Social choice theory and data science

Social Choice Theory and Data Science: The Beginning of a Beautiful Friendship

Prepared for "The Future of Economic Design"

Ugur Ozdemir
University of Edinburgh

Abstract

In this note, I advocate the development of a symbiotic relationship between the theoretical world of economic design and that of quantitative data analysis. I illustrate both directions of this symbiosis by looking at how the axiomatic approach of social choice theory can provide foundations for the quantitative methodological choices and, how data science tools can help testing assumptions of formal models. For the former, I focus on the two important stages of any empirical research design: measurement and methods selection. For the latter, I assert that machine learning algorithms can be employed to test assumptions regarding to the nature of individual preferences. The spatial model of electoral competition serves as a workhorse example throughout the paper.

Suppose a researcher would like to use the spatial model of voting in order to shed some light on the nature of electoral competition in a given country. She is interested in comparing the game theoretical estimates of party positions on the ideological space with the empirical estimates and, in understanding how important issue voting is in determining voters' choices. For this sort of an analysis, she needs to have the following: i) policy space on which the electoral competition on taking place, ii) an estimate of the empirical distribution of the voter ideal points iii) estimates of party positions on that space.

For the first two, the researcher will use data from a representative survey experiment. The idea is to use the issue opinion questions in this survey to determine the latent ideological dimensions in the polity and respondents' scores on these dimensions. There are many different feature extraction and dimensionality reduction techniques available for this purpose. Looking at the earlier literature, the researcher decides that she should use either principal component analysis or

common factor analysis. She runs both algorithms and observes that the results differ substantively. This does not come as a surprise once she looks at a bit more closely to these techniques because even if both principal component analysis and factor analysis aim to reduce the dimensionality of a set of data, the approaches taken to do so are very much different.[Jolliffe, 2010] They assume completely different models relating the underlying variables to latent dimensions. So which method is the most appropriate? Can there be theoretical reasons to choose one method over the other? Which axioms does each of these methods satisfy? This is, after all, an aggregation problem, and this is the point where, I argue, social choice theory can help with.

The existing research on the axiomatisation of the data analysis methods is quite limited and focuses entirely on the clustering algorithms. One of the reasons for this is that this strand of work is dominated by computer scientists and data scientists and, has not received attention from social scientists yet. Very much like social choice theory, this literature started with an impossibility result. [Kleinberg, 2002] proposed three axioms[1] and showed that there can be no clustering method which satisfies all of these at the same time.

[Ben-David and Ackerman, 2009] criticise Kleinberg's result by arguing that impossibility is not an inherent feature of clustering, but rather it is an artefact of the specific formalism. In contrast to Kleinberg's setup, they propose to focus on the clustering quality measures as the object to be axiomatised rather than clustering methods and provide several clustering quality measures all satisfying the proposed axioms. Another "positive" result for clustering is by [Zadeh and Ben-David, 2009] who relax one of Kleinberg's axioms and show that there exists a unique clustering method which satisfy their axioms.

As mentioned above, the existing literature on the theoretical foundations of statistical methods is entirely on clustering procedures and there is no work on other data analysis methods frequently used by social scientists, such as feature extraction our researcher used, or classification (clustering with known, predefined categories, i.e, *supervised* clustering.). Social choice theory

[1]Here are Kleinberfg's axioms stated rather informally:

- Clustering function is not sensitive to changes in the units of distance measurement.
- Any desired clustering structure should be attainable by some distance measure.
- If we reduce distances within the clusters and enlarge distances between the clusters then the clustering structure should not change.

can contribute to providing guidelines to help empirical social scientist make better and more conscious methodological choices. This is particularly important for empirical social sciences where theoretical justification is expected to be more important than better predictions. In practice however, researchers make this choice in quite an ad-hoc manner with reasons such as, "freely available code", "ease of use", "it has worked for another paper" etc. With the increasing "super large N" datasets, this becomes even a bigger problem, because if you have a sufficiently large dataset, you can reach almost any conclusion with a "smart" choice of methods.

Let's now go back to the our spatial model example. Since our researcher now has an empirical distribution of the voters, the next step is to get the estimates for party positions on the same policy space. One standard way of doing this is to use expert surveys. In these surveys, a group of experts is asked to locate the policy positions of political parties on different issues. So she decides to use the most commonly used one in the literature [Bakker et al., 2015]. This survey includes the scores from individual experts. But what should be the method of aggregation? Simply taking the average across all experts as it is done by almost everyone? That seems to be a bit problematic because, looking at the distributions of the scores at the expert level, she observes a considerable level of disagreement among experts. Moreover, there is significant variation in this disagreement both at the party and issue dimension level. Note that statistical inference problem in expert surveys differs quite substantially from that in public opinion surveys because we are not aggregating information over a randomly selected sample [Benoit et al., 2006]. That means we cannot use the standard "confidence intervals around the mean" to deal with uncertainties.

There are in fact some statistics used in the psychometrics to assess levels of agreement and reliability among experts such as intraclass correlation. [James et al., 1984] But again, none of these measures are theoretically justified. So, we have another aggregation problem which can certainly benefit from a theoretical foundation, in particular, from the axiomatic approach that social choice has been founded on. Expert surveys are just one example of composite indices used in social sciences. There many other similar measures social scientists use such as measures of democracy, inequality, poverty, power and environmental pollution responsibility. Just like the axiomatisation of methods, axiomatisation of measurement has not received much attention from social scientists as well. An interesting recent exception is [Patty and Penn, 2015] who does argue that formal

theory, social choice theory in particular, is the heart of measurement and present an axiomatic analysis of network centrality measures to demonstrate the idea.

In order to illustrate the opposite direction of the symbiotic relationship, i.e., how data science can help us test our theoretical assumptions, let's go back to our example once again. Our researcher is now ready to run a multinomial logistic regression on voter choice[2], the distance between the voters and the parties being the independent variable of interest together with some sociodemographic variables as controls. She is however, a skeptical type and enjoys questioning her theoretical assumptions as much as the methodological ones. One thing catches her eye is the Euclidean preferences assumption she made. This assumption is operationalised through the choice of the metric used to measure the distances between the political parties and the individuals. By employing the Euclidean metric, she realizes that she is assuming circular indifference curves. Is this really a reasonable assumption? How can we test this. The brand new tools of machine learning can help us test this assumption and help us choose "the best distance metric" [L. Yang, 2006]. We can use supervised metric learning algorithms on a subsample of our dataset (training data) to see which metric best explains the underlying behaviour, and use that metric to run the regression analysis instead of simply assuming the Euclidean metric.

Conclusion

We have transitioned into a world of complex, multidimensional data which changed the lives of empirical social scientists dramatically. This has increased the need for new analytical and statistical techniques and, aggregated measures to reduce the inherent complexity and discover the patterns buried in the data. In fact, an entirely new research field -that of data science- was born. This world is mainly dominated by computer scientists and statisticians, but it is transforming from a set of mysterious techniques to an essential toolkit for a much broader social science community.

[2]Why logistic regression? Note that, similar to our investigation regarding to principal component analysis versus factor analysis above, we might want to question this methodological choice as well. After all, logistic regression is nothing but a classification algorithm. There are other methods such as random forests or support vector machines suitable for the same task. So, the axiomatic study of quantitative methods can go as far as to include regression analysis.

In this note, I pointed into a direction for a rather surprising collaboration between social choice theory and data science. I argued that social choice theory can help bridging the gap between empirical social scientists and statisticians in order to strengthen the theoretical basis of the social scientific enquiry through development of an axiomatic understructure. This is increasingly becoming important as ad-hoc choices of methods and measures with no theoretical justification can transform empirical social sciences into data mining in the age of big data. I further argued that brand-new data science algorithms can provide opportunities to test the assumptions of theoretical models.

I see this as the beginning of a beautiful friendship between two seemingly unrelated disciplines and looking forward to being a part of this exciting journey.

References

[Bakker et al., 2015] Bakker, R., De Vries, C., Edwards, E., Hooghe, L., Jolly, S., Marks, G., Polk, J., Rovny, J., Steenbergen, M., and Vachudova, M. A. (2015). Measuring party positions in europe: The chapel hill expert survey trend file, 1999–2010. *Party Politics*, 21(1):143–152.

[Ben-David and Ackerman, 2009] Ben-David, S. and Ackerman, M. (2009). Measures of clustering quality: A working set of axioms for clustering. In *Advances in neural information processing systems*, pages 121–128.

[Benoit et al., 2006] Benoit, K., Laver, M., et al. (2006). *Party policy in modern democracies*. Routledge.

[James et al., 1984] James, L. R., Demaree, R. G., and Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of applied psychology*, 69(1):85.

[Jolliffe, 2010] Jolliffe, I. T. (2010). *Principal Component Analysis*. Springer.

[Kleinberg, 2002] Kleinberg, J. (2002). An impossibility theorem for clustering. In *NIPS*, volume 15, pages 463–470.

[L. Yang, 2006] L. Yang, R. J. (2006). Distance metric learning: a comprehensive survey. Technical report, Michigan State University.

[Patty and Penn, 2015] Patty, J. W. and Penn, E. M. (2015). Analyzing big data: social choice and measurement. *PS: Political Science & Politics*, 48(01):95–101.

[Zadeh and Ben-David, 2009] Zadeh, R. B. and Ben-David, S. (2009). A uniqueness theorem for clustering. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 639–646. AUAI Press.