Inferring Unobserved Firm Networks

Jesse Tweedle

Abstract

Use data and machine learning to infer unobserved firm-firm trading networks. Given data on firm characteristics, and a detailed geographic trading network, infer the unobserved firm-firm trading network that matches the Canadian national accounts.

1 Equations

The first set of equations to match is

$$\sum_{r=1}^{R} I_r a_{ri} + \sum_{j=1}^{N} s_j g_{ji} = s_i, \quad i = 1, \ldots, N \tag{1}$$

$$I \cdot a_{\cdot i} + s \cdot g_{\cdot i} = s_i, \quad i = 1, \ldots, N \tag{2}$$

Write $Z = (I, s) = (I_1, \ldots, I_R, s_1, \ldots, s_N)$, which is length $R+N$. Write $s = c$ to match notation in ML papers.

$$X = \begin{bmatrix} Z & 0 & \ldots & 0 \\ 0 & Z & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & Z \end{bmatrix}$$

Then X has dimensions $N \times (R+N)N$.

Let y be a combined vectorized A and G, so that y is a vector with length $(R+N)N$.

$$y = (a_{11}, a_{21}, \ldots, a_{R1}, g_{11}, g_{21}, \ldots, g_{N1}, \ldots, a_{1N}, \ldots, a_{RN}, g_{1N}, \ldots, g_{NN})$$

Or, write $a_{\cdot i}$ as the i-th column of A, and so on.

$$y = (a_{\cdot 1}, g_{\cdot 1}, \ldots, a_{\cdot N}, g_{\cdot N})$$

Then the benchmarking equations are:

$$Xy = c \tag{3}$$

Or,

$$\underbrace{\begin{bmatrix} I & s & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & I & s & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & I & s \end{bmatrix}}_{N \times (RN+N^2)} \underbrace{\begin{bmatrix} a_{\cdot 1} \\ g_{\cdot 1} \\ \vdots \\ a_{\cdot N} \\ g_{\cdot N} \end{bmatrix}}_{(RN+N^2)\times 1} = \underbrace{\begin{bmatrix} s_1 \\ \vdots \\ s_N \end{bmatrix}}_{N \times 1}$$

And the minimization problem is:

$$\min_y ||Xy - c|| + \lambda ||y|| \tag{4}$$

This is an underdetermined system, since the number of variables is much more than the number of "observations" (which, in this case, are firm characteristics and later, national accounts). The estimated y gives the most sparse implied network that matches the national accounts. Density can be increased by using elastic-net (which combines $l1$ and $l2$ regularization). To use the geographic trade network, we assume the implied firm-firm network identified from that as a subset of the true network, and leave those edges out of the penalty $||y||$.

Now use Lasso algorithm to solve this problem.

Short term strategy: set $\beta = 0.5$, draw random s_i, random firm locations r_i, calculate regional income $I_r = \sum_{i \in r} \beta s_i$, then construct X and c, then run glmnet for different R, N, skewness in s and so on. See if anything works.

Use the implied A and G to solve the equations and see how close it gets. Check sparseness, check network measures, and so on. Does each firm have at least one customer, at least one supplier? Make sure $g_{ii} = 0$ somehow?