9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

On Learning Agent-Based Models from Data

Keywords: agent-based models, machine learning, opinion dynamics, economics

Motivation and Setup

Agent-Based Models (ABMs) are used in several fields to study the evolution of complex systems from micro-level assumptions. These models encode the causal mechanism that drives the dynamical process, and have the advantage of being easy to interpret. However, they do not exploit the availability of data so their predictive power is limited. In addition, parameter calibration and model selection are manual and difficult tasks. Finally, ABMs typically can not estimate agent-specific (or "micro") state variables.

Here, we showcase a protocol to learn the latent micro-level variables of an ABM from data. The first step of our protocol is to cast the ABM into a probabilistic generative model, characterized by a computationally-tractable likelihood. This transformation follows two general design principles: balance of stochasticity and data availability, and replacement of unobservable discrete choices with differentiable approximations. Then, our protocol proceeds by maximizing the likelihood of the latent variables via a gradient-based expectation maximization algorithm.

We demonstrate our protocol by applying it in two different contexts: (*i*) an opinion dynamics ABM which models the evolution of the opinion of users interacting on social media, (*ii*) an ABM of the housing market, in which agents with different incomes bid higher prices to live in high-income neighborhoods. We show that the resulting models enable accurate estimation of the latent variables while preserving the general behavior of the ABM. In addition, our estimates improve the out-of-sample forecasting power of the model, compared to traditional methods.

Agent-Based Models

OD: Opinion Dynamics. We use the opinion dynamics model devised by Jager and Amblard [1]. It extends the classical bounded confidence model by distinguishing between positive and negative interactions, with opposite effects on the opinions of the involved agents. In reality, neither the opinion of a single agent nor the *sign* of the interactions are easily observable. The only observables available are unsigned interactions and some kind of action (e.g., posting on a subreddit or using a specific hashtag) which represent a noisy proxy for latent opinions.

HM: Housing Market. We use the housing market ABM presented by Pangallo et al. [2]. The ABM describes the housing market of a city composed of L locations or neighborhoods, each with a number of indistinguishable homes, inhabited by agents. Each agent belongs to an income class k, out of K income classes, each characterized by an income Y_k. At each time step, individual agents choose a neighborhood to purchase a home if they act as buyers, or put their home on sale if they act as sellers. The *attractiveness* of each neighborhood regulates how likely an agent is to bid for that location and is one of the fundamental building blocks. The model assumes that the higher the income of residents, the more attractive a neighborhood is. Matching between individual buyers searching in a neighborhood and sellers in the same neighborhood is modeled as a *continuous double auction*. The social composition of the city evolves as a byproduct of transactions, as high-income buyers may replace low-income sellers and lead to the gentrification of some neighborhoods. Only the number of transactions in the neighborhoods and the average price are observable, while the housing demand and the composition of the neighborhoods are latent.

Inference

We propose an inference mechanism for fitting the traces of a these ABMs, by maximizing the likelihood of their generative translation. The inference algorithm uses online expectation maximization to learn the latent parameters of the model. Given a set of observables, our algorithm can estimate the micro-level latent trajectories of the variables of interest: individual opinions and the sign of interactions for the OD model, and demand and composition of the neighborhood in the HM one. This algorithm maximizes the likelihood of the observed data at each time step, and repeats the process for a number of epochs to allow the bac-propagation of the effect of later observations to the initial conditions.

Results

The HM model allows us to show that our likelihood-based paradigm can be applied to very complex ABMs (see Figure 1). On this model, we perform a range of synthetic experiments: we generate data traces, and we use our inference algorithm applied to the observable part of each trace to estimate the unobservable state. First, we find that our model is able to recover the unobservable state of agents with sufficient accuracy (Figure 2a). Second, we find that such estimates lead to predictions of future traces that are more accurate than traditional methods of forecasting with ABMs (Figure 2b), such as the method of moments applied to observable time series, and a proportional baseline based on an approximation of the system.

The OD model, being more simple, allows us to test additional claims. Here, we consider different scenarios, synthetically generated by different macro parameters (Figure 3). Given its observables—without knowledge of the opinions of each agent—our model can identify the most likely set of macro parameters used to generate a data trace, thus allowing for testing of hypotheses. Furthermore, we apply our OD model to real-world data from user interactions on Reddit. We are able to recover the latent opinion of users and subreddits, and their trajectory in time. The distance in latent opinion space between a user and a subreddit is in fact predictive of the number of upvotes that a user receives on the given subreddit (Figure 4): as expected, the higher the distance, the lower the score.

Discussion

We have proposed to reformulate ABMs into a probabilistic generative guise to enable automatic inference of latent variables and parameters of the model. This type of model retains the benefits of agent-based ones (i.e., causal interpretation) while introducing the ability to perform model selection and hypothesis testing on real data. Our protocol can be seen as an alternative to black-box data assimilation methods (see Figure 1 for the HM example). Moreover, it forces the modeler to lay bare the assumptions of the model, to think about the inferential process, and to spot potential identification problems. This feature allows fine-grained analysis of real individuals with the same techniques used to describe ABMs. In other words, we are able to empirically quantify and verify the assumptions of ABMs at an individual level.

References

[1] W. Jager and F. Amblard. Uniformity, bipolarization and pluriformity captured as generic stylized behavior with an agent-based simulation model of attitude change. *Computational & Mathematical Organization Theory*, 10:295–303, 2005. (Cited on 1)

[2] M. Pangallo, J.-P. Nadal, and A. Vignes. Residential income segregation: A behavioral model of the housing market. *Journal of Economic Behavior & Organization*, 159:15–35, 2019. (Cited on 1)

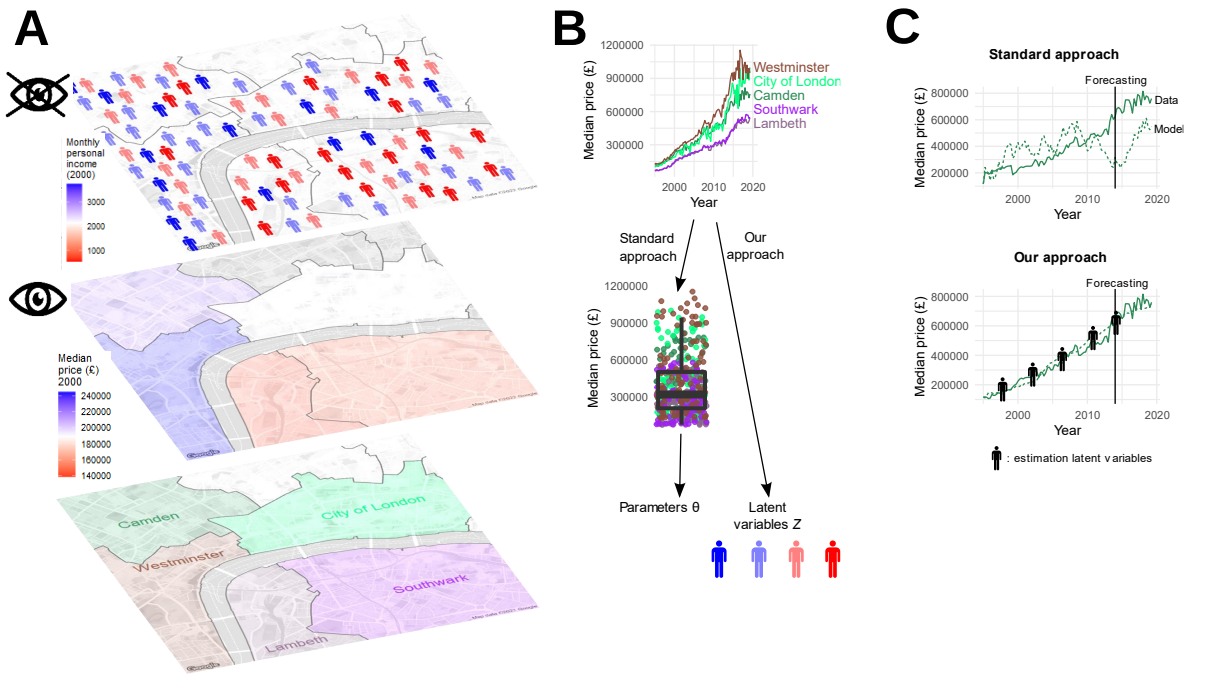
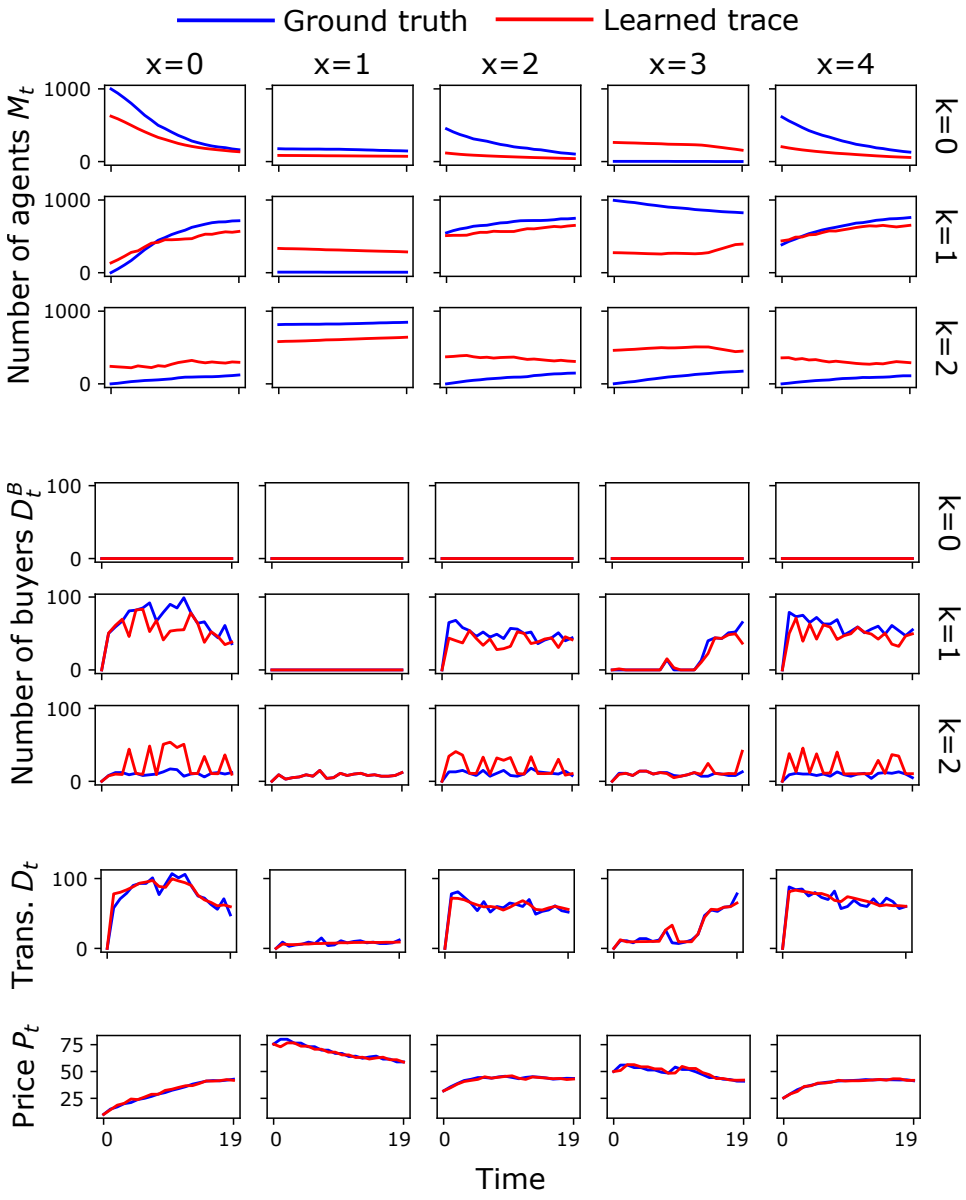
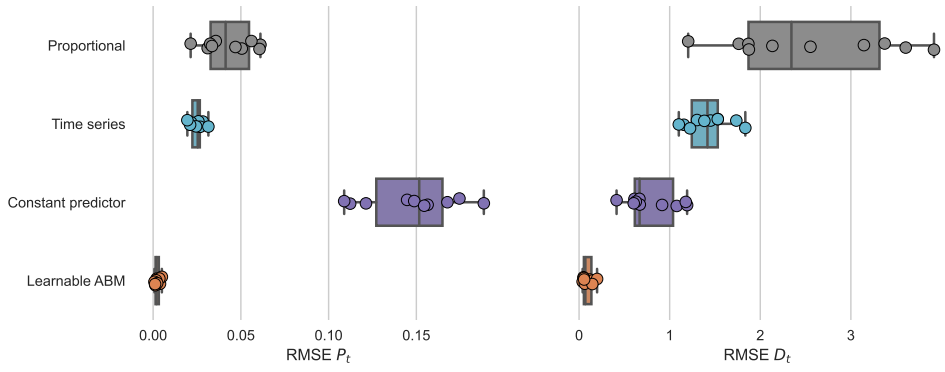


Figure 1: Our approach compared to a standard approach towards calibrating an ABM of the housing market. (A) Focusing on the boroughs in the center of a city (bottom), we consider the yearly average of transaction prices (middle) as observed variable, and the distribution of agent incomes (top) as latent variable. (B) For each borough, we observe a time series of transaction prices. In standard calibration, modelers typically calibrate some parameters Θ by computing the moments of prices across boroughs and years, minimizing the distance with the same moments in model-generated time series. In our approach, instead, we calibrate the evolution of latent variables Z by leveraging all information contained in time series, rather than reducing this information to specific summary statistics. (C) In the model, prices depend on agent incomes. Thus, since in the standard approach agent incomes are not calibrated, model-generated time series are bound to diverge, even if prices are initialized as in the data. With our approach, as we repeatedly estimate incomes, we can make model-generated time series track empirical ones.



(a)



(b)

Figure 2: Figure (a) shows estimates for model variables compared to the traces generated with the original housing market ABM, in a single experiment, chosen as the median experiment in terms of estimation quality. In each plot, the X axis represents the model time steps. Figure (b) shows forecasting error for our method compared to alternative benchmarks, as the RMSE of the P_t and D_t time series. Whiskers extend from the minimum to the maximum value, while boxes show the two central quartiles.)

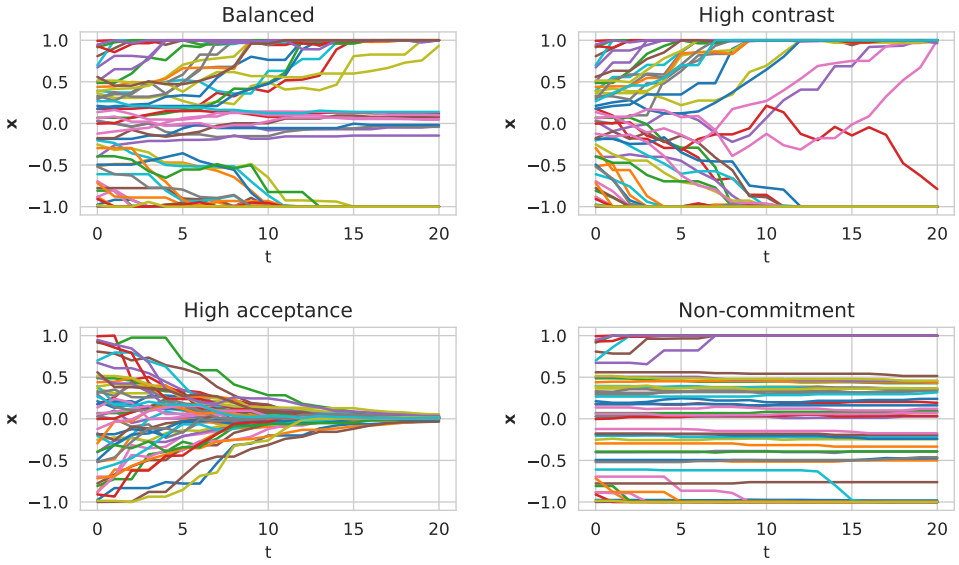


Figure 3: Examples of synthetic data traces generated in different scenarios, as opinion trajectories along time.

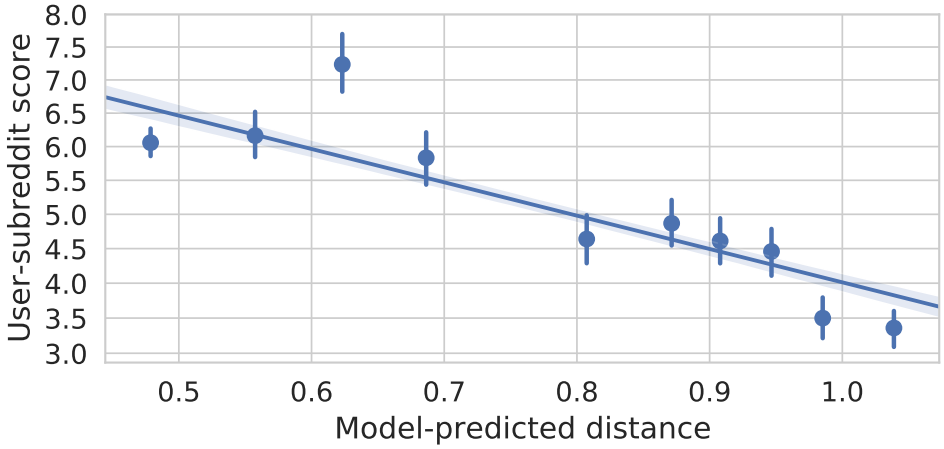


Figure 4: Univariate regression between user-subreddit average upvotes and user-subreddit distance in the opinion space as inferred by our model. Correlation coefficient is negative (-0.127) and highly significant ($p < 10^{-6}$).