Master thesis proposal, spring 2017

Bayesian Variable Selection
in
Linear Mixed Models

Bertil Wegmann and Linda Wänström

14 November 2016

1 Introduction

Data on subjects within groups (e.g. children in families, students in classes) are commonly collected for analysis in epidemiology, biology, neuroimaging, sociology, psychology and economic sciences. The subjects-within-groups data consists of measurements collected for each subject within each group. This implies within-group dependence in the measurements, which justifies the regression coefficients in a regression model to be group-specific. A regression model with only group-specific regression coefficients is called a random effects model. However, there might also be variables that are correlated with the population-averaged response, which implies fixed effects components that are not group-specific. Therefore, a regression model for subjects-within-groups data typically consists of a mixture of both fixed and random effects components to form a linear mixed-effects model.

2 The linear mixed-effects model

Consider first an ordinary multiple linear regression model with only fixed effects:

$$y_{ij} = \alpha + \beta x_{ij}^{f} + \epsilon_{ij},$$

where y_{ij} is the observed response for subject j in group i, β is the vector of regression coefficients $(\beta_1, \beta_2, \ldots, \beta_k)'$ to the vector of k explanatory variables $x_{ij}^{f} = (x_{1ij}, x_{2ij}, \ldots, x_{kij})'$ for the fixed effects and $\epsilon_{ij} \sim N\left(0, \sigma_{ij}^2\right)$. In this model the effect from each explanatory variable is assumed to be the same for each group i. Now, add random effects to the linear regression model to account for group-specific regression coefficients. Then, the model becomes a linear mixed-effects model, defined by

$$y_{ij} = \beta x_{ij}^{f} + \alpha_i + \beta_i x_{ij}^{r} + \epsilon_{ij},$$

where α_i is the group-specific intercepts in the model and $\beta_i = (\beta_{1i}, \beta_{2i}, \ldots, \beta_{gi})$ is the vector of group-specific regression coefficients to the vector of g explanatory variables $x_{ij}^{r} = (x_{1ij}, x_{2ij}, \ldots, x_{gij})'$ for the random effects. At the second level of the model structure the group-specific regression coefficients α_i and β_i can be modeled as a linear combination of group-specific variables, which forms the hierarchical modeling structure.

3 Bayesian variable selection in linear mixed-effects models

The linear mixed-effects model can be analysed by using both frequentistic and Bayesian methods. However, even if one can obtain accurate p-values for likelihood ratio tests or score tests with frequentistic methods, it is not clear how to use such methods to account for uncertainty in selecting the best subset of explanatory variables for the random and fixed effects. The rescue is fortunately Bayesian methods by a number of practical advantages. Bayesian methods do not need to rely on asymptotic approximations to the marginal likelihood or to the distribution of the likelihood ratio test statistic. In addition, Bayesian methods allows to incorporate prior information to account for full model uncertainty that assigns each model prior and posterior probabilities.

There is a rich literature on both Bayesian and frequentistic methods for selecting explanatory variables in the linear regression model with only fixed effects, but variable selection for the linear mixed-effects model has received much less attention. Nevertheless, to fully account for within-group dependence in subjects-within-groups studies, one needs to rely on estimating a mixed-effects model. Bayesian variable selection is a Bayesian method for selecting the best subset of explanatory variables in a regression model. This area of research has received a lot of attention lately, but there have only been a few studies in Bayesian variable selection for the fixed and random effects in the linear mixed-effects model.

Most work in linear mixed-effects models have focused on the variance of the random-effects, where a very small variance around 0 for the posterior distribution of a group-specific regression coefficient implies no random-effect. However, this approach has received most attention for high-dimensional problems and might be too simplified for a low-dimensional random effect. Therefore, it has been suggested to implement other methods for Bayesian variable selection by using priors such as shrinkage or spike-and-slab smoothing priors.

4 Goal of the master thesis

There are several venues for this master thesis to take form. The ultimate goal would be to develop a new method for Bayesian variable selection in linear mixed-effects models for at least low-dimensional random-effects. Other goals of the thesis include extending current methods to the better or comparing current methods in this field to identify prons and cons of the current methods. The Bayesian variable selection methods will be both contrasted on real and simulated data.

The real dataset is a stratified random sample of males and females in the U.S. in 1979 (NLSY79) in which the children of the original families have been assessed biannually since 1986. A cross-section of the data, containing the children, could be used for this project. Because the children are grouped within families, models that account for this dependency among siblings, such as mixed models, should be used. The dataset includes a vast number of variables both on child-level and on family-level, such as child IQ, mother IQ, family socioeconomic status, age of mother at first child, number of siblings of the child, urban/rural residence, race, sex etc.