Estimating diffusion and adoption parameters in networks

New estimation approaches for the latent-diffusion-observed-adoption model

Stephan, L.S.

Publication date
2021

Chapter 1

Introduction

1.1 Introduction

In this thesis I propose three distinct estimation methods suitable for a specific yet widely applicable model of social network interaction: the "latent-diffusion-observed-adoption model". This model features a hidden network diffusion process that exposes the individual to a choice once she is newly reached by it. The network on which the signal propagates, the seeding of the diffusion process and affirmative choices constitute the observed data.

The fact that the network interaction is hidden entails a curse of dimensionality: as the number of agents and time periods grow large, there is a multitude of scenarios by which the signal could have been spread that are all in accordance with the observed data. With the number of these diffusion scenarios growing exponentially in the number of agents involved, estimation is highly challenging. In each of the following three chapters, I propose a different estimation method to tackle this challenge. Chapters two and three both use the Maximum Likelihood Estimation (MLE) method while chapter four identifies the parameters by moment-based estimation. In chapter two a reduction in dimensionality is achieved by restricting the modelled time horizon (this will be referred to as the "Two-period Estimator"). In chapter three I propose a trimmed MLE that neglects some and focuses on the remaining diffusion scenarios (the "Trimming Estimator"). Chapter four identifies and uses easily calculable individual specific moments in a GMM-style estimation neglecting moment conditions that are computationally prohibitively hard to evaluate (the "Moment-based Estimators").

These methods are thus viable alternatives to the simulation-based estimation methods previously employed to this model.

1.2 Literature

Networks have been studied extensively in a broad range of academic disciplines, including sociology, epidemiology, biology, mathematics and computer science. They have increasingly also attracted the attention of economic researchers in recent years. There are several reasons for this. First, technological progress and the enhanced networking opportunities it entails has completely reshaped the landscape for economic activities. It has provided both consumers and producers with a totally new scope of action and interaction outside of traditional, established marketplaces. Economists still have to - empirically and theoretically - gain a better understanding what this implies for economic theory and policy. Second, while the liberal fundamentals of economic science have long remained unquestioned, recent years have seen more willingness among economists to challenge some of the key underlying assumptions of their science. There is thus more willingness to incorporate social networks as a form of non-contractual, privileged relationship or a form of social or human capital into economic models. Third, the aforementioned technological progress has also entailed a tremendous increase in data on social networks available to researchers, thus opening the door for interesting empirical projects. Fourth, the world has recently faced challenges that tragically highlight the impact of social networks on not just the economy.

The importance economists attribute to the study of social networks is most strikingly demonstrated by the fact that Econometrica featured one or several articles on social networks in practically all of the journal's issues in the past five volumes. Network models have come to use in practically all fields of applied economic research from development, transport, regional and environmental economics to finance and business studies.

Network models entail a threefold challenge for econometricians. First, the breakdown of the assumption of independence across individual observations limit the applicability of standard procedures for estimation or hypothesis testing. Second, network models are highly nonlinear in nature such that closed-form solutions for the parameters of interest are oftentimes not available. Third, estimation generally requires large amounts of tedious computations.

The economic literature on social networks can broadly be classified into network formation and network interaction models. Network formation models can further be classified into random graph models and micro-founded models. The former set up a stochastic model that produce network graphs that closely mimic the main characteristics of real-world social networks (such as degree heterogeneity, small average path length, homophily and high levels of clustering). The latter start with utility-

maximizing rational agents that choose their linking behaviour optimally. For an overview of these models see de Paula (2020) and Jackson (2004). Network interaction models on the other hand model phenomena that take place mediated by social networks. Most parsimoniously, Spatial Auto-Regression (SAR) models acknowledge the importance of network interaction but abstain from further modelling how and why this interaction takes place. Instead, they assume that the impact of neighbouring entities on each other is linear in means such that the neighbours' outcomes, observed or unobserved characteristics can be included as a regressand, determining each individual's outcome. This leads to a simultaneous-equation regression model that is usually estimated by Instrumental Variable (IV) techniques. For a survey of SAR models, see Anselin (2001), Anselin and Bera (1998), for an application to the recent challenges, see Krisztin, Piribauer, and Woegerer (2020). Parametric models have also been proposed to model Network interaction. Among these, diffusion models play a prominent role. These models have emerged in computer science and biology, but gain popularity in economic modelling: see Singh, Singh, Kumar, Shakya, and Biswas (2019), Cowan (2005). So far, most work on network diffusion has focused on purely random models and there is very little work on microfounded diffusion models. For an overview of network diffusion models, see Jackson and Yariv (2005) and Valente (2005). The "latent-diffusion-observed-adoption" model falls in this category.

Many authors have investigated random diffusion on networks under the assumption that the diffusion process is observed. Oftentimes, the focus has been on analysing aggregate characteristics such as convergence to a potential steady state or the impact of network topology: see Valente (1996) as an example. One prominently applied model, the "SIR" (susceptible-infected-recovered) model, is closely related to the model that I employ in this thesis, the main differences being the unobservability of the diffusion process (that is, the spreading of the signal remains hidden), the existence of a second stochastic and partly observed process (the adoption, observed if the choice to adopt is affirmative) and the absence of recovery (that is, infected agents stay infected forever) in the model to be used by me. The observability of diffusion is a reasonable assumption in some socio-economic contexts, in particular when collecting the diffusion data can be done relatively cheaply. This includes research projects that use diffusion data routinely collected by web or phone applications. However, when data has to be accumulated by surveys, retracing the diffusion would be prohibitively costly. Relatively little work has been done on the analysis of models with latent diffusion processes. Furthermore, much research in this area has been conducted on theoretical modelling but not so much on thereafter estimating the resulting models. In this thesis I hope to contribute towards bridging this gap.

1.3 The Model, the Data and Previous Estimation

The "latent-diffusion-observed-adoption" model is a random diffusion model that stands out by its simplicity yet widespread applicability. In this model, a signal propagates through the Network. This signal can be thought of i.a. as a message, a virus, a piece of information or a debt claim. The researcher can only observe who initially obtained the signal and knows who is linked to whom. The network interaction, i.e. the spread of the signal, however, remains hidden to the researcher. Once the signal is newly received, it potentially triggers a change in an outcome variable. If so, this change is observed. This outcome could be an observed action, choice, illness or business default. However, only changes in outcomes are observed, thus it is ambiguous whether agents that do not change their outcomes did so deliberately or by lack of alternative (i.e. they never obtained the signal). The researcher uses the observed network, changes in outcomes and the signal initiation to infer the probability of signal spreading (the diffusion rate) and outcome change (the adoption rate).
The model can easily be enriched to include determinants or even a micro-foundation for either the adoption or the diffusion process. It can also be adapted to various circumstances by varying the distributional assumptions of the two random processes involved. In each case, the computational methods developed and code written for this thesis remain applicable.
The model was employed by Banerjee, Chandrasekhar, Duflo, and Jackson (2013) and accordingly, the real-world data used in this thesis also stems from this study and from Jackson, Rodriguez-Barraquer, and Tan (2012). Furthermore, Monte Carlo studies are used to examine the estimators' properties.
This model is computationally highly challenging to estimate due to the diffusion process being unobserved. The aforementioned study has used simulation methods to circumvent this problem. Due to the highly nonlinear nature of the model however, computation times are extremely long and it is oftentimes not feasible to execute the number of replications that would be desirable from a statistical standpoint. Investigation into faster, more efficient estimation techniques for this model is thus highly beneficial.
Standard software packages perform relatively poorly on network models. Further, the large amount of memory and computation power needed made it necessary for me to carry out my estimations on a supercomputer or computer cluster. Using a software package thus implied limiting my options since a cluster or supercomputer needed to have the correct software installed. A non-negligible part of the work on this thesis thus consisted of writing efficient code and self-executable scripts for estimation of

the diffusion models. The increase in speed when moving from software packages to low-level languages has been in the area of ten to twenty-fold.

1.4 Thesis Outline

The information contained in the network and the initiation is used to establish the distribution of the outcome variable, which is used in all proposed estimation methods. With the diffusion variables being unobserved, I obtain the distribution of the outcome by means of the Law of Total Probability, (i.e. by integrating out the latent diffusion variables). However, due to the dependencies of the diffusion variables across individuals, setting up the exact outcome distribution would imply summing up over close to an infinite number of terms. The aim of this thesis is to propose estimators that nonetheless facilitate parameter identification and hypothesis testing without relying on simulation methods.

1.4.1 The "Two-period Estimator"

In chapter two, I achieve tractability by restricting the time horizon to two periods. The complications that arise when multiple signal exchanges are considered arise due to the fact that agents involved in earlier exchanges may or may not receive the signal and may or may not pass it on. However, when only one round of diffusion is considered, these problems do not arise and exact ML estimation is feasible. Using the real-world data mentioned above, I estimate various nested models and test which channels impact diffusion and adoption. My results indicate that richer models including adoption determinants fit the data significantly better. Individuals with adopting friends are more likely to adopt themselves as are poorer individuals. The diffusion parameter is significantly different for adopters and non-adopters. These results thus have important practical implications when the aim is to increase the overall adoption rate: it may be worthwhile to specifically target poorer individuals and stimulate early adoption.

1.4.2 The "Trimming Estimator"

As the intractability arises from the fact that the information scenarios resulting from multiple rounds of diffusion are too numerous to be considered, dimension reduction can also be achieved by considering not all, but only a subset of the possible diffusion scenarios. The curse of dimensionality arises from "Potentially Informed Individuals" (PIIs), that is, agents who could have obtained the signal, but who never adopt. For

these individuals, the data does not allow the researcher to distinguish whether or not the signal has been obtained. The trimming strategy proposed in chapter three identifies those individuals for whom both scenarios (obtaining and not obtaining the signal) are close to equally likely. All remaining individuals, who consequently exhibit a large probability to either have been hit by the signal or not will be trimmed to the respective more likely scenario. Putting it differently, the number of senders must be sufficiently large for signal transmission and once the signal reception rate exceeds a second threshold, the signal is obtained with certainty. A Monte Carlo Study indicates that the trimming estimator converges to the full MLE relatively quickly when the fraction of individuals trimmed is in the range of roughly one third of all PIIs. The trimming strategy is thereafter applied to some real villages. While the number of PIIs in the real data is too large to be in the range where convergence occurs, the trimming results nonetheless highlight convincingly the improvement of the estimate when the trimming value increases.

1.4.3 The "Moment-based Estimators"

If the high dimensionality prevents ML estimation, moment-based estimation is a feasible alternative. Moment-based estimation with over-identification allows for the possibility to restrict the set of moments used for estimation to those that can readily be calculated. Excluding some information (the moments that are hard to compute) is inconsequential for parameter identification as long as the moments that are used for estimation are more numerous than the model parameters and are correctly specified. Precisely, the individual's probability to receive the signal can be evaluated by means of some shorthand formulas yet only in the very first period in which she can potentially receive the signal. These formulas provide a close-form expression for the signal reception probabilities that hence circumvent the challenge of having to integrate out the latent variables of the agent's neighbours. The computation of the unconditional expected value of the individual's outcome in the following period is thus easily executed, giving rise to individual-specific moment conditions. In Chapter four, two estimators that differ with respect to how the individual moments are employed in the sample objective function are proposed and their theoretical properties are investigated. A Monte Carlo study highlights the supremacy of the Non-aggregated estimator, thus showing once more that correlation across individuals within the same village is non-negligible and does not cancel out at the aggregate level. The estimators are subsequently employed to the real data.

1.5 Conclusion

Each of the following chapters proposes an estimation strategy applicable to the "latent-diffusion-observed-adoption" model. These estimators distinguish themselves with respect to how they achieve a dimension reduction. The limited research conducted on estimating latent diffusion on social networks so far and the applicability of the model to various economic settings make this research important and interesting.