Discrete Optimization for Robust Feature Selection

Luca Insolia
EMbeDS, Sant'Anna School of Advanced Studies, Pisa, 56127, Italy
IASI "Antonio Ruberti", National Research Council, 00185 Rome, Italy
Luca.insolia@santannapisa.it

Abstract. Modern regression problems are increasingly complex and often comprise a large number of features. The larger a study, the higher the likelihood that a substantial portion of the features may be redundant and/or contain outlying values that can hinder classical estimation methods. Focusing on linear models, we contribute to this area developing a general framework for simultaneous feature selection and outlier detection based on *mixed-integer programming* techniques [1]. It is robust against outliers in the response and/or the design matrix, and provides optimality guarantees from both optimization and theoretical standpoints. This is also extended to classification problems through the logistic regression model and *mixed-integer conic programming* techniques [2]. Its generalizations to clusterwise regression problems, where non-outlying data points belong to different regression structures, are also developed. Furthermore, computational efficiency is of utmost importance in these applications. We thus develop approximate estimation methods that can provide sub-optimal, high-quality solutions in a very efficient manner. We show the superior performance of our proposals compared to existing methods through simulations and real-world applications related to the "Omics" sciences and entomology.

Keywords: mixed-integer programming; robust regression; sparse estimation

References

[1] Luca Insolia, Ana Kenney, Francesca Chiaromonte, and Giovanni Felici. Simultaneous feature selection and outlier detection with optimality guarantees. *Biometrics*, Accepted manuscript, 2021. doi: https://doi.org/10.1111/biom.13553.

[2] Luca Insolia, Ana Kenney, Martina Calovi, and Francesca Chiaromonte. Robust variable selection with optimality guarantees for high-dimensional logistic regression. *Stats*, 4(3):665–681, 2021. doi: 10.3390/stats4030040.