable selection process. This holds even if the reference model does not perfectly resemble the true data generating process. Clearly, the reference model approach requires a sensible model. The construction of such a model should follow proper modelling workflow (see, e.g. Gelman et al. 2020). Better predictive models imply better selection results. The goodness of a reference model comes from its predictive ability which should be assessed via proper validation methods. Our analyses indicate that the substantial reduction of variance attributable to noise is usually more important than small potential bias due to model misspecification. When the goal is ranking the variables

Table 3 Computational time (in seconds) of body fat example of Sect. 2.1

Data	Reference model	Search path	Projpred	Step.lm	Ref + Step.lm
Body fat	38	15	53		
		0.08		0.08	
	38	0.08			38
+ Noisy variables	170	78	247		
		12		12	
	170	12			182

The reported times for each each method on the three last columns are sums of the time to fit the reference model (second column) and to explore the search path (third column)

for importance to perform variable selection, it is known that large variance is generally more harmful than small bias (Piironen and Vehtari 2017a).
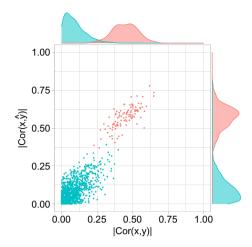
In general, there is no restriction on the type of model the reference models should be, and a sensible reference model does not even need to be Bayesian necessarily. However, Bayesian methods can help in some of those situations where MLE procedures struggle. It is known that as the number of parameters in the model increases, the MLE estimator is dominated by shrinkage estimators (Stein 1956; Stein and James 1961; Parmigiani and Inoue 2009; Efron 2011). The use of a prior in Bayesian inference automatically incorporates some kind of shrinkage in the Bayesian estimator under a given loss function (see, e.g. Rockova et al. 2012). For the sake of our study, we rely on simple data structure which can be well described by linear regression models. However, more complex data can arise in practice, for example, including hierarchical structures. In these cases, MLE inference tends to become cumbersome, whereas the Bayesian framework provides a natural way to convey uncertainties and make inference through the joint posterior distribution of all parameters.

We argue that, regardless of how the reference model is set up and used in the inference procedure, it can be always seen as acting as a filter on the observed data. Furthermore, regardless of what specific model selection method is used, a reference model can be used instead of raw data during the selection process to improve the stability and selection performance. Our results indicate that the core reason why the reference model based methods perform well is the reference model itself, rather than the specific way of using it. In general, the less data we have and the more complex the estimated models are, the higher is the benefit of using a reference models as the basis for variable selection.

If one of the models to be compared is the full model, which can be used as a reference model, there is no additional cost of using a reference model as it was estimated as part of the analysis anyway. Sometimes, including all the available variables in an elaborate model can be computationally demanding. In such a case, even simpler screening or dimensionality reduction techniques, as for example the supervised principal components (Bair et al. 2006; Piironen and Vehtari 2018), can produce useful reference models (Piironen et al. 2020). As an example, Table 3 reports

Fig. 2 Sample correlation plot of each variable (relevant in red, non-relevant in blue) with the target variable y and the latent variable f respectively on the x- and y-axes. Simulated data are generated according to (1) with parameters $n = 70$, $\rho = 0.3$, $k = 100$, $p = 1000$

the computational time for the body fat data example of Sect. 2.1. For reference model approaches, i.e. projpred and ref+step.lm, the main time consuming operation is to fit the reference model. The burden of it depends clearly on the specific application and type of model used.

2.3 Why the reference model helps

A good predictive model is able to filter part of the noise present in the data. The noise is the main source of the instability in the selection and tends to obscure the relevance of the variables in relation to the target variable of interest. We demonstrate it with the following simple explanatory example taken from Piironen et al. (2020). The data generation mechanism is

$$
\begin{aligned}
f &\sim N(0, 1) \\
Y|f &\sim N(f, 1) \\
X_j|f &\overset{iid}{\sim} N(\sqrt{\rho}f, 1 - \rho) \qquad j = 1, \ldots, k \\
X_j|f &\overset{iid}{\sim} N(0, 1) \qquad j = k + 1, \ldots, p,
\end{aligned}
\tag{1}
$$

where f is the latent variable of interest of which Y is a noisy observation. The first k variables are strongly related to the target variable Y and correlated among themselves. Precisely, ρ is the correlation among any pair of the first k variables, whereas $\sqrt{\rho}$ and $\sqrt{\rho/2}$ are the level of correlation between any relevant variable and, respectively, f and Y. If we had an infinite amount of observations, the sample correlation would be equal to the true correlation between X_j and Y. However, even in this ideal asymptotic regime, this correlation would still remain a biased indicator of the true relevance of each variable (represented by the correlation between X_j and f) due to the intrinsic noisy nature of Y.

When using a reference model, we first obtain predictions for f using all the variables $\{X_j\}_{j=1}^p$ taking into account that we have only observed the noisy representation

Y of *f*. If our model is good, we are able to describe *f* better than *Y* itself can, which improves the accuracy in the estimation of the relevance of the variables. Figure 2 illustrates this process in the form of a scatter plot of (absolute) correlations of the variables with *Y* against the corresponding correlations with the predictions of a reference model (in this case, the posterior predictive means of model (7); see Sect. 4.3). Looking at the marginal distributions, we see that using a reference model to filter out noise in the data, the two groups of variables (relevant and non-relevant) can be distinguished much better than when the correlation is computed using the observed noisy data directly.
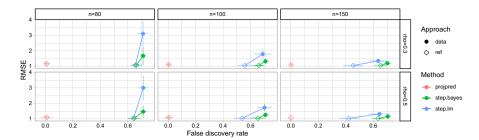
3 Minimal subset variable selection

In the body fat example above, our two simultaneous goals were to obtain good predictive performance and to select a smaller number of variables. When the goal is to select a minimal subset of variables, which have similar predictive performance as the full model, we call it *minimal subset variable selection*. This minimal subset might exclude variables which have some predictive information about the target but, given the minimal subset, these variables are not able to provide such additional information that would improve predictive performance in a substantial manner. The usual reason for this is that the relevant variables which are not in the minimal subset are highly correlated with variables already in the minimal subset. Such nature of the minimal subset makes the solution not unique, except in the particular case of completely orthogonal predictors. The non-uniqueness of the solution is not a problem per se, as different samples of the data are expected to give possible different minimal subsets but with same predictive power. However, it makes it difficult to define a proper concept of stability of the selection. We will return to a problem of finding all the variables with some predictive power in Sect. 4.

3.1 Simulation study 1

Using the data generating mechanism (1), we simulate data sets of different sizes with a relatively large number of variables $p = 70$, with $k = 20$ of them being predictive. We compare the minimal subset variable selection performance of the projection predictive approach (which uses a reference model and it is referred to as projpred), a Bayesian stepwise backward selection with and without a reference model, and maximum likelihood stepwise backward selection with and without a reference model (steplm). The following is a summary of the implementation of the compared methods:

- **projpred**: the projective prediction approach is used. The reference model is a Bayesian linear regression model using the first five supervised principal components (Piironen and Vehtari 2018) as predictors and the full posterior predictive distribution as the basis of the projection. The search heuristic is forward search and the predictive performance is estimated via 10-fold cross-validation.

Fig. 3 Simulation study 1: root mean square error (RMSE) against false discovery rate in the minimal subset variable selection with one standard deviation error bars. Use of a reference model reduces RMSE of the selected model for all the methods. The projpred approach has the smallest RMSE and false discovery ratio
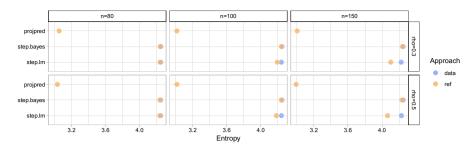
The selection continues until the predictive performance is close to the predictive performance of the reference model. See more details in Appendix A.

- **Bayesian stepwise selection (without a reference model)**: at each step, the fitted model is a Bayesian linear regression using the regularised horseshoe prior and the variable excluded is the one with the highest Bayesian p-value defined as $\min\{P(\theta \leq 0|D), P(\theta > 0|D)\}$, where D stays for the observed data. The selection continues if the reduced model has an elpd score higher (i.e., better) than the current model.
- **Bayesian stepwise selection (with a reference model)**: the reference model is the same as for projpred, but only point predictions (posterior predictive means) $\hat{y}$ are used to replace the target variable y. The same Bayesian p-value selection strategy as in the data based Bayesian stepwise selection is used.
- **steplm (without a reference model)**: stepwise selection using AIC as in the body fat example.
- **steplm (with a reference model)**: the reference model is the same as for projpred, but only point predictions (posterior predictive means) $\hat{y}$ are used to replace the target variable y. The same AIC selection strategy as in the data based steplm is used.
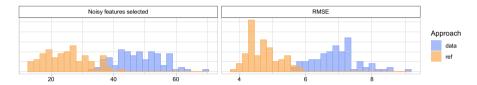
In Fig. 3, the predictive performance is shown for different values of n and ρ in terms of RMSE and the false discovery rate (FDR, the ratio of the number of non-relevant selected variables over the number of selected variables) of the selected submodel, averaged after 100 data simulations. Use of a reference models greatly reduces the overfitting in the selection process, and thus produces much lower test set RMSE of the selected model for all the methods. projpred has the lowest FDR and RMSE. Using a reference model improves also FDR of steplm stepwise Bayesian linear regression significantly.

When the variable selection is repeated with different simulated data sets, there is some variability in the selected variables. We measure the stability of variable selection by computing the entropy of the observed distribution of the included variables over different models. The smallest entropy would be obtained if the approach always selected the same set of variables, and the largest entropy would be observed

Fig. 4 Simulation study 1: entropy score in the minimal subset variable selection. The projpred approach has much smaller entropy score than the other approaches

Fig. 5 Body fat example: stepwise backward selection with and without using a reference model. The x-axis denotes the number of selected irrelevant variables on the left and the out-of-sample RMSE on the right-hand side based on 100 bootstrap samples. The reference approach reduces the number of noisy variables selected and the out-of-sample RMSE

if the approach would always select different sets of variables. Therefore, lower entropy corresponds to a more stable selection. Highly correlated predictive variables may happen to be selected alternately, thus making stability estimation of the selection a non-trivial task. Entropy can not distinguish the interchangeability due to correlation from instability. Thus, such a measure should be considered as a relative, and not as an absolute, measure of stability. Figure 4 shows the entropy scores for the different compared methods. The use of a reference model improves the stability of steplm in variable selection slightly, while it makes little difference for the Bayesian linear regression. The projpred approach turns out to be far more stable than all other methods. This is likely due to projpred being based on a better decision theoretical formulation which (1) takes into account the full predictive distribution and not just a point estimate and (2) projects the reference model posterior to the submodel instead of using a simple refit of submodels.

3.2 Body fat example: part 2

Here, we repeat the selection of Sect. 2.1 via stepwise backward regression. In this case, the overall number of variables (original plus noisy) is 100, as it was in the last part of Sect. 2.1. We compare results with and without using a simple reference model approach outlined in (7) with steplm. Figure 5 shows the number of irrelevant variables included in the final model and the out-of-sample root mean square error (RMSE). Results are based on 100 bootstrap samples on the whole dataset, and the

Table 4 Body fat example: means and standard deviations (between brackets) of the results shown in Fig. 5

Method	RMSE	Selected noisy
Stepwise selection	6.9 (0.7)	48 (7)
Reference model + stepwise selection	4.7 (0.5)	25 (7)

predictive performance is tested on the observations excluded at each bootstrap sample. We observe that the reference model reduces the number of irrelevant variables included in the final model. This leads to less overfitting and thus to improved out-of-sample predictive performance in terms of RMSE. The reference model approach applied to the stepwise backward regression achieves outstanding improvements considering its simplicity, although it does not reach the goodness of the much more sophisticated projective prediction approach (see results of Sect. 2.1).

3.3 Reference model's effect and the projection predictive approach

The previous experiments allowed us to study the impact of using reference models in variable selection and to disentangle their influence from that of the actual variable selection method, specifically in the minimal subset selection case. The simulation study based on artificial data of Sect. 3.1 shows clear improvements in terms of predictive performance of the selected model and false discovery rate when the selection is based on a reference model, regardless of the actual variable selection method applied. These results are confirmed also in the real word data example with the body fat dataset. The stepwise selection achieves far better selection when coupled with the reference model approach (see Fig. 5 and Table 4). However, the projection predictive approach remains the best method in any of the experiments we run and on all the performance indexes we measured. Although we designed a reference model approach for general selection method, the projection predictive approach is a principled and validated way to do the selection. Indeed, the purpose of the former is only for fair comparisons in our study rather than a ready-to-use selection method.

4 Complete variable selection

An alternative to minimal subset variable selection is *complete variable selection*, in which the goal is to find all relevant variables that have some predictive information about the target. In complete variable selection, it is possible that there are theoretically relevant variables, but given finite noisy data we are not able to infer their relevance. The projection predictive approach was originally designed for the minimal subset variable selection, but we will test a simple iterative variant for the complete variable selection case in this section. In addition, we analyse the benefits of the reference model approach in combination with three other methods which have been specifically designed for complete variable selection. As the criteria of selection performance, we

evaluate the average false discovery rate and the average sensitivity (i.e., the ratio of the number of relevant selected variables over the total number of relevant variables). We also provide a comparison of the stability of the selection by means of a stability measure proposed by Nogueira et al. (2017), which goes from 0 to 1 and a higher value means a more stable selection.

4.1 Iterative projections

Projection predictive approach has been originally designed for minimal subset selection. For the comparison purposes, we modify the projection predictive approach for complete variable selection by using it iteratively. Applying the straightforward implementation of projpred, we are able to select a minimal subset of variables, which yield to a model with a predictive performance comparable to the full model's predictive performance. The iterative projection repeats the projpred selection for different iterations, at each time excluding the variables selected in the previous iterations from the search. At each iteration, the selected submodel size corresponds to the one having a predictive performance close enough to the baseline model, which in this iterative version is the submodel with the highest predictive score explored at the current iteration. This translates in the following stopping rule at each iteration:

$$\min\{i \in \{0, \dots, p\} : P(\mathrm{elpd}_i - \mathrm{elpd}_{\mathrm{best}} > 0) \geq \alpha\}, \tag{2}$$

where i indexes the submodel size, "best" stands for the best predictive explored submodel at the considered iteration, and the probability is computed from a normal approximation of the uncertainty in the cross-validation performance comparison using the mean and standard error of the elpd difference between reference and submodel (Vehtari et al. 2017; Piironen et al. 2020; Sivula et al. 2020). The algorithm terminates when the empty model (only intercept) satisfies the stopping rule (see Algorithm 1). The choice of the hyperparameter α is non-trivial, and we have observed sensitivity of the selection to such a choice, mainly when using cross-validation with a small number of observations or not very predictive variables. We have used the stopping rule recommended by Piironen et al. (2020) for the minimal subset selection, but other stopping rules would be possible in the iterative case and may be worth further research. In our experiments, we chose the default value used in the `projpred` R-package, that is, $\alpha = 0.16$. The choice of α to determine the submodel size is discussed in Piironen et al. (2020). One possible way to proceed, but just as rule of thumb, is to choose $\alpha = 0.16$, which correspond to requiring that the submodel predictive score is at most at one standard deviation distance from the best submodel predictive score.

Algorithm 1 Automated iterative projections

Output: $R := \{\text{selected variables}\}$
  $F := \{\text{set of variables}\}$
  $R = \{\varnothing\}$
  Fit reference model
  while $F \neq \{\emptyset\}$ **do**
    projection.HeuristicSearch()
    projection.elpdEstimate()
    $S = \min\{\text{sub} : P(\text{elpd}_{\text{sub}} - \text{elpd}_{\text{best}} > 0) \geq \alpha)\}$
    if $S = \{\varnothing\}$ **then**
      break
    else
      $R = R \cup S$
      $F = F \setminus S$
    end if
  end while

In the experiments shown in the next sections, we include an additional iterative method, which we refer to as 'iterative lasso'. It consists of the same iterative algorithm as iterative projpred except for not using any reference model, but the lasso method for variable selection, instead. That is, it uses the observed target values instead of predictions of the reference model. The comparison with iterative lasso can help to disentangle the effects of the iterative procedure and the usage of a reference model in complete feature selection.

4.2 Alternative complete variable selection methods

We consider three alternative complete selection methods: the control of the local false discovery rate (Efron 2008, 2012), the empirical Bayes median (Johnstone and Silverman 2004), and the selection by posterior credible intervals.

The control of the local false discovery rate consists of testing the z-values $\{z_j\}_{j=1}^{p}$ of a normal mean problem (explained in Sect. 4.3) on whether they belong to the theoretical null distribution f_0 (i.e., the null hypothesis H_0 meaning no relevance) against the alternative hypothesis distribution f_1. In our case, f_0 corresponds to the standard normal distribution (see expression (6)). The quantity of interest is the local false discovery rate (loc.fdr) defined as:

$$\text{loc.fdr}(z) = P(H_0|z) = \frac{f_0(z)\pi_0}{f(z)}, \tag{3}$$

where π_0 is the prior probability of H_0 and $f(z) = \pi_0 f_0(z) + (1 - \pi_0)f_1(z)$ is the marginal distribution of the z-values. The latter is estimated using splines with 7 degrees of freedom. We select variables with local false discovery rate below 0.2, which is

suggested by Efron (2012) as it corresponds to a Bayes factor larger than 36 (assuming $\pi_0 \geq 0.9$). The results of the comparison are not sensitive to the specific value. To estimate π_0 from the data, we use the default setting provided by the R-package locfdr (Efron et al. 2015).

The empirical Bayes median approach consists of fitting a Bayesian model with a prior composed by a mixture of a delta spike in zero and a heavy-tailed distribution. We use the implementation in the R-package EbayesThresh (Silverman et al. 2017). As suggested by Johnstone and Silverman (2004), we use a Laplace distribution resulting in a thresholding property, that is, there exists a threshold value such that all the data under that threshold have posterior median equal to zero. Therefore, the selection is done by selecting only those parameters whose posterior median is different from zero. The hyperparameter of the Laplace distribution and the mixing weight of the prior are estimated by marginal maximum likelihood.

The selection by 90% posterior credible intervals is done using the regularised horseshoe prior (Piironen and Vehtari 2017b) and selecting those variables whose posterior distribution does not include zero in the interval between the 5% and the 95% quantiles.

All of these methods provide a complete selection approach, and we compare their performance with and without using a reference model. That is, in the data condition, we apply the method on the original data y while, in the reference model condition, we replace y by their mean predictions $\hat{y}$ based on the reference model.

4.3 Simulation study 2

The iterative projection applies straightforwardly to data, whereas to investigate the performance of the three alternative complete selection approaches, we are going to use simulations based on the normal means problem. The normal means problem consists of estimating the (usually sparse) vector of means of a vector of normally distributed observations. The dimensionality of the vector of means is denoted by p and $\{z_j\}_{j=1}^{p}$ is the vector of observations of the random variables $\{Z_j\}_{j=1}^{p}$. The task is to estimate the latent variables $\{\theta_j\}_{j=1}^{p}$ of the following model:

$$Z_j|\theta_j, \sigma \overset{ind}{\sim} \mathrm{N}(\theta_j, \sigma^2), \quad j = 1, \dots, p. \tag{4}$$

This is equivalent to a linear regression where the design matrix is the identity matrix with the number of variables being equal to the number of observations. This formulation can be found in practice, for example, in the analysis of microarray data, where a large set of genes are tested in two groups of patients labeled as positive or negative to some disease (Efron 2008, 2012). The objective of the analysis is to select the subset of genes statistically relevant to the disease. One common way to proceed is to compute the two-sample t-statistic for every gene separately. After normalising these statistics, they then become the data Z in the normal means problem (4). For further details, see the examples by Efron (2008) and Efron (2012).

In our experiments, we retrieve the normal means problem from the sample correlations between the target and the variables using the Fisher z-transformation

(Hawkins 1989). Suppose we have a continuous target random variable Y, a set of p continuous variables $\{X_j\}_{j=1}^p$, and denote $\rho_j = \text{Cor}(Y, X_j)$. Further, suppose we have observed n statistical units and define r_j the sample correlation between the observations of the target variables $\{y_i\}_{i=1}^n$ and the j-th variable $\{x_{ij}\}_{i=1}^n$. Finally, we refer to the Fisher z-transformation function $\tanh^{-1}(\cdot)$ as $T_F(\cdot)$. Assuming each pair (Y, X_j) to be bivariate normally distributed, the corresponding transformed correlations are approximately normally distributed with known variance:

$$T_F(r_j) \overset{ind}{\sim} \text{N}\left(T_F(\rho_j), \frac{1}{n-3}\right), \quad j = 1, \dots, p. \tag{5}$$

Therefore, rescaling the quantities $T_F(r_j)$ by $\sqrt{n-3}$ and denoting the results as z_j, we have the formulation (4) of the normal means problem, this time with unit variance:

$$Z_j | \theta_j \overset{ind}{\sim} \text{N}(\theta_j, 1), \quad j = 1, \dots, p. \tag{6}$$

In this case, the quantities of interest θ_j are equal to $\sqrt{n-3}\, T_F(\rho_j)$.

In our simulations, we use different levels of correlation $\rho \in \{0.3, 0.5\}$ and numbers of observations $n \in \{50, 70, 100\}$. The total number of variables p and the number of relevant variables k are fixed to $p = 1000$ and $k = 100$, respectively. In general, the lower ρ and n, the more challenging the variable selection is. For this example, Piironen et al. (2020) proposed to use a reference model which a Bayesian linear regression using the first five supervised principal components (SPC) as variables and imposing a hierarchical prior on their coefficients:

$$\begin{aligned}
Y_i | \boldsymbol{\beta}, \sigma^2, \boldsymbol{u}_i &\overset{ind}{\sim} \text{N}(\boldsymbol{u}_i^T \boldsymbol{\beta}, \sigma^2) && i = 1, \dots, n \\
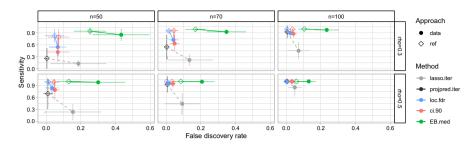\beta_j | \tau &\overset{iid}{\sim} \text{N}(0, \tau^2) && j = 1, \dots, 5 \\
\tau &\sim t_4^+(0, s_{max}^{-2}) && \\
\sigma &\sim t_3^+(0, 10). &&
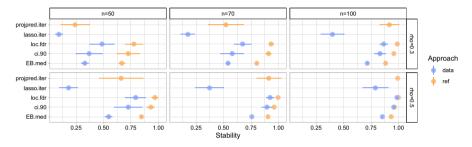\end{aligned} \tag{7}$$

In the above, u_{ij} represents the j-th SPC evaluated at observation i, and s_{max} denotes the sample standard deviation of the largest SPC. The SPCs are computed using the R-package dimreduce (https://github.com/jpiironen/dimreduce) setting the screening threshold parameter at $0.6 s_{max}$. In our experiments, the results are not sensitive to the specific choice of the screening threshold, yet a more principled approach would be to use cross-validation to select the threshold as done by Piironen et al. (2020).

Figure 6 shows the average sensitivity on the vertical axis and the average false discovery rate on the horizontal axis based on 100 data simulations for the different combinations of n and ρ. The best selection performance is on the top-left corner of each plot, as it implies the lowest false discovery rate and the highest sensitivity. We see that for all tested selection methods, the use of a reference model improves the selection performance, as it reduces the false discovery rate (shifting to the left) and increases the sensitivity (shifting upwards). In accordance with what can be expected, the larger data set size (n) and the higher the true correlations (ρ), the

Fig. 6 Simulation study 2: complete variable selection sensitivity against false discovery rate based on 100 data simulations with one standard deviation error bars. The reference approach improves sensitivity and reduces false discovery rate for all methods
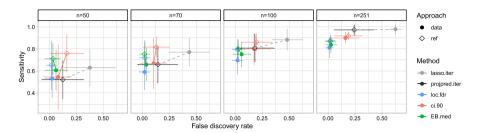

Fig. 7 Simulation study 2: complete variable selection stability estimates with 95% intervals based on 100 data simulations. The reference approach improves stability for all methods

easier the selection is. Thus, for easier selection scenarios, the benefits of the reference model are smaller since the raw data already provide enough information to identify the relevant variables. The iterative projpred has good false discovery rate in all cases, and the sensitivity is good except when the number of observations and the correlation level are small. It performs better than the iterative lasso selection in all simulated scenarios. For all these variable selection methods, tuning the corresponding method parameters could affect the balance between FDR and sensitivity, and thus the methods could made to produce more similar results. This was not done in this experiment, as the main point was to show that all methods can perform better with a reference model.

Figure 7 shows the estimates of the stability measure proposed by Nogueira et al. (2017) with 0.95 confidence intervals based on 100 simulations. Such a measure takes to account the variability of the subset of the selected variables at each simulation (originally at each bootstrap sample), modelling the selection of each variable as a Bernoulli process. Further details are available in Nogueira et al. (2017). Use of a reference model improves the stability of all selection methods. The improvement is larger when the problem is more difficult (small n and ρ). In addition, we observe less uncertainty in the stability estimates for the reference approach (i.e., smaller width of the 95% intervals), which can be still connected to the overall stability of the procedure. As in Fig. 6, the iterative projection does not perform well in the hardest scenarios.
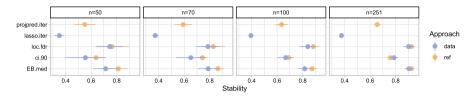
Fig. 8 Body fat example with noisy variables: complete variable selection sensitivity against false discovery rate based on 100 bootstrap samples with one standard deviation error bars. The improvement from using the reference approach is small (except that the projpred is much better than lasso)

4.4 Body fat example: part 3

We conclude our complete selection experiments using the body fat dataset one more time. As earlier, we add noisy uncorrelated variables to the original data to get a total of 100 variables. Since we do not have a ground truth available with regard to the original variables of the data, we assume it is reasonable to consider all of them relevant, at least to some degree. The artificially added variables are naturally irrelevant by construction. We compute correlations between each variable and the target variable, that is the amount of fat, and transform them by Fisher-Z-transformation. The original assumption in order for (5) to hold is that the variables are jointly normally distributed. In our experience the normal approximation in (5) is still reasonable, but after rescaling by $\sqrt{n-3}$ we do not fix the variance to be one, and instead estimate it from the data. We compare the iterative projection, the control of the local false discovery rate (loc.fdr), the empirical Bayes median (EB.med) and the selection by posterior credible intervals at level 90% (ci.90). In order to vary the difficulty of the selection, we bootstrap subsamples of different sizes, going from $n = 50$ up to $n = 251$ (i.e., the full size of the data). For each condition, results are averaged over 100 bootstrap samples of the respective size.

Figure 8 shows the sensitivity against the false discovery rate. In almost all of the bootstrapped subsamples, the reference model improves the selection both in terms of sensitivity and false discovery rate. When $n = 50$, we observe worse false discovery rates, yet by a lower amount compared to the gain in sensitivity. Again, we observe that the benefits are more evident as the selection becomes more challenging (i.e., lower number of observations). The great performance of projpred in minimal subset selection is not carried over for the complete variable selection with iterative projpred (even changing the reference model with a full encompassing linear regression with regularised horseshoe prior), and the methods specifically designed for the complete variable selection perform better. However, we still observe a better selection with respect to iterative lasso in any of the examined scenarios, mainly in terms of false discovery rate. Figure 9 shows the stability results using the measure by Nogueira et al. (2017). The benefits of the reference model are here marginal, with only small improvements.

Fig. 9 Body fat example with noisy variables: complete variable selection stability estimates with 0.95 confidence intervals based on 100 bootstrap samples. The improvement from using the reference approach is small (except that the projpred is much better than lasso)

In this example, we have used the reference model defined as a linear regression over some supervised principal components, because it is natural for a large number of correlating variables, and has fairly good predictive performance plus it is computationally efficient. We do not argue that this is always the best choice, and more sophisticated models can lead to even better results. Here the purpose of the experiments were to motivate the use of reference models in general, and as we needed to average results over a lot of repetitions per simulation condition, we preferred such a comparably simple and computationally fast reference model.

5 Conclusion

In this paper, we demonstrated the benefits of using a reference model to improve variable selection, or more generally, model reduction. We have motivated and explained the general benefits of a reference model regardless of the method it is applied in combination with. Specifically, we have seen how the reference model acts as an approximation of the data generation mechanism through its predictive distribution. Such approximation is generally less noisy than the sample estimation available purely from the observed data, leading to the main benefits of the reference model approach. In our comparisons, we have analysed the effect of a reference model in the form of a filter on the observed target values on top of different widely used variable selection methods. Overall, using a reference model leads to more accurate and stable selection results independently of the specific selection method. These benefits apply to a large family of different methods, all involving a reference model in one way or the other. Some of these approaches have been present in the literature for some time (e.g., see references in Vehtari and Ojanen 2012; Piironen et al. 2020) but often without a clear explanation of why they are actually favourable and how they connect to other related approaches. We hope that the present paper can fill some of these gaps by providing a unifying framework and understanding of reference models.

We argue that, whenever it is possible to construct a reasonable reference model, it should be employed on top of the preferred selection procedure or as an integral part of more complex methods, for example, the projective prediction approach (Piironen et al. 2020). Note that one of the main challenges in many real world application will consist in devising a sensible reference model itself and assessing its predictive performance.

To build good predictive reference models, which are specifically tuned to the data and problem at hand, we recommend them to be developed using a robust Bayesian modelling workflow, for instance, as outlined by Gelman et al. (2020) and Gabry et al. (2019).

Another main result of this paper is that the projective prediction approach shows superior performance in minimal subset variable selection compared to alternative methods, whether or not these methods make use of a reference model. That is, while the reference model is certainly one important aspect of the projective prediction approach, it is not the only reason for its superior performance. Rather, by incorporating the full uncertainty of the posterior predictive distribution into the variable selection procedure (instead of just using point estimates) and using a principled cross-validation method, projective predictions combine several desirable variables into a single procedure (Piironen et al. 2020). In summary, we would strongly recommend using projective predictions for minimal subset variable selection if possible and computationally feasible. However, if this is not an option in a given situation, we would in any case recommend using a reference model on top of the chosen variable selection method.

The projective prediction approach was not designed for the complete variable selection. We tested a simple iterative version of projpred, with mixed results, and the methods specifically designed for complete variable selection (especially loc.fdr) performed better in our experiments. It is left for future research to develop a better projective prediction approach for complete variable selection problems.

All Bayesian models in this paper have been implemented in the probabilistic programming language `Stan` (Carpenter et al. 2017) and fit via dynamic Hamiltonian Monte Carlo (Hoffman and Gelman 2014; Betancourt 2017), through the R-packages `rstan` (Stan 2019) and `rstanarm` (Goodrich et al. 2019). Graphics elaborations have been done using `ggplot2` (Wickham 2016) and the `tidyverse` framework (Wickham et al. 2019).

Appendix 1: Projective predictions

The projective prediction (projpred) approach was developed and is thoroughly described by Piironen et al. (2020). In this appendix, we provide a high level description of the method so that readers do not need to study paper by Piironen et al. (2020) in detail to understand the main ideas behind projpred.

The parameter distribution of a given candidate submodel is denoted by π and the induced predictive distribution by $q_\pi(\tilde{y})$. We would like to choose π so that $q_\pi(\tilde{y})$ maximises some predictive performance utility, for example, the expected log-predictive density (elpd) defined as:

$$\text{elpd}[q_\pi] = \int \log q_\pi(\tilde{y}) p_t(\tilde{y}) d\tilde{y}, \tag{A1}$$

where $p_t(\tilde{y})$ denotes the (usually unknown) true generating mechanism of future data $\tilde{y}$. If we refer to the posterior predictive distribution of a reference model with $p(\tilde{y}|D)$, where D stands for the data on which we conditioned on, we can approximate (A1) using $p(\tilde{y}|D)$ instead of the true data generation mechanism $p_t(\tilde{y})$. The

maximisation of the elpd using the reference model's predictive distribution is equivalent to the minimisation of the Kullback-Leibler (KL) divergence from the reference model's predictive distribution to the submodel's predictive distribution:

$$\max_{\pi} \int \log q_{\pi}(\tilde{y}) p(\tilde{y}|D) \, \mathrm{d}\tilde{y} \;\Leftrightarrow\; \min_{\pi} \mathrm{KL}[p(\tilde{y}|D) \,\|\, q_{\pi}(\tilde{y})] \tag{A2}$$

The term on the right-hand side of Eq. (A2) describes what is referred to as the projection of the predictive distribution, which is the general idea behind the projection predictive approach (see Piironen et al. 2020). We now summarise the workflow of the projection predictive approach in the particular case of the draw-by-draw projection (original formulation by Dupuis and Robert 2003), following Piironen et al. (2020). Suppose we have observed n statistical units with target values $\{y_i\}_{i=1}^{n}$ and a set of observed variables for which we want to obtain a minimally relevant subset. Then, the main steps are the following:

1. Devise and fit a reference model. Let $\{\boldsymbol{\theta}_*^s\}_{s=1}^{S}$ be the set of S draws from the reference model's posterior.
2. Rank the variables according to their relevance using some heuristics and consider as candidate submodels only those which preserve this order, starting from including only the highest ranked variable. The submodels are then naturally identified by their model size. This step is not strictly necessary but reduces the number of submodels considered in the following steps and thus reduces computation time.
3. For each submodel π selected in Step 2, project each of the reference model's posterior draws $\boldsymbol{\theta}_*^s$ as follows:

$$\boldsymbol{\theta}_{\perp}^s = \operatorname*{argmin}_{\boldsymbol{\theta}^s \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \mathrm{KL}\big[p(\tilde{y}_i|\boldsymbol{\theta}_*^s) \,\|\, q_{\pi}(\tilde{y}_i|\boldsymbol{\theta}^s)\big], \tag{A3}$$

   where $p(\tilde{y}_i|\boldsymbol{\theta}_*^s)$ stands for the predictive distribution of the reference model with parameters fixed at $\boldsymbol{\theta}_*^s$ and conditioning on all the variable values related to the statistical unit (identified by the subscript i), whereas $q_{\pi}(\tilde{y}_i|\boldsymbol{\theta}^s)$ is the predictive distribution of the submodel. The projected draws $\boldsymbol{\theta}_{\perp}^s$ then present the projected posterior for the submodel.
4. For each submodel (size), test the predictive performance for a chosen predictive utility score, for example, via cross-validation. Fast cross-validation can be performed using approximate leave-one-out cross-validation via Pareto-smoothed importance-sampling (PSIS-LOO-CV; Vehtari et al. 2017).
5. Choose the smallest submodel (size) that is sufficiently close to the reference model's predictive utility score. The results in this paper were not sensitive to the specific choice of how "sufficiently close" is defined, and we used the same definition as (Piironen et al. 2020).

In general, Expression (A3) is not an easy optimisation problem However, in the special case of the submodels being generalised linear models with a likelihood coming from the exponential family, (A3) reduces to a maximum likelihood estimation

problem, which can be easily solved (Dupuis and Robert 2003). For further details on the projective prediction workflow and implementation, see the paper by Piironen et al. (2020).

References

Akaike H (1974) A new look at the statistical model identification selected papers of Hirotugu Akaike. Springer, pp 215–222

Bair E, Hastie T, Paul D, Tibshirani R (2006) Prediction by supervised principal components. J Am Stat Assoc 101(473):119–137

Betancourt M (2017) A conceptual introduction to Hamiltonian Monte Carlo. arXiv:1701.02434

Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Riddell A (2017) Stan: a probabilistic programming language. J Stat Softw 76(1):1–32

Catalina A, Bürkner PC, Vehtari A (2020) Projection predictive inference for generalized linear and additive multilevel models. arXiv:2010.06994

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

Dupuis JA, Robert CP (2003) Variable selection in qualitative models via an entropic explanatory power. J Stat Plan Inference 111(1–2):77–94

Efron B (2008) Microarrays, empirical Bayes and the two-groups model. Stat Sci 23(1):1–22

Efron B (2011) Tweedie's formula and selection bias. J Am Stat Assoc 106(496):1602–1614

Efron B (2012) Large-scale inference: empirical Bayes methods for estimation, testing, and prediction. Cambridge University Press, Cambridge

Efron B, Turnbull B, Narasimhan B (2015) locfdr: Computes local false discovery rates https://CRAN.R-project.org/package=locfdr. R package version 1.1-8

Faraggi D, LeBlanc M, Crowley J (2001) Understanding neural networks using regression trees: an application to multiple myeloma survival data. Stat Med 20(19):2965–2976

Gabry J, Simpson D, Vehtari A, Betancourt M, Gelman A (2019) Visualization in Bayesian workflow. J R Stat Soc Ser A (Stat Soc) 182(2):389–402

Gelman A, Vehtari A, Simpson D, Margossian CC, Carpenter B, Yao Y, Modrák M (2020) Bayesian workflow. arXiv:2011.01808

Goodrich B, Gabry J, Ali I Brilleman S (2019) rstanarm: Bayesian applied regression modeling via Stan. https://mc-stan.org/rstanarm. R package version 2.19.3

Harrell FE (2015) Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Springer, Berlin

Hawkins D (1989) Using U statistics to derive the asymptotic distribution of Fisher's Z statistic. Am Stat 43(4):235–237

Heinze G, Wallisch C, Dunkler D (2018) Variable selection—a review and recommendations for the practicing statistician. Biom J 60(3):431–449

Hoffman MD, Gelman A (2014) The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. J Mach Learn Res 15(1):1593–1623

Johnson RW (1996) Fitting percentage of body fat to simple body measurements. J Stat Educ 4(1)

Johnstone IM, Silverman BW (2004) Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. Ann Stat 32(4):1594–1649

Lindley DV (1968) The choice of variables in multiple regression. J Roy Stat Soc Ser B (Methodol) 30(1):31–53

Nogueira S, Sechidis K, Brown G (2017) On the stability of feature selection algorithms. J Mach Learn Res 18(1):6345–6398

Parmigiani G, Inoue L (2009) Decision theory: principles and approaches, vol 812. Wiley, New York

Paul D, Bair E, Hastie T, Tibshirani R (2008) "Preconditioning" for feature selection and regression in high-dimensional problems. Ann Stat 36(4):1595–1618

Piironen J, Vehtari A (2015) Projection predictive variable selection using Stan + R. arXiv:1508.02502

Piironen J, Vehtari A (2016) Projection predictive model selection for Gaussian processes. In: 2016 IEEE 26th international workshop on machine learning for signal processing (MLSP)

Piironen J, Vehtari A (2017a) Comparison of Bayesian predictive methods for model selection. Stat Comput 27(3):711–735

Piironen J, Vehtari A (2017b) Sparsity information and regularization in the horseshoe and other shrinkage priors. Electron J Stat 11(2):5018–5051

Piironen J, Vehtari A (2018) Iterative supervised principal components. In: Storkey A, Perez-Cruz F (eds) Proceedings of the 21st international conference on artificial intelligence and statistics, vol 84, pp 106–114

Piironen J, Paasiniemi M, Vehtari A (2019) projpred: projection predictive feature selection. http://mc-stan.org/projpred, http://discourse.mc-stan.org/

Piironen J, Paasiniemi M, Vehtari A (2020) Projective inference in high-dimensional problems: prediction and feature selection. Electron J Stat 14(1):2155–2197

R Core Team (2018) R: a language and environment for statistical computing Vienna, Austria. https://www.R-project.org/

Rockova V, Lesaffre E, Luime J, Löwenberg B (2012) Hierarchical Bayesian formulations for selecting variables in regression models. Stat Med 31(11–12):1221–1237

Silverman BW, Evers L, Xu K, Carbonetto P, Stephens M (2017) Ebayesthresh: empirical bayes thresholding and related. https://CRAN.R-project.org/package=EbayesThresh. R package version 1.4-12

Sivula T, Magnusson, M Vehtari A (2020) Uncertainty in Bayesian leave-one-out cross-validation based model comparison. arXiv:2008.10296

Stan Development Team (2019) RStan: the R interface to Stan. http://mc-stan.org/. R package version 2.19.2

Stein C (1956) Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Proceedings of the third Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics

Stein C, James W (1961) Estimation with quadratic loss. In: Proceedings of the 4th Berkeley symposium mathematical statistics probability, vol 1, pp 361–379

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Roy Stat Soc Ser B (Methodol) 58(1):267–288

Vehtari A, Ojanen J (2012) A survey of Bayesian predictive methods for model assessment, selection and comparison. Stat Surv 6:142–228

Vehtari A, Gelman A, Gabry J (2017) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat Comput 27(5):1413–1432

Venables WN, Ripley BD (2013) Modern applied statistics with s-plus. Springer, Berlin

Wickham H (2016) ggplot2: elegant graphics for data analysis. Springer, New York

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Yutani H (2019) Welcome to the tidyverse. J Open Source Softw 4(43):1686