Bayesian Adaptive Selection under Prior Ignorance*

Tathagata Basu TATHAGATABASUMATHS@GMAIL.COM
Matthias C. M. Troffaes MATTHIAS.TROFFAES@DURHAM.AC.UK
Jochen Einbeck JOCHEN.EINBECK@DURHAM.AC.UK
Department of Mathematical Sciences, Durham University, UK

Regression analysis is an important part of statistical modelling and is used in many different real-life problems. In this context, we consider a problem to be high dimensional if the number of co-variates present in the model is more than the total number of observations. Let $y := (y_1, \ldots, y_n)^T$ denote the vector of n real valued responses and $\mathbf{x} := [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$ denote the corresponding $n \times p$-dimensional design matrix of the predictors. Then for a vector of regression coefficients $\beta := (\beta_1, \ldots, \beta_p)^T$, a linear model is given by:

$$y = \mathbf{x}\beta + \varepsilon \tag{1}$$

where $\varepsilon := (\varepsilon_1, \ldots, \varepsilon_n)^T$ is a vector of the noises which are assumed to be normally distributed. For high dimensional problems, $p > n$ leads to potential difficulties in the parameter estimation and we wish to perform a variable selection to overcome this.

Variable selection problems are well investigated in the Bayesian paradigm. These Bayesian variable selection methods are usually developed on the basis of posterior contraction rates of the regression coefficients. However, these posterior contract rates are based on the asymptotic behaviours of the model and often focused only on learning from the data and neglect prior elicitation. This can be problematic as high dimensional problems often lack the necessary information to perform a Bayesian analysis and expert opinions may differ. Robust Bayesian approach uses a set of priors instead of a single prior, which helps us to accommodate available information on the selection of a predictor in an efficient manner and therefore we propose the following spike and slab prior [2] to specify the j-th regression coefficient:

$$\beta_j \mid \gamma_j, \sigma^2 \sim \mathscr{N}\left(0, \sigma^2\tau^2_{\gamma_j}\right); \qquad \gamma_j \mid q_j \sim \operatorname{Ber}(q_j); \qquad q_j \sim \operatorname{Beta}(s\alpha_j, s(1-\alpha_j)); \qquad \sigma^2 \sim \operatorname{InvGamma}(a,b) \tag{2}$$

where $\alpha_j \in \mathscr{P}_j \subseteq (0,1)$. Here γ_j works as a selection indicator for the j-th regression coefficient. We fix a sufficiently small τ_0^2 so that the probability mass is concentrated around zero when $\gamma_j = 0$ and fix $\tau_1^2 \geq 1$ when $\gamma_j = 1$. We incorporate our prior information about the selection of a co-variate through $\mathscr{P}_j$. For complete ignorance, we use a near vacuous case to specify this $\mathscr{P}_j$, so that $\mathscr{P}_j = [\varepsilon, 1-\varepsilon]$, where $1 \gg \varepsilon > 0$.

Our hierarchical model gives us closed-form expressions for the conditional distributions of the modelling parameters and therefore we can easily compute the posteriors through Gibbs sampling [1]. For variable selection we use a robust decision rule on the posterior odds for the selection indicators. We consider γ_j

$$\text{to be inactive if, } \sup_{\alpha_j \in \mathscr{P}_j} \frac{P(\gamma_j = 1 \mid y)}{P(\gamma_j = 0 \mid y)} < 1; \quad \text{to be active if, } \inf_{\alpha_j \in \mathscr{P}_j} \frac{P(\gamma_j = 1 \mid y)}{P(\gamma_j = 0 \mid y)} > 1; \quad \text{to be indecisive otherwise.} \tag{3}$$

This robust decision rule allows a predictor to be indecisive during variable selection. This allows us to interpret the underlying imprecision on two different levels: one being the imprecision in variable selection and the other in model fitting through the set of posteriors for the regression coefficients. To illustrate this, we use synthetic datasets and real datasets and compare with other with Bayesian methods for variable selection.

References

[1] Alan E. Gelfand and Adrian F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990. ISSN 01621459. URL http://www.jstor.org/stable/2289776.

[2] Hemant Ishwaran and J. Sunil Rao. Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, 33(2):730–773, 2005. ISSN 0090-5364. doi: 10.1214/009053604000001147.

* This work is funded by the European Commission's H2020 programme, through the UTOPIAE Marie Curie Innovative Training Network, H2020-MSCA-ITN-2016, Grant Agreement number 722734.