BAYESIAN ENRICHED POST-SELECTION MODELS FOR HIGH DIMENSIONAL DATA

DANIEL HARVEY LIDDELL, DURHAM UNIVERSITY, DEPARTMENT OF MATHEMATICAL SCIENCES

ABSTRACT

Over the past few years, regularisation approaches have become the key methods for analysing high dimensional data. However, being able to obtain accurate estimates and predictions alongside statistical inference remains a major and problematic challenge. Thus, in this research we introduce a Bayesian enriched post-selection method for analysing high-dimensional data. The main idea of this approach, is to first use a variable selection procedure to split the model into the selected and unselected covariates. We then use these two sets to construct a post-selection model that includes an asymptotically accurate approximation of the unselected covariates. This is then used to enrich the selected model and eliminate bias in high dimensions.

Keywords: High dimensional data, Sparsity, Bayesian regularisation, Post-selection model.

INTRODUCTION

In this research, we consider the high dimensional linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad p >> n.$$

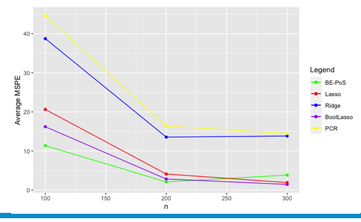
We shall take this model and apply the lasso regression technique (Tibshirani, 1996) to split the model into an *core part* $\mathbf{X}_s\boldsymbol{\beta}_s$, and *ignored part* $\mathbf{X}_u\boldsymbol{\beta}_u$. From here, we apply relevant prior distributions to both the core covariates $\boldsymbol{\beta}_s$, as well as an asymptotically valid approximation to the ignored part. Any valid approximation can be used here, but in this research we use $\mathbf{P}_k\mathbf{V}_k^T\boldsymbol{\theta}_k$, where $\mathbf{P}_k$ and $\mathbf{V}_k$ are found using the singular value decomposition of $\mathbf{X}_u$. This allows us to form a posterior distribution from which we can make relevant estimations, such as point estimates $\boldsymbol{\beta}_s^*$. From these estimates, we can also make valid predictions, and evaluate their performance.

IMPLEMENTATION

The posterior distribution which we wish to find for our data has no analytic form. This is due to the use of a horseshoe prior (Carvalho et al. 2008) for the approximation to the ignored covariates $\boldsymbol{\theta}_k$. This prior is ilicited due to the preferable performance of the horseshoe when the data is sparse, as it will be due to the fact these covariates are not selected by the lasso. We thus implement a Markov Chain Monte Carlo simulation with 1000 burn iterations and 5000 simulated iterations. The programming language R is used to run these simulations, with the outputted posterior distribution being used to find estimates for selected covariates $\boldsymbol{\beta}_s^*$ and posterior variance σ^*.

FURTHER RESULTS

We also evaluate the prediction performance of this new BE-PoS method, by calculating the relevant mean squared prediction error (MSPE). We compare this to a number of existing methods as illustrated in the figure below. We observed that, especially when n increases, the prediction error of the new method outperforms that of the other methods.



SIMULATION STUDIES

We conduct simulations to evaluate the finite-sample properties of the newly devised method. We generate 100 datasets from a high dimensional linear regression model, and evaluate results for $n \in \{100, 200, 300\}$ and $p \in \{500, 1000, 5000, 10000\}$. We calculate the average length of the credible intervals devised by the new method, with a segment of the full results given in the table below

n	5000	10000
100	5.81	5.08
200	3.25	3.95
300	2.54	2.32

As we see, as n increases, the length of the credible intervals decreases quickly towards zero, which is a very desirable statistical property. We also calculate the mean absolute bias of the new method (BE-PoS) compared to existing methods the lasso and ridge regression. The results show that the bias of this new method is much lower than of existing methods, as is represented by the figure below. This emphasises the preferable estimation performance of this new method when compared to the methods used currently.

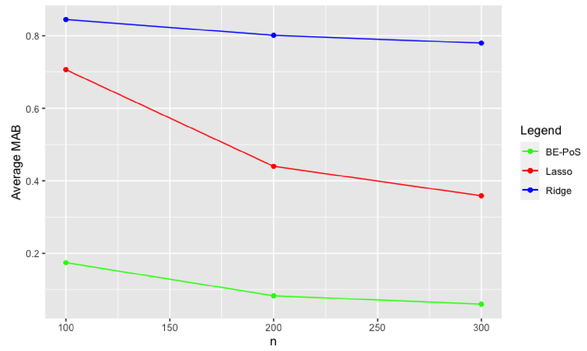


Figure 1: Example mean absolute bias between different methods

CONCLUSION

The Bayesian enriched post-selection method is a new approach for estimation, prediction and statistical inference for high dimensional data. The newly proposed method performs well and is superior to the methods evaluated throughout our simulations. When considering the whole results, we see that this conclusion is particularly true when the data is non-sparse (for example $s_0 = 50$) and when we have high dimensions (such as $p = 10000$). The reason for this performance is due to the incorporation of new information from the unselected covariates being used to enhance the original model leftover from the variable selection procedure. Our empirical results demonstrate the good performance of the post-selection credible intervals, with the length of such intervals decreasing asymptotically to zero being a particularly useful result. Furthermore, the general nature of the method allows for different priors and variable selection procedures to be used if we are dealing with a specific type of data. This avenue for future research is very interesting, and will allow for the method to be improved and generalised to provide valid and preferable performance under a wide range of circumstances and models.

REFERENCES

Carvalho, C. M., Polson, N. G. and Scott, J. G. (2008). Handling Spar- sity via the Horseshoe. Journal of Machine Learning Research, W and CP 5 73– 80.

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

FUTURE RESEARCH

The proposed approach is general due to the fact any variable selection procedure for high dimensional data can be used. Our procedural choices are centred around the data which we encounter being sparse, however under certain conditions, different choices of prior distributions could also be used. Relevant examples which could be implemented include the elastic net for variable selection, and normal priors for all covariates. The idea of Bayesian enriched post-selection models can be further extended to other important classes of models, including additive models and generalised linear models. Post-selection classification can then be used in these settings instead.

CONTACT INFORMATION

Web https://laidlawscholars.network/users/daniel-harvey-liddell

Email daniel.h.liddell@durham.ac.uk

Supervisor Reza Drikvandi

Email reza.drikvandi@durham.ac.uk