TIME SERIES APPROACH TO THE EVOLUTION OF NETWORKS: PREDICTION AND ESTIMATION

ANNA BYKHOVSKAYA

Abstract. The paper analyzes non-negative multivariate time series which we interpret as weighted networks. We introduce a model where each coordinate of the time series represents a given edge across time. The number of time periods is treated as large compared to the size of the network. The model specifies the temporal evolution of a weighted network that combines classical autoregression with non-negativity, a positive probability of vanishing, and peer effect interactions between weights assigned to edges in the process. The main results provide criteria for stationarity vs. explosiveness of the network evolution process and techniques for estimation of the parameters of the model and for prediction of its future values.

Natural applications arise in networks of fixed number of agents, such as countries, large corporations, or small social communities. The paper provides an empirical implementation of the approach to monthly trade data in European Union. Overall, the results confirm that incorporating non-negativity of dependent variables into the model matters and incorporating peer effects leads to the improved prediction power.

Date: January 6, 2020.

The author would like to thank Donald Andrews, Vadim Gorin, Bruce Hansen, Peter Phillips, Jack Porter, and Larry Samuelson for valuable comments, suggestions, and encouragement throughout the duration of the project.

1. Introduction

1.1. Motivation. This paper studies non-negative time series. We are especially interested in multivariate time series, whose components can be interpreted as edges of some network (if a network is composed of only two vertices, then we are dealing with a basic one-dimensional time series). The novelty of our view on networks lies in the time series approach. Citing Diebold and Yılmaz (2014), "the network and multivariate time series literatures have much to learn from each other", and our paper makes a step in this direction. We develop a model and a consistent estimator of its parameters. The three main features of our model are non-negativity of weights associated with edges, positive probability of vanishing of each edge, and possibility for edges to affect each other in the multivariate setting.

Non-negative time series often arise in different socio-economic settings. For example, production decisions by firms cannot be negative. Similarly, technology adoption by firms or countries is a non-negative variable, which evolves over time. Macroeconomic variables such as exchange rates, T-bill rates and so on are non-negative by definition. Finally, if we look at the number of passengers travelling from country A to country B at a given day, it is again either zero or positive.

One can note that in almost all of the above examples zero occurs with positive probability: there are times when firms decide not to produce at all, or two countries do not sell some good to each other, or a group of people do not call each other and so on. Thus, if one wants to find a suitable model for such non-negative processes, allowing it to take zero values with positive probability is an important condition. However, this feature cannot be captured by linear models or by their simplest modifications such as GARCH.

Moreover, we look not only at a one-dimensional time series, but also at multivariate ones. For instance, instead of considering the traffic flow from one given country to another, we can track the whole set of countries, and get a non-negative weighted network, which governs the amount of travel. In the same spirit, we can construct a time-varying network of trade between different countries or firms or a social network, which measures the amount of communication between people. In this paper we focus on networks (or weighted graphs) as a main example of a non-negative structure. In contrast to other cases, networks may potentially involve interactions between the coordinates corresponding to their edges. For example, if a firm A is in search of a provider of some intermediate good, it may contact firm B, with which it has well-established trading relationship. Even if B cannot provide such intermediate good, it may recommend firm C, with which B in turn has a well-established relationship. Robinson and Stuart (2006) illustrate that the whole network of past alliances in the biotechnology industry affects the structure and size of alliance agreements between any two given firms in the industry in the future. (For more examples on the role of social and economic networks in real life see, for instance, Jackson (2010).)

Often one may need to be able to predict a structure of a network as it evolves. It may be important to know whom is better to target with ads, election campaigns, or other type of information. Countries may need to know the future trading patterns to decide production of which types of goods to support. We also refer to Holme and Saramäki (2012) for an interdisciplinary review highlighting the importance of the temporal structure of networks. The availability of time series data for social and economic networks, such as call detail records and international and financial trade data has grown tremendously. In this paper we use such time-series data to estimate parameters of network evolution and to make predictions. When one goes to the data, working with networks as a multivariate time series (when the number of time periods is large compared to the size of the network) and not as a cross-section or panel makes a difference between our paper and what is currently done in networks literature (see e.g. Bramoullé et al., eds (2016)). One natural example of applicability of our setting is when agents are countries or large companies (as there are not so many of them). For social communities this is also of relevance, as the dynamics of small groups of people might be very different from the large ones (see e.g., Palla et al. (2007)).

1.2. Results. In this paper we deal with multivariate non-negative time series, which we interpret as weighted networks (although alternative interpretations and apllications are also possible and some of them are presented in Section 7). Each edge is associated with a number, which evolves with time. We want to capture three essential aspects of networks. Those properties are non-negativity of weights, possibility not to have an edge between any two nodes with positive probability, and possibility for the past of the network to affect all edges today. These three features lead us to modelling the evolution of a network as a non-linear process

$$y_{ijt} = [\alpha_{ij} + \beta_{ij} y_{ijt-1} + \gamma_{ij} z_{ijt-1} + u_{ijt}]_+, \tag{1}$$

where y_{ijt} is the weight of the edge $(i \to j)$, $[\cdot]_+$ stands for positive part, α_{ij}, β_{ij}, γ_{ij} are unknown coefficients, which need to be estimated, u_{ijt} is a random error, which is independent across time t, but may be correlated along edges (i, j), and z_{ijt-1} is a peer effect or an interaction term. That is, z_{ijt-1} is a function (assumed to be explicitly known) of past periods of the network. It allows edges to affect each other in the future. Note that we do not impose linearity on z. It may be a non-linear function giving us a lot of flexibility in how we model peer-effects. Moreover, the positive part per se creates nonlinearity, which leads to technical difficulties. The equation (1) can be obtained as a solution to a certain utility maximization problem, as we outline in Section 2. One can add more regressors to Eq. (1) and most of our results continue to hold in the extended setting (see Remarks 1 and 2). We note that there is a large number of much more sophisticated models for the network evolution in the literature (e.g. Pin and Rogers (2015) where agents play a

prisoner's dilemma every period and choose connections based on the observed outcomes), yet the estimation procedures become more advanced and model-specific with each layer of complexity. It becomes preferable to address the basic case, so that the general principles and challenges can be identified.

Our setting allows for arbitrary heteroskedastisity in the errors u_{ijt}, that is, arbitrary correlation across individuals. This differs a lot from what is feasible when one relies on cross-sectional variation instead of the variation across time. The classical approach to networks (n is large and $T = 1$) assumes either dyadic error structure where $Cov(u_{ijt}, u_{k\ell t}) = 0$ if $i \neq k, \ell$ and $j \neq k, \ell$ (e.g., Graham (2017), Graham (2019)) or other type of decreasing correlation, so that one can average over individuals and use law of large numers and central limit theorem.

In principle, an alternative basic approach could be to consider a latent censored model, see e.g., Wei (1999) and references therein. That is to assume that there is some unobserved underlying process y_t, which is allowed to take negative values. The researcher instead observes the truncated process of the form $y_t^* = y_t \mathbf{1}(y_t > 0)$. In our case there need not exist such an underlying process. Moreover, the censored structure makes the observed process non-Markovian. Thus, it is hard to make predictions. Our approach, in contrast, guarantees, a Markovian structure (in a state space taking into account several time lags). So it is easy to make predictions based on our model.

Let us describe our main findings. First, we show a sufficient condition for the model to be stationary. The model does not belong to any known class and requires a different treatment. We normalize z_{ijt} in a way which insures that peer effects do not grow faster than their maximal argument (see Section 2.3 for more details). We prove that if $\max(0, \beta_{ij}) + |\gamma_{ij}|$ is uniformly over i, j bounded by some constant smaller than one, then the process is strongly mixing and converges (when started from any initial conditions) to a stationary distribution as $t \to \infty$. For the special case of the model where $\gamma_{ij} \equiv 0$, we provide the full classification of asymptotic behavior of the process: the change from stationary to explosive behavior is at $\alpha_{ij} = 0$, $\beta_{ij} = 1$. At that boundary case, the process rescaled by $\sqrt{T}$ converges to the absolute value of a Brownian motion as $T \to \infty$. One distinction with the usual linear autoregression model is that the $\alpha_{ij} < 0$, $\beta_{ij} = 1$ case is stationary due to the presence of positive part.

We remark that the non-linearities created by $[\cdot]_+$ and by z_{ijt} make asymptotic analysis challenging. The arbitrary peer-effect function $z_{ij}(\cdot)$ significantly complicates the setting, leading to a potentially non-Lipschitz dependence on the past. For example, the triangular peer-effect function $z_{ijt} = \sum_k \frac{\sqrt{y_{ikt} y_{kjt}}}{n-2}$, which we use in the empirical example, is not Lipschitz. Thus, the classical arguments for stationarity, which are based on the contraction mapping, are not applicable and one has to come up with a different approach. To overcome those

problems, we develop a proof based on the large deviations principle for $\{y_{ijt}\}$ and analyze the expected time it takes for the process to jump to zero. Then we apply renewal theorem to obtain mixing and stationarity independent of initial conditions. As far as we know, our stationarity results are new; the only relevant papers seems to be Jong and Herrera (2011), Hahn and Kuersteiner (2010), and Michel and de Jong (2018). Among other results, in these papers sufficient conditions for the existence of a stationary solution are found for the special case of Eq. (1) dealing with a one-dimensional process with z linearly depending on the past periods. This particular stationary solution (rather than a general one) is further shown to be strongly mixing.

Let us stress that in contrast to classical autoregression, where a continuous distribution of the errors is important to establish strong mixing (see Withers (1981) and Andrews (1984) for examples of non-strongly mixing AR processes with discrete errors), in our setting we only need the errors to have support that is unbounded from below. Thus, the distribution is not required to be continuous. This also differs from Hahn and Kuersteiner (2010), who require a continuous distribution of the errors to obtain strong mixing.

Second, we discuss how to estimate the parameters of the model. The nonlinearities due to censoring and peer effects significantly complicate estimation. The ordinary least squares (OLS) estimator turns out to be inconsistent in our setting. This matches a similar inconsistency for censored regression models, c.f. discussion at the end of Section 4.2 in Amemiya (1984). Our approach builds on and extends the least absolute deviations (LAD) method, which was used in the context of censored regression in Powell (1984). We obtain an estimator by minimizing with respect to α_{ij}, β_{ij}, γ_{ij} the sum of the absolute differences between y_{ijt} and $[\alpha_{ij}+\beta_{ij}y_{ijt-1}+\gamma_{ij}z_{ijt-1}]_+$. Theorems from Powell (1984) are not applicable in our setting because the independence assumptions from that paper do not hold. Thus, novel ideas are required.

Some of our results cover explosive cases and we show that while the large T asymptotics of the system changes drastically, the LAD estimator is still consistent. To our knowledge, we are the first to analyze LAD in the explosive setting. Our findings are in line with results on the consistency of OLS in the explosive autoregressive model (see White (1958) and Anderson (1959) for the model without a constant and Wang and Yu (2015) for the model with an intercept.)

We remark that, in general, minimization of absolute deviations in models with a positive part is a non-convex problem and designing numerical algorithms requires special care (see e.g. Khan and Powell (2001)). Similarly, asymptotic normality in the censored cross-section model in Powell (1984) relies on certain continuity properties of a function of the true parameter value and the error distribution, which are hard to check. In contrast, we find that whenever z has non-negative support and the true α_{ij}, β_{ij}, γ_{ij} are all positive, neither of

these problems exist in our setting: the optimization problem is convex, asymptotic normality does not rely on any additional conditions, and the asymptotic variance of the LAD estimator is given by a simple formula.

We also show in the Supplementary Material how to correct the OLS procedure to restore consistency, yet in doing so one needs to ignore a lot of observations, and, thus, the accuracy of the estimation decreases significantly. Further, if we assume the errors u_{ijt} are Gaussian, then we can explicitly write down the likelihood function. We prove that in this case the maximum likelihood estimator (MLE) is consistent. However, for the semiparametric case when the distribution of u_{ijt} is not specified, there is no guarantee that the Gaussian MLE is consistent.

Being able to consistently estimate γ_{ij} allows us to analyze the importance of peer effects in the network formation process. This relates our paper to the vast literature on estimating peer effects in various network models (e.g. see Blume et al. (2011) for review). Several approaches rely on specific functional forms to ensure a consistent estimate (e.g. triangular peer effects in Graham (2016), which resemble our triangles example after Assumption 5 in Section 2.3), while others introduce model free notions for the strength of peer effects (such as connectedness in Diebold and Yılmaz (2015)). However, we have not seen in the literature estimation techniques based on variation across time in non-linear Markov evolution model, as in our work.

Finally, we propose an objective quality measurement of the model by looking at the absolute value of the prediction error at time $t + T'$ for the model estimated from the data from time t and up to time $t + T' - 1$ for all $t = 0, \dots, T - T'$, where T' is the size of the window used for estimation. Let us emphasize that the use of absolute values instead of the more standard squared differences is important here: for the latter the optimal prediction (even if α_{ij}, β_{ij}, γ_{ij} are exactly known) depends on the distribution of errors, while for the former we can speak about model-independent prediction power.

To illustrate our methodology, we apply the framework to monthly trade data between European Union countries for pharmaceutical products. We use the basic prediction "tomorrow=today" as a benchmark for comparing the prediction power, and the interactions (peer effects) are modelled by a function based on the "friend of my friend is my friend" principle. In our experiments, any model-based estimation techniques lead to an improvement of the power over benchmark case. The basic linear OLS procedure, which ignores the positivity of weights, leads to the worst results among model-based estimations. The MLE estimator performs better, and the LAD estimator leads to the best results. The addition of peer effects z_{ijt} (vs. setting $\gamma_{ij} = 0$ in the model specification) also leads to improved prediction power. Results of the Diebold-Mariano test (Diebold and Mariano (1995)) and

the tests for the significance of peer effects, which are reported in Section 6, support the above conclusions.

1.3. Outline of the paper. Section 2 presents the model and the main equation of interest. All assumptions are stated in that section, as well as sufficient conditions for stationarity. Section 3 discusses the special case of the model when there are no peer effects, so that evolution of each edge is a separate process. The full classification in terms of stationary/explosive behavior is established in this case. Section 4 discusses estimation of the model, while Section 5 proposes the method to measure predictive power. Section 6 applies the model to the trade of pharmaceutical products in European Union. Section 7 discusses extensions of our setting. Finally, Section 8 concludes. All proofs, unless otherwise noted, are in the Appendix. Supplementary Material contains results on OLS and MLE estimation and additional lemmas used to prove the main theorems.

2. Model

2.1. Set up. We analyze a multivariate time series, which we interpret as a network. The network consists of n vertices and evolves across time. The network is observed over T periods. In our asymptotic results we assume that n is fixed and T goes to infinity. Thus, we are dealing with many repeated observations of a small network.

Our model allows both for undirected and directed networks. The main example of the former is a social weighted network, where nodes represent people and edges represent how much time they spend together. E.g., one can use phone call data as a measure of friendship (the more two people text or talk to each other, the closer their relationship is). For the case of a directed network, the applications are mostly for firms or countries and trade between them. Looking separately at exports and imports, we get a directed network.

The equation of interest is

$$y_{ijt} = \left[\alpha_{ij} + \beta_{ij} y_{ijt-1} + \sum_{\kappa=1}^{K} \gamma_{ij}^{\kappa} p_{ij}^{\kappa} \left(\left\{ y_{kls} \right\}_{\substack{\text{k,l=1,...,n} \\ \text{s=t-H,...,t-1}}} \right) + u_{ijt} \right]_{+}, \tag{2}$$

where u_{ijt} is a random error, $t = 1, \ldots, T$ stands for time, and $i, j = 1, \ldots, n$ stand for agents (in the undirected case $i < j$, in the directed case $i \neq j$). So that y_{ijt} can be interpreted as either how much i and j talk at time t or the amount of trade from i to j.

In Eq. (2) $\alpha_{ij} \in \mathbb{R}$ represents homophily, i.e. how similar i and j are. The larger α_{ij} is, the stronger is the link connecting i and j. We allow β_{ij} and γ_{ij}^{κ}, $\kappa = 1, \ldots, K$ to be of any sign. The coefficient β_{ij} measures the dependence on the own past. The larger β_{ij} is, the more the link between i and j yesterday affects its weight today. Coefficients γ_{ij}^{κ} capture the dependence on the peer effects/interactions p_{ij}^{κ}. Functions p_{ij}^{κ} serve as various aggregators

of the past structure of the network in a way that affects the current state. The multitude of peer-effect terms gives a lot of flexibility and allows one to capture very general, diverse forms of interactions. A lot of features of networks can be captured throught p_{ij}^{κ}. Finally, the positive part in Eq. (2) creates nonlinearity and leads to a positive mass at zero.

The model is initialized at $t = 1 - H, \dots, 0$ by arbitrary values (possibly random). To be more precise, we assume that as T goes to infinity, H does not grow and $y_{ijt} = O(1)$ for $t = 1 - H, \dots, 0$.

For simplicity, in the rest of the paper we will focus on the special case of Eq. (2) where $K = 1$,

$$y_{ijt} = \left[\alpha_{ij} + \beta_{ij} y_{ijt-1} + \gamma_{ij} p_{ij} \left(\left\{ y_{kls} \right\}_{\substack{\text{k,l=1,...,n} \\ \text{s=t-H,...,t-1}}} \right) + u_{ijt} \right]_+ . \tag{3}$$

Focusing on $K = 1$ allows us to shorten the notation. As noted in the Remarks 1 and 2, our results still hold in the extended setting of Eq. (2).

2.2. Maximization problem. In this subsection we present a stylized game theoretical model that leads to our equation of interest, Eq. (3). It is another justification of Eq. (3) along with the primary one, which is to capture a number of essential properties of networks (non-negativity of edges, positive probability of vanishing of each edge, and interactions between edges, which affect the whole network).

Consider a world with n myopic agents (people/firms/countries/etc.) with quadratic adjustment costs. Agents can interact with each other over time. Time is discrete and goes from 1 to T. Every period, each agent i chooses how much time to spend with or how much to trade with each other agent j. The decision is based on two components: costs and benefits.

Benefits are characterized by a per unit gain of $\alpha_{ij} + u_{ijt}$. Here α_{ij} is a constant, while u_{ijt} is random component that is independent across time. Thus, y units of communication/trade leads to a benefit of $y(\alpha_{ij} + u_{ijt})$.

The second component is a quadratic adjustment cost function. Agents get disutility whenever there are deviations from some target expected level of communication/trade. The target is composed from an own past and a peer-effect or interactions component. The interaction term aggregates the whole structure of the network for up to H periods. That is, we assume that agent i by choosing to devote y units to agent j has to pay

$$\frac{1}{2} \left(y - \beta_{ij} y_{ijt-1} - \gamma_{ij} p_{ij} \left(\left\{ y_{kls} \right\}_{\substack{\text{k,l=1,...,n} \\ \text{s=t-H,...,t-1}}} \right) \right)^2 ,$$

where

$$p_{ij}\left(\left\{y_{kls}\right\}_{\substack{\text{k,l=1,...,n}\\ \text{s=t-H,...,t-1}}}\right):\mathbb{R}_{+}^{n^2H}\to\mathbb{R}$$

represents peer effects/interactions function, which depends on H previous periods.

The interpretation is that β_{ij} represents a rate at which stock/relationship depreciates/appreciates. If $\beta_{ij}<1$, then the agent is introverted and tends to decrease communication; while if $\beta_{ij}>1$, the agent is an extrovert, who tends to expand communication. The interpretation of β_{ij} for firms corresponds to production depreciation (i.e. technology wears out) or production appreciation (better technology management over time increases production). The coefficient γ_{ij} indexes the sensitivity of the reference level with respect to peer effects/interactions, which are, in turn, represented by the function p_{ij}. The peer effects function depends on H past periods of the whole network, and captures interactions between different edges y_{kls} across time.

Agents have separate maximization problems for each time period t and with each peer j. Agent i solves the following maximization problem at day t with respect to agent j:

$$\max_{y\geq 0}\left[y(\alpha_{ij}+u_{ijt})-\frac{1}{2}\left(y-\beta_{ij}y_{ijt-1}-\gamma_{ij}p_{ij}\left(\left\{y_{kls}\right\}_{\substack{\text{k,l=1,...,n}\\ \text{s=t-H,...,t-1}}}\right)\right)^2\right]. \tag{4}$$

The solution to the maximization problem (4) is

$$y_{ijt}^{*}=\left[\alpha_{ij}+\beta_{ij}y_{ijt-1}+\gamma_{ij}p_{ij}\left(\left\{y_{kls}\right\}_{\substack{\text{k,l=1,...,n}\\ \text{s=t-H,...,t-1}}}\right)+u_{ijt}\right]_{+},$$

which leads to the network evolution process described by Eq. (3).

2.3. Assumptions. We need to impose some assumptions on the error distribution and on the peer effects function.

Assumption 1. *The vector* $\{u_{ijt}\}_{i,j=1,...,n}$ *is i.i.d. over* t.

Assumption 1 allows errors to have arbitrary correlation across edges of the network and ensures that those correlations do not vary over time. This difference substantially from classical approach to networks, where they are modeled as cross-section or short panel. In those cases the correlation across individuals must either form a dyadic relationship or decrease with some measure of a distance, so that one can average over individuals and apply law of large numbers. In our setting n is assumed to be fixed, and instead the variation across T is used to identify the network formation process. Thus, our setting does not require to rely on "decreasing across n correlations" and can accommodate any type of cross-sectional interdependence.

Assumption 2. *The vector* $\{u_{ijt}\}_{i,j=1,\dots,n}$ *has support s.t.* $\mathbb{P}(u_{ijt} < -M \ \forall i,j) > 0$ *for all* $M > 0$.

Assumption 2 is used to show stationarity. It implies that errors jointly take large negative values with positive probability. The consequence is that for any values of the process $\{y_{ijt}\}_{i,j}$ at time t, there is a positive probability that at time $t+1$ the process jumps to zero. This observation is crucial for Theorems 2 and 3, as it ensures that for specific range of parameters the process forgets its past in a finite time and restarts from scratch.

Eq. (3) has one degree of freedom, so we need to impose a normalization assumption on the error term u_{ijt}. We consider two different normalization assumptions: zero mean or zero median, which are stated below.

Assumption 3 (Normalization of the mean.)**.** *For all* i,j,t, $\mathbb{E}u_{ijt} = 0$.

Assumption 4 (Normalization of the median.)**.** *For all* i,j,t, $med(u_{ijt}) = 0$.

Alternative Assumptions 3 and 4 only lead to differences in α_{ij}:

$$\alpha_{ij}^{E} = \alpha_{ij}^{\text{med}} + \mathbb{E}u_{ijt}^{\text{med}},$$

where α_{ij}^{E} is an intercept under Assumption 3 and α_{ij}^{med} and $\mathbb{E}u_{ijt}^{\text{med}}$ are an intercept and a mean of the error under Assumption 4.

Assumption 5 (Peer effects do not grow faster than their maximal argument.)**.**
$p : \mathbb{R}_{+}^{n^2H} \to \mathbb{R}$ *is such that there exists a constant* $\mathcal{A} \in \mathbb{R}$ *for which*

$$\left| p\left(\left\{ y_{kls} \right\}_{\substack{k,l=1,\dots,n \\ s=t-H,\dots,t-1}} \right) \right| \leq \mathcal{A} + \max_{\substack{k,l \\ s=t-H,\dots,t-1}} y_{kls}.$$

Let us present some examples of possible peer effect functions.

- *Maximum*:

$$p_{ij}\left(\left\{ y_{kls} \right\}_{\substack{k,l=1,\dots,n \\ s=t-H,\dots,t-1}} \right) = \max_{\substack{(k,l)\neq(i,j) \\ s=t-H,\dots,t-1}} y_{kls}.$$

  This function represents the largest possible stimulus to increase trade or communication. This can be interpreted as a steadily expanding economy.

- *Minimum*:

$$p_{ij}\left(\left\{ y_{kls} \right\}_{\substack{k,l=1,\dots,n \\ s=t-H,\dots,t-1}} \right) = \min_{\substack{(k,l)\neq(i,j) \\ s=t-H,\dots,t-1}} y_{kls}.$$

  This function corresponds to the smallest, but still non-zero influence from others. That is, if some edge jumps to zero, it pushes the other edges in that direction. Alternatively, if all edges have positive weights, the peer effect term is still positive and helps to maintain a non-zero edge between i and j.

- *Linear*:

$$p_{ij}\left(\left\{y_{kls}\right\}_{\substack{k,l=1,\dots,n\\ s=t-H,\dots,t-1}}\right) = \sum_{\substack{\text{k,l=1,}\dots\text{,n}\\ \text{r=1,}\dots\text{,H}}} \lambda_{klr} y_{klt-r},$$

where $\{\lambda_{klr}\}_{\substack{k,l=1,\dots,n\\ r=1,\dots,H}}$ are known and $\lambda_{klr} \geq 0$, $\sum_{\substack{\text{k,l=1,}\dots\text{,n}\\ \text{r=1,}\dots\text{,H}}} \lambda_{klr} \leq 1$.

The linear function represents an intermediate point between the previous examples.

- *Triangles*:

$$p_{ij}\left(\left\{y_{kls}\right\}_{\substack{k,l=1,\dots,n\\ s=t-H,\dots,t-1}}\right) = \sum_{k\neq i,j} \frac{\sqrt{y_{ikt-H} y_{kjt-H}}}{n-2}.$$

The square root makes the order of the peer effect term to be the same as the order of the autoregressive term, y_{ijt}. Thus, Assumption 5 is satisfied.

Triangular peer effects have the most interesting functional form, and we use it in our empirical application. The interpretation is that if i and k are strongly connected, and k and j are too, then there is a higher probability of a connection between i and j in the future. Note that the product makes it important that both connections are present. If k and j are not connected, we cannot expect k to "introduce j to i". Thus, we look at all triangles which have (i, j) as one of their legs. Two strong links in such triangles are expected to strengthen the third leg, (i, j). This can be summed up as "friend of my friend is my friend". Such peer effects may be present in social interactions, interactions between firms or countries, and so on.

The triangular peer effect function is related to the network formation process in Graham (2016), where the presence of an edge between two nodes depends on the number of triangules containig those two nodes in the past. In Graham (2016) edges do not have weights, so the number of triangles containing i and j at time t is $\sum_k D_{ikt} D_{kjt}$, where $D_{ikt} = 1$ when there is an edge between i and k at time t. A similar statistics $\max_k D_{ikt} D_{kjt}$ is used in Leung and Moon (2019) to generate network clustering.

2.4. Stationarity. Theorem 1, which is proved in the Appendix, provides sufficient conditions under which the network does not explode. Non-exploding does not guarantee convergence to a stationary distribution as formally the process may have cycles. Yet, it is enough for our estimation and prediction approaches to work. Moreover, if we additionally assume that $\{u_{ijt}\}_{i,j}$ has unbounded from below support (Assumption 2), then stationarity holds (Theorem 2).

Theorem 1. *Suppose that Assumptions 1 and 5 are satisfied, $\mathbb{E}u_{ijt}$ exists for all i, j, t, and there exists a constant $C \in (0,1)$ such that $\max(0, \beta_{ij}) + |\gamma_{ij}| < C$ for all i, j. Then the multivariate process $\{y_{ijt} : i, j = 1 \dots, n\}_{t \geq 1}$ does not explode (i.e. there exists a constant C_1 such that $\mathbb{E}y_{ijt} < C_1 < \infty$ for all i, j, t).*

Definition. *The process $\bar{y}_t = \{y_{ijt}\}_{i,j}$ is strongly mixing if for arbitrary Borel sets Δ_1, Δ_2*

$$\lim_{t \to \infty} |\mathbb{P}(\bar{y}_s \in \Delta_1, \bar{y}_{t+s} \in \Delta_2) - \mathbb{P}(\bar{y}_s \in \Delta_1)\mathbb{P}(\bar{y}_{t+s} \in \Delta_2)| = 0.$$

Theorem 2. *Suppose that Assumptions 1, 2, and 5 are satisfied, $\mathbb{E}u_{ijt}^4 < \infty$ for all i, j, t, and there exists a constant $C \in (0,1)$ such that $\max(0, \beta_{ij}) + |\gamma_{ij}| < C$ for all i, j, then the multivariate process $\{y_{ijt} : i, j = 1 \dots, n\}_{t \geq 1}$ is strongly mixing and converges to a stationary process.*

Remark 1. In the extended setting of Eq. (2), Assumption 5 should hold for each p_{ij}^{κ}, and $|\gamma_{ij}|$ in Theorem 2 should be replaced by the sum $\sum_{\kappa=1}^{K} |\gamma_{ij}^{\kappa}|$.

A striking feature of Theorem 2 is that we do not need the error distribution to be continuous to get strong mixing. This differs from the linear case (see Withers (1981) and Andrews (1984) for examples of $AR(1)$ processes which are not strongly mixing). The reason is that in our setting the expected time until the process jumps to be identically zero ($y_{ijt} = 0$ for all i, j) is finite. Thus, the process forgets the initial condition in finite time.

Example 1. The fact that the peer effects function p_{ij} can depend only on a fixed number of time periods is crucial. For example, suppose that we have only one equation ($n = 2$) which is initialized at $y_0 = 0$, and the error process u_t has unbounded support from above. Further suppose $\alpha = \beta = 0$, $\gamma = 0.5$ and $z_t := p_{ij}(y_t, \dots, y_0) = \max(y_t, \dots, y_0)$, so that Assumption 5 is satisfied. Then $\beta + |\gamma| = 0.5 < 1$, but the process y_t is explosive. To see this, let us analyze the behavior of y_t and z_t.

By definition, $z_0 = y_0 = 0$, so that $y_1 = [u_1]_+$ and $z_1 = \max(0, u_1) \geq 0$. Thus, $y_2 = [0.5z_1 + u_2]_+ \geq [u_2]_+$ and $z_2 = \max(0, y_1, y_2) \geq \max(0, u_1, u_2)$. Similarly, $y_3 = [0.5z_2 + u_3]_+ \geq [u_3]_+$ so that $z_3 \geq \max(0, u_1, u_2, u_3)$. Applying induction, for any t we get $z_t \geq \max(0, u_1, \dots, u_t)$. Therefore, $z_t \xrightarrow[t \to \infty]{a.s.} \infty$, as the support of u_t is unbounded from above and the maximum of an infinite number of random variables with unbounded support diverges. Because $y_t = [0.5z_{t-1} + u_t]_+$, y_t also goes to infinity almost surely.

3. Special case: No interactions

In this section we consider a special case, where the connection between i and j at time t depends only on its past. That is, past interactions between $k \neq i, j$ and l do not influence

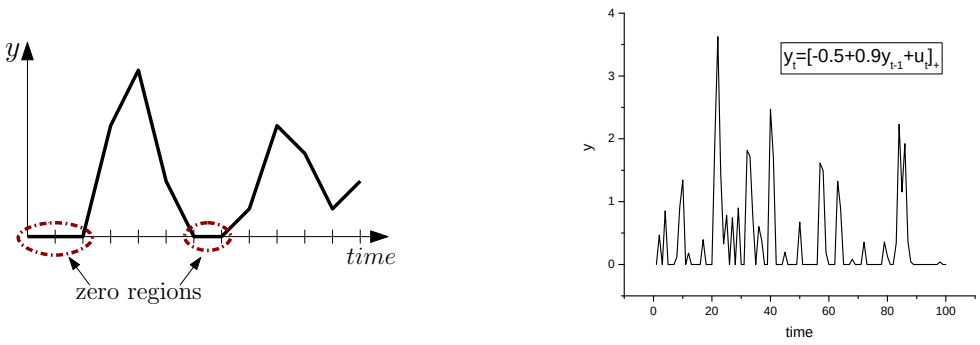

FIGURE 1. A typical sample path for $y_t = [\alpha + \beta y_{t-1} + u_t]_+$, $u_t \sim i.i.d.(0, \sigma^2)$.

y_{ijt}. Thus, the model reduces to $\frac{n(n-1)}{2}$ separate equations of the form

$$y_t = [\alpha + \beta y_{t-1} + u_t]_+, \; u_t \sim i.i.d.(0, \sigma^2). \tag{5}$$

A typical sample path for Eq. (5) is shown in Figure 1. When the process hits zero, it stays at zero for some time, then goes to an "AR(1)-excursion", until it becomes negative. Positive part in (5) then forces y_t to become zero instead, and everything starts again.

The following theorem provides a full classification of stationary/explosive behavior in the case of no interactions. In contrast to classical autoregression which has no positive part, when $\beta = 1$ the process still converges to a stationary distribution when $\alpha < 0$.

Theorem 3. *(Classification Theorem) Let Assumption 1 hold. Under the assumptions that u_t has unbounded support from below, $\mathbb{E}u_t = 0$, and $\mathbb{E}u_t^4 < \infty$,*

- *If $\beta < 1$, then y_t is strongly mixing and converges to a stationary process*[1]*;*
- *If $\beta = 1$, $\alpha < 0$, then y_t is strongly mixing and converges to a stationary process;*
- *If $\beta > 1$, then y_t is divergent: $y_t \xrightarrow{a.s.} \infty$;*
- *If $\beta = 1$, $\alpha > 0$, then y_t is divergent: $y_t \xrightarrow{a.s.} \infty$;*
- *If $\beta = 1$, $\alpha = 0$, then y_t is mean-divergent: $\mathbb{E}y_t \to \infty$. The proper scaling limit is*

$$\frac{1}{\sqrt{T}} y_{\lfloor Tr \rfloor} \xrightarrow{d} \sigma |W(r)|, \; r \in [0,1], \;\; as \; T \to \infty,$$

where $W(\cdot)$ is a standard Brownian motion and $\mathbb{E}u_t^2 = \sigma^2$.

A visual summary of the results in Theorem 3 is shown in Figure 2, where the evolution of y_t is illustrated for different values of α and β.

Theorem 3 is proved in the Section A of the Appendix (Theorems A.1, A.2, A.3, and A.5). The stationarity part of the proof relies on the large deviations principle and the renewal theorem. The idea is to show that the expected time until the process reaches zero is finite. Then one can apply the renewal theorem to get the limiting distribution. Interestingly, just

[1]Formally this means that the finite-dimensional distributions of the process $\{y_{t+\tau}\}_{\tau \in \mathbb{Z}}$, converge to those of a stationary in τ process as $t \to \infty$.

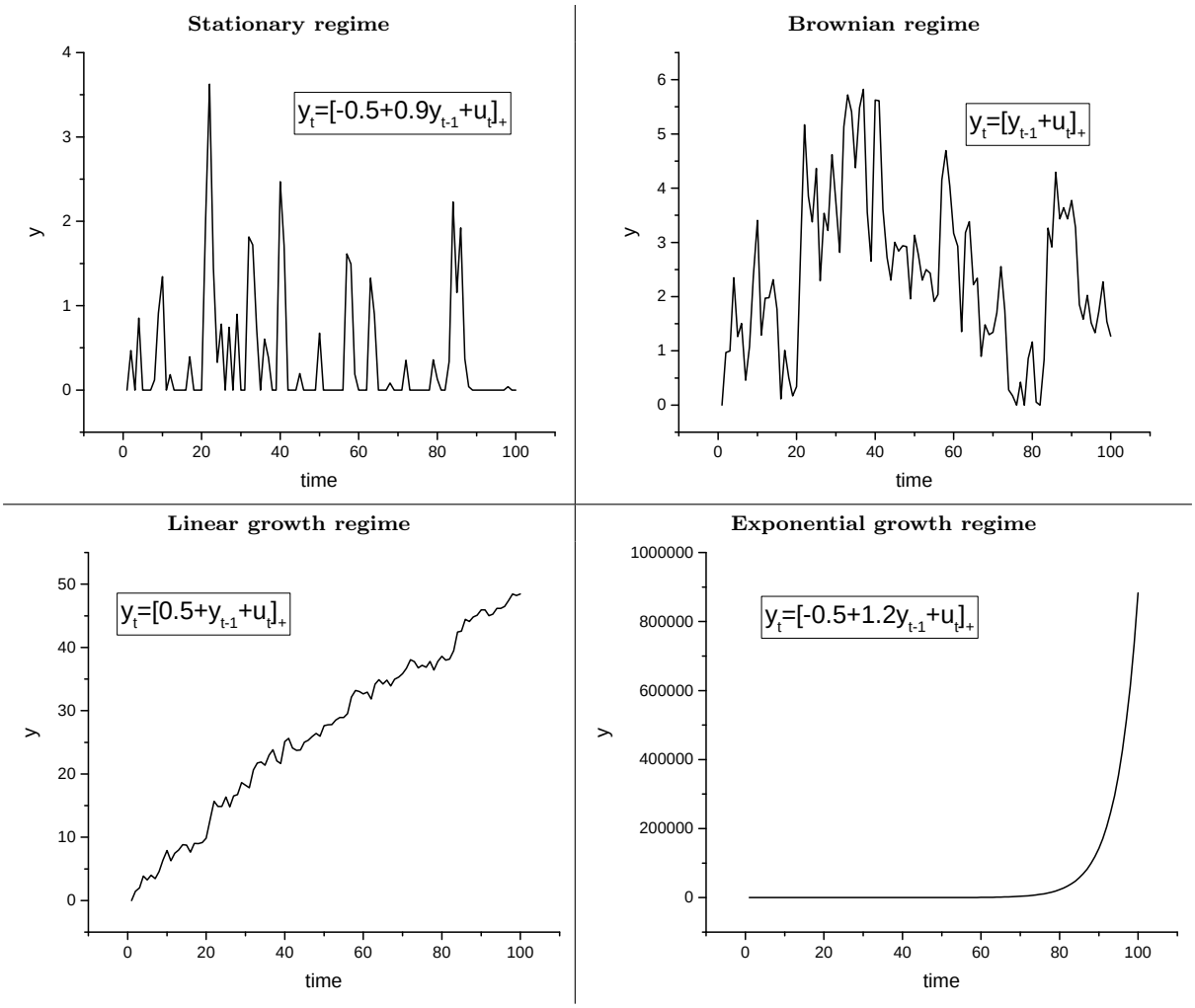

FIGURE 2. Illustration of Theorem 3.

the knowledge that the process hits zero with probability one is not enough. In particular, when $\alpha = 0, \beta = 1$ we get a standard unit-root process while y_t is positive. This process always hits zero. However, it does not converge to a stationary distribution, and as Theorem 3 shows, the process $y_t = [y_{t-1} + u_t]_+$ has exploding mean. Thus, there is a discontinuity between the stationary and explosive regions. This is similar to what is observed in the classical linear autoregressive case.

The limiting distribution in the stationary case of Theorem 3 is complicated and cannot be written explicitly as a function of α, β, and the distribution of u. It can only be obtained numerically. However, what can be calculated is the expected time the process y_t spends at zero once it hits zero. As shown in Lemma 4, it equals $\frac{1}{1-F_u(-\alpha)}$, where F_u is the cumulative distribution function of u. Thus, the smaller is α, the longer are the intervals of zeros.

Lemma 4. *Once the process y_t hits zero, the expected time it spends until finally jumping to a positive value is $\frac{1}{1-F_u(-\alpha)}$, where F_u is the cdf of u_t.*

Proof. If $y_t = 0$, then $y_{t+1} > 0$ when $u_{t+1} > -\alpha$. Therefore, after each zero with probability $F_u(-\alpha)$ the process remains at zero and with the remaining probability it becomes positive. Thus, we get a sequence of Bernoulli random variables, where the expected time until the first tail observation is

$$1 \cdot (1 - F_u(-\alpha)) + 2 \cdot F_u(-\alpha)(1 - F_u(-\alpha)) + 3 \cdot F_u^2(-\alpha)(1 - F_u(-\alpha)) + \ldots = \frac{1}{1 - F_u(-\alpha)}. \quad \square$$

4. Estimation

We now return back to the general model summarized in Eq. (3). The two main difficulties in our model are the interactions between the outcome variables and the non-separable errors. We can overcome the former by estimating the following model

$$y_t = [\alpha + \beta y_{t-1} + \gamma z_{t-1} + u_t]_+. \tag{6}$$

To be more specific, for each pair (i, j) we define

$$z_{ijt-1} = p_{ij}\left(\left\{y_{kls}\right\}_{\substack{\text{k,l=1,...,n} \\ \text{s=t-H,...,t-1}}}\right)$$

and ignore the fact that z_{ijt} are functions of lags of y_{klt}, $k, l = 1, \ldots, n$. That is, after we calculate the values of z_{ijt}, we are not going to use the fact that those values were obtained from $\{y_{klt}\}_{k,l=1,\ldots,n}$ and its lages. Instead we treat z_{ijt} as any other regressors. For each edge $(i \to j)$ we separately estimate Eq. (6) with $y_t = y_{ijt}$, $z_t = z_{ijt}$.

The natural question is whether we lose predictive power by treating each equation independently or not. Generally it is not clear and the answer should depend on the class of estimators we consider. Yet, there is a reason to believe that we do not lose a lot. Gourieroux and Monfort (1980) show that for the linear models like vector autoregressions if the errors are independent across components of y_t (in our case across edges) or the matrices of regressors span the same subspace for each component of y_t, then single-equation generalized least squares (GLS) is equivalent to overall GLS.

In the following subsections we present an approach to estimating the model (6) and discuss the properties of the estimator. Unless otherwise noted, we assume that y_t is strongly mixing and converges to a stationary distribution, as in Theorems 2 and 3.

Remark 2. All of the results in this section can be straightforwardly generalized to the case of multiple peer effects terms. I.e., to the model with K regressors $z_{t-1}^1, \ldots, z_{t-1}^K$

$$y_t = [\alpha + \beta y_{t-1} + \sum_{k=1}^{K} \gamma^k z_{t-1}^k + u_t]_+,$$

where coefficients α, β, $\{\gamma^k\}_{k=1}^K$ are unknown and have to be estimated. For example, for the above model, the analogue of the matrix M_R defined in Eq. (8) and used in Theorems 6 and 8 is a $K+2$ by $K+2$ matrix composed of all second moments of the vector $(1, y_t, z_t^1, \ldots, z_t^K)\mathbf{1}(\alpha + \beta y_t + \sum_{k=1}^K \gamma^k z_t^k \geq R)$.

4.1. $\mathbf{L_1}$ Estimation. For this and the following subsection we assume that errors have strictly positive density at zero, $f_u(0)$.

The least absolute deviations (LAD) estimator is the solution to the following minimization problem

$$\min_{a,b,c} \sum_{t=1}^{T} |y_t - [a + by_{t-1} + cz_{t-1}]_+|. \tag{7}$$

The LAD estimation procedure for the case of censored regression was first proposed by Powell (1984). He considers a cross-section model $y_t = [x_t'\beta + u_t]_+$. We cannot directly use his proofs, as they require u_t to be independent of all x_s, which does not hold for $s > t$ in an autoregressive model.

The minimization problem (7) is convex when $a > 0$, $b \geq 0$, $c \geq 0$, and $z \geq 0$. Thus, the numerical solution is a global maximum when the true $\alpha > 0$, $\beta \geq 0$, $\gamma \geq 0$. It turns out, as Theorems 5 and 7 show, that the LAD estimators are consistent and asymptotically normal.

Moreover, $[a + by_{t-1} + cz_{t-1}]_+ \equiv a + by_{t-1} + cz_{t-1}$ when a is positive and b, c and z_{t-1} are non-negative (y is non-negative by assumption). Thus, in this case the positive part in the minimization problem never binds, and the additional complicated condition on the stationary distribution and true parameter values from Powell (1984) does not arise in our case.

In what follow (y, z) denotes the distributional limit of (y_t, z_t) as $t \to \infty$.

Theorem 5. *When $\alpha > 0$, $\beta \geq 0$, $\gamma \geq 0$, the peer effect function is non-negative (i.e., $z_t \geq 0$ for all t), u_t has a continuous density at 0, and the random variables $1, y, z$ are linearly independent, the LAD estimator is consistent:*

$$\begin{pmatrix} \hat{\alpha}_{LAD} \\ \hat{\beta}_{LAD} \\ \hat{\gamma}_{LAD} \end{pmatrix} \xrightarrow{\mathbb{P}} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} \text{ as } T \to \infty.$$

Remark 3. The linear independence condition holds automatically whenever the peer effect function $p_{ij}(\cdot)$ depends on an argument other that y_{ijt-1} in a non-degenerate way. Indeed, in this case y_t contains only error u_{ijt}, while z_t also contains errors u_{klt} for $(k, l) \neq (i, j)$ (if z_t only depends on lagged peer effects, then it is independent of u_{ijt}). Thus, there is uncertainty, which cannot be removed by taking linear combinations.

If we do not impose positivity, we still can get consistency if the matrix M_R is nonsingular for some $R > 0$, where

$$M_R = \mathbb{E}\left[\begin{pmatrix} 1 & y & z \\ y & y^2 & yz \\ z & yz & z^2 \end{pmatrix} \mathbf{1}(\alpha + \beta y + \gamma z \geq R)\right], \tag{8}$$

and the random function $\Delta \mapsto \mathbf{1}(\alpha + \beta y + \gamma z + \Delta > 0)$ is continuous with probability one at the true parameter values and at $\Delta = 0$. When $\alpha = 0$ the latter condition indeed causes a problem, because $y = 0$ with positive probability. If z is also zero, then the indicator jumps depending on the sign of Δ. On the other hand if $\alpha > 0$, $\beta, \gamma \geq 0$ and $z > 0$ with probability 1, then this function is identical 1 and we return to the setting of Theorem 5.

The matrix M_R is singular if $\beta = \gamma = 0$, $\alpha < R$. In this case the indicator equals zero, so the matrix is identically zero. The other case is if $\beta = 0$, $\gamma > 0$ and z never takes values above $\frac{R-\alpha}{\gamma}$, so that again the matrix M_R is identically zero. When $\beta > 0$, the peer effect function $p_{ij}(\cdot)$ does not depend on y_{ijt}, and the random variable z is non-constant, the matrix is nonsingular. For instance, the triangular peer effect functions p_{ij} do not depend on y_{ijt}. Minimum and maximum functions, if taken over all edges except the given edge $(i \to j)$, also do not depend on y_{ijt}. Similarly linear functions satisfy this as long as the corresponding weight $\lambda_{ijt} = 0$. Thus, in all those examples M_R is non-singular for $\beta > 0$.

The idea is that if M_R is singular, then there exists a non-zero vector $(\lambda_1, \lambda_2, \lambda_3)$ such that $\lambda_1 \mathbf{1}(\alpha + \beta y + \gamma z \geq R) + \lambda_2 y \mathbf{1}(\alpha + \beta y + \gamma z \geq R) + \lambda_3 z \mathbf{1}(\alpha + \beta y + \gamma z \geq R) \equiv 0$. That is, when $\alpha + \beta y + \gamma z \geq R$, we must have $\lambda_1 + \lambda_2 y + \lambda_3 z = 0$. However, as $\beta > 0$ and z does not depend on y, we can perturb y a bit and get $y' = y + \varepsilon$, $\varepsilon > 0$, in which case the indicator is still non-zero, but the second equality fails unless $\lambda_2 = 0$. If $\lambda_2 = 0$, then we must have $z = -\lambda_1/\lambda_3$ whenever the indicator equals one. This again is impossible, when z is not a fixed constant.

Theorem 6. *Suppose that $(\alpha, \beta, \gamma) \in \Theta$, where Θ is some compact space in $\mathbb{R}^3$. When M_R is nonsingular for some $R > 0$ at the true parameter values and the random function $\Delta \mapsto \mathbf{1}(\alpha + \beta y + \gamma z + \Delta > 0)$ is continuous with probability one at the true parameter values and at $\Delta = 0$, the LAD estimator is consistent:*

$$\begin{pmatrix} \hat{\alpha}_{LAD} \\ \hat{\beta}_{LAD} \\ \hat{\gamma}_{LAD} \end{pmatrix} \xrightarrow[T\to\infty]{\mathbb{P}} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix}.$$

Remark 4. Note that if $R_1 > R_2$ and M_{R_1} is nonsingular, then M_{R_2} is also nonsingular. Similarly, if M_{R_2} is singular, then so is M_{R_1}. The reason is that if there exists a non-zero vector $(\lambda_1, \lambda_2, \lambda_3)$ such that $\lambda_1 + \lambda_2 y + \lambda_3 z = 0$ when $\alpha + \beta y + \gamma z \geq R_2$, then the same holds

for $\alpha + \beta y + \gamma z \geq R_1$, as $R_1 > R_2$. Thus, there is a bound $\bar{R} \in \mathbb{R} \cup \{+\infty\}$ such that M_R is nonsingular for any $R < \bar{R}$ and singular for any $R > \bar{R}$.

Remark 5. For consistency, the condition on the error density $f_u(0) > 0$ can be weakened to

$$\mathbb{P}(u \in [-\varepsilon, 0)) > 0,\ \mathbb{P}(u \in (0, \varepsilon]) > 0 \text{ for any } \varepsilon > 0.$$

That is, u_t must be in the left and right neighbourhoods of zero with positive probability. From the proof of Theorem 6, we only need $\min\left(\int_{-\tau}^{0}(\tau+u)dF_u(u), \int_{0}^{\tau}(\tau-u)dF_u(u)\right)$ to be positive for any $\tau > 0$, which is satisfied in this case.

On the other hand, for the following Theorems 7 and 8 such weakening of the condition $f_u(0) > 0$ leads to a change in the asymptotic theory, and we do not address it in the present paper.

Asymptotic normality in the positive case does not require any additional conditions, as is shown in Theorem 7. This is in line with consistency result for the positive case (Theorem 5).

Theorem 7. *When $\alpha > 0$, $\beta \geq 0$, $\gamma \geq 0$, the peer effect function is non-negative (i.e., $z_t \geq 0$ for all t), and u_t has a continuous density at 0, and the random variables $1, y, z$ are linearly independent, the LAD estimator is asymptotically normal:*

$$\sqrt{T}\begin{pmatrix} \hat{\alpha}_{LAD} - \alpha \\ \hat{\beta}_{LAD} - \beta \\ \hat{\gamma}_{LAD} - \gamma \end{pmatrix} \xrightarrow[T\to\infty]{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \frac{1}{4f_u^2(0)}\begin{pmatrix} 1 & \mathbb{E}y & \mathbb{E}z \\ \mathbb{E}y & \mathbb{E}y^2 & \mathbb{E}yz \\ \mathbb{E}z & \mathbb{E}yz & \mathbb{E}z^2 \end{pmatrix}^{-1}\right).$$

Similar to the general consistency result, we get asymptotic normality if the random function

$$\Delta \mapsto \mathbf{1}(\alpha + \beta y + \gamma z + \Delta > 0)$$

is continuous with probability 1 at the point corresponding to the true parameter values and $\Delta = 0$.

Theorem 8. *Suppose that $(\alpha, \beta, \gamma) \in \Theta$, where Θ is some compact space in $\mathbb{R}^3$. When M_0 is nonsingular at the true parameter values and the random function $\Delta \mapsto \mathbf{1}(\alpha+\beta y+\gamma z+\Delta > 0)$ is continuous with probability one at the true parameter values and at $\Delta = 0$, the LAD estimator is asymptotically normal:*

$$\sqrt{T}\begin{pmatrix} \hat{\alpha}_{LAD} - \alpha \\ \hat{\beta}_{LAD} - \beta \\ \hat{\gamma}_{LAD} - \gamma \end{pmatrix} \xrightarrow[T\to\infty]{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \frac{1}{4f_u^2(0)}M_0^{-1}\right).$$

Remark 6. Note that we do not impose a condition on the existence of nonsingular M_R for some $R > 0$ in Theorem 8. This is because non-degeneracy of M_0 together with continuity of the indicator $\mathbf{1}(\alpha + \beta y + \gamma z + \Delta > 0)$ implies the existence of $R > 0$, for which M_R is also nonsingular.

4.2. LAD in the explosive case. For the model without peer effects Theorem 3 provides a full classification of the asymptotic behavior of y_t. Using this theorem it is possible to establish consistency of the LAD estimator in the model without peer effects not only under a stationary distribution, but also for explosive and mean-explosive scenarios. This corresponds to cases when $\beta > 1$ or $\beta = 1$, $\alpha \geq 0$.

Theorem 9. *Suppose $\gamma = 0$ (no peer effects) and suppose that $(\alpha, \beta) \in \Theta$, where Θ is some compact space in $\mathbb{R}^2$. Then if u_t has a continuous density at 0, the LAD estimator is consistent for $\beta = 1$, $\alpha \geq 0$ and $\beta > 1$:*

$$\begin{pmatrix} \hat{\alpha}_{LAD} \\ \hat{\beta}_{LAD} \end{pmatrix} \xrightarrow[T\to\infty]{\mathbb{P}} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}.$$

We conjecture that consistency holds for any value of γ. Because the proof of Theorem 9 uses a separate argument depending on the asymptotic behavior of y_t, we cannot extend it to the model with peer effects. To be more precise, when $\gamma = 0$ the model has three types of explosive behavior: exponential growth regime ($\beta > 1$), linear growth regime ($\beta = 1, \alpha > 0$), and Brownian regime ($\beta = 1, \alpha = 0$). The process y_t under different regimes is shown in Figure 2. Yet, it is unclear what the analogue of these regimes is when $\gamma \neq 0$.

4.3. Discussion: LAD vs OLS. Interestingly, although the model is non-linear we still can treat it as linear to get consistent L_1 or LAD estimates. Yet, the same approach does not work with L_2 or OLS. The following example illustrates that OLS leads to inconsistency bias, while LAD does not.

Example 2. Suppose $y_t = [\alpha + u_t]_+$, $a \geq 0$, $\text{med}(u_t) = \mathbb{E}u_t = 0$. Then the LAD estimate solves

$$\min_{\hat{\alpha}} \sum_t |y_t - \hat{\alpha}|.$$

The solution to the minimization problem is the sample median, that is $\hat{\alpha}_{LAD} = \text{med}(y_1, \ldots, y_T)$. As $T \to \infty$ the sample median converges to the median of the stationary distribution of y_t. The median of y_t equals α, because with probability 0.5 the error u_t is positive, so that $y_t = \alpha + u_t \geq \alpha$, and with probability 0.5 the error u_t is negative, so that either $y_t = 0 \leq \alpha$ or $y_t = \alpha + u_t \leq \alpha$. Thus, with probability 0.5 $y_t \geq \alpha$ and with probability 0.5 $y_t \leq \alpha$. So $\hat{\alpha}_{LAD} \xrightarrow[T\to\infty]{\mathbb{P}} \text{med}(y) = \alpha$.

However, the results are different if we minimize an L_2 norm instead of an L_1 norm ignoring positive part. The solution to

$$\min_{\hat{\alpha}} \sum_t (y_t - \hat{\alpha})^2$$

is the sample mean, $\hat{\alpha}_{OLS} = \frac{\sum_t y_t}{T}$. As $T \to \infty$ the sample mean converges to the actual mean of y_t, so that

$$\hat{\alpha}_{OLS} \xrightarrow[T\to\infty]{\mathbb{P}} \mathbb{E}y_t = \mathbb{E}[\alpha + u_t]_+ = \int_{-\alpha}^{\infty} (\alpha + u) f_u(u) du = \alpha(1 - F(-\alpha)) + \int_{-\alpha}^{\infty} u f_u(u) du \neq \alpha.$$

The intuition is that median is more robust to truncation at zero: if the median of a process is positive, it does not matter if we replace negative values with zero and vice versa. Yet, the mean is significantly shifted by such procedure.

In the Supplementary Material we show how to correct the OLS procedure to get consistent estimates. The idea is to use "identification from infinity" approach similar to Chamberlain (1986).

5. Prediction

In this section we again treat equations for each edge separately. After the parameters of the model are estimated, one can do predictions. The model is Markovian (in a state space taking into account H time lags), thus, can be easily used for predictions.

To measure predictive power, we use a rolling window approach. That is, we choose some number $T' < T$, and estimate the model based on observations $t, \dots, t + T' - 1$. For each $t = 1, \dots, T - T'$ we calculate the difference between the predicted value $\hat{y}_{t+T'}$ and the actual value $y_{t+T'}$. Thus, we get a measure of how well we can predict the data:

$$R_{abs} = \frac{1}{T - T'} \sum_{t=1}^{T-T'} |y_{t+T'} - \hat{y}_{t+T'}|. \tag{9}$$

The smaller R_{abs} is, the better predictions we have on average. Similarly, we can also sum over all pairs (i, j) to get a prediction measure over the whole network.

There are two reasons why we use absolute deviations, i.e. the L_1 norm and not the more usual L_2 norm. First, as the estimation relies on minimizing the L_1 norm, it is more consistent to also use the same norm to evaluate predictions. Second, the optimal prediction in L_2 norm is $\int_{-\alpha-\beta y_{t-1}-\gamma z_{t-1}}^{\infty} (\alpha + \beta y_{t-1} + \gamma z_{t-1} + u) f_u(u) du$. Thus, it crucially depends on the distribution of the error term, f_u. However, optimal prediction in L_1 is $[\alpha + \beta y_{t-1} + \gamma z_{t-1}]_+$, as shown below. That is, it does not depend on the distribution of the error, and is more convenient to work with.

Remark 7. To see that optimal prediction in L_1 is $[\alpha + \beta y_{t-1} + \gamma z_{t-1}]_+$, define $\Delta\hat{y}_{t+1} = \hat{y}_{t+1} - \alpha - \beta y_t - \gamma z_t$ and write

$$\begin{aligned} \int |y_{t+1} - \hat{y}_{t+1}| f(u)du &= \int |[\alpha + \beta y_t + \gamma z_t + u]_+ - \hat{y}_{t+1}| f(u)du \\ &= \int |\max(u, -\alpha - \beta y_t - \gamma z_t) - \Delta\hat{y}_{t+1}| f(u)du. \end{aligned} \tag{10}$$

Because $\arg\min_C \mathbb{E}|v - C| = \text{med}(v)$, and

$$\max(u, -\alpha - \beta y_t - \gamma z_t) = \begin{cases} u, & u \geq -\alpha - \beta y_t - \gamma z_t, \\ -\alpha - \beta y_t - \gamma z_t, & u < -\alpha - \beta y_t - \gamma z_t, \end{cases},$$

$$\text{med}(\max(u, -\alpha - \beta y_t - \gamma z_t)) = \begin{cases} 0, & \alpha + \beta y_t + \gamma z_t \geq 0, \\ -\alpha - \beta y_t - \gamma z_t, & \alpha + \beta y_t + \gamma z_t < 0, \end{cases}$$

minimizing Eq. (10), gives

$$\Delta\hat{y}_{t+1} = \begin{cases} 0, & \alpha + \beta y_t + \gamma z_t \geq 0, \\ -\alpha - \beta y_t - \gamma z_t, & \alpha + \beta y_t + \gamma z_t < 0 \end{cases}$$

and $\hat{y}_{t+1} = [\alpha + \beta y_t + \gamma z_t]_+$.

6. Empirical Application

It is important to understand the process of the formation of an international trade network of various goods. A good forecast about the future amount of trade is often crucial for numerous policy decisions such as, for example, how much ships/planes/etc. to allocate to the transportation of a given good. The distinguishing feature of trade data is that it is associated with a lot of zeros, as not all countries trade with each other at each moment of time (see e.g. Table 1 in Dueñas, M. and Fagiolo (2013)). This motivates us to use trade data to analyze the performance of the techniques developed in the paper.

We apply our model to monthly exports of pharmaceutical products. The data is obtained from the Eurostat COMEXT database (European Commission (accessed May 1, 2019)). Pharmaceutical industry represents one of the largest industrial sector in the EU and provides a sizable, positive contribution to the EU trade balance (e.g. see Section II in Gambardella et al. (2000) for the discussion based on the Eurostat data). Pharmacutical industry has a two-digit code 30 as labeled by the Harmonized Commodity Description and Coding Systems (HS). The time period of observation is from January of 1999 until February of 2018, so that $T = 230$. We choose 12 European Union (EU) countries. Those are the countries which joined the EU first. Thus, each pair of countries represents an edge in the network, and we apply the techniques from previous sections to each such edge.

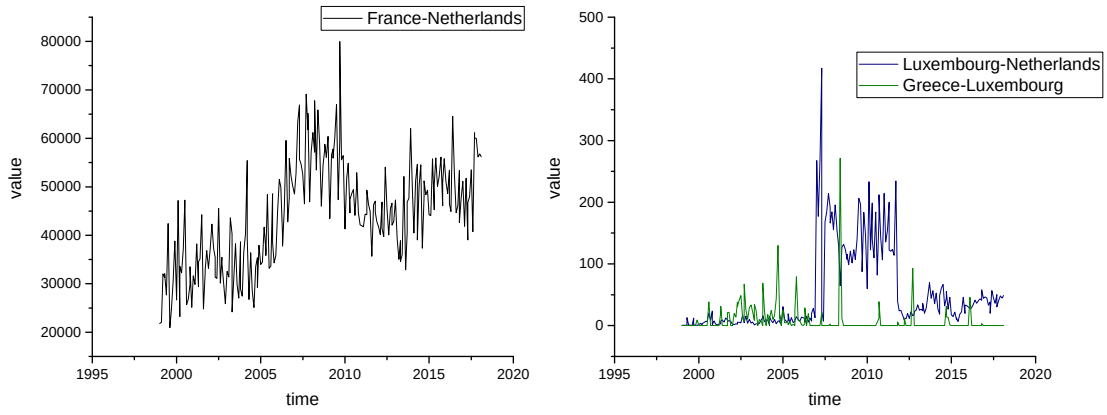

FIGURE 3. Exports of pharmaceutical products between EU countries. Value is in 1000 Euro.

Figure 3 shows exports of pharmaceutical products between France and Netherlands, Luxembourg and Netherlands, and Greece and Luxembourg. We can see that although the first graph does not have zeros, the second has periods of no trade for both pairs of countries. Zero trade is consistent with our model.

To measure predictive power we use a rolling window of size 200. Our consistency results rely on $T \to \infty$, thus, we want to have as large as possible window size. Additionally, we need more than one window to see how good are our predictions. This leads us to the choice of window size $T' = 200$, which is large enough and still can be rolled enough times[2]. The increment between successive rolling windows is 1 period. For each window we make a 1-period-ahead forecast. That is, based on observations from t to $t + 199$ we predict what happens at $t + 200$ for $t = 1, \dots, 30$. This gives us 30 overlapping windows and, thus, 30 forecasts. We report the average predictive absolute error over them. For testing in the next subsection we need more windows and with that aim in mind we also report results for a rolling window of size 115, which gives 115 overlappind windows (from t to $t + 114$ for $t = 1, \dots, 115$). The reason is that successive windows are very correlated, and to do testing we want to diminish the effect from correlation. Thus, we need more windows over which the test is going to average. In other words, there is a trade-off between maximizing the window size to improve the quality of estimates and maximizing the number of windows (hence, decreasing the size of each) to improve the statistical significance of testing. Without any additional information on this trade-off, we choose the values of two parameters to be equal, leading to window size $T' = 115$. It is impossible to efficiently present the results for all 132 edges individually, so instead we sum over all edges and report the total result.

The benchmark prediction is "today equals tomorrow", i.e., $\hat{y}_{t+1} = y_t$. Another option is to ignore positivity and treat the model as linear, i.e., $y_t = \alpha + \beta y_{t-1} + u_t$ (OLS in Tables

[2] $T' = 195$ or $T' = 205$ would be as good and the precise value of T is ad hoc in this sense.

Estimation method	R_{abs} in billion euros	
	window size $T' = 200$	window size $T' = 115$
LAD	1.9865	1.8977
MLE	2.0036	1.8847
OLS	2.0037	1.8848
"today"	2.0316	1.9096

TABLE 1. Prediction errors under different estimation techniques. Smaller numbers mean better quality of prediction. $R_{abs} = \frac{1}{T-T'} \sum_{i,j} \sum_{t=1}^{T-T'} |y_{ij,t+T'} - \hat{y}_{ij,t+T'}|$, where $y_{ij,t+T'}$ is the true value and $\hat{y}_{ij,t+T'}$ is the prediction.

Model and method	R_{abs} in billion euros	
	window size $T' = 200$	window size $T' = 115$
LAD w. p.e.	1.8934	1.8315
MLE w. p.e.	1.9215	1.8274
OLS w. p.e.	1.9217	1.8279
LAD w/o p.e.	1.9865	1.8977
MLE w/o p.e.	2.0036	1.8847
OLS w/o p.e.	2.0037	1.8848
"today"	2.0316	1.9096

TABLE 2. Prediction errors with and without peer effects under different estimation techniques. Smaller numbers mean better quality of prediction. $R_{abs} = \frac{1}{T-T'} \sum_{i,j} \sum_{t=1}^{T-T'} |y_{ij,t+T'} - \hat{y}_{ij,t+T'}|$, where $y_{ij,t+T'}$ is the true value and $\hat{y}_{ij,t+T'}$ is the prediction.

1 and 2). We compare these two approaches with the LAD estimator without peer effects. Results are shown in Table 1. We can see that for the larger rolling window model-based predictors outperform both alternatives. Moreover, the LAD estimate outperforms the MLE. Same pattern remains present if we add peer effects, as shown in Table 2. However, for the smaller window size 115 all estimates lead to poor results. This suggests that the available T is at the border of applicability of our methods. Yet, in the next paragraph we will see that the addition of peer effects improves the results for small T as well. The fact that LAD performs significantly better than the MLE for the larger window size suggests that the error distribution may be far from normal.

We use the triangular peer effect function: $z_{ijt-1} = p_{ij}\left(\{y_{k\ell t-1}\}_{k,\ell}\right) = \sum_k \frac{\sqrt{y_{ikt-1}y_{kjt-1}}}{n-2}$. Table 2 shows that adding peer effects reduces the prediction error both under LAD and MLE estimation approaches. Same happens if we ignore positivity, assume linear model, and estimate it by OLS. This suggests the presence of peer effects in the data. The scatter

Estimated model	R_{abs} in billion euros	
	window size $T' = 200$	window size $T' = 115$
5 lags	1.6731	1.6650
4 lags+p.e. at $t-4$	1.6733	1.6437
4 lags+p.e. at $t-1$	1.6797	1.6511
1 lag+p.e. at $t-4$	1.8569	1.8172
1 lag+p.e. at $t-1$	1.8934	1.8315
1 lag	1.9865	1.8977
"today"	2.0316	1.9096

TABLE 3. Prediction errors with different lags and peer effects. Smaller numbers mean better quality of prediction. $R_{abs} = \frac{1}{T-T'} \sum_{i,j} \sum_{t=1}^{T-T'} |y_{ij,t+T'} - \hat{y}_{ij,t+T'}|$, where $y_{ij,t+T'}$ is the true value and $\hat{y}_{ij,t+T'}$ is the prediction.

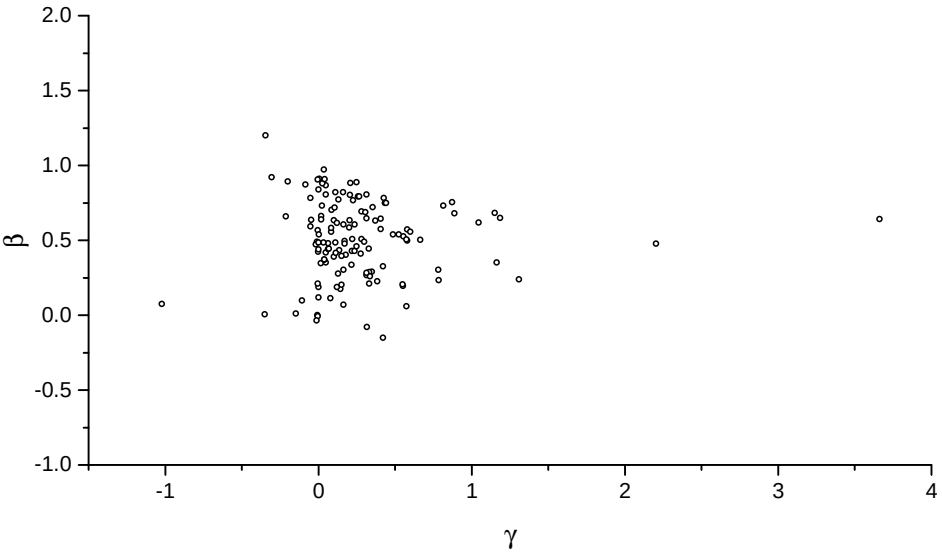

FIGURE 4. LAD estimates for 132 country pairs for the model

$$y_t = [\alpha + \beta y_{t-1} + \gamma z_{t-1} + u_t]_+.$$

plot of the LAD estimates $(\beta_{ij}, \gamma_{ij})$ is shown in Figure 4. The mean value of β is 0.5037 and the mean value of γ is 0.2652.

In general, including irrelevant regressors in the model leads to more noise when one does prediction. Yet, one can expect that the use of additional relevant regressors may improve the prediction results reported in Table 2. It turns out, as Table 3 shows, that using four lags and one peer effect regressor or using five lags gives the best results ($z_{ijt-1} = \sum_k \frac{\sqrt{y_{ikt-4} y_{kjt-4}}}{n-2}$). Also using peer effects evaluated at $t-4$ leads to better performance than using peer effects

	Number of rejections	
	5%	10%
"today" vs. 1 lag	38	47
"today" vs. 1 lag+p.e. at $t-1$	45	59
"today" vs. 1 lag+p.e. at $t-4$	53	60
1 lag vs. 1 lag+p.e. at $t-1$	39	51
1 lag vs. 1 lag+p.e. at $t-4$	49	60
5 lags vs. 4 lags+p.e. at $t-1$	16	22
5 lags vs. 4 lags+p.e. at $t-4$	15	27

TABLE 4. Two-sided Diebold-Mariano test of equal predictability.

evaluated at $t-1$. This suggests that peer effects can help to predict the future, though the optimal functional form of the peer effect function is unclear.

6.1. Tests. When we were analyzing the performance of our techniwues and comparing predictive errors (Tables 1, 2, and 3), we averaged over all edges of the networks. In the current subsection, we look at how good the prediction is for each given edge. For each edge we compute the Diebold-Mariano test (Diebold and Mariano (1995)) to compare the predictive accuracy under different model specifications. Here we use the rolling window of size 115, as the test requires a large number of predictions. This gives us 115 overlapping windows (from t to $t+114$ for $t=1,\dots,115$). We consider 1-period-ahead forecasts under various models. We compare the model without peer effects against the benchmark prediction "today=tomorrow" and against the model with peer effects. There are 132 edges in total, and generally for between one third and one half of them we reject the null of identical predictive accuracy. This is shown in Table 4. Similarly, Table 5 shows the results of the one-sided Diebold-Mariano tests for different model specifications. Almost all of the rejections in the former table correspond to rejections in the latter table. That is, when we compare two possible model specifications, the one with the positive part and peer effects dominates. This strengthens the importance of incorporating positivity and peer effects. On the other hand, testing the model with 5 lags versus the model with peer effects seems to show much less difference in predictability.

Additionally, for the model with 1 lag and peer effects evaluated at time $t-1$ we calculate t-statistics to analyze significance of the peer effects for each edge. That is, for each edge $(i \to j)$ we calculate $\frac{\sqrt{T}\hat{\gamma}_{ij}}{\hat{std}(\gamma_{ij})}$. We use asymptotic variance from Theorem 8 to estimate $\hat{std}(\gamma_{ij})$. The matrix of second moments M_0 is estimated by its sample analogue, while density at zero, $f_{u_{ij}}(0)$, is estimated by uniform kernel. The bandwidth is chosen so that approximately half of the observations are used. On 5% level peer effects are significant for approximately two thirds of edges. If we instead consider peer effects evaluated at time $t-4$, the number of

	Number of rejections	
	5%	10%
"today" worse than 1 lag	41	51
"today" worse than 1 lag+p.e. at $t-1$	52	62
"today" worse than 1 lag+p.e. at $t-4$	57	65
1 lag worse than 1 lag+p.e. at $t-1$	38	46
1 lag worse than 1 lag+p.e. at $t-4$	50	57
5 lags worse than 4 lags+p.e. at $t-1$	11	21
5 lags worse than 4 lags+p.e. at $t-4$	21	23

TABLE 5. One-sided Diebold-Mariano test.

significant coefficients goes slightly up (92 out of 132 versus 83 out of 132). This reinforces results of the Diebold-Mariano tests reported in Table 5.

7. Extensions

In Section 2 we specified the model in terms of evolution of edges of a network. The specification allows for both directed and directed networks. Besides, our framework is also suitable for non-negative panels where n is fixed and T is large and we expect some interactions across units. E.g. alcohol consumption by classmates or behavior of various financial markets. In this case we can view cross-section units (classmates or financial markets) as vertices in a network and analyze their evolution. If $y_{iit} \equiv y_{it}$ represents the evolution of a characteristic of agent i, such as amount of alcohol consumed by agent i at day t, we can use our techniques to estimate and predict future y_{it}s. That is, we can analyze how the alcohol consumption of one's peers affects one decision to drink or how a crush in one markets affects the behavior of other markets. Mathematically, this corresponds to the non-negative multivariate time series $y_{it_{i,t}}$:

$$y_{it} = \left[\alpha_i + \beta_i y_{it-1} + \gamma_i p_i \left(\left\{ y_{js} \right\}_{\substack{\text{j=1,...,n} \\ \text{s=t-H,...,t-1}}} \right) + u_{it} \right]_+, \tag{11}$$

where, as before, $p_i(\cdot)$ is a peer-effect function. When one analyzes financial markets, we may expect that a crush in one of them pushes the others also to go down. Thus, a peer effect function of the form $\min_{i,j} y_{ijt}$ may be useful.

Furthermore, our setting allows for joint analysis of the evolution of edges and vertices. If we consider both Eq. (3) and Eq. 11 jointly, then we capture both y_{ijt}, $i \neq j$ and $y_{iit} \equiv y_{it}$. Moreover, peer effect functions p_{ij} and p_i can depend both on y_{kls} and y_{kks} with suitable modification of the Assumption 5. In this extended setting Theorems 1, 2, 5, 6, 7, 8, and 9 continue to hold.

8. Conclusion

This paper presents a novel approach to modeling and estimating networks. Estimation does not require knowledge of the error distribution, thereby making the whole process more attractive to use. Instead of using the variation across individuals, one can use variation across time to identify parameters of the network. In this approach, we treat networks as multivariate time series. The main advantage is that we allow the realization of each edge today to depend on the whole structure of the graph in the previous time period, and not only on the properties of two nodes, which are connected by that edge. Moreover, the Markov form of the equations makes them convenient for doing predictions. As the empirical example suggests, our model does, indeed, help to predict the future. Overall, the results confirm that incorporating non-negativity of the dependent variables into the model matters and incorporating peer effects leads to the improved predictive power.

In the future it would be interesting to apply the model to different data sets. Phone call data for a small group of individuals and technology adoption by countries from one another seem like natural candidates.

From a theoretical point of view, it would be interesting to investigate in more depth the discontinuity in the asymptotics in the model without peer effects. The behaviour of the process $y_t = [y_{t-1} + u_t]_+$ differs dramatically from what one gets by shifting α from zero or β from one slightly. Thus, finding a way to unify the cases in the neighbourhood of the point $(\alpha = 0, \beta = 1)$ in the spirit of Phillips (1987) may be helpful from a practical point of view. Yet, it is a challenging problem. The proper scaling limit of such a process is complicated, because it involves the computation of the time the process spends at zero in the limit.

Appendix A. Stationarity/Explosiveness.

This subsection presents proofs on stationary/explosive behavior of the process y_{ijt} as $t \to \infty$.

Proof of Theorem 1. To simplify notation, let us denote $z_{ijt-1} = p\left(\left\{y_{kls}\right\}_{\substack{(k,l)\neq(i,j)\\ s=t-H,\ldots,t-1}}\right)$. Then

$$
\begin{aligned}
y_{ijt} &= [\alpha_{ij} + \beta_{ij} y_{ijt-1} + \gamma_{ij} z_{ijt-1} + u_{ijt}]_+ \\
&\leq \max(0, \beta_{ij}) y_{ijt-1} + |\gamma_{ij} z_{ijt-1}| + \max(0, \alpha_{ij} + u_{ijt}) \\
&\leq (\max(0, \beta_{ij}) + |\gamma_{ij}|) \max_{k,l,s=t-H,\ldots,t-1} y_{kls} + |\mathcal{A}| + \max(0, \alpha_{ij} + u_{ijt}) \\
&\leq C \max_{k,l,s=t-H,\ldots,t-1} y_{kls} + |\mathcal{A}| + \max(0, \alpha_{ij} + u_{ijt}).
\end{aligned} \tag{12}
$$

Denote $v_t = |\mathcal{A}| + \max\limits_{k,l,s=t-H,\ldots,t-1} \max(0, \alpha_{kl} + u_{kls})$. Then from Eq. (12) we get

$$\max_{i,j} y_{ijt} \leq C \max_{k,l,s=t-H,\ldots,t-1} y_{kls} + v_t. \tag{13}$$

We are going to show that

$$\max_{s=mH+1,\ldots,mH+H}^{i,j} y_{ijs} \leq C \max_{s=(m-1)H+1,\ldots,mH}^{i,j} y_{ijs} + w_m,$$

where $w_m = v_{mH+1} + \ldots + v_{mH+H}$.

By Eq. (13),

$$\max_{i,j} y_{i,j,mH+1} \leq C \max_{i,j;s=(m-1)H+1,\ldots,mH} y_{ijs} + v_{mH+1}.$$

Applying Eq. (13) twice (for $t = mH + 2$ and $t = mH + 1$) we get

$$\begin{aligned}
\max_{i,j} y_{i,j,mH+2} &\leq C \max_{i,j;s=(m-1)H+2,\ldots,mH+1} y_{ijs} + v_{mH+2} \\
&\leq C \max(C \max_{i,j;s=(m-1)H+1,\ldots,mH} y_{ijs} + v_{mH+1}, \max_{i,j;s=(m-1)H+2,\ldots,mH} y_{ijs}) + v_{mH+2} \\
&\leq C \max_{i,j;s=(m-1)H+1,\ldots,mH} y_{ijs} + v_{mH+1} + v_{mH+2},
\end{aligned}$$

because $C < 1$ and $v_t \geq 0$.

We can redo the same for $t = mH + 3, \ldots, mH + H$, so that

$$\max_{i,j} y_{i,j,mH+r} \leq C \max_{i,j;s=(m-1)H+1,\ldots,mH} y_{ijs} + v_{mH+1} + \ldots + v_{mH+r}.$$

Thus,

$$\max_{s=mH+1,\ldots,mH+H}^{i,j} y_{ijs} \leq C \max_{s=(m-1)H+1,\ldots,mH}^{i,j} y_{ijs} + w_m,$$

where $w_m = v_{mH+1} + \ldots + v_{mH+H} \geq 0$.

Iterative back-substitution leads to

$$\begin{aligned}
&\max_{s=mH+1,\ldots,mH+H}^{i,j} y_{ijs} \leq C \max_{s=(m-1)H+1,\ldots,mH}^{i,j} y_{ijs} + w_m \\
&\leq w_m + C w_{m-1} + C^2 \max_{s=(m-2)H+1,\ldots,(m-1)H}^{i,j} y_{ijs} \leq \sum_{s=0}^{m-1} C^s w_{m-s} + C^m \max_{s=1,\ldots,H}^{i,j} y_{ijs}.
\end{aligned} \tag{14}$$

Note that $\mathbb{E}w = H\mathbb{E}v < C_1 < \infty$, as H and number of agents n are finite. Then from Eq. (14)

$$\mathbb{E}y_{ijt} \leq \mathbb{E} \max_{s=H\lfloor \frac{t-1}{H} \rfloor+1,\ldots,H\lfloor \frac{t-1}{H} \rfloor+H}^{k,l} y_{kls} \leq \frac{\mathbb{E}w}{1-C} + const < C_2 < \infty. \quad \square$$

In the following set of theorems we show sufficient condition for stationarity for the general model (Theorem 2) and the full characterization of stationary/explosive limiting behavior of y_t depending on α, β (Theorem 3) for the model without peer effects.

In Section 3.1 in Supplementary material we show that

- If Assumptions 1, 2, and 5 are satisfied, $\mathbb{E}u_{ijt}^4 < \infty$ for all i, j, t, and for all i, j $\max(0, \beta_{ij}) + |\gamma_{ij}| < C < 1$, then $\mathbb{E}$(time until graph is empty for H periods) is finite. That is, the expected time until $y_{ijt} = \ldots = y_{ij,t+H-1} = 0$ for all i, j is finite.
- If Assumptions 1 and 2 are satisfied for the model without γ, $\mathbb{E}u_t^4 < \infty$, and $\beta < 1$ or $\alpha < 0$, $\beta = 1$, then $\mathbb{E}$(length until zero) is finite.

Theorem A.1. *Suppose that Assumptions 1, 2, and 5 are satisfied and $\mathbb{E}u_{ijt}^4 < \infty$ for all i, j, t. For the model without γ, if $\beta < 1$ or $\beta = 1, \alpha < 0$, then y_t is strongly mixing and converges to a stationary distribution. For the model with γ, if for some C $\max(0, \beta_{ij}) + |\gamma_{ij}| < C < 1$ for all i, j, then y_{ijt} is strongly mixing and converges to a stationary distribution*[3]*.*

Proof. Let us first show convergence to a stationary distribution. The proof follows the lines of section $XI.8$ in Feller (2008). Let us show the proof for the model without peer effects.

From Lemmas $S6$ and $S8$ in the Supplementary Material, we know that $\mathbb{E}$(length until zero) is finite. Let us denote the expected time until zero by μ. Thus, with probability one the process reaches zero. The continuation of a process after it reaches zero is a probabilistic replica of the whole process started from the previous zero.

For any Borel set Δ denote by $P_\Delta(t)$ the probability that $y_{t+s} \in \Delta$ given that s is the (finite) time before the process first hits zero. The process y_t is by definition strongly Markov, so that such probability does not depend on s.

We are going to show that for any Borel set Δ there exists $\lim_{t\to\infty} P_\Delta(t) = P_\Delta$ such that $P_\Delta \geq 0$, $P_{\mathbb{R}_+} = 1$, and P_Δ is countably additive. This would imply that the one-point distribution of y_t converges to a limit at $t \to \infty$; by the Markov property the latter further implies the desired convergence of all finite-dimensional distributions of $\{y_{t+\tau}\}_{\tau\in\mathbb{Z}}$.

Define by S_1 the first time of hitting zero, by S_2 the second time of hitting zero, etc. Also define $q_\Delta(t) = \mathbb{P}(S_1 > t, y_t \in \Delta)$. Then

$$q_\Delta(t) + q_{\mathbb{R}_+\setminus\Delta}(t) = 1 - F(t),$$

where F is a distribution of time between two consequent moments of hitting zero $(S_{n+1} - S_n)$.

Because the probability that $y_{t+s} \in \Delta$ given $S_1 = s$ does not depend on s, we can write

$$P_\Delta(t) = \mathbb{P}(S_1 > t,\, y_t \in \Delta) + \mathbb{P}(S_1 \leq t,\, y_t \in \Delta) = q_\Delta(t) + \int_0^t P_\Delta(t - y)F(dy).$$

The function $q_\Delta(t)$ is directly integrable since it is dominated by the monotone integrable function $1 - F$. Therefore by the renewal theorem $\lim_{t\to\infty} P_\Delta(t) = \frac{1}{\mu}\sum_{t=0}^{\infty} q_\Delta(t) \geq 0$.

[3]Formally this means that the finite-dimensional distributions of the process $\{y_{t+\tau}\}_{\tau\in\mathbb{Z}}$, converge to those of a stationary in τ process as $t \to \infty$.

By definition, $q_{\mathbb{R}_+}(t) = 1 - F(t)$ so that $\lim\limits_{t\to\infty} P_{\mathbb{R}_+}(t) = \frac{1}{\mu}\sum\limits_{t=0}^{\infty}(1-F(t)) = \frac{\mu}{\mu} = 1$. Similarly, for any Δ, $q_\Delta(t) \le 1 - F(t)$, and, thus, $\lim\limits_{t\to\infty} P_\Delta(t) \le 1$. Finally, we need to check that P_Δ is countably additive. This follows from the fact that for a countable number of pairwise disjoint sets Δ_i,

$$q_{\cup_i\Delta_i}(t) = \mathbb{P}(S_1 > t, y_t \in \cup_i\Delta_i) = \sum_i q_{\Delta_i}(t).$$

Thus, $\lim\limits_{t\to\infty} P_{\cup\Delta_i}(t) = \frac{1}{\mu}\sum\limits_{t=0}^{\infty} q_{\cup\Delta_i}(t) = \frac{1}{\mu}\sum\limits_{t=0}^{\infty}\sum\limits_i q_{\Delta_i}(t) \ge 0 = \sum\limits_i \lim\limits_{t\to\infty} P_{\Delta_i}(t)$.

The proof for the model with peer effects is the same, with the only difference that now y_t is a vector $\{y_{ijt}\}_{i,j}$. Such process is strongly Markov in extended space, where the element is a vector $\{y_{ij,t-1}, \dots, y_{ij,t-H}\}_{i,j}$. (That is, we divide the time scale into blocks of length H.) Moreover, by Lemma $S6$ in Supplementary Material $\mathbb{E}$(length until zero) is finite.

Let us now show strong mixing. All bounds on $\mathbb{E}$(length until zero) were uniform with respect to the choice of starting point y_0 (i.e. if $y_0 \in [0, M]$, then there exists a constant such that $\mathbb{E}$(length until zero) $<$ *const*). (See Eq. 25 and Eq. 27 in Supplementary Material.) Thus, because in the proof above $\lim\limits_{t\to\infty} P_{\cup\Delta_i}(t)$ is uniform over the choice of y_0 in a bounded set

$$\lim_{t\to\infty} |\mathbb{P}(y_s \in \Delta_1, y_{t+s} \in \Delta_2) - \mathbb{P}(y_s \in \Delta_1)\mathbb{P}(y_s \in \Delta_1, y_{t+s} \in \Delta_2)| = 0. \quad \square$$

Theorem A.2. *If $\beta = 1$, $\alpha > 0$ or $\beta > 1$, then y_t is divergent ($y_t \xrightarrow[t\to\infty]{a.s.} \infty$).*

Proof. We need to show that $\forall M \ge 0\ \mathbb{P}(\lim\limits_{t\to\infty} y_t < M) = 0$.

Assume first that $\alpha > 0$ and note, that because taking positive part of a random variable can only increase it, we get

$$y_t = [\alpha + \beta y_{t-1} + u_t]_+ \ge \alpha + \beta y_{t-1} + u_t \ge \alpha + y_{t-1} + u_t.$$

Thus, if $\alpha > 0$, then $y_t \ge \alpha t + y_0 + \sum\limits_{s=1}^{t} u_s$, and

$$\text{If } y_t < M \text{ then } \alpha t + y_0 + \sum_{s=1}^{t} u_s < M \text{ and } \frac{1}{t}\sum_{s=1}^{t} u_s < \frac{M - y_0}{t} - \alpha.$$

By the strong law of large numbers, $\frac{1}{t}\sum\limits_{s=1}^{t} u_s \xrightarrow[t\to\infty]{a.s.} \mathbb{E}u_t = 0$. Therefore, $\mathbb{P}\left(\lim\limits_{t\to\infty} \frac{1}{t}\sum\limits_{s=1}^{t} u_s = 0\right) = 1$ and $\forall\varepsilon > 0\ \mathbb{P}\left(\lim\limits_{t} \frac{1}{t}\sum\limits_{s=1}^{t} u_s < -\varepsilon\right) = 0$.

Fix $\varepsilon = \frac{2\alpha}{3}$ and T' such that $\frac{M-y_0}{T'} < \frac{\alpha}{3}$. Then

$$\mathbb{P}\left(\lim_{t\to\infty} y_t < M\right) \le \mathbb{P}\left(\lim_{t\to\infty} \frac{1}{t}\sum_{s=1}^{t} u_s < \frac{M-y_0}{t} - \alpha\right) \le \mathbb{P}\left(\lim_{t\to\infty} \frac{1}{t}\sum_{s=1}^{t} u_s < -2\alpha/3\right) = 0.$$

So that $\mathbb{P}\left(\lim_{t\to\infty} y_t < M\right) \xrightarrow[t\to\infty]{} 0$.

Now suppose that $\alpha \leq 0$ and $\beta > 1$. We first will show convergence in probability. We are going to show that

$$\forall M \quad \mathbb{P}(y_t \in [M, M+1]) \xrightarrow[t\to\infty]{} 0.$$

Then $\mathbb{P}(|y_t| < M) \to 0$. As $y_t \geq 0$, we only need to consider intervals in $\mathbb{R}_+$.

Consider events $S_t(M) = \begin{cases} y_t \in [M, M+1] \\ y_{t+s} > M+1 \ \forall s \geq 1. \end{cases}$ Such events are disjoint for $t \neq t'$ ($S_t \cap S_{t'} = \emptyset$ when $t \neq t'$). Thus,

$$\sum_{t=1}^{\infty} \mathbb{P}(S_t) \leq 1. \tag{15}$$

Choose $A > 1$ such that

$$\left(\mathbb{E}u^4 + \left(\mathbb{E}u^2\right)^2\right) \sum_{s=1}^{\infty} \frac{3s^2}{(As-1)^4} < \frac{1}{2}. \tag{16}$$

Also choose $\varepsilon > 0$ such that $\beta - \varepsilon > 1$ and M' such that $\varepsilon M' > |\alpha| + A$. Thus, conditional on being in $S_t(M')$, $\forall s \geq 1$

(17)

$$y_{t+s} = [\alpha + \beta y_{t-1} + u_t]_+ \geq \alpha + \beta y_{t+s-1} + u_{t+s} \geq A + (\beta - \varepsilon) y_{t+s-1} + u_{t+s} \geq As + y_t + \sum_{r=1}^{s} u_{t+r}.$$

If $y_t \geq M$, then from Eq. (17), if $A + u_{t+1} > 1$, we get $y_{t+1} > M+1$. If also $2A + u_{t+1} + u_{t+2} > 1$, then similarly $y_{t+2} > M+1$. Thus, if $Ap + \sum_{r=1}^{p} u_{t+r} > 1$ for all $p = 1, \ldots, s$ we get $y_{t+s} > M+1$. Let us calculate the probability that $Ap + \sum_{r=1}^{p} u_{t+r} > 1$ for all $p \geq 1$.

$$\mathbb{P}\left(As + \sum_{r=1}^{s} u_{t+r} \leq 1\right) = \mathbb{P}\left(\sum_{r=1}^{s} u_{t+r} \leq 1 - As\right) \leq \mathbb{P}\left(\left|\sum_{r=1}^{s} u_{t+r}\right| \leq As - 1\right)$$

$$\leq \frac{\mathbb{E}\left(\sum_{r=1}^{s} u_{t+r}\right)^4}{(As-1)^4} = \frac{s\mathbb{E}u^4 + 3s(s-1)\left(\mathbb{E}u^2\right)^2}{(As-1)^4} \leq \frac{3s^2\left(\mathbb{E}u^4 + \left(\mathbb{E}u^2\right)^2\right)}{(As-1)^4},$$

where we used the Markov inequality for $\left|\sum_{r=1}^{s} u_{t+r}\right|^4$ to bound probability by expectation.

Therefore,

$$\mathbb{P}\left(Ap+\sum_{r=1}^{p}u_{t+r}>1\ \ \forall p\geq 1\right)\geq 1-\sum_{s=1}^{\infty}\mathbb{P}\left(As+\sum_{r=1}^{s}u_{t+r}\leq 1\right)$$
$$\geq 1-\sum_{s=1}^{\infty}\frac{3s^2\left(\mathbb{E}u^4+\left(\mathbb{E}u^2\right)^2\right)}{(As-1)^4}\geq 1-\frac{1}{2}=\frac{1}{2}. \tag{18}$$

Thus, for $M\geq M'$

$$\mathbb{P}(S_t(M))\geq \mathbb{P}\left(y_t\in[M,M+1],\ Ap+\sum_{r=1}^{p}u_{t+r}>1\ \ \forall p\geq 1\right)$$
$$=\mathbb{P}\left(y_t\in[M,M+1]\right)\mathbb{P}\left(Ap+\sum_{r=1}^{p}u_{t+r}>1\ \ \forall p\geq 1\right)\geq 0.5\mathbb{P}\left(y_t\in[M,M+1]\right). \tag{19}$$

Plugging Eq. (19) into Eq. (15), we get $\sum_{t=1}^{\infty}\mathbb{P}\left(y_t\in[M,M+1]\right)\leq 2$. Thus, the series converges, so it must be that $\mathbb{P}\left(y_t\in[M,M+1]\right)\xrightarrow[t\to\infty]{}0$.

We have shown that for $M\geq M'$, $\mathbb{P}\left(y_t\in[M,M+1]\right)\xrightarrow[t\to\infty]{}0$. Let us show that this also holds for $M<M'$.

We know that for all $M\geq M'$, $\lim_t\mathbb{P}\left(y_t\in[M,M+1]\right)=0$. Thus, also $\lim_t\mathbb{P}\left(y_t\in[M',\beta M'+1]\right)=0$. Suppose by contradiction that

$$\lim_t\mathbb{P}\left(y_t\in[\beta^{-1}M',M']\right)>0.$$

But then with positive probability $u_{t+1}\in[-\alpha,-\alpha+1]$ and with positive probability $y_{t+1}=[\alpha+\beta y_t+u_{t+1}]_+\in[M',\beta M'+1]$, so that we get a contradiction. Thus, $\lim_t\mathbb{P}\left(y_t\in[\beta^{-1}M',M']\right)=0$. We can repeat the argument with $\beta^{-1}M'$ instead of M', then with $\beta^{-2}M'$ instead of $\beta^{-1}M'$ and so on. Thus, we get that $\mathbb{P}\left(y_t\in(0,M]\right)\xrightarrow[t\to\infty]{}0$ for all $M>0$ (as $\beta^{-k}\to 0$). If $\lim_t\mathbb{P}\left(y_t=0\right)>0$, then by a similar argument we must have that $\lim_t\mathbb{P}\left(y_t\in[1,2]\right)>0$, which is a contradiction. (Take $u_{t+1}\in[-\alpha+1,-\alpha+2]$.)

Therefore, $\mathbb{P}\left(y_t\in[M,M+1]\right)\xrightarrow[t\to\infty]{}0$ for all M and $y_t\xrightarrow{\mathbb{P}}\infty$.

Now let us show that $y_t\xrightarrow{a.s.}\infty$. Note that we can choose $A(k)$ in Eq. (16) such that

$$\left(\mathbb{E}u^4+\left(\mathbb{E}u^2\right)^2\right)\sum_{s=1}^{\infty}\frac{3s^2}{(A(k)s-1)^4}<\frac{1}{k}$$

and $M(k)$ such that $\varepsilon M(k)>|\alpha|+A(k)$. Then if $y_t>M(k)$, for Eq. (18) with A replaced by $A(k)$, we get that $y_{t+s}>M(k)$ for all $s\geq 1$ with probability at least $1-\frac{1}{k}$.

Because $y_t \xrightarrow{\mathbb{P}} \infty$, for any M $\mathbb{P}(y_t > M) \xrightarrow[t\to\infty]{} 1$. Thus, for any M and $\delta > 0$ there exists T such that $\mathbb{P}(y_T > M) > 1-\delta$. Therefore, if $M > M(k)$, $\lim\limits_t y_t > M$ with probability of at least $(1-\delta)(1-1/k)$. Because δ is arbitrary, we must have $\mathbb{P}(\lim\limits_t y_t > M) \geq 1-1/k$ for any $M > M(k)$. Because k can be chosen arbitrary, we must have $\mathbb{P}(\lim\limits_t y_t > M) = 1$ for any M. (Note that if $M' > M''$, then $\mathbb{P}(\lim\limits_t y_t > M'') \geq \mathbb{P}(\lim\limits_t y_t > M')$). Thus, $y_t \xrightarrow{a.s.} \infty$. □

Theorem A.3. *If $\beta = 1$, $\alpha = 0$, then y_t is mean-divergent ($\mathbb{E}y_t \xrightarrow[t\to\infty]{} \infty$).*

Proof. Because u_t has full support, with positive probability $u_t < -y_{t-1}$. Thus,

$$\mathbb{E}y_t = \mathbb{E}[y_{t-1} + u_t]_+ > \mathbb{E}(y_{t-1} + u_t) = \mathbb{E}y_{t-1},$$

and $\mathbb{E}y_t$ is a strictly increasing sequence of t. Therefore, either $\mathbb{E}y_t \to \infty$ or $\mathbb{E}y_t \to const$. Suppose that the latter is true. Then by Markov's inequality for any $C > 0$,

$$\mathbb{P}(y_t \geq C) \leq \frac{\mathbb{E}y_t}{C} \leq \frac{\lim\limits_t \mathbb{E}y_t}{C}.$$

Let us choose C such that $\frac{\lim\limits_t \mathbb{E}y_t}{C} \leq \frac{1}{2}$. Thus, for any t, $\mathbb{P}(y_t \geq C) \leq \frac{1}{2}$ and $\mathbb{P}(y_t < C) \geq \frac{1}{2}$.

$$\mathbb{E}y_t = \mathbb{E}[y_{t-1} + u_t]_+(\mathbf{1}\{y_{t-1} \geq C\} + \mathbf{1}\{y_{t-1} < C\}), \tag{20}$$

$$\mathbb{E}[y_{t-1} + u_t]_+\mathbf{1}\{y_{t-1} \geq C\} \geq \mathbb{E}(y_{t-1} + u_t)\mathbf{1}\{y_{t-1} \geq C\} = \mathbb{E}y_{t-1}\mathbf{1}\{y_{t-1} \geq C\}, \tag{21}$$

(22)

$$\mathbb{E}[y_{t-1} + u_t]_+\mathbf{1}\{y_{t-1} < C\} = \mathbb{E}[y_{t-1} + \max(-C, u_t)]_+\mathbf{1}\{y_{t-1} < C\}$$

$$\geq \mathbb{E}(y_{t-1} + \max(-C, u_t))\mathbf{1}\{y_{t-1} < C\} = \mathbb{E}y_{t-1}\mathbf{1}\{y_{t-1} < C\} + \mathbb{P}(y_{t-1} < C)\mathbb{E}\max(-C, u_t).$$

Because $\mathbb{P}(y_t < C) \geq \frac{1}{2}$ and $\mathbb{E}\max(-C, u_t) > \mathbb{E}u_t = 0$, combining Eq. (20), (21), and (22), we get

$$\mathbb{E}y_t \geq \mathbb{E}y_{t-1}\mathbf{1}\{y_{t-1} \geq C\} + \mathbb{E}y_{t-1}\mathbf{1}\{y_{t-1} < C\} + \mathbb{P}(y_{t-1} < C)\mathbb{E}\max(-C, u_t) \geq \mathbb{E}y_{t-1} + C_1.$$

where $C_1 = \mathbb{P}(y_{t-1} < C)\mathbb{E}\max(-C, u_t) > 0$. Thus, $\mathbb{E}y_t \geq tC_1 + const$, and $\mathbb{E}y_t \to \infty$. □

Lemma A.4. *If $\beta = 1$, $\alpha = 0$, then we can equivalently rewrite the evolution of y_t as follows*

$$y_t = [y_{t-1} + u_t]_+ = y_0 + \sum_{s=1}^{t} u_s + \sup_{r=0,\dots,t}\left[-y_0 - \sum_{s=1}^{r} u_s\right]_+.$$

Proof. Define $z_t = y_0 + \sum\limits_{s=1}^{t} u_s + \sup\limits_{r=0,\dots,t}\left[-y_0 - \sum\limits_{s=1}^{r} u_s\right]_+$, $z_0 = y_0$. Note that by definition z_t is always non-negative, as when $y_0 + \sum\limits_{s=1}^{t} u_s$ becomes negative, we are adding its absolute value

or even a larger positive number ($\sup_{r=0,\dots,t}\left[-y_0-\sum_{s=1}^{r}u_s\right]_+$). Let us show that $z_t=y_t$ for all t. Let us proceed by induction.

By definition $z_0=y_0\geq 0$. Let us look at $t=1$. If $y_0+u_1\geq 0$, then $\sup_{r=0,1}\left[-y_0-\sum_{s=1}^{r}u_s\right]_+=0$ and $z_1=y_0+u_1=y_1$. If $y_0+u_1<0$, then $y_1=[y_0+u_1]_+=0$ and $\sup_{r=0,1}\left[-y_0-\sum_{s=1}^{r}u_s\right]_+=-y_0-u_1>0$. Thus, $z_t=y_0+u_1+(-y_0-u_1)=0=y_1$.

Suppose that $z_t=y_t$ for all $t\leq t'$. Let us prove that $z_{t'+1}=y_{t'+1}$. First, suppose that $\exists p\in\{0,\dots,t'\}$ such that $\left[-y_0-\sum_{s=1}^{p}u_s\right]_+\geq\left[-y_0-\sum_{s=1}^{t'+1}u_s\right]_+$. Thus, either $t+1$ is not an argmaximum or it is not a unique argmaximum over $\{0,\dots,t'+1\}$. In that case,

$$\sup_{r=0,\dots,t'}\left[-y_0-\sum_{s=1}^{r}u_s\right]_+=\sup_{r=0,\dots,t'+1}\left[-y_0-\sum_{s=1}^{r}u_s\right]_+$$

and

$$\begin{aligned}z_{t'+1}&=y_0+\sum_{s=1}^{t'+1}u_s+\sup_{r=0,\dots,t'+1}\left[-y_0-\sum_{s=1}^{r}u_s\right]_+\\&=y_0+\sum_{s=1}^{t'}u_s+u_{t'+1}+\sup_{r=0,\dots,t'}\left[-y_0-\sum_{s=1}^{r}u_s\right]_+=z_{t'}+u_{t'+1}=y_{t'}+u_{t'+1}\geq 0,\end{aligned}$$

where we used the induction hypothesis and the observation that $z_t\geq 0$ for all t

Thus,

$$y_{t'+1}=[y_{t'}+u_{t'+1}]_+=y_{t'}+u_{t'+1}=z_{t'+1}.$$

Now suppose that $\forall p\in\{0,\dots,t'\}$, $\left[-y_0-\sum_{s=1}^{p}u_s\right]_+<\left[-y_0-\sum_{s=1}^{t'+1}u_s\right]_+$. Thus, $\left[-y_0-\sum_{s=1}^{t'+1}u_s\right]_+>0$ and

$$z_{t'+1}=y_0+\sum_{s=1}^{t'+1}u_s+\sup_{r=0,\dots,t'+1}\left[-y_0-\sum_{s=1}^{r}u_s\right]_+=y_0+\sum_{s=1}^{t'+1}u_s+\left(-y_0-\sum_{s=1}^{t'+1}u_s\right)_+=0.$$

In turn,

$$\begin{aligned}y_{t'+1}=[y_{t'}+u_{t'+1}]_+&=\left[y_0+\sum_{s=1}^{t'}u_s+\sup_{r=0,\dots,t'}\left[-y_0-\sum_{s=1}^{r}u_s\right]_+u_{t'+1}\right]_+\\&=\left[y_0+\sum_{s=1}^{t'+1}u_s+\sup_{r=0,\dots,t'}\left[-y_0-\sum_{s=1}^{r}u_s\right]_+\right]_+=0=z_{t'+1},\end{aligned}$$

as $y_0 + \sum_{s=1}^{t'+1} u_s < 0$ and $\left[-y_0 - \sum_{s=1}^{p} u_s\right]_+ < -y_0 - \sum_{s=1}^{t'+1} u_s$ for all $p \in \{0, \dots, t'\}$.

Therefore, $z_{t'+1} = y_{t'+1}$. By induction we get that $y_t = z_t$ for all t. □

Theorem A.5. *If $\beta = 1$, $\alpha = 0$, then for all $r \in (0, 1]$, $\frac{1}{\sqrt{T}} y_{\lfloor rT \rfloor} \xrightarrow[T\to\infty]{d} \sigma |W(r)|$, where $\sigma = \mathbb{E}u^2$ and W is a standard Brownian motion.*

Proof. By Lemma A.4, y_t can be alternatively written as

$$y_t = y_0 + \sum_{s=1}^{t} u_s + \sup_{p=0,\dots,t} \left[-y_0 - \sum_{s=1}^{p} u_s\right]_+ .$$

This is Skorokhod transformation for $y_0 + \sum_{s=1}^{t} u_s$, which is a continuous transformation. Thus,

$$\frac{1}{\sqrt{T}} y_{\lfloor rT \rfloor} = \frac{1}{\sqrt{T}} y_0 + \frac{1}{\sqrt{T}} \sum_{s=1}^{\lfloor rT \rfloor} u_s + \sup_{p=0,\dots,\lfloor rT \rfloor} \left[-\frac{y_0}{\sqrt{T}} - \frac{1}{\sqrt{T}} \sum_{s=1}^{p} u_s\right]_+$$

By functional central limit theorem $\frac{1}{\sqrt{T}} \sum_{s=1}^{\lfloor rT \rfloor} u_s \xrightarrow[T\to\infty]{d} \sigma W(r)$. So that using the continuity of Skorokhod transformation,

$$\frac{1}{\sqrt{T}} y_{\lfloor rT \rfloor} \xrightarrow[T\to\infty]{d} \sigma W(r) + \sup_{p\in[0,r]} \left[-\sigma W(p)\right]_+ \stackrel{d}{=} \sigma |W(r)|,$$

where the last equation, which gives equivalence in distribution, was first proved in Lévy (1948). (See, for example, Section 3.6.C in Karatzas and Shreve (2012).) □

Appendix B. Properties of the estimators.

All properties of estimators are proved in this section. We use the following lemma to justify cases, where sums are approximated by expectations.

Lemma B.1. *If y_t converges to a stationary process in Theorem 2 and Theorem 3 and there exists $\mathbb{E}|u|^{4+\varepsilon}$ for some $\varepsilon > 0$, then as $T \to \infty$*

$$\frac{1}{T} \sum_{t=1}^{T} y_t \xrightarrow{\mathbb{P}} \mathbb{E}y,\ \frac{1}{T} \sum_{t=1}^{T} y_t^2 \xrightarrow{\mathbb{P}} \mathbb{E}y^2,\ \frac{1}{T} \sum_{t=1}^{T} y_t z_t \xrightarrow{\mathbb{P}} \mathbb{E}yz,\ \frac{1}{T} \sum_{t=1}^{T} z_t \xrightarrow{\mathbb{P}} \mathbb{E}z,\ \frac{1}{T} \sum_{t=1}^{T} z_t^2 \xrightarrow{\mathbb{P}} \mathbb{E}z^2,$$

where expectations are taken with respect to the stationary distribution.

Proof. Note that if the error u_t has a moment of order k, then y_t also has a moment of order k.

Let us show that $Cov(y_s, y_{s+t}) \xrightarrow[t\to\infty]{} 0$.

$$\begin{aligned} Cov(y_s, y_{s+t}) &= \mathbb{E}y_s y_{t+s} - \mathbb{E}y_s \mathbb{E}y_{t+s} = \mathbb{E}y_s y_{t+s}\mathbf{1}(y_s < M, y_{t+s} < M) - \mathbb{E}y_s \mathbb{E}y_{t+s} \\ &+ \mathbb{E}y_s y_{t+s}\mathbf{1}(y_s \geq M, y_{t+s} < M) + \mathbb{E}y_s y_{t+s}\mathbf{1}(y_s < M, y_{t+s} \geq M) \\ &+ \mathbb{E}y_s y_{t+s}\mathbf{1}(y_s \geq M, y_{t+s} \geq M) \end{aligned}$$

As y_t is mixing and converges in distribution to a random variable, its limit is independent of y_s,

$$\mathbb{E}y_s y_{t+s}\mathbf{1}(y_s < M, y_{t+s} < M) \to \mathbb{E}y_s\mathbf{1}(y_s < M)\mathbb{E}y_{s+t}\mathbf{1}(y_{s+t} < M).$$

As M goes to infinity, $\mathbb{E}y_s\mathbf{1}(y_s < M) \to \mathbb{E}y_s$ for all s. Thus, by choice of M large enough we can make $\lim\limits_{t\to\infty} (\mathbb{E}y_s y_{t+s}\mathbf{1}(y_s < M, y_{t+s} < M) - \mathbb{E}y_s\mathbb{E}y_{t+s})$ arbitrarily close to zero. Moreover,

$$\begin{aligned} &\mathbb{E}y_s y_{t+s}\mathbf{1}(y_s \geq M, y_{t+s} < M) \leq \mathbb{E}\frac{y_s^{1+\varepsilon/2}}{M^{\varepsilon/2}} y_{t+s}\mathbf{1}(y_s \geq M, y_{t+s} < M) \\ &\leq \frac{1}{M^{\varepsilon/2}}\mathbb{E}(y_s^{2+\varepsilon} + y_{t+s}^2)\mathbf{1}(y_s \geq M, y_{t+s} < M) \leq \frac{1}{M^{\varepsilon/2}}\mathbb{E}(y_s^{2+\varepsilon} + y_{t+s}^2) \xrightarrow[M\to\infty]{} 0. \end{aligned}$$

Similarly,

$$\mathbb{E}y_s y_{t+s}\mathbf{1}(y_s < M, y_{t+s} \geq M) \leq \frac{1}{M^{\varepsilon/2}}\mathbb{E}(y_s^2 + y_{t+s}^{2+\varepsilon}) \xrightarrow[M\to\infty]{} 0.$$

Finally,

$$\begin{aligned} &\mathbb{E}y_s y_{t+s}\mathbf{1}(y_s \geq M, y_{t+s} \geq M) \leq \mathbb{E}(y_s^2 + y_{t+s}^2)\mathbf{1}(y_s \geq M, y_{t+s} \geq M) \\ &\leq \frac{1}{M^{\varepsilon}}\mathbb{E}(y_s^{2+\varepsilon} + y_{t+s}^{2+\varepsilon})\mathbf{1}(y_s \geq M, y_{t+s} \geq M) \leq \frac{1}{M^{\varepsilon}}\mathbb{E}(y_s^{2+\varepsilon} + y_{t+s}^{2+\varepsilon}) \xrightarrow[M\to\infty]{} 0. \end{aligned}$$

Thus, by choice of M large enough, $\mathbb{E}y_s y_{t+s}\mathbf{1}(y_s \geq M, y_{t+s} < M) + \mathbb{E}y_s y_{t+s}\mathbf{1}(y_s < M, y_{t+s} \geq M) + \mathbb{E}y_s y_{t+s}\mathbf{1}(y_s \geq M, y_{t+s} \geq M)$ can be made arbitrarily close to zero. That is, $Cov(y_s, y_{s+t}) \to 0$ as $t \to \infty$. This means that the variance of $\frac{1}{T}\sum\limits_{t=1}^{T} y_t$ goes to zero as $T \to \infty$, so that law of large numbers holds for y_t.

To see that the variance of $\frac{1}{T}\sum\limits_{t=1}^{T} y_t$ goes to zero as $T \to \infty$, note that

$$\mathbb{V}\sum_{t=1}^{T} y_t = \mathbb{E}\left(\sum_{t=1}^{T} y_t - T\mathbb{E}y\right)^2 = \sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}(y_t - \mathbb{E}y)(y_s - \mathbb{E}y) \leq \sum_{s=1}^{T}\left(\sum_{t=0}^{T} Cov(y_s, y_{s+t})\right)$$

Because $Cov(y_s, y_{s+t})$ goes to zero as $t \to \infty$, $\sum\limits_{t=0}^{T} Cov(y_s, y_{s+t}) = o(T)$ (otherwise terms would not disappear as $t \to \infty$). Thus, $\sum\limits_{s=1}^{T}\left(\sum\limits_{t=0}^{T} Cov(y_s, y_{s+t})\right) = o(T^2)$ (there are T terms each of order $o(T)$). That is, $\mathbb{V}\sum\limits_{t=1}^{T}\frac{1}{T}y_t = \frac{o(T^2)}{T^2} = o(1)$ and the law of large numbers holds for y_t.

The same logic applies to the four other limits. Corresponding moments of z_t are finite, as z_t is bounded by the maximum of finite number of variables $\{y_{ijs}\}_{s=t,\dots,t-H+1}$ (Assumption 5). □

Proof of Theorem 5. Denote the LAD estimate as $(\hat{\alpha}, \hat{\beta}, \hat{\gamma})$. The estimate is the solution to minimization problem (7). The corresponding first order conditions are

$$\sum_{t=1}^{T} \begin{pmatrix} 1 \\ y_{t-1} \\ z_{t-1} \end{pmatrix} \operatorname{sgn}\left(y_t - a - by_{t-1} - cz_{t-1}\right) = 0. \tag{23}$$

Define $\Delta\alpha = \hat{\alpha} - \alpha$, $\Delta\beta = \hat{\beta} - \beta$, $\Delta\gamma = \hat{\gamma} - \gamma$, $v_{t+1} = \max(u_{t+1}, -\alpha - \beta y_t - \gamma z_t)$. Then, noting that $y_t = \alpha + \beta y_{t-1} + \gamma z_{t-1} + v_t$, we can rewrite Eq. (23) as

$$\sum_{t} \begin{pmatrix} 1 \\ y_{t-1} \\ z_{t-1} \end{pmatrix} \operatorname{sgn}\left(v_{t+1} - \Delta\alpha - \Delta\beta y_t - \Delta\gamma z_t\right) = 0. \tag{24}$$

Dividing Eq. (24) by T and applying the law of large numbers, we get

$$\mathbb{E} \begin{pmatrix} 1 \\ y_{t-1} \\ z_{t-1} \end{pmatrix} \operatorname{sgn}\left(v_{t+1} - \Delta\alpha - \Delta\beta y_t - \Delta\gamma z_t\right) = 0. \tag{25}$$

Let us linearize the expectations in Eq. (25).

$$\begin{aligned} &\mathbb{E}\operatorname{sgn}\left(v_{t+1} - \Delta\alpha - \Delta\beta y_t - \Delta\gamma z_t\right) \\ &= \int_y \int_z \int_v \operatorname{sgn}\left(v - \Delta\alpha - \Delta\beta y - \Delta\gamma z\right) f_{y,z,v}(y,z,v) dv dz dy \\ &= \int_y \int_z \int_{\Delta\alpha+\Delta\beta y+\Delta\gamma z}^{\infty} f_{y,z,v}(y,z,v) dv dz dy - \int_y \int_z \int_{-\infty}^{\Delta\alpha+\Delta\beta y+\Delta\gamma z} f_{y,z,v}(y,z,v) dv dz dy \\ &= 1 - 2\mathbb{E}_{y,z} F_{v|y,z}(\Delta\alpha + \Delta\beta y + \Delta\gamma z | y, z). \end{aligned} \tag{26}$$

Taylor expanding around $\Delta\alpha + \Delta\beta y + \Delta\gamma z = 0$, we can rewrite Eq. (26) as

$$1 - 2\mathbb{E}_{y,z}\left(F_{v|y,z}(0|y,z) + f_{v|y,z}(0|y,z)(\Delta\alpha + \Delta\beta y + \Delta\gamma z | y, z)\right).$$

As $v_{t+1} = \max(u_{t+1}, -\alpha - \beta y_t - \gamma z_t)$ and $-\alpha - \beta y_t - \gamma z_t < 0$ for $\alpha > 0$, $\beta \geq 0$, $\gamma \geq 0$, the density of v_{t+1} has a mass point at $-\alpha - \beta y_t - \gamma z_t$ and then coincides with the density of u_{t+1}. Thus, $F_{v|y,z}(0|y,z) = F_u(0) = 0.5$ and $f_{v|y,z}(0|y,z) = f_u(0)$. To sum up, the first order condition with respect to α becomes

$$-2f_u(0)\mathbb{E}(\Delta\alpha + \Delta\beta y + \Delta\gamma z) = 0. \tag{27}$$

Let us now analyze the second line of Eq. (25) in a similar fashion.

(28)

$$\mathbb{E}y_t \text{sgn}\left(v_{t+1} - \Delta\alpha - \Delta\beta y_t - \Delta\gamma z_t\right) = \int_y \int_z \int_v y \text{sgn}\left(v - \Delta\alpha - \Delta\beta y - \Delta\gamma z\right) f_{y,z,v}(y,z,v) dv dz dy$$
$$= \mathbb{E}y - 2\mathbb{E}_{y,z} y F_{v|y,z}(\Delta\alpha + \Delta\beta y + \Delta\gamma z | y, z)$$
$$\approx \mathbb{E}y - 2\mathbb{E}_{y,z} y \left(F_{v|y,z}(0|y,z) + f_{v|y,z}(0|y,z)(\Delta\alpha + \Delta\beta y + \Delta\gamma z|y,z)\right)$$
$$= -2f_u(0)\mathbb{E}y(\Delta\alpha + \Delta\beta y + \Delta\gamma z).$$

Thus, first order condition with respect to β becomes

$$-2f_u(0)\mathbb{E}y(\Delta\alpha + \Delta\beta y + \Delta\gamma z) = 0. \tag{29}$$

Similarly, the first order condition with respect to γ becomes

$$-2f_u(0)\mathbb{E}z(\Delta\alpha + \Delta\beta y + \Delta\gamma z) = 0. \tag{30}$$

Combining Eq. (27), (29), and (30), we get

$$\begin{pmatrix} 1 & \mathbb{E}y & \mathbb{E}z \\ \mathbb{E}y & \mathbb{E}yz & \mathbb{E}yz \\ \mathbb{E}z & \mathbb{E}yz & \mathbb{E}z^2 \end{pmatrix} \begin{pmatrix} \Delta\alpha \\ \Delta\beta \\ \Delta\gamma \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Matrix $\begin{pmatrix} 1 & \mathbb{E}y & \mathbb{E}z \\ \mathbb{E}y & \mathbb{E}yz & \mathbb{E}yz \\ \mathbb{E}z & \mathbb{E}yz & \mathbb{E}z^2 \end{pmatrix}$ is nonsingular. If it were singular, it would have an eigenvector $(\lambda_1, \lambda_2, \lambda_2)$ corresponding to the zero eigenvalue. Then the random variable $\lambda_1 + \lambda_2 y_t + \lambda_3 z_t$ must have zero second moment. That is, $\lambda_1 + \lambda_2 y_t + \lambda_3 z_t$ must be zero, which contradicts conditions of the theorem.

Thus, the only solution to the above system is $\Delta\alpha = 0$, $\Delta\beta = 0$, $\Delta\gamma = 0$. That is, the LAD estimator is consistent and converges to the true value of the parameters as $T \to \infty$. □

Proof of Theorem 6. This proof follows the lines of Powell (1984).

Define $\theta = (\alpha, \beta, \gamma)'$, $x_t = (1,\ y_{t-1},\ z_{t-1})'$, $\mathcal{F}_t = \{y_t, z_t, y_{t-1}, z_{t-1}, \ldots\}$, and

$$S_T(\theta) = \frac{1}{T}\sum_t |y_t - [\alpha + \beta y_{t-1} + \gamma z_{t-1}]_+| = \frac{1}{T}\sum_t \left|[x_t'\theta_0 + u_t]_+ - [x_t'\theta]_+\right|,$$

where θ_0 corresponds to the true value of θ.

We want to show that $S_T(\theta) - S_T(\theta_0)$ is uniformly bounded away from zero for large T and $||\theta - \theta_0|| > \varepsilon$ for any $\varepsilon > 0$. Then $\hat{\theta}_{LAD} \underset{\mathbb{P}}{\to} \theta_0$.

Let us write

$$Q_T(\theta) := S_T(\theta) - S_T(\theta_0) = \frac{1}{T}\sum_t \Bigg[|[x_t'\theta_0 + u_t]_+ - [x_t'\theta]_+| - |[x_t'\theta_0 + u_t]_+ - [x_t'\theta_0]_+| \\ - \mathbb{E}\left(|[x_t'\theta_0 + u_t]_+ - [x_t'\theta]_+| - |[x_t'\theta_0 + u_t]_+ - [x_t'\theta_0]_+| \,|\mathcal{F}_{t-1}\right)\Bigg] \\ + \frac{1}{T}\sum_t \mathbb{E}\left(|[x_t'\theta_0 + u_t]_+ - [x_t'\theta]_+| - |[x_t'\theta_0 + u_t]_+ - [x_t'\theta_0]_+| \,|\mathcal{F}_{t-1}\right) \tag{31}$$

Let us analyze the first summation in Eq. (31). Define

$$s_t(\theta, x_t) := |[x_t'\theta_0 + u_t]_+ - [x_t'\theta]_+| - |[x_t'\theta_0 + u_t]_+ - [x_t'\theta_0]_+|$$

and note that $s_t(\theta, x_t) - \mathbb{E}(s_t(\theta, x_t)|\mathcal{F}_{t-1})$ is a martingale difference sequence. Thus,

$$\mathbb{E}\left(\sum_t \left(s_t(\theta, x_t) - \mathbb{E}(s_t(\theta, x_t)|\mathcal{F}_{t-1})\right)\right)^2 = \sum_t \mathbb{E}\left(s_t(\theta, x_t) - \mathbb{E}(s_t(\theta, x_t)|\mathcal{F}_{t-1})\right)^2.$$

$$\mathbb{E}\left(s_t(\theta, x_t) - \mathbb{E}(s_t(\theta, x_t)|\mathcal{F}_{t-1})\right)^2 = \mathbb{E}\Bigg(|[x_t'\theta_0 + u_t]_+ - [x_t'\theta]_+| - |[x_t'\theta_0 + u_t]_+ - [x_t'\theta_0]_+| \\ - \int \left(|[x_t'\theta_0 + u]_+ - [x_t'\theta]_+| - |[x_t'\theta_0 + u]_+ - [x_t'\theta_0]_+|\right) f(u)du\Bigg)^2,$$

which is a function of stationary distribution of y_t and z_t and parameter θ. Because $\theta \in \Theta$, which is compact, and second moments of y and z exist, $\mathbb{E}\left(s_t(\theta, x_t) - \mathbb{E}(s_t(\theta, x_t)|\mathcal{F}_{t-1})\right)^2$ is bounded by some constant C (which depends on first two moments of y and z). Therefore, by Markov inequality, for any $a > 0$

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_t \left(s_t(\theta, x_t) - \mathbb{E}(s_t(\theta, x_t)|\mathcal{F}_{t-1})\right)\right| > a\right) \leq \frac{\mathbb{E}\left|\sum_t \left(s_t(\theta, x_t) - \mathbb{E}(s_t(\theta, x_t)|\mathcal{F}_{t-1})\right)\right|^2}{T^2a^2} \\ = \frac{T\mathbb{E}\left(s_t(\theta, x_t) - \mathbb{E}(s_t(\theta, x_t)|\mathcal{F}_{t-1})\right)^2}{T^2a^2} \leq \frac{C}{Ta^2} \xrightarrow[T\to\infty]{} 0.$$

Thus,

$$\frac{1}{T}\sum_t \left(s_t(\theta, x_t) - \mathbb{E}(s_t(\theta, x_t)|\mathcal{F}_{t-1})\right) \xrightarrow[T\to\infty]{\mathbb{P}} 0$$

uniformly over θ.

Let us now analyze the second summation in Eq. (31). We are going to show that $\mathbb{E}\left(|y_t - [x_t'\theta]_+| - |y_t - [x_t'\theta_0]_+| \,|\mathcal{F}_{t-1}\right)$ is always non-negative.

Because $y_t = [x_t'\theta_0 + u_t]_+$,

$$\mathbb{E}(|y_t - [x_t'\theta]_+||\mathcal{F}_{t-1}) = \mathbf{1}(x_t'\theta < 0) \int\limits_{-x_t'\theta_0}^{\infty} (x_t'\theta_0 + u) f_u(u) du + \mathbf{1}(x_t'\theta \geq 0) \left(x_t'\theta F_u(-x_t'\theta_0) + \int\limits_{-x_t'\theta_0}^{\infty} |u - x_t'(\theta - \theta_0)| f_u(u) du \right)$$

and

$$\mathbb{E}(|y_t - [x_t'\theta_0]_+||\mathcal{F}_{t-1}) = \mathbf{1}(x_t'\theta_0 < 0) \int\limits_{-x_t'\theta_0}^{\infty} (x_t'\theta_0 + u) f_u(u) du + \mathbf{1}(x_t'\theta_0 \geq 0) \left(x_t'\theta_0 F_u(-x_t'\theta_0) + \int\limits_{-x_t'\theta_0}^{\infty} |u| f_u(u) du \right).$$

Therefore, omitting the derivations, we get

$$\begin{aligned} \mathbb{E}\left(|y_t - [x_t'\theta]_+| - |y_t - [x_t'\theta_0]_+| \,|\mathcal{F}_{t-1}\right) &= 2\mathbf{1}(x_t'\theta_0 \geq 0, x_t'\theta < 0) \int\limits_{-x_t'\theta_0}^{0} (x_t'\theta_0 + u) f_u(u) du \\ &+ 2\mathbf{1}(x_t'\theta_0 < 0, x_t'\theta \geq 0) \left(\int\limits_{-x_t'\theta_0}^{x_t'(\theta-\theta_0)} (x_t'(\theta - \theta_0) - u) f_u(u) du + \int\limits_{0}^{-x_t'\theta_0} x_t'\theta f_u(u) du \right) \\ &+ 2\mathbf{1}(x_t'\theta_0 \geq 0, x_t'\theta \geq 0) \int\limits_{0}^{x_t'(\theta-\theta_0)} (x_t'(\theta - \theta_0) - u) f_u(u) du \geq 0, \end{aligned} \tag{32}$$

as every function under integral is non-negative over the domain of integration.

Because all terms in the Eq. (32) are non-negative,

$$\begin{aligned} \mathbb{E}\left(|y_t - [x_t'\theta]_+| - |y_t - [x_t'\theta_0]_+| \,|\mathcal{F}_{t-1}\right) &\geq 2\mathbf{1}(x_t'\theta_0 \geq 0, x_t'\theta < 0) \int\limits_{-x_t'\theta_0}^{0} (x_t'\theta_0 + u) f_u(u) du \\ &+ 2\mathbf{1}(x_t'\theta_0 \geq 0, x_t'\theta \geq 0) \int\limits_{0}^{x_t'(\theta-\theta_0)} (x_t'(\theta - \theta_0) - u) f_u(u) du. \end{aligned}$$

Moreover, for $R \geq 0$ such that M_R is nonsingular and any $\tau \in (0, R]$,

$$\mathbb{E}\left(|y_t - [x_t'\theta]_+| - |y_t - [x_t'\theta_0]_+| \,|\mathcal{F}_{t-1}\right)$$

$$\geq 2\mathbf{1}(x_t'\theta_0 \geq R, x_t'\theta < 0)\mathbf{1}(|x_t'(\theta-\theta_0)| \geq \tau) \int_{-\tau}^{0} (\tau+u) f_u(u) du$$

$$+ 2\mathbf{1}(x_t'\theta_0 \geq R, x_t'\theta \geq 0)\mathbf{1}(|x_t'(\theta-\theta_0)| \geq \tau) \int_{0}^{\tau} (\tau-u) f_u(u) du$$

$$\geq 2\min\left(\int_{-\tau}^{0} (\tau+u) f_u(u) du, \int_{0}^{\tau} (\tau-u) f_u(u) du \right) \mathbf{1}(x_t'\theta_0 \geq R)\mathbf{1}(|x_t'(\theta-\theta_0)| \geq \tau).$$

Thus,

$$\frac{1}{T}\sum_t \mathbb{E}\left(|y_t - [x_t'\theta]_+| - |y_t - [x_t'\theta_0]_+| \,|\mathcal{F}_{t-1}\right)$$

$$\geq \frac{2}{T}\min\left(\int_{-\tau}^{0} (\tau+u) f_u(u) du, \int_{0}^{\tau} (\tau-u) f_u(u) du \right) \sum_t \mathbf{1}(x_t'\theta_0 \geq R)\mathbf{1}(|x_t'(\theta-\theta_0)| \geq \tau).$$

As T goes to infinity, $\frac{1}{T}\sum_t \mathbf{1}(x_t'\theta_0 \geq R)\mathbf{1}(|x_t'(\theta-\theta_0)| \geq \tau)$ converges to

$$\mathbb{E}\mathbf{1}(x_t'\theta_0 \geq R,\, |x_t'(\theta-\theta_0)| \geq \tau) = \mathbb{P}(x_t'\theta_0 \geq R,\, |x_t'(\theta-\theta_0)| \geq \tau)$$
$$= \mathbb{P}(x_t'\theta_0 \geq R)\mathbb{P}(|x_t'(\theta-\theta_0)| \geq \tau |\, x_t'\theta_0 \geq R).$$

Because M_R is nonsingular, $\mathbb{P}(x_t'\theta_0 \geq R) > 0$ (otherwise indicator in M_R will always be zero, so that the matrix M_R will be identically zero).

We are going to prove that

$$\mathbb{P}(|x_t'(\theta-\theta_0)| \geq \tau_0 |\, x_t'\theta_0 \geq R) \geq C_1 > 0,$$

where $\tau_0 = const \cdot ||\theta-\theta_0||^2$ and $||\cdot||$ denotes L_2 norm. This will imply that the sum of conditional expectations (Eq. (32)) is bounded from zero uniformly in $||\theta-\theta_0||$. Thus, initial summation (Eq. (31)) is also uniformly bounded from zero, so that as T goes to infinity, for any $\theta \neq \theta_0$ with probability tending to one we have

$$\frac{1}{T}\sum_t \mathbb{E}\left(|y_t - [x_t'\theta]_+| - |y_t - [x_t'\theta_0]_+| \,|\mathcal{F}_{t-1}\right) \geq const(||\theta-\theta_0||) > 0$$

and

$$\lim_{T\to\infty} \mathbb{P}(Q_T(\theta) \geq const(||\theta-\theta_0||)) = 1,$$

where constant is increasing in $||\theta-\theta_0||$.

In contrast, for any T, $Q_T(\theta_0) = 0$. Thus, we must have that LAD estimate $\hat{\theta}_T$ converges to θ_0 as $T \to \infty$, as $Q_T(\theta)$ is bounded from zero for $\theta \neq \theta_0$ and T large enough.

Let us show that $\mathbb{P}(|x_t'(\theta-\theta_0)| \geq \tau_0 |\, x_t'\theta_0 \geq R) \geq C_1 > 0$, where $\tau_0 = const \cdot ||\theta-\theta_0||^2$. Define by $\lambda_{\min}$ the minimal eigenvalue of matrix $\mathbb{E}(x_t x_t' | x_t'\theta_0 > R)$. It is non-zero, because

the matrix is nonsingular. Then

$$\mathbb{E}\left(|x_t'(\theta-\theta_0)|^2\,|x_t'\theta_0\geq R\right)\geq||\theta-\theta_0||^2\lambda_{\min},$$

as such conditional expectation corresponds to the value of the quadratic form $\mathbb{E}(x_tx_t'|x_t'\theta_0>R)$ on vector $\theta-\theta_0$.

Choose $\varepsilon>0$ such that $\varepsilon<||\theta-\theta_0||^2\lambda_{\min}$ and note that for any $A>0$

$$\begin{aligned}
&\mathbb{E}\left(|x_t'(\theta-\theta_0)|^2\,|x_t'\theta_0\geq R\right)=\mathbb{E}\left(|x_t'(\theta-\theta_0)|^2\mathbf{1}(|x_t'(\theta-\theta_0)|^2)<A\,|x_t'\theta_0\geq R\right)\\
&+\mathbb{E}\left(|x_t'(\theta-\theta_0)|^2\mathbf{1}(|x_t'(\theta-\theta_0)|^2\geq A)\,|x_t'\theta_0\geq R\right)\\
&\leq\mathbb{E}\left(|x_t'(\theta-\theta_0)|^2\mathbf{1}(|x_t'(\theta-\theta_0)|^2<A)\,|x_t'\theta_0\geq R\right)\\
&+\mathbb{E}\left(\frac{|x_t'(\theta-\theta_0)|^4}{A}\mathbf{1}(|x_t'(\theta-\theta_0)|^2\geq A)\,|x_t'\theta_0\geq R\right)\\
&\leq\mathbb{E}\left(|x_t'(\theta-\theta_0)|^2\mathbf{1}(|x_t'(\theta-\theta_0)|^2<A)\,|x_t'\theta_0\geq R\right)+\frac{\mathbb{E}\left(|x_t'(\theta-\theta_0)|^4\,|x_t'\theta_0\geq R\right)}{A},
\end{aligned}$$

where the fourth moment exists, because u_t has fourth moment. Choose $A(\varepsilon)$ such that $A(\varepsilon)>\frac{||\theta-\theta_0||^4\mathbb{E}\left(||x_t||^4\,|x_t'\theta_0\geq R\right)}{\varepsilon}$. Then

$$\frac{\mathbb{E}\left(|x_t'(\theta-\theta_0)|^4\,|x_t'\theta_0\geq R\right)}{A}<\frac{\varepsilon\mathbb{E}\left(|x_t'(\theta-\theta_0)|^4\,|x_t'\theta_0\geq R\right)}{||\theta-\theta_0||^4\mathbb{E}\left(||x_t||^4\,|x_t'\theta_0\geq R\right)}\leq\varepsilon,$$

as $\mathbb{E}\left(|x_t'(\theta-\theta_0)|^4\,|x_t'\theta_0\geq R\right)\leq||\theta-\theta_0||^4\mathbb{E}\left(||x_t||^4\,|x_t'\theta_0\geq R\right)$. Thus,

$$\begin{aligned}
\mathbb{E}\left(|x_t'(\theta-\theta_0)|^2\mathbf{1}(|x_t'(\theta-\theta_0)|^2<A(\varepsilon))\,|x_t'\theta_0\geq R\right)&\geq\mathbb{E}\left(|x_t'(\theta-\theta_0)|^2\,|x_t'\theta_0\geq R\right)-\varepsilon\\
&\geq||\theta-\theta_0||^2\lambda_{\min}-\varepsilon>0.
\end{aligned}$$

Finally,

$$\begin{aligned}
&\mathbb{E}\left(|x_t'(\theta-\theta_0)|^2\mathbf{1}(|x_t'(\theta-\theta_0)|^2<A(\varepsilon))\,|x_t'\theta_0\geq R\right)\\
&=\mathbb{E}\left(|x_t'(\theta-\theta_0)|^2\mathbf{1}(|x_t'(\theta-\theta_0)|^2<A(\varepsilon))\mathbf{1}(|x_t'(\theta-\theta_0)|\geq\tau)\,|x_t'\theta_0\geq R\right)\\
&+\mathbb{E}\left(|x_t'(\theta-\theta_0)|^2\mathbf{1}(|x_t'(\theta-\theta_0)|^2<A(\varepsilon))\mathbf{1}(|x_t'(\theta-\theta_0)|<\tau)\,|x_t'\theta_0\geq R\right)\\
&\leq A(\varepsilon)\mathbb{P}(|x_t'(\theta-\theta_0)|\geq\tau|\,x_t'\theta_0\geq R)+\tau,
\end{aligned}$$

so that for $\tau\in(0,||\theta-\theta_0||^2\lambda_{\min}-\varepsilon)$,

$$\mathbb{P}(|x_t'(\theta-\theta_0)|\geq\tau|\,x_t'\theta_0\geq R)\geq\frac{||\theta-\theta_0||^2\lambda_{\min}-\varepsilon-\tau}{A(\varepsilon)}=C_1>0.$$

Fixing $\varepsilon=\frac{1}{2}||\theta-\theta_0||^2\lambda_{\min}$, $A(\varepsilon)=4\frac{||\theta-\theta_0||^2\mathbb{E}\left(||x_t||^4\,|x_t'\theta_0\geq R\right)}{\lambda_{\min}}$, $\tau=\frac{1}{4}||\theta-\theta_0||^2\lambda_{\min}$. Then

$$\mathbb{P}(|x_t'(\theta-\theta_0)|\geq\tau|\,x_t'\theta_0\geq R)\geq\frac{\lambda_{\min}^2}{16\mathbb{E}\left(||x_t||^4|x_t'\theta_0\geq R\right)}=C_1>0.$$

If $||\theta' - \theta_0|| > ||\theta - \theta_0||$, then

$$\begin{aligned}&\mathbb{P}(|x_t'(\theta' - \theta_0)| \geq \frac{1}{4}||\theta - \theta_0||^2 \lambda_{\min}| \, x_t'\theta_0 \geq R) \\ &\geq \mathbb{P}(|x_t'(\theta' - \theta_0)| \geq \frac{1}{4}||\theta' - \theta_0||^2 \lambda_{\min}| \, x_t'\theta_0 \geq R) \geq \frac{\lambda_{\min}^2}{16\mathbb{E}\left(||x_t||^4 | x_t'\theta_0 \geq R\right)}.\end{aligned}$$ □

Proof of Theorem 7. Denote the LAD estimate as $(\hat{\alpha}, \hat{\beta}, \hat{\gamma})$. We are going to use results from the proof of Theorem 5. The difference is that now we will apply the martingale central limit theorem instead of the law of large numbers. Define the filtration $\mathcal{F}_t = \{y_t, z_t, y_{t-1}, z_{t-1}, \ldots\}$. Then

$$\xi_t := \begin{pmatrix} \text{sgn}\left(v_{t+1} - \Delta\alpha - \Delta\beta y_t - \Delta\gamma z_t\right) - \mathbb{E}\left(\text{sgn}\left(v_{t+1} - \Delta\alpha - \Delta\beta y_t - \Delta\gamma z_t\right)|\mathcal{F}_t\right) \\ y_t\text{sgn}\left(v_{t+1} - \Delta\alpha - \Delta\beta y_t - \Delta\gamma z_t\right) - \mathbb{E}\left(y_t\text{sgn}\left(v_{t+1} - \Delta\alpha - \Delta\beta y_t - \Delta\gamma z_t\right)|\mathcal{F}_t\right) \\ z_t\text{sgn}\left(v_{t+1} - \Delta\alpha - \Delta\beta y_t - \Delta\gamma z_t\right) - \mathbb{E}\left(z_t\text{sgn}\left(v_{t+1} - \Delta\alpha - \Delta\beta y_t - \Delta\gamma z_t\right)|\mathcal{F}_t\right) \end{pmatrix}$$

is a martingale difference with respect to filtration $\mathcal{F}_t$.

Let us calculate the corresponding asymptotic covariance matrix.

First, observe that

$$\mathbb{E}\left(\text{sgn}\left(v_{t+1} - \Delta\alpha - \Delta\beta y_t - \Delta\gamma z_t\right)|\mathcal{F}_t\right) = 1 - 2\mathbb{P}_v\left(v_{t+1} < \Delta\alpha + \Delta\beta y_t + \Delta\gamma z_t\right).$$

When $\Delta\alpha + \Delta\beta y_t + \Delta\gamma z_t \approx 0$, the probability that v_{t+1} is smaller than $\Delta\alpha + \Delta\beta y_t + \Delta\gamma z_t$ equals to $1/2$, as $\text{med}(v) = 0$. Thus, $\mathbb{E}\left(\text{sgn}\left(v_{t+1} - \Delta\alpha - \Delta\beta y_t - \Delta\gamma z_t\right)|\mathcal{F}_t\right) \approx 0$ for $\Delta\alpha + \Delta\beta y_t + \Delta\gamma z_t \approx 0$.

Second, note that $\text{sgn}^2\left(v_{t+1} - \Delta\alpha - \Delta\beta y_t - \Delta\gamma z_t\right) \equiv 1$, so that

- $\mathbb{E}\left(y_t\text{sgn}^2\left(v_{t+1} - \Delta\alpha - \Delta\beta y_t - \Delta\gamma z_t\right)|\mathcal{F}_t\right) \equiv y_t$,
- $\mathbb{E}\left(z_t\text{sgn}^2\left(v_{t+1} - \Delta\alpha - \Delta\beta y_t - \Delta\gamma z_t\right)|\mathcal{F}_t\right) \equiv z_t$,
- $\mathbb{E}\left(y_t z_t\text{sgn}^2\left(v_{t+1} - \Delta\alpha - \Delta\beta y_t - \Delta\gamma z_t\right)|\mathcal{F}_t\right) \equiv y_t z_t$,
- $\mathbb{V}\left(\text{sgn}\left(v_{t+1} - \Delta\alpha - \Delta\beta y_t - \Delta\gamma z_t\right)|\mathcal{F}_t\right) = 1 - 0 = 1$,
- $\mathbb{V}\left(y_t\text{sgn}\left(v_{t+1} - \Delta\alpha - \Delta\beta y_t - \Delta\gamma z_t\right)|\mathcal{F}_t\right) = y_t^2(1-0) = y_t^2$,
- $\mathbb{V}\left(z_t\text{sgn}\left(v_{t+1} - \Delta\alpha - \Delta\beta y_t - \Delta\gamma z_t\right)|\mathcal{F}_t\right) = z_t^2(1-0) = z_t^2$.

Under stationarity $\frac{1}{T}\sum_t y_t \to \mathbb{E}y$, $\frac{1}{T}\sum_t z_t \to \mathbb{E}z$, $\frac{1}{T}\sum_t y_t^2 \to \mathbb{E}y^2$, $\frac{1}{T}\sum_t y_t z_t \to \mathbb{E}yz$, $\frac{1}{T}\sum_t z_t^2 \to \mathbb{E}z^2$. We want to apply the martingale CLT. Thus, we need to check the Lindeberg condition. Because

$$\sum_{t=1}^{T} \mathbb{V}\left(\frac{1}{\sqrt{T}}\xi_t|\mathcal{F}_t\right) = \begin{pmatrix} 1 & \frac{1}{T}\sum_t y_t & \frac{1}{T}\sum_t z_t \\ \frac{1}{T}\sum_t y_t & \frac{1}{T}\sum_t y_t^2 & \frac{1}{T}\sum_t y_t z_t \\ \frac{1}{T}\sum_t z_t & \frac{1}{T}\sum_t y_t z_t & \frac{1}{T}\sum_t z_t^2 \end{pmatrix},$$

the variance matrix is $O(1)$. So we need to check that for any numbers λ_1, λ_2, λ_3 and $\varepsilon > 0$

$$\sum_{t=1}^{T} \mathbb{E}\left(\frac{1}{T}(\lambda_1\xi_{1t} + \lambda_2\xi_{2t} + \lambda_3\xi_{3t})^2 \mathbf{1}\left(\frac{1}{\sqrt{T}}|\lambda_1\xi_{1t} + \lambda_2\xi_{2t} + \lambda_3\xi_{3t}| > \varepsilon\right)\right) \xrightarrow[T\to\infty]{} 0. \tag{33}$$

Because y_t and z_t are stationary, Eq. (33) reduces to

$$\mathbb{E}(\lambda_1\xi_{1t} + \lambda_2\xi_{2t} + \lambda_3\xi_{3t})^2 \mathbf{1}\left(\frac{1}{\sqrt{T}}|\lambda_1\xi_{1t} + \lambda_2\xi_{2t} + \lambda_3\xi_{3t}| > \varepsilon\right) \xrightarrow[T\to\infty]{} 0,$$

which is true as $(1, y_t, z_t)$ and, thus, $\xi_t = (\xi_{1t}, \xi_{2t}, \xi_{3t})'$ and any linear combination of ξ_t's coordinates, have finite second moments.

So by the martingale CLT,

$$\frac{1}{\sqrt{T}}\begin{pmatrix} \sum_t \left[\operatorname{sgn}\left(v_{t+1} - \Delta\alpha - \Delta\beta y_t - \Delta\gamma z_t\right) - \mathbb{E}\left(\operatorname{sgn}\left(v_{t+1} - \Delta\alpha - \Delta\beta y_t - \Delta\gamma z_t\right)|\mathcal{F}_t\right)\right] \\ \sum_t \left[y_t\operatorname{sgn}\left(v_{t+1} - \Delta\alpha - \Delta\beta y_t - \Delta\gamma z_t\right) - \mathbb{E}\left(y_t\operatorname{sgn}\left(v_{t+1} - \Delta\alpha - \Delta\beta y_t - \Delta\gamma z_t\right)|\mathcal{F}_t\right)\right] \\ \sum_t \left[z_t\operatorname{sgn}\left(v_{t+1} - \Delta\alpha - \Delta\beta y_t - \Delta\gamma z_t\right) - \mathbb{E}\left(z_t\operatorname{sgn}\left(v_{t+1} - \Delta\alpha - \Delta\beta y_t - \Delta\gamma z_t\right)|\mathcal{F}_t\right)\right] \end{pmatrix} \xrightarrow[T\to\infty]{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \mathbb{E}y & \mathbb{E}z \\ \mathbb{E}y & \mathbb{E}y^2 & \mathbb{E}yz \\ \mathbb{E}z & \mathbb{E}yz & \mathbb{E}z^2 \end{pmatrix}\right). \tag{34}$$

By Eq. (34), first order conditions (23) can be rewritten as

$$\begin{aligned} &\mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \mathbb{E}y & \mathbb{E}z \\ \mathbb{E}y & \mathbb{E}y^2 & \mathbb{E}yz \\ \mathbb{E}z & \mathbb{E}yz & \mathbb{E}z^2 \end{pmatrix}\right) \\ &+ \frac{1}{\sqrt{T}}\begin{pmatrix} \sum_t \mathbb{E}\left(\operatorname{sgn}\left(v_{t+1} - \Delta\alpha - \Delta\beta y_t - \Delta\gamma z_t\right)|\mathcal{F}_t\right) \\ \sum_t \mathbb{E}\left(y_t\operatorname{sgn}\left(v_{t+1} - \Delta\alpha - \Delta\beta y_t - \Delta\gamma z_t\right)|\mathcal{F}_t\right) \\ \sum_t \mathbb{E}\left(z_t\operatorname{sgn}\left(v_{t+1} - \Delta\alpha - \Delta\beta y_t - \Delta\gamma z_t\right)|\mathcal{F}_t\right) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}. \end{aligned} \tag{35}$$

Using approximations (27), (29), and (30) for the second term in Eq. (35) and taking the limit as $T \to \infty$ to approximate sums with expectations, we get

$$\mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \mathbb{E}y & \mathbb{E}z \\ \mathbb{E}y & \mathbb{E}y^2 & \mathbb{E}yz \\ \mathbb{E}z & \mathbb{E}yz & \mathbb{E}z^2 \end{pmatrix}\right) - 2f_u(0)\sqrt{T}\begin{pmatrix} \Delta\alpha + \Delta\beta\mathbb{E}y + \Delta\gamma\mathbb{E}z \\ \Delta\alpha\mathbb{E}y + \Delta\beta\mathbb{E}y^2 + \Delta\gamma\mathbb{E}yz \\ \Delta\alpha\mathbb{E}z + \Delta\beta\mathbb{E}yz + \Delta\gamma\mathbb{E}z^2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Thus,

$$\sqrt{T}\begin{pmatrix} \Delta\alpha \\ \Delta\beta \\ \Delta\gamma \end{pmatrix} = \sqrt{T}\begin{pmatrix} \hat{\alpha} - \alpha \\ \hat{\beta} - \beta \\ \hat{\gamma} - \gamma \end{pmatrix} \xrightarrow[T\to\infty]{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \frac{1}{4f_u^2(0)}\begin{pmatrix} 1 & \mathbb{E}y & \mathbb{E}z \\ \mathbb{E}y & \mathbb{E}y^2 & \mathbb{E}yz \\ \mathbb{E}z & \mathbb{E}yz & \mathbb{E}z^2 \end{pmatrix}^{-1}\right). \quad \square$$

Proof of Theorem 8. Let us first linearize expectations of the first order conditions in spirit of the proof of Theorem 5.

$$\sum_{t=1}^{T} \begin{pmatrix} 1 \\ y_{t-1} \\ z_{t-1} \end{pmatrix} \operatorname{sgn}\left(y_t - a - by_{t-1} - cz_{t-1}\right) \mathbf{1}(a + by_{t-1} + cz_{t-1} > 0) = 0.$$

Thus, we are left with expectations of the form

$$\begin{aligned} &\mathbb{E}_{y,z} \begin{pmatrix} 1 \\ y \\ z \end{pmatrix} (1 - 2F_{v|y,z}(\Delta\alpha + \Delta\beta y + \Delta\gamma z|y, z)) \\ &\qquad \cdot \mathbf{1}(\alpha + \beta y_{t-1} + \gamma z_{t-1} + \Delta\alpha + \Delta\beta y_{t-1} + \Delta\gamma z_{t-1} > 0) = 0. \end{aligned} \tag{36}$$

Instead of Taylor expansion, we are going to use the direct formula: $f(x+\Delta x)g(x+\Delta x) = f(x)g(x)+f(x)(g(x+\Delta x)-g(x))+g(x+\Delta x)(f(x+\Delta x)-f(x))$. Thus, the first expectation in Eq. (36) can be rewritten as

(37)

$$\begin{aligned} &\mathbb{E}_{y,z}(1 - 2F_{v|y,z}(0|y, z))\mathbf{1}(\alpha + \beta y_{t-1} + \gamma z_{t-1} > 0) \\ &- 2\mathbb{E}_{y,z} f_{0|y,z}(0|y, z)\mathbf{1}(\alpha + \beta y_{t-1} + \gamma z_{t-1} + \Delta\alpha + \Delta\beta y_{t-1} + \Delta\gamma z_{t-1} > 0)(\Delta\alpha + \Delta\beta y_{t-1} + \Delta\gamma z_{t-1}) \\ &+ \mathbb{E}_{y,z}(1 - 2F_{v|y,z}0|y, z))\Big(\mathbf{1}(\alpha + \beta y_{t-1} + \gamma z_{t-1} + \Delta\alpha + \Delta\beta y_{t-1} + \Delta\gamma z_{t-1} > 0) \\ &- \mathbf{1}(\alpha + \beta y_{t-1} + \gamma z_{t-1} > 0)\Big) \\ &\approx -2\mathbb{E}_{y,z} f_u(0)\mathbf{1}(\alpha + \beta y_{t-1} + \gamma z_{t-1} > 0)(\Delta\alpha + \Delta\beta y + \Delta\gamma z), \end{aligned}$$

where we used the fact that $F_{v|y,z}(0|y, z) = F_u(0) = 0.5$ and $f_{v|y,z}(0|y, z) = f_u(0)$ when $\alpha + \beta y_{t-1} + \gamma z_{t-1} > 0$, and continuity of $\mathbf{1}(\alpha + \beta y_{t-1} + \gamma z_{t-1} + \Delta\alpha + \Delta\beta y_{t-1} + \Delta\gamma z_{t-1} > 0)$ with respect to $\Delta\alpha + \Delta\beta y_{t-1} + \Delta\gamma z_{t-1} = 0$.

Similarly, the other two expectations in Eq. (36) can be rewritten as

$$-2\mathbb{E}_{y,z} f_u(0) y_{t-1}\mathbf{1}(\alpha + \beta y_{t-1} + \gamma z_{t-1} > 0)(\Delta\alpha + \Delta\beta y_{t-1} + \Delta\gamma z_{t-1})$$

and

$$-2\mathbb{E}_{y,z} f_u(0) z_{t-1}\mathbf{1}(\alpha + \beta y_{t-1} + \gamma z_{t-1} > 0)(\Delta\alpha + \Delta\beta y_{t-1} + \Delta\gamma z_{t-1}).$$

The rest of the proof follows the lines of Theorem 7: we add and subtract conditional expectations from each term in the first order conditions. We then apply the martingale CLT for the difference and use the linearizations above (Eq. (37) and two others) for the expectations. The only difference is that everything is multiplied by an indicator $\mathbf{1}(\alpha+\beta y+$

$\gamma z + \Delta\alpha + \Delta\beta y + \Delta\gamma z > 0$). Thus,

$$\sqrt{T}\begin{pmatrix}\Delta\alpha\\ \Delta\beta\\ \Delta\gamma\end{pmatrix} = \sqrt{T}\begin{pmatrix}\hat{\alpha}-\alpha\\ \hat{\beta}-\beta\\ \hat{\gamma}-\gamma\end{pmatrix}$$

$$\xrightarrow[T\to\infty]{d} \mathcal{N}\left(\begin{pmatrix}0\\0\\0\end{pmatrix}, \frac{1}{4f_u^2(0)}\left(\mathbb{E}\left[\begin{pmatrix}1 & y & z\\ y & y^2 & yz\\ z & yz & z^2\end{pmatrix}\mathbf{1}(\alpha+\beta y+\gamma z>0)\right]\right)^{-1}\right). \qquad \square$$

Proof of Theorem 9. Denote the LAD estimate as $(\hat{\alpha}, \hat{\beta})$. The estimate is the solution to minimization problem

$$\min_{a,b}\sum_t |y_{t+1} - [a+by_t]_+|\,. \tag{38}$$

The corresponding first order conditions are

$$\sum_t \operatorname{sgn}(y_{t+1}-a-by_t)\,\mathbf{1}(a+by_t>0) = 0, \tag{39}$$

$$\sum_t y_t \operatorname{sgn}(y_{t+1}-a-by_t)\,\mathbf{1}(a+by_t>0) = 0. \tag{40}$$

Note that formally the summations may not equal to zero, as they involve discrete increments. Thus, we need to find point, where the summations switch signs from minus to plus.

The proof differs depending on the behavior of y_t ($\beta > 1$ or $\beta = 1$, $\alpha > 0$ or $\beta = 1$, $\alpha = 0$).

Let us analyze Eq. (39). Define $\xi_t = \operatorname{sgn}(y_{t+1}-a-by_t)\,\mathbf{1}(a+by_t>0)$. Then Eq. (39) can be rewritten as

$$\sum_t \mathbb{E}(\xi_t|y_t) + \sum_t (\xi_t - \mathbb{E}(\xi_t|y_t))\,. \tag{41}$$

The second term, $\sum_t (\xi_t - \mathbb{E}(\xi_t|y_t))$ is of order $O(\sqrt{T})$. This follows from the fact that

$$\mathbb{E}\sum_t (\xi_t - \mathbb{E}(\xi_t|y_t)) = 0$$

and

$$\mathbb{E}\left(\sum_t (\xi_t - \mathbb{E}(\xi_t|y_t))\right)^2 = \sum_t \mathbb{E}\,(\xi_t - \mathbb{E}(\xi_t|y_t))^2 < const\cdot T,$$

as each $\xi_t \in \{-1, 0, 1\}$.

Define $\Delta\alpha = \hat{\alpha}-\alpha$, $\Delta\beta = \hat{\beta}-\beta$, $v_{t+1} = \max(u_{t+1}, -\alpha-\beta y_t)$. Then

$$\begin{aligned}\mathbb{E}(\xi_t|y_t) &= \mathbb{E}\left(\operatorname{sgn}(v_{t+1} - \Delta\alpha - \Delta\beta y_t)\mathbf{1}(\Delta\alpha + \Delta\beta y_t + \alpha + \beta y_t > 0)|y_t\right)\\ &= (1 - 2F_v(\Delta\alpha + \Delta\beta y_t))\mathbf{1}(\Delta\alpha + \Delta\beta y_t + \alpha + \beta y_t > 0).\end{aligned}$$

We want to linearize $\mathbb{E}(\xi_t|y_t)$. When $\beta > 1$ or $\alpha > 0, \beta = 1$ we know from Theorem A.2, that $y_t \xrightarrow{a.s.} \infty$. Thus, for T' large enough, starting from $t > T'$ we get $\alpha + \beta y_t \gg 0$, so that the indicator does not bind.

• Case I: $\beta > 1$.

Assume that $\Delta\beta y_t \approx 0$, so that we can linearize $\mathbb{E}(\xi_t|y_t)$ around $\Delta\alpha + \Delta\beta y_t = 0$. Because y_t goes to infinity, we cannot assume $\Delta\beta y_t \approx 0$ for all t. However, we can assume that $\Delta\beta y_t \approx 0$ for $t < T - \sqrt{T}$. At the end of the proof we will find solution to first order conditions, which indeed satisfies this assumption, and has $\Delta\beta y_T = O(1)$ (so that $\Delta\beta y_{T-\sqrt{T}} = o(1)$) and $\Delta\alpha = o(1)$.

Thus, for $t \in (T', T - \sqrt{T})$,

$$\begin{aligned}(1 - 2F_v(\Delta\alpha + \Delta\beta y_t))\mathbf{1}(\Delta\alpha + \Delta\beta y_t + \alpha + \beta y_t > 0) &= (1 - 2F_v(\Delta\alpha + \Delta\beta y_t))\\ &\approx -2f_u(0)(\Delta\alpha + \Delta\beta y_t),\end{aligned}$$

so that

$$\sum_t \mathbb{E}(\xi_t|y_t) \approx -2Tf_u(0)\Delta\alpha - 2f_u(0)\Delta\beta \sum_t y_t + O(\sqrt{T}),$$

where we again used the fact that $\xi_t \in \{-, 1, 0, 1\}$ to bound terms with $t \notin (T', T - \sqrt{T})$.

Therefore, Eq. (39) can be rewritten as

$$-2Tf_u(0)\Delta\alpha - 2f_u(0)\Delta\beta \sum_{t=T'}^{T-\sqrt{T}} y_t + O(\sqrt{T}) = 0. \tag{42}$$

Let us now analyze the second first order condition, Eq. (40).

Suppose that $\beta > 1$. Then terms corresponding to $t \approx T$ dominate the summation, as they have the largest y_{t+1}'s, and the indicator is no longer binding. By Lemma $S9$ in the Supplementary Marerial we know that almost surely there exists T_0 such that $y_{t+1} > \beta' y_t$ for some $\beta' \in (1, \beta)$ and any $t \geq T_0$. Thus, $y_{T-k} \leq \beta'^{-k} y_T$. Simultaneously we assume $y_{T_0} > 1$ to avoid pathologies.

Choose an additional integer τ to be fixed later and split the sum in Eq. (40) into three:

$$\begin{aligned}\sum_{t=1}^{T_0} y_t \mathbf{1}(\hat{\alpha} + \hat{\beta} y_t > 0)\operatorname{sgn}\big(u_t - \Delta\alpha - \Delta\beta\big) + \sum_{t=T_0+1}^{T-\tau} y_t \operatorname{sgn}\big(u_t - \Delta\alpha - \Delta\beta y_t\big)\\ + \sum_{t=T-\tau+1}^{T} y_t \operatorname{sgn}\big(u_t - \Delta\alpha - \Delta\beta\big)\end{aligned} \tag{43}$$

The first sum is bounded by a (random) number M as $T \to \infty$. Now choose (random, independent from T) τ such that $(\beta')^\tau > \frac{2}{1-1/\beta'}$ and $(\beta')^\tau > 2M$. In this case whenever all the signs in the third sum are positive, (43) is positive, and whenever all the signs in the third sum are negative, (43) is negative. Indeed, the first sum is bounded by M, and the very last term with $t = T$ is at least twice larger due to our choice of τ and the bound $y_{T_0} > 1$ that we started from. Since y_t grows faster than geometric series with denominator β' for $t > T_0$, the second sum can be bounded from above by $y_{T-\tau}$ multiplied by geometric series with denominator $1/\beta'$. Hence, by our choice of τ and inequality $y_T > (\beta')^\tau y_{T-\tau}$, the last term with $t = T$ is at least twice as large as the sum.

The conclusion is that the value of $\Delta\beta$ lies between the maximum value of b which makes all $\operatorname{sgn}\big(u_t - \Delta\alpha - by_t\big)$ positive for $t = T-\tau+1, \dots, T$, and the minimum value of b which makes all $\operatorname{sgn}\big(u_t - \Delta\alpha - by_t\big)$ negative for $t = T-\tau+1, \dots, T$. Thus, looking at the points, where $\operatorname{sgn}\big(u_t - \Delta\alpha - by_t\big)$ changes, we obtain

$$\min_{t=T-\tau-1,\dots,T} \frac{u_t - \Delta\alpha}{y_t} \le \Delta\beta \le \max_{t=T-\tau-1,\dots,T} \frac{u_t - \Delta\alpha}{y_t}. \tag{44}$$

Expressing $\Delta\beta$ via $\Delta\alpha$ using (42) and plugging into (44), we get

$$\frac{1}{T}\left(\sum_{t=1}^{T-\sqrt{T}} y_t\right) \min_{t=T-\tau-1,\dots,T} \frac{u_t - \Delta\alpha}{y_t} \le O\left(T^{-\frac{1}{2}}\right) - \Delta\alpha \le \frac{1}{T}\left(\sum_{t=1}^{T-\sqrt{T}} y_t\right) \max_{t=T-\tau-1,\dots,T} \frac{u_t - \Delta\alpha}{y_t}. \tag{45}$$

The last inequality implies that $\Delta\alpha \to 0$ almost surely as $T \to \infty$, since $\frac{\sum_{t=T'}^{T-\sqrt{T}} y_t}{y_{t'}}$ stays bounded as $T \to \infty$ for all $t' = T-\tau+1, \dots, T$ (the numerator is at most $y_{T-\sqrt{T}}$ multiplied by a geometric series with denominator $1/\beta'$ and the denominator is at least $y_{T-\sqrt{T}}\beta'^{\sqrt{T}-\tau}$, so that their ratio is bounded by $1/(1-\beta'^{-1})$). In fact, we see that the speed of decay is $\frac{1}{T}$. Since $\Delta\alpha \to 0$, it stays bounded, and therefore, (44) implies that $\Delta\beta \to 0$ as fast as $1/y_T$, i.e. exponentially fast.

Finally, the solution to the first order conditions is a global minimum, not a local one, because y_t grows exponentially fast in t ($y_{t+1} > \beta' y_t$). Thus, for all large t, $\max(a + by_t, 0) = a + by_t$. This implies that most of the terms in (38) are convex functions of a and b (if the parameters are taken in a compact set), and therefore, the solution to the minimization problem can be found (up to a small error) as a point satisfying the first order conditions.

• Case II: $\beta = 1$, $\alpha > 0$.

In this case we do not need to consider separately observations with $t > T - \sqrt{T}$, and we will find solution to first order conditions, which has $\Delta\beta y_T = o(1)$ and $\Delta\alpha = o(1)$. Thus,

for $t > T'$,

$$(1 - 2F_v(\Delta\alpha + \Delta\beta y_t))\mathbf{1}(\Delta\alpha + \Delta\beta y_t + \alpha + \beta y_t > 0) = (1 - 2F_v(\Delta\alpha + \Delta\beta y_t)) \approx -2f_u(0)(\Delta\alpha + \Delta\beta y_t),$$

and Eq. (39) can be rewritten as

$$-2Tf_u(0)\Delta\alpha - 2f_u(0)\Delta\beta \sum_{t=T'}^{T} y_t + O(\sqrt{T}) = 0. \tag{46}$$

Let us now analyze the second first order condition, Eq. (40). Define

$$\eta_t = y_t \mathrm{sgn}(y_{t+1} - (\hat{\alpha} + \hat{\beta} y_t)).$$

Then for $t > T'$ the sum in Eq. (40) can be rewritten as

$$\sum_t \mathbb{E}(\eta_t | y_t) + \sum_t (\eta_t - \mathbb{E}(\eta_t | y_t)),$$

where $\mathbb{E}\sum_t(\eta_t - \mathbb{E}(\eta_t|y_t)) = 0$ and

$$\mathbb{E}\left(\sum_t (\eta_t - \mathbb{E}(\eta_t|y_t))\right)^2 = \mathbb{E}\sum_t (\eta_t - \mathbb{E}(\eta_t|y_t))^2 = \mathbb{E}\sum_t \left(\eta_t^2 - (\mathbb{E}(\eta_t|y_t))^2\right)$$
$$= \mathbb{E}\sum_t y_t^2 \left(1 - \left(\mathbb{E}(\mathrm{sgn}(y_{t+1} - (\hat{\alpha} + \hat{\beta} y_t))|y_t)\right)^2\right) \leq \mathbb{E}\sum_t y_t^2,$$

because $\mathrm{sgn} \in [-1, 1]$.

Random variable y_t grows linearly in t (i.e. $y_T = O(T)$). To see this, note that $y_t \geq \alpha + y_{t-1} + u_t \geq \alpha t + y_0 + \sum_{s=1}^{t} u_s$ and $y_t \leq \alpha + y_{t-1} + |u_t| \leq \alpha t + y_0 + \sum_{s=1}^{t} |u_s|$, so that

$$\alpha \xleftarrow[t\to\infty]{} \alpha + y_0/t + \frac{1}{t}\sum_{s=1}^{t} u_s \leq y_t/t \leq \alpha + y_0/t + \frac{1}{t}\sum_{s=1}^{t} |u_s| \xrightarrow[t\to\infty]{} \alpha + \mathbb{E}|u_t|.$$

Thus, $\mathbb{E}\sum_t y_t^2$ is of order $\sum_t t^2$, so that $\mathbb{E}\sum_t y_t^2 = O(T^3)$ and $\sum_t(\eta_t - \mathbb{E}(\eta_t|y_t)) = O(T^{3/2})$.

Define $v_{t+1} = \max(u_{t+1}, -\alpha - \beta y_t)$. We are left with analyzing

$$\sum_t \mathbb{E}(\eta_t|y_t) = \sum_t y_t \mathbb{E}(\mathrm{sgn}(y_{t+1} - (\hat{\alpha} + \hat{\beta} y_t))|y_t) = \sum_t y_t \mathbb{E}(\mathrm{sgn}(v_{t+1} - \Delta\alpha - \Delta\beta y_t)|y_t)$$
$$= \sum_t y_t (1 - 2F_{v_{t+1}|y_t}(-\Delta\alpha - \Delta\beta y_t)).$$

Taylor expanding $1 - 2F_{v_{t+1}|y_t}(-\Delta\alpha - \Delta\beta y_t)$ around $\Delta\alpha + \Delta\beta y_t = 0$ and using the fact that $F_{v|y}(0) = F_u(0) = 0.5$, $f_{v|y}(0) = f_u(u)$ as for $t > T'$ $\alpha + \beta y_t > 0$, we get

$$\sum_t \mathbb{E}(\eta_t|y_t) = -2f_u(0)\sum_t y_t(\Delta\alpha + \Delta\beta y_t + o(\Delta\alpha + \Delta\beta y_t)),$$

so that Eq. (40) can be rewritten as

$$\Delta\alpha \sum_{t=T'}^{T} y_t + \Delta\beta \sum_{t=T'}^{T} y_t^2 = O(T^{3/2}). \tag{47}$$

Solving Eq. (46) and (47), and using the fact that y_t grows linearly in T so that $\sum\limits_t y_t = O(\sum\limits_t t) = O(T^2)$, we get $\Delta\alpha = O(T^{-0.5})$, $\Delta\beta = O(T^{-1.5})$. Thus, $\Delta\alpha \xrightarrow[T\to\infty]{} 0$, $\Delta\beta \xrightarrow[T\to\infty]{} 0$, and the LAD estimator is consistent. Similarly to the Case I, the solution to the first order conditions is a global minimum because y_t grows linearly in t. Thus, for all large t, $\max(a + by_t, 0) = a + by_t$. This implies that most of the terms in (38) are convex functions of a and b (if the parameters are taken in a compact set), and therefore, the solution to the minimization problem can be found (up to a small error) as a point satisfying the first order conditions.

• Case III: $\beta = 1$, $\alpha = 0$.

Let us calculate the order of terms, where indicator binds. By Theorem A.5, $\frac{1}{\sqrt{T}} y_{\lfloor Ts \rfloor} \to \sigma|W(s)|$ for any $s \in (0, 1]$, where $\sigma = \mathbb{E}u_t^2$ and W is a standard Brownian motion. Thus, the indicator can be rewritten as $\mathbf{1}\left(b\frac{y_{\lfloor Ts \rfloor}}{\sqrt{T}} > -\frac{a}{\sqrt{T}}\right)$.

As T goes to infinity, $\frac{a}{\sqrt{T}}$ goes to zero and $b\frac{y_{\lfloor Ts \rfloor}}{\sqrt{T}}$ goes to $b\sigma|W(s)|$. The (random) time Brownian motion spends inside interval $[-\varepsilon, \varepsilon]$ goes to zero almost surely as $\varepsilon \to 0$ (see Section 3.6 in Karatzas and Shreve (2012)). Thus, for $b > 0$ the time $b\sigma|W(s)|$ is smaller than $-\frac{a}{\sqrt{T}}$ goes to zero as $T \to \infty$, i.e. it is $o(1)$. Therefore, $\sharp\{t : b\frac{y_{\lfloor Ts \rfloor}}{\sqrt{T}} \le -\frac{a}{\sqrt{T}}\} = o(T)$.

Note also that $b \le 0$ cannot solve the minimization problem (38). When $b < 0$, $b\sigma|W(s)| \le 0$ and the indicator starts to bind all the time, so that we get $\sum\limits_t |y_{t+1}| = O(T^{3/2})$, as $\frac{1}{T\sqrt{T}} \sum\limits_t |y_{t+1}| \to \sigma \int\limits_0^1 |W(s)|ds$. Yet, $\sum\limits_t |y_{t+1} - y_t| = \sum\limits_t |u_t| = O(T)$. When $b = 0$, we again get $\sum\limits_t |y_{t+1} - [a]_+| = O(T^{3/2})$ as $\frac{1}{T\sqrt{T}} \sum\limits_t |y_{t+1} - [a]_+| = \frac{1}{T} \sum\limits_t |y_{t+1}/\sqrt{T} - [a]_+/\sqrt{T}| \to \sigma \int\limits_0^1 |W(s)|ds$.

Thus, we can linearize Eq. (39) to get

$$-2Tf_u(0)\Delta\alpha - 2f_u(0)\Delta\beta \sum_t y_t + o(T) + O(\sqrt{T}) = 0. \tag{48}$$

To analyze the second first order condition, Eq. (40), we proceed as in Case II. The only difference is that now we use Theorem A.5 to calculate the order of different sums involving y_t. That is, $\frac{1}{T^2}\mathbb{E}\sum\limits_t y_t^2 \xrightarrow[T\to\infty]{} \mathbb{E}\sigma^2 \int_0^1 W^2(s)ds$ implying $\mathbb{E}\sum\limits_t y_t^2 = O(T^2)$ and $\frac{1}{T^{3/2}}\mathbb{E}\sum\limits_t y_t \xrightarrow[T\to\infty]{} \mathbb{E}\sigma \int_0^1 |W(s)|ds$ implying $\mathbb{E}\sum\limits_t y_t = O(T^{3/2})$. Thus, the second first order condition, Eq. (40) can be rewritten as

$$\Delta\alpha O(T^{3/2}) + \Delta\beta O(T^2) + o(T^{3/2}) = O(T). \tag{49}$$

Solving Eq. (48) and (49), we get $\Delta\alpha = o(1)$, $\Delta\beta = o(T^{-0.5})$. Thus, $\Delta\alpha \xrightarrow[T\to\infty]{} 0$, $\Delta\beta \xrightarrow[T\to\infty]{} 0$, and the LAD estimator is consistent. The solution to the first order conditions is a global minimum because the indicator in the maximization problem binds in $o(T)$ terms. This implies that the dominant terms in (38) are convex functions of a and b (if the parameters are taken in a compact set), and therefore, the solution to the minimization problem can be found (up to a small error) as a point satisfying the first order conditions. □

References

Amemiya, T., "Tobit models: A survey," *Journal of Econometrics*, 1984, *24* (1–2), 3–61.

Anderson, T.W., "On Asymptotic Distributions of Estimates of Parameters of Stochastic Difference Equations," *The Annals of Mathematical Statistics*, 1959, *30* (3), 676–678.

Andrews, D.W.K., "Non-Strong Mixing Autoregressive Processes," *Journal of Applied Probability*, 1984, *21* (4), 930–934.

Blume, L.E., W.A. Brock, S.N. Durlauf, and Y.M. Ioannides, *Identification of social interactions*, Vol. 1, North-Holland, 2011. In Handbook of social economics.

Bramoullé, Y., A. Galeotti, and B. Rogers, eds, *The Oxford handbook of the economics of networks*, Oxford University Press, 2016.

Chamberlain, G., "Asymptotic efficiency in semi-parametric models with censoring," *Journal of Econometrics*, 1986, *32* (2), 189–218.

Diebold, F.X. and K. Yılmaz, "On the network topology of variance decompositions: Measuring the connectedness of financial firms," *Journal of Econometrics*, 2014, *182* (1), 119–134.

_____ **and** _____, *Financial and macroeconomic connectedness: A network approach to measurement and monitoring*, Oxford University Press, 2015.

_____ **and R.S. Mariano**, "Comparing predictive accuracy," *Journal of Business & economic statistics*, 1995, *13* (3), 253–63.

Dueñas, M. and G. Fagiolo, "Modeling the international-trade network: a gravity approach," *Journal of Economic Interaction and Coordination*, 2013, *8* (1), 155–178.

European Commission, "Eurostat," `https://ec.europa.eu/eurostat/` accessed May 1, 2019.

Feller, W., *An introduction to probability theory and its applications*, Vol. 2, John Wiley & Sons, 2008.

Gambardella, A., L. Orsenigo, and F. Pammolli, "Global competitiveness in pharmaceuticals: a European perspective," *MPRA Paper No.* 15965, 2000.

Gourieroux, C. and A. Monfort, "Sufficient linear structures: Econometric applications," *Econometrica*, 1980, *48* (5), 1083–1097.

Graham, B.S., "Homophily and transitivity in dynamic network formation," *Working paper*, 2016.

_____, "An econometric model of network formation with degree heterogeneity," *Econometrica*, 2017, *85* (4), 1033–1063.

_____, "Dyadic regression," *Working paper*, 2019.

Hahn, J. and G. Kuersteiner, "Stationarity and mixing properties of the dynamic Tobit model," *Economics Letters*, 2010, *107* (2), 105–111.

Hayashi, F., *Econometrics*, Princeton University Press, 2000.

Holme, P. and J. Saramäki, "Temporal networks," *Physics reports*, 2012, *519* (3), 97–125.

Jackson, M.O., *Social and economic networks*, Princeton University Press, 2010.

Jong, R. and A.M. Herrera, "Dynamic censored regression and the Open Market Desk reaction function," *Journal of Business & Economic Statistics*, 2011, *29* (2), 228–237.

Karatzas, I. and S. Shreve, *Brownian motion and stochastic calculus*, Springer Science & Business Media, 2012.

Khan, S. and J.L. Powell, "Two-step estimation of semiparametric censored regression models," *Journal of Econometrics*, 2001, *103* (1–2), 73–110.

Leung, M.P. and H.R. Moon, "Normal Approximation in Large Network Models," *arXiv preprint arXiv:*1904.11060, 2019.

Lévy, P., *Processus Stochastiques et Mouvement Brownien*, Gauthier-Villars,Paris, 1948.

Michel, J. and R.M. de Jong, "Mixing properties of the dynamic Tobit model with mixing errors," *Economics Letters*, 2018, *162*, 112–115.

Palla, G., A.L. Barabási, and T. Vicsek, "Quantifying social group evolution," *Nature*, 2007, *446* (7136), 664–667.

Phillips, P.C.B., "Towards a unified asymptotic theory for autoregression," *Biometrika*, 1987, *74* (3), 535–547.

Pin, P. and B.W. Rogers, "Cooperation, punishment and immigration," *Journal of Economic Theory*, 2015, *160*, 72–101.

Powell, J.L., "Least absolute deviations estimation for the censored regression model," *Journal of Econometrics*, 1984, *25* (3), 303–325.

Robinson, D.T. and T.E. Stuart, "Network effects in the governance of strategic alliances," *The Journal of Law, Economics, & Organization*, 2006, *23* (1), 242–273.

Wang, X. and J. Yu, "Limit theory for an explosive autoregressive process," *Economics Letters*, 2015, *126*, 176–180.

Wei, S.X., "A Bayesian approach to dynamic Tobit models," *Econometric Reviews*, 1999, *18* (4), 417–439.

White, J.S., "The Limiting Distribution of the Serial Correlation Coefficient in the Explosive Case," *The Annals of Mathematical Statistics*, 1958, *29* (4), 1188–1197.

Withers, C.S., "Central limit theorems for dependent variables. *I*," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 1981, *57* (4), 509–534.

(Anna Bykhovskaya) University of Wisconsin-Madison
E-mail address: anna.bykhovskaya@wisc.edu