The reference model approach in feature selection problems

Federico Pavone, Juho Piironen, Paul-Christian Bürkner and Aki Vehtari *

June 28, 2019

Abstract

When a reference model is used to guide the variable selection problem, it acts as a noise-filter. We show how this translate into higher stability and goodness of the selection. We devise a simple reference model approach that can be used on top of any feature selection procedure and use the normal means problem as a benchmark comparing different methods using or not the reference model. We include in our comparisons state-of-the-art complete selection procedures showing improved stability, false discovery rate and sensitivity of the selection, and also Bayesian methods controlling the amount of posterior shrinkage and bias. In addition to that we show in a real-world example the benefits achieved by the reference model through the projection predictive approach compared to the selection via stepwise backward regression.

1 Introduction

In statistical applications, one of the main steps of the modelling workflow is covariate or feature selection, which is a special case of model reduction. The necessity of carrying out the selection arises from different problematics. Sometimes the leading objective of the analysis is to predict future observations, thus the main focus in building the model is on the predictive performance. Yet a high dimensional parameter space often leads to dangers as overfitting or other technical problems, as expensive costs to collect all the features for future observations or computational and memory burdens (e.g. when we are operating on limited resources), and thus a sparser representation of the model would be preferred. In addition to that, simpler models bring also advantages in terms of interpretability. Some other times, variable selection can be also seen as an inferential tool: the interest is on knowing which variables are meaningful, or relevant, in relation to the target variable of interest. This typically happens in a *small n large p* scenario, as in many microarray data analysis in medicine or biology. A very large number of genes is usually analysed in two groups of individuals labeled as *positive* or *negative* to some disease; the goal is to spot the subset of explanatory genes in order to perform further experiments (see examples in Efron, 2012).

Either the target of the selection is the minimal subset of feature with good enough predictive ability, or the whole subset of relevant variables, the selection process is going to be based on assumptions regarding the "true" data generation mechanism. Making assumptions means making models, thus we identify two families of feature selection, or more in general model reduction, approaches depending on how the data generation mechanism is modelled: the data based and the reference model based ones. The former includes all those methods that use the observed empirical distribution as model approximation of the data generation mechanism, as for example happens in the Lasso selection Tibshirani (1996) or in the stepwise backward/forward regression, where the selected submodels balance sparsity with predictive ability on the observed data. The reference model based methods use instead the predictive distribution of a full-encompassing model, i.e. the reference model, as an approximation of the data generation mechanism. Such an idea is widely present in the literature with different names; for example, Harrell (2015) refers to the reference model as a full model that can be thought as the "gold standard" and shows some examples of how it can be used to seek a sparser approximation. He highlights some of the benefits as the possibility to calculate the accuracy with which the submodel approximates the best model and the inheritance of the shrinkage when it is applied to the full model. 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 covariates are difficult to be understood in terms of relevance. Faraggi

*Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, Finland.

et al. (2001) propose to combine the benefits of the neural networks, which become what we call reference models, with the interpretability of regression trees, in order to provide a framework to make inferences about risk groups. 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 response quantities y.
The reference model approach has been used for long time also in the Bayesian framework, an example the pioneering work by Lindley (1968). Following Goutis and Robert (1998), Dupuis and Robert (2003) project the posterior distribution of the reference model in each candidate model by minimisation of the Kullback-Liebler divergence, and then select the smallest one with enough explanatory power avoiding any prior elicitation for the submodels. In the special case of candidate models being GLMs from the exponential family, the optimisation problem corresponds to the task of finding the maximum likelihood estimate using the expected pointwise predictions of the reference model instead of the observed target values. Different extensions have followed Dupuis and Robert (2003), as Nott and Leng (2010) that add an l_1 penalty to select the best approximating model and Tran et al. (2012), who select the best approximating model by optimising the Kullback-Liebler divergence between the predictive distribution of the reference model and the approximating model plus an l_1 penalty on the submodel coefficients. A recent review of these methods together with some methodological improvements, as clustered projection and selection by comparison of a predictive performance utility, is given by Piironen et al. (2018).

In this paper we bring motivations for the use of a reference model. A properly designed model is able to clean part of the noise present in the data, thus to provide an improved and more stable selection process. As an example, we show the improvements in the selection when using the projection predictive approach versus the stepwise regression, with a real world dataset. We argue that, however the reference model is used, it can be always seen acting as a filter on the observed data. Following this idea, we devise a simple reference model approach that can be applied on top of any selection procedure. We thus show the reference model benefits, regardless of what specific procedure is applied, using as a benchmark the normal means estimation problem. We apply different state-of-the-art methods comparing in each case the results basing the selection on the data or on the reference model. In case of full Bayesian methods, we report results in terms of shrinkage and accuracy of the posterior estimates, whereas when a full selection is carried out, we compute quantities as false discovery rate, sensitivity and a measure of the stability of the selection. The results indicate how the core reason why the reference model based methods perform well is the reference model itself, rather than the specific way of using it. We do recognise that all these methods diversify how the subset of variables is actually chosen and which properties it has, thus it remains fundamental to investigate and develop such different procedures. We hope that this work can increase the attractiveness on this family of methods and, therefore, it can be of interest to a large number of practitioners.
We would like also the reader to note that besides the fact that we consider a Bayesian framework in our discussion, the reference model can and is built also in frequentist settings, as it is done in some of the references previously provided.

The paper is structured as follows. In Section 2 we introduce the idea of the reference model, its benefits with examples, including the projection predictive approach, and how it can be used as a filter on data in a simple way. In Section 3 we use different methods for feature selection comparing the selection results using or not the reference model in the framework of the normal means problem. Finally, in Section 4 we present our conclusions.

2 Why the reference model helps

The main assumption under any reference model approach is to be operating in a $\mathcal{M}$-*complete* framework (Bernardo and Smith, 2009; Vehtari and Ojanen, 2012), i.e. to be able to construct a model, that is the reference model, whose description of a future observation distribution is comparable to the "true" data generation mechanism. In practice, any full-encompassing model which has a good predictive performance is a candidate reference model. General modelling guidelines apply also for the reference model, except no variable selection is carried out at this stage.

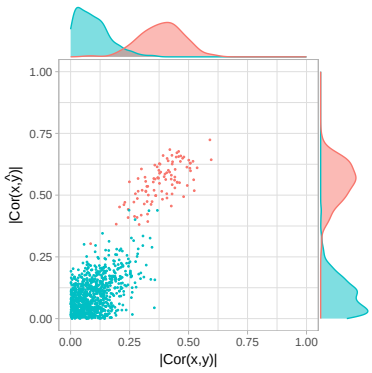

Figure 1: Sample correlation plot of each feature (relevant in red, non-relevant in blue) with the target variable y and the latent variable f respectively on the x and y axis. Data are generated according to procedure (1) *with parameters* $n = 70$, $\rho = 0.3$, $k = 100$, $p = 1000$.

A good predictive model is able to clean part of the noise present in the data. The noise is the main source of the instability and tends to obscure the relevance of the covariates in relation to the target variable of interest. Indeed, let us consider the following explanatory example taken from Piironen et al. (2018). The data generation mechanism for each statistical unit is the following:

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

f is the latent variable of interest of which Y is a noisy observation. The first k covariates are significantly related to the target variable Y and correlated among themselves. Precisely, ρ is the correlation among any pair of the first k covariates, whereas $\sqrt{\rho}$ and $\sqrt{\rho}/2$ are the level of correlation between any significant covariate and respectively f and Y. If we had an infinite amount of observations, the sample correlation would be equal to the true correlation, though, even in this ideal asymptotic regime, the correlation computed would be still a biased indicator of the true relevance of each variable with respect to the latent variable of interest due to the intrinsic noisy nature of Y. When using a reference model, we are trying to model f through the covariates $\{X_j\}_{j=1}^p$ taking in account that what we observe (i.e. Y) is corrupted by noise. Thus, if our model is good, we are able to describe f better than the simply observed variable Y and to use such information to assess the relevance of the features. Indeed, Figure 1 shows the scatter plot of the absolute value of the sample correlation of each feature with the point prediction by a reference model, in this case see model (15), against the absolute value of the sample correlation of each feature with the noisy observations $\{y\}_{i=1}^n$. Looking at the marginal distribution, we observe that using a reference model the two groups of features, i.e. relevant and non-relevant, have a minor overlap, and thus can be better distinguished, with respect to when the correlation is computed using directly the observed data. We can think about the correlation as a general indicator of the relevance of a feature.

In Figure 1 the reference model acts as a noise-filter on the observed data through its predictions. This simple idea is what happens every time a reference model is used, as it is evident for example in the preconditioning by Paul et al. (2008) or in the draw-by-draw projection for GLMs by Dupuis and Robert (2003).
As an another example, let us consider a candidate submodel in a feature selection process and let us refer with π to the parameter distribution of the submodel and with $q_\pi(\tilde{y})$ to the induced predictive distribution. We would like to choose π in order that the latter maximises some predictive performance utility, as the expected log-predictive density (elpd) defined as:

$$\text{elpd}[q_\pi] = \int \log q_\pi(\tilde{y}) p_t(\tilde{y}) d\tilde{y} \tag{2}$$

If we refer to the posterior predictive distribution of a reference model with $p_{\text{ref}}(\tilde{y}|D)$, where D stands for the data on which we condition on, we can approximate (2) using $p_{\text{ref}}(\tilde{y}|D)$ instead of the true, unknown, data generation mechanism $p_t(\tilde{y})$. The maximisation of the elpd using the reference model predictive distribution is equivalent to the minimisation of the KL divergence between the reference model predictive distribution and the submodel one:

$$\max_{\pi} \int \log q_\pi(\tilde{y}) p_{\text{ref}}(\tilde{y}|D) d\tilde{y} \leftrightarrow \min_{\pi} \text{KL}[p_{\text{ref}}(\tilde{y}|D)||q_\pi(\tilde{y})] \tag{3}$$

The term on the right-hand side of Equation (3) describes what is referred as the projection of the predictive distribution, which is the general idea behind the projection predictive approach. Again the reference model is acting as a filter, or substitute, on data.

Relying on the idea of the reference model as a noise-filter, a simple reference model approach, which can be used on top of any feature selection procedure, consists in substituting the target observations with a point prediction of the reference model, e.g. posterior predictive mean. Afterwards, any method for feature selection can be used. Note that we do not argue that such reference model approach should be used as a feature selection tool in real applications; there are many other and more sophisticated ones available in the literature. The reason of this simple approach is to have a fair comparison between using a reference model or not with exactly the same selection procedure. Other kinds of comparison, as the one in Section 2.1, work well as motivational examples, but fail in terms of generality, resulting in a not principled comparison since the presence of the reference model is not the only thing differentiating the two procedures. In Section 3 we compare some well-known methods from the literature using both such reference model approach and the original data, observing at the end of the selection higher stability and quality of the selection thanks to the reference model.

2.1 Benefits of the reference model: an example with the projection predictive approach

As a motivational example to the study of the consequences of the use of a reference model in feature selection problems, we show its application in the form of the projection predictive approach, which uses the reference model to project the posterior samples in the submodels. The R-package *projpred* (available at https://CRAN.R-project.org/package=projpred) implements the projection in case of submodels in the family of the generalised linear models and, additionally, it provides a framework to choose the optimal submodel size through a predictive utility comparison between the submodels and the reference model.
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. (2018). Assume to have n statistical units with target values $\{y_i\}_{i=1}^n$ and a set of observed covariates, 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 posterior samples of the vector parameter of the model.
2. Rank the covariates using a heuristics and consider as candidate submodels the nested ones, starting to include the best ranked feature. The submodel are naturally identify by their model size.
3. For each submodel, project the reference model samples as follows:

$$\boldsymbol{\theta}_\perp^s = \underset{\boldsymbol{\theta}\in\Theta}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} \text{KL}\left[p(\tilde{y}_i|\boldsymbol{\theta}_*^s) \,||\, p(\tilde{y}_i|\boldsymbol{\theta}^s)\right] \quad s = 1, .., S \tag{4}$$

   where $p(\tilde{y}_i|\boldsymbol{\theta}_*^s)$ stands for the model distribution of the reference model with parameters fixed at the sample value $\boldsymbol{\theta}_*^s$ and conditioning on all the covariates related to the statistical unit (identified by the subscript i in the variable $\tilde{y}_i$), whereas $p(\tilde{y}_i|\boldsymbol{\theta}^s)$ the same but for the submodel. The projected samples correspond to the submodel parameter distribution.
4. For each submodel (size), test the predictive performance.

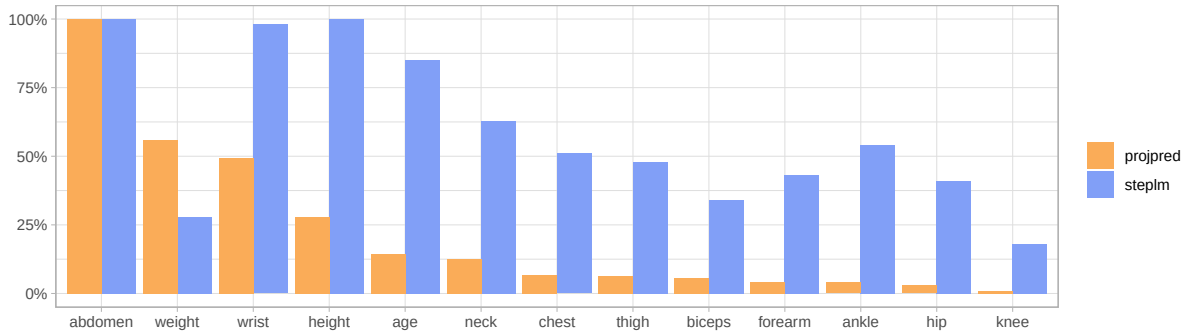

Figure 2: Bootstrap inclusion frequencies after 1000 bootstrap samples.

5. Choose the smallest submodel (size) that is sufficiently close to the reference model prediction utility score.

Expression (4) is not an easy optimisation problem, but in the special case of submodels in the GLMs family, (4) reduces to a maximum likelihood estimation problem, which can be easily optimised (e.g. using iteratively reweighted least squares alogirthm, IRLS). For further details on *projpred* workflow and implementation see Piironen et al. (2018).

We show the benefits of the reference model in the form of the projection approach (projpred) with respect to the stepwise backward regression (steplm) using the body fat data (Johnson, 1996). The analysis is mainly taken from Vehtari's R-notebook available at https://avehtari.github.io/modelselection/bodyfat.html, adjusting only the number of bootstrap repetitions to 1000 to have a consistent comparison with results from Heinze et al. (2018). The target variable of interest is the amount of body fat, which is measured with a complex and expensive procedure, and the original covariates are 13 anthropometric measurements (e.g. height, weight), some of which are highly correlated among themselves and thus provide some challenges in the model selection, for a total of 251 observations. The goal is to find the simplest model which is able to describe, that is to predict, sufficiently well the amount of fat.

Heinze et al. (2018) report the results using the stepwise backward elimination with a significance level of 0.157 with AIC selection, fixing abdomen and height to be always included in the model. We implement the selection via projection: the reference model includes all the covariates using as a prior the regularised horseshoe, submodels are explored using a forward search and the predictive utility is the expected log-predictive density (elpd) estimated using PSIS-LOO (Vehtari et al., 2017). In order to automatise the procedure, we select the submodel size as the smallest one which has an elpd score higher than the reference model score with probability 0.05. Figure 2 shows the bootstrap inclusion frequencies combining the results from Heinze et al. (2018) for the stepwise backward elimination and our experiments for the projection. The projection predictive approach has only two variables with inclusion frequencies above 50% (abdomen is the only included always), the third most included is 'wrist' at 49.5% and the fourth is 'height' at 27.8%, while the stepwise regression has seven features above 50%, of which abdomen and height are fixed. Such a lower stability of the stepwise regression can be observed also in the bootstrap model selection frequencies reported in Table 1. The first five selected models have a cumulative frequency of almost 70% with projpred, whereas with the stepwise regression only of 11.8%. In addition to that we note that the size of the selected model is much smaller when using projpred.

Table 2 reports the predictive performances, in terms of root mean square error, of the full model and the correspondent selected model using both the projection predictive approach and the stepwise backward elimination. We also repeat the selection after adding uncorrelated noisy features up to a total of 100 covariates and report in such a case both the size of the selected model and the number of selected noisy features. Whilst with the original data we do not observe any gain using projpred in terms of predictive performance (we do observe a smaller size for the selected model, however with a much higher computational cost), when noisy features are added the projection is able to notably improve the predictive performance

M	projpred	Freq (%)	steplm	Freq (%)
1	abdomen, weight	34.2	abdomen, height, wrist, age, chest, biceps	3.2
2	abdomen, wrist	13.3	abdomen, height, wrist, age, neck, forearm, thigh, hip	2.9
3	abdomen, weight, wrist	7.6	abdomen, height, wrist, age, forearm, chest	1.9
4	abdomen, height	7.4	abdomen, height, wrist, age, neck, forearm, chest	1.9
5	abdomen, height, wrist	6.7	abdomen, height, wrist, age, neck, forearm, chest, thigh, hip	1.9
6	abdomen, age, wrist	2.9	abdomen, height, wrist, age, neck, chest, biceps	1.8
7	abdomen, height, neck	2.2	abdomen, height, wrist, age, neck, thigh, biceps, hip	1.6
8	abdomen, age, height, wrist	1.2	abdomen, height, wrist, age, neck, forearm	1.5
9	abdomen, weight, thigh	1.1	abdomen, height, wrist, age, neck, biceps	1.5
10	abdomen, weight, neck	1.0	abdomen, height, wrist, age, beck, forearm, chest, biceps	1.4

Table 1: Bootstrap model selection frequencies after 1000 bootstrap samples.

	rmse.full	rmse.sel	size.sel	noisy.sel
projpred	4.4	4.5	2	-
steplm	4.4	4.5	7	-
projpred	*4.4	*4.5	2	0
steplm	5.3	5.5	38	32

Table 2: Model predictive performances with original data (first two rows) and adding noisy features (last two rows). Results with an asterisk using 10-fold CV, otherwise 20-fold CV.

keeping the same submodel size, whereas the stepwise regression selects a much larger submodel including a large number (32 over 38) of noisy features.

2.2 Computational burden

We can consider the reference model approach as a family of methods, thus providing an answer to the question what the computational burden is, is not possible, because not unique. It is possible to identify a common and inevitable overhead though, that is the cost of fitting the reference model. In the simple reference model approach we have described at the end of Section 2, it is actually the only added cost. In general, a full Bayesian model is advisable in order to have the best predictive performance, however including all the available features can be often quite computationally demanding. In such a case a first possible speed-up could be using screening or dimensionality reduction techniques, as for example the supervised principal components (Bair et al., 2006; Piironen and Vehtari, 2018). However, in some cases this does not always result in an appreciable speed-up, as when there are many predictive and not correlated features and, thus, many of them have to be included in the reference model anyway. Other alternatives could be fixing some parameters with their maximum marginal likelihood value (typically referred as empirical Bayes) or even use a frequentist approach, if predictions are good enough.

The comparison example between projpred and steplm shows that such additional computational burden can be worth it. In the case of the projection predictive approach it results in sparser submodels, which can result in saving time and general costs using the selected model for future predictions. Table 2 shows that there could be even an improvement in the prediction ability, thus justifying more additional computational costs in the selection. More benefits obtainable through the reference model are provided in the experiments of Section 3.

3 A comparison in the normal means problem framework

In this Section we show the benefits of the use of a reference model using as a benchmark the normal means estimation problem, which consists in estimating the vector of means, usually sparse, of a normal vector of observations. Let us call p the dimensionality of the vector of means and let $\{z_j\}_{j=1}^p$ be the vector of observations of the random variables $\{Z_j\}_{j=1}^p$, the normal means problem consists in estimating the latent variables $\{\theta_j\}_{j=1}^p$ of the following model:

$$Z_j|\theta_j, \sigma \overset{ind}{\sim} N(\theta_j, \sigma^2) \quad j = 1, .., p \tag{5}$$

Note that it is equivalent to a linear regression where the design matrix is the identity one. This formulation can be retrieved in different ways, a common example is given by microarray data, where a large set of genes are usually tested in two groups of patient labelled as positive or negative to some disease. The objective of the analysis is to spot the whole subset of statistically relevant genes to the disease and one possible way to proceed is to compute the two-samples t-statistics for each gene and after combining it with the cumulative density function of the standard distribution, the resulting data can be used in the problem formulation (5). For further details see the examples in Efron (2008, 2012).
In our examples we retrieve the normal means problem from the sample correlation using the Fisher z-transformation approximation (Hawkins, 1989). Suppose to have a continuous target random variable Y and a set of p continuous covariates $\{X_j\}_{j=1}^p$ and let us call $\rho_j = Cor(Y, X_j)$. Suppose to observe 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 covariate $\{x_{ij}\}_{i=1}^n$. Let finally refer to the function $\tanh^{-1}(\cdot)$ as $T_F(\cdot)$. It holds the following modelling approximation (assuming each pair (Y, X_j) normally distributed):

$$T_F(r_j) \overset{ind}{\sim} N(T_F(\rho_j), \frac{1}{n-3}) \quad j = 1, .., p \tag{6}$$

Therefore rescaling the quantities $T_F(r_j)$ by $\sqrt{n-3}$ and referring to the results as variables z_j, we find again formulation (5) with unit variance:

$$Z_j|\theta_j \overset{ind}{\sim} N(\theta_j, 1) \quad j = 1, .., p \tag{7}$$

Note that now the quantity of interest θ_j stands for $\sqrt{n-3}\,T_F(\rho_j)$. There are different ways to proceed with the selection from the normal means problem, we consider in our comparison both frequentist and Bayesian state-of-the-art methods. From a full-Bayesian perspective, a Bayesian regression model is fitted to the normal means problem typically using a sparsifying prior as in Bhattacharya et al. (2015), Bhadra et al. (2017) and Johnstone and Silverman (2004). In case of continuous priors, as the Dirichlet-Laplace (DL) and the horseshoe+ priors, there is not a canonical way to complete the selection since the posterior probability of an exactly zero signal is always zero; in their experiments Bhattacharya et al. (2015) use k-means clustering over the posterior samples using the DL prior, whereas Bhadra et al. (2017) focus the analysis on the obtained shrinkage and the correct estimation of the sparse signals using the horseshoe+ prior. We include in our experiments the DL and the regularised horseshoe prior (Piironen and Vehtari, 2017), which has shown in our opinion better scalability with respect to the horseshoe+, comparing the shrinkage and the correct estimation of the sparse signals using or not the reference model approach. The quantity of interests that we consider are the average SSE (sum of square errors) and the average SESE (sum of expected square errors), defined as follows:

$$\text{SSE} = \frac{1}{n-3}\sum_{j=1}^{p}(\hat{\theta}_j - \theta_j^0)^2 \tag{8}$$

$$\text{SESE} = \frac{1}{n-3}\sum_{j=1}^{p}\int_{\mathbb{R}}(\theta_j - \theta_j^0)^2 p(\theta_j|z_1, .., z_p)d\theta_j \tag{9}$$

$$= \frac{1}{n-3}\sum_{j=1}^{p}\mathbb{E}_{\theta_j}[(\theta_j - \theta_j^0)^2|z_1, .., z_p] \tag{10}$$

where $\theta_j^0 = \sqrt{n-3}\,T_F(\rho_j)$ is the true value, $\hat{\theta}_j$ is the posterior median of each parameter θ_j and (10) is a sum of posterior expectations. We divide the errors by $(n-3)$ to have quantities independent from the number of statistical units. Note that the quantity SESE takes in account also the uncertainty, thus the shrinkage, of the posterior distributions, whereas the SSE measures only the bias in the point estimates. Note also that any kind of point estimate could be used to compute the SSE, here we follow Bhattacharya et al. (2015) using the posterior median. The results that we show later in Section 3.1.1 are not sensible to such a choice. We also take into account the posterior mean for the shrinkage factor (κ_j) for each parameter θ_j, which for

Method	Alias	Type	Quantities of interest
Regularised Horseshoe	RHS	Shrinkage prior	SSE, SESE, κ
Dirichlet-Laplace	DL	Shrinkage prior	SSE, SESE, κ
Local false discovery rate	loc.fdr	Complete selection	fdr, sensitivity, stability
Median estimator	EB.med	Complete selection	fdr, sensitivity, stability
Credible intervals	ci.90	Complete selection	fdr, sensitivity, stability

Table 3: Summary of the methods used in the comparison.

model (7) with a hierarchical shrinkage prior of the type:

$$Z_j|\theta_j \overset{ind}{\sim} N(\theta_j, 1) \qquad j = 1, .., p \tag{11}$$

$$\theta_j|\tau, \nu_j \overset{ind}{\sim} N(0, \nu_j^2\tau^2) \tag{12}$$

is defined as:

$$\kappa_j = \frac{1}{1 + \tau^2\nu_j^2} \tag{13}$$

and, thus, in our case it becomes:

$$\kappa_j^{\text{RHS}} = \frac{1}{1 + \tau^2\lambda_j^2}, \qquad \kappa_j^{\text{DL}} = \frac{1}{1 + \psi\phi_j^2\tau^2} \tag{14}$$

respectively for the regularised horseshoe and the Dirichlet-Laplace priors. In the left hand-side expression, τ and λ_j stand for the global and local shrinkage parameters of the regularised horseshoe prior (notation and further details in Piironen and Vehtari, 2017), whereas on the right hand-side $\psi\phi_j^2\tau^2$ is the variance term of the Dirichlet-Laplace prior of parameter θ_j, following the same notation as Bhattacharya et al. (2015).

Johnstone and Silverman (2004) use a prior of spike-and-slab type (Mitchell and Beauchamp, 1988) given by a mixture of a delta distribution in zero and a heavy-tailed distribution; when the latter is a Laplace distribution they show an interesting thresholding property of the median estimator and, thus, present a complete selection procedure using their prior. We include in our comparison such method together with the control of the local false discovery rate (Efron, 2012, 2008) and a simple selection procedure using inclusion probability at 0.9 for posterior credible intervals using the regularised horseshoe prior (ci.90). All these methods provide a complete selection procedures, we compare the result using or not a reference model on top of the procedure evaluating the average false discovery rate (i.e. ratio of the number of non-relevant selected features over the number of selected features) and the average sensitivity (i.e. ratio of the number of relevant selected features over the total number of relevant features). We also provide a comparison of the stability of the selection using the measure proposed by Nogueira et al. (2017). Table 3 summarises the methods used in the comparison.

3.1 Simulated data

Here we use simulated data in order to control the difficulty of the selection problem. We rely on the scheme (1) using different levels of correlation (ρ) and number observation (n), precisely p and k are fixed respectively at 1000 and 100, whereas $n \in \{50, 70, 100\}$ and $\rho \in \{0.3, 0.5\}$. Lower the correlation level and the number of observations are, more challenging the selection is. As mentioned, this data generation mechanism was already proposed by Piironen et al. (2018), who also devise a reference model sufficiently good in predictions. It consists in a linear regression using the first five supervised principal components (SPCs) (Bair et al., 2006; Piironen and Vehtari, 2018) as follows:

$$\begin{aligned} Y_i|\boldsymbol{\beta}, \sigma^2, \boldsymbol{u}_i &\overset{ind}{\sim} N(\boldsymbol{u}_i^T\boldsymbol{\beta}, \sigma^2) && i = 1, .., n \\ \beta_j|\tau &\overset{iid}{\sim} N(0, \tau^2) && j = 1, .., 5 \\ \tau &\sim t_4^+(0, s_{max}^{-2}) \\ \sigma &\sim t_3^+(0, 10) \end{aligned} \tag{15}$$

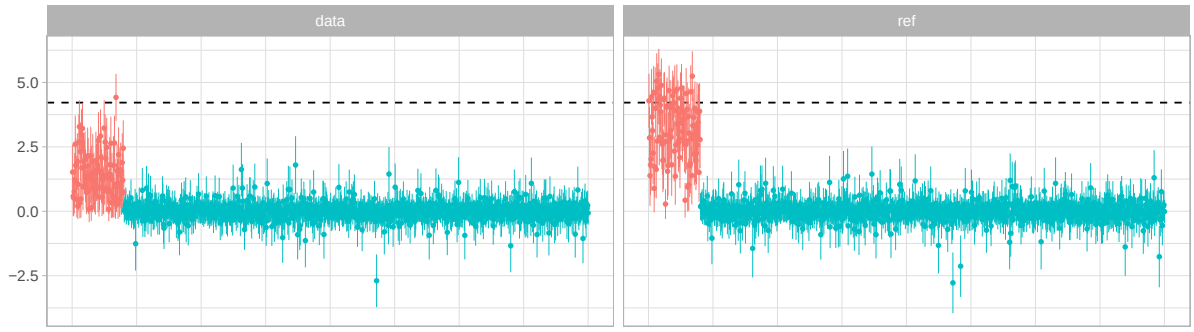

Figure 3: Posterior intervals for the parameters of the normal means problem ($n = 50$, $\rho = 0.3$) using the regularised horseshoe prior. Posterior means and one standard deviation intervals depicted. Respectively in red and in blue relevant and non-relevant features. Horizontal dashed line corresponds to the true value of the mean for the relevant features.

where u_{ij} stays for the j-th SPC of observation i, and s_{max} denotes the sample standard deviation of the largest SPC. We use as screening threshold for the SPCs the value $0.6s_{max}$, which has shown satisfying predictive performance. In our experiments the results are not very sensible to the chosen value, yet a more principled approach would be to use cross-validation to select the threshold as in Piironen et al. (2018).
The normal means problem formulation is obtained as explained in Section 3 by Fisher-transformation. We label as "ref" the reference approach, which consists in using the reference model on top of the computation of the z-values, and "data" the standard approach based on the observed data only. In Section 3.1.1 we analyse in the context of full-Bayes methods the different shrinkage and signal estimation provided by the use of a reference model or not. Section 3.1.2 shows results for the complete selection procedures.

3.1.1 Shrinkage and signal analysis

In the framework of full-Bayes methods we analyse the effect of the reference model using two continuous hierarchical shrinkage priors: the Dirichlet-Laplace (Bhattacharya et al., 2015) and the regularised horseshoe prior (Piironen and Vehtari, 2017). Although the horseshoe+ prior by Bhadra et al. (2017) is a state-of-the-art prior for sparse signal analysis in the context of the normal means problem, we decided to include the regularised horseshoe due to its faster MCMC inference, since our purpose in this paper is not to study the best prior choice, but rather compare and show the benefits of using a reference model when carrying out variable selection. We used an uniform prior on the interval $(1/p, 1/2)$ for the Dirichlet hyperparameter of the Dirichlet-Laplace prior, whereas regarding the reguelarised horseshoe prior we followed the indication provided in Piironen and Vehtari (2017), setting the prior guess of the effective number of parameters equal to 1 to achieve the highest shrinkage. Yet we have not observed any sensibility of the results due to such a choice. Since, as already mentioned, continuous priors do not usually give an automatic selection procedure, in this Section we study the amount of shrinkage and the bias of the posterior median looking at the SESE and the SSE (see definitions (8) and (10)), together with the posterior mean of the shrink factors.

Figure 3 shows an example of posterior distributions for the model (7) using the regularised horseshoe prior. Respectively on the right and on the left the results using or not the reference model approach. It is evident how the reference model helps in tearing apart the signals from the noise, resulting in less biased estimates. This is quantified in Figure 4 where average SESE and SSE are depicted after 100 simulations. We observe that the reference approach (in orange) has a clear lower error estimate for both the SESE and the SSE. As n and ρ grow, the amount of error diminish, as expected, due to a larger amount of information of the relevance of each variable provided by the data. We also note that the two priors seem to achieve very similar results, regardless of the approach used.

Figure 5 shows the posterior means for the shrinkage factors of each parameter using the regularised horseshoe

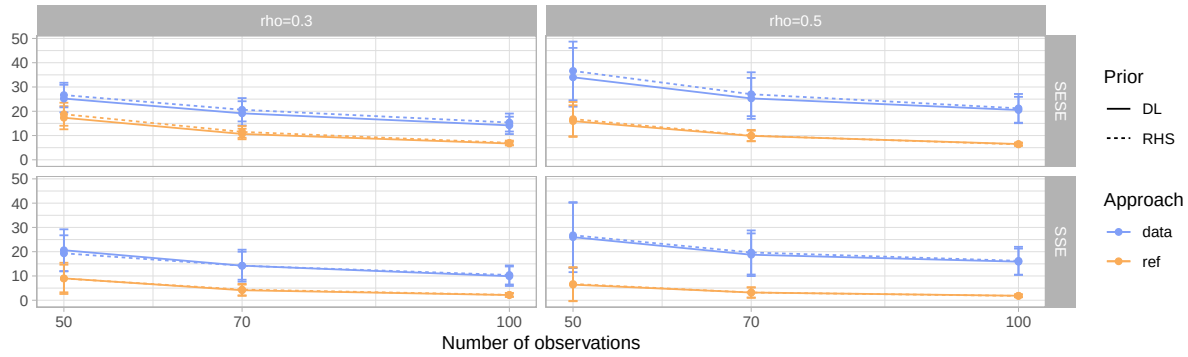

Figure 4: Average SESE and SSE with one standard deviation error bars after 100 simulations.

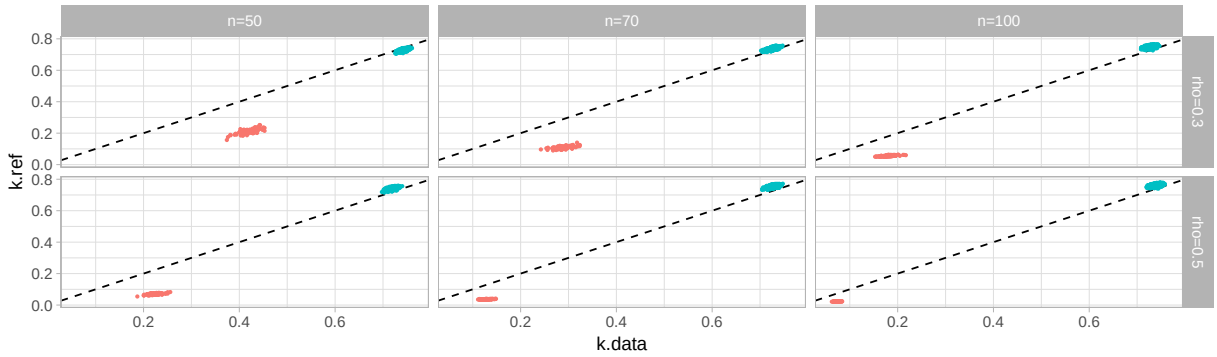

Figure 5: Average posterior mean values for the shrinkage factors using the regularised horseshoe prior after 100 simulations. Respectively in red and in blue relevant and non-relevant variables.

prior, averaged after 100 data simulations. Relevant and non-relevant features are respectively highlighted in red and blue. The ideal shrinkage would have $\kappa = 0$ when the parameter is a signal, whereas $\kappa = 1$ when it is noise. Respectively on the y-axis and on the x-axis are depicted the shrinkage factors using the reference model approach or not. If the reference model did not make any difference, we would expect the points to lie on the diagonal. Non-relevant features happen to lie on average on the diagonal for all the simulated scenario, meaning not evident benefits coming from the reference model. Although when considering the shrinkage amount for the relevant ones, we note that the points always lie under the diagonal, meaning that the reference model is able to shrink less the true signals. Such benefit is proportional to the difficulty of the problem, that is it increases with lower correlation and number of observations. Note that the perfect distinction of the two clusters of features (relevant and non-relevant) observable on both the marginal distributions of the shrinkage factors is due to the averaging over the 100 data simulations. In a single data realisation there would be more noise and the clusters would be closer and more overlapping. The results for the Dirichlet-Laplace prior are very similar and they are reported in the Appendix 4.

3.1.2 Complete selection analysis

As complete selection procedures we consider 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 is applied through the R-package *locfdr*. It consists in testing the z-values $\{z_j\}_{j=1}^{p}$ to belong to the theoretical null distribution f_0 (the null hypothesis, meaning no relevance) against the alternative hypothesis distribution f_1. In our case f_0 correspond to the standard normal distribution, see expression (7). The quantity of interest is the local false discovery rate (loc.fdr)

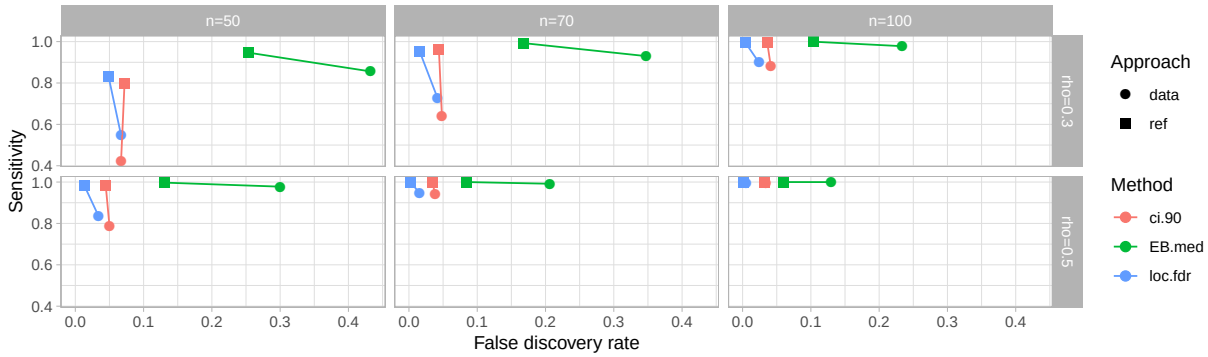

Figure 6: Sensitivity against false discovery rate after 100 data simulations.

defined as:

$$loc.fdr(z) = P(H_0|z) = \frac{f_0(z)\pi_0}{f(z)} \tag{16}$$

where π_0 is the prior probability of the null hypothesis 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 and we select features with local false discovery rate under the value 0.2. Such a value is commonly suggested only because it corresponds to a Bayes factor large than 36 (assuming $\pi_0 \geq 0.9$), other values can be considered and in our experience do not affect the results of the comparison. We use the default setting to estimate π_0 from the data provided by the package.
The empirical Bayes median procedure is given by the R-package *EbayesThresh* and consists in fitting a Bayesian model with a prior composed by a mixture of a delta in zero and a heavy-tailed distribution. We use, as suggested, a Laplace distribution resulting in a thresholding property, i.e. there exists a threshold value such that all the data under that threshold have posterior median equal to zero. Therefore the selection is done including 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 posterior credible intervals is done using the regularised horseshoe prior and selecting those features whose posterior distribution does not include zero in the interval between the 0.05-th and the 0.95-th quantiles.

Figure 6 reports the average sensitivity (on the y-axis) versus the average false discovery rate (x-axis) after 100 data simulations for the different combination of n and ρ. For each method (different colours) the square and the circle dots respectively correspond to using or not the reference approach. The best selection performance is on the top-left corner of each plot, meaning lowest false discovery rate and highest sensitivity. It is possible to note that regardless of the method used, the use of a reference model improves the goodness of the selection diminishing the false discovery rate (left shifting) and/or augmenting the sensitivity (up shifting). In accordance with what expected, higher the number of observations and the correlation are, easier the selection is and, thus, the benefits of the reference model are less notable, since the unprocessed data already give enough information to identify the significant variables.

Figure 7 shows the estimates of the stability measure proposed by Nogueira et al. (2017) with 0.95 confidence intervals after 100 simulations. Such a measure takes in account the variability of the subset of the selected features at each simulation (originally at each bootstrap sample), modelling the selection of each variable as a Bernoulli process. Further details are available in their paper as asymptotic normal distribution, based on which confidence intervals are computed. The reference model helps improving the stability of the selection: again, such benefit is notable when the problem is more difficult (small n and ρ). We observe also less uncertainty in the stability estimates for the reference approach (i.e. the width of the 0.95 confidence intervals), which can be still connected to the overall stability of the procedure.

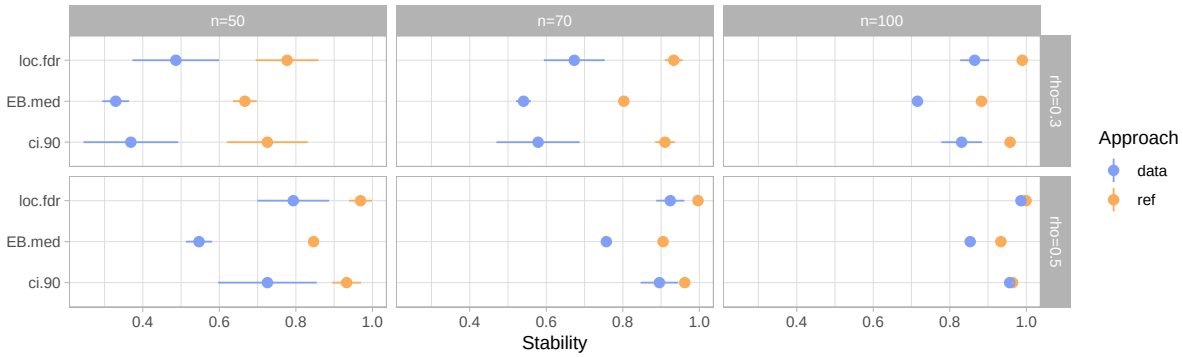

Figure 7: Stability estimates with 0.95 confidence intervals after 100 data simulations.

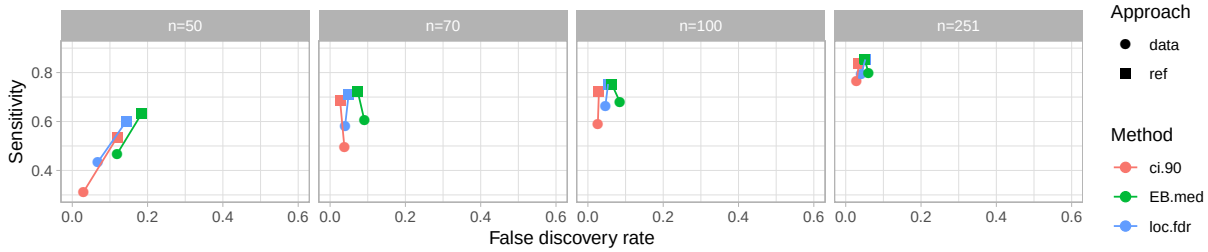

Figure 8: Sensitivity against false discovery rate after 100 data simulations.

3.2 Real world data: the body fat dataset

We conclude our experiments using the real world dataset body fat, already described in the comparison between the projection predictive approach and the stepwise regression in Section 2.1. Consistently with the previous comparisons, we design the variable selection by normal means problem. We add to the original data noisy uncorrelated covariates up to a total of 1000 features. These covariates are normally distributed and scaled as the original ones. We compute correlations between each variable and the target variable, that is the amount of fat, and transform them by Fisher transformation. The original assumption, in order that (6) holds, is that the variables are jointly normally distributed. In our experience the normal approximation in (6) is still reasonable, but after rescaling by $\sqrt{n-3}$ we do not fix the variance to be one, instead we estimate it from the data. The methods we compare are those used in Section 3.1.2: 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 0.9 (ci.90). The maximum likelihood estimator is used to estimate the variance of the null hypothesis in the local false discovery rate (see Efron, 2012, Chap. 6), whereas the median absolute deviation from zero for the empirical Bayes median method. The selection via posterior credible intervals is done using a regularised horseshoe prior on the parameters for the means, while a standard log-normal prior for the variance. In order to vary the difficulty of the selection, we bootstrap subsamples of different sizes, going from $n = 50$ up to $n = 251$. Each time results are averaged over 100 bootstrap subsamples.

Figure 8 shows the sensitivity against the false discovery rate. Since we do not have a ground truth regarding the original covariates of the data, we believe it is reasonable to consider all of them relevant, at least at some degree. We thus consider as non-relevant all the artificially added ones. In almost all the bootstrapped subsamples, the reference model improves both in terms of sensitivity and false discovery rate. When $n = 50$, we observe a worsening in terms of false discovery rate, yet in a lower amount compared to the gain in sensitivity. We observe again that the benefits are more evident as the selection is more challenging (i.e. less number of observations). Figure 9 shows the stability results, still using the measure provided by Nogueira et al. (2017). The benefits of the reference model are here marginal: small improvements can be observed, but overall not very notable.

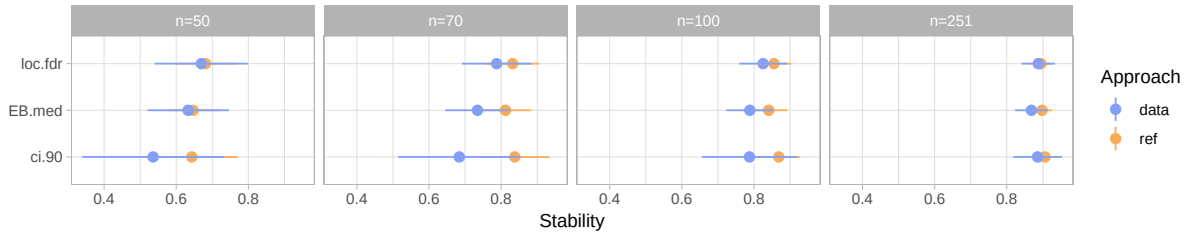

Figure 9: Stability estimates with 0.95 confidence intervals after 100 data simulations.

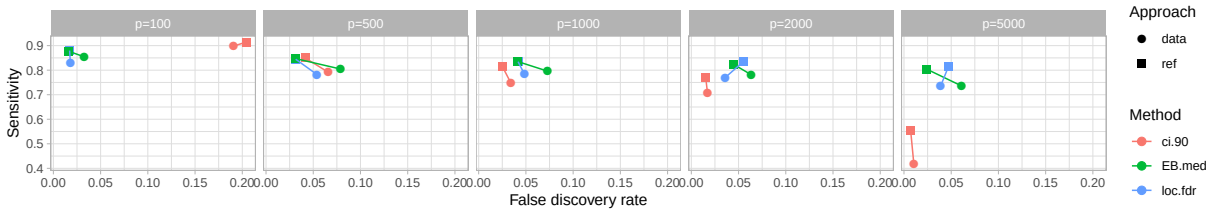

Figure 10: Sensitivity against false discovery rate after 100 data simulations.

As an additional experiment, we run the selection bootstrapping from the whole dataset, but varying the overall number of covariates from $p = 100$ up to $p = 5000$. Figures 10 and 11 respectively show the sensitivity against the false discovery rate and the stability results. As in the previous experiment, we note that the reference model has a larger impact as the complexity of the problem increases: previously as n decreases, now as p increases. Note the different scale of the sensitivity and the false discovery rate axes. We observe an odd behaviour of the selection via credible intervals (ci.90): when $p = 100$ and $p = 5000$ the results are strongly different from the other methods, both in terms of sensitivity/false discovery rate and stability.

Finally, we repeate the selection of Section 2.1 implementing the stepwise backward selection by AIC as implemented in the *step* function in R. We have not fixed any variable to be always included in the model and the overall number of covariates (original plus noisy) is 100. We compare results using or not the reference model on top of the procedure; here we have used again the reference model given by the first five supervised principal components. Figure 12 shows the number of noisy features included in the final model and the out-of-sample root mean square error (RMSE). Results are after 100 bootstrap simulations and the predictive performance is tested on the observations excluded at each bootstrap sample. We observe that the reference model reduces the number of noisy features included in the final model. This leads to less overfit to the data and, thus, improved out-of-sample predictive performance in terms of RMSE. Also in this case, we see that there is not a notable difference in terms of the stability of the two approaches, which in this case is qualitatively assessed through the variability of the histogram. We can not use the stability measure provided by Nogueira et al. (2017), since it does not take in account possibly swaps in the set of selected features due to high correlation between the covariates, which in this case can occur. Our reference model approach applied to the stepwise backward selection achieves outstanding improvements considering its simplicity, yet it does not reach the goodness of more thoughtful implementations, as projpred (see results of Section 2.1).

We would like the reader to note that in this example we have still used the reference model defined as a linear regression over some supervised principal components, because of its fairly good predictive performance and speed to fit. We do not argue that this is always a good choice, more sophisticated and well-thought models can lead to even better results. Since in our case the sake of the analysis was only explanatory in terms of the comparison we have been carrying out through the paper and needed to average results over many (100) simulations, we preferred to stick to such model.

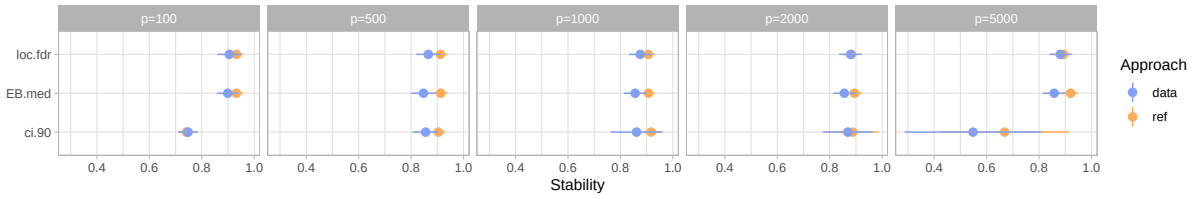

Figure 11: Stability estimates with 0.95 confidence intervals after 100 data simulations.

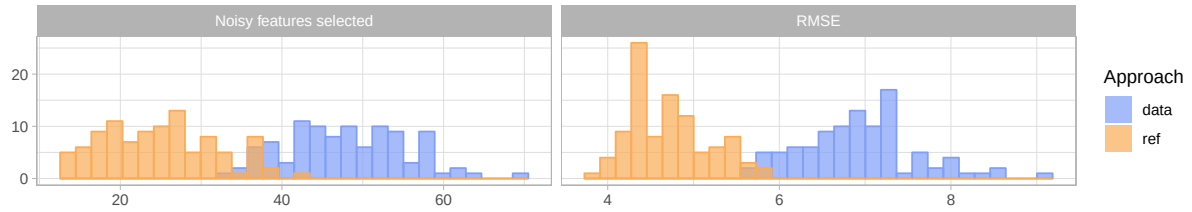

Figure 12: Stepwise backward selection using or not the reference model: number of noisy feature included in the model and RMSE after 100 bootstrap samples.

4 Conclusion

This paper has discussed the importance of using a reference model when dealing with feature selection, or more in general model reduction. We want to highlight that the goal of this paper is not to provide a specific algorithm or procedure to carry out the selection, but rather to bring motivations and attentions to the general use of a reference model.

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 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 variable selection methods widely used. The overall benefits we have showed independently of the specific method applied, consisting in stability and goodness of the selection, motivate what we refer to as the reference model approach for the model reduction framework. Such approach translates into a family of different methods, some of which have been present in the literature for several years.

In a real-world data application (body fat data), we have applied one of these methods, namely the projection predictive approach as indicated by Piironen et al. (2018), in opposition to the selection via stepwise backward regression with the results provided by Heinze et al. (2018). The selection via projection resulted to be more stable and led to a sparser representation. After adding additional noisy covariates, the selected submodel showed also higher predictive performance compared to the one selected by the stepwise backward regression, which included several non-relevant features.

Whenever it is possible to come up with a reasonable reference model, our suggestion is to employ it when selecting models, because of its nature as baseline of the comparisons and its ability to improve the selection. Note that the main challenge in many real world application will consist in devising the reference model itself and assessing its predictive performance, for which we suggest the user to rely on robust modelling workflow indications (see e.g. Gabry et al., 2019).

All Bayesian models have been implemented using Stan (Carpenter et al., 2017), which uses HMC-NUTS sampler (Hoffman and Gelman, 2014) and allows full posterior Bayesian inference. The code to run all the experiments is available at https://github.com/fpavone/ref-approach-paper.

References

Eric Bair, Trevor Hastie, Debashis Paul, and Robert Tibshirani. Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473):119–137, 2006.

José M Bernardo and Adrian FM Smith. *Bayesian Theory*, volume 405. John Wiley & Sons, 2009.

Anindya Bhadra, Jyotishka Datta, Nicholas G Polson, Brandon Willard, et al. The horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis*, 12(4):1105–1131, 2017.

Anirban Bhattacharya, Debdeep Pati, Natesh S Pillai, and David B Dunson. Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512):1479–1490, 2015.

Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 2017.

Jérome A Dupuis and Christian P Robert. Variable selection in qualitative models via an entropic explanatory power. *Journal of Statistical Planning and Inference*, 111(1-2):77–94, 2003.

Bradley Efron. Microarrays, empirical Bayes and the two-groups model. *Statistical Science*, 23(1):1–22, 2008.

Bradley Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2012.

David Faraggi, Michael LeBlanc, and John Crowley. Understanding neural networks using regression trees: an application to multiple myeloma survival data. *Statistics in medicine*, 20(19):2965–2976, 2001.

Jonah Gabry, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(2):389–402, 2019.

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

Frank E Harrell. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015.

DL Hawkins. Using U statistics to derive the asymptotic distribution of Fisher's Z statistic. *The American Statistician*, 43(4):235–237, 1989.

Georg Heinze, Christine Wallisch, and Daniela Dunkler. Variable selection–a review and recommendations for the practicing statistician. *Biometrical Journal*, 60(3):431–449, 2018.

Matthew D Hoffman and Andrew Gelman. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

Roger W Johnson. Fitting percentage of body fat to simple body measurements. *Journal of Statistics Education*, 4(1), 1996.

Iain M Johnstone and Bernard W Silverman. Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32(4):1594–1649, 2004.

Dennis V Lindley. The choice of variables in multiple regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(1):31–53, 1968.

Toby J Mitchell and John J Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988.

Sarah Nogueira, Konstantinos Sechidis, and Gavin Brown. On the stability of feature selection algorithms. *The Journal of Machine Learning Research*, 18(1):6345–6398, 2017.

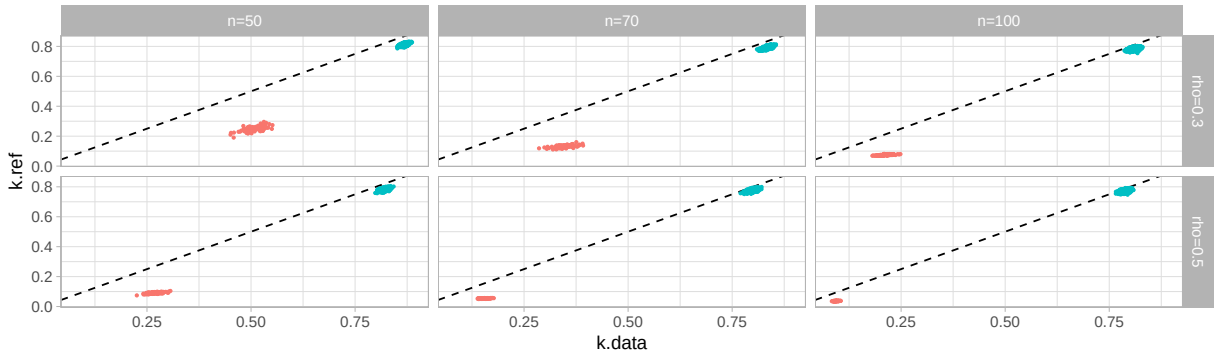

Figure 13: Average posterior mean values for the shrinkage factors using the Dirichlet-Laplace prior after 100 simulations. Respectively in red and in blue relevant and non-relevant variables.

David J Nott and Chenlei Leng. Bayesian projection approaches to variable selection in generalised linear models. *Computational Statistics & Data Analysis*, 54(12):3227–3241, 2010.

Debashis Paul, Eric Bair, Trevor Hastie, and Robert Tibshirani. "Preconditioning" for feature selection and regression in high-dimensional problems. *The Annals of Statistics*, 36(4):1595–1618, 2008.

Juho Piironen and Aki Vehtari. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2):5018–5051, 2017.

Juho Piironen and Aki Vehtari. Iterative supervised principal components. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 106–114, 2018.

Juho Piironen, Markus Paasiniemi, and Aki Vehtari. Projective inference in high-dimensional problems: prediction and feature selection. *arXiv preprint arXiv:1810.02406*, 2018.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Minh-Ngoc Tran, David J Nott, and Chenlei Leng. The predictive lasso. *Statistics and computing*, 22(5): 1069–1084, 2012.

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

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432, 2017.

Appendix A: Shrinkage factors with Dirichlet-Laplace prior

Figure 13 shows the posterior mean shrinkage factors using the Dirichlet-Laplace prior for the normal means problem in the simulated data study of Section 3. The result is qualitatively equal to the regularised horseshoe prior's one (see Figure 5).