INFERENCE ON SOCIAL EFFECTS WHEN THE NETWORK IS SPARSE AND UNKNOWN

ERIC GAUTIER AND CHRIS ROSE

We consider the following simultaneous equations model

$$\boldsymbol{Y}(\boldsymbol{A}-\boldsymbol{I}_N)+\sum_{k=1}^{K}\boldsymbol{X}_{(k)}(\boldsymbol{C}_{(k)}+\boldsymbol{\eta}_{(k)})+\boldsymbol{\alpha}'\otimes\mathbf{1}_T+\mathbf{1}_N'\otimes\boldsymbol{\lambda}+\boldsymbol{\epsilon}=\mathbf{0}\ , \tag{0.1}$$

where $\boldsymbol{Y}$ and $(\boldsymbol{X}_{(k)})_{k=1}^K$ are observed random matrices of dimension $T\times N$, $\boldsymbol{\epsilon}$ is an unobserved random matrix of dimension $T\times N$ of error terms, $\boldsymbol{I}_N$ is the $N\times N$ identity matrix, $\mathbf{1}_T$ is the $T\times 1$ vector where all entries are equal to 1, $'$ stands for the transpose and $\otimes$ for the Kronecker product. The parameters in this model are: $\boldsymbol{A}$ and $(\boldsymbol{C}_k)_{k=1}^K$ which are $N\times N$ matrices, $(\boldsymbol{\eta}_{(k)})_{k=1}^K$ are diagonal $N\times N$ matrices, $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$ which are respectively of dimension $N\times 1$ and $T\times 1$. This is a linear of model of social interactions. The matrices $\boldsymbol{A}$ and $(\boldsymbol{C}_{(k)})_{k=1}^K$ have zeros on the diagonal and account respectively for the endogenous effects and exogenous or contextual effects. N is the number of individuals in the network and T is the number of observations of the individuals interacting. Typically T is a number of time periods.

The matrix $\boldsymbol{A}$ accounts for endogenous social effect while the matrices $(\boldsymbol{C}_{(k)})_{k=1}^K$ account for exogenous social effects. Typical low dimensional parameters of interest in this model are

$$a=\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\boldsymbol{A}_{i,j}$$

$$\forall k=1,\dots,K,\ c_{(k)}=\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\boldsymbol{C}_{(k)i,j},\ \eta_{(k)}=\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\eta}_{(k)i,i}\ .$$

We maintain the assumption that the network is sparse, namely the matrices $\boldsymbol{A}$ and $(\boldsymbol{C}_{(k)})_{k=1}^K$ have many zeros but the identity of the zeros is unknown. This implies that the number of true nonzero parameters is small but because we don't know their identity the actual number of parameters is very

Keywords: Instrumental variables, endogeneity, panels, simultaneous equations, heterogeneity, peer effects, networks, *STIV* estimator, sensitivity characteristics, high-dimensional regression, convex programming, confidence sets, variable selection, unknown variance, sign consistency.

Eric Gautier acknowledges financial support from the grants ANR-13-BSH1-0004 and ERC POEMH.

large. Because of this, we propose a method that accommodates a large variety of shape restrictions that are often maintained in the literature.

The parameter $\boldsymbol{\alpha}$ accounts for individual characteristics (observable or not) which are fixed over time. Like in a panel data model with fixed effects, we will "difference out" the term $\boldsymbol{\alpha}' \otimes \mathbf{1}_T$ which could therefore be random and arbitrarily dependent with $\boldsymbol{Y}$, $\boldsymbol{X}$, $\boldsymbol{\lambda}$, and $\boldsymbol{\epsilon}$.

The term $\mathbf{1}_N' \otimes \boldsymbol{\lambda}$ accounts for time varying effects. The vector $\boldsymbol{\lambda}$ can be of the form $\boldsymbol{\lambda} = \boldsymbol{Z}\boldsymbol{\delta}$ where $\boldsymbol{Z}$ is a $T \times L$ matrix such that the columns are known functions of time and observed variables which only vary with time. This second set of variables are typically attributes which are common to the individuals (*e.g.*, characteristics of a school or similar demographic characteristics). The vector $\boldsymbol{\delta}$ accounts for the so-called correlated effects (see, *e.g.*, Manski (1993)) which can be a parameter of interest. The vector $\boldsymbol{\lambda}$ can more generally be of the form $\boldsymbol{\lambda}_t = f(\boldsymbol{Z}_t, \boldsymbol{U}_t, t)$ where the function f is unknown and $\boldsymbol{U}_t$ is a vector of unobservable characteristics which are common to the individuals in the network. There, $\boldsymbol{U}$ can be arbitrarily correlated with $\boldsymbol{Z}$, t $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\epsilon}$, and we cannot recover $\boldsymbol{\lambda}$. The presence of $\mathbf{1}_N' \otimes \boldsymbol{\lambda}$ with unobservable common characteristics also impedes the identification of the endogenous and exogenous effects. We will therefore also difference out the term $\mathbf{1}_N' \otimes \boldsymbol{\lambda}$ in that case.

For these two cases we consider that the errors $(\boldsymbol{\epsilon}_t)_{t=1}^T$ are independent of $\boldsymbol{X}$ but impose the stronger assumption that $(\boldsymbol{\epsilon}_{t,i})_{\substack{t=1,\dots,T\\ i=1,\dots,N}}$ are i.i.d. homoscedastic standard normals to handle the second case. The propose method estimates this system of simultaneous equations using a simple convex program and generalizes the STIV estimator of Gautier and Tsybakov (2011). Inference is obtained by solving simple linear programs and is robust to identification. This is in sharp contrast with NP-hard methods proposed to estimate directed acyclic graph which precludes simultaneity but do not include the supplemental interaction term involving the exogenous $\boldsymbol{X}$ matrix (see, *e.g.*, van de Geer and Bühlmann (2013)).

References

[1] Bramoullé, Y., Djebbari, H., and B. Fortin (2009): "Identification of Peer Effects Through Social Networks ". *Journal of Econometrics,* 150, 41–55. Gautier, E., and A. Tsybakov (2011, 2014): "High-dimensional Instrumental Variables Regression and Confidence Sets". Preprint arXiv:1105.2454v4.

[2] Manski, C. F. (1993): "Identification of Endogenous Social Effects: The Reflection Problem". *Review of Economic Studies,* 60, 531–542.

[3] van de Geer, S., and P. Bühlmann (2013): "ℓ_0-penalized Maximum Likelihood For Sparse Directed Acyclic Graphs". *The Annals of Statistics*, 41, 536–567.

Toulouse School of Economics, 21 allée de Brienne, 31000 Toulouse, France.
E-mail address: eric.gautier@tse-fr.eu

Toulouse School of Economics, 21 allée de Brienne, 31000 Toulouse, France.
E-mail address: chris.rose@tse-fr.eu