Performance comparison of the discrete choice models of consumer choice

Exploration of the Econometrics and Machine Learning models' performances in the presence of heterogeneous preferences and random effects utilities

Research master thesis

Nikita Gusarov
Master 2
MIASHS C2ES (UGA)

Under supervision of:
Iragaël Joly, HDR (GAEL, UGA, Grenoble INP)
Beatrice Roussillon, MCF (GAEL, UGA)

Université Grenoble Alpes
Faculté d'Economie et Gestion - FEG

2019 - 2020

Abstract: This works is a cross-disciplinary study of econometrics and machine learning (ML) models applied to consumer choice modelling. To breach the interdisciplinary gap an integrated simulation and theory-testing framework is proposed. It incorporates all essential steps from hypothetical setting generation to the comparison of various performance metrics.

The flexibility of the framework in theory-testing and models comparison over economics and statistical indicators is illustrated based on the work of Michaud, Llerena and Joly (2012). Two datasets are generated using the predefined utility functions simulating the presence of homogeneous and heterogeneous individual preferences for alternatives' attributes. Then, three models issued from econometrics and ML disciplines are estimated and compared.

This study shows the proposed methodological approach's efficiency, successfuly capturing the differences between the models issued from different fields given the homogeneous or heterogeneous consumer preferences.

Key words: Consumer Choice, Preference Studies, Willingness to Pay, Econometrics, Data Science, Machine Learning, Classification Techniques, Synthetic Datasets

Author: Nikita Gusarov (UGA)

Under supervision of: Iragaël Joly, HDR (GAEL, UGA, Grenoble INP); Beatrice Roussillon, MCF (GAEL, UGA)

Abstrait: Ce travail est une étude interdisciplinaire des modèles d'économétrie et d'apprentissage automatique (ML) appliqués à la modélisation des choix des consommateurs. Pour briser la frontière interdisciplinaire, un cadre intégré pour tester des théorie est proposé. Il intègre toutes les étapes essentielles de la génération de paramètres hypothétiques à la comparaison de diverses mesures de performance.

La flexibilité du cadre dans les tests de théorie et la comparaison de modèles par rapport aux indicateurs économiques et statistiques est illustrée à partir des travaux de Michaud, Llerena et Joly (2012). Deux ensembles de données sont générés à l'aide des fonctions d'utilité prédéfinies simulant la présence de préférences individuelles homogènes et hétérogènes pour les attributs des alternatives. Trois modèles issus des disciplines économétrie et ML sont ensuite estimés et comparés.

Cette étude montre l'efficacité de l'approche méthodologique proposée, en captant avec succès les différences entre les modèles issus de différents domaines compte tenu des préférences homogènes ou hétérogènes des consommateurs.

Mots clés: Choix du consommateur, Études de Préférences, Consentements à Payer, Économétrie, Science des Données, Apprentissage Automatique, Techniques de Classification, Données Synthétiques

Acknowledgements

This work was accomplished with financial aid from Multidisciplinary Institute in Artificial Intelligence (MIAI), supported by Sihem Amer-Yahia, head of the SLIDE team at the LIG laboratory.

I would like to express my gratitude for the administrative and technical support from the Grenoble Informatics Laboratory (LIG) and Grenoble Applied Economics Laboratory (GAEL), which helped to fulfil this work during COVID-19 crisis.

Credits for dataset generation algorithm go to Amirreza Talebijamalabad, first year master student at Grenoble INP, who worked on the theory of the artificial datasets generation.

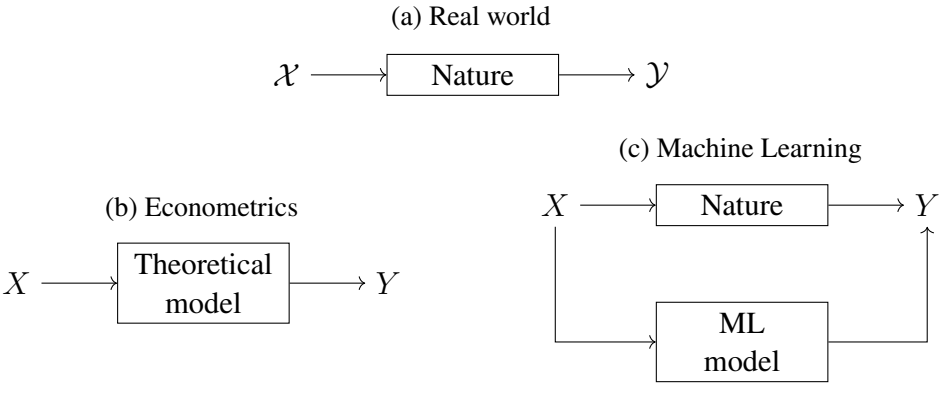
Summary

Introduction

The advances in statistical learning, data analysis and data science of the past decades have resulted in propagation of *Machine Learning* (ML) techniques to different applied fields, including social and human sciences. Nowadays, it is impossible to imagine a field of science that is not benefiting from the fruits of statistical learning. The works of De Palma et al. (2011) and Cascetta (2009) on transportation modelling, the publications of Molina and Garip (2019) dedicated to sociology problematic, the articles of Coussement, Benoit, and Poel (2010) concerning marketing decisions, actuary analysis studies (Denuit and Trufin (2019), Denuit and Hainaut (2019)) or even psychology with an example of Baayen et al. (2017) work reflect the literal omnipresence of the newly developed techniques.

However, there exist two completely distinct approaches to applying statistical learning, as described by Breiman and others (2001) and latter by Athey and Imbens (2019): the *Machine Learning* which focuses on the predictive qualities (figure 1c) and *Econometrics* which attempts to decipher the underlying properties of the data (figure 1b). In economics, where the research is focused on hidden patterns exploration, the scientific community prefers to implement the traditional econometrics techniques using the more advanced statistical models only in some special cases or as some assistance tools (Athey 2018). This discrepancy is explained by the fact that econometrics, contrary to traditional ML paradigm focusses on the accessibility of results. Consequently, many of the advanced ML techniques rarely appear in economics publications because of their believed lack of interpretability and excessive complexity in application. Nevertheless, some multidisciplinary scientists make attempts to breach this wall between *ML* and *Econometrics*: Varian (2014), Mullainathan and Spiess (2017) or, among the most recent, Athey and Imbens (2019). Their advances are mostly focused on resolving the general interdisciplinary tool-set integration questions, without considering the application specific details. Nevertheless, in the attempt to breach the interdisciplinary barrier the details reveal themselves to be of utmost importance in the solution of the problem.

Figure 1: The different paradigms



There have already been a multitude of studies comparing the performances of different econometric and ML models in various real world scenarios: the study of machine learning methods to model the

car ownership demand estimation of Paredes et al. (2017), for example; or the use of decision trees in microeconomics of Brathwaite, Vij, and Walker (2017). However, there's no known to us work incorporating at least all the baseline models, as it would require an unimaginable amount of efforts to accomplish. For instance, in the literature the performance of competing models are studied according to several absolutely alien criteria: in terms of the quality of data adjustments, in terms of predictive capacity, as well as in terms of the quality of the economic and behavioural indicators derived from estimates and, finally, according to their algorithmic efficiency and computational costs. None of the known to us articles manages to incorporate all these aspects into their benchmarks, limiting their studies only with several performance criteria.

These various aspects, greatly impact the performance of particular models or algorithms, although some of them are often ignored by the researchers. Not only there exists inconsistency in the targeted performance metrics in the contemporary models' comparisons, but there is also omnipresent problems of theoretical background choice, dataset selection or model's specifications. For example, speaking about the datasets used to support their findings, many researchers explore the impacts of different specifications on the same observed or simulated choice situation (Munizaga and Alvarez-Daziano (2005), Fiebig et al. (2010), McCausland and Marley (2013), Bouscasse, Joly, and Peyhardi (2019)) as it appears to be the most theoretically reliable procedure. However, there is still no established unified methodology documenting this field.

From this unambiguity in the scientific community the main problematic of this work arises. It is particularly important to establish a common framework for performance comparison of the discrete choice models be they from the econometrics or ML tool-set. However, this task cannot be accomplished outside a precise context, which will potentially impose some limitations over the models' structure, as well as influence the choice of performance metrics. In economics the discrete choice models are extensively used for consumer choice analysis (Anderson, De Palma, and Thisse 1992), willingness to pay derivation (Michaud, Llerena, and Joly 2012) and other preference studies. The field specific theories and traditional research objectives frame and define this study's scope.

From the economics perspective there exist three major points of interest to be taken into account. First, there is a strong interest in economics to explore the different behavioural set-ups, under different settings and assumptions. Secondly, given the different choice situations there is a potential need to test how the available mathematical models, potentially sensitive to the tested behavioural hypotheses or dependent on these hypotheses, perform in a given context. Last, but not least, a comprehensive implementation of a performance evaluation methodology, combining reproducibility and control of experimental conditions, should be introduced in the proposed framework.

Consumer choice

The economic decision theory derives mostly from the random utility theory (RUM) of McFadden (1974) and more recently of McFadden (2001), that were recently challenged by alternative visions such as random regret minimisation theory (RRM) of Chorus (2010), with a related relative advantage

maximisation theory (RAM) of Leong and Hensher (2015), or even quantum decision theory (QDT) of Yukalov and Sornette (2017), which offers a wide range of tools for modelling under uncertainty.

These different theories address various aspects of the decision making process, under different suppositions and incorporating different biases. For example, one of the basic assumptions of the traditional choice theory is the transitivity of choice, meaning there exists a strict hierarchy of individual preferences among alternatives. This assumption may be unsuitable for real world choice situation and lead to potential bias, which is addressed by quantum decision theory. QDT manages to bypass this shortcoming and incorporate non-transitivity of choices into the framework. There exist a multitude of other behavioural elements unexplained by the most traditional models that may be incorporated into the decision making framework, such as loss aversion for example, that could be addressed with random regret minimisation theory.

There is a particular interest in detecting the differences in the models' performances depending on the choice context and the assumed decision-making framework. It is important, because different consumer behaviour in the individual choice context result in different choice distributions, which may affect the models' performances. In economics RUM theory is nowadays one of the most used choice settings in the individual decision modelling. Nevertheless, there still exist some unexplored limitations, that such theoretical framework may impose over the estimation techniques, as well as to what potential biases a model's misspecification may lead.

Mathematical models

In general any classification technique may be used to model individual decisions, although nearly every model has some restrictions and limitations, which may largely affect its performances in a given context.

Usually the choice of model is rarely discussed in applied studies, as the researchers tend to use either the simplest model possible or attempt to implement one particular model of interest ignoring some times the other possible choices. For example, many traditional econometrics studies, given a multiple choice problem context, use a multinomial logistic regression (MNL) or even simplify the problem to a binary case, allowing to implement even more traditional models such as binary logit or binary probit models. However, there exists a multitude of particular cases in modelling individual choices, that require specific techniques to be implemented. A family of duration models may be used to model the individual decisions over time (Vitetta (2016)); network modelling that allows to incorporate spatial and social dependencies for the explored data (Brock and Durlauf (2003)); preference learning techniques aiming to explore the positioning of different alternatives by an individual (Tsoukiàs and Viappiani (2013), Pigozzi, Tsoukiàs, and Viappiani (2016)) and many other advanced techniques from *machine learning* field such as neural networks or support vector machines.

An incorrect choice of the modelling technique may have a strong impact on the derived target values leading to some erroneous conclusions in the end. For example, an incorrectly estimated willingness to pay for a particular product may lead to significant losses. When conducting an applied research study

one should always be conscious of the eventual biases introduced by the choice of the model and the eventual consequences of these choices. Some models are not suitable to be implemented on a particular set of data, while others are unable to provide necessary information about the relationships within a particular dataset or derive the particular target values of interest.

Taking into account the implications of RUM theory, there exists a particular interest to make the focus on the state of art econometric discrete choice models (Agresti (2013), Agresti (2007), Baltagi (2008), Train (2009), McFadden (2001), McFadden (1974)) as well as their counterparts used in ML (Hastie, Tibshirani, and Friedman (2009), Kotsiantis, Zaharakis, and Pintelas (2006)). A comparison of some simple models against more complex ones may reveal the trade-off between precise estimates and the resources invested.

Data

Different sources of data are available for a researcher, that could be divided into two groups (Japkowicz and Shah 2011): *field datasets*, which are gathered through an experiment or collected from the real world observations or real world uncontrolled experiment; and *synthetic datasets*, which are artificially generated by the researcher to suit his needs and respect some particular limitations. Although this variability of dataset choices not that evident in the context of applied studies, there is an ongoing debate concerning the eventual impacts of data choice on the models' performances and resulting metrics.

Given a task of performance evaluation and comparison for different algorithms or mathematical models there is always a difficult choice of the data type to be used in the study. Both of the mentioned above dataset types have their advantages and disadvantages and require a particular attention. However, having for objective the theory- and model-testing framework construction there is a strong interest to use the artificially generated data in order to have as much control as possible over the situation.

The framework and context

Given these three key elements we propose an integrated simulation and theory-testing framework which will encompass all the different aspects of the model comparison task. The steps to be integrated into such framework encompass many theoretical questions starting with the underlying theoretical assumptions and ending with the choice of correct performance metrics. Consequently, this work attempts to fill the gap between two statistical paradigms: *econometrics* and *machine learning*, taking into account the key elements, among which the different combinations of decision theory assumptions, dataset generation procedures, mathematical models and target performance measures. The problematic arises from the insufficient points of contact among researchers from different fields of applications, as well as insufficiently unified methodology to put into relations the different approaches. A work that uses unified knowledge from several disciplines might be highly beneficial for the scientific community as it will lie a foundation and provide support for future applied studies. Following the logic of Athey (2018) and Mullainathan and Spiess (2017) the project will attempt to merge the essentials of ML and econometrics paradigms, retaining their key concepts in the context of consumer choice problem.

We propose to use an applied paper in econometrics of choice modelling to facilitate understanding

of the field of application and tools. This means not that we will attempt to replicate the results, but rather to use the context provided in the work for demonstration of the proposed hypothesis-testing framework. We select the article of Michaud, Llerena, and Joly (2012) as our reference paper, because of the advantages to work directly with the authors of the paper. The work of Michaud, Llerena, and Joly (2012) is focused on investigation of consumers' willingness to pay (WTP) for environmental attributes of a non-food agricultural products, taking roses as example. Authors constructed an experimental framework to derive the premium the testing subjects were ready to pay for such environmental attributes as lower carbon imprint and ecological labelling, certifying the source of the environmentally friendly practices. That study explored individual preferences for roses with an eco-label and a carbon footprint using discrete choice modelling techniques and real economic incentives resulting in real purchases of roses. The gathered dataset was analysed with a mixed logit model demonstrating notorious premiums for both attributes. We will benefit of the obtained results to demonstrate all of the complexity of a proposed theory-testing framework, its functionality and perspectives.

The present report is divided into two main parts. The first section presents the chosen context for this work followed by short presentations of all the theoretical aspects which play their major roles in this study, tracing at the same time parallels with the context. The second part presents the results of all the results step-by-step, demonstrating the functionality of the designed framework. Each of the sections has an identical logical structure of presentation of the framework's components in the successive order: starting with the behavioural modelling and data related questions, directly followed by the models' presentation and the performance measures. The final section concludes.

Agresti, Alan. 2007. *An Introduction to Categorical Data Analysis, Second Edition*.

———. 2013. *Categorical Data Analysis, Third Edition*.

Allaire, JJ, and François Chollet. 2020. *Keras: R Interface to 'Keras'*. https://CRAN.R-project.org/package=keras.

Allaire, JJ, and Yuan Tang. 2020. *Tensorflow: R Interface to 'Tensorflow'*. https://CRAN.R-project.org/package=tensorflow.

Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2018. *Rmarkdown: Dynamic Documents for R*. https://CRAN.R-project.org/package=rmarkdown.

Anderson, Simon P, Andre De Palma, and Jacques-Francois Thisse. 1992. *Discrete Choice Theory of Product Differentiation*. MIT press.

Athey, Susan. 2018. "The Impact of Machine Learning on Economics." Book. In *The Economics of Artificial Intelligence: An Agenda*, by Ajay Agrawal, Joshua Gans, and Avi Goldfarb, 507–47. National Bureau of Economic Research; University of Chicago Press. https://doi.org/https://doi.org/10.7208/chicago/9780226613475.001.0001.

Athey, Susan, and Guido W. Imbens. 2019. "Machine Learning Methods That Economists Should Know About." *Annual Review of Economics* 11 (1): 685–725. https://doi.org/10.1146/annurev-economics-080217-053433.

Baayen, Harald, Shravan Vasishth, Reinhold Kliegl, and Douglas Bates. 2017. "The Cave of Shadows: Addressing the Human Factor with Generalized Additive Mixed Models." *Journal of Memory and Language* 94: 206–34. https://doi.org/https://doi.org/10.1016/j.jml.2016.11.006.

Baltagi, Badi. 2008. *Econometric Analysis of Panel Data*. John Wiley & Sons.

Bouscasse, Hélène, Iragaël Joly, and Jean Peyhardi. 2019. "A new family of qualitative choice models: An application of reference models to travel mode choice." *Transportation Research Part B: Methodological* 121 (C): 74–91. https://doi.org/10.1016/j.trb.2018.12.010.

Brathwaite, Timothy, Akshay Vij, and Joan L Walker. 2017. "Machine Learning Meets Microeconomics: The Case of Decision Trees and Discrete Choice." *arXiv Preprint arXiv:1711.04826*.

Breiman, Leo, and others. 2001. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." *Statistical Science* 16 (3). Institute of Mathematical Statistics: 199–231.

Brock, William, and Steven Durlauf. 2003. "Multinomial Choice with Social Interactions." NBER Technical Working Papers 0288. National Bureau of Economic Research, Inc. https://EconPapers.repec.org/RePEc:nbr:nberte:0288.

Cascetta, Ennio. 2009. *Transportation Systems Analysis: Models and Applications*. Vol. 29. Springer Science & Business Media.

Chorus, Caspar G. 2010. "A New Model of Random Regret Minimization." *European Journal of Transport and Infrastructure Research* 10 (2).

Coussement, Kristof, Dries F. Benoit, and Dirk Van den Poel. 2010. "Improved Marketing Decision Making in a Customer Churn Prediction Context Using Generalized Additive Models." *Expert Systems with Applications* 37 (3): 2132–43. https://doi.org/https://doi.org/10.1016/j.eswa.2009.07.029.

Croissant, Yves. 2020. *Mlogit: Multinomial Logit Models*. https://CRAN.R-project.org/package=mlogit.

Denuit, Michel, and Donatien Hainaut. 2019. *Effective Statistical Learning Methods for Actuaries Iii: Neural Networks and Extentions*. Springer.

Denuit, Michel, and Julien Trufin. 2019. *Effective Statistical Learning Methods for Actuaries I: GLMs and Extentions*. Springer.

De Palma, André, Robin Lindsey, Emile Quinet, and Roger Vickerman. 2011. *A Handbook of Transport Economics*. Edward Elgar Publishing.

Fiebig, Denzil, Michael Keane, Jordan Louviere, and Nada Wasi. 2010. "The Generalized Multinomial Logit Model: Accounting for Scale and Coefficient Heterogeneity." *Marketing Science* 29 (3): 393–421. https://EconPapers.repec.org/RePEc:inm:ormksc:v:29:y:2010:i:3:p:393-421.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

Heinzen, Ethan, Jason Sinnwell, Elizabeth Atkinson, Tina Gunderson, and Gregory Dougherty. 2020. *Arsenal: An Arsenal of 'R' Functions for Large-Scale Statistical Summaries*. https://CRAN.R-project.org/package=arsenal.

Hlavac, Marek. 2018. *Stargazer: Well-Formatted Regression and Summary Statistics Tables*. https://CRAN.R-project.org/package=stargazer.

Japkowicz, Nathalie, and Mohak Shah. 2011. *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press. https://doi.org/10.1017/CBO9780511921803.

Kotsiantis, Sotiris, I. Zaharakis, and P. Pintelas. 2006. "Machine Learning: A Review of Classification and Combining Techniques." *Artificial Intelligence Review* 26 (November): 159–90. https://doi.org/10.1007/s10462-007-9052-3.

Leong, Waiyan, and David A. Hensher. 2015. "Contrasts of Relative Advantage Maximisation with Random Utility Maximisation and Regret Minimisation." *Journal of Transport Economics and Policy (JTEP)* 49 (1): 167–86. https://www.ingentaconnect.com/content/lse/jtep/2015/00000049/00000001/art00010.

McCausland, William J., and A.A.J. Marley. 2013. "Prior Distributions for Random Choice Structures." *Journal of Mathematical Psychology* 57 (3): 78–93. https://doi.org/https://doi.org/10.1016/j.jmp.2013.05.001.

McFadden, Daniel. 1974. "The Measurement of Urban Travel Demand." *Journal of Public Economics* 3 (4): 303–28. https://doi.org/https://doi.org/10.1016/0047-2727(74)90003-6.

———. 2001. "Economic Choices." *The American Economic Review* 91 (3). American Economic Association: 351–78. http://www.jstor.org/stable/2677869.

Michaud, Celine, Daniel Llerena, and Iragael Joly. 2012. "Willingness to pay for environmental attributes of non-food agricultural products: a real choice experiment." *European Review of Agricultural Economics* 40 (2): 313–29. https://doi.org/10.1093/erae/jbs025.

Microsoft, and Steve Weston. 2020. *Foreach: Provides Foreach Looping Construct*. https://CRAN.R-project.org/package=foreach.

Molina, Mario, and Filiz Garip. 2019. "Machine Learning for Sociology." *Annual Review of Sociology* 45 (1): 27–45. https://doi.org/10.1146/annurev-soc-073117-041106.

Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31 (2): 87–106. https://doi.org/10.1257/jep.31.2.87.

Munizaga, Marcela A., and Ricardo Alvarez-Daziano. 2005. "Testing Mixed Logit and Probit Models by Simulation." *Transportation Research Record* 1921 (1): 53–62. https://doi.org/10.1177/0361198105192100107.

Paredes, Miguel, Erik Hemberg, Una-May O'Reilly, and Chris Zegras. 2017. "Machine Learning or Discrete Choice Models for Car Ownership Demand Estimation and Prediction?" In *2017 5th Ieee International Conference on Models and Technologies for Intelligent Transportation Systems (Mt-Its)*, 780–85. IEEE.

Pigozzi, Gabriella, Alexis Tsoukiàs, and Paolo Viappiani. 2016. "Preferences in Artificial Intelligence." *Annals of Mathematics and Artificial Intelligence* 77 (3-4). Springer Verlag: 361–401. https://doi.org/10.1007/s10472-015-9475-5.

R Core Team. 2018a. *Foreign: Read Data Stored by 'Minitab', 'S', 'Sas', 'Spss', 'Stata', 'Systat', 'Weka', 'dBase', ...* https://CRAN.R-project.org/package=foreign.

———. 2018b. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Talebijamalabad, Amirreza, Nikita Gusarov, and Iragael Joly. 2020. *Sdcm: Simulation of Discrete Choice Models*.

Train, Kenneth E. 2009. *Discrete Choice Methods with Simulation*. Cambridge university press.

Tsoukiàs, Alexis, and Paolo Viappiani. 2013. "Tutorial on preference handling." In *ACM Conference on Recommender System (RecSys)*, 497–98. Hong Kong, China. https://doi.org/10.1145/2507157.2508065.

Varian, Hal R. 2014. "Big Data: New Tricks for Econometrics." *Journal of Economic Perspectives* 28 (2): 3–28. https://doi.org/10.1257/jep.28.2.3.

Vitetta, Antonino. 2016. “A Quantum Utility Model for Route Choice in Transport Systems.” *Travel Behaviour and Society* 3. Elsevier: 29–37.

Wickham, Hadley. 2019. *Tidyverse: Easily Install and Load the 'Tidyverse'*. https://CRAN.R-project.org/package=tidyverse.

Xie, Yihui. 2020a. *Knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://CRAN.R-project.org/package=knitr.

———. 2020b. *Tinytex: Helper Functions to Install and Maintain Tex Live, and Compile Latex Documents*. https://CRAN.R-project.org/package=tinytex.

Yukalov, Vyacheslav I, and Didier Sornette. 2017. “Quantum Probabilities as Behavioral Probabilities.” *Entropy* 19 (3). Multidisciplinary Digital Publishing Institute: 112.