Finite sample behaviour of MLE in network autocorrelation models

Michele La Rocca, Giovanni C. Porzio, Maria Prosperina Vitale and Patrick Doreian

Abstract This work evaluates the finite sample behaviour of ML estimators in network autocorrelation models, a class of auto-regressive models studying the network effect on a variable of interest. Through an extensive simulation study, we examine the conditions under which these estimators are normally distributed in the case of finite samples. The ML estimators of the autocorrelation parameter have a negative bias and a strongly asymmetric sampling distribution, especially for high values of the network effect size and the network density. In contrast, the estimator of the intercept is positively biased but with an asymmetric sampling distribution. Estimators of the other regression parameters are unbiased, with heavy tails in presence of non-normal errors. This occurs not only in randomly generated networks but also in well-established network structures.

Key words: network effect model, density, network topology, non-normal distribution.

1 Introduction

Network autocorrelation models (NAMs, 3, 4, 5) deal with the presence of individual units embedded within social structures. They represent a class of auto-

Michele La Rocca, Maria Prosperina Vitale
Dept. of Economics and Statistics, University of Salerno (Italy), email: larocca@unisa.it, mvitale@unisa.it

Giovanni C. Porzio
Dept. of Economics and Law, University of Cassino and Southern Lazio (Italy), email: porzio@unicas.it

Patrick Doreian
Faculty of Social Sciences, University of Ljubljana (Slovenia); Dept. of Sociology, University of Pittsburgh (US), email: pitpat@pitt.edu

regressive models used to study the effect of a network on an outcome variable of interest when the data points are interdependent. Specifically, we can refer to the "social influence" (or contagion) mechanism in which the social relations among individuals provide a foundation for predicting actor behaviors given the behaviours of other actors in the network in which they are embedded (12).

Among the models proposed in the literature to address social influence on individual behavior, NAMs propose an approach dealing with, simultaneously, network effects and individual attributes. Nevertheless despite the clear advantages over other conventional approaches, it is known that in these models the estimated autocorrelation parameter has a finite sample negative bias, the amount of which is positively related with the network density (13, 14).

Our contribution aims at describing the *whole* finite sample distribution of the Maximum Likelihood Estimators (MLE) of the autocorrelation and regression parameters. Through an extensive simulation study, we investigate the conditions under which, MLEs are normally distributed in case of the finite samples. The finite sample distributions are evaluated with respect to the network density and topology, the distribution of error terms, and the strength of the autocorrelation parameter (i.e., the network effect size).

We focus on three research questions:

- *What is the whole sampling distribution of the network effect estimator*?
- *What are the finite sample distributions of the regression coefficient estimators?*
- *What are the consequences of the errors not being normally distributed?*

The remaining of the paper is organized as follows. Section 2 presents a brief review on NAMs. The Monte Carlo simulation study used to deal with the aforementioned research questions is described in Section 3. Section 4 reports the main results, while Section 5 concludes with some final remarks.

2 A brief review of network autocorrelation models

Two types of network autocorrelation models are available within the literature (3): the network effects model and the network disturbances model. In the first case, interdependencies between actors are modelled through the inclusion of an autocorrelation parameter in the dependent term, while in the second case interdependencies are included in the disturbance term. Here the focus is on the network effects model which allows individual outcome to be directly associated with neighbours' levels of outcome by including the network effect as a weight matrix (11).

More formally, let $\mathbf{y}$ be a $(n \times 1)$ vector of values of a dependent (endogenous) variable for n individuals making up a network, let $\mathbf{X}$ represent the $(n \times p)$ matrix of values for the n individuals on p covariates (including a unit vector for the intercept term), and let $\mathbf{W}$ be the $(n \times n)$ network weight matrix whose elements, w_{ij}, measure the influence actor j has on actor i. The network effects model is defined as:

$$\mathbf{y} = \rho \mathbf{W}\mathbf{y} + \mathbf{X}\beta + \varepsilon$$

where ρ is the network autocorrelation parameter referred as the strength of the social influence mechanism in a network, β is a $(p \times 1)$ vector of regression parameters, and the error terms ε are assumed to be independently normally distributed with zero means and equal variances, $\varepsilon \sim N(0, \sigma^2\mathbf{I})$.

This class of models represents a popular tool for conducting social network analysis. First adopted to describe social influence mechanism (12), it has been recently applied in many social science fields (see e.g. 2, 7, 9, 15) and it has been extended to the study of multiple networks (6, 16) and to the presence of two-mode networks (10).

From a methodological point of view, recent contributions focused on the bias of the MLE of the network autocorrelation parameter ρ. They discovered a systematic negative bias, whose magnitude increases with the network density (13). In addition, it has been found that this bias does not depend on network size, numbers of exogenous variables in the model, and whether the network weight matrix $\mathbf{W}$ was normalized or reported in raw form. The bias also does not depend on the presence of well-established network structures (e.g., scale-free and small world configurations), although it is especially pronounced at extremely low-density levels in the star network (14).

Recently, rather than look for more conditions in which network autocorrelation parameter is underestimated, Wang *et al*. (17) investigated the likelihood of identifying a statistically significant network effect. They show that NAM well controls for Type I error rates, that the statistical power is a nonlinear function of ρ and of the network size, and that network density and structured networks have little impact on statistical power. With respect to this latter aspect, Faber *et al*. (8) showed that the average degree of a random network impacts the power of tests.

However, as highlighted within the introduction, knowing the full sampling distributions of MLE estimates is needed. This is the focus of our study. Its design and results are described in the next sections.

3 Simulation design

An extensive Monte Carlo study (5000 MC replications) was used to assess the whole finite sampling distribution of MLEs. To accomplish this, the following conditions have been varied: *i)* the network density (Δ), *ii)* the network autocorrelation parameter (ρ), *iii)* the network topology ($\mathbf{W}$), and *iv)* the error distributions (ε).

Two covariates were considered, and data were generated according to the following network autocorrelation model:

$$y_i = \rho W y_i + \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, \qquad i = 1, 2 \ldots, n.$$

Corresponding to the simulation design used by (13, 14), the elements in the simulation scheme were set as follows. Covariates were randomly generated according to a standard normal distribution; without loss of generality, all the β's were set equal to the same constant value ($\beta_0 = \beta_1 = \beta_2 = 2$); only positive values of ρ were considered ($0.00 \leq \rho \leq 0.90$), accounting for low to high network effect size. The error were independently randomly generated with constant variance ($Var[\varepsilon_i] = 1$) according to three different schemes. First, as is done usually in such studies, errors were standard normally generated. Second, to consider the effect of asymmetric errors, a standard log-normal was considered. Finally, to consider a completely non-standard distribution, the error term was derived by generating data from an equal weight mixture of distributions of a (centered) chi-square with one degree of freedom and a Student's t with 4 degrees of freedom.

We consider a sample size of 50 nodes. The network weight matrix $\mathbf{W}$ was row normalized, randomly generated at each run. The network density Δ took values $0.05 \leq \Delta \leq 0.80$. Beyond the Erdos-Renyi random graphs (E-Rs) adopted as baseline model, two other kinds of topologies were considered to assess the evidence of non-random behaviours in the formation of network ties. Specifically, the scale-free (1) and the small world (18) network configurations where taken into account so that the effect of well-established network structure can be evaluated, as reported in (14). In the first case, preferential attachment defines the tie formation mechanism. This mechanism accounts for the tendency to be linked with the best connected nodes (i.e., nodes with the highest degree). Hence a scale-free structure emerges when nodes degree distribution follows a power law distribution, and a "star" network structure appears (i.e. one node is linked to all the others and no other connections are present among the remaining nodes). The small-world configuration presents instead a high node connectivity with low average distance among regions of the network. More specifically, the concurrent presence of dense local clustering (measured by clustering coefficient) with short network distances (measured by the average path length) is observed.

4 Results

We examine the observed sampling distributions of (r-ρ), (b_0-β_0), (b_k-β_k), where r, b_0 and b_k are the ML estimators of ρ, β_0, and β_k, respectively, (k= 1,2). First, somewhat to our surprise, all three of the examined network structures provided substantially the same results. One implication is that the operation of network effects in NAMs do not rest on the global structure of the network. For this reason, only results for the E-R random graph are discussed further.

A series of parallel boxplots are reported in each picture illustrating the ML finite distribution for different levels of network effects size, densities, and error distributions.

The sampling distributions obtained for (r-ρ) are shown in Figure 1. Each frame in the figure corresponds to simulation results obtained for a fixed value of the pop-

ulation parameter ρ (ρ= 0, 0.1, 0.2,. . . , 0.9). Boxplots show results with respect to five level of density (horizontal axes, Δ= 0.05,. . . , 0.8), and three error distributions (normal, lognormal, mixture, in this order).

As expected, sampling distributions are negatively biased, a result increasing with ρ and Δ. Normality does not seem to hold: for low values of ρ and Δ, heavy tails appear. For higher values of both parameters, distributions are quite strongly asymmetric. Finally, it seems that differences in the error distributions have a minor effect on the resulting estimator distributions.

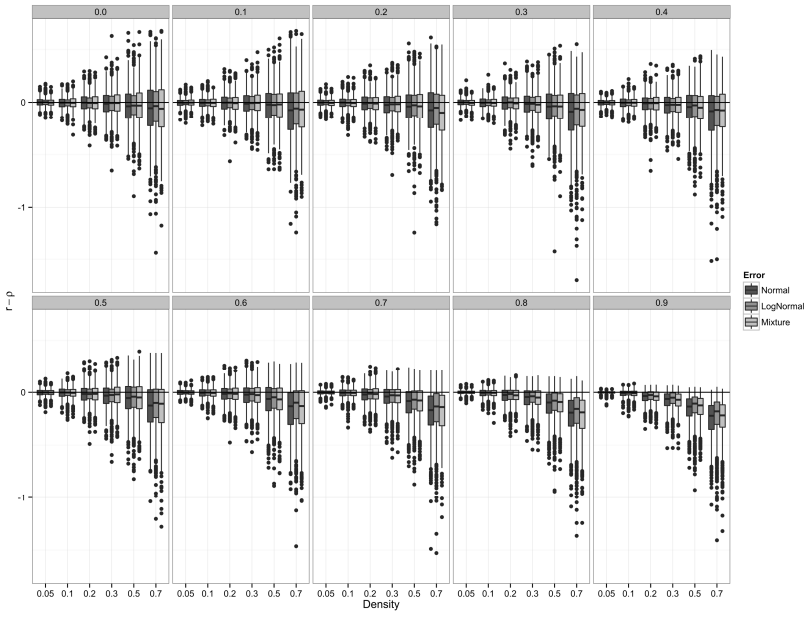

Fig. 1 Boxplots of the sampling distribution r-ρ (vertical axis) for E-R random graphs. The density values are reported on the horizontal axis. Results are reported for different values of ρ.

Adopting the same graphical structure, results obtained for the sampling distributions of the regression coefficient estimators are reported in Figures 2 and 3 [for $(b_0 - \beta_0)$ and $(b_1 - \beta_1)$, respectively]. Results for the sampling distributions of $(b_2 - \beta_2)$ are not reported as they are the same to those observed for $(b_1 - \beta_1)$.

As for the regression coefficient estimators, some different behaviours arise. The intercept estimator distributions mirrors the distribution of the autocorrelation parameter estimator: it is positively biased, with such a bias increasing with ρ and Δ, with analogous effects in terms of asymmetries and heavy tails. On the other hand, results suggest that the estimators of the other regression coefficients are unbiased, with heavy tails in the presence of non-normal errors.

Overall, it seems that where non-normality of the estimator distributions arises as a consequence of a certain degree of autocorrelation and density, this effect over-

whelms the effect due to non-normality of the errors. However, where distributions are unbiased and normal, non-normality of the errors plays a more substantial role. To conclude, ML estimators of the autocorrelation parameter and of the intercept are not normally distributed in case of small sample size, even in presence of normally distributed errors. Furthermore, the network density has some effect on the variability of the estimators. On the other hand, it seems that other features of the network topologies, in the main, have little effects.

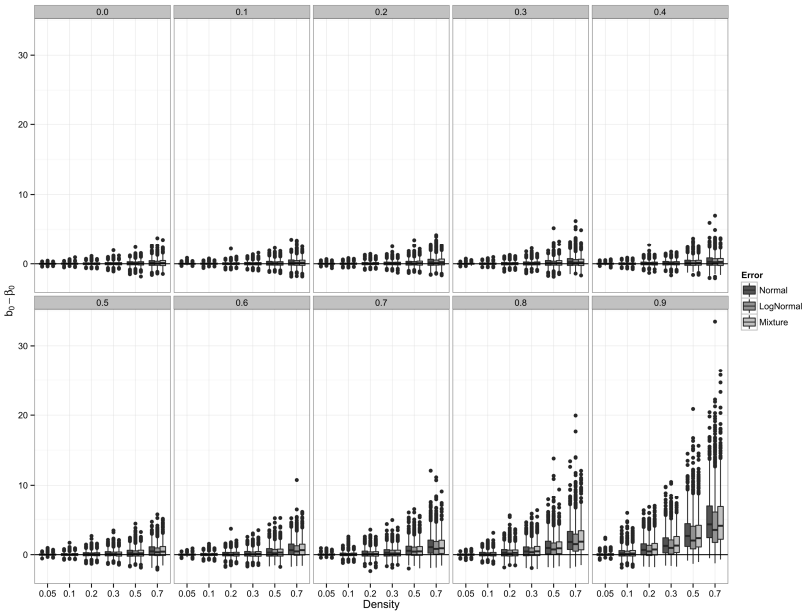

Fig. 2 Boxplots of the sampling distribution of $b_0 - \beta_0$ (vertical axis) for E-R random graphs. The density values are reported on the horizontal axis. Results are reported for different values of ρ.

5 Discussion and conclusions

The present contribution has shown that the ML estimator of ρ in NAMs not only contains a systematic negative bias, as expected, but also that its distribution is typically non-normal and asymmetric.

According to our results, for high values of the autocorrelation parameter ρ and network density Δ, the sampling distribution of the autocorrelation parameter is negatively biased and quite strongly asymmetric. On the other hand, the sampling distribution of regression coefficients is positively biased and asymmetric for the estimator of the intercept, and unbiased and with heavy tails in presence of non-

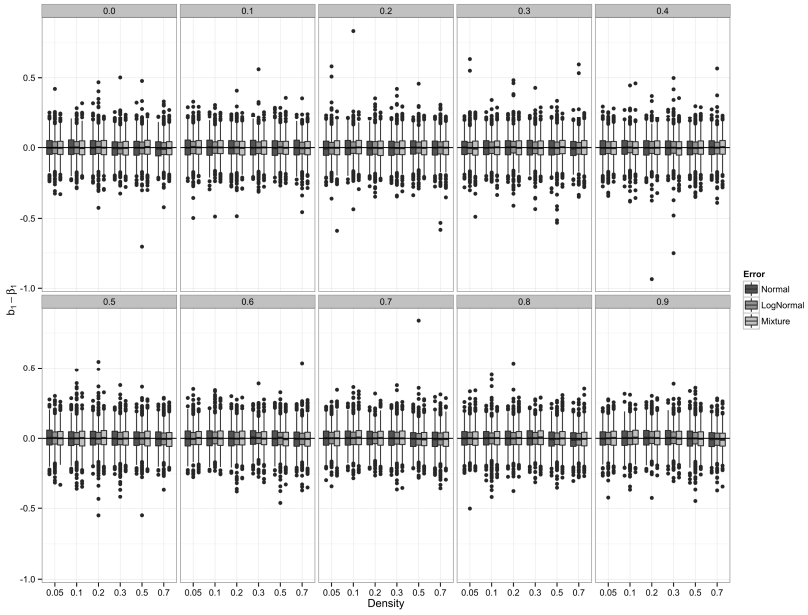

Fig. 3 Boxplots of the sampling distribution of $b_1 - \beta_1$ (vertical axis) for E-R random graphs. The density values are reported on the horizontal axis. Results are reported for different values of ρ.

normal errors for the other regression coefficients. This occurs not only in randomly generated networks but in well-established network structures as well.

Furthermore, the non-normality and the asymmetry is not confined strictly to networks with high density. At least in small world networks, these features exist also for very low levels of density. This suggests will be worthwhile to extend this analysis in order to study the performances of other related estimation tools, such as the finite sample confidence intervals built on the corresponding asymptotic theory. The authors intend to report on that issue elsewhere.

References

[1] Albert, R., Barabási, A.-L.: Statistical mechanics of complex networks. Rev. Mod. Phys. **74**, 47-97 (2002)

[2] De Nooy, W.: Communication in natural resource management: agreement between and disagreement within stakeholder groups. Ecol. Soc. 18(2): 44 (2013) Available via DIALOG. http://dx.doi.org/10.5751/ES-05648-180244

[3] Doreian, P.: Linear models with spatially distributed data: spatial disturbances or spatial effects? Sociol. Methods Res. **9**, 29-60 (1980)

[4] Doreian, P., Teuter, K., Wang, C.H.: Network autocorrelation models: some Monte Carlo results. Sociol. Methods Res. **13**, 155-200 (1984)
[5] Dow, M.M., Burton, M.L., White, D.R.: Network autocorrelation: a simulation study of a foundational problem in regression and survey research. Soc. Networks **4**, 169-200 (1982)
[6] Dow, M.M.: Galton's problem as multiple network autocorrelation effects cultural trait transmission and ecological constraint. Cross Cult. Res. **41**, 336-363 (2007)
[7] Dow, M.M., Eff, E.A.: Global, regional, and local network autocorrelation in the standard cross-cultural sample. Cross Cult. Res. **42**, 148-171 (2008)
[8] Farber, S., Páez, A., Volz, E.: Topology and dependency tests in spatial and network autoregressive models. Geogr. Anal. **41**, 158-180 (2009)
[9] Franzese Jr, R.J., Hays, J.C., Kachi, A., Alvarez, R.M., Freeman, J.R., Jackson, J.E.: Modeling history dependence in network-behavior coevolution. Polit. Anal. **20**, 175-190 (2012)
[10] Fujimoto, K., Chou, C.P., Valente, T.W.: The network autocorrelation model using two-mode data: Affiliation exposure and potential bias in the autocorrelation parameter. Soc. networks **33**, 231-243 (2011)
[11] Leenders, R.T.A: Modeling social influence through network autocorrelation: constructing the weight matrix, Soc. Networks **24**, 21-47 (2002)
[12] Marsden, P.V., Friedkin, N.E.: Network Studies of Social Influence. Sociol. Methods Res. **22**, 127-151 (1993)
[13] Mizruchi M.S., Neuman, E.J.: The effect of density on the level of bias in the network autocorrelation model. Soc. Networks **30**, 190–200 (2008)
[14] Neuman, E.J., Mizruchi, M.S.: Structure and bias in the network autocorrelation model. Soc. Networks **32**, 290-300 (2010)
[15] Vitale, M.P., Porzio, G.C., Doreian, P.: Examining the effect of social influence on student performance through network autocorrelation models. J. Appl. Stat. **43**, 115-127 (2016)
[16] Zhang, B., Thomas, A.C., Doreian, P., Krackhardt, D., Krishnan, R.: Contrasting multiple social network autocorrelations for binary outcomes with applications to technology adoption. ACM T. Man. Inf. Syst. **3**, 1-21 (2013)
[17] Wang, W., Neuman, E.J., Newman, D.A.: Statistical power of the social network autocorrelation model. Soc. Networks **38**, 88-99 (2014)
[18] Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature **393**, 440-442 (1998)