1. PHD PROJECT DESCRIPTION (4000 characters max., including the aims and work plan)

Project title: Model selection in high-dimensional statistics

1.1. Project goals

In the paper [4] we introduced the model selection algorithm, which is consistent in generalized linear models. In the project we plan to adapt it to crucial statistical models:
- the Cox model from the survival analysis,
- graphical models,
- models working with groups of variables.

We will study the algorithm theoretically (by mathematical theorems with rigorous proofs) and experimentally (on simulated and real data sets).

1.2. Outline

High-dimensional statistics refers to investigating data sets, in which a number of predictors (variables) is (much) greater than a number of observations, so a number of unknown parameters of constructed models clearly exceeds a number of observed objects. Therefore, we face a difficult problem, comparable to the school example of a system of linear equations in which there are more unknowns than equations. The main task in this situation is variable selection, i.e. selecting those variables which are informative and rejecting those that are irrelevant to an observed phenomenon. High-dimensional data sets are now common in many branches of science, for instance in biology, chemistry or genetics. The analysis of such difficult data sets cannot be based on classical statistical methods, but requires new, effective and computationally efficient tools.

In the project we will consider model selection, which has a wider context than variable selection. For instance, it also contains learning a structure of a graph (i.e. to find edges in a graph knowing only values at vertices). The graphical models are often used to recognize correlations between genes, enzymes or in social networks. On the other hand, we will also study groupwise selection, which is necessary when working with highly correlated variables or categorical predictors. High-dimensionality of these problems relates to a fact that a number of considered parameters (a number of variables, groups

of variables or possible edges) in data sets significantly exceeds a number of observations.

Among many approaches to high-dimensional model selection one can distinguish a large group of methods based on penalized estimation [1]. The main representative of these methods is LASSO (Least Absolute Shrinkage and Selection Operator) [5]. There are many papers confirming good properties of LASSO in estimation and prediction, but they also indicate that this algorithm behaves poorly in model selection. Namely, it selects a model, which is usually too large and requires the restrictive conditions for selection consistency [6]. Thus, it became indisputable that LASSO should be viewed as the first step of a more complex method. Many statistical procedures were developed, whose purpose was to improve LASSO. The main refinements are: thresholded LASSO, adaptive LASSO or algorithms with nonconvex penalties. Recently, we have developed another approach [4], which combines LASSO with information criteria. Namely, first one computes LASSO and orders its nonzero coefficients. Then one chooses the final model, which minimizes a Generalized Information Criterion in a nested family induced by the ordering.

In the project we plan to extend the algorithm from [4] to crucial statistical models as the Cox model, graphical models or models working with groups of variables. It requires specific tools and methods, which are appropriate for each individual extension. For instance, in the Cox model we will work in a continuous time scenario, so we will rely strongly on the martingale theory. Obviously, graphical models also require more sophisticated methods than regression models. We plan to extend methods from [2], which work only for discrete graphs. Finally, groupwise selection demands the tools, which we have introduced in the recent paper [3].

1.3. **Work plan** is the same as **Project goals**

1.4. **Literature** *(max. 10 listed, as a suggestion for a PhD candidate)*

[1] Buhlmann, P., van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer, New York.

[2] Miasojedow, B., Rejchel, W. (2018). *Sparse estimation in Ising model via penalized Monte Carlo methods*, Journal of Machine Learning Research, 19:1-26.

[3] Nowakowski, S. Pokarowski, P. Rejchel, W., Sołtys A. (2023). *Improving Group Lasso*

for high-dimensional categorical data, In: Mikyska, J., de Mulatier, C., Paszynski, M., Krzhizhanovskaya, V.V., Dongarra, J.J., Sloot, P.M. (eds). Computational Science - ICCS 2023. Lecture Notes in Computer Science, 14074: 455-470.Springer, Cham.

[4] Pokarowski, P., Rejchel, W., Sołtys, A., Frej, M., Mielniczuk, J. (2022). *Improving Lasso for model selection and prediction*, Scandinavian Journal of Statistics, 49: 831-863.

[5] Tibshirani, R. (1996). *Regression shrinkage and selection via the lasso*. Journal of the Royal Statistical Society, Series B, 58:267-288.

[6] Zhao, P., Yu, B. (2006). *On model selection consistency of Lasso*. Journal of Machine Learning Research, 7:2541-2563.

1.5. Required initial knowledge and skills of the PhD candidate

- analytical thinking,
- readiness to self-learning,
- a good knowledge in mathematics: linear algebra, analysis, probability,
- a good knowledge in mathematical statistics: linear models, generalized linear models, penalized estimation.

1.6. Expected development of the PhD candidate's knowledge and skills

- becoming a researcher in mathematics,
- being able to provide appropriate theoretical studies and to confirm it using experimental investigations,
- having a substantial knowledge in mathematical statistics and machine learning.