Lecture 2: Multinomial choice and simulation method

Jean-François Houde
UW-Madison

November 10, 2021

Alternatives models with non-IIA preferences

- IIA-like substitution patterns are not unique to Logit.
- Any model with IID and full-support shocks will produce similar substitution/sorting patterns (e.g. IID multinomial probit).

Alternatives models with non-IIA preferences

- IIA-like substitution patterns are not unique to Logit.
- Any model with IID and full-support shocks will produce similar substitution/sorting patterns (e.g. IID multinomial probit).
- How to relax the IIA assumption? Introduce correlation between utility shocks.
- Three approaches:
  - Nested-logit distribution (or Generalized Extreme Value)
  - Random-coefficients (i.e. β_i's)
  - Multinomial Probit with correlated errors

Mixed-Logit Model

- Random-utility model with *random-coefficients:*

$$V_{ij} = X_{ij}\beta_i + \epsilon_{ij}, \qquad \text{Where } \beta_i \sim f(\beta; \theta)$$

  - θ are parameters determining the distribution of β_i
  - A few common examples:

$$\text{Normal RC: } \beta_i \sim N(\mu, \Sigma)$$

$$\text{Normal RC with demographics: } \beta_i = Z_i\gamma + \eta_i, \quad \eta_i \sim N(0, \Sigma)$$

$$\text{Finite mixture: } \beta_i \in \{\beta_1, \ldots, \beta_K\} \text{ and } \Pr(\beta_i = \beta_k) = \omega_k$$

Mixed-Logit Model

- Random-utility model with *random-coefficients:*

$$V_{ij} = X_{ij}\beta_i + \epsilon_{ij}, \qquad \text{Where } \beta_i \sim f(\beta;\theta)$$

  - θ are parameters determining the distribution of β_i
  - A few common examples:

$$\text{Normal RC: } \beta_i \sim N(\mu, \Sigma)$$

$$\text{Normal RC with demographics: } \beta_i = Z_i\gamma + \eta_i, \quad \eta_i \sim N(0, \Sigma)$$

$$\text{Finite mixture: } \beta_i \in \{\beta_1, \ldots, \beta_K\} \text{ and } \Pr(\beta_i = \beta_k) = \omega_k$$

- Mixed-logit conditional choice probabilities:

$$\Pr(y_i = j | X_{i0}, \ldots, X_{iJ}, \theta) = \int \frac{\exp(X_{ij}\beta)}{1 + \sum_{j=1}^{J} \exp(X_{ij}\beta)} f(\beta;\theta) d\beta \equiv P_{ij}$$

Substitution Patterns

- *Elasticity of substitution:* Change in attribute $X_{ij,l}$ on the probability of choosing option k

$$\begin{aligned}\eta^{l}_{jk}(X_i) &= \frac{X_{ij,l}}{P_{ik}} \int -\beta_{j,l} P_{ij}(\beta) P_{ik}(\beta) f(\beta) d\beta \\ &= \int -X_{ij,l}\beta_{j,l} P_{ij}(\beta) \underbrace{\frac{P_{ik}(\beta)}{P_{ik}}}_{\neq 1} f(\beta) d\beta \neq -X_{ij,l}\beta_{j,l} P_{ij}\end{aligned}$$

- Therefore, the elasticity of substitution between (j, k) is not strictly proportional to probability of choosing j
  - Violation of IIA

Substitution Patterns

- *Elasticity of substitution:* Change in attribute $X_{ij,l}$ on the probability of choosing option k

$$\begin{aligned}\eta^{l}_{jk}(X_i) &= \frac{X_{ij,l}}{P_{ik}} \int -\beta_{j,l} P_{ij}(\beta) P_{ik}(\beta) f(\beta) d\beta \\ &= \int -X_{ij,l}\beta_{j,l} P_{ij}(\beta) \underbrace{\frac{P_{ik}(\beta)}{P_{ik}}}_{\neq 1} f(\beta) d\beta \neq -X_{ij,l}\beta_{j,l} P_{ij}\end{aligned}$$

- Therefore, the elasticity of substitution between (j, k) is not strictly proportional to probability of choosing j
  - Violation of IIA
- The magnitude of $\eta^{l}_{jk}(X_i)$ depends on the correlation between $P_{ij}(\beta)$ and $P_{ik}(\beta)$ across β's
  - If $P_{ij}(\beta)$ and $P_{ik}(\beta)$ are positively correlated $\rightarrow$ close substitutes
  - In other words, two products are close substitutes if they are popular among the same "types" of individuals (i.e. βs)

Example: Quality Ladder

- Consumer surplus:

$$V_{ij} = X_{ij}\beta + \epsilon_{ij} - \alpha p_{ij}$$

and $\ln \alpha \sim N(\bar{\alpha}, \sigma_\alpha^2)$.

Example: Quality Ladder

- Consumer surplus:

$$V_{ij} = X_{ij}\beta + \epsilon_{ij} - \alpha p_{ij}$$

  and $\ln \alpha \sim N(\bar{\alpha}, \sigma_\alpha^2)$.

- Elasticity of substitution:

$$\eta_{jk}^{p}(X_i) = \frac{p_{ij}}{P_{ik}} \int \alpha P_{ij}(\alpha) P_{ik}(\alpha) f(\alpha; \theta) d\alpha$$

- Substitution patterns:
  - Luxury products ($\uparrow p_{ij}$) are purchased by individuals with small α
  - Entry products ($\downarrow p_{ij}$) are purchased by individuals with high α
  - $\eta_{jk}^{p}(X_i) > \eta_{jl}^{p}$ if (j, k) have *similar* prices (unlike (j, l))

Multinomial probit model

- Similar substitution patterns can be obtained using multinomial probit model with correlated shocks
- Example 1: Random utility with 4 options

$$V_{ij} = X_{ij}\beta + \epsilon_{ij} \quad \epsilon_i \sim (0, \Sigma) \text{ and } j = 0, \ldots, 3$$

- Option 1 is chosen if:

$$A_{i1}(\beta) = \left\{ \epsilon_{i0}, \ldots, \epsilon_{i3} \middle| \underbrace{V_{i1} > V_{i2}}_{\epsilon_{i2} - \epsilon_{i1} < (X_{i1} - X_{i2})\beta} , \underbrace{V_{i1} > V_{i3}}_{\epsilon_{i3} - \epsilon_{i1} < (X_{i1} - X_{i3})\beta} , \underbrace{V_{i1} > 0}_{\epsilon_{i0} - \epsilon_{i1} < (X_{i0} - X_{i1})\beta} \right\}$$

- Other partitions of ϵ's characterize the choice of other options: $A_{ij}(\beta)$ for $j = 0, \ldots, 3$

Multinomial probit model

- Example 2: Dynamic probit

$$Y_{it} = \begin{cases} 1 & \text{If } X_{it}\beta + \epsilon_{it} > 0, \text{ and } (\epsilon_{i1}, \ldots, \epsilon_{iT}) \sim N(0, \Sigma) \\ -1 & \text{Else.} \end{cases}$$

  E.g.: $\epsilon_{it} = \rho\epsilon_{it-1} + \eta_{it}$ where $\eta_{it} \sim N(0,1)$ and $\epsilon_{i1} \sim N(0, \frac{1}{(1-\rho)^2})$ (i.e. stationary distribution).
- As in the previous example, the policy function is defined by a series of thresholds. Let $y_i = \{y_{i1}, \ldots, y_{iT}\}$ denotes a sequence of binary choices.
- The partition of ϵ that rationalize y_i is:

$$A_i(y_i|\beta) = \left\{ \epsilon_{i1}, \ldots, \epsilon_{iT} \middle| \epsilon_{i1} > y_{i1}X_{i1}\beta, \ldots, \epsilon_{iT} > y_{iT}X_{iT}\beta \right\}$$

Multinomial probit model

- In both examples, the probability of observing a choice y_i is given by:

$$\Pr(y_i|X_i,\beta) = \int_{\epsilon \in A_i(y_i|\beta)} \phi(\epsilon_i|\Sigma) d\epsilon$$

- This is a more complicated integration problem, since we need to integrate over a truncated support ($\neq$ mixed-logit model)

More complicated integration problems...

- In other cases, integration is very very complicated, because it is not feasible to analytically define the region of integration
- Example: "Pure" characteristics random-utility model

$$y_i = j \text{ if } X_{ij}\beta_i > X_{ik}\beta_i, \quad \forall k \neq j$$

where $\beta_i \sim N(\mu, \Sigma)$

- In this model, consumers choose different options because they value observed characteristics differently.
- Characterizing the region of integration $A_i(y_i)$ is more difficult (linear programming problem).
- In order to evaluate the likelihood of observing y_i we need to integrate over an "unknown" distribution:

$$\Pr(y_i|X_i, \mu, \Sigma) = \int \sum_j 1(y_i = j) \times 1(X_{ij}\beta_i > X_{ik}\beta_i, \forall k \neq j)\phi(\beta_i; \mu, \Sigma)$$

Estimation: Simulated MLE *vs* GMM

- Estimation of Mixed-Logit and multinomial Probit models must rely on approximation methods to calculate the choice-probabilities

$$\hat{P}_{ij}(\theta) = \sum_r w_r \Pr(i|\beta_r)$$

where $\beta_r \sim f(\beta; \theta)$ is a random-coefficient.

- Two common estimators:

$$\text{Simulated MLE: } \max_\theta \sum_i \sum_j d_{ij} \ln \hat{P}_{ij}(\theta)$$

$$\text{Simulated MM: } \sum_i \sum_j \left[d_{ij} - \hat{P}_{ij}(\theta) \right] z_{ij} = 0$$

where $d_{ij} = 1(y_i = j)$ and z_{ij} is a vector of "instruments".

Estimation: Simulated MLE *vs* GMM

- **Tradeoff:** Consistency
  - With fixed-R, SML is not consistent. Why? Simulation error enters non-linearly in the model (because of the "log")
  - SMM does not have this problem. Why? Simulation error is linear, and orthogonal to the instrument z_{ij}.
  - References: McFadden (1989) and Pakes & Polard (1989)

Estimation: Simulated MLE *vs* GMM

- **Tradeoff:** Consistency
  - With fixed-R, SML is not consistent. Why? Simulation error enters non-linearly in the model (because of the "log")
  - SMM does not have this problem. Why? Simulation error is linear, and orthogonal to the instrument z_{ij}.
  - References: McFadden (1989) and Pakes & Polard (1989)
- **Tradeoff:** Efficiency
  - MSM is not efficient unless z_{ij} is chosen "optimally". In this example, the most efficient instrument is:

$$\sum_i \sum_j \left[d_{ij} - \hat{P}_{ij}(\theta) \right] z_{ij} = \sum_i \sum_j \left[d_{ij} - \hat{P}_{ij}(\theta) \right] \frac{\partial \ln P_{ij}(\theta)}{\partial \theta} = 0 \text{ [Score]}$$

This is not feasible without introducing an error, since $\frac{\partial \ln P_{ij}(\theta)}{\partial \theta}$ needs to be approximated via simulation (breaks consistency)

Estimation: Simulated MLE *vs* GMM

- **Tradeoff:** Consistency
  - With fixed-R, SML is not consistent. Why? Simulation error enters non-linearly in the model (because of the "log")
  - SMM does not have this problem. Why? Simulation error is linear, and orthogonal to the instrument z_{ij}.
  - References: McFadden (1989) and Pakes & Polard (1989)
- **Tradeoff:** Efficiency
  - MSM is not efficient unless z_{ij} is chosen "optimally". In this example, the most efficient instrument is:

$$\sum_i \sum_j \left[d_{ij} - \hat{P}_{ij}(\theta) \right] z_{ij} = \sum_i \sum_j \left[d_{ij} - \hat{P}_{ij}(\theta) \right] \frac{\partial \ln P_{ij}(\theta)}{\partial \theta} = 0 \text{ [Score]}$$

    This is not feasible without introducing an error, since $\frac{\partial \ln P_{ij}(\theta)}{\partial \theta}$ needs to be approximated via simulation (breaks consistency)
- **Recommendation:** Use MLE when the integral can be approximated reasonably well. Use MSM otherwise. How to choose z_{ij}?
  - Non-linear functions of $\{X_{i1}, \ldots, X_{iJ}\}$

Numerical integration methods

- Quadrature methods
- How to draw over "known" density?
  - Pseudo-random numbers
  - Halton sequences
- How to draw from "unknown" densities?
  - Accept/Reject
  - Importance sampling
  - MCMC

Quadrature Methods

- **Newton-Cotes Methods**:
  - Bounded interval: $\epsilon \in (a, b)$
  - Uniform grid: $\epsilon_r = a + (r-1)h$ (fix step h)
  - Linear approximation:

$$\int_{\epsilon_r}^{\epsilon_{r+1}} m(\epsilon) f(\epsilon) d\epsilon \approx \frac{h}{2} \left(m(\epsilon_r) f(\epsilon_r) + m(\epsilon_{r+1}) f(\epsilon_{r+1}) \right) = \text{Trapezoid}$$

Quadrature Methods

- **Newton-Cotes Methods**:
  - Bounded interval: $\epsilon \in (a, b)$
  - Uniform grid: $\epsilon_r = a + (r-1)h$ (fix step h)
  - Linear approximation:

$$\int_{\epsilon_r}^{\epsilon_{r+1}} m(\epsilon) f(\epsilon) d\epsilon \approx \frac{h}{2} \left(m(\epsilon_r) f(\epsilon_r) + m(\epsilon_{r+1}) f(\epsilon_{r+1}) \right) = \text{Trapezoid}$$

  - Summing across intervals:

$$\bar{m} \approx \sum_{r=1}^{R} m(\epsilon_r) f(\epsilon_r) w_r$$

    where $w_1 = w_R = h/2$ and $w_r = h$.
  - A more accurate can be obtained using a quadratic approximation (Simpson's rule)
  - The choice of h is crucial: Adaptive quadrature iterates on R until $\bar{m}_h$ stops changing.

Quadrature Methods

- **Gaussian Methods**:
  - When the function $m(\cdot)$ is smooth, it is more efficient to "space" the grid points strategically
  - The weights and nodes are chosen such that the polynomial approximation is satisfied:

$$\int x^k w(x) dx = \sum_r^R w_r x_k^k, \quad k = 0, \ldots, 2R-1$$

  - Standard softwares have pre-computed nodes/weights for standard polynomial degrees
  - See Judd (1999) for nodes/weights tables

Quadrature Methods

- **Gaussian Methods**:
  - When the function $m(\cdot)$ is smooth, it is more efficient to "space" the grid points strategically
  - The weights and nodes are chosen such that the polynomial approximation is satisfied:

$$\int x^k w(x) dx = \sum_r^R w_r x_k^k, \quad k = 0, \ldots, 2R-1$$

  - Standard softwares have pre-computed nodes/weights for standard polynomial degrees
  - See Judd (1999) for nodes/weights tables
  - Given a quadrature degree R, integral can be evaluated:

$$\bar{m} \approx \sum_r m(\epsilon_r) f(\epsilon_r) w_r$$

  - This is the preferred approach for most low-dimensional econometrics problem. Works very well when $m(\epsilon)$ is smooth.

Quadrature Methods

- **Drawback:** Curse of dimensionality
  - Each dimension requires a small number of nodes (e.g. 10)
  - Multidimensional integral can be approximated using the tensor product of each grid
  - Number of evaluations: R^d
  - Feasible for $d \leq 3$
- **Recent advances:** Sparse Grid Integration
  - Alleviate the curse of dimensionality problem
  - References: Heiss & Winschel (2010) and Skrainka & Judd (2011)
  - Download wrids/weights packages: http://www.sparse-grids.de

Simulation Methods

- *Key idea:* Generate a sequence of *pseudo-random* numbers
  - What is *pseudo-random* number generator? Algorithm that generates a *deterministic* sequence of numbers starting from a *seed* value
  - Better alternative: Halton sequence

Simulation Methods

- *Key idea:* Generate a sequence of *pseudo-random* numbers
  - What is *pseudo-random* number generator? Algorithm that generates a *deterministic* sequence of numbers starting from a *seed* value
  - Better alternative: Halton sequence
- Sampling from a univariate distribution $f(\epsilon)$:
  1. Sample a uniform pseudo-random number: $u_r \sim U[0,1]$
  2. Invert the CDF of ϵ:

$$\epsilon_r = F^{-1}(u_r)$$

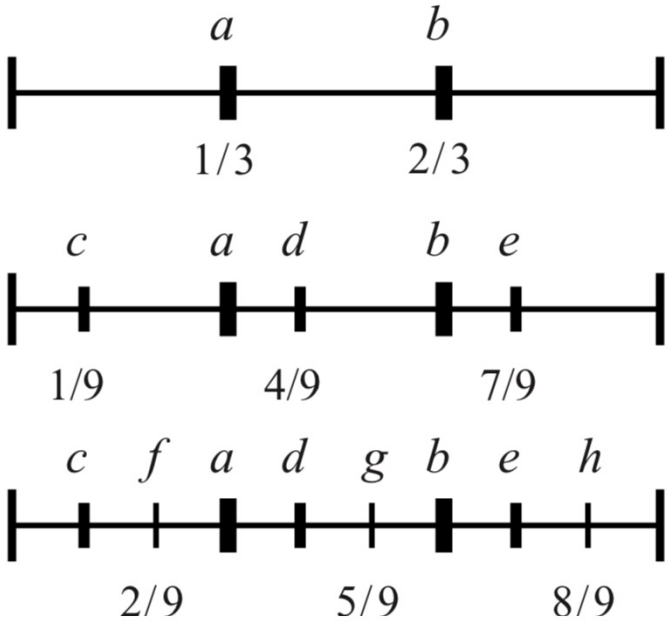
  Where $F^{-1}(u_r)$ is the quantile function of $f(\cdot)$
- Examples:

$$\text{Normal distribution: } F^{-1}(u) = \mu + \sigma \underbrace{\Phi^{-1}(u)}_{\text{Std. Normal}}$$

$$\text{Exponential: } F(\epsilon) = 1 - \exp(-\epsilon/\sigma) \rightarrow F^{-1}(u) = -\sigma \ln(1-u)$$

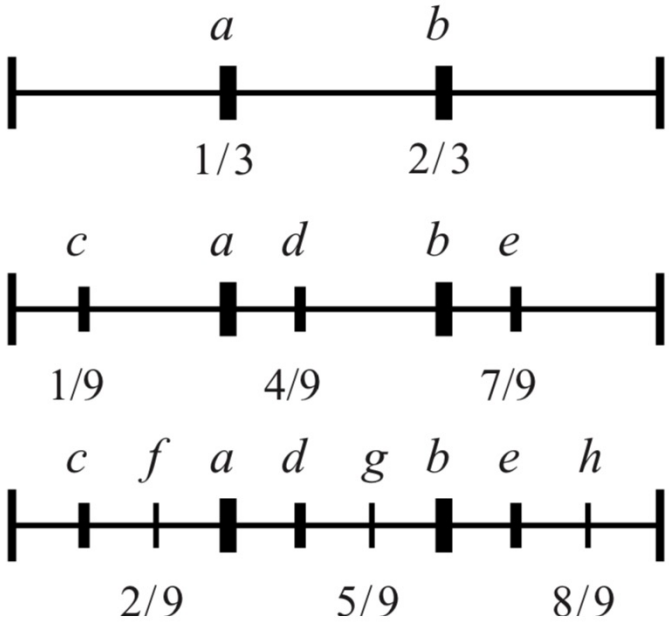
Halton Sequences: General Idea

- **Goal:** Generate a *deterministic* sequence of ϵ that have better coverage than standard pseudo-random number sequence.
- **Example:** Halton sequence of degree 3 (must be a prime number)
  1. Divide $[0, 1]$ in three equal segments $(1/3, 2/3)$
  2. Divide each subsegments in three
  3. ...



Halton Sequences: General Idea

- **Goal:** Generate a *deterministic* sequence of ϵ that have better coverage than standard pseudo-random number sequence.
- **Example:** Halton sequence of degree 3 (must be a prime number)
  1. Divide $[0, 1]$ in three equal segments $(1/3, 2/3)$
  2. Divide each subsegments in three
  3. ...

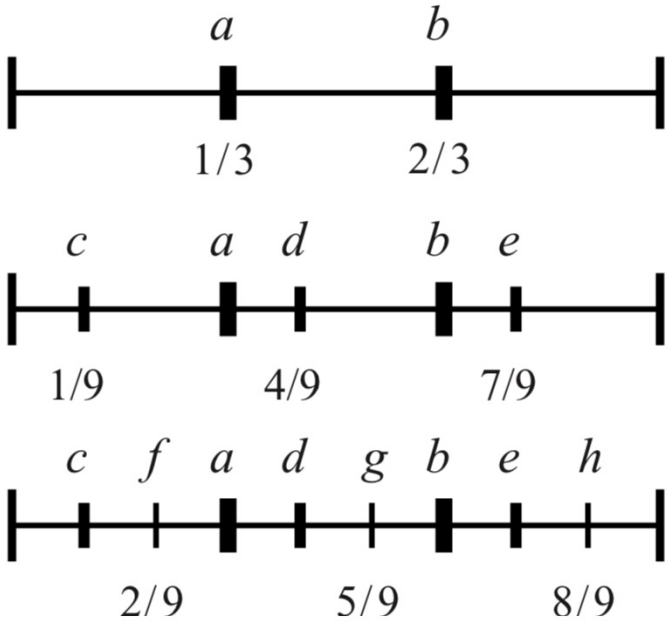


- **Pseudo-code:**
  - Initial sequence $s_0 = \{0\}$
  - Sequence t: $s_{t+1} = \{s_t, s_t + (1/3)^t, s_t + (2/3)^t\}$

  s_t = vector containing the numbers from the previous sequence.

Halton Sequences: General Idea

- **Goal:** Generate a *deterministic* sequence of ϵ that have better coverage than standard pseudo-random number sequence.
- **Example:** Halton sequence of degree 3 (must be a prime number)
  1. Divide $[0, 1]$ in three equal segments $(1/3, 2/3)$
  2. Divide each subsegments in three
  3. ...



- **Pseudo-code:**
  - Initial sequence $s_0 = \{0\}$
  - Sequence t: $s_{t+1} = \{s_t, s_t + (1/3)^t, s_t + (2/3)^t\}$

  s_t = vector containing the numbers from the previous sequence.
- Example (degree 3):
  - First sequence: s_1 contains 3 elements
  - Second sequence: s_2 contains 9 elements
  - Second sequence: s_3 contains 27 elements

Halton Sequences: Mixed-Logit

- **Key advantage:** Halton draws have better coverage than pseudo-random draws, which means that it requires fewer draws to achieve the same precision.
- **Example:** Mixed-Logit choice probability

$$P_{ij} = \int P_{ij}(\beta) f(\beta) d\beta, \text{ where } \dim(\beta) = K$$

  - N = Number of individuals (i) and R = Number of simulation draws
  - For each random-coefficient, generate a Halton sequence of length $M + N \cdot R$
  - Use different prime numbers for each parameter, and discard the first M elements (e.g. 10)
  - The first R draws are used from individual 1, $R + 1$ to $2R$ are used for individual 2, etc.

Sampling from truncated and multivariate distributions

- Sampling from truncated and multivariate distributions are the two more challenging problems

Sampling from truncated and multivariate distributions

- Sampling from truncated and multivariate distributions are the two more challenging problems
- **Special case 1:** Univariate truncated distribution. Ex:

$$f(\epsilon; a, b) = \frac{1}{K} f(\epsilon) \text{ If } a < \epsilon < b$$

Where $K = \int_a^b f(\epsilon) d\epsilon$ is the normalization constant.

Sampling from truncated and multivariate distributions

- Sampling from truncated and multivariate distributions are the two more challenging problems
- **Special case 1:** Univariate truncated distribution. Ex:

$$f(\epsilon; a, b) = \frac{1}{K} f(\epsilon) \text{ If } a < \epsilon < b$$

  Where $K = \int_a^b f(\epsilon) d\epsilon$ is the normalization constant.
- When ϵ is a scalar, we can construct a random sample by drawing from the truncated uniform distribution
  1. Sample uniform r.v.: $u_r \sim U[0, 1]$
  2. Transform u into a quantile in (a, b): $q(u_r) = F(a) + u_r [F(b) - F(a)]$
  3. Invert unconditional CDF: $\epsilon_r = F^{-1}(q(u_r))$

Sampling from truncated and multivariate distributions

- Sampling from truncated and multivariate distributions are the two more challenging problems
- **Special case 1:** Univariate truncated distribution. Ex:

$$f(\epsilon; a, b) = \frac{1}{K} f(\epsilon) \text{ If } a < \epsilon < b$$

  Where $K = \int_a^b f(\epsilon) d\epsilon$ is the normalization constant.
- When ϵ is a scalar, we can construct a random sample by drawing from the truncated uniform distribution
  1. Sample uniform r.v.: $u_r \sim U[0, 1]$
  2. Transform u into a quantile in (a, b): $q(u_r) = F(a) + u_r [F(b) - F(a)]$
  3. Invert unconditional CDF: $\epsilon_r = F^{-1}(q(u_r))$
- Importantly, ϵ_r is in (a, b) by construction.

Sampling from truncated and multivariate distributions

- **Special case 2:** Multivariate normal (non-truncated). Ex: $\epsilon \sim N(\mu, \Sigma)$

$$\epsilon = \mu + L\eta \text{ where } \Sigma = LL' \text{ and } \eta \sim N(0, I)$$

Sampling from truncated and multivariate distributions

- **Special case 2:** Multivariate normal (non-truncated). Ex: $\epsilon \sim N(\mu, \Sigma)$

$$\epsilon = \mu + L\eta \text{ where } \Sigma = LL' \text{ and } \eta \sim N(0, I)$$

- This leads to a recursive formulation ($K = 3$):

$$\begin{aligned}
\epsilon_1 &= \mu_1 + L_{11}\eta_1 \\
\epsilon_2 &= \mu_2 + L_{21}\eta_1 + L_{11}\eta_2 \\
\epsilon_3 &= \mu_3 + L_{31}\eta_1 + L_{32}\eta_2 + L_{33}\eta_3
\end{aligned}$$

- Multivariate sampling:
  1. Sample K standard-normal variables (iid): η_{rk}
  2. Rescale the variables using μ and L

Simulation Methods: How to sample from "unknown" densities

- We are interested in methods to approximate via simulations:

$$\rightarrow \bar{m} = \int_a^b m(\epsilon) f(\epsilon; a, b) d\epsilon \approx \sum_r m(\epsilon_r) w_r$$

  where $f(\epsilon; a, b)$ is the conditional density of ϵ.
- We will cover four approaches:
  - Accept/Reject
  - Importance sampling
  - MCMC

Accept/Reject Sampling Method

- **Case:** $\dim(\epsilon) > 1$ and $\epsilon_i \sim f(\epsilon; a, b)$
- The "naive" method is to sample R numbers from the *unconditional* distribution $f(\epsilon)$ an apply the truncation rule:
  1. Sample $\epsilon_r \sim f(\epsilon)$
  2. If $a < \epsilon_r < b$, keep. Else reject.

Accept/Reject Sampling Method

- **Case:** $\dim(\epsilon) > 1$ and $\epsilon_i \sim f(\epsilon; a, b)$
- The "naive" method is to sample R numbers from the *unconditional* distribution $f(\epsilon)$ an apply the truncation rule:
  1. Sample $\epsilon_r \sim f(\epsilon)$
  2. If $a < \epsilon_r < b$, keep. Else reject.
- Let $\mathcal{A}_R(a, b)$ denotes the set of "accepted" ϵ_r. If $\#\mathcal{A}_R(a, b) \to \infty$ we can consistently approximate the integral:

$$\bar{m} \approx \frac{1}{\#\mathcal{A}_R(a, b)} \sum_{r \in \mathcal{A}(a,b)} m(\epsilon_r)$$

Accept/Reject Sampling Method

- **Case:** $\dim(\epsilon) > 1$ and $\epsilon_i \sim f(\epsilon; a, b)$
- The "naive" method is to sample R numbers from the *unconditional* distribution $f(\epsilon)$ an apply the truncation rule:
  1. Sample $\epsilon_r \sim f(\epsilon)$
  2. If $a < \epsilon_r < b$, keep. Else reject.
- Let $\mathcal{A}_R(a, b)$ denotes the set of "accepted" ϵ_r. If $\#\mathcal{A}_R(a, b) \to \infty$ we can consistently approximate the integral:

$$\bar{m} \approx \frac{1}{\#\mathcal{A}_R(a, b)} \sum_{r \in \mathcal{A}(a,b)} m(\epsilon_r)$$

- **Upside:** *Always* work. Even when we cannot evaluate $f(\epsilon; a, b)$
- **Drawbacks:**
  - The number of draws cannot be fixed ex-ante: $E[\#\mathcal{A}_R(a, b)] = K \times R$ and K is unknown
  - Otherwise we obtain a very noisy estimate of $\bar{m}$
  - Particularly problematic when K is small (e.g. rare events)

Importance Sampling

- **Case:** Difficult to sample directly from $f(\epsilon)$.

Importance Sampling

- **Case:** Difficult to sample directly from $f(\epsilon)$.
- IS formulation:

$$\begin{aligned} \bar{m} &= \int m(\epsilon)\frac{f(\epsilon)}{g(\epsilon)}g(\epsilon)d\epsilon \\ &\approx \frac{1}{R}\sum_r m(\epsilon_r)\frac{f(\epsilon_r)}{g(\epsilon_r)} = \sum_r m(\epsilon_r)w_r \end{aligned}$$

  where $\{\epsilon_r\}_{r=1,\ldots,R}$ is a random sample drawn from $g(\epsilon)$.
- Restrictions:
  - Density $g(\epsilon)$ must have the same support as $f(\epsilon)$
  - Make sense only if it is easy to sample from $g(\cdot)$

Importance Sampling

- **Case:** Difficult to sample directly from $f(\epsilon)$.
- IS formulation:

$$\begin{aligned} \bar{m} &= \int m(\epsilon)\frac{f(\epsilon)}{g(\epsilon)}g(\epsilon)d\epsilon \\ &\approx \frac{1}{R}\sum_r m(\epsilon_r)\frac{f(\epsilon_r)}{g(\epsilon_r)} = \sum_r m(\epsilon_r)w_r \end{aligned}$$

  where $\{\epsilon_r\}_{r=1,\ldots,R}$ is a random sample drawn from $g(\epsilon)$.
- Restrictions:
  - Density $g(\epsilon)$ must have the same support as $f(\epsilon)$
  - Make sense only if it is easy to sample from $g(\cdot)$
- **Example:** Truncated normal distribution with correlated errors.
  - $f(\epsilon) = \phi(\epsilon; \mu, \Sigma)/K$
  - $g(\epsilon)$: Truncated normal distribution over (a, b) with independent errors

Example: Multinomial Probit

- Example with 4 choices:

$$V_{ij} = X_{ij}\beta + \epsilon_{ij} \quad \epsilon_i \sim (0, \Sigma)$$

  See lecture 1 for discussion on how to standardize Σ.

- The probability of choosing option 4 (arbitrariry):

$$\begin{aligned} \Pr(d_{i4} = 1) &= \Pr(\epsilon_{ij} - \epsilon_{i4} < -(X_{ij} - X_{i4})\beta, \forall k \neq 4) \\ &= \int_{A_1 \cap A_2 \cap A_3} dF(\epsilon_{i1} - \epsilon_{i4}, \epsilon_{i1} - \epsilon_{i3}, \epsilon_{i3} - \epsilon_{i4}) \\ &\quad \text{where } A_j = \{\epsilon_{ij} - \epsilon_{i4} < -(X_{ij} - X_{i4})\beta\} \end{aligned}$$

- Redefine the variables in difference relative to 4:

$$\begin{aligned} &\nu_{ij} = \epsilon_{ij} - \epsilon_{i4} \\ &X^*_{ij} = X_{ij} - X_{i4} \\ &\nu \sim N(0, \Sigma), \quad \Sigma_{3\times 3} = LL' \\ &\nu = L\eta, \quad \eta \sim N(0, I). \end{aligned}$$

Example: Multinomial Probit

- Three approaches
  - Monte-Carlo simulation with Accept-Reject
  - SML with the GHK simulator (truncated normal)
  - Bayesian with MCMC simulator
- Accept-Reject:
  1. Draw R vectors $\eta_i^r \sim (0, I)$,
  2. Keep draw r if $C'\eta_i^r \in A_1 \cap A_2 \cap A_3$. Otherwise reject.
  3. Compute simulated choice probability:

$$\widehat{\Pr(d_{i4})} = \frac{\text{Number accepted draws}}{R} \tag{1}$$

- Problems and limitations of A/R:
  - Non-smooth simulator (i.e. cannot use gradient methods)
  - Require a high number of draws to avoid $Pr(d_{i4}) = 0$.
  - If the dimension of integration is large: infeasible

GHK Method

- Recall that:

$$\nu_1 = L_{11}\eta_1, \quad \nu_2 = L_{21}\eta_1 + L_{11}\eta_2, \quad \nu_3 = L_{31}\eta_1 + L_{32}\eta_2 + L_{33}\eta_3$$

GHK Method

- Recall that:

$$\nu_1 = L_{11}\eta_1, \quad \nu_2 = L_{21}\eta_1 + L_{11}\eta_2, \quad \nu_3 = L_{31}\eta_1 + L_{32}\eta_2 + L_{33}\eta_3$$

- In order to compute $\widehat{\Pr}(d_{i4} = 1)$ we proceed sequentially:
  1. **Draw ν_{i1}^r:** Compute $\Phi_{i1} = \Pr(\nu_{i1} < -X_{i1}^*\beta)$. Draw η_1^r from a truncated normal: $\eta \sim \Phi(-\frac{X_{i1}^*\beta}{L_{11}})$

GHK Method

- Recall that:

$$\nu_1 = L_{11}\eta_1, \quad \nu_2 = L_{21}\eta_1 + L_{11}\eta_2, \quad \nu_3 = L_{31}\eta_1 + L_{32}\eta_2 + L_{33}\eta_3$$

- In order to compute $\widehat{\Pr}(d_{i4} = 1)$ we proceed sequentially:
  1. **Draw ν_{i1}^r:** Compute $\Phi_{i1} = \Pr(\nu_{i1} < -X_{i1}^*\beta)$. Draw η_1^r from a truncated normal: $\eta \sim \Phi(-\frac{X_{i1}^*\beta}{L_{11}})$
  2. Draw ν_{i2}^r from a truncated normal (conditional on ν_{i1}^r):

$$\nu_2^r = L_{21}\eta_1^r + L_{22}\eta_2 < -X_{i2}^*\beta$$
$$\eta_2^r \sim \Phi\left(\frac{-X_{i2}^*\beta - L_{21}\eta_{i1}^r}{L_{22}}\right) \equiv \Phi_{i2}$$

     Compute $\nu_{i2}^r = L_{21}\eta_{i1}^r + L_{22}\eta_{i2}^r$.
  3. Compute Φ_{i3} similarly.

GHK Method

- Recall that:

$$\nu_1 = L_{11}\eta_1, \quad \nu_2 = L_{21}\eta_1 + L_{11}\eta_2, \quad \nu_3 = L_{31}\eta_1 + L_{32}\eta_2 + L_{33}\eta_3$$

- In order to compute $\widehat{\Pr}(d_{i4} = 1)$ we proceed sequentially:
  1. **Draw ν_{i1}^r:** Compute $\Phi_{i1} = \Pr(\nu_{i1} < -X_{i1}^*\beta)$. Draw η_1^r from a truncated normal: $\eta \sim \Phi(-\frac{X_{i1}^*\beta}{L_{11}})$
  2. Draw ν_{i2}^r from a truncated normal (conditional on ν_{i1}^r):

$$\nu_2^r = L_{21}\eta_1^r + L_{22}\eta_2 < -X_{i2}^*\beta$$
$$\eta_2^r \sim \Phi\left(\frac{-X_{i2}^*\beta - L_{21}\eta_{i1}^r}{L_{22}}\right) \equiv \Phi_{i2}$$

     Compute $\nu_{i2}^r = L_{21}\eta_{i1}^r + L_{22}\eta_{i2}^r$.
  3. Compute Φ_{i3} similarly.
  4. Compute $\widehat{\Pr}(d_{i4} = 1)$:

$$\hat{P}_{i4} = \frac{1}{R}\sum_m \Phi\left(\frac{-X_{i1}^*\beta}{L_{11}}\right) \times \Phi\left(\frac{-X_{i2}^*\beta - L_{12}\eta_{i1}^r}{L_{22}}\right) \times \Phi\left(\frac{-X_{i3}^*\beta - L_{31}\eta_{i1}^r - L_{32}\eta_{i2}^r}{L_{33}}\right)$$

GHK Method

- Advantages of the GHK:
  - Highly accurate even with high dimension integrals
  - Differentiable: Can use standard Quasi-Newton methods
  - Require fewer draws than MC-AR
  - Applicable to panel data (i.e. serial correlation in ϵ)
- **References:** Geweke (1991), Keane (1994), Hajiavassiliou et al. (1994)

Bayesian Method

- **Central idea:** Inference about the model parameters θ can be performed using Bayes' rule:

$$p(\theta|Y,X) = \frac{k(\theta)l(Y,X|\theta)}{l(Y,X)} \propto k(\theta)l(Y,X|\theta)$$

where $l(Y,X|\theta)$ is the model likelihood, $k(\theta)$ is the prior density over θ, and $l(Y,X)$ is the marginal density of observed data (normalizing constant)

  - For a "frequentist", $p(\theta|Y,X)$ is approximated by the asymptotic distribution of θ
  - For a "bayesian", $p(\theta|Y,X)$ is the posterior distribution of the parameters

Bayesian Method

- **Central idea:** Inference about the model parameters θ can be performed using Bayes' rule:

$$p(\theta|Y,X) = \frac{k(\theta)l(Y,X|\theta)}{l(Y,X)} \propto k(\theta)l(Y,X|\theta)$$

  where $l(Y,X|\theta)$ is the model likelihood, $k(\theta)$ is the prior density over θ, and $l(Y,X)$ is the marginal density of observed data (normalizing constant)
  - For a "frequentist", $p(\theta|Y,X)$ is approximated by the asymptotic distribution of θ
  - For a "bayesian", $p(\theta|Y,X)$ is the posterior distribution of the parameters
- In many econometrics applications, we select an "uninformative" prior $k(\theta)$ (or flat), and the posterior distribution is used much like the asymptotic distribution (e.g. confidence interval, test-statistics, etc)

Bayesian Method

- **Central idea:** Inference about the model parameters θ can be performed using Bayes' rule:

$$p(\theta|Y,X) = \frac{k(\theta)l(Y,X|\theta)}{l(Y,X)} \propto k(\theta)l(Y,X|\theta)$$

  where $l(Y,X|\theta)$ is the model likelihood, $k(\theta)$ is the prior density over θ, and $l(Y,X)$ is the marginal density of observed data (normalizing constant)
  - For a "frequentist", $p(\theta|Y,X)$ is approximated by the asymptotic distribution of θ
  - For a "bayesian", $p(\theta|Y,X)$ is the posterior distribution of the parameters
- In many econometrics applications, we select an "uninformative" prior $k(\theta)$ (or flat), and the posterior distribution is used much like the asymptotic distribution (e.g. confidence interval, test-statistics, etc)
- **Simulation:** Rather than calculating the posterior directly (impossible), we approximate $p(\theta|Y,X)$ using Monte-Carlo Markov-Chain (MCMC) methods

Example 1: Binomial Probit

- Radom utility for option 1:

$$V_{ij} = X_i\beta + \epsilon_i = (X_{i1} - X_{i0})\beta + \epsilon_{i1} - \epsilon_{i0}$$

and $\epsilon_i \sim N(0, 1)$

Example 1: Binomial Probit

- Radom utility for option 1:

$$V_{ij} = X_i\beta + \epsilon_i = (X_{i1} - X_{i0})\beta + \epsilon_{i1} - \epsilon_{i0}$$

  and $\epsilon_i \sim N(0, 1)$
- **Data augmentation:** Treat the latent utility V_i as part of the parameter vector

$$p(V_1, \ldots, V_n, \beta | Y, X) = p(V|\beta, Y, X)p(\beta|Y, X)$$

  where $p(V|\beta, Y, X)$ is the joint density of latent utilities given choices.

Example 1: Binomial Probit

- Radom utility for option 1:

$$V_{ij} = X_i\beta + \epsilon_i = (X_{i1} - X_{i0})\beta + \epsilon_{i1} - \epsilon_{i0}$$

  and $\epsilon_i \sim N(0, 1)$
- **Data augmentation:** Treat the latent utility V_i as part of the parameter vector

$$p(V_1, \ldots, V_n, \beta | Y, X) = p(V|\beta, Y, X)p(\beta|Y, X)$$

  where $p(V|\beta, Y, X)$ is the joint density of latent utilities given choices.
- **Key idea:** Under the assumption that ϵ_i is $N(0, 1)$ (iid across individuals), and that individuals maximize utilities, it is very simple to draw from $p(V|\beta, Y, X)$

$$p(V_i|\beta, Y_i, X_i) = \frac{\phi(V_i - X_i\beta)}{1 - \Phi(-X_i\beta)} = \text{Truncated normal}$$

- This naturally leads to a Gibbs sampler

Example 1: Binomial Probit

- Prior distribution: $\beta \sim N(\bar{\beta}, A^{-1})$. Diffuse prior: A is set to a small value.

Example 1: Binomial Probit

- Prior distribution: $\beta \sim N(\bar{\beta}, A^{-1})$. Diffuse prior: A is set to a small value.
- **Markov chain:** Starting point (β^0, V^0)

Example 1: Binomial Probit

- Prior distribution: $\beta \sim N(\bar{\beta}, A^{-1})$. Diffuse prior: A is set to a small value.
- **Markov chain:** Starting point (β^0, V^0)
  1. **Data augmentation:** Conditional on β^{t-1}, sample latent utility from truncated normal

$$V_i \sim \begin{cases} \frac{\phi(V_i - X_i\beta)}{1-\Phi(-X_i\beta^{t-1})} & \text{If } Y_i = 1 \\ \frac{\phi(V_i - X_i\beta)}{\Phi(-X_i\beta^{t-1})} & \text{If } Y_i = 0 \end{cases}$$

Example 1: Binomial Probit

- Prior distribution: $\beta \sim N(\bar{\beta}, A^{-1})$. Diffuse prior: A is set to a small value.
- **Markov chain:** Starting point (β^0, V^0)
  1. **Data augmentation:** Conditional on β^{t-1}, sample latent utility from truncated normal

$$V_i \sim \begin{cases} \frac{\phi(V_i - X_i\beta)}{1-\Phi(-X_i\beta^{t-1})} & \text{If } Y_i = 1 \\ \frac{\phi(V_i - X_i\beta)}{\Phi(-X_i\beta^{t-1})} & \text{If } Y_i = 0 \end{cases}$$

  2. **Parameter updating:** Conditional on "realized" random utilities for each i we have a standard linear regression

$$V_i^t = X_i\beta + \epsilon_i$$

$$\beta^t \sim N\left(\hat{\beta}, (X'X + A)^{-1}\right), \quad \hat{\beta} = (X'X)^{-1}(X'V^t + A\bar{\beta})$$

Example 1: Binomial Probit

- Prior distribution: $\beta \sim N(\bar{\beta}, A^{-1})$. Diffuse prior: A is set to a small value.
- **Markov chain:** Starting point (β^0, V^0)
  1. **Data augmentation:** Conditional on β^{t-1}, sample latent utility from truncated normal

$$V_i \sim \begin{cases} \frac{\phi(V_i - X_i\beta)}{1-\Phi(-X_i\beta^{t-1})} & \text{If } Y_i = 1 \\ \frac{\phi(V_i - X_i\beta)}{\Phi(-X_i\beta^{t-1})} & \text{If } Y_i = 0 \end{cases}$$

  2. **Parameter updating:** Conditional on "realized" random utilities for each i we have a standard linear regression

$$V_i^t = X_i\beta + \epsilon_i$$

$$\beta^t \sim N\left(\hat{\beta}, (X'X + A)^{-1}\right), \quad \hat{\beta} = (X'X)^{-1}(X'V^t + A\bar{\beta})$$

- This process is repeated $t \to \infty$. After dropping the first M draws, $\{\beta^t\}$ is a random sample from the posterior distribution (e.g. mean/standard deviation $\approx \beta^{mle}$ and s.e.).

Example 2: Multinomial Probit

- The same insights can be applied to multinomial model with correlated errors.
- Random utility with 4 options:

$$V_{ij} = X_{ij}\beta + \epsilon_{ij} \quad \epsilon_i \sim (0, \Sigma) \text{ and } j = 1, \ldots, 3$$

  as before $V_{i0} = 0$.
- **Data Augmentation:** Given $(\beta^{t-1}, \Sigma^{t-1})$, option 1 is chosen if

$$V_{i1} > V_{i2}, \quad V_{i1} > V_{i3}, \quad V_{i1} > 0$$

  and $V_i \sim N(X_i\beta^t, \Sigma)$. This means that we can draw V's using the GHK procedure:
  1. Draw ϵ_{i1} such that: $\epsilon_1 > -X_{i1}\beta^t$
  2. Draw ϵ_{i2} such that: $\epsilon_1^t - \epsilon_2 > X_{i2}\beta^{t-1} - X_{i1}\beta^{t-1}$
  3. Draw ϵ_{i3} such that: $\epsilon_1^t - \epsilon_3 > X_{i3}\beta^{t-1} - X_{i1}\beta^{t-1}$

Example 2: Multinomial Probit

- **Parameter updating:** Given $\{V_{i1}^t, \ldots, V_{i3}^t\}_{i=1,\ldots,n}$, we can "estimate" (β, Σ) using standard regression (i.e. we have a system of 3 linear equations)

$$V_{ij}^t = X_{ij}\beta + \epsilon_{ij}, \quad \epsilon_i \sim N(0, \Sigma)$$

Example 2: Multinomial Probit

- **Parameter updating:** Given $\{V_{i1}^t, \ldots, V_{i3}^t\}_{i=1,\ldots,n}$, we can "estimate" (β, Σ) using standard regression (i.e. we have a system of 3 linear equations)

$$V_{ij}^t = X_{ij}\beta + \epsilon_{ij}, \quad \epsilon_i \sim N(0, \Sigma)$$

Bayesian updating step for β^{t+1}:

$$\beta^{t+1} \sim N\left(\hat{\beta}, (X'\Omega^{-1}X + A)^{-1}\right)$$

$$\text{and } \hat{\beta} = (X'\Omega^{-1}X + A)^{-1}(X'\Omega^{-1}V^t + A\bar{\beta})$$

where $\Omega = \Sigma^t \otimes I_N$.

Example 2: Multinomial Probit

- **Parameter updating:** Given $\{V_{i1}^t, \ldots, V_{i3}^t\}_{i=1,\ldots,n}$, we can "estimate" (β, Σ) using standard regression (i.e. we have a system of 3 linear equations)

$$V_{ij}^t = X_{ij}\beta + \epsilon_{ij}, \quad \epsilon_i \sim N(0, \Sigma)$$

Bayesian updating step for β^{t+1}:

$$\beta^{t+1} \sim N\left(\hat{\beta}, (X'\Omega^{-1}X + A)^{-1}\right)$$

$$\text{and } \hat{\beta} = (X'\Omega^{-1}X + A)^{-1}(X'\Omega^{-1}V^t + A\bar{\beta})$$

where $\Omega = \Sigma^t \otimes I_N$.
Bayesian updating step for Σ^{t+1}:

$$\Sigma^{t+1^{-1}} \sim W(v + N, (V + S)^{-1}), \; S = \sum_i \epsilon_i \epsilon_i'$$

where (v, V) are priors on Σ^{-1}, and $W(\cdot)$ is the Wishart distribution (i.e. "multivariate" version of the chi-square)