

Is there a future for traditional stochastic models (in business and industry) in the AI and ML era?

Fabrizio Ruggeri

Istituto di Matematica Applicata e Tecnologie Informatiche

Consiglio Nazionale delle Ricerche

Via Alfonso Corti 12, I-20133, Milano, Italy, European Union

fabrizio@mi.imati.cnr.it

www.mi.imati.cnr.it/fabrizio/

AN ASMBI COLLABORATIVE PAPERS

- The talk is based on a forthcoming collaborative paper by many researchers in the field who discuss, and illustrate through examples, when the "traditional" stochastic models should be preferred to AI and ML methods and vice versa
- Actually, sentences from the paper are presented and commented
- The paper showcases examples from business and industry, since it will be published in *Applied Stochastic Models in Business and Industry*
- So far the contributors are:

David Banks, Marcos Escobar, Nicholas Fisher/William Cleveland, Paolo Giudici, Roger Hoerl/Dennis Lin, Ron Kenett/Nalini Ravishanker/Marco Reis, Wai Keung Li/Philip Yu, Jean-Michel Poggi, Marco Reis, Gilbert Saporta, Piercesare Secchi, Rituparna Sen, Ansgar Steland, and Zhanpan Zhang

FEW QUOTES

- *We can stop looking for models. We can analyse the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.*
(Christopher Anderson, computer scientist)
- *All models are wrong, but some are useful.*
(George P. Box)
- *All models are wrong, and increasingly you can succeed without them.*
(Peter Norvig, computer scientist and former director of research at Google)
- *The goal of science is not to make predictions. It is also offering an image of reality, a conceptual framework for thinking about things. This ambition has made scientific thinking effective. If the goal of science were only predictions, Copernicus would not have discovered anything compared to Ptolemy: his astronomical predictions were no better than Ptolemy's. But Copernicus found a key to rethink everything and understand better.*
(Carlo Rovelli, physicist, text translated from Italian by Piercesare Secchi)

SOME QUESTIONS

- Is there a future for traditional stochastic models (in business and industry) in the AI and ML era?
- Would stochastic modelling still be useful with the impressive performance of ML methods?
- Is it still justified to push research and teaching in stochastic modelling?
- What can stochastic modelling contribute to ML and vice versa?
- Are there areas where stochastic modelling performs better and others where ML does?
- How does AI affect (and is affected by) stochastic modelling and ML?
- What are "traditional" and "new" stochastic models? Regression vs. deep neural network?

SOME PRELIMINARY COMMENTS

- The advances in analytics under headings of AI, ML or deep learning have changed the approach to data analysis. Part of this change followed the big data, sensor technology and computational capabilities advances.
- A proper methodology for assessing the right balance between purely data driven empirical methods, physics-based models and probabilistic stochastic models is requiring further development. This is necessary for the effective and robust achievement of data driven knowledge.
- This requires substantial revisions in educational curricula where statistics and data science methods need to have stronger ties with physics and engineering, and rely on software such as R, Python, JMP.
- Papacharalampous et al (2019) compare forecast performances of the two approaches and found that ML methods do not differ dramatically from the stochastic, while none of the methods under comparison dominates the other. From a scientific point of view both tools would be valuable to an investigator who would like to seek a better understanding of his/her problem.

SOME PRELIMINARY COMMENTS

- In contrast to deterministic modelling, which many ML methods employ, a key advantage of statistical modelling is its ability to quantify uncertainty. This quantification is crucial for enhancing model reliability and supporting decision-making in real-world applications.
- In addition, statistical models generally offer better explainability compared to black-box ML methods. A thorough understanding of model behaviour enhances transparency and builds trust in high-risk domains such as healthcare and finance.
- Often, sample size is another important factor in determining the selection of appropriate modelling approaches. It is well known that ML, particularly deep learning methods, require a large amount of data to achieve satisfactory performance.
- When physical simulation is computationally slow and/or experimental data collection is costly, practitioners will need to explore alternative approaches. For example, Bayesian networks can be an effective method for addressing inverse design problems and time-dependent dynamic issues.

SOME PRELIMINARY COMMENTS

- If the question is predicting the occurrence of a future event, e.g., future conditional mean of a stochastic process, and there is a large dataset with many variables, then ML methods may be a first choice.
- However, in many investigations, predictions may not be the only objective and one would like to have more understanding of the underlying data generating process.
- It is true that all models may be only approximations at best but a good one should be able to give us insight into why the data are behaving in a certain way.
- In such cases, classical stochastic modelling may provide more insights into the data generating mechanism, especially if the number of observations is larger than the number of variables.
- The principles of randomization, replication and blocking need to be taken into account for any designed experiment. For sample surveys, it is extremely important to ensure that common sources of bias like selection, response/non-response etc. do not creep in. Given the particular scenario, stochastic modelling provides a guiding principle on data collection for the most efficient utilisation of resources geared towards answering the business and industry question at hand.

A WORLD IN EVOLUTION

- A simple example motivated by finance, but easily extrapolated to other areas, is the evolution of stock prices in the last 60 years.
- The price of a share is not only dictated by rational economic forces, but it is also a by-product of human individual and collective behaviour, desires, passion, wishes, and mistakes; all this summarised on a single number at any given time.
- The modelling of stocks had to adapt/evolve from a quite robust and simple Gaussian model in the 60's to, e.g., the inclusion of jumps in the 70's (Levy models), random local volatility as detected in the 80's (i.e., GARCH, CEV* models), stochastic volatility in the 90's motivated by the need to trade volatilities and the emergence of the volatility index (VIX), and more recently the presence of stochastic volatility of volatility and stochastic skewness on prices as captured by the new indexes VVIX and SKEW respectively. The history for multi-asset modelling is even richer.
- The increase in complexity will only continue, requiring more advanced models (based on ML and AI?) to properly explain the dynamics of single and multiple stocks.
- Would this mean the end of stochastic modelling?

*Constant elasticity of variance

STOCHASTIC MODELLING: EXAMPLES

- Many consider logistic regression a stochastic model since it produces individual or collective probabilities of default, but this is an abuse of the term, as it is very difficult to define a generative model of default, and it is actually an empirical approach.
- The method is the industry standard for credit scoring and fulfills regulatory requirements (at least in EU), thanks to the simplicity of its formulation and interpretation (e.g. on macroeconomic variables like GDP and interest rates affecting the corporate probability of default).
- Numerous attempts to use more complex models have not led to significant progress as quoted in, e.g., Bazzana et al (2024): *This paper uses a large sample of small Italian companies to compare the performance of various ML classifiers and a more traditional logistic regression approach. In particular, we perform feature selection, use the algorithms for default prediction, evaluate their accuracy, and find a more suitable threshold as a function of sensitivity and specificity. Our outcomes suggest that ML is slightly better than logistic regression. However, the relatively small performance gain is insufficient to conclude that classical statistical classifiers should be abandoned, as they are characterized by more straightforward interpretation and implementation.*

STOCHASTIC MODELLING: EXAMPLES

- Stochastic modelling is valuable when considering massive amount of data on individual electricity consumption provided by new metering technologies and smart grids, for load profiling and modelling at different scales of the electricity network.
- A methodology based on a mixture of high-dimensional regression models is used to perform clustering of individual customers. It accounts for forecasting, combined with the partitioning of the electrical signal into successive curves to consider it as functional data.
- The method extracts nice features from individual consumption data with little information (2 days) and no other prior, focusing on the discovery of clusters corresponding to different regression models, which could then be used directly for profiling, but can also be useful for forecasting purposes,
- The statistical approach allows a deeper analysis of the use of the internal objects of the method from a practical perspective, focusing not only on the results of the method but also on the by-products of the method, providing visualisation tools to understand the estimates and facilitate interpretation.

STOCHASTIC MODELLING: EXAMPLES

- In the context of business, there is great potential for Statistics in computational advertising.
- One way that stochastic processes arise is in contract fulfilment for showing online advertisements.
- When a user visits a website (e.g., cnn.com), it triggers a complex sequence of activity lasting less than ten milliseconds.
- There is a virtual auction among demand-side platforms who must decide
 1. whether to bid on showing an ad to the user
 2. which ad to show
 3. how much to bid
- Typically, the demand-side platform knows quite a lot about the user: gender, approximate age, approximate income, marital status, location, and previous purchase history

STOCHASTIC MODELLING: EXAMPLES

- The whole topic of survival analysis has been developed to handle censored and truncated observations efficiently. Without stochastic modelling such topics will remain completely out of reach.
- There are works about the use of deep learning (Wiegrebe et al, 2024) especially exploiting unstructured or high-dimensional data such as images, text or omics data, but not much exists about censored and truncated observations as confirmed by the regularly updated website <https://survival-org.github.io/DL4Survival/>
- The corporate probability of default is an important quantity for regulatory purposes as it measures how much risk a bank is taking overall. A ML algorithm will be useful in identifying which customers are more liable to default. But when the interest is in the overall probability of default, which is a population parameter, concepts of population and stochastic modelling naturally need to be taken into account.

STOCHASTIC MODELLING: EXAMPLES

- Even for point predictions of future time series the knowledge of the underlying probability model can provide many insights about interpretation and quality of the prediction.
- Wong and Li (2000) considered a three-component mixture autoregressive model for the first difference y_t of the IBM stock price data from 1961 to 1962 with 369 observations and easily obtained one-step ahead predictive distributions.
- It was observed that the predictive distributions for different time points would exhibit distinct bimodality when the volatility of y_t was high. In contrast unimodality was observed when the market was less volatile.
- In other words, the market would have higher chances of a sharp increase or decrease when volatility was high.
- In such a case a point forecast of the future time series would not be informative but knowledge of bimodality about the predictive distribution and the fitted mixture model would be useful to the investigator and risk manager.

MACHINE LEARNING

- The main advantage of ML methods is in prediction tasks and this is most common in regression, clustering, classification and time series or spatial forecasting settings.
- ML models are boosting AI applications in many domains, such as finance, health-care, and automotive.
- This is mainly due to their advantage, in terms of predictive accuracy, with respect to “classic” statistical learning models.
- However, although complex ML models may reach high predictive performance, their predictions are not explainable, and have an intrinsic black-box nature.
- Furthermore, these models may not be robust, and may use data that are not representative, thereby generating biases and discriminations
- The success of advanced ML methods in areas such as image and face recognition is spectacular, although there are a few shortcomings in terms of robustness, where the modification of a single pixel can dramatically change the prediction, but the intelligibility of the models was hardly questioned until recently.

MACHINE LEARNING

- There are attempts to make the algorithms explainable: this is the XAI, with methods that seek to open the black box, with new measures of variable importance or the use of surrogate models to explain a decision using trees or local linear models.
- Some researchers are even calling for black boxes to be abandoned in favour of simple models.
- It should be noted that works on feature importance in ML, with Shapley measures for example, have renewed the classic problems which already showed that even simple models are not so easy to interpret.
- As well as being transparent, algorithms need to be fair and non-discriminatory when applied to human groups.
- Algorithmic fairness is a major area of development in computer science, but one in which statisticians are not yet very involved.
- Transparency or explicability are not enough: if a model is not causal, which is in line with a definition of stochastic modelling, and is based solely on correlations, it can lead to erroneous conclusions.

MACHINE LEARNING: EXAMPLES

- We shall someday have all vehicles on the road being networked and autonomous, and then there will be an ocean of travel data.
- We will not need differential equation models for traffic flow since we shall have the empirical process in great detail and the need for process modelling will be much less, although the stochastic process itself will remain important.
- Considering data on flows of networked vehicles, they are actually the superposition of observations from multiple and perhaps simpler data generation mechanisms, like daily commuters, long-haul trucking, school buses, holiday and weekend travel.
- An analyst might use simple stochastic models to describe each component, and then attempt to decompose the complex empirical process into its constituent parts.
- We have a superposition of results from many distinct data generation mechanisms, some of which are well understood, some which are partially understood, and some of which may represent new ones.

MACHINE LEARNING: EXAMPLES

- The inverse design problem, aimed to identify input values producing desired outputs, has long been a challenge in natural science and engineering areas, due to the many-to-one relationships between inputs and outputs that are commonly embedded in the data.
- Recently, several deep learning-based methods have been developed which tackle the problem in a more direct way: one of them is conditional invertible neural network (cINN), which benefits from both mathematical and statistical principles.
- First, cINN is a generative probabilistic model which stochastically generates posterior samples of inputs X given specified outputs Y , i.e. $P(X|Y = y)$.
- Second, the network architecture of cINN is carefully designed to enable the computation of a Jacobian determinant.
- Furthermore, it is critical to effectively select a subset of posterior samples of inputs for the follow-up validation study, which may incorporate user constraints and potentially involve an optimisation process.

MACHINE LEARNING: EXAMPLES

- Sequential aggregation of individual predictions aims to predict y_t , given past values up to t : y_1, \dots, y_{t-1} , using K experts whose only known information consists of their immediate predictions of y_t and the history of their predictions.
- The idea is to optimally combine the predictions by adjusting the weights (assigned to each expert) at each step based on instantaneous losses (e.g. quadratic), with no stochastic modelling.
- Audebert et al (2016) considered daily average concentration of PM10 in a location in Normandy, as well as the corresponding forecasts given by 10 different statistical models
- The sequential prediction strategy significantly improves the performance of the best expert, both in terms of errors and alerts.

STOCHASTIC MODELS AND MACHINE LEARNING

- Fernandez-Delgado et al (2014) compared 179 classifiers over 121 (non-large-scale) data sets from the UCI ML classification database, and found that parallel random forest (mostly due to a statistician, Breiman), performs best, followed by support vector machines. However, when no training data is available at the design phase or when relatively small amounts of data need to be analysed, stochastic modelling still shines.
- There is also an increasing number of applications, such as embedded systems (HW+SW) and implanted medical devices, which can only access very limited computational power. To provide them with ML capabilities, one may employ randomised neural networks, which can be trained with extremely low computational costs, combined with computationally efficient statistical methods to assess uncertainty, such as fixed-length confidence intervals to determine required sample sizes.
- Emulators are (usually) Gaussian process approximations to the agent-based model and offer fast and practical solutions that are sufficiently faithful to it to provide useful guidance. Remarkably, emulators can give posterior distributions over the discrepancy function, which indicates for which regions of the input space the emulator does a poor job of matching the agent-based model (computer simulations used to study the interactions between people, things, places, and time).

STOCHASTIC MODELS AND MACHINE LEARNING

- Although various state-of-the-art AI/ML approaches do not rely on explicit stochastic modelling, this is not the rule.
- Indeed, many ML problems can greatly benefit from stochastic modelling expertise and require results from Probability and Statistics for their improvement.
- For example, in industrial quality control, when setting up inspection processes, manually selecting key characteristics requires substantial resources and efforts.
- An auto-encoding neural network was used to learn such features from training data, with correlated features identified and evaluated by means of a risk analysis.
- The overall analysis includes pre-processing, design of net parameters, specification of the dependence measure, and combines ML as well as human expertise from Statistics and Engineering.

STOCHASTIC MODELS AND MACHINE LEARNING

- Active learning is a highly relevant learning problem in industry, dealing with the automatised sequential exploration of a sample space to identify the safe region where a system can optimally operate.
- It aims at reducing the number of labelled examples needed to achieve a certain accuracy by selecting the most informative safe examples from a large unlabelled dataset and determining their labels from a costly additional experiment.
- When it comes to dynamic systems, for example when a robot explores an environment, exploration takes place along trajectories instead of single points, and then the whole trajectory needs to be safe.
- Gaussian processes are an attractive approach to model the sequential data collection mechanism as well as the prediction uncertainty at a new point. For this purpose the Borel-TIS inequality is used instead of Monte Carlo simulations.
- It turns out that the combination of stochastic modelling and results from statistical theory and probability theory allows to produce a state-of-the-art ML method.

ARTIFICIAL INTELLIGENCE

- AI is susceptible to how a question is posed
- AI is limited in its ability to separate the good from the bad, in terms of the resources it uses to carry out its 'reasoning'. These resources may include resources generated by AI and derived from unsound resources or algorithms.
- The recent European regulation on AI (*AI Act*) aims to regulate the use of AI with a set of requirements of trustworthiness for AI applications, to be embedded in a risk management model. The requirements established for high-risk applications in the AI Act can be classified in four main variables to measure: Security, Accuracy, Fairness and Explainability. All of them need a set of consistent metrics that can establish not only if, but also how much, the requirements are satisfied over time.
- Regarding the role stochastic modelling can play in the AI era, one can distinguish between specialised ML approaches competing with stochastic methods, e.g.,
 - prediction using deep neural networks vs. (non)parametric regression
 - AI systems which automatically plan and model a statistical problem and then analyse real data vs. traditional computer-aided modelling and analysis done by a trained human analyst

ARTIFICIAL INTELLIGENCE

- Any answer to the initial question needs to predict to some extent the future development of AI, and thus may fail, but one can make an educated guess.
- The rapid progress of LLNs (Liquid Neural Networks) and their fine-tuning to specific domains and tasks allows to set up interacting AI agents to generate, check and curate output.
- It seems that software development is the field where this approach is most advanced and already provides convincing results.
- Here one agent outputs source code based on a user prompt, and a further specialised agent interacting with the user performs code checking and outputs directives for source code revisions, in order to eliminate errors and improve the program.
- It is clear that the concept of several AI agents, which interact among each other and with the user to solve certain problems, will become widespread and has a substantial potential for better and less error-prone AI.

ARTIFICIAL INTELLIGENCE

- LLNs are trained from massive internet data which contains huge amounts of source code in many programming languages and this is certainly the reason for their capabilities in generating computer programs.
- Probability, Theoretical Statistics and large parts of Applied Statistics and stochastic modelling are exact sciences and follow relatively simple grammars or, at least, they can be put into a formal language following a simple grammar.
- Therefore, one can expect that future AI systems have capabilities in these areas which are comparable to their skills in software development, thus going beyond the already remarkable functionality of Open AI's Code Interpreter.
- It is likely that future AI systems substantially simplify the development and application of stochastic models and statistical analyses and provide access to a large amount of knowledge.
- This might have devastating effects on the job market for graduates, since then only a few experts are needed to supervise the AI.

ARTIFICIAL INTELLIGENCE

- However, that development will be probably relatively slow.
 - Compared to software, there are less data for training of a Statistics agent.
 - Whereas software is written in a formal language, this does not strictly apply to scientific papers and books which form the training corpus.
 - Although not completely transparent, the state-of-the-art AI systems use lots of manual input from human experts.
- When it comes to Statistics, this needs well trained experts which are not available on a large scale, and it is questionable whether providers will invest here.
- Thus, as long as human input is needed, the capabilities in such fields will substantially lag behind areas such as programming.
- AI systems capable of producing valid stochastic models, methods for their analysis and derivations of their properties might result as a side product of efforts to create AI systems which can solve scientific research questions addressing hot topics such as finding cancer treatments, understanding the human brain or finding the grand unified theory of physics.

ARTIFICIAL INTELLIGENCE

- Nevertheless, it is an open question whether future AI systems will be smart enough to produce valid novel scientific results going beyond correct and nice sounding science prose, and whether they will be superior to humans in identifying and resolving challenging problems arising in modelling and analysing real-world data.
- Examples for the latter are causality (versus association), confounders and colliders, bias and multiple testing.
- For example, in order to identify causality, controlled experiments and randomization are the methods of choice, and available data and those obtained from observational studies often suffer from confounders (variables affecting response and regressor) and colliders (variables affected by response and regressor), which lead to distorted estimates of causal effects.
- Generally, blind analysis of available data collections may lead to severe biases which quickly result in discrimination and harm for people.

MY OWN INTERESTS

- Bayesian networks to split complex problems in many simpler ones, exploiting information and keeping interpretability
- Adversarial classification as an example of adversarial ML based on a proper Bayesian statistical approach
- AI-driven expert elicitation for prior choice in high dimensional problems

CONCLUSIONS

- It is worth noting that both the statistical modelling and ML communities are rapidly evolving. Therefore, a good practice for problem solving is to leverage multiple approaches and assess their performance and usability. Integrating insights from multiple approaches addresses the problem from different perspectives, thereby guiding users toward a clear path for continuous improvement.
- There is no any antagonism between statistical modelling and ML, which in some respects is its 21st century version, just as the emergence of Computer Science led to the development of Multivariate Statistics in the second half of the 20th century. ML has enriched the statistician's toolbox and, above all, has provided him/her with the fundamental concept of generalisation and the need to go beyond simply fitting a model on the basis of training data alone.
- For their part, statisticians can provide ML practitioners with their knowledge of biases, their sense of data (missing, aberrant, etc.) and their culture to avoid reinventing techniques such as categorical data encoding or principal component analysis! But statisticians must not be afraid to enter this new field, otherwise they will be marginalised.

CONCLUSIONS

- Statistical modelling remains a powerful tool for problem formulation and solution development, and ML can yield more insightful outcomes when it is integrated into the problem-solving process.
- Statistics contributed and contributes a lot to the development of ML and AI, since recognising the importance to formulate the problem of learning from data as an inference on the underlying data generating process (e.g. random sample or time series) is fundamental for a deeper understanding of ML methods.
- Statistics provides a more advanced and complete general framework for data analysis and is highly relevant for any AI development since AI applications often neglect issues such as evaluation of uncertainty, interpretability, model stability and reproducibility, or issues such as confounding, especially in areas where ML is regarded as state-of-the-art.
- This certainly applies to problems dealing with large-scale data such as image classification, which require enormous computational resources and are almost exclusively researched by computer scientists. However, for other task where small to moderately large samples are sufficient, the situation can be different.