Causal Learning in Economics

Part II: SL+ UL

Mingli Chen

University of Warwick

May 2, 2025

Previous examples were all pretty much supervised learning

- have data on a set of features $X_1, ..., X_p$ and an outcome Y
- goal to build a model for $Y|X_1, ..., X_p$ that captures important features but doesn't overfit
- model assessment pretty natural

Unsupervised learning:

- have data on a set of features $X_1, ..., X_p$
- want to extract interesting low-dimensional information from data
  - E.g. is there a useful way to visualize/present complicated or high-dimensional data? (EDA/fancy descriptive statistics)
  - E.g. is there a useful dimension reduction of X that can be taken before applying other methods?
  - E.g. are there sets of similar variables or observations?

Challenge of unsupervised learning is that it's very subjective (in many situations) and hard to assess

Some unsupervised techniques pretty familiar to economists:

- density estimation
- factor models/PCA

Other approaches may be less familiar:

- clustering
- topic models
- graphical models

Factor Models and PCA

$$X_t = \Lambda F_t + \epsilon_t$$

where X_t is the observed data, Λ is the factor loading matrix, F_t are latent factors, and ϵ_t is the error term.

- X_t: $N \times 1$ vector
- F_t: $K \times 1$ vector
- Λ: $N \times K$ matrix
- e_t: $N \times 1$ vector

PCA finds principal components by maximizing variance:

$$\text{PC}_k = \arg \max_{\|v\|=1} \text{Var}(Xv), \quad \text{subject to orthogonality constraints}$$

PCA and Factor Models (Comments)

Stock and Watson (2002) "Forecasting Using Principal Components From a Large Number of Predictors" JASA provides conditions under which PCA is consistent for factors from factor model (in appropriate sense) in large N and T setting

PCA (and factor analysis) often used as a pre-processing step to do dimension reduction before applying other supervised learning technique for forecasting (e.g. Stock and Watson's motivation)

- If you have target in mind, why not do factor extraction jointly with learning?
- Partial least squares, Supervised Principal Components, ... try to do this
- PCA inherently a high-dimensional operation, might want further regularization, e.g. sparse PCA

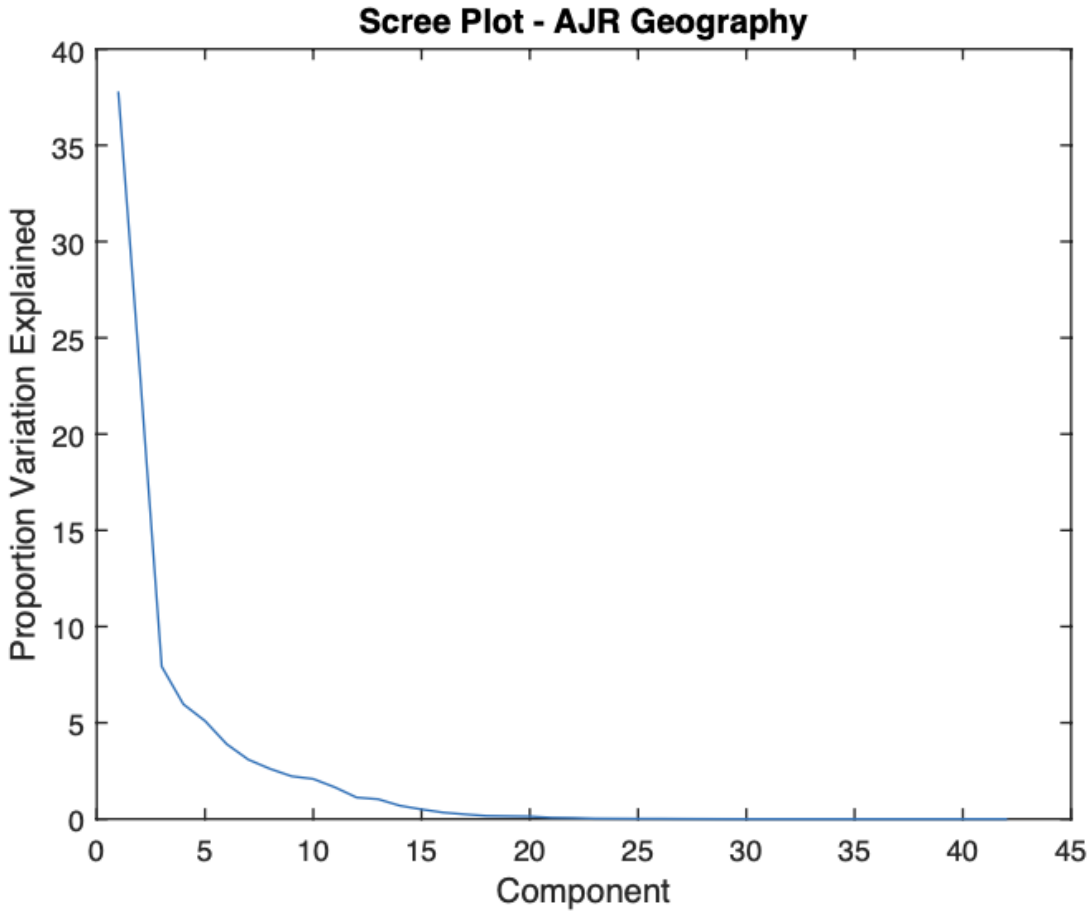
PCA in AJR Institutions IV Example

- Equation of interest:

  $$\log(\text{GDP per capital}_i) = \alpha(\text{Protection from Expropriation}_i) + x_i'\beta + \varepsilon_i$$

- Want to control for persistent variables related to institutions and GDP development.
- Leading candidate: Geography (Geographic Determinism).
- Let's look at principal components in the geography variables

Scree Plot for Geography PCs



Looks like most of the variation captured by at most 20 PC - probably need far fewer.

2SLS Estimates Using PCs as Controls

- **Baseline (just latitude):**
  - First-stage: -0.5487 (0.1659)
  - $\hat{\alpha}$: 0.9252 (0.2095)
- **PC 1:**
  - First-stage: -0.3542 (0.1732)
  - $\hat{\alpha}$: 1.2151 (0.4759)
- **PC 1-5:**
  - First-stage: -0.2931 (0.1656)
  - $\hat{\alpha}$: 1.1119 (0.5160)
- **PC 1-20:**
  - First-stage: -0.0658 (0.2435)
  - $\hat{\alpha}$: 2.1202 (6.4129)
- Question: How many PCs are optimal?

Clustering

- Groups similar objects into clusters based on features, without predefined labels.
- **K-Means Clustering**: Minimizes within-cluster variance:

$$\min_{C_1,\ldots,C_K} \sum_{k=1}^{K} \sum_{i \in C_k} \|x_i - \mu_k\|^2$$

  where C_k are clusters, μ_k is the centroid of cluster k, and x_i are data points.
- **Hierarchical Clustering**: Builds a tree (dendrogram) by merging or splitting clusters based on a distance metric (e.g., Euclidean).

Topic Modeling

- Topic modeling is a form of clustering for discrete data, widely used in text analysis.
- Core building block: A multimodal mixture model.
- **Model Equation**:

$$\mathrm{x}_i \sim \mathrm{MN}\left(\omega_{i1}\theta_1 + \omega_{i2}\theta_2 + \ldots + \omega_{iK}\theta_K, m_i\right)$$

  - x_i: Data for document i (e.g., all tokens in document i).
  - K: Number of potential topics.
  - θ_k: Probability of word j in topic k $[\theta_k = \{\theta_{k1}, \ldots, \theta_{kJ}\}']$.
  - ω_{ik}: Probability that document i belongs to topic k.
  - $m_i = \sum_j x_{ij}$: Total words in document i.

Interpretation

- The word vector x_i in each document follows a multinomial distribution with probabilities as a mixture of topics.
- θ_k vectors represent topic phrase probabilities, with $\sum_{j=1}^{J} \theta_{kj} = 1$.
  - Example: A banking topic might have high probabilities for "money", "interest rate", "loan", etc.
- Document weights ω_{ik} are probabilities, with $\sum_{k=1}^{K} \omega_{ik} = 1$.
  - Example: A paper on the economics of bank runs would have a high probability for the banking topic.
- Unlike traditional clustering, each word comes from a topic, and each document is a mixture of topics.

Casual Panel

FE, IFE, MC

$$\begin{cases} Y_{it}(0) \\ Y_{it}(1) = Y_{it}(0) + \tau_{it} \end{cases}$$

- FE

$$Y_{it}(0) = X_{it}\beta + Z_i\theta_t + \delta_t + \alpha_i + \varepsilon_{it}$$

- IFE

$$Y_{it}(0) = X_{it}\beta + Z_i\theta_t + \lambda_t\mu_i + \varepsilon_{it}$$

- MC assumes that the non-treated potential outcome matrix Y(0) can be approximated by L (we omit covariates and additive fixed effects for simplicity):

$$\mathrm{Y}(0) = \mathrm{L} + \varepsilon, E[\varepsilon|\mathrm{L}] = 0,$$

Comparison

- Factor model (interactive fixed effects approach)
- Nuclear norm penalisation

Recall that for a matrix $A \in \mathbb{R}^{m \times k}$ its nuclear norm is given by

$$\|A\|_* = \sum_{i=1}^{\min\{m,k\}} \sigma_i,$$

where $\sigma_1, \ldots, \sigma_{\min\{m,k\}}$ are the singular values of A.

MC, nuclear norm penalisation

- As with IFE, L can be expressed as the product of two k-dimension matrices: $L = \Lambda F$
- Different from IFE, however, instead of estimating factors F and factor loadings Λ separately, the MC estimator seek to directly estimate L by solving the following minimization problem:

$$\hat{\mathsf{L}} = \arg\min_{\mathsf{L}} \left[\sum_{(i,t)\in\mathcal{O}} \frac{(Y_{it} - L_{it})^2}{|\mathcal{O}|} + \theta \|\mathsf{L}\|_* \right],$$

  in which $\mathcal{O} = (i, t)$, $D_{it} = 0$ and $|\mathcal{O}|$ is the number of elements in $\mathcal{O}$. $\|\mathsf{L}\|$ is the chosen matrix norm of $\|\mathsf{L}\|$ and θ is a tunning parameter.
- It's a "matrix version" of LASSO where the nuclear norm (ℓ_1 norm on singular values) is the regularization.