Model selection in colored graphical Gaussian models - Bayesian and frequentist perspectives.

For the development of multivariate data analysis it is necessary to create new specialized probabilistic tools that allow statistical analysis of the data. An example of a situation where ”classical” probabilistic and statistical tools fail is the study of genetic data, where the number of observations (individuals) is usually much smaller than the number of variables (genes). This is a typical situation in modern machine learning tasks. For this type of modeling to be possible, it is necessary to reduce the dimension of the model parameters.

Such a reduction can be achieved by imposing an appropriate conditional independence structure, thus defining the so-called *graphical model*. Graphical models are a fundamental tool used to formulate ML problems and are on the intersection of statistics, computer science, probability theory and graph theory. They form the basis of modern statistical methods used in areas such as medical diagnosis, image and speech recognition, natural language processing and many others.

In graphical models, vertices of a graph represent random variables, while its edges correspond to relationships between these variables. A conditional independence structure (a Markov structures) is imposed on the joint distribution of the random vector. Such structures are conveniently described by graphs and allow to significantly reduce the dimensionality of the problem - the number of parameters should then be related to the size of the largest clique in the graph, and not to all variables as in the classical setting. However, if the graph describing the Markovian structure is not sparse enough (the size of the largest clique still far exceeds the sample size) then such procedure does not allow for a reliable estimation of the covariance matrix. In such a situation one can use *colored graphical models*. In color graphical models, in addition to the Markov structure, certain symmetries are imposed on the concentration or partial correlation matrices. Three types of such symmetries were introduced by Højsgaard and Lauritzen to describe situations where some entries of concentration or partial correlation matrices are approximately equal These equalities are conveniently represented by a graph with colored vertices and edges. Addition symmetry to the conditional independence restrictions, reduces the number of parameters to estimate. This is especially useful when parsimony is needed, i.e. when the number of variables is substantially larger than the number of observations.

The main challenge in this field is the problem of model selection i.e. choosing the model (graph and its coloring) that best fits the observed data. There are many approaches to this problem, but few of them can be applied to high-dimensional data. Moreover, existing methods often lack theoretical foundations and offer no statistical guarantees. In our work, we intend to address two strands related to the model selection problem for colored graph models. We will study Bayesian and frequentist approaches. The results obtained will contribute both to the development of research methods and statistical tools.