-MSc thesis project proposal-
Variable selection in mixed model

Adeline Leclercq-Sanson, LJK (adeline.leclercq-samson@imag.fr)
Sophie Lambert-Lacroix, TIMC (sophie.lambert-lacroix@imag.fr)

Biological or physiological data may have a group structure. Let us detail two main examples: 1/ Measuring the same biological or physiological time process from a population of subjects creates longitudinal data with groups that are defined by the subjects. 2/ Observing repeatedly an experiment realized with different inputs (e.g. a neuron stimulated with different current levels) creates grouped data that are defined by the experimental conditions.

Our main statistical objective is to estimate a trend or a behavior from these repeated data with a group structure. The second objective may be to study the influence of covariates (variable selection).

Statistics on these grouped data is difficult because data within the group are not independent. Then standard statistical methods do not apply to the whole dataset. Each group could be analyzed separately but the number of parameters to estimate increases quickly with the number of groups and covariates. Moreover, one could be interested in estimating both the inter-group and intra-group variabilities that could only be modeled analyzing data from all the groups simultaneously.

This can be done by introducing random parameters in the statistical model, yielding the framework of mixed models that contain fixed and random parameters [1]. Random parameters model the inter-group variability and share the same probability distribution. The intra-group variability is modeled by the usual regression error. In mixed model, only the (parametric) distribution of the random parameters is estimated, and not every group parameter. When this distribution is Gaussian, the number of parameters to be estimated is then substantially reduced. This is why mixed models are so popular to analyze longitudinal, hierarchical or group structured data, especially in biology or clinical trials.

However, statistical inference is complex in mixed models because of the heterogeneity of the data. We focus in this project on maximum likelihood estimation and its derivatives.

Assuming Gaussian random effects and Gaussian error model, the likelihood is explicit but not convex in the variance parameters. Estimation is well established when the number of covariates p is small. For fixed p, Bondell have proposed sparse estimation of both the fixed and the random effects penalizing the likelihood with a lasso penalty (L1 norm) of the fixed effects and the variance parameters [2]. They prove the consistency of the estimators. The likelihood being not convex in the variance parameters, they propose a penalized EM algorithm to compute the parameter estimator. The tuning parameter of the penalty is selected by BIC on a preselected grid. However, their implementation does not always work because their criterion is not properly defined. We would like to propose alternatives by using a penalized EM algorithm that corresponds to a properly defined criterion to be optimized. We want also to study the properties of the estimator but in the diverging number of parameters context. That is the number of parameters should be large and grow with the sample size.

Remark.- Possibility of PhD.

References

[1] J. Pinheiro and D. Bates. Mixed-effect models in S and Splus. Springer-Verlag, 2000.

[2] H. Bondell, A. Krishna, and S. Ghosh. Joint variable selection for fixed and random effects in linear mixed-effects models. Biometrics, 66:1069--1077, 2010.