Spatial approximation of sparsely sampled networks

Miklós Koren
CEU, MTA KRTK and CEPR

Motivation

Motivation

- Interaction among firms is important for performance.
  - Spatial economics: trade costs, Marshallian externalities.
  - Urban economics evidence: firms in dense areas are more productive.
  - Network methods: input-output linkages matter.
- How to measure the strength of interaction and its effect on performance?
- Quantifiable spatial economics models: geography of space matters, but we can quantify its impact.

Buyer-seller networks

- N buyers and K sellers.
- How do buyer-seller links correlate with firm performance? (No causal analysis so far.)
- Conceptual/computational problems:
  - there are $N \times K$ potential buyer-seller links.
  - links may vary over time.

Model

A balls-and-bins model of link reporting

- How much did firm n buy from firm k last month (week, day, hour, minute, second)?

A balls-and-bins model of link reporting

- How much did firm n buy from firm k last month (week, day, hour, minute, second)?
- As Armenter and Koren (2013) show, not much can be solved by time aggregation.
- Instead, model
$$\Pr(G_{nk} = 1).$$

Poisson process for link reporting

$$\Pr(G_{nk,t,t+h} = 1) = 1 - e^{-\lambda_{nk} h} \approx \lambda_{nk} h$$

Dimension reduction

- Both statistical learning and theorizing are about *dimension reduction*: map complex problem into fewer dimensions.
- Can we make this explicit for the buyer-seller problem?

A spatial model

- Buyer n is located at $X_n \in \mathbb{R}^M$. $(1 < M \ll K)$
- Seller k is located at $Y_k \in \mathbb{R}^M$. $(1 < M \ll N)$

Buyer-seller links

- Probability of a link depends on (squared) Euclidean distance

$$-\sum_{m=1}^{M}(x_{nm}-y_{km})^2 = -\sum_{m=1}^{M} x_{nm}^2-\sum_{m=1}^{M} y_{km}^2+\sum_{m=1}^{M} x_{nm}y_{km} = \mu_n+\nu_k+\sum_{m=1}^{M} x_{nm}y_{km}$$

Matrix factorization

- But the last term is exactly as in matrix factorization models. (Recommendation engines, Netflix prize)

$$\Pr(G_{nk} = 1 | X_n, Y_k) = \mu_n + \nu_k + X_n Y_k$$

$$\mathbf{G} \approx \mu + \nu + \mathbf{XY}$$

- This approximates G with a low-rank matrix (e.g., singular value decomposition).
- When $M = 1$, $X, Y \approx$ Pagerank, worker-firm fixed effects.

Firm performance

- Node-level performance,
$$Q_n = f(X_n).$$
- This encompasses spatial amenities (good X) and agglomeration (good neighbors).

Questions

Answerable

- How does firm position in network account for inequality in firm performance?

$$dQ_n \approx \sum f_m dX_{nm}$$

- What are X_ns correlated with?

Not yet

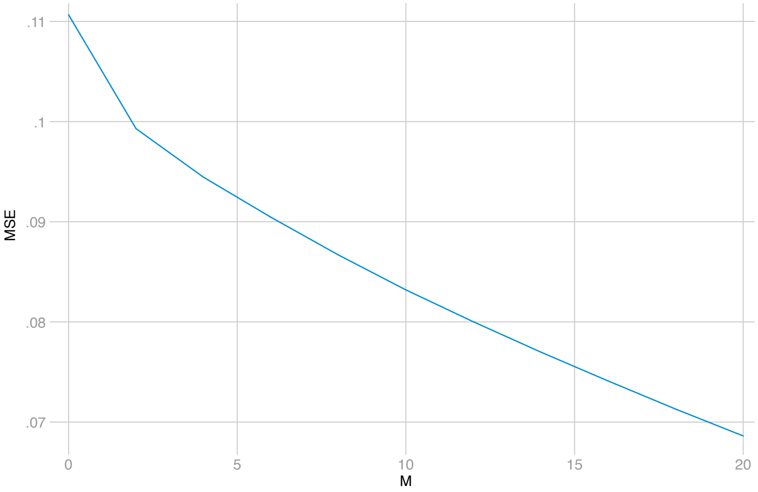
- How much of it is amenities vs agglomeration? (X vs G)
- What if firms choose X_n endogenously?
- Do returns to X_n change in equilibrium? (optimal transport?)

Applications

Procurement network

- Buyers, $N = 8300$.
- Sellers, $K = 27400$.
- Edges: 86900.
- Work with 100×500 matrix for now.
- Train: 2010..2014, test: 2015..2017

Goodness of fit



Prediction

- MSE across years: 0.2022
- MSE from M=20: 0.1962
- Cross-validated M: 2
- Cross-validated MSE: 0.1912