Discrete Choice Models

Kosuke Imai

Princeton University

POL573 Quantitative Analysis III
Fall 2016

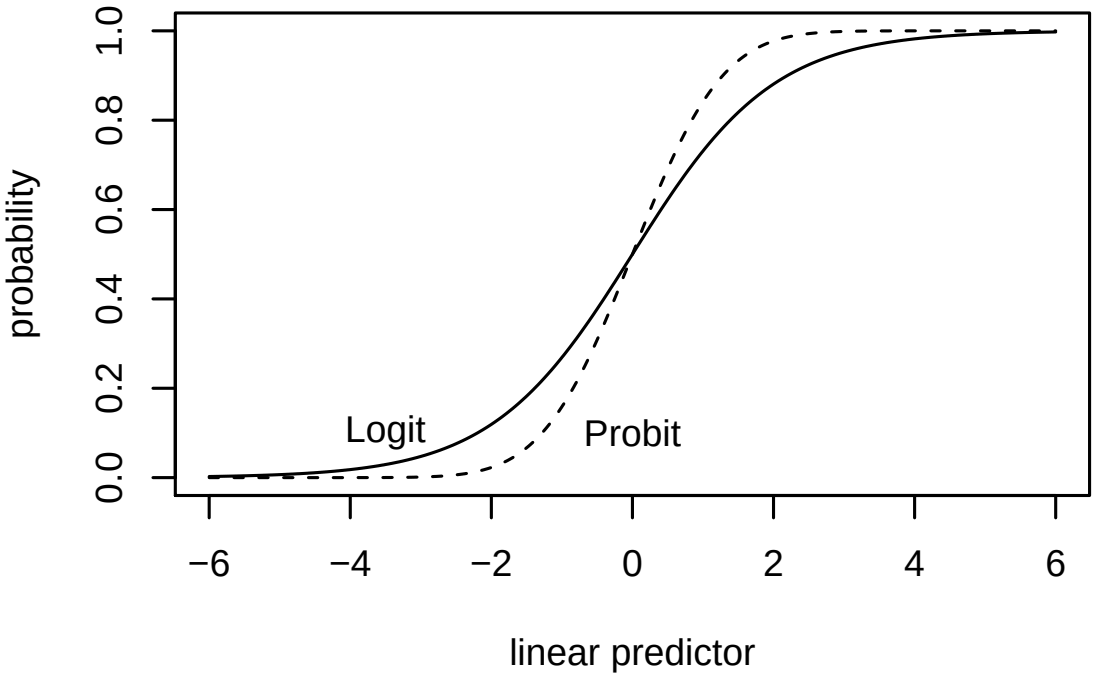
Recall Binary Logit and Probit Models

- Logit and probit models for binary outcome $Y_i \in \{0, 1\}$:

$$Y_i \overset{\text{indep.}}{\sim} \text{Bernoulli}(\pi_i)$$

$$\pi_i = \frac{\exp(X_i^\top \beta)}{1 + \exp(X_i^\top \beta)} = \frac{1}{1 + \exp(-X_i^\top \beta)}$$

- Logit function: $\text{logit}(\pi_i) \equiv \log(\pi_i/(1-\pi_i)) = X_i^\top \beta$
- Probit function: $\Phi^{-1}(\pi_i) = X_i^\top \beta$



- monotone increasing
- symmetric around 0
- maximum slope at 0
- logit coef. $=$ probit coef. $\times 1.6$

Latent Variable Interpretation

- The latent variable or the "Utility": Y_i^*
- The Model:

$$\begin{aligned} Y_i &= \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{if } Y_i^* \leq 0 \end{cases} \\ Y_i^* &= X_i^\top \beta + \epsilon_i \quad \text{with} \quad \mathbb{E}(\epsilon_i) = 0 \end{aligned}$$

- Logit: $\epsilon_i \overset{\text{i.i.d.}}{\sim} \text{logistic}$ (the density is $\exp(-\epsilon_i)/\{1+\exp(-\epsilon_i)\}^2$)
- Probit: $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$
- The variance of Y_i^* is not identifiable
- The "cutpoint" is not identifiable

Inference with the Logit and Probit Models

- Likelihood and log-likelihood functions:

$$L_n(\beta \mid Y, \mathbf{X}) = \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$$

$$l_n(\beta \mid Y, \mathbf{X}) = \sum_{i=1}^{n} \{ Y_i \log \pi_i + (1 - Y_i) \log(1-\pi_i) \}$$

- Logit model:
  - Score function: $s_n(\beta) = \sum_{i=1}^{n}(Y_i - \pi_i)X_i$
  - Hessian: $\mathbf{H}_n(\beta) = -\sum_{i=1}^{n} \pi_i(1-\pi_i)X_i X_i^\top \leq 0$
  - Approximate variance: $\mathbb{V}(\hat{\beta}_n \mid \mathbf{X}) \approx \{\sum_{i=1}^{n} \pi_i(1-\pi_i)X_i X_i^\top\}^{-1}$
  - Globally concave

Calculating Quantities of Interest

- Logistic regression coefficients are NOT quantities of interest
- Predicted probability: $\pi(x) = \Pr(Y = 1 \mid X = x) = \frac{\exp(x^\top \beta)}{1+\exp(x^\top \beta)}$
- Attributable risk (risk difference): $\pi(x_1) - \pi(x_0)$
- Relative risk: $\pi(x_1)/\pi(x_0)$
- Odds and odds ratio: $\frac{\pi(x)}{1-\pi(x)}$ and $\frac{\pi(x_1)/\{1-\pi(x_1)\}}{\pi(x_0)/\{1-\pi(x_0)\}}$
- Average Treatment Effect:

$$\mathbb{E}\{\Pr(Y_i = 1 \mid T_i = 1, X_i) - \Pr(Y_i = 1 \mid T_i = 0, X_i)\}$$

- MLE: plug in $\hat{\beta}_n$
- Asymptotic distribution: the Delta method (a bit painful!)

$$\sqrt{n}(\hat{\pi}(x) - \pi(x)) \xrightarrow{D} \mathcal{N}\left(0, \frac{\pi(x)^2}{\{1 + \exp(x^\top \beta_0)\}^2} x^\top \Omega(\beta_0)^{-1} x\right)$$

Application 1: Case-Control Design

- Research design mantra: "Don't select on dependent variable"
- But, sometimes, we want to select on dependent variable (e.g., rare events)
  - civil war, campaign contribution, protest, lobbying, etc.

- The standard case-control (choice-based sampling) design:
  1. Randomly sample "cases"
  2. Randomly sample "controls"
- Under this design, $\Pr(Y_i = 1)$ is known and hence $\Pr(Y_i = 1 \mid X_i)$ is non-parametrically identifiable
- When $\Pr(Y_i = 1)$ is unknown, the odds ratio is still nonparametrically identifiable

- The design extends to the "contaminated" control

Application 2: Ideal Point Estimation

- Originally developed for educational testing: measuring the "ability" of students based on their exam performance
- Political science application: measuring ideology using rollcalls
- Poole and Rosenthal; Clinton, Jackman and Rivers

- Naive approach: count the number of correct answers
- The problem: some questions are easier than others

- The model:

$$\Pr(Y_{ij} = 1 \mid x_i, \alpha_j, \beta_j) \quad = \quad \text{logit}^{-1}(\alpha_j - \beta_j x_i)$$

  where
  - x_i: "ideal point"
  - α_j: difficulty parameter
  - β_j: discrimination parameter
- The key assumption: dimensionality

- Connection to the special theory of voting
  - Quadratic random utilities:

$$\begin{aligned} U_i(\text{yea}) &= -\|x_i - \zeta_j\|^2 + \eta_{ij} \\ U_i(\text{nay}) &= -\|x_i - \psi_j\|^2 + \nu_{ij} \end{aligned}$$

    where $\eta_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ and $\nu_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \omega^2)$
  - Latent utility differential:

$$\begin{aligned} Y_{ij}^* &= U_i(\text{yea}) - U_i(\text{nay}) \\ &= 2(\zeta_j - \xi_j)^\top x_i - \zeta_j^\top \zeta_j + \psi_j^\top \psi_j + \eta_{ij} - \nu_{ij} \\ &= \beta_j^\top x_i - \alpha_j + \epsilon_{ij} \end{aligned}$$

- Identification: scale, rotation
- Estimation: EM algorithm, Markov chain Monte Carlo
- Various extensions to survey, speech, etc.

Ordered Outcome

- The outcome: $Y_i \in \{1, 2, \ldots, J\}$ where $Y_i = 1 \leq Y_i = 3$, etc.
- Assumption: there exists a underlying unidimensional scale
- 5-level Likert scale:

- Ordered logistic regression model:

$$\Pr(Y_i \leq j \mid X_i) = \frac{\exp(\tau_j - X_i^\top \beta)}{1 + \exp(\tau_j - X_i^\top \beta)}$$

for $j = 1, \ldots, J$, which implies,

$$\pi_j(X_i) \equiv \Pr(Y_i = j \mid X_i) = \frac{\exp(\tau_j - X_i^\top \beta)}{1 + \exp(\tau_j - X_i^\top \beta)} - \frac{\exp(\tau_{j-1} - X_i^\top \beta)}{1 + \exp(\tau_{j-1} - X_i^\top \beta)}$$

- Normalization for identification (X_i includes an intercept):
  $\tau_0 = -\infty < \tau_1 = 0 < \tau_2 < \cdots < \tau_{J-1} < \tau_J = \infty$
- Generalization of binary logistic regression

Latent Variable Representation

- Random "utility": $Y_i^* = X_i^\top \beta + \epsilon_i$ where $\epsilon_i \overset{\text{i.i.d.}}{\sim}$ logistic
- If $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$, then the model becomes ordered probit

$$\pi_j(X_i) \;=\; \Phi(\tau_j - X_i^\top \beta) - \Phi(\tau_{j-1} - X_i^\top \beta)$$

- Normalization for variance
- The observation mechanism:

$$Y_i \;=\; \begin{cases} 1 & \text{if} \quad -\infty = \tau_0 \;<\; Y_i^* \;\leq\; \tau_1, \\ 2 & \text{if} \quad \tau_1 = 0 \;<\; Y_i^* \;\leq\; \tau_2, \\ \vdots & \qquad\qquad\quad \vdots \\ J & \text{if} \quad \tau_{J-1} \;<\; Y_i^* \;<\; \tau_J = \infty \end{cases}$$

Inference and Quantities of Interest

- Likelihood function:

$$L(\beta, \tau \mid Y, X) = \prod_{i=1}^{n} \prod_{j=1}^{J} \left\{ \frac{\exp(\tau_j - X_i^\top \beta)}{1 + \exp(\tau_j - X_i^\top \beta)} - \frac{\exp(\tau_{j-1} - X_i^\top \beta)}{1 + \exp(\tau_{j-1} - X_i^\top \beta)} \right\}^{\mathbf{1}\{Y_i = j\}}$$

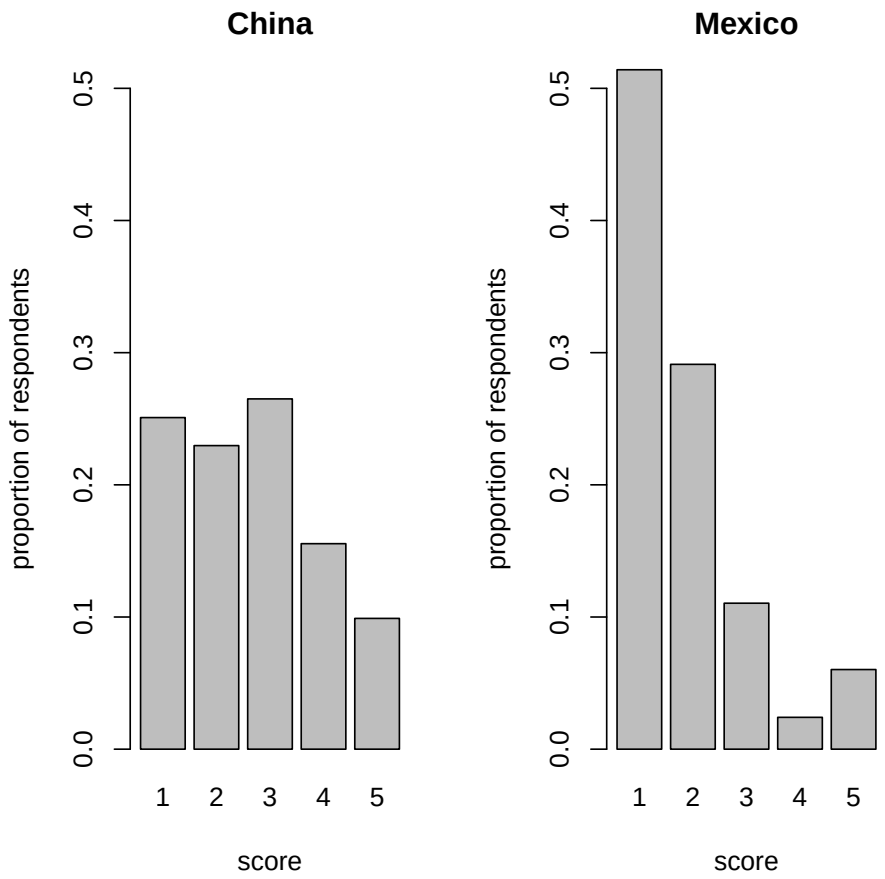
- β itself is difficult to interpret
- Directly calculate the predicted probabilities and other quantities of interest
- Suppose $J = 3$ and $\beta > 0$. Then,

$$\begin{aligned} \frac{\partial}{\partial X_i} \Pr(Y_i = 1 \mid X_i) &< 0 \\ \frac{\partial}{\partial X_i} \Pr(Y_i = 3 \mid X_i) &> 0 \\ \frac{\partial}{\partial X_i} \Pr(Y_i = 2 \mid X_i) &\ ? \ 0 \end{aligned}$$

Differential Item Functioning in Survey Research

- Different respondents may interpret the same questions differently
  - Cross-national surveys $\Longrightarrow$ cultural differences
  - Vague questions $\Longrightarrow$ more room for different interpretation
- Such measurement bias is called differential item functioning (DIF)

- 2002 WHO survey in China and Mexico:
  How much say do you have in getting the government to address issues that interest you?
  1. no say at all
  2. little say
  3. some say
  4. a lot of say
  5. unlimited say



Anchoring to Reduce DIF

- Item Response Theory (IRT) and NOMINATE
- How to bridge across chambers, over time, different actors?
- Key idea: anchoring responses using the same items
- King *et al.* (2004) APSR: anchoring vignettes

Alison lacks clean drinking water. She and her neighbors are supporting an opposition candidate in the forthcoming elections that has promised to address the issue. It appears that so many people in her area feel the same way that the opposition candidate will defeat the incumbent representative.

Jane lacks clean drinking water because the government is pursuing an industrial development plan. In the campaign for an upcoming election, an opposition party has promised to address the issue, but she feels it would be futile to vote for the opposition since the government is certain to win.

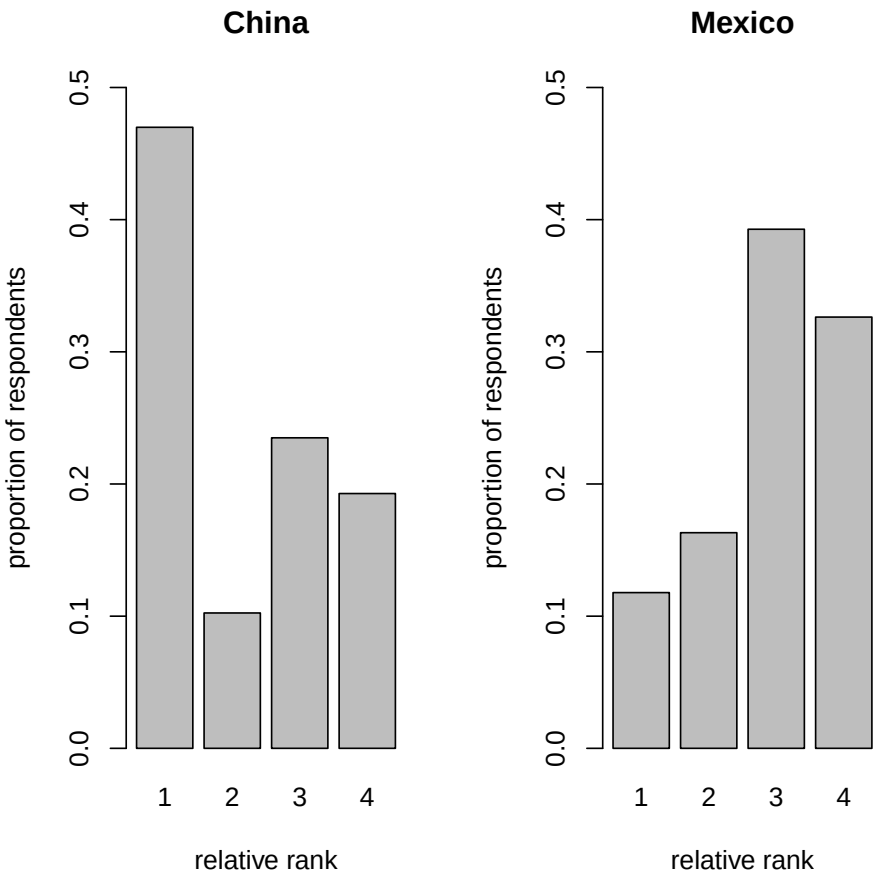
Moses lacks clean drinking water. He would like to change this, but he can't vote, and feels that no one in the government cares about this issue. So he suffers in silence, hoping something will be done in the future.

- The respondent was then asked to assess each vignette in the same manner as the self-assessment question.

- *How much say does Alison/Jane/Moses in getting the government to address issues that interest him/her?*

  1. no say at all
  2. little say
  3. some say
  4. a lot of say
  5. unlimited say

- Plot relative rank of self against vignettes: $4 \geq \text{Alison} > 3 \geq \text{Jane} > 2 \geq \text{Moses} > 1$



Multinomial Outcome

- $Y_i \in \{1, 2, \ldots, J\}$ as before but is not ordered!
- A generalization of binary/ordered logit/probit
- Example: vote choice (abstein, vote for dem., vote for rep.)
- Multinomial logit model:

$$\begin{aligned} \pi_j(X_i) &\equiv \Pr(Y_i = j \mid X_i) \\ &= \frac{\exp(X_i^\top \beta_j)}{\sum_{k=1}^{J} \exp(X_i^\top \beta_k)} \\ &= \frac{\exp(X_i^\top \beta_j)}{1 + \sum_{k=1}^{J-1} \exp(X_i^\top \beta_k)} \end{aligned}$$

- $\beta_J = 0$ for identification: $\sum_{k=1}^{J} \pi_k = 1$

Latent Variable Representation

- Observation mechanism and model:

$$Y_i = j \iff Y_{ij}^* = \max(Y_{i1}^*, Y_{i2}^*, \ldots, Y_{iJ}^*)$$
$$Y_{ij}^* = X_i^\top \beta_j + \epsilon_{ij}$$

- ϵ_{ij} has Type I extreme-value distribution:

$$F(\epsilon) = \exp\{-\exp(-\epsilon)\}$$
$$f(\epsilon) = \exp\{-\epsilon - \exp(-\epsilon)\}$$

- McFadden's Proof:

$$\Pr(Y_i = j \mid X_i) = \prod_{j' \neq j} \Pr(Y_{ij}^* > Y_{ij'}^* \mid X_i) = \prod_{j' \neq j} \Pr\{\epsilon_{ij'} < \epsilon_{ij} + X_i^\top (\beta_j - \beta_{j'})\}$$
$$= \int_{-\infty}^{\infty} \left[\prod_{j' \neq j} F\{\epsilon_{ij} + X_i^\top (\beta_j - \beta_{j'})\} \right] f(\epsilon_{ij}) d\epsilon_{ij}$$
$$= \frac{\exp(X_i^\top \beta_j)}{\sum_{j'=1}^{J} \exp(X_i^\top \beta_{j'})}$$

Conditional Logit Model

- A further generalization:

$$\pi_j(X_{ij}) \equiv \Pr(Y_i = j \mid X_{ij}) = \frac{\exp(X_{ij}^\top \beta)}{\sum_{k=1}^{J} \exp(X_{ik}^\top \beta)}$$

- Subject-specific and choice-specific covariates, and their interactions
- Multinomial logit model as a special case:

$$X_{i1} = \begin{pmatrix} X_i \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \quad X_{i2} = \begin{pmatrix} 0 \\ X_i \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \quad \cdots, \quad X_{iJ} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ X_i \end{pmatrix}$$

- Some restrictions are necessary for identification: for example, one cannot include a different intercept for each category

Multinomial Probit Model

- IIA (Independence of Irrelevant Alternatives)
- Multinomial/Conditional logit

$$\frac{\Pr(Y_i = j \mid X_{ij})}{\Pr(Y_i = j' \mid X_{ij'})} = \exp\{(X_{ij} - X_{ij'})^\top \beta\}$$

- blue vs. red bus; Chicken vs. Fish
- MNP: Allowing for the dependence among errors

$$Y_i = j \iff Y_{ij}^* = \max(Y_{i1}^*, Y_{i2}^*, \ldots, Y_{iJ}^*)$$

$$\underbrace{Y_i^*}_{J\times 1} = \underbrace{X_{ij}^\top}_{J\times K}\underbrace{\beta}_{K\times 1} + \underbrace{\epsilon_i}_{J\times 1} \quad \text{where} \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \underbrace{\Sigma}_{J\times J})$$

Identification and Inference

- Two additional steps for identification:
  1. Subtract the Jth equation from the other equations:

$$W_i = Z_i^\top \beta + \eta_i \quad \text{where} \quad \eta_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Lambda)$$

     where $W_{ij} = Y_{ij}^* - Y_{iJ}^*$, $Z_{ij} = X_{ij} - X_{iJ}$, and $\Lambda = [I_{J-1}, -\mathbf{1}_{J-1}]\Sigma[I_{J-1}, -\mathbf{1}_{J-1}]$
  2. Set $\Lambda_{11} = 1$

- Likelihood function for unit i who selects the Jth category:

$$L(\beta, \Lambda \mid X_i, Y_i) = \int_{-\infty}^{-Z_{i1}^\top \beta} \cdots \int_{-\infty}^{-Z_{i,J-1}^\top \beta} f(\eta_i \mid \Lambda) d\eta_{i1} \cdots d\eta_{i,J-1}$$

- High-dimensional integration $\longrightarrow$ Bayesian MCMC

Other Discrete Choice Models

- Nested Multinomial Logit: Modeling the first choice $j \in \{1, 2, \ldots, J\}$ and the second choice given the first $k \in \{1, 2, \ldots, K_j\}$

$$\Pr(Y = (j, k) \mid X_i) = \frac{\exp(X_{ijk}^\top \beta)}{\sum_{j'=1}^{J} \sum_{k'=1}^{K_{j'}} \exp(X_{ij'k'} \beta)}$$

- Multivariate Logit/Probit: Modeling multiple correlated choice $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iJ})$ where $Y_{ij} = \mathbf{1}\{Y_{ij}^* > 0\}$ and

$$Y_i^* = X_i^\top \beta + \epsilon_i \quad \text{where} \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma)$$

where $\Sigma_{jj} = 1$ and $\Sigma_{jj'} = \rho_{jj'}$ for all j and $j' \neq j$

Optimization Using the *EM* Algorithm

- The Expectation and Maximization algorithm by Dempster, Laird, and Rubin: Google scholar 30,000 citations!
- Useful for maximizing the likelihood function with missing data
- Pedagogical reference: S. Jackman (*AJPS*, 2000)

- Goal: maximize the observed-data log-likelihood, $l_n(\theta \mid Y_{obs})$
- The *EM* algorithm: Repeat the following steps until convergence
  1. *E*-step: Compute

$$Q(\theta \mid \theta^{(t)}) \equiv \mathbb{E}\{l_n(\theta \mid Y_{obs}, Y_{mis}) \mid Y_{obs}, \theta^{(t)}\}$$

     where $l_n(\theta \mid Y_{obs}, Y_{mis})$ is the complete-data log-likelihood
  2. *M*-step: Find

$$\theta^{(t+1)} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, Q(\theta \mid \theta^{(t)})$$

- The *ECM* algorithm: *M*-step replaced with multiple conditional maximization steps

Monotone Convergence Property

- The observed-data likelihood increases each step:

$$l_n(\theta^{(t+1)} \mid Y_{obs}) \geq l_n(\theta^{(t)} \mid Y_{obs})$$

- "Proof":
  1. $l_n(\theta \mid Y_{obs}) = \log f(Y_{obs}, Y_{mis} \mid \theta) - \log f(Y_{mis} \mid Y_{obs}, \theta)$
  2. Taking the expectation w.r.t. $f(Y_{mis} \mid Y_{obs}, \theta^{(t)})$

$$l_n(\theta \mid Y_{obs}) = Q(\theta \mid \theta^{(t)}) - \int \log f(Y_{mis} \mid Y_{obs}, \theta) f(Y_{mis} \mid Y_{obs}, \theta^{(t)}) dY_{mis}$$

  3. Finally,

$$\begin{aligned} & l_n(\theta^{(t+1)} \mid Y_{obs}) - l_n(\theta^{(t)} \mid Y_{obs}) \\ = \; & Q(\theta^{(t+1)} \mid \theta^{(t)}) - Q(\theta^{(t)} \mid \theta^{(t)}) \\ & + \int \log \frac{f(Y_{mis} \mid Y_{obs}, \theta^{(t)})}{f(Y_{mis} \mid Y_{obs}, \theta^{(t+1)})} f(Y_{mis} \mid Y_{obs}, \theta^{(t)}) dY_{mis} \\ \geq \; & 0 \end{aligned}$$

- Stable, no derivative required

Application 1: A Finite Mixture Model

- Used for flexible modeling and clustering in statistics
- Building block for many advanced models
- Can be used to test competing theories (Imai & Tingley *AJPS*)
- M competing theories, each of which implies a statistical model $f_m(y \mid x, \theta_m)$ for $m = 1, \ldots, M$
- The data generating process:

$$Y_i \mid X_i, Z_i \quad \sim \quad f_{Z_i}(Y_i \mid X_i, \theta_{Z_i})$$

  where Z_i is the *latent* variable indicating the theory which generates observation i
- Probability that an observation is generated by theory m:

$$\pi_m(X_i, \psi_m) \quad = \quad \Pr(Z_i = m \mid X_i)$$

Observed-Data and Complete-Data Likelihoods

- The observed-data likelihood function:

$$L_{obs}(\Theta, \Psi \mid \{X_i, Y_i\}_{i=1}^N) = \prod_{i=1}^N \left\{ \sum_{m=1}^M \pi_m(X_i, \psi_m) f_m(Y_i \mid X_i, \theta_m) \right\},$$

- The complete-data likelihood function:

$$\begin{aligned} & L_{com}(\Theta, \Psi \mid \{X_i, Y_i, Z_i\}_{i=1}^N) \\ = & \prod_{i=1}^N \pi_{Z_i}(X_i, \psi_{Z_i}) f_{Z_i}(Y_i \mid X_i, \theta_{Z_i}) \\ = & \prod_{i=1}^N \prod_{m=1}^M \{\pi_m(X_i, \psi_m) f_m(Y_i \mid X_i, \theta_m)\}^{\mathbf{1}\{Z_i = m\}} \end{aligned}$$

Estimation via the EM Algorithm

- The E–Step:

$$\begin{aligned}
\zeta_{i,m}^{(t-1)} &= \Pr(Z_i = m \mid \Theta^{(t-1)}, \Psi^{(t-1)}, \{X_i, Y_i\}_{i=1}^N) \\
&= \frac{\pi_m(X_i, \psi_m^{(t-1)})\, f_m(Y_i \mid X_i, \theta_m^{(t-1)})}{\sum_{m'=1}^M \pi_{m'}(X_i, \psi_{m'}^{(t-1)})\, f_{m'}(Y_i \mid X_i, \theta_{m'}^{(t-1)})}
\end{aligned}$$

  and thus, the Q function is,

$$\sum_{i=1}^N \sum_{m=1}^M \zeta_{i,m}^{(t-1)} \left\{ \log \pi_m(X_i, \psi_m^{(t)}) + \log f_m(Y_i \mid X_i, \theta_m^{(t)}) \right\}$$

- The M–Step: $2 \times M$ weighted regressions!

Application 2: Sample Selection Model

- Non-random sampling: $S_i = 1$ if unit i is in the sample, $S_i = 0$ otherwise
- The outcome model: $Y_i = X_i^\top \beta + \epsilon_i$
- The selection model: $S_i = \mathbf{1}\{S_i^* > 0\}$ with $S_i^* = X_i^\top \gamma + \eta_i$
- Selection bias:

$$\mathbb{E}(Y_i \mid X_i, S_i = 1) = X_i^\top \beta + \mathbb{E}(\epsilon_i \mid X_i, \eta_i > -X_i^\top \gamma) \neq \mathbb{E}(Y_i \mid X_i)$$

- Inverse Mill's ratio (under normality $\begin{pmatrix} \epsilon_i \\ \eta_i \end{pmatrix} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}\right]$):

$$\mathbb{E}(\epsilon_i \mid X_i, \eta_i > -X_i^\top \gamma) = \rho\sigma \frac{\phi(X_i^\top \gamma)}{\Phi(X_i^\top \gamma)}$$

- Sample selection as a specification error
- Exclusion restriction needed for "nonparametric" identification
- Sensitive to changes in the assumptions: Y_i is unobserved for $S_i = 0$

Application to the Heckman's Selection Model

- Bivariate Normal as a complete-data model:

$$\begin{pmatrix} Y_i \\ S_i^* \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} X_i^\top \beta \\ X_i^\top \gamma \end{pmatrix}, \begin{pmatrix} \sigma^2 & \sigma\rho \\ \sigma\rho & 1 \end{pmatrix} \right)$$

- Factoring the bivariate normal

$$\underbrace{\mathcal{N}(X_i^\top \gamma,\ 1)}_{f(S_i^*|X_i)} \times \underbrace{\mathcal{N}(X_i^\top \beta + \rho\sigma(S_i^* - X_i^\top \gamma),\ \sigma^2(1-\rho^2))}_{f(Y_i|S_i^*,X_i)}$$

- The E–Step:
  - Sufficient statistics: Y_i, Y_i^2 for $S_i = 0$, and S_i^*, ${S_i^*}^2$, $Y_i S_i^*$ for all
  - Compute the conditional expectation of sufficient statistics given all observed data and parameters
- The M–Step:
  - Run two regression with the results from the E-step
  - Regress S_i^* on X_i; Regress Y_i on S_i^* and X_i

Application 3: Topic Modeling

- Clustering documents based on topics:
  - word or term: basic unit of data
  - document: a sequence of words
  - corpus: a collection of documents
- Supervised and unsupervised learning approaches

- The "bag-of-words" assumption: term-document matrix
- $\texttt{tf} - \texttt{idf}(w, d) = \texttt{tf}(w, d) \times \texttt{idf}(w)$
  1. term frequency $\texttt{tf}(w, d)$: frequency of term w in document d
  2. document frequency $\texttt{df}(w)$: # of documents that contain term w
  3. inverse document frequency $\texttt{idf}(w) = \log \frac{N}{\texttt{df}(w)}$ where N is the total number of documents
- $\texttt{tf} - \texttt{idf}(w, d)$ is:
  1. highest when w occurs many times within a small number of documents
  2. lower when w occurs fewer times in a document or occurs in many documents

Latent Dirichlet Allocation (LDA)

- Probabilistic modeling $\Longrightarrow$ statistical inference
- Notation:
  - Documents: $i = 1, \ldots, N$
  - Number of words in document i: M_i
  - A sequence of words in document i: $W_{i1}, \ldots, W_{iM_i}$
  - Number of unique words: K
  - Latent topic for the jth word in document i: $Z_{ij} \in \{1, \ldots, L\}$
- "topic" $=$ a probability distribution over word frequencies
- a document belongs to a mixture of topics: mixed-membership

- The Model:

$$
\begin{aligned}
W_{ij} \mid Z_{ij} = z &\overset{\text{indep.}}{\sim} \text{Multinomial}(K, \beta_z) \\
Z_{ij} &\overset{\text{indep.}}{\sim} \text{Multinomial}(L, \theta_i) \\
\theta_i &\overset{\text{i.i.d.}}{\sim} \text{Dirichlet}(\alpha)
\end{aligned}
$$

1. distribution of words within each topic β_z
2. topic mixture for each document θ_i

Likelihood Inference

- Likelihood function:

$$
\begin{aligned}
& p(W \mid \alpha, \beta) \\
= & \int \sum_{Z} p(W, Z, \theta \mid \alpha, \beta) d\theta \\
= & \prod_{i=1}^{N} \int p(\theta_i \mid \alpha) \sum_{z} \prod_{j=1}^{M_i} p(W_{ij} \mid Z_{ij} = z, \beta_z) p(Z_{ij} = z \mid \theta_i) d\theta_i
\end{aligned}
$$

- Log-likelihood function:

$$
\sum_{i=1}^{N} \log \int p(\theta_i \mid \alpha) \sum_{z} \prod_{j=1}^{M_i} p(W_{ij} \mid Z_{ij} = z, \beta_z) p(Z_{ij} = z \mid \theta_i) d\theta_i
$$

- Maximize it to obtain ML or Bayesian estimate of (α, β)
- Markov chain Monte Carlo

Variational Inference

- Approximation: maximize the lower bound of log-likelihood
- The variational distribution with the factorization assumption:

$$q(\theta, Z \mid \gamma, \psi) \;=\; \prod_{i=1}^{N} q(\theta_i \mid \gamma_i) \times \prod_{i=1}^{N} \prod_{j=1}^{M_i} q(Z_{ij} \mid \psi_{ij})$$

  where (γ, ψ) are variational parameters
- Choose (γ^*, ψ^*) s.t. $q(\theta, Z \mid \gamma^*, \psi^*)$ is "closest" to $p(\theta, Z, W \mid \alpha, \beta)$
- Kullback-Leibler divergence:

$$\mathbb{E}_q \left\{ \log \frac{q(\theta, Z \mid \gamma, \psi)}{p(\theta, Z, W \mid \alpha, \beta)} \right\} \;\geq\; 0$$

- Advantage: easy and fast
- Disadvantage: approximation may be poor
  - appropriateness of factorization assumption
  - choice of factorization distributions

Variational EM for LDA

- The lower bound for the log-likelihood function:

$$
\begin{aligned}
& \log p(W \mid \alpha, \beta) \\
= \; & \log \int \sum_{Z} p(\theta, Z, W \mid \alpha, \beta) d\theta \\
= \; & \log \int \sum_{Z} \frac{p(\theta, Z, W \mid \alpha, \beta)}{q(\theta, Z \mid \gamma, \psi)} q(\theta, Z \mid \gamma, \psi) d\theta \\
\geq \; & \mathbb{E}_q \left\{ \log \frac{p(\theta, Z, W \mid \alpha, \beta)}{q(\theta, Z \mid \gamma, \psi)} \right\} \quad \text{(Jensen's inequality)} \\
= \; & -\mathbb{E}_q \left\{ \log \frac{q(\theta, Z \mid \gamma, \psi)}{p(\theta, Z, W \mid \alpha, \beta)} \right\}
\end{aligned}
$$

- Maximizing the lower bound $=$ minimizing the KL divergence

- E-step:

$$
\begin{aligned}
& \text{The lower bound} \\
= \; & \mathbb{E}_q\{\log p(\theta, Z, W \mid \alpha, \beta)\} - \mathbb{E}_q\{\log q(\theta, Z \mid \gamma, \psi)\} \\
= \; & \mathbb{E}_q\{\log p(\theta \mid \alpha)\} + \mathbb{E}_q\{\log p(Z \mid \theta)\} + \mathbb{E}_q\{\log p(W \mid Z, \beta)\} \\
& -\mathbb{E}_q\{\log q(\theta \mid \gamma)\} - \mathbb{E}_q\{\log q(Z \mid \psi)\}
\end{aligned}
$$

  Evaluate each expectation analytically

- The M–step: maximize this function using coordinate ascent
  1. maximize the bound w.r.t. (γ, ψ) given (α, β)
  2. maximize the bound w.r.t. (α, β) given (γ, ψ)

Derivation of the Variational Distribution

- Consider optimization with respect to $q(\theta)$:

$$\begin{aligned} \text{ELBO} &= \mathbb{E}_\theta[\mathbb{E}_Z\{\log p(\theta, W, Z \mid \alpha, \beta)\}] - \mathbb{E}_\theta\{\log q(\theta \mid \gamma)\} + \text{const.} \\ &= \mathbb{E}_\theta\{\log \tilde{p}(\theta, W \mid \alpha, \beta) - \log q(\theta \mid \gamma)\} + \text{const.} \end{aligned}$$

  where $\log \tilde{p}(\theta, W \mid \alpha, \beta) = \mathbb{E}_Z\{\log p(\theta, W, Z \mid \alpha, \beta)\} + \text{const.}$

- This is a negative KL divergence between $\tilde{p}(\theta, W \mid \alpha, \beta)$ and $q(\theta \mid \gamma)$
- We can maximize this quantity when $q(\theta \mid \gamma) = \tilde{p}(\theta, W \mid \alpha, \beta)$
- Thus, the optimal variational distribution is given by:

$$q(\theta \mid \gamma) = \mathbb{E}_Z\{\log p(\theta, W, Z \mid \alpha, \beta)\} + \text{const.}$$

- We can do the same for $q(Z \mid \psi)$
- This is a general result for any model