25th International Congress on Modelling and Simulation, Darwin, NT, Australia, 9 to 14 July 2023
mssanz.org.au/modsim2023

A novel Bayesian variable selection algorithm for multiple linear regression

L.M. Rodrigo [a,b], **R. Kohn** [a,d], **S. Cripps** [b,c] **and M.J. Cleary** [a,b]

[a] *ARC Training Centre in Data Analytics for Resources and Environments, South Eveleigh, Australia*
[b] *School of Aerospace, Mechanical and Mechatronic Engineering, The University of Sydney, Australia*
[c] *Human Technology Institute, University of Technology Sydney, Australia*
[d] *School of Business, University of New South Wales, Australia*
Email: urod2079@uni.sydney.edu.au

Abstract: Decision making in natural resources management is driven by model complexity and data diversity rather than by data volume. Indeed, while new sensor technology and increased computational power have improved our ability to capture and store data, our understanding of these systems has not increased accordingly. This is because the amount of truth remains fixed, irrespective of the amount of data we collect, and the challenges are the identification of useful data and the construction of models that aid our understanding of the natural world. Variable selection is one of the most common strategies in identifying useful data by determining which variables are important to understand the mechanism behind such complex natural systems.

This study proposes a novel Bayesian variable selection technique for natural resource problems in a high dimensional setting where data is sparse. Our proposed method is motivated by the spike and slab prior. The spike and slab prior places a mixture distribution over regression coefficients by the introducing a vector of discrete indicator variables, the purpose of which is to allow the regression coefficients to be identically zero, with some probability. In this study, we modify this prior by replacing the discrete indicator variable by a variable with a Beta distribution and placing a hyper-prior over the parameters of the Beta distribution. We show that the spike and slab prior is recovered in the limit that the parameters of the Beta distribution approach 0. Additionally, it is shown that different values of these parameters produce other well-known priors used for variable selection such as the Horseshoe prior. We estimate the model using MCMC, where the transition kernel is a Metropolis-Hasting step with a mixed proposal distribution, which is a combination of a deterministic move and a random walk move.

We also study the frequentist properties of this technique via simulation and show that our new method outperforms a model which uses the Horseshoe prior as well as the model with the spike and slab prior. We demonstrate the efficiency of the new approach in identifying the key predictors of monthly rainfall in NSW.

Keywords: *Bayesian variable selection, multivariate Gaussian mixture prior, inclusion probability of model coefficients, multiple linear regression*