Sparse Bayesian Modelling for Categorical Predictors

Daniela Pauger

JKU Linz

The usual strategy to include a categorical covariate in a regression type model is to define one of the levels as the baseline and to introduce dummy variables for all other levels. As this can result in a high-dimensional vector of regression effects, methods which allow sparser representation of the effect of categorical covariates are required. In contrast to metric predictors, sparsity for a categorical predictor cannot only be achieved by restricting regression coefficients to zero but also when two or more of its levels have the same effect. Routine application of variable selection methods is therefore inappropriate as these allow only selection of single regression coefficients.

We achieve a sparse representation of the effect of a nominal predictor by defining informative prior distributions. The specification of a spike and slab prior on level effect differences allows classification of these differences as (practically) zero or non-zero. Thus, we can decide whether

1. a categorical predictor has no effect at all,
2. some (all) level effects are non-zero and/or
3. some (all) categories can be fused as they have essentially the same effect on the response.

Additionally we consider a modification of the standard spike-and slab prior where the spike at zero is combined with a slab distribution which is a location mixture distribution. Model-based clustering of the effects during MCMC allows to detect levels which have essentially the same effect size.

We demonstrate the performance of the developed methods in simulation studies and for real data.