Reference model in feature selection problems

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

May 21, 2019

Abstract

We carry out a comparison of different methods of feature selection using or not what we call a reference model on top of the procedure, that is a model that describes well our data. We measure the performance and the stability of each method using both the standard approach, that simply relies on data, and the reference model one. We include different procedures in the comparison as a selection through high probability density posterior credibility intervals, a selection made by controlling the Q-value and the local false discovery rate. Finally, we include in the comparison also a real world data example. Results show an increased stability and a better performance of the selection in favour of the reference model approach.

1 Introduction

In statistical applications, one of the main steps of the modelling workflow is covariate or feature selection. From a full Bayesian perspective, when the goal of the analysis is to do prediction, no selection should be carried out and integration over all parameters uncertainties, i.e. posterior distributions, is the recommended procedure. However, a high dimensional parameter space brings different difficulties as computation burden and overfitting. For such reason different feature selection procedures have been proposed both in the frequentist and Bayesian setting (e.g. see Rockova et al. (2012), Vehtari and Ojanen (2012) and Piironen and Vehtari (2017a)). Sometimes the task of the analysis is not prediction, but rather inference on which variables are significantly related to the phenomenon of interest. A common case is microarray data in medicine or biology, where for example for a very large number of genes, expression levels are measured for two sets of patients labeled as *positive* or *negative* to some disease (typically a *small n large p* scenario). The main goal of the analysis is to spot which subset of these genes is significantly related to the disease in order to perform further experiments. A possible way to proceed is to perform hypothesis testing to assess the significance of each gene, based on some modelling assumptions. Efron (2012) refers to this as multiple hypothesis testing problem. Such example could be still considered as feature selection, intended this time as the problem of "identifying *all* features that are predictive about (that is, statistically related to) the target variable" (Piironen et al., 2018).

Regardless which kind of selection problem we are trying to solve, we consider examples of the reference model approach all those methods that make use of a full-encompassing model, i.e. the reference one, to carry out the selection. The idea of using a reference model is already present in some works from the 60s as Lindley (1968). It is also the basis for the projection predictive approach (Dupuis and Robert, 2003; Piironen et al., 2018), from which we have taken the hint of using the reference model as a filter (see Section 2). The reference model, if properly designed, is a less noisy representation of the true data generation mechanism with respect to the observed data distribution. Therefore in a variable selection context it improves the selection process in terms of quality and stability of the selected subset of features, whether the goal is the smallest model with high enough predictive performance or the whole subset of relevant features.

In this paper we try to show such benefits with:

- a real-world data example comparing the projection predictive approach with the stepwise backward regression selection,

---

- a structured comparison in the framework of the normal means estimation problem using different state-of-the-art methods with or not a reference model on top of the procedure, with simulated data and a real-world data example.

The paper is structured as follows. In Section 2 we introduce the idea of the reference model, its benefits with examples and how we used it as a filter on our data in a very 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   Reference model approach to feature selection

The reference model approach has his roots in the $\mathcal{M}$-*complete* framework (Bernardo and Smith, 2009; Vehtari and Ojanen, 2012) and started to be used in the feature selection context already starting from the 60s with Lindley (1968). There is not a specific definition of what a reference model is; ideally any full-encompassing model which has a good predictive performance is a candidate reference model. General guidelines on how to devise it are the same about any model in a data analysis context, expect for the fact that no selection is carried out for the reference model. Piironen et al. (2018) showed that through the projection predictive approach the reference model can helps in terms of stability and model size in the selection process. In this Section we aim to give some hints on why this happens and on how we can generally rely on a reference model approach to feature selection with a simple example.

Consider the following data generation mechanism for each statistical unit, already used in Piironen et al. (2018):

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

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. 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 (10), 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 group of features, i.e. relevant and non-relevant, have a minor overlap with respect to when the correlation is computed using directly the observed data. Therefore the significance of each covariate is more clearly described when those are compared to a less noisy representation of the true phenomenon of interest, that is the prediction done by a reference model.

What is shown in Figure 1 is related also to the projection predictive approach. In case of a submodel belonging to the generalised linear model family, it happens that the draw-by-draw projection coincide with the maximum likelihood estimator with observed data substituted with reference model's predictions (Piironen et al., 2018). It is possible to see the idea of the reference model as a filter on our data also from the projection of the predictive distribution itself. Let us call $p_{\text{ref}}(\tilde{y}|D)$ the posterior predictive distribution for a new observation $\tilde{y}$ for the reference model and let $\pi(\theta)$ and $q_\pi(\tilde{y})$ be respectively the submodel parameter and the induced predictive distributions. The projection of the predictive distribution allows to find the optimal parameter distribution π^* minimising the KL discrepancy among the induced predictive distribution of the submodel and the reference model one:

$$
\pi^* \quad \text{s.t.} \quad \text{KL}[p_{\text{ref}}(\tilde{y}|D)||q_{\pi^*}(\tilde{y})] = \min_\pi \text{KL}[p_{\text{ref}}(\tilde{y}|D)||q_\pi(\tilde{y})]
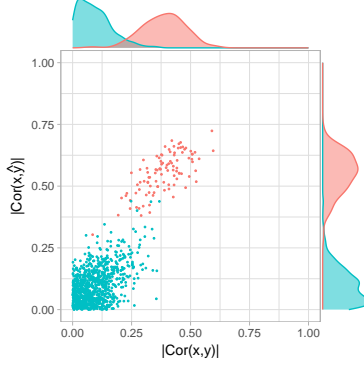\tag{2}
$$

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. Simulation parameters: $n = 70$, $\rho = 0.3$, $k = 100$, $p = 1000$.

Consider now the predictive performance of such submodel, a common utility in Bayesian context is the expected log-predictive density defined as:

$$\text{elpd}[q_\pi] = \int \log q_\pi(\tilde{y}) p_t(\tilde{y}) d\tilde{y} \tag{3}$$

where p_t is the distribution of a new observation under the true data generation mechanism. Maximising the elpd of a model is equivalent to minimising the KL-divergence of its predictive distribution from the true data generation mechanism. Therefore the projection corresponds to maximising the elpd using as distribution of the true data generation mechanism the posterior predictive distribution of the reference model, i.e.:

$$\min_\pi \text{KL}[p_{\text{ref}}(\tilde{y}|D)||q_\pi(\tilde{y})] \leftrightarrow \max_\pi \int \log q_\pi(\tilde{y}) p_{\text{ref}}(\tilde{y}|D) d\tilde{y} \tag{4}$$

Equation (4) clearly shows the action of the reference model as a filter on our target observations, substituting the unknown true data generation mechanism, that is usually estimated with the data empirical distribution, with the best known approximation that is available, i.e. the reference model posterior predictive distribution.

A simple, but effective, reference model approach that 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. 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

- correct bootstrap results for projpred –> 1000 bootstrap samples
- thus correct the comment on the table and on the barplot!!!!

The projection predictive approach 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 provides a framework to choose the optimal submodel size through a predictive utility comparison between the submodels and the reference model. When the goal is prediction, the choice is usually on the smallest submodel that is sufficiently close to the reference model prediction utility score. Further details on the projection predictive approach can be found in Piironen et al. (2018).
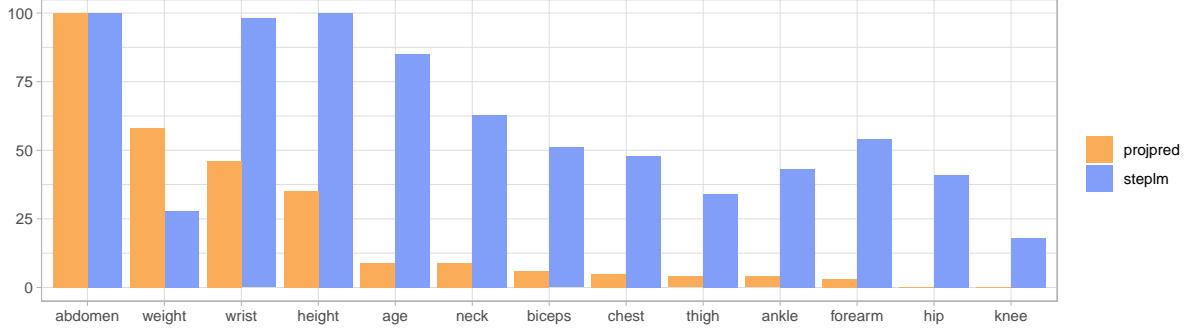
Figure 2: Bootstrap inclusion frequencies after 1000 bootstrap samples.

M	projpred	Freq	steplm	Freq
1	abdomen, weight	37	abdomen, height, wrist, age, chest, biceps	32
2	abdomen, wrist	10	abdomen, height, wrist, age, neck, forearm, thigh, hip	29
3	abdomen, height	10	abdomen, height, wrist, age, forearm, chest	19
4	abdomen, height, wrist	9	abdomen, height, wrist, age, neck, forearm, chest	19
5	abdomen, weight, wrist	8	abdomen, height, wrist, age, neck, forearm, chest, thigh, hip	19
6	abdomen, chest, height, wrist	3	abdomen, height, wrist, age, neck, chest, biceps	18
7	abdomen, height, neck, wrist	2	abdomen, height, wrist, age, neck, thigh, biceps, hip	16
8	abdomen, age, wrist	2	abdomen, height, wrist, age, neck, forearm	15
9	abdomen, age, height, neck, thigh, wrist	2	abdomen, height, wrist, age, neck, biceps	15
10	abdomen, chest	1	abdomen, height, wrist, age, beck, forearm, chest, biceps	14

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

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, 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 which has an elpd score higher than the reference model with probability 0.05. Figure 2 shows the bootstrap inclusions 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), while the stepwise regression has seven features, 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 74% with projpred, whereas with the stepwise regression only of 11.5%. 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. While with the original data we do not observe any gain using projpred in terms of predictive

4

	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 (top table) and adding noisy features (bottom table). Results with an asterisk using 10-fold CV, otherwise 20-fold CV.

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

3   A comparison in the normal means problem framework

In this Section we show the benefits of the use of a reference model. We consider as reference problem 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, the normal means problem can be written as follows:

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

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 et al. (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; 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, 2017b), 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:

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

$$\text{SESE} = \sum_{j=1}^{p} \mathbb{E}_{\theta_j}[(\theta_j - \theta_j^0)^2 | \mathbf{z}] \tag{9}$$

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 (9) is a sum of posterior expactations. 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. Johnstone et al. (2004) use a mixture of a delta distribution in zero and a heavy-tailed distribution, when this 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 using the R-package *EbayesThresh* together with the control of the local false discovery rate (Efron, 2012, 2008), using the R-package *locfdr*, and a simple selection procedure using inclusion probability at 0.9 for the HDP posterior credible intervals using the regularised horseshoe prior. 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).

3.1   Simulated data

First we use simulated data in order to vary 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 observation 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, 2017c) as follows:

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

where $\boldsymbol{u}_i$ stays for the i-th SPC and s_{max} denotes the sample standard deviation of the largest SPC. All model are implemented in Stan (Carpenter et al., 2017). 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
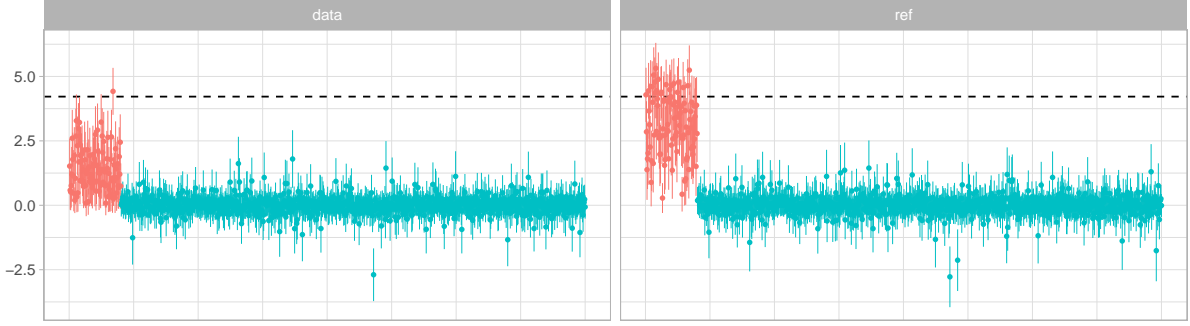
Figure 3: Posterior intervals for the parameters of the normal means problem ($n = 50$, $\rho = 0.3$) using the regularised horseshoe prior. Posterior means and one standard deviation intervals depicted. Respectively in red and in blue relevant and non-relevant features. Horizontal dashed line corresponds to the true value of the mean for the relevant features.

prior (Piironen and Vehtari, 2017b). 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 $(1/p, 1/2)$ for the Dirichlet hyperparameter of the Dirichlet-Laplace prior, whereas regarding the reguelarised horseshoe prior we followed the indication provided in Piironen and Vehtari (2017b) setting the prior guess of the effective number of parameters equal to 1 to achieve the highest shrinkage.

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 (9)). In the case of the regularised horseshoe prior, we also take into account the posterior mean for the shrinkage factor for each parameter θ_j, which for model (7) is defined as:

$$\kappa_j = \frac{1}{1 + \tau^2 \lambda_j^2} \tag{11}$$

with τ and λ_j respectively global and local shrinkage parameters, for further details see Piironen and Vehtari (2017b).

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, when using the reference model, whereas it unexpectedly increases when it is not used (WHY??). We also note that the two priors seem to achieve very similar results, regardless of the approach used.

Figure 5 shows the posterior medians for the shrinkage factors of each parameter, 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. In almost every simulated scenario, the reference model approach achieves a better shrinkage, resulting in a lower posterior median value for κ for the relevant features (in red) and a higher one for the non-relevant ones. The only exception is the case $n = 50$ and $\rho = 0.3$, where we observe that on average the reference model approach slightly underestimate the shrinkage factor for the non-relevant variables with the respect to the approach without a reference model. However, the shrinkage factor for the relevant features is much better estimated
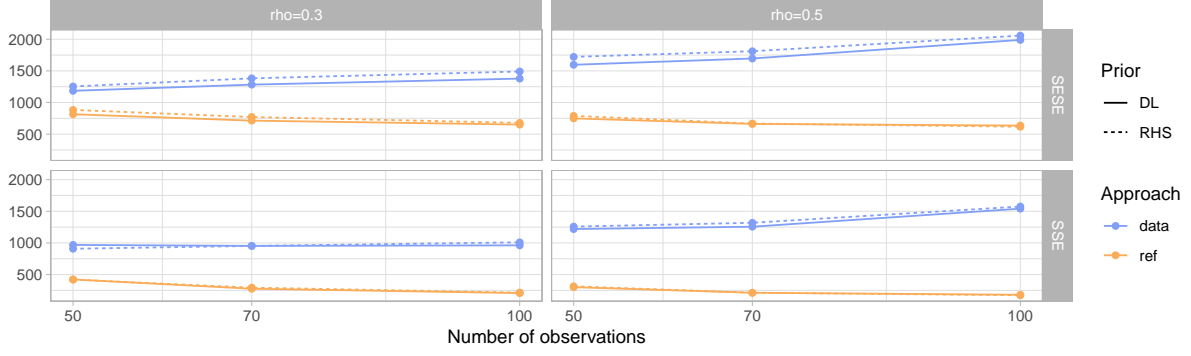
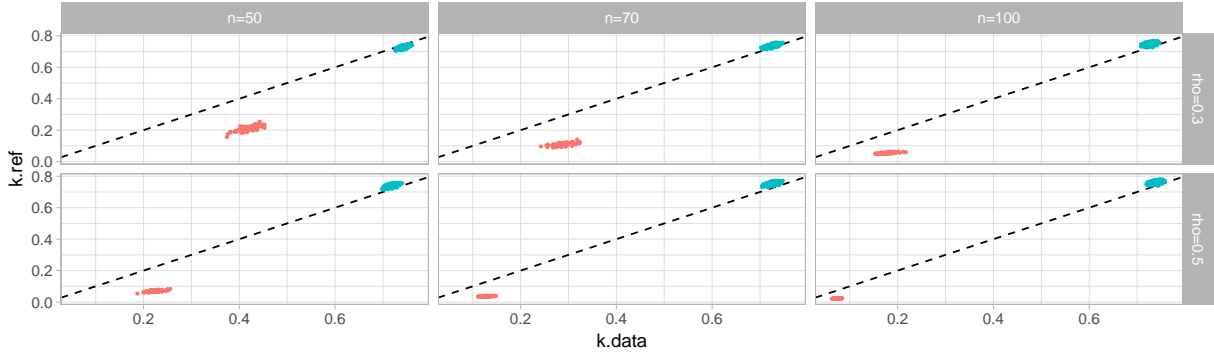*Figure 4: Average SESE and SSE after 100 simulations.*

Figure 5: Average posterior median values for the shrinkage factors after 100 simulations. Respectively in red and in blue relevant and non-relevant variables.

using a reference model with a difference in the order of 0.2 units, whereas for the non-relevant features the relative underestimation is only about 0.05.

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 et al., 2004) and the selection by posterior credible intervals. The control of the local false discovery rate is applied through the R-package *locfdr* using splines with 7 degrees of freedom to fit the marginal density and selecting features with local false discovery rate under the value 0.2. Such a value is the common suggested one just for the fact that it corresponds to a Bayes factor large than 36, other values can be considered and in our experience do not affect the results of the comparison. 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 highest probability density posterior credible interval does not include zero at the level 0.9.

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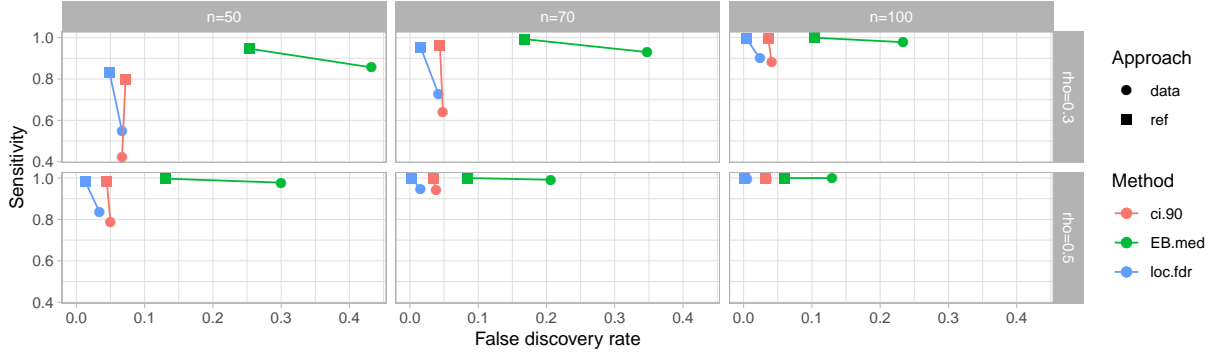
Figure 6: Sensitivity against false discovery rate after 100 data simulations.
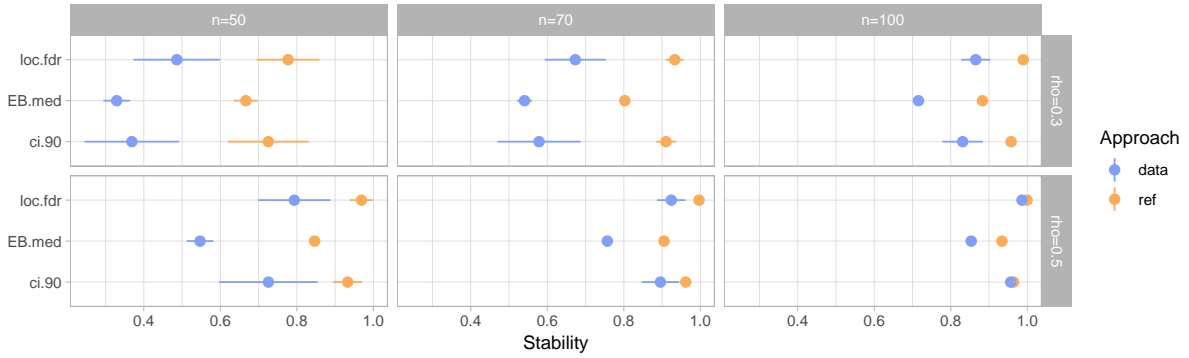

Figure 7: Stability estimates with 0.95 confidence intervals after 100 data simulations.

performance is on the top left corner, which means lowest false discovery rate and highest sensitivity. It is possible to note that regardless of the method used for the selection, the use of a reference model improves the quality diminishing the false discovery rate (left shifting) and/or augmenting the sensitivity (up shifting). In accordance with what we would expect, higher the number of observation and the correlation are, easier the selection is and thus the benefit 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.

3.2 Real world data

- Bodyfat data
- Results for local false discovery rate

4 Conclusions and discussion

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.

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.

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

Iain M Johnstone, Bernard W Silverman, et al. 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.

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.

Juho Piironen and Aki Vehtari. Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, 27(3):711–735, 2017a.

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

Juho Piironen and Aki Vehtari. Iterative supervised principal components. *arXiv preprint arXiv:1710.06229*, 2017c.

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

Veronika Rockova, Emmanuel Lesaffre, Jolanda Luime, and Bob Löwenberg. Hierarchical Bayesian formulations for selecting variables in regression models. *Statistics in Medicine*, 31(11-12):1221–1237, 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.