Item Response Theory Models

Philip D. Waggoner

MACS 40500: Computational Methods for American Politics

November 12, 2019

Lecture Outline

Our Guiding Question

Our Guiding Question

How can we measure something we know exists and influences observed behavior, but we can't directly observe?

Challenges in Estimating Ideal Points via NOMINATE

- Large-N algorithm requiring tons of data to converge

Challenges in Estimating Ideal Points via NOMINATE

- Large-N algorithm requiring tons of data to converge
  - NOMINATE experiences challenges in cases with fewer than 50 legislators, where there are few recorded roll calls, or both

Challenges in Estimating Ideal Points via NOMINATE

- Large-N algorithm requiring tons of data to converge
  - NOMINATE experiences challenges in cases with fewer than 50 legislators, where there are few recorded roll calls, or both
  - Drops lopsided roll calls but in doing so, loss in discrimination between the most exteme legislators on either side

Challenges in Estimating Ideal Points via NOMINATE

- Large-N algorithm requiring tons of data to converge
  - NOMINATE experiences challenges in cases with fewer than 50 legislators, where there are few recorded roll calls, or both
  - Drops lopsided roll calls but in doing so, loss in discrimination between the most exteme legislators on either side
  - Legislators with fewer than 25 roll calls are dropped in NOMINATE–can be problematic in certain settings

Challenges in Estimating Ideal Points via NOMINATE

- Large-N algorithm requiring tons of data to converge
  - NOMINATE experiences challenges in cases with fewer than 50 legislators, where there are few recorded roll calls, or both
  - Drops lopsided roll calls but in doing so, loss in discrimination between the most exteme legislators on either side
  - Legislators with fewer than 25 roll calls are dropped in NOMINATE–can be problematic in certain settings
- NOMINATE uses ad hoc constraints to identify the model (polarity and unit hypersphere)

Challenges in Estimating Ideal Points via NOMINATE

- Large-N algorithm requiring tons of data to converge
  - NOMINATE experiences challenges in cases with fewer than 50 legislators, where there are few recorded roll calls, or both
  - Drops lopsided roll calls but in doing so, loss in discrimination between the most exteme legislators on either side
  - Legislators with fewer than 25 roll calls are dropped in NOMINATE–can be problematic in certain settings
- NOMINATE uses ad hoc constraints to identify the model (polarity and unit hypersphere)
- Does not enable incorporation of auxiliary information

Lecture Outline

The Classic Case

- *How can we measure learning ability/intelligence using responses to multiple choice test instruments?*

The Classic Case

- *How can we measure learning ability/intelligence using responses to multiple choice test instruments?*
- **The Problem**: we can see the questions students get correct, but the responses are only *related* to intelligence (e.g., easy vs. hard; math vs. reading vs. problem-solving, etc.)

The Classic Case

- *How can we measure learning ability/intelligence using responses to multiple choice test instruments?*
- **The Problem**: we can see the questions students get correct, but the responses are only *related* to intelligence (e.g., easy vs. hard; math vs. reading vs. problem-solving, etc.)
- **The Task**: recover student intelligence from sets of test responses

The Classic Case

- *How can we measure learning ability/intelligence using responses to multiple choice test instruments?*
- **The Problem**: we can see the questions students get correct, but the responses are only *related* to intelligence (e.g., easy vs. hard; math vs. reading vs. problem-solving, etc.)
- **The Task**: recover student intelligence from sets of test responses
- IRT was designed to solve this problem

Since then...

Since then...

- How does ideology inform voting behavior in legislatures?
  - MCs vote according to many considerations, most notably, *what they believe (ideology)*
  - But we can't simply "read" their ideologies
  - So we want to recover their ideologies in some way, just like the testing case

Since then...

- How does ideology inform voting behavior in legislatures?
  - MCs vote according to many considerations, most notably, *what they believe (ideology)*
  - But we can't simply "read" their ideologies
  - So we want to recover their ideologies in some way, just like the testing case
- Recovering SCOTUS ideology from case decisions
  - *Attitudinalism*
  - If this theory is correct, then we should be able to look at SCOTUS decision records and recover some measure of ideology

Item Response *Theory*

- The SCOTUS case, e.g. $\rightsquigarrow$ *theory bound*

Item Response *Theory*

- The SCOTUS case, e.g. $\rightsquigarrow$ *theory bound*
  - Measurements only can only be recovered by assuming people act based on a theory

Item Response *Theory*

- The SCOTUS case, e.g. $\rightsquigarrow$ *theory bound*
  - Measurements only can only be recovered by assuming people act based on a theory
  - We are interested in measuring whether characteristics are consistent across behavior on the basis of that theory

Item Response *Theory*

- The SCOTUS case, e.g. $\rightsquigarrow$ *theory bound*
  - Measurements only can only be recovered by assuming people act based on a theory
  - We are interested in measuring whether characteristics are consistent across behavior on the basis of that theory
- Thus, the measures from IRT are *not atheoretic*, which limits *how* they can be used

Item Response *Theory*

- The SCOTUS case, e.g. $\rightsquigarrow$ *theory bound*
  - Measurements only can only be recovered by assuming people act based on a theory
  - We are interested in measuring whether characteristics are consistent across behavior on the basis of that theory
- Thus, the measures from IRT are *not atheoretic*, which limits *how* they can be used
  - We can't measure assuming some theory is true, and then use that measure to test the theory (e.g., $Y \rightarrow Y$)

Item Response *Theory*

- Yet, within these limitations, IRT measures can be extremely useful

Item Response *Theory*

- Yet, within these limitations, IRT measures can be extremely useful
- The intuition: we look at a bunch of cases of different types of decisions, and then pick latent characteristics of people and their decisions that are most consistent with the observed patterns we see in their decisions

Item Response *Theory*

- Yet, within these limitations, IRT measures can be extremely useful
- The intuition: we look at a bunch of cases of different types of decisions, and then pick latent characteristics of people and their decisions that are most consistent with the observed patterns we see in their decisions
- Theoretically similar to MLE

Item Response Theory: The Rasch Model

- Basic Rasch, 1 parameter model (1PL) model

Item Response Theory: The Rasch Model

- Basic Rasch, 1 parameter model (1PL) model
- The idea $\rightsquigarrow$ a latent dimension that influences observed behavior

Item Response Theory: The Rasch Model

- Basic Rasch, 1 parameter model (1PL) model
- The idea $\rightsquigarrow$ a latent dimension that influences observed behavior
- There are two things on this scale:

Item Response Theory: The Rasch Model

- Basic Rasch, 1 parameter model (1PL) model
- The idea $\rightsquigarrow$ a latent dimension that influences observed behavior
- There are two things on this scale: the decision maker and the items

Item Response Theory: The Rasch Model

- Basic Rasch, 1 parameter model (1PL) model
- The idea $\rightsquigarrow$ a latent dimension that influences observed behavior
- There are two things on this scale: the decision maker and the items
- Each person looks at their position and then the possible vote and asks, "should I do this or not?"

Item Response Theory: The Rasch Model

- Basic Rasch, 1 parameter model (1PL) model
- The idea $\rightsquigarrow$ a latent dimension that influences observed behavior
- There are two things on this scale: the decision maker and the items
- Each person looks at their position and then the possible vote and asks, "should I do this or not?"
- Negative positions on the scale mean you shouldn't support the proposal; positive positions mean you should

Item Response Theory: The Rasch Model

- Basic Rasch, 1 parameter model (1PL) model
- The idea $\rightsquigarrow$ a latent dimension that influences observed behavior
- There are two things on this scale: the decision maker and the items
- Each person looks at their position and then the possible vote and asks, "should I do this or not?"
- Negative positions on the scale mean you shouldn't support the proposal; positive positions mean you should
- By implication, the greater that distance, the more you benefit from saying yes or no (positive or negative, respectively)

Item Response Theory: The Rasch Model

- The location of the item forms a cut point where people on one side should say yes, and people on the other side should say no

Item Response Theory: The Rasch Model

- The location of the item forms a cut point where people on one side should say yes, and people on the other side should say no
- While we could say this cut-point denotes a *true* cutpoint, in reality decisions are likely probabilistic

Item Response Theory: The Rasch Model

- The location of the item forms a cut point where people on one side should say yes, and people on the other side should say no
- While we could say this cut-point denotes a *true* cutpoint, in reality decisions are likely probabilistic
- So instead we say, your probability of saying yea on a vote is determine by some link function (normal, e.g.), and is a function of your distance to the vote cutpoint $(\theta_i - \beta)$

Item Response Theory: The Rasch Model

- The location of the item forms a cut point where people on one side should say yes, and people on the other side should say no
- While we could say this cut-point denotes a *true* cutpoint, in reality decisions are likely probabilistic
- So instead we say, your probability of saying yea on a vote is determine by some link function (normal, e.g.), and is a function of your distance to the vote cutpoint $(\theta_i - \beta)$
- If the result is positive, then you have a greater chance of making *yea* decision; if *negative*, then you have less chance of making yea decision

Item Response Theory: The Rasch Model

- With no normalization, at 0 there's a 50% of saying yes; as the $(\theta - \beta)$ index declines, you have a lesser chance of saying yes, and increased index means a greater chance of saying yes

Item Response Theory: The Rasch Model

- With no normalization, at 0 there's a 50% of saying yes; as the $(\theta - \beta)$ index declines, you have a lesser chance of saying yes, and increased index means a greater chance of saying yes

- That's 1PL:

Item Response Theory: The Rasch Model

- With no normalization, at 0 there's a 50% of saying yes; as the $(\theta - \beta)$ index declines, you have a lesser chance of saying yes, and increased index means a greater chance of saying yes
- That's 1PL: we map positions into decision probabilities, where positions include positions for both the decision makers and also of the indifference, "difficulty" points (where $0.5 =$ indifferent between options)

Item Response Theory: The Rasch Model

- With no normalization, at 0 there's a 50% of saying yes; as the $(\theta - \beta)$ index declines, you have a lesser chance of saying yes, and increased index means a greater chance of saying yes
- That's 1PL: we map positions into decision probabilities, where positions include positions for both the decision makers and also of the indifference, "difficulty" points (where 0.5 = indifferent between options)
- Thus, we are interested in calculating predictions of being over or below that line

Identification

- But just calculating these probabilities isn't enough; this brings us to identification constraints in the Rasch model

Identification

- But just calculating these probabilities isn't enough; this brings us to identification constraints in the Rasch model

- **The Problem**: Technically the Rasch model is unidentified, which means that each parameter doesn't have a unique solution; thus, there needs to be a *unique solution* rather than many potential solutions

Identification

- But just calculating these probabilities isn't enough; this brings us to identification constraints in the Rasch model

- **The Problem**: Technically the Rasch model is unidentified, which means that each parameter doesn't have a unique solution; thus, there needs to be a *unique solution* rather than many potential solutions

- **The Solution**: We can force idenitification by forcing priors for the individual locations; this anchors the position of the locations, which concentrates estimates around the (normal) prior (hence, *Bayesian* IRT models; more in a bit)

Identification

- But just calculating these probabilities isn't enough; this brings us to identification constraints in the Rasch model

- **The Problem**: Technically the Rasch model is unidentified, which means that each parameter doesn't have a unique solution; thus, there needs to be a *unique solution* rather than many potential solutions

- **The Solution**: We can force idenitification by forcing priors for the individual locations; this anchors the position of the locations, which concentrates estimates around the (normal) prior (hence, *Bayesian* IRT models; more in a bit)

- **Another Solution**: in multiple parameters models (e.g., 2PL), we set a few representative actors as fixed values to anchor the latent scale, which allows for *relative* locations which solves the identification problem

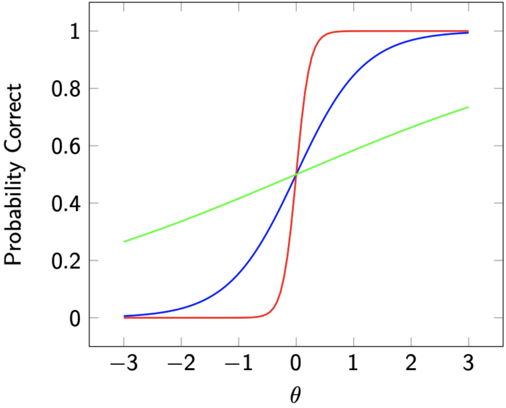
2PL

$$P(y_i = 1|\theta;\beta) = \frac{exp(D(\theta - \alpha))}{1 + exp(D(\theta - \beta))}$$

where

- θ is the ability (latent trait)
- α is the item discrimination parameter
- β is the item difficulty parameter
- D is a scaling factor, where set to ≈ 1.7 results in the logistic model behaving like standard normal case (thus you sometimes see the Rasch model as $P(y_i = 1|\theta;\beta) = \Phi(\theta - \beta)$)

2PL: Difficulty & Discrimination



Summing Up the Idea Behind IRT

- Subjects are assumed to possess some latent level of ability (or ideology, or democracy, etc.) that we believe can be measured through a series of items (e.g., test scores, survey questions, etc.)

Summing Up the Idea Behind IRT

- Subjects are assumed to possess some latent level of ability (or ideology, or democracy, etc.) that we believe can be measured through a series of items (e.g., test scores, survey questions, etc.)

- Each subject is predicted to answer the question *correctly* (or in a conservative/liberal manner, or in a way indicating higher levels of democracy, etc.) if they have a level of ability above some threshold; and *incorrectly* otherwise

Summing Up the Idea Behind IRT

- Subjects are assumed to possess some latent level of ability (or ideology, or democracy, etc.) that we believe can be measured through a series of items (e.g., test scores, survey questions, etc.)
- Each subject is predicted to answer the question *correctly* (or in a conservative/liberal manner, or in a way indicating higher levels of democracy, etc.) if they have a level of ability above some threshold; and *incorrectly* otherwise
- The threshold (indifference point) between "correct" and "incorrect" is known as the item **difficulty** parameter

Summing Up the Idea Behind IRT

- Subjects are assumed to possess some latent level of ability (or ideology, or democracy, etc.) that we believe can be measured through a series of items (e.g., test scores, survey questions, etc.)
- Each subject is predicted to answer the question *correctly* (or in a conservative/liberal manner, or in a way indicating higher levels of democracy, etc.) if they have a level of ability above some threshold; and *incorrectly* otherwise
- The threshold (indifference point) between "correct" and "incorrect" is known as the item **difficulty** parameter
- IRT models involve the joint estimation of parameters: **difficulty** (1PL) and **discrimination** (2PL)

Item Characteristic Curve

- The Item Characteristic Curve is the primary concept of IRT

Item Characteristic Curve

- The Item Characteristic Curve is the primary concept of IRT
  - Consider the trait θ

Item Characteristic Curve

- The Item Characteristic Curve is the primary concept of IRT
  - Consider the trait θ
  - We have a binary item

Item Characteristic Curve

- The Item Characteristic Curve is the primary concept of IRT
  - Consider the trait θ
  - We have a binary item
  - We encode $Y_i = 1$ if the question is answered “correctly” and 0 otherwise

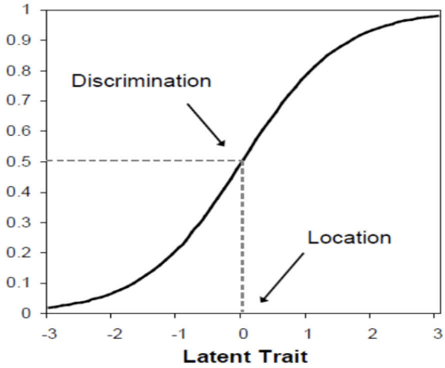
Item Characteristic Curve

- The Item Characteristic Curve is the primary concept of IRT
  - Consider the trait θ
  - We have a binary item
  - We encode $Y_i = 1$ if the question is answered "correctly" and 0 otherwise
  - We relate $P(Y_i = 1)$ to θ, assuming a monotonic, increasing function (e.g., logit)

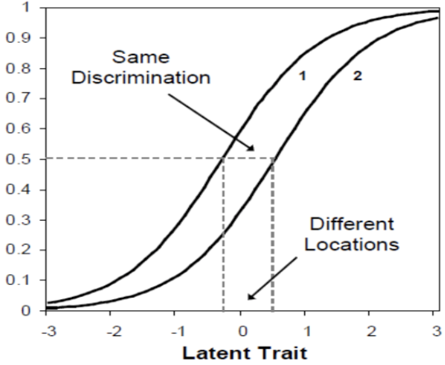
Item Characteristic Curve

- The Item Characteristic Curve is the primary concept of IRT
  - Consider the trait θ
  - We have a binary item
  - We encode $Y_i = 1$ if the question is answered "correctly" and 0 otherwise
  - We relate $P(Y_i = 1)$ to θ, assuming a monotonic, increasing function (e.g., logit)
- Thus, the ICC connects or links the subject's probability of success on an item to the trait measured by the set of test items

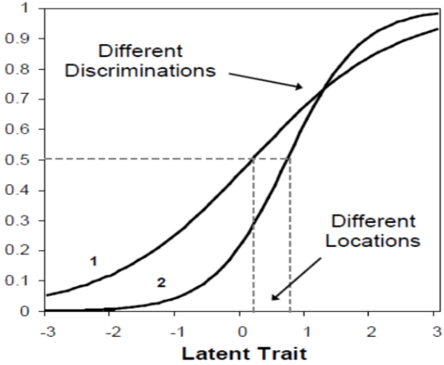
ICC: The intuition



ICC: The intuition



ICC: The intuition



Lecture Outline

Frequentist Approaches

- In all statistics, we have observed data y and unknown parameters θ (and possible inclusion of fixed covariates x)

Frequentist Approaches

- In all statistics, we have observed data y and unknown parameters θ (and possible inclusion of fixed covariates x)
- Frequentist approaches treat observed data as random variables

Frequentist Approaches

- In all statistics, we have observed data y and unknown parameters θ (and possible inclusion of fixed covariates x)
- Frequentist approaches treat observed data as random variables
- Underlying parameter is fixed

Frequentist Approaches

- In all statistics, we have observed data y and unknown parameters θ (and possible inclusion of fixed covariates x)
- Frequentist approaches treat observed data as random variables
- Underlying parameter is fixed
- Coin flip is a random experiment; underlying probability of heads is fixed

Frequentist Approaches

- In all statistics, we have observed data y and unknown parameters θ (and possible inclusion of fixed covariates x)
- Frequentist approaches treat observed data as random variables
- Underlying parameter is fixed
- Coin flip is a random experiment; underlying probability of heads is fixed
- Data we see are the product of a single trial of repeatable experiment

Bayesian Approaches

- Bayesians treat parameters as random variables

Bayesian Approaches

- Bayesians treat parameters as random variables
- Data are taken as given

Bayesian Approaches

- Bayesians treat parameters as random variables
- Data are taken as given
- Using Bayes' rule, we update beliefs about the parameter values, conditional on the data we get to see

An Example: The Linear Model

$$y_i|\beta, \sigma^2, x_i \sim N(x_i\beta, \sigma^2) \tag{1}$$

An Example: The Linear Model

$$y_i|\beta, \sigma^2, x_i \sim N(x_i\beta, \sigma^2) \tag{1}$$

- How can we estimate this model?

An Example: The Linear Model

$$y_i|\beta, \sigma^2, x_i \sim N(x_i\beta, \sigma^2) \quad (1)$$

- How can we estimate this model?
- OLS, ML, Bayesian

An Example: The Linear Model

$$y_i|\beta, \sigma^2, x_i \sim N(x_i\beta, \sigma^2) \tag{1}$$

- How can we estimate this model?
- OLS, ML, Bayesian
- if Gauss-Markov holds $[X'X]^{-1}X'y$ is BLUE

A Maximum Likelihood Approach

$$\mathcal{L}(\beta, \sigma^2 | y) = \prod_{i=1}^{n} \phi \left(\frac{y_i - x\beta}{\sigma} \right) \tag{2}$$

A Maximum Likelihood Approach

$$\mathcal{L}(\beta, \sigma^2 | y) = \prod_{i=1}^{n} \phi\left(\frac{y_i - x\beta}{\sigma}\right) \tag{2}$$

- We maximize $\mathcal{L}(\cdot)$ with respect to the parameters to get ML estimate

A Maximum Likelihood Approach

$$\mathcal{L}(\beta, \sigma^2 | y) = \prod_{i=1}^{n} \phi\left(\frac{y_i - x\beta}{\sigma}\right) \tag{2}$$

- We maximize $\mathcal{L}(\cdot)$ with respect to the parameters to get ML estimate
- Which parameter values were most likely to have produced the observed data?

A Maximum Likelihood Approach

$$\mathcal{L}(\beta, \sigma^2 | y) = \prod_{i=1}^{n} \phi\left(\frac{y_i - x\beta}{\sigma}\right) \tag{2}$$

- We maximize $\mathcal{L}(\cdot)$ with respect to the parameters to get ML estimate
- Which parameter values were most likely to have produced the observed data?
- We recognize that our parameter estimates are based on a single sample, so we use the sampling distribution to compute standard errors

A Bayesian Approach

$$\mathcal{P}(\theta|y) = \frac{\mathcal{P}(y|\theta)\mathcal{P}(\theta)}{\mathcal{P}(y)} \tag{3}$$

A Bayesian Approach

$$\mathcal{P}(\theta|y) = \frac{\mathcal{P}(y|\theta)\mathcal{P}(\theta)}{\mathcal{P}(y)} \tag{3}$$

- Where ML ignores the prior, Bayesian approaches take advantage of it

A Bayesian Approach

$$\mathcal{P}(\theta|y) = \frac{\mathcal{P}(y|\theta)\mathcal{P}(\theta)}{\mathcal{P}(y)} \tag{3}$$

- Where ML ignores the prior, Bayesian approaches take advantage of it
- Posterior translates likelihood function into probability distribution over unknown parameters

A Bayesian Approach

$$\mathcal{P}(\theta|y) = \frac{\mathcal{P}(y|\theta)\mathcal{P}(\theta)}{\mathcal{P}(y)} \tag{3}$$

- Where ML ignores the prior, Bayesian approaches take advantage of it
- Posterior translates likelihood function into probability distribution over unknown parameters
- We summarize that posterior distribution to calculate quantities of interest

Bayesian Linear Model

$$\mathcal{L}(\beta, \sigma^2 | y) \propto \mathcal{P}(y | \beta, \sigma^2) = \prod_{i=1}^{n} \phi\left(\frac{y_i - x\beta}{\sigma}\right) \times \mathcal{P}(\beta)\mathcal{P}(\sigma^{-2}) \tag{4}$$

Bayesian Linear Model

$$\mathcal{L}(\beta, \sigma^2 | y) \propto \mathcal{P}(y | \beta, \sigma^2) = \prod_{i=1}^{n} \phi\left(\frac{y_i - x\beta}{\sigma}\right) \times \mathcal{P}(\beta)\mathcal{P}(\sigma^{-2}) \tag{4}$$

- $\mathcal{P}(\beta)$ is multivariate normal density and $\mathcal{P}(\sigma^{-2})$ is a Gamma density

Bayesian Linear Model

$$\mathcal{L}(\beta, \sigma^2|y) \propto \mathcal{P}(y|\beta, \sigma^2) = \prod_{i=1}^{n} \phi\left(\frac{y_i - x\beta}{\sigma}\right) \times \mathcal{P}(\beta)\mathcal{P}(\sigma^{-2}) \qquad (4)$$

- $\mathcal{P}(\beta)$ is multivariate normal density and $\mathcal{P}(\sigma^{-2})$ is a Gamma density
- The idea is that we generate a posterior by combining the likelihood and multiplying it by the prior

Bayesian Coin Flip

- We hold a coin in our hand. *We assume no knowledge* about the coin

Bayesian Coin Flip

- We hold a coin in our hand. *We assume no knowledge* about the coin
- We collect data: in 10 flips the coin comes up heads 7 times

Bayesian Coin Flip

- We hold a coin in our hand. *We assume no knowledge* about the coin
- We collect data: in 10 flips the coin comes up heads 7 times
- In the frequentist world, we would calculate the probability of seeing 7 heads in 10 flips when flipping a fair coin and assess whether we can reject the null hypothesis of fairness

Bayesian Coin Flip

- We hold a coin in our hand. *We assume no knowledge* about the coin
- We collect data: in 10 flips the coin comes up heads 7 times
- In the frequentist world, we would calculate the probability of seeing 7 heads in 10 flips when flipping a fair coin and assess whether we can reject the null hypothesis of fairness
- In the Bayesian world, we update our belief about the nature of the coin based on the data (e.g., 0.7 vs. 0.5)

Bayesian Coin Flip

- We hold a coin in our hand. *We assume no knowledge* about the coin
- We collect data: in 10 flips the coin comes up heads 7 times
- In the frequentist world, we would calculate the probability of seeing 7 heads in 10 flips when flipping a fair coin and assess whether we can reject the null hypothesis of fairness
- In the Bayesian world, we update our belief about the nature of the coin based on the data (e.g., 0.7 vs. 0.5)
- Our prior is a uniform distribution $\mathcal{P}(\theta) = 1$ for $0 \leq \theta \leq 1$

Bayesian Coin Flip

- We hold a coin in our hand. *We assume no knowledge* about the coin
- We collect data: in 10 flips the coin comes up heads 7 times
- In the frequentist world, we would calculate the probability of seeing 7 heads in 10 flips when flipping a fair coin and assess whether we can reject the null hypothesis of fairness
- In the Bayesian world, we update our belief about the nature of the coin based on the data (e.g., 0.7 vs. 0.5)
- Our prior is a uniform distribution $\mathcal{P}(\theta) = 1$ for $0 \leq \theta \leq 1$
- Our likelihood is $\mathcal{P}(\theta|y) \propto \left(\begin{array}{c} 10 \\ 7 \end{array} \right) \theta^7 (1-\theta)^3$

Lecture Outline

Lecture Outline

Comparing to Other Unfolding Techniques

- Parametric NOMINATE and IRT treat errors differently than a nonparametric classification procedure

Comparing to Other Unfolding Techniques

- Parametric NOMINATE and IRT treat errors differently than a nonparametric classification procedure

- NOMINATE and IRT assume errors are: iid, normally distributed, and less likely to occur as the distance from the cutting plane (or difficulty parameter) increases; Classification treats errors equally and only tries to minimize the total number of classification errors

Comparing to Other Unfolding Techniques

- Parametric NOMINATE and IRT treat errors differently than a nonparametric classification procedure
- NOMINATE and IRT assume errors are: iid, normally distributed, and less likely to occur as the distance from the cutting plane (or difficulty parameter) increases; Classification treats errors equally and only tries to minimize the total number of classification errors
- Classification is preferred in circumstances where there are low error rates and where the distribution of the errors is unknown

Comparing to Other Unfolding Techniques

- Parametric NOMINATE and IRT treat errors differently than a nonparametric classification procedure
- NOMINATE and IRT assume errors are: iid, normally distributed, and less likely to occur as the distance from the cutting plane (or difficulty parameter) increases; Classification treats errors equally and only tries to minimize the total number of classification errors
- Classification is preferred in circumstances where there are low error rates and where the distribution of the errors is unknown
- If the error rate is too small, the parametric methods push ideal points to the edges of the space to maximize the log-likelihoods

Estimation

- MCMC constructs a set of random draws from the posterior distribution for each parameter

Estimation

- MCMC constructs a set of random draws from the posterior distribution for each parameter
- Using the sample from the posterior, the point estimates can be taken as the mean of the posterior distribution

Estimation

- MCMC constructs a set of random draws from the posterior distribution for each parameter
- Using the sample from the posterior, the point estimates can be taken as the mean of the posterior distribution
- Thus, the advantage of Bayesian methods is the joint estimation of ideal points and parameters

Estimation

- MCMC constructs a set of random draws from the posterior distribution for each parameter
- Using the sample from the posterior, the point estimates can be taken as the mean of the posterior distribution
- Thus, the advantage of Bayesian methods is the joint estimation of ideal points and parameters
- At a minimum, vague priors are required for identification, or...

Estimation

- MCMC constructs a set of random draws from the posterior distribution for each parameter
- Using the sample from the posterior, the point estimates can be taken as the mean of the posterior distribution
- Thus, the advantage of Bayesian methods is the joint estimation of ideal points and parameters
- At a minimum, vague priors are required for identification, or...
- There is a set of fixed legislators to identify the model

Estimation

- MCMC constructs a set of random draws from the posterior distribution for each parameter
- Using the sample from the posterior, the point estimates can be taken as the mean of the posterior distribution
- Thus, the advantage of Bayesian methods is the joint estimation of ideal points and parameters
- At a minimum, vague priors are required for identification, or...
- There is a set of fixed legislators to identify the model
- This means we can estimate a 2PL model on the roll call matrix and interpret the results in terms of spatial voting as it relates to any proposal

Estimation

- MCMC constructs a set of random draws from the posterior distribution for each parameter
- Using the sample from the posterior, the point estimates can be taken as the mean of the posterior distribution
- Thus, the advantage of Bayesian methods is the joint estimation of ideal points and parameters
- At a minimum, vague priors are required for identification, or...
- There is a set of fixed legislators to identify the model
- This means we can estimate a 2PL model on the roll call matrix and interpret the results in terms of spatial voting as it relates to any proposal
- In sum, there are many ways to estimate these models: MCMC (`ideal()`): time consuming, but can derive the *complete* posterior; EM (`emIRT()`): fast, but requires bootstrapping to get standard errors because we are *approximating* the posterior; or even MLE

Summing Up: A Caveat

Summing Up: A Caveat

- IRT models are useful, but they aren't magic bullets: they assume there is some true theory of decision making

Summing Up: A Caveat

- IRT models are useful, but they aren't magic bullets: they assume there is some true theory of decision making
- Estimates are only as good as the model, e.g.,

Summing Up: A Caveat

- IRT models are useful, but they aren't magic bullets: they assume there is some true theory of decision making
- Estimates are only as good as the model, e.g.,
  - *Strategic voting*: all models assume elites are voting on their locations relative to cut point line
  - But *if* elites don't vote according to their own preferences, but based on all kinds of things like logrolling, etc., then we could get misleading estimates of ideal points
  - In other words, can't predict Y with Y

Lecture Outline

Coming Up

- Thursday: Application in `R` – SCOTUS Case Decisions