Big Data, Machine Learning, and Artificial Intelligence: Methods Lectures and Applications

Susan Athey, Professor, Stanford University
Department of Economics

Sunday, August 30, 2020
Longitudinal Data and Matrix Factorization
Motivating Example

- California's anti-smoking legislation (Proposition 99) took effect in 1989.

- **What is the causal effect of the legislation on smoking rates in California in 1989?**

- We **observe** smoking rates in California in 1989 given the legislation. We need to **impute** the **counterfactual** smoking rates in California in 1989 had the legislation not been enacted.

- We have data in the absence of smoking legislation in California prior to 1989, and for other states both before and in 1989. (and other variables, but not of essence)
Setup for Causal Inference with Panel Data

Set Up: we observe (in addition to covariates):

\[
Y = \begin{pmatrix}
Y_{11} & Y_{12} & Y_{13} & \ldots & Y_{1T} \\
Y_{21} & Y_{22} & Y_{23} & \ldots & Y_{2T} \\
Y_{31} & Y_{32} & Y_{33} & \ldots & Y_{3T} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
Y_{N1} & Y_{N2} & Y_{N3} & \ldots & Y_{NT}
\end{pmatrix}
\]  
(realized outcome).

\[
W = \begin{pmatrix}
1 & 1 & 0 & \ldots & 1 \\
0 & 0 & 1 & \ldots & 0 \\
1 & 0 & 1 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & 1 & \ldots & 0
\end{pmatrix}
\]  
(binary treatment).

- rows of \( Y \) and \( W \) correspond to physical units, columns correspond to time periods.
Setup for Causal Inference with Panel Data

In terms of potential outcome matrices $Y(0)$ and $Y(1)$:

$$Y(0) = \begin{pmatrix} ? & ? & \checkmark & \ldots & ? \\
\checkmark & \checkmark & ? & \ldots & \checkmark \\
? & \checkmark & ? & \ldots & \checkmark \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
? & \checkmark & ? & \ldots & \checkmark \end{pmatrix} \quad Y(1) = \begin{pmatrix} \checkmark & \checkmark & ? & \ldots & \checkmark \\
? & \checkmark & ? & \ldots & ? \\
\checkmark & ? & ? & \ldots & ? \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\checkmark & \checkmark & ? & \ldots & ? \end{pmatrix}$$

$Y_{it}(0)$ observed iff $W_{it} = 0$, $Y_{it}(1)$ observed iff $W_{it} = 1$.

In order to estimate the average treatment effect for the treated, (or other average, e.g., overall average effect)

$$\tau = \frac{\sum_{i,t} W_{it} (Y_{it}(1) - Y_{it}(0))}{\sum_{i,t} W_{it}},$$

We need to **impute** the missing potential outcomes in $Y(0)$ (and in $Y(1)$ for other estimands).
Matrix Factorization

Analyst picks $K$. Use various methods to find factor matrices that minimize mean squared error of prediction versus true outcome.

If matrix has some structure ($\text{rank} < N,T$), can be well approximated by a factor model.
Outcome Model

\[ Y_{it} = L_{it} + \sum_{p=1}^{P} \sum_{q=1}^{Q} X_{ip} H_{pq} Z_{qt} + \gamma_i + \delta_t + V_{it} \beta + \varepsilon_{it} \]

- We do not necessarily need the fixed effects \( \gamma_i \) and \( \delta_t \), these can be subsumed into \( L \). It is convenient to include the fixed effects given that we regularize \( L \).
\[ \mathbf{L}_{N \times T} = \mathbf{S}_{N \times N} \mathbf{\Sigma}_{N \times T} \mathbf{R}_{T \times T} \]

\( \mathbf{S}, \mathbf{R} \) unitary, \( \mathbf{\Sigma} \) is rectangular diagonal with entries \( \sigma_i(\mathbf{L}) \) that are the **singular values**. Rank of \( \mathbf{L} \) is number of non-zero \( \sigma_i(\mathbf{L}) \).

\[ \| \mathbf{L} \|_F^2 = \sum_{j=1}^{\min(N,T)} \sigma_j^2(\mathbf{L}) = \sum_{i,t} |L_{it}|^2 \quad \text{(Frobenius, like ridge)} \]

\[ \Rightarrow \quad \| \mathbf{L} \|_* = \sum_{j=1}^{\min(N,T)} \sigma_j(\mathbf{L}) \quad \text{(nuclear norm, like LASSO)} \]

\[ \| \mathbf{L} \|_R = \sum_{j=1}^{\min(N,T)} 1_{\sigma_j(\mathbf{L}) > 0} \quad \text{(Rank, like subset selection)} \]
Following Candès & Recht, 2009; Candès & Plan (2010) we regularize using nuclear norm:

$$\min_{L} \frac{1}{|\emptyset|} \sum_{(i,t) \in \emptyset} (Y_{it} - L_{it})^2 + \lambda_L \|L\|_*$$

For the general case we estimate $H$, $L$, $\delta$, $\gamma$, and $\beta$ as

$$\min_{H,L,\delta,\gamma} \frac{1}{|\emptyset|} \sum_{(i,t) \in \emptyset} \left( Y_{it} - L_{it} - \sum_{p=1}^{P} \sum_{q=1}^{Q} X_{ip} H_{pq} Z_{qt} - \gamma_i - \delta_t - V_{it}\beta \right)^2$$

$$+ \lambda_L \|L\|_* + \lambda_H \|H\|_{1,e}$$

We choose $\lambda_L$ and $\lambda_H$ through cross-validation.
Methods for Causal Panel Models

  ◦ Compare to “horizontal,” “vertical,” or two-way fixed effects regressions
  ◦ Matrix completion works well as matrix size changes from thin to wide (best in intermediate cases)

Synthetic Difference-in-Differences (Athey et al 2018):
  ◦ Weighted difference-in-differences, to put more weight on similar time periods and units
Illustration I: Stock Market Data

We use daily returns for 2453 stocks over 10 years (3082 days). We create sub-samples by looking at the first $T$ daily returns of $N$ randomly sampled stocks for pairs of $(N, T)$ such that $N \times T = 4900$, ranging from fat to thin: $(N, T) = (10, 490), \ldots, (70, 70), \ldots, (490, 10)$.

Given the sample, we pretend that half the stocks are treated at the mid point over time, so that 25% of the entries in the matrix are missing in a particular block.

\[
Y_{N \times T} = \begin{pmatrix}
\checkmark & \checkmark & \checkmark & \checkmark & \ldots & \checkmark \\
\checkmark & \checkmark & \checkmark & \checkmark & \ldots & \checkmark \\
\checkmark & \checkmark & \checkmark & \checkmark & \ldots & \checkmark \\
\checkmark & \checkmark & \checkmark & \checkmark & \ldots & ? \\
\checkmark & \checkmark & \checkmark & \checkmark & \ldots & ? \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\checkmark & \checkmark & \checkmark & ? & \ldots & ? \\
\end{pmatrix}
\]
Matrix Factorization

I x J choice matrix \approx I x K matrix of user factors \times K x J matrix of item factors

Analyst picks K. User various methods to find factor matrices that minimize mean squared error of predicts versus true choice matrix. If matrix has some structure (rank < I,J), can be well approximated by a factor model.
ML and Structural Models: Shopping Application

Combine structural model with matrix factorization techniques and computational methods from ML

Scanner data from supermarket
  ◦ Product hierarchy (category, class, subclass, UPC)
  ◦ Prices change Tuesday evening
  ◦ Study 123 high-frequency categories with 1263 UPCs
    ◦ Multiple UPCs per category
    ◦ Typically purchase only one UPC per trip in category
    ◦ Independent price changes
    ◦ Not too much seasonality
    ◦ 333,000 shopping trips for ~2000 consumers over 20 months

Economic Goals:
  ◦ Optimal pricing
  ◦ Benefits of personalization versus simpler segmentation

Methodological Goals:
  ◦ Contrast off-the-shelf ML, off-the-shelf econometrics with combined models
  ◦ Tune and test models for counterfactual performance

Joint work with Rob Donnelly, David Blei, Fran Ruiz
Structural Model

Mixed logit
• User $u$, product $i$, time $t$

$$
\mu_{uit} = \nu_{ui} + \beta X_i - \alpha_u p_{it}
$$

$$
U_{uit} = \mu_{uit} + \epsilon_{uit}
$$

• If $\epsilon_{uit}$ i.i.d. Type I EV, then

$$
\Pr(Y_{uit} = i) = \frac{\exp(\mu_{uit})}{\sum_j \exp(\mu_{uji})}
$$

• Counterfactuals
  • Out of stock
  • Price changes

Matrix Factorization

$$
egin{bmatrix}
U \times K
\end{bmatrix}
\times
egin{bmatrix}
K \times I
\end{bmatrix}
\approx
egin{bmatrix}
U \times I
\end{bmatrix}
$$
Structural Model + Factorization

Mixed logit
• User $u$, product $i$, time $t$

$$\mu_{uit} = \nu_{ui} + \kappa X_i - \alpha_u p_{it}$$
$$U_{uit} = \mu_{uit} + \epsilon_{uit}$$

• If $\epsilon_{uit}$ i.i.d. Type I EV, then

$$\Pr(Y_{uit} = i) = \frac{\exp(\mu_{uit})}{\sum_j \exp(\mu_{ujt})}$$

• Counterfactuals
  • Out of stock
  • Price changes

Mixed logit + factors
• User $u$, product $i$, time $t$

$$\mu_{uit} = \beta_u \theta_i + \kappa_u X_i - \rho_u \alpha_i p_{it}$$

• Add in nesting for outside good
  • Implement as two-stage estimation with inclusive value (McFadden)
  • Also factorization of outside good
MCMC based Bayesian approaches
- Difficult scaling

Variational Bayes
- Choose parameterized family of distributions to approximate posterior
- Find parameters that minimize KL-divergence to true posterior

Challenges for our problem
- Nonlinear form of choice probabilities
- Time-varying prices and availability

Tricks
- Approximations and bounds
- Stochastic gradient descent (SGD)
  - Predict: lots of applications of SGD!!
Model Comparisons

Compare alternative models, show that Nested Factorization performs best in representative scenarios as well as in scenarios with price changes

Nested Factorization
- All categories estimated in single model
- Items substitutes within category, independent across
- Tuned on held-out validation set

Hierarchical Poisson Factorization (HPF)
- All items in single model, each item independent of others
- A form of matrix factorization allowing for covariates
- Ignores prices
- Scales easily

Category by category logits
- Mixed logit (random coefficients)
- Nested Logit
- With various controls (demographic, etc.)

Logits with HPF Factors
- Include user-item prediction from HPF model
Validation of Structural Parameter Estimates

Compare Tues-Wed change in price to Tues-Wed change in demand, in test set
Break out results by how price-sensitive (elastic) we have estimated consumers to be