A model robust design approach for optimally subsampling big data

- Subsampling methods have been proposed as a computationally efficient approach to analyse big data.
- A key question in subsampling is how to obtain an informative subset based on the questions being asked of the big data.
- For this purpose, random sampling has been proposed based on subsampling probabilities determined via methods from optimal experimental design.
- Drawback of this approach is that the subsampling probabilities can rely on an assumed model for the big data.
- We propose a model robust approach, where a set of models is instead considered, and the subsampling probabilities are evaluated based on the weighted average of the probabilities that would be obtained if each model was considered singularly.
- Our model robust approach outperforms current subsampling methods through providing informative data for estimating a range of potential models for the big data according to the simulation study and real-world applications.

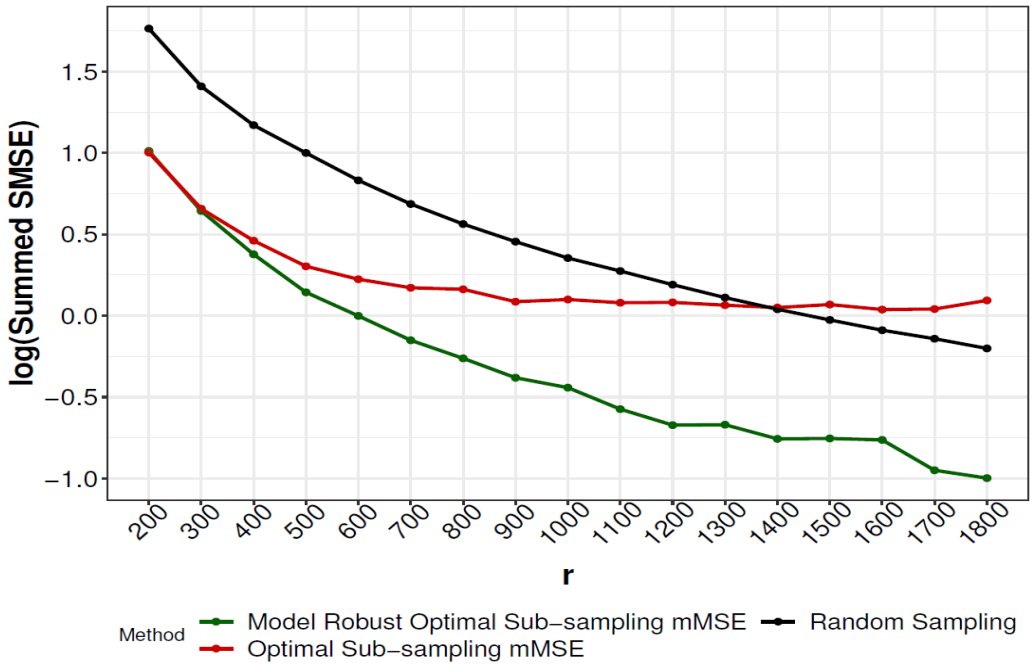

Fig 1: Logarithm of summed simulated mean squared error over the available models for logistic regression applied on the “Skin segmentation” data.