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

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

Which I rewrite as (and add important other restrictions)

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

$$a_{ri} \geq 0, \quad a_{ri} - 1 \leq 0 \tag{3}$$

$$g_{ij} \geq 0, \quad g_{ij} - 1 \leq 0 \tag{4}$$

$$g_{ii} = 0 \tag{5}$$

$$\sum_{i \in N} a_{ri} = 1 \tag{6}$$

$$\sum_{j \in N} g_{ij} = 1 - \beta_i \tag{7}$$

$$X_{mc} = \begin{bmatrix} M & 0 & \ldots & 0 & s & 0 & \ldots & 0 \\ 0 & M & \ldots & 0 & 0 & s & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & M & 0 & 0 & \ldots & s \end{bmatrix}$$

Then X has dimensions $N \times (R+N)N$ (this may be a problem later—it's sparse but still has exactly $(R+N)N$ non-zero entries, which is too many if

N is any economically reasonable number). mc stands for market clearing, and represents Equation (2).

There are also $R+N$ equations that require the expenditure shares add up to 1.

$$X_a = \begin{bmatrix} 1 & 0 & \dots & 0 & \dots & 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & \dots & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \dots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & \dots & 0 & 0 & \dots & 1 \end{bmatrix} = \underbrace{\begin{bmatrix} I_R \dots I_R \end{bmatrix}}_{R \times RN}$$

And X_g is similar,

$$X_g = \underbrace{\begin{bmatrix} I_N \dots I_N \end{bmatrix}}_{N \times N^2}$$

Combine those two into Z

$$Z = \begin{bmatrix} X_a & 0 \\ 0 & X_g \end{bmatrix}$$

And then construct X as

$$X = \begin{bmatrix} X_{mc} \\ Z \end{bmatrix}$$

And

$$c = \begin{bmatrix} s \\ 1_R \\ 1-\beta \end{bmatrix}$$

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

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

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

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

We have the three things we need: c is $(N+R+N)\times 1$, y is $(RN+N^2)\times 1$, and X is $(N+R+N) \times (RN+N^2)$, and the set of equations to solve is:

$$Xy = c \tag{8}$$

Or,

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

And the minimization problem is:

$$\min_y ||Xy - c||_2 + \lambda ||y||_1 \tag{9}$$

Or:

$$\min_y ||Xy - c||_2 + \lambda \left((1-\alpha)\frac{1}{2}||y||_{\ell 2}^2 + \alpha ||y||_{\ell 1} \right) \tag{10}$$

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 $\ell 1$ and $\ell 2$ 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 `glmnet` to solve this problem. Simulation testing: set R, N, set , draw random β_i and s_i, random firm locations r_i, calculate regional income $I_r = \sum_{i \in r} \beta_i s_i$, then construct X and c, then run glmnet for different R, N, skewness in s and so on. See if anything works.

2 First try

It's not bad. If not sparse, everything looks pretty great, every equation is satisfied or close to it. When it's sparse (say 5%), the sizes aren't terrible ($0.9\ R^2$ for non-zero predicted sizes, and $0.99\ R^2$ for firms above the median). Rank is almost always preserved, but the bottom of the distribution curves disastrously. The other problem: the sum equations (6) and (7) don't seem to hold ever. One solution might be post-processing of the A and G.