On the Evaluation of Economic Theories using Machine Learning Techniques

Jesper Armouti-Hansen

June 24, 2024

The *completeness* of a given parametric model, as introduced by Fudenberg et al. (2022), reveals (i) how large a fraction of the predictable variation of the data the model captures, and (ii) how large a gain in predictive performance the model achieves compared to a simple baseline model. In turn, this shows the potential gains by either considering more complex functional forms or by adding additional terms to the model. The current definition (i) assumes a data abundant environment and (ii) allows for subjective definitions of the baseline. We show how to perform the estimation in a data sparse environment and propose a standardization of the baseline to be the lowest reasonable estimate. Motivated by the estimation of mixtures of utility models (See e.g., Bruhin et al. 2010), we expand on the completeness concept to allow for unobserved heterogeneity. Using panel data from the lab containing experimental observations of binary dictator games and reciprocity games from Bruhin et al. (2019), we show how to apply these extensions.

JEL codes: C52, C53, D11–D12

Keywords: theory evaluation, machine learning, random utility, mixture models, social preferences, matrix factorization, latent factor models, recommender systems

Armouti-Hansen: Institute for Applied Microeconomics, University of Bonn; Email: armoutihansen@uni-bonn.de. An earlier version of the paper titled "Evaluating the Completeness of Social Preference Theories" was part of my PhD thesis at the University of Cologne. I am grateful to Dirk Sliwka and Marco Mariotti for helpful comments and suggestions. I also thank Adrian Bruhin, Ernst Fehr and Daniel Schunk for sharing their code and data, as well as Arne Risa Hole for sharing Stata modules on non-linear mixed logit estimation. The code is available in the GitHub repository: https://github.com/armoutihansen/EETML. Please note that the data is not publicly available.

1. Introduction

Laboratory data on choices provide the means to test whether actual decision-making matches with proposed theories of choice. In particular, the data makes it possible for us to investigate the extent to which parameterized theories, in their proposed functional form, are able to predict individuals' choices. Such an investigation sheds light on two important points that provide insights on the vary nature of decision-making on the considered domain. Firstly, given a theory's included behavioral motives, it allows us to conclude how well the theory is able to predict choices, compared to how well a theory could have predicted on the considered domain. In turn, this allows us to conclude (i) whether the proposed functional form is optimal and (ii) how much better a theory could perform by considering more complex functional forms. Secondly, it allows us to conclude the extent to which the behavioral motives included in the model matter for decision-making. On the domain of other-regarding preferences[1], such motives might, for instance, be inequity aversion or reciprocity. These two points can be sufficiently addressed in a representative agent setting.[2] Given a proposed theory of choice that matches actual decision-making is found, a third point worth investigating relates to heterogeneity of choices. In its most general form, a proposed theory allows for individual specific parameter values. However, as with the functional form of our theories, we tend to prefer sparsity in the variation in parameter values across individuals within a specific theory. In particular, we would like to know if the heterogeneity in choices across individual can be captured in the given theory by allowing for a relatively small number of types that are economically distinguishable.

In this paper, we address these three points on the domain of other-regarding preferences. We address them by evaluating simple parameterized social preference theories using data on binary dictator games and reciprocity games from Bruhin et al. (2019). The social preference theories are designed in a way that gradually increases the complexity by sequentially adding behavioral motives. Our starting point is a linear preference model, in which the decision maker (DM) only cares about her own payoff. By sequentially adding more motives, our end point is a model that includes potentially inequity aversion (or, alternatively, differentiated altruism) and both negative and positive reciprocity.[3]

The insights mentioned above stem from the predictive performance of the models. However, merely looking at the performance of a model does not reveal the whole picture. Firstly, when we construct theories, we are likely - perhaps intentionally - not including every potential motive that may influence choice. Secondly, the act of choosing might be inherently random on its own. Thus, conditional on the included variables in the theory, we should expect some randomness in choice, leading to less than perfect predictions. It

[1] 1
[2] 2
[3] 3

follows that to evaluate the predictive capability of a given model, we need a measure that informs us on how well we could optimally predict, conditional on the included variables used in the models. Such a measure would directly show us the potential improvement, in terms of predictive performance, an alternative formulated theory could bring. Hence, this also allows for the comparison of two models, such that the improvement of including a behavioral motive becomes clear. At the same time, a lower boundary, in terms of predictive performance, is needed to inform us whether the model is able to capture the relationship between the features and choice to a sufficient degree. For instance, if the performance of a model appear relatively close to optimal, it is not clear whether it sufficiently captures structure in the data unless we compare it to a naive measure that makes predictions without feature-based information, such as the unconditional average.

In order to conduct the analysis of the two aforementioned points on the representative agent level, we first translate the social preference models into prediction rules by subsuming a random utility framework in the same manner as Bruhin et al. (2019). Subsequently, we apply the concept of a model's /completeness/ as proposed by Fudenberg et al. (2022). In our setting, a given parameterized model's completeness is calculated by the improvement in predictive capability that the model brings compared to a naive benchmark model, relative to the largest possible improvement in predictive capability in the data. The naive benchmark model in our setting is based on a simple unconditional average.[4] Such a baseline provides the best estimate when no features are used. Hence, we should expect that any proposed theory that uses information to some extent will predict at least as good. To estimate the best possible predictions, we use what is referred to as the "lookup table" algorithm in Fudenberg et al. (2022). In essence, the prediction in any given game, which is uniquely defined by its features, is the average choice within that game by the subjects. We show how to estimate this algorithm in a simple regression framework, in which we can use regularization to investigate whether we have enough data, and we compare the result to conventional machine learning (ML) algorithms, which allow for (i) a non-parametric and flexible estimation of the predictive patterns in the data to see if the aforementioned algorithm indeed is optimal in the specific setting (ii) model interpretation techniques to investigate how the features are linked to choices.

Our findings show, that a full linear model that includes all the considered other-regarding behavioral motives, achieves a relatively high completeness of approximately 82%.

We subsequently extend the setting by allowing, in each model, heterogeneity in the parameters as in Bruhin et al. (2019). That is, in each of the models we allow for the existence of a finite number of types, each characterized by their own set of parameter values. To evaluate the completeness of a model in this setting, we propose and explore two

[4] 4

extensions of the original definition of completeness. The first variant that we introduce is what we call a model's /unrestricted completeness/. Here we evaluate the predictability of a heterogeneous model by comparing its predictive performance to a fully flexible ML model that uses the subject identifier as a feature. This will provide us with an indication on how well a parametric theory consisting of a parsimonious representation of individuals, in the form of types, predicts compared to a fully flexible non-parametric model that may adjust its predictions to any of the subjects. The second variant is what we refer to as a model's /within-type completeness/. Here we evaluate the completeness of a model by estimating the completeness within each type that the given model proposes. Specifically, we compare the predictive capability within the type of a given model to that of an ML model, as the estimate of the optimal predictive performance, and to that of a simple model, as the naive benchmark. Besides allowing us to estimate the partial impact of a given behavioral motive, this will allow us to infer (i) whether there is substantial variation in a model's predictive capability across the types, and (ii) whether, for some types, a more complex social preference model is needed to fully capture the within-type behavior.

The unrestricted completeness results indicate that a linear social preference model with only three types is able to capture most of the individual variation in the data. In particular, the completeness estimates range between approximately 85% and 88%. Our within-type completeness results on this domain suggest the existence of three types in all the considered models, with two relatively large ones and one minority type. The behavior of subjects belonging to the first of the large types, which can be characterized by strong other-regarding preferences, seems to be well predicted by linear social preference models, with completeness estimates ranging between 88% and 93%. The behavior of the second-type subjects is characterized by modest other-regarding preferences. However, the linear social preference theories are only able to achieve a within-type completeness of between 60% and 65%. This indicates that a more complex theory is needed to fully capture this type's behavior. Finally, for the minority type, we find that choices are very random, in the sense that only using the subjects' own payoffs for prediction leads to relative poor predictions. However, due to the type's small size, we do not have enough power to estimate the within-type completeness.

As a final point of investigation, we increase the scope for heterogeneity by allowing for an infinite number of types in each of the social preference models. We do this by imposing assumptions on the distribution of the parameters in each of the models, allowing us to overcome the limitations due to only having a finite data sample. We once again calculate the unrestricted completeness of the models by comparing these to ML models that use the subject identifier as a feature. The results will shed light on the potential reduction in predictive capability we should expect by summarizing the population into a finite number of types as opposed to allowing for individual parameter values f_θ.

The results here show that ...

In summary, this paper expands the domain of the proposed concept of model completeness proposed by Fudenberg et al. (2022) in two important ways. Firstly, it expands its application to the domain of social preferences. Secondly, as heterogeneity is near impossible to argue against, we here show how the concept can be applied once allowing for subject heterogeneity in choices.

Roadmap. The remainder of the paper is organized as follows. The next section sheds light on related literature and distinguishes our approach. The section that follows describes the setup. In this section, the primitives of our investigation will be defined, the data that we are using will be described, and the social preference models will be presented and translated into parametric models that predict the probability of a decision maker choosing one allocation over the other. Section ? describes our estimation strategy for evaluating the completeness of the models on the aggregate level and when allowing for heterogeneity, as well as for the evaluation of within-type completeness. In Section ?, we present the findings of our investigation, first on the aggregate level, and subsequently in the heterogeneous setting. Section ? discusses our findings. Finally, Section ? concludes.

2. Related Literature

Contribution to the literature on social preferences.

Contribution to the literature on theory evaluation.

Contribution to the literature on structural modelling.

3. Setup

In this section, we firstly lay out the primitives central to our investigation, based on that of Fudenberg et al. (2021), and present the concept of a parametric model's completeness, as proposed by Fudenberg et al. (2022) on the representative agent level. Secondly, we extend the concept, allowing for heterogeneity.

3.1. Primitves

3.2. Data

3.3. Utility Models

3.4. Parametric Models

4. Estimation

4.1. Data Splitting Strategy

4.2. Homogenous Estimation

Parametric Models.

Homogenous Completeness.

4.3. Heterogeneous Estimation

Parametric Models.

Heterogeneous Completeness.

Within-Type Completeness.

5. Results

5.1. Homogenous Completeness

5.2. Heterogeneous Completeness

5.3. Within-Type Completeness

6. Discussion

7. Conclusion

References

Bruhin, Adrian, Helga Fehr-Duda, and Thomas Epper. 2010. "Risk and Rationality: Uncovering Heterogeneity in Probability Distortion." *Econometrica* 78 (4): 1375–1412.

Bruhin, Adrian, Ernst Fehr, and Daniel Schunk. 2019. "The Many Faces of Human Sociality: Uncovering the Distribution and Stability of Social Preferences." *Journal of the European Economic Association* 17 (4): 1025–1069.

Fudenberg, Drew, Wayne Gao, and Annie Liang. 2021. "How Flexible Is That Functional Form? Measuring the Restrictiveness of Theories." In *Proceedings of the 22nd ACM Conference on Economics and Computation*, EC '21: 497–498. New York, NY, USA: Association for Computing Machinery.

Fudenberg, Drew, Jon Kleinberg, Annie Liang, and Sendhil Mullainathan. 2022. "Measuring the Completeness of Economic Models." *Journal of Political Economy* 130 (4): 956–990.

Appendix A. Mathematical Derivations

Appendix B. Figures and Tables

B.1. Homogenous Completeness with CES Utility Functions

B.2. Heterogeneous Completeness with CES Utility Functions