14.385 Nonlinear Econometric Analysis

(Causal) matrix completion

Maximilian Kasy

Department of Economics, MIT

Fall 2022

Outline

- Setup: Filling in missing entries in a matrix Y.
  1. Recommender systems.
  2. Counterfactual outcomes in panel data.
- Recap: Singular value decompositions and principal components.
- Empirical risk minimization methods.
- Nearest neighbor methods.
- Missingness assumptions.
- Reweighting.
- The synthetic nearest neighbor algorithm.

Takeaways for this part of class

- Typical assumption:
  $\boldsymbol{Y}$ is a sum of a low rank matrix $\boldsymbol{A}$ and idiosyncratic noise $\boldsymbol{E}$.
- Any matrix has a singular value decomposition.
  Principal components correspond to the largest singular values.
- Popular methods for matrix completion:
  1. *Empirical risk minimization*.
  2. *Nearest neighbors*.
- Standard methods suffer from bias with non-random missingness.
- For (conditionally) missing at random data,
  reweighting can provide a solution.
- An algorithm for more general missingness is *Synthetic nearest neighbors*.

Setup

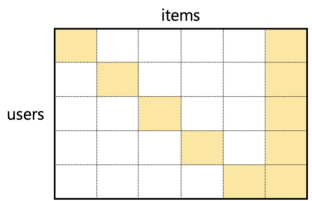
- Random matrices with m rows, n columns:
  - Latent matrix A.
  - Error matrix E, where $E[E|A] = 0$.
  - Outcome matrix $Y = A + E$, thus $E[Y|A] = A$.
  - Observability matrix D.
  - Probability of observability: $P = E[D]$.
- Goal: Estimate the entries of the latent matrix A based on observations $(Y_{ij} \cdot D_{ij}, D_{ij})$.
- Typical loss function:

$$\frac{1}{m \cdot n} \sum_{i,j} \left(\hat{A}_{ij} - A_{ij} \right)^2 .$$

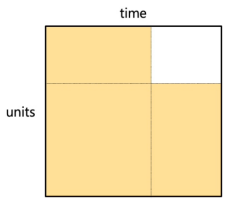
Interpretations

- Recommender systems:
  - Rows i index individuals, columns j index movies.
  - Y_{ij} are movie ratings by individual i for movie j.
  - D_{ij} are indicators for whether a movie was rated by an individual.
  - Goal: Recommend movies that would receive high ratings.
- Panel data causal inference (cf. synthetic controls):
  - Rows i index cross-sectional units, columns j index time-periods.
  - Y_{ij} are potential outcomes absent treatment.
  - D_{ij} are indicators for untreated units.
  - Goal: Recover missing Y_{ij} to recover causal effects.

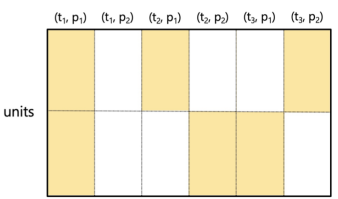
Typical patterns of missingness



(a) Recommender systems.



(b) Panel data.



(c) Sequential decision-making.

Reminder: Singular value decomposition

- Any real valued matrix A with m rows, n columns, rank $k = \mathrm{rank}(A)$ can be decomposed as

$$A = U \cdot \Sigma \cdot V = \sum_{l=1}^{k} \sigma_l \cdot u_l \cdot v_l'.$$

  - U is an $m \times k$ matrix with orthonormal columns u_l.
  - V is an $n \times k$ matrix with orthonormal columns v_l.
  - Σ is a $k \times k$ diagonal matrix with entries σ_l of decreasing magnitude.

- Special case: Diagonalization of square matrices A.
- Consider the largest singular values σ_l:
  - Principal components: $\sigma_l \cdot u_l$.
  - Low-rank approximation: $A \approx \sum_{l=1}^{\kappa} \sigma_l \cdot u_l \cdot v_l'$, where $\kappa < k$.

Empirical risk minimization (ERM) methods

- Minimize average prediction error for observed outcomes:

$$\hat{A} = \underset{a}{\operatorname{argmin}} \sum_{i,j} D_{ij} \cdot \left(a_{ij} - Y_{ij}\right)^2 + \lambda \cdot Reg(a).$$

- Here ***Reg*** is one of several possible regualarization penalties, λ is a tuning parameter.
- Popular choice: Nuclear norm (or trace norm).

$$Reg(a) = \operatorname{tr}(\sqrt{a' \cdot a}) = \sum_{l} \sigma_l(a).$$

  The $\sigma_l(a)$ are the singular values of a.

  $\Rightarrow$ Lasso penalty for the singular values of $\hat{A}$.

  $\Rightarrow$ *SoftImpute* algorithm

- Variant: Rather than penalizing $\hat{A}$, constrain $\hat{A}$ to be low rank.

Nearest neighbor methods

- Consider a specific i,j with $D_{i,j} = 0$.
- Find a set $\mathcal{I}$ of k rows i', such that
  1. $D_{i'j} \neq 0$.
  2. Row i' is "similar" to row i.
- "Similar" often means a small distance of the vector of observed values,

$$\sum_{j'} D_{ij'} \cdot D_{i'j'} \cdot \left(Y_{ij'} - Y_{i'j'}\right)^2.$$

- Impute an estimate for Y_{ij} as

$$\hat{Y}_{ij} = \frac{1}{k} \sum_{i' \in \mathcal{I}} Y_{i'j}.$$

Missingness assumptions

1. **Missing completely at random** (MCAR):

$$D_{ij}|Y \sim^{iid} Ber(p).$$

2. **Missing at random** (MAR):

$$D_{ij}|Y,X \sim Ber(P_{ij}(X)),$$

independently across i,j,
where X are observable controls, and $P_{ij}(X) > 0$.

3. **Missing not at random** (MNAR):
   - D and Y are not independent,
   - D_{ij} and $D_{i'j'}$ are not independent,
   - $P_{ij} = 0$ is allowed.

Practice problem

- Suppose MCAR holds.
  Consider any empirical risk minimization (ERM) algorithm.
  What is the expectation of the objective function for such an algorithm?

- Suppose MAR holds.
  How could you modify empirical risk minimization, to avoid biases?

Reweighting under MAR

- Suppose that P takes the form

$$P_{ij} = g\left(X_i \cdot \beta_x + W_j \cdot \beta_w + \delta_i + \gamma_j\right),$$

  where $g(\cdot)$ is a link function; e.g. the logistic $g(x) = \frac{\exp(x)}{1+\exp(x)}$.

- We can estimate P by logistic regression of D_{ij} on X_i, W_j, and row and column fixed effects.

- Reweighted ERM:

$$\hat{A} = \underset{a}{\operatorname{argmin}} \sum_{i,j} \frac{D_{ij}}{\hat{P}_{ij}} \cdot \left(a_{ij} - Y_{ij}\right)^2 + \lambda \cdot Reg(a).$$

A restricted form of MNAR

Assummptions:

1. **Low rank factor model**:
   $\text{rank}(A) = k < \min(m, n)$, so that

$$A = \sum_{l=1}^{k} \sigma_l \cdot u_l \cdot v_l'.$$

2. **Selection on latent factors**:

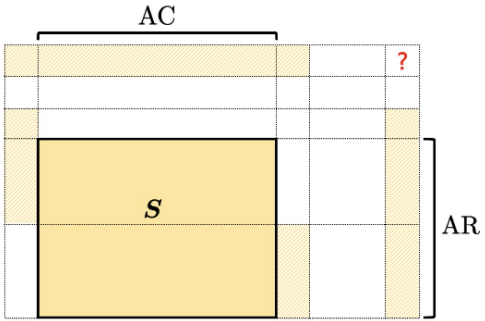
$$E[E|U, V, D] = 0.$$

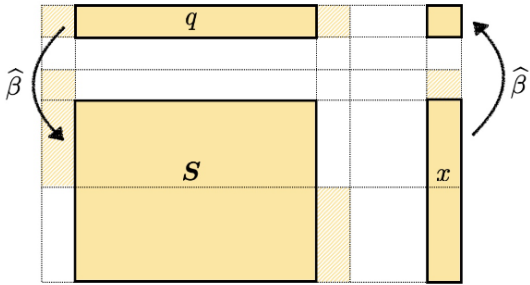
3. **Linear span inclusion**:
   Any set of k rows of U has full rank.

Identification

- Assumption 3 could be weakened, but holds generically.
- Immediate implication of these assumptions:
  - Fix a pair (i,j).
  - Let $\mathscr{I}$ be such that $D_{i'j} = 1$ for all $i' \in \mathscr{I}$ and $|\mathscr{I}| \geq k$.
  - Then there is a β such that
$$u_{i,.} = \sum_{i' \in \mathscr{I}} \beta_{i'} \cdot u_{i',.}.$$
  - Furthermore,
$$A_{ij} = \sum_{i' \in \mathscr{I}} \beta_{i'} \cdot E[Y_{i'j}|U,V,D].$$
- Thus:
  - Suppose we could estimate β.
  - Then we could impute
$$\hat{A}_{ij} = \sum_{i' \in \mathscr{I}} \beta_{i'} \cdot Y_{i'j}.$$

Synthetic nearest neighbors





Synthetic nearest neighbors (1)

Algorithm proposed by Agarwal et al. (2021);

1. Fix tuning parameter $\kappa \in \mathbb{N}$ (rank of approximations).
2. Consider some (i,j) for which we want to estimate A_{ij}.
3. Find a set of rows and columns AR and AC such that

$$D_{i'j'} = D_{i'j} = D_{ij'} = 1$$

for all $i' \in AR$ and $j' \in AC$.
Let S be the submatrix of Y corresponding to rows AR, columns AC.

4. Find the singular value decomposition

$$S = \sum_{l=1}^{k} \sigma_i \cdot \hat{u}_i \cdot \hat{v}_i'$$

Synthetic nearest neighbors (2)

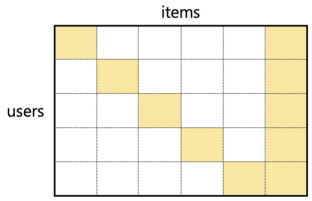
5. Estimate
$$\hat{\beta} = \left(\sum_{l=1}^{\kappa} \frac{1}{\sigma_i} \cdot \hat{u}_i \cdot \hat{v}_i' \right) \cdot A_{i,AC}.$$
(Note we are truncating the sum at κ.)

6. Impute
$$\hat{A}_{ij} = \sum_{i' \in \mathscr{J}} \beta_{i'} \cdot Y_{i'j}.$$

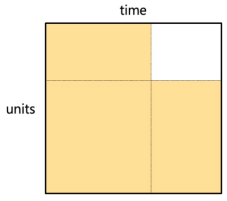
7. Repeat for different rows AC, columns AR, and average.

Role of columns and rows could be switched, without affecting the estimate.

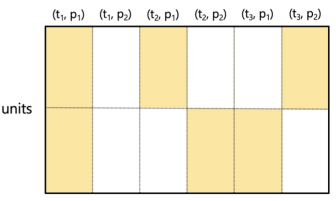
Typical patterns of missingness



(a) Recommender systems.



(b) Panel data.



(c) Sequential decision-making.

References

- *Athey, S., Bayati, M., Doudchenko, N., Imbens, G., and Khosravi, K. (2021). Matrix completion methods for causal panel data models.* Journal of the American Statistical Association*, 116(536):1716–1730*
- *Agarwal, A., Dahleh, M., Shah, D., and Shen, D. (2021). Causal matrix completion.* arXiv preprint arXiv:2109.15154