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

Project title: The *Screening- Selection* algorithm in high-dimensional statistics

1.1 Project goals We plan to establish selection consistency of the Screening-Selection algorithm in

-the Cox model from survival analysis,

-graphical models,

-models with factor predictors,

-misspecified models.

We will do it theoretically (by mathematical theorems with rigorous proofs) and experimentally (on simulated and real data sets). We will also provide a software, which enables practitioners to use algorithms.

1.2 Outline

The analysis of large-scale data sets is a fundamental challenge in statistics and machine learning. High-dimensional model selection is one of the most intensively studied topics. In the project we consider model selection in two contexts: the first one is variable selection (i.e. to find variables explaining the observed phenomenon). The second approach is to learn the structure of a graph (i.e. to find edges in the graph knowing only values at vertices). The latter is often used to recognize correlations between genes, enzymes or in social networks. High-dimensionality of the problem relates to the fact that the number of considered variables (the number of possible edges, respectively) in the large-scale data sets exceeds significantly the 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. It was noticed that Lasso finds the ,,true'' model only if restrictive conditions are satisfied[1]. Many procedures were developed to improve it: adaptive Lasso, thresholded Lasso or methods with non-convex penalties. The common drawback of them is that they are non-constructive, i.e. they need values of unknown parameters to work well. In practice, this problem is overcome using cross-validation methods, but it makes them computationally complex.

Recently, the constructive procedure, called the *Screening-Selection (S-S)* algorithm, has been proposed for high-dimensional linear models in [3]. It was improved and extended in [4]. The procedure has two steps: in the first one we compute Lasso and order its nonzero coefficients according to their decreasing absolute values. In the second step we choose the final model, which minimizes the Generalized Information Criterion in a nested family induced by ordering. It was established in [4] that this procedure is computationally efficient and works better than its competitors.

In the project we plan to extend the S-S algorithm to crucial statistical models as the Cox model, graphical models, models with factor predictors or misspecified models. It requires specific tools and methods appropriate for each individual extension. For instance, in the Cox model we work in the continuous time scenario, so we will rely strongly on the martingale theory. Obviously, graphical models require more sophisticated methods than regression models. We plan to extend methods from [2], which works only for discrete graphs. Finally, the analysis of data with factor predictors demands that the algorithm is able to discard irrelevant predictors as well as "merge" unessential factor levels.

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

1.4 Literature

1. P Buhlmann, S vdGeer. Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer, New York, 2011

2. B Miasojedow, W Rejchel. Sparse estimation in Ising Model via penalized Monte Carlo methods, J MACH LEARN RES, 19:1-26, 2018

3. P Pokarowski, J Mielniczuk. Combined l1 and greedy l0 penalized least squares for linear model selection. J MACH LEARN RES, 16:961-992, 2015

4. P Pokarowski, W Rejchel, A Sołtys, M Frej, J Mielniczuk. Improving Lasso for model selection and prediction, arXiv:1907.03025, 2019

1.5 Required initial knowledge and skills of the PhD candidate

-analytical thinking

-readiness to self-study

-good knowledge in mathematics: linear algebra, analysis, probability

-basic knowledge in mathematical statistics: linear models, generalized linear models

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

-becoming a researcher in mathematics

-being able to provide appropriate theoretical study and to confirm it using experimental investigation

-substantial knowledge in mathematical statistics and machine learning

4) W. Rejchel (2017). "Oracle inequalities for ranking and U-processes with Lasso penalty", Neurocomputing, vol. 239, p. 214–222

f. List of promoted doctoral students: their titles, last names, titles of doctoral dissertations, names of universities, year and field of graduation

None

g. Description of previous (and potential) scientific cooperation with other academic centers in the last 5 years (1 page)

During last five years my scientific cooperation has been mainly related to statisticians from Warsaw. First, I obtained the "FUGA" grant from the National Science Centre and between 2014 and 2016 I was on a post-doc position at University of Warsaw. In the current project we plan to study the S-S algorithm in various important statistical models. This procedure was introduced by P. Pokarowski and J. Mielniczuk in the context of high-dimensional linear models. The improvement and extension of this procedure was the main scientific problem of my second post-doc (2017-2018) at University of Warsaw. It was within the "OPUS" grant from the National Science Centre obtained by P. Pokarowski. Now I am an investigator in another "OPUS" grant, whose principal investigator is B. Miasojedow (2018-2021). Few years ago I also started collaborating with M. Bogdan from University of Wrocław.

Below I give the list of papers, which were obtained during this cooperation and which are closely related to the problems considered in the current project. The potential cooperation is also described.

a) B. Miasojedow, W. Rejchel. Sparse estimation in Ising Model via penalized Monte Carlo methods, J MACH LEARN RES, 19:1-26, 2018.
It applies successfully the S-S algorithm to discrete graphs. In the project we plan to obtain similar results for continuous graphs. We have obtained initial experimental results on this problem with P. Pokarowski and D. Ambroziak. We will continue this cooperation in the future to complete experimental results and to obtain theoretical ones.

b) P. Pokarowski, W. Rejchel, A. Sołtys, M. Frej, J. Mielniczuk. Improving Lasso for model selection and prediction, arXiv:1907.03025, 2019.
It improves the S-S algorithm and extends it to generalized linear models and a quite general class of *M*-estimators. Initial experimental results on the first and third problem from the section "Project goals" are obtained with P. Pokarowski and A. Sołtys. We plan to complete them and obtain theoretical ones.

c) W. Rejchel, M. Bogdan. Rank-based Lasso - efficient methods for high-dimensional robust model selection, arXiv:1905.05876, 2019.
d) K. Furmańczyk and W. Rejchel. Prediction and variable selection in high-dimensional

misspecified binary classification, Entropy, 2020, accepted.
Papers c) and d) contain results on estimation consistency of Lasso estimators in misspecified models. So, they relate to the first step of the S-S algorithm. They can be viewed as preliminary and encouraging results concerning the forth point from the section "Project goals". In the future cooperation we plan to extend them to the S-S algorithm.

4. DECLARATION OF TECHNICAL/EXPERIMENTAL/FINANCIAL RESOURCES SUFFICIENT AND NECESSARY TO COMPLETE THE PROJECT

I declare that the project can be completed using technical and experimental resources, which are presently available to PhD students on the Faculty of Mathematics and Computer Science NCU, for instance the access to the library or a computer. Completing the project does not require any additional technical and experimental resources.

We plan to apply for additional financial resources from NCU (for instance, from ID-UB) or from outside NCU (for instance, from NCN). They could be used, for example, to enable the PhD student to participate in domestic and international scientific conferences.

5. DECLARATION CONCERNING THE AUTHORSHP OF PROJECT IDEA

I declare that the author of the idea for the doctoral project is:

…… Aleksander Zaigrajew and Wojciech Rejchel……

6. DECLARATION CONERNING THE POSSIBILITY OF PUBLISHING THE CONTENT OF THE PROJECT

I declare that the description of the project submitted do the contest from point 1. can be published on the website of Doctoral School of Exact and Natural Sciences, Nicolaus Copernicus University in Toruń.

Toruń, 14.05.2020

place, date signature of project submitter