Addressing the Rashomon Effect through ranking aggregation

C. Sessa[†,1], E. Borgonovo[2], A. Cillo[§,1], G.P. Crespi[§,1], X. Lu[3]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Dec. 2022 – Dec. 2025

[1] LIUC Business University, Castellanza, Italy
{csessa, acillo, pcrespi}@liuc.it

[2] Department of Decision Sciences, Bocconi University, Milan, Italy
emanuele.borgonovo@unibocconi.it

[3] SKEMA Business School, Université Côte d'Azur, Paris, France
xuefei.lu@skema.edu

Abstract

When dealing with prediction problems, analysts rely on variable importance measures and global sensitivity measures to understand the predictive power of variables and uncover the relationships in the data [10]. When the data generating process (DGP) is unknown, analysts typically train machine learning models to use as surrogates, and derive explanations for the patterns in the data computing the variable importance of the best performing model. The validity of this approach is threatened by the Rashomon Effect [2], whereby multiple models achieve similar predictive accuracy but offer different and sometimes conflicting explanations for the underlying patterns. Indeed, the Rashomon Set [5] – the collection of all almost-optimal prediction models – can be seen both as a challenge and an opportunity for analysts: while this adds uncertainty to inference, it also allows for broader exploration of potential explanations.

A number of studies have succeeded in framing a procedure to compute or approximate the Rashomon Set for some specific model classes [11, 12, 4, 9]. Few attempts, however, have been made to explain the relationships in the data by exploiting the whole Rashomon Set [5, 4]. In this work, we propose a novel methodological framework that leverages all the models in the Rashomon Set to produce more reliable and consistent insights into variable importance. Our idea is to view the Rashomon Set for a dataset as a collection of agents, each expressing its own possibly different preference for the features, much like how different experts may offer varying interpretations of the same data. The strength of this preference corresponds to the importance of each variable for the prediction, quantified through an importance measure. By transforming the importance vectors for all the models into rankings and then aggregating them, our method allows analysts to generate a consensus ranking which reflects the preferences of the entire Rashomon Set, offering a comprehensive view on the mechanisms in the data. We draw upon the established literature on ranking aggregation techniques [3, 6, 8] to combine the individual importance rankings into a unified ranking that is robust to model variability.

The proposed framework complements existing variable importance measures and provides analysts with a powerful tool to handle model multiplicity in practical applications. We validate our methodology using both simulated data from known DGPs and real-world datasets, to demonstrate how the framework reconciles conflicting signals from multiple models and produces an

importance ranking of variables that is more aligned with the true DGP. We test different aggregation techniques to show how the choice of the technique impacts the consensus ranking. Furthermore, we provide theoretical results on the structure of the Rashomon Set for the specific class of linear regression models. In particular, we clarify the connection between the coefficients of linear models in the Rashomon Set and the permutation importance measure [1], a widely used measure in machine learning, exploring its relation to total indices [7].

Short biography (PhD student)

Claudia Sessa is a PhD Candidate at LIUC Business University in Castellanza, Italy. Previously, she obtained a MSc in Data Science and Business Analytics (cum laude) and a BSc in Economics, Management and Computer Science, both from Bocconi University in Milan. Her research lies at the intersection of operations research and machine learning.

References

[1] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[2] Leo Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3), August 2001.

[3] Wade D. Cook. Distance-based and ad hoc consensus models in ordinal preference ranking. *European Journal of Operational Research*, 172(2):369–385, July 2006.

[4] Jiayun Dong and Cynthia Rudin. Exploring the cloud of variable importance for the set of all good models. *Nature Machine Intelligence*, 2(12):810–824, 2020.

[5] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

[6] Dorit S. Hochbaum and Asaf Levin. Methodologies and algorithms for group-rankings decision. *Management Science*, 52(9):1394–1408, September 2006.

[7] Toshimitsu Homma and Andrea Saltelli. Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering & System Safety*, 52(1):1–17, April 1996.

[8] Xue Li, Xinlei Wang, and Guanghua Xiao. A comparative study of rank aggregation methods for partial and top ranked lists in genomic applications. *Briefings in Bioinformatics*, 20(1):178–189, August 2017.

[9] Kota Mata, Kentaro Kanamori, and Hiroki Arimura. Computing the collection of good models for rule lists. In *Proc. the 18th International Conference on Machine Learning and Data Mining (MLDM 2022)*, 2022.

[10] Brian D. Williamson, Peter B. Gilbert, Noah R. Simon, and Marco Carone. A general framework for inference on algorithm-agnostic variable importance. *Journal of the American Statistical Association*, 118(543):1645–1658, January 2022.

[11] Rui Xin, Chudi Zhong, Zhi Chen, Takuya Takagi, Margo Seltzer, and Cynthia Rudin. Exploring the whole rashomon set of sparse decision trees. *Advances in Neural Information Processing Systems*, 35:14071–14084, 2022.

[12] Chudi Zhong, Zhi Chen, Jiachang Liu, Margo Seltzer, and Cynthia Rudin. Exploring and interacting with the set of good sparse generalized additive models. *Advances in Neural Information Processing Systems*, 36, 2024.