A (quick) Introduction to Structural Modeling

Winter 25/26

Niels Boissonnet, Bielefeld

Motivation

Assume you have data from an economic experiment

And you have a certain idea ("model") that seems to fit behavior

Motivation

Assume you have data from an economic experiment

And you have a certain idea ("model") that seems to fit behavior

- Can you estimate model parameters (say a utility function) to compare them across experiments?

Motivation

Assume you have data from an economic experiment

And you have a certain idea ("model") that seems to fit behavior

- Can you estimate model parameters (say a utility function) to compare them across experiments?
- Is your model significantly better than another one's?

Motivation

Assume you have data from an economic experiment

And you have a certain idea ("model") that seems to fit behavior

- Can you estimate model parameters (say a utility function) to compare them across experiments?
- Is your model significantly better than another one's?
- Does it fit all subjects better or should you distinguish subject types?

Motivation

Assume you have data from an economic experiment

And you have a certain idea ("model") that seems to fit behavior

- Can you estimate model parameters (say a utility function) to compare them across experiments?
- Is your model significantly better than another one's?
- Does it fit all subjects better or should you distinguish subject types?
- How can you use existing data to predict behavior in related decision problems?

Motivation

Assume you have data from an economic experiment

And you have a certain idea ("model") that seems to fit behavior

- Can you estimate model parameters (say a utility function) to compare them across experiments?
- Is your model significantly better than another one's?
- Does it fit all subjects better or should you distinguish subject types?
- How can you use existing data to predict behavior in related decision problems?

We attempt all these things in "structural analyses" of experimental data

Motivation

Assume you have data from an economic experiment

And you have a certain idea ("model") that seems to fit behavior

- Can you estimate model parameters (say a utility function) to compare them across experiments?
- Is your model significantly better than another one's?
- Does it fit all subjects better or should you distinguish subject types?
- How can you use existing data to predict behavior in related decision problems?

We attempt all these things in "structural analyses" of experimental data

- A complete model of behavior with stochastic choice, (social/risk) preferences, strategic beliefs, possibly controlling for subject heterogeneity

Objective of this course

- My objective is to provide you with a first glimpse of what a structural analysis looks like
- Structural Analysis: Estimate the parameters of a behavioral model, using actual data and accounting for noise
- This complements the lecture "Verhaltensorientierte Entscheidungstheorie", which does not discuss how to handle noise
- And it sets the ground for subsequent lectures on social preferences and behavior in games, which extensively deal with noise

Content

- A simple model of stochastic choice: logit
- Model likelihood Computation, numerical maximization
- Standard errors Simple, Huber-Sandwich, bootstrap
- Parameter testing SE-based
- LR testing Vuong nested/nonnested, Schennach-Wilhelm
- Model selection AIC, BIC information criteria
- Model testing LR testing against actual or hypothetical models

Organization

Organization

Webpage Lernraum

Email niels.boissonnet@uni-bielefeld.de

Office hours by appointment, room V8–227

Proceedings of the course

Course schedule:

- **Oct 22 / Oct 29 / Nov 5** – Lectures on structural modeling (2–3 sessions)
- **By then** – Form teams of up to 2 members
- **Nov 19** – Informal check-in: discuss data sources and your progress if needed.
- **Dec 3 & Jan 14** – Interim presentations.
- Interim presentation of progress in early January
- **Jan 24** – Final presentations.
- **End of January** – Submit a brief report documenting your analysis and code.

What is Structural Modeling

A (quick) Introduction to Structural Modeling

Defining Structural Modeling

- Reiss and Wolak (2007)

  "Today economists refer to models that combine explicit economic theories with statistical methods as structural econometric models."

- Nevo and Whinston (2010)

  "Structural modeling attempts to use data to identify the parameters of an underlying economic model, based on models of individual choice or aggregate relations derived from them."

Key idea: Structural models *embed economic theory* into the statistical framework, enabling interpretation of estimated parameters in terms of preferences, technologies, and equilibria.

Structural vs. Non-Structural Approaches

Non-structural (reduced-form) modeling:

- Focuses on statistical relationships between variables.
- Parameters may lack a direct economic interpretation.
- Useful for prediction and descriptive analysis.

Structural modeling:

- Starts from explicit economic theory (utility, equilibrium, etc.).
- Parameters have a clear interpretation (e.g., risk aversion, elasticity).
- Useful for counterfactuals and policy analysis.

Contrast: Structural $\Rightarrow$ theory-driven, interpretable, policy-relevant.
Non-structural $\Rightarrow$ data-driven, flexible, predictive.

Steps in Constructing a Structural Model

Step 1: Specify an economic theory

- Define the *setting*: options of choice, risk, or a game.
- Identify *exogenous variables*: e.g. constraints in choice.
- Specify the *objective function*: typically utility maximization.
- Choose an *equilibrium concept*: Nash equilibrium, behavioral equilibrium, etc.

Step 2: Translate into an econometric model

- Account for deviations between model and data:
  - Researcher uncertainty (unobservables).
  - Agent uncertainty about the environment (mixed strategies, rational inattention).
  - Optimization error by agents (e.g. quantal response equilibrium).
  - Measurement error in observed variables.

Estimating a Structural Model

Step 3: Estimate the econometric model

- Choose functional forms and distributional assumptions.
- Select an estimation method (e.g. maximum likelihood, GMM, Bayesian).
- Compute parameter estimates.
- Conduct specification and robustness tests.

Key idea: Structural modeling links theory $\rightarrow$ econometric model $\rightarrow$ estimation, letting us test hypotheses about behavior and policy.

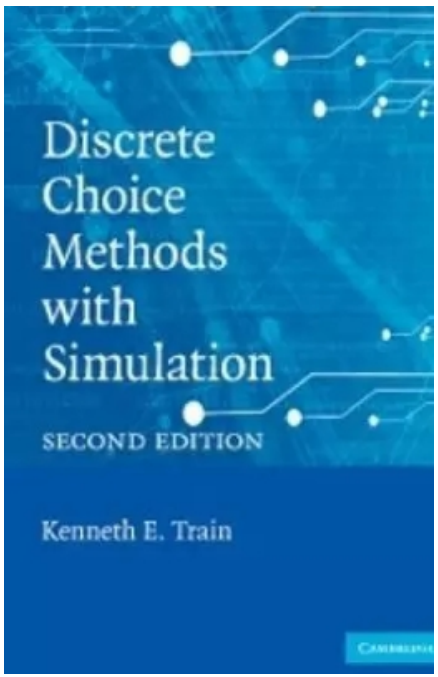
Precious resources

Computing environment:

Code samples in *language R* are provided to illustrate the techniques

- open source, flexible language, many existing extension packages
- R `http://ftp5.gwdg.de/pub/misc/cran/`

Learning Resources:



- **Book:** *Discrete Choice Methods with Simulation*
  https://eml.berkeley.edu/books/choice2.html
- **Youtube Channel:** ResEcon 703 - UMass Amherst
  (https://www.youtube.com/@ResEcon703UMassAmherst)

Covered Topics: Modeling Choice

Covered Topics: Modeling Choice

Introduction to Structural Analysis

- Goals of structural modeling
- Comparison with non-structural (reduced-form) analysis

Covered Topics: Modeling Choice

Introduction to Structural Analysis

- Goals of structural modeling
- Comparison with non-structural (reduced-form) analysis

Random Utility and the Logit Model

- Interpretation of the Random Utility Model
- The Logit model and its key properties
- Behavioral foundations of the Logit model
- Common criticisms and limitations

Covered Topics: Modeling Choice

Introduction to Structural Analysis

- Goals of structural modeling
- Comparison with non-structural (reduced-form) analysis

Random Utility and the Logit Model

- Interpretation of the Random Utility Model
- The Logit model and its key properties
- Behavioral foundations of the Logit model
- Common criticisms and limitations

Maximum Likelihood Estimation

- Numerical maximization methods
- Standard errors and statistical inference
- Model selection and comparison
- Likelihood-ratio tests

Optional Topics: Modeling subject heterogeneity

Optional Topics: Modeling subject heterogeneity

Heterogeneity I: Random Parameters (Mixed Logit)

- Definition and motivation
- Capturing correlation in unobserved preferences
- Estimation methods (Train, 2003)

Optional Topics: Modeling subject heterogeneity

Heterogeneity I: Random Parameters (Mixed Logit)

- Definition and motivation
- Capturing correlation in unobserved preferences
- Estimation methods (Train, 2003)

Heterogeneity II: Discrete Subject Types (Finite Mixture Models)

- Definition and interpretation
- Applications in behavioral and choice models
- Estimation via the EM algorithm
- Model selection using information criteria
- McLachlan and Peel (2000), Bilmes (1998)

Optional Topics: Modeling subject heterogeneity

Heterogeneity I: Random Parameters (Mixed Logit)

- Definition and motivation
- Capturing correlation in unobserved preferences
- Estimation methods (Train, 2003)

Heterogeneity II: Discrete Subject Types (Finite Mixture Models)

- Definition and interpretation
- Applications in behavioral and choice models
- Estimation via the EM algorithm
- Model selection using information criteria
- McLachlan and Peel (2000), Bilmes (1998)

Dynamic Heterogeneity: Changing Types (Markov Switching Models)

- Definition and intuition
- Estimation via the Forward–Backward algorithm (Frühwirth-Schnatter, 2006)

Materials provided

I will provide

- R codes
- Videos from previous lectures
- Topics and papers on which you can work
- Further techniques I'm not going to cover her.

Random utility—An econometric model of utility Theory

A (quick) Introduction to Structural Modeling

Random utility

In this course, **Step 1** and **Step 2** are

Step 1: Specify and economic theory

- setting =alternatives to choose and constraints on choice
- objective function/ equilibrium=maximization of utility

Step 2: Translate into an econometric model

- DMs observed choices look stochastics
- Random utility

An interpretation of the Random Utility Model

An interpretation of the Random Utility Model

1. A DM chooses among J (finite) alternatives from which he obtains a certain level of utility

An interpretation of the Random Utility Model

1. A DM chooses among J (finite) alternatives from which he obtains a certain level of utility
2. He knows this level of utility, and chooses the best
   - Choose alternative $i \in J$, if and only if $U(i) > U(j)$, for all $j \neq i$

An interpretation of the Random Utility Model

1. A DM chooses among J (finite) alternatives from which he obtains a certain level of utility
2. He knows this level of utility, and chooses the best
   - Choose alternative $i \in J$, if and only if $U(i) > U(j)$, for all $j \neq i$
3. **Researcher don't** observe all aspects determining the DM's utility

An interpretation of the Random Utility Model

1. A DM chooses among J (finite) alternatives from which he obtains a certain level of utility
2. He knows this level of utility, and chooses the best
   - Choose alternative $i \in J$, if and only if $U(i) > U(j)$, for all $j \neq i$
3. **Researcher don't** observe all aspects determining the DM's utility

An interpretation of the Random Utility Model

1. A DM chooses among J (finite) alternatives from which he obtains a certain level of utility
2. He knows this level of utility, and chooses the best
   - Choose alternative $i \in J$, if and only if $U(i) > U(j)$, for all $j \neq i$
3. **Researcher don't** observe all aspects determining the DM's utility
   - she observes some attributes of alternatives x_j

An interpretation of the Random Utility Model

1. A DM chooses among J (finite) alternatives from which he obtains a certain level of utility
2. He knows this level of utility, and chooses the best
   - Choose alternative $i \in J$, if and only if $U(i) > U(j)$, for all $j \neq i$
3. **Researcher don't** observe all aspects determining the DM's utility
   - she observes some attributes of alternatives x_j
   - she specifies a *representative utility* $V_j = V(x_j)$

An interpretation of the Random Utility Model

1. A DM chooses among J (finite) alternatives from which he obtains a certain level of utility
2. He knows this level of utility, and chooses the best
   - Choose alternative $i \in J$, if and only if $U(i) > U(j)$, for all $j \neq i$
3. **Researcher don't** observe all aspects determining the DM's utility
   - she observes some attributes of alternatives x_j
   - she specifies a *representative utility* $V_j = V(x_j)$
   - the true utility is presumed to be

$$U(j) = V_j + \varepsilon_j$$

where ε_j is a random variable that captures unobserved aspects of the DM's utility

An interpretation of the Random Utility Model

1. A DM chooses among J (finite) alternatives from which he obtains a certain level of utility
2. He knows this level of utility, and chooses the best
   - Choose alternative $i \in J$, if and only if $U(i) > U(j)$, for all $j \neq i$
3. **Researcher don't** observe all aspects determining the DM's utility
   - she observes some attributes of alternatives x_j
   - she specifies a *representative utility* $V_j = V(x_j)$
   - the true utility is presumed to be

$$U(j) = V_j + \varepsilon_j$$

     where ε_j is a random variable that captures unobserved aspects of the DM's utility
   - the distribution of $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_J)$ is assumed to be $f(\varepsilon)$

An interpretation of the Random Utility Model

1. A DM chooses among J (finite) alternatives from which he obtains a certain level of utility
2. He knows this level of utility, and chooses the best
   - Choose alternative $i \in J$, if and only if $U(i) > U(j)$, for all $j \neq i$
3. **Researcher don't** observe all aspects determining the DM's utility
   - she observes some attributes of alternatives x_j
   - she specifies a *representative utility* $V_j = V(x_j)$
   - the true utility is presumed to be

$$U(j) = V_j + \varepsilon_j$$

   where ε_j is a random variable that captures unobserved aspects of the DM's utility
   - the distribution of $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_J)$ is assumed to be $f(\varepsilon)$

Can this be interpreted in another way?

Choice Probability

Given V the predicted choice probability is then

$$\begin{aligned} P(i) &= \Pr(U(i) > U(j),\ \forall j \neq i) \\ &= \Pr(V_i + \varepsilon_i > V_j + \varepsilon_j,\ \forall j \neq i) \\ &= \Pr(\varepsilon_j - \varepsilon_i < V_i - V_j,\ \forall j \neq i) \\ &= \int \mathbb{1}_{\{\varepsilon_j - \varepsilon_i < V_i - V_j,\ \forall j \neq i\}} f(\varepsilon)\, d\varepsilon. \end{aligned}$$

The integral takes a closed form only for certain specifications of $f(\cdot)$.

- e.g Logit

Choice Probability

Given V the predicted choice probability is then

$$P(i) = \Pr(U(i) > U(j),\ \forall j \neq i)$$

$$= \Pr(\varepsilon_j - \varepsilon_i < V_i - V_j,\ \forall j \neq i)$$
$$= \int \mathbb{1}_{\{\varepsilon_j - \varepsilon_i < V_i - V_j,\ \forall j \neq i\}} f(\varepsilon)\, d\varepsilon.$$

The integral takes a closed form only for certain specifications of $f(\cdot)$.

- e.g Logit

Choice Probability

Given V the predicted choice probability is then

$$P(i) = \Pr(U(i) > U(j),\ \forall j \neq i)$$

$$= \int \mathbb{1}_{\{\varepsilon_j - \varepsilon_i < V_i - V_j,\ \forall j \neq i\}} f(\varepsilon)\, d\varepsilon.$$

The integral takes a closed form only for certain specifications of $f(\cdot)$.

- e.g Logit

Choice Probability

Given V the predicted choice probability is then

$$\begin{aligned} P(i) &= \Pr(U(i) > U(j),\ \forall j \neq i) \\ &= \Pr(V_i + \varepsilon_i > V_j + \varepsilon_j,\ \forall j \neq i) \\ &= \Pr(\varepsilon_j - \varepsilon_i < V_i - V_j,\ \forall j \neq i) \\ &= \int \mathbb{1}_{\{\varepsilon_j - \varepsilon_i < V_i - V_j,\ \forall j \neq i\}} f(\varepsilon)\, d\varepsilon. \end{aligned}$$

The integral takes a closed form only for certain specifications of $f(\cdot)$.

- e.g Logit

Logit model

Multinomial logit

The probability of choosing $j \in J$ is: $\Pr(j) = \frac{\exp\{\lambda \cdot u(j)\}}{\sum_{j' \in J} \exp\{\lambda \cdot u(j')\}}$ for some $\lambda \geq 0$

The logit model results from assuming that

1. Each ε_j is independently and identically distributed
2. ...according to a distribution Gumbel (or type I extreme value)

$$f(\varepsilon_j) = e^{-\epsilon_j} \cdot e^{-e^{-\epsilon_j}} \qquad F(\varepsilon_j) = e^{-e^{-\epsilon_j}}$$

3. with a variance of $\frac{\pi^2}{6}$

A simple model of stochastic choice

Multinomial logit

The probability of $j \in J$ is $\Pr(j) = \frac{\exp\{\lambda \cdot u(j)\}}{\sum_{j' \in J} \exp\{\lambda \cdot u(j')\}}$ for some $\lambda \geq 0$

A simple model of stochastic choice

Multinomial logit

The probability of $j \in J$ is $\Pr(j) = \frac{\exp\{\lambda \cdot u(j)\}}{\sum_{j' \in J} \exp\{\lambda \cdot u(j')\}}$ for some $\lambda \geq 0$

Main properties (detailed discussion later)

- λ is referred to as precision, this will often be the parameter we try to estimate
- For $\lambda = 0$, $P(j) = \frac{1}{J}$, i.e the DM choice is uniformly random
- For $\lambda \to \infty$

$$P(j) > 0 \iff u(j) \geq u(i), \forall i \in J$$

  the DM choice utility maximizing $\lambda \to \infty$
- The logit model can be axiomatized with "simple" behavioral axioms

Digression: Luce's Characterization of the Logit model

Suppose we observe for each $I \subseteq J$ and each alternative $i \in J$, we observe a choice probability $P(j|I)$.

IIA: for all $I, I' \subseteq J$, $i, j \in I \cap I'$, $\frac{P(j|I)}{P(i|I)} = \frac{P(j|I')}{P(i|I')}$.

- Adding or removing other "irrelevant" alternatives $l, k...$ doesn't change how likely to pick i relative to j.

Theorem

P satisfies **IIA** *if and only if* $\Pr(j|I) = \frac{\exp\{\lambda \cdot u(j)\}}{\sum_{j' \in I} \exp\{\lambda \cdot u(j')\}}$ *for some* $\lambda > 0$.

Digression: Problem with the logit model

From Debreu, assume we three options for traveling bb= blue bus, rb=red bus, and a car c.

The DM does not care about the color of the bus

- so $P(bb|T) = P(rb|T) > 0$ for all $T \ni bb, rb$
- But we should also expect $P(bb|\{c, bb\}) = P(bb|\{c, bb, rb\})$

This is impossible under the logit model, for otherwise

Digression: Problem with the logit model

From Debreu, assume we three options for traveling bb= blue bus, rb=red bus, and a car c.

The DM does not care about the color of the bus

- so $P(bb|T) = P(rb|T) > 0$ for all $T \ni bb, rb$
- But we should also expect $P(bb|\{c, bb\}) = P(bb|\{c, bb, rb\})$

This is impossible under the logit model, for otherwise

$$\frac{P(bb|\{c, bb\})}{P(c|\{c, bb\})} = \frac{P(bb|\{c, bb, rb\})}{P(c|\{c, bb, rb\})} \implies 1 = \frac{P(c|\{c, bb\})}{P(c|\{c, bb, rb\})}$$

Similarly, $\frac{P(c|\{c,rb\})}{P(c|\{c,bb,rb\})} = 1$ so that $P(c|\{c, bb, rb\}) = 1$

$$P(rb|\{c, bb, rb\}) = 0.$$

Similate some logit choices

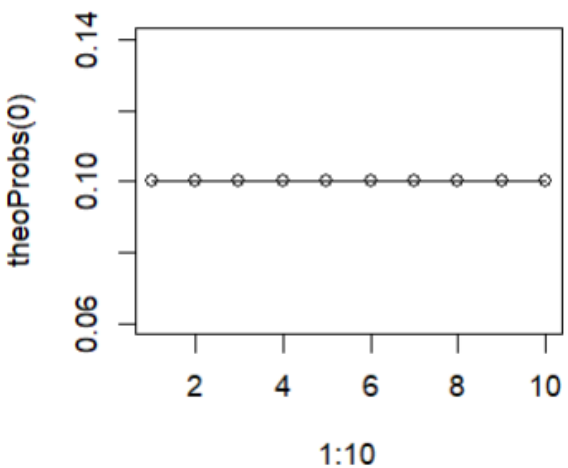
Utility function $u(a) = 5 - |5.5 - a|$ from a choice set choice from set $A = \{1, 2, \ldots, 10\}$

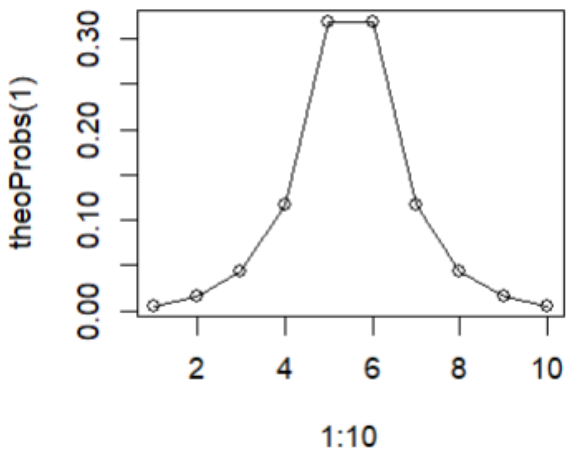
Similate some logit choices

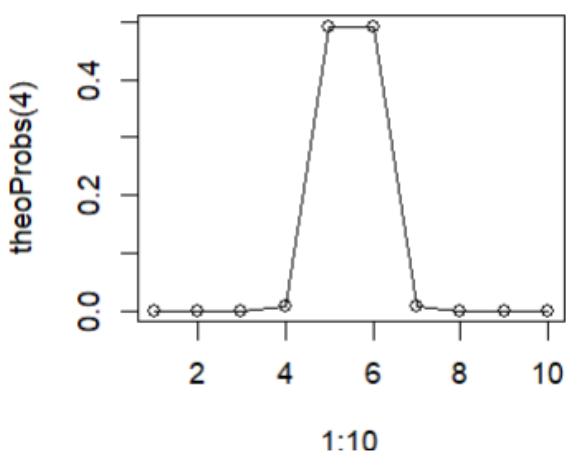
Utility function $u(a) = 5 - |5.5 - a|$ from a choice set choice from set $A = \{1, 2, \ldots, 10\}$

Compute utility and theoretical probabilities for $\lambda = 1$

```
u <- 5-abs(5.5 - 1:10)
LogitProbs <- function(lambda) exp(lambda * u) / sum (exp(lambda * u))
plot(1:10,LogitProbs(0),type="l")
plot(1:10,LogitProbs(1),type="l")
plot(1:10,LogitProbs(4),type="l")
```







Maximum likelihood

A (quick) Introduction to Structural Modeling

Maximum likelihood: Intuition

Three draws from a normal distribution $\mathcal{N}(\mu, \sigma^2)$, with unknown μ and σ^2

$$\mathbf{y} = \{48.7, 50.9, 48.8\}$$

Is it likelier that **y** be generated from $\mathcal{N}(0, 1)$ or $\mathcal{N}(50, 1)$?

- $\mathcal{N}(50, 1)$ is clearly more likely

How to define what is the most likely data generating process?

- The prob. density that $\mathcal{N}(\mu, 1)$ generates 48.7: $\frac{1}{\sqrt{2\pi}} e^{-(48.7-\mu)^2}$
- Also regarded as the probability that μ given that 48.7 is observed
- Joint density of observing **y**, $\prod_{x \in \mathbf{y}} \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \right)$
- **Idea of Maximum likely:** Maximize this joint density
  - Select the model parameters such that "observations are most likely"

Maximum Likelihood

Idea Select the model parameters such that "observations are most likely"

Notation Set of observations $\mathbf{y}$,
Probability of observation $y \in \mathbf{y}$ if model has parameters θ is $\Pr(y|\theta)$

Maximum Likelihood

Idea Select the model parameters such that "observations are most likely"

Notation Set of observations $\mathbf{y}$,
Probability of observation $y \in \mathbf{y}$ if model has parameters θ is $\Pr(y|\theta)$

Likelihood function $L(\theta|\mathbf{y})$ describes the probability that a model with parameters θ generates observations $\mathbf{y}$; LL is the log-likelihood

$$L(\theta|\mathbf{y}) = \prod_{y \in \mathbf{y}} \underbrace{\Pr(y|\theta)}_{\text{density if continuous}} \qquad LL(\theta|\mathbf{y}) = \sum_{y \in \mathbf{y}} \ln \underbrace{\Pr(y|\theta)}_{\text{density if continuous}}$$

Maximum Likelihood

Idea Select the model parameters such that "observations are most likely"

Notation Set of observations $\mathbf{y}$,
Probability of observation $y \in \mathbf{y}$ if model has parameters θ is $\Pr(y|\theta)$

Likelihood function $L(\theta|\mathbf{y})$ describes the probability that a model with parameters θ generates observations $\mathbf{y}$; LL is the log-likelihood

$$L(\theta|\mathbf{y}) = \prod_{y \in \mathbf{y}} \underbrace{\Pr(y|\theta)}_{\text{density if continuous}} \qquad LL(\theta|\mathbf{y}) = \sum_{y \in \mathbf{y}} \ln \underbrace{\Pr(y|\theta)}_{\text{density if continuous}}$$

ML estimator $\theta_{ML} = \arg\max_\theta LL(\theta|\mathbf{y})$

Maximum Likelihood an example

Three draws from a Poisson distribution with density $f(y|\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}$

$$\mathbf{y} = \{2, 0, 1\}$$

The log likelihood function is

$$\log \frac{e^{-\lambda}\lambda^2}{2} - \lambda + \log(e^{-\lambda}\lambda) = 3\log\lambda - 3\lambda - \log 2$$

At the maximum we have

$$\frac{3}{\lambda} = 3 \implies \lambda = 1$$

The ML estimator is $\lambda_{ML} = 1$

However, in general we cannot obtain a reduced form as here and we need to approach it through R!

Properties of Maximum likelihood

Under certain regularity conditions the ML estimator has nice properties:

- **Consistent**: θ_{ML} converges in probability to the true parameter
  - As the sample size increases it becomes vanishingly close to the true parameter
- **Asymptotic Efficiency:** Among all consistent estimators, the ML estimator achieves the lowest possible asymptotic variance.
- **Asymptotic Normality:** The asymptotic distribution of the ML estimator follows a normal law
  - $\theta_{ML}^n \approx \mathcal{N}(\theta_0, \frac{1}{n} I(\theta_0)^{-1})$ when n is large
  - θ_0 is the true parameter(s)
  - $I(\theta_0) = -\mathbb{E}\left(\frac{\partial LL(\theta_0)}{\partial \theta_0 \partial \theta_0'}\right)$ (more certainty about θ_{ML} when LL has more curvature).

A Maximization Algorithm `nlm()` Works

`nlm()` implements a variant of the **Newton–Raphson method** for minimizing a (smooth) objective function $f(\theta)$.

It requires to specify:

- The function
- A starting point
- Other arguments

It works like this

1. Idea: Find the parameter vector θ^* such that the gradient is zero:

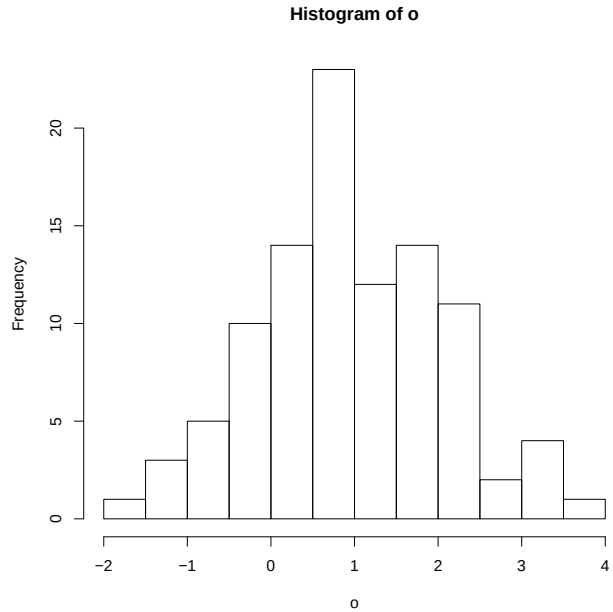
$$\nabla f(\theta^*) = 0.$$

2. Iterative update rule: $\theta_{t+1} = \theta_t - \underbrace{H(\theta_t)^{-1}}_{\text{Inverse of the Hessian}} \underbrace{\nabla f(\theta_t)}_{\text{Gradient}},$

Numerical likelihood maximization (est-mu.R)

Numerical likelihood maximization (est-mu.R)

Example: Compute the estimate mean μ of normally distributed observations



Observations from $\mathcal{N}(1, 1)$

```
o <- rnorm (100, mean=1, sd=1)
```

Estimate μ, assuming $\sigma = 1$

```
ll <- function (m, obs) -sum (log (dnorm (obs, mean=m, sd=1)))
opt1 <- nlm (ll, p=0, obs=o)
```

Numerical likelihood maximization (est-mu.R)

Estimate μ, assuming σ

```
ll <- function (m,p,obs) -sum (log (dnorm (obs, mean=p[1], p[2])))
opt1 <- nlm (ll, p=c(0,1),obs=o)
```

- First line defines the negative log-likelihood with μ and σ as free parameters.
- Second line calls `nlm` with initial values for μ and σ.

Estimate precision λ of logit choices (est-lambda.R)

Estimate precision λ of logit choices (est-lambda.R)

Multinomial logit choice from set $A = \{1, 2, \ldots, 10\}$

Utility function $u(a) = 5 - |5.5 - a|$ is known, precision λ is unknown

Estimate precision λ of logit choices (est-lambda.R)

Multinomial logit choice from set $A = \{1, 2, \ldots, 10\}$

Utility function $u(a) = 5 - |5.5 - a|$ is known, precision λ is unknown

Compute utility and theoretical probabilities for $\lambda = 1$

```
u <- 5-abs(5.5 - 1:10)
LogitProbs <- exp(1 * u) / sum (exp(1 * u))
```

Estimate precision λ of logit choices (est-lambda.R)

Multinomial logit choice from set $A = \{1, 2, \ldots, 10\}$

Utility function $u(a) = 5 - |5.5 - a|$ is known, precision λ is unknown

Compute utility and theoretical probabilities for $\lambda = 1$

```
u <- 5-abs(5.5 - 1:10)
LogitProbs <- exp(1 * u) / sum (exp(1 * u))
```

Sample choices for these probabilities

```
o <- sample (1:10, size=100, replace=TRUE, prob=LogitProbs)
```

Estimate precision λ of logit choices (est-lambda.R)

Multinomial logit choice from set $A = \{1, 2, \ldots, 10\}$

Utility function $u(a) = 5 - |5.5 - a|$ is known, precision λ is unknown

Compute utility and theoretical probabilities for $\lambda = 1$

```
u <- 5-abs(5.5 - 1:10)
LogitProbs <- exp(1 * u) / sum (exp(1 * u))
```

Sample choices for these probabilities

```
o <- sample (1:10, size=100, replace=TRUE, prob=LogitProbs)
```

Maximize the likelihood function

```
ll <- function (lambda, obs) {
    p <- exp(lambda * u) / sum (exp(lambda * u))
    min (1e6, -sum (log (p[obs])))
}

opt <- nlm (ll, p=2, obs=o)
```

Practical Issues in Optimization

Why optimization can be tricky

- Newton–Raphson methods (e.g. `nlm()`) may fail if:
  - the log-likelihood is not concave
  - the problem is multi-modal (several local optima-notably when we control for heterogeneity)
- Alternative algorithms:
  - `newuoa()` (package `minqa`): gradient-free
  - `cmaes()`: evolutionary, global search

Checking convergence

- Inspect the gradient (≈ 0 at optimum)
- Inspect the Hessian (not too close to 0, reasonable curvature)
- Strategy: first run a global optimizer, then refine with Newton–Raphson

Practical Tips for Robust Estimation

Parameter scaling

- Rescale parameters so all are of similar order (≈ 1)
- Ensures a 10% change is comparable across dimensions
- Especially useful for gradient-free algorithms

Starting values

- Use a grid search to find good initial guesses
- Reduces risk of convergence to poor local optima

Constraints

- Many models require restrictions:
  - $\lambda > 0$ (precision parameters), $\rho_k \geq 0$ (weights)
- Enforce constraints by:
  - using constrained optimizers (`nlminb()`, `BOBYQA`)
  - reparameterizing / bounding within likelihood

Standard errors and inference

A (quick) Introduction to Structural Modeling

Standard Errors in ML (est-se.R)

Key idea: After estimating parameters by ML, we want to know how precise they are (i.e., compute standard errors).

- measures how much an estimate would vary if you repeated the experiment (or sampling) many times.

Asymptotic result: $\theta^n_{ML} \approx \mathcal{N}(\theta_0, \frac{1}{n} I(\theta_0)^{-1})$

- $I = -\mathbb{E}\left(\frac{\partial LL(\theta_0)}{\partial \theta_0 \partial \theta_0'}\right)$=Fisher Information (curvature of the log-likelihood).
- Variance of MLs: I^{-1}.
- Standard errors = square root of diagonal of I^{-1}.

Practical shortcut: we don't know θ_0, but we can approximate it

$$I(\theta_0) \approx -\frac{1}{n} H(\hat{\theta}_{ML}) \implies I(\theta_0)^{-1} \approx -n \cdot H(\hat{\theta}_{ML})^{-1}$$

where H = Hessian of loglikelihood (matrix of 2nd derivatives) at the optimum.

Computing SEs in R (est-se.R)

Estimating SE of normal distribution (μ, σ) of a Normal distribution

```
o <- rnorm(100, mean=1, sd=1)   #data
l12 <- function(p, obs) -sum(log(dnorm(obs, mean=p[1], sd=p[2])))
opt3 <- nlm(l12, p=c(0,1), hessian=TRUE, obs=o)
se <- sqrt(diag(solve(opt3$hessian)))  #standard errors
```

- `hessian=TRUE`: compute Hessian at optimum.
- `solve()`: invert Hessian $\rightarrow$ variance-cov matrix.
- `sqrt(diag(...))`: extract standard errors.

Comment: It would be better for the accuracy of the alogorithm to impose $sd > 0$

- Can start with $sd = log(p[2])$, maximize, and then scale the obtained standar error $exp(se[2])$

When Usual SEs Fail

- Standard errors from the **Hessian** are valid only if:
  - The model is correctly specified, i.e. the assumed distribution is the true data generating process
  - Observations are independent and identically distributed (i.i.d.)
- If these assumptions fail, standard errors may be **biased or too small**.
- Common problems:
  - Heteroskedasticity (unequal error variance)
  - Serial or clustered dependence
  - Model misspecification

Robust (Huber–White) "Sandwich" SEs

$$\widehat{\mathrm{Var}}(\hat{\theta}_{ML}) = H(\hat{\theta}_{ML})^{-1}\, S(\hat{\theta}_{ML})\, H(\hat{\theta}_{ML})^{-1}$$

- $H(\hat{\theta}_{ML})$: Hessian (matrix of 2nd derivatives) — the "bread."
  - tells you how curved the log-likelihood is (how precisely the model pins down the parameter).
- $S(\hat{\theta}_{ML})$: outer product of score vectors — the "meat."
  - $S(\hat{\theta}) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} s_i(\hat{\theta})\, s_i(\hat{\theta})^\top\right]$, where $s_i(\theta) = \nabla_\theta \log f(y_i|\theta)$
  - tells you how variable the information (scores) really is across observations
- Called a **sandwich estimator** because of this structure.
- Provides valid inference even when model assumptions fail.
- In R, use functions from packages like `sandwich` or `lmtest`.

Bootstrapping Standard Errors

Idea: An alternative to using the Hessian is to *bootstrap* the standard errors.

- Does not rely on strong parametric assumptions—preferable in small samples
- Useful with clustered or repeated observations

Case resampling procedure:

1. Resample data with replacement (at subject level).
2. Compute the statistic of interest on each resample.
3. Repeat many times ($R \approx 500{-}2000$).
4. Empirical distribution $\rightsquigarrow$ estimate SEs.

Bootstrap Example in R

Estimate (μ, σ) for Normal data using bootstrap.

```
estF <- function(data, indices) {
    nlm(ll2, p=c(1,1), obs=data[indices])$estimate
}

library(boot)
b <- boot(data = o, statistic = estF, R = 1000)

print(b)
```

- `boot()` handles resampling and repetitions.
- `R = 1000` = number of bootstrap replications.
- `print(b)` shows bootstrap estimates and SEs.

Model selection: Bayesian Information Criterion

Key idea:

Adding parameters usually improves fit $\Rightarrow$ risk of overfitting.

Model selection: Bayesian Information Criterion

Key idea:

Adding parameters usually improves fit $\Rightarrow$ risk of overfitting.

Setup:**BIC** penalizes model complexity.

- Compare models M_1 and M_2 with log-likelihoods LL_1, LL_2
- Subjects $s \in S$; individual likelihoods $l_{1,s}$, $l_{2,s}$
- Parameters: d_1 (for M_1), d_2 (for M_2)

Model selection: Bayesian Information Criterion

Key idea:

Adding parameters usually improves fit $\Rightarrow$ risk of overfitting.

Setup:**BIC** penalizes model complexity.

- Compare models M_1 and M_2 with log-likelihoods LL_1, LL_2
- Subjects $s \in S$; individual likelihoods $l_{1,s}$, $l_{2,s}$
- Parameters: d_1 (for M_1), d_2 (for M_2)

Bayesian Information Criterion (Schwarz, 1978)

$$\mathrm{BIC} = \underbrace{-2\,LL}_{\textit{Fit term}} + \underbrace{d\,\log S}_{\textit{Penalty term}}$$

Model selection: Bayesian Information Criterion

Key idea:

Adding parameters usually improves fit $\Rightarrow$ risk of overfitting.

Setup:**BIC** penalizes model complexity.

- Compare models M_1 and M_2 with log-likelihoods LL_1, LL_2
- Subjects $s \in S$; individual likelihoods $l_{1,s}$, $l_{2,s}$
- Parameters: d_1 (for M_1), d_2 (for M_2)

Bayesian Information Criterion (Schwarz, 1978)

$$\mathrm{BIC} = \underbrace{-2\,LL}_{\textit{Fit term}} + \underbrace{d\,\log S}_{\textit{Penalty term}}$$

- Smaller BIC $\Rightarrow$ preferred model.
- *Remark*: $d \log S$ grows with parameters and sample size

Model selection: Bayesian Information Criterion

Key idea:

Adding parameters usually improves fit $\Rightarrow$ risk of overfitting.

Setup:**BIC** penalizes model complexity.

- Compare models M_1 and M_2 with log-likelihoods LL_1, LL_2
- Subjects $s \in S$; individual likelihoods $l_{1,s}$, $l_{2,s}$
- Parameters: d_1 (for M_1), d_2 (for M_2)

Bayesian Information Criterion (Schwarz, 1978)

$$\mathrm{BIC} = \underbrace{-2\,LL}_{\textit{Fit term}} + \underbrace{d\,\log S}_{\textit{Penalty term}}$$

- Smaller BIC $\Rightarrow$ preferred model.
- *Remark*: $d \log S$ grows with parameters and sample size

Takeaway: BIC balances goodness-of-fit and parsimony.

Model Comparison via Likelihoods

Basic idea

Compare two models M_1 and M_2 using their log-likelihoods (LL_1, LL_2).

- **Nested models:** $M_1 \subset M_2$

$$D = 2(LL_2 - LL_1) \overset{H_0}{\sim} \chi^2(d_2 - d_1)$$

- **Non-nested models:** use Vuong test

$$Z = \frac{LR/n}{\hat{\omega}/\sqrt{n}} \overset{H_0}{\sim} \mathcal{N}(0,1)$$

where $LR = LL_2 - LL_1$ and $\hat{\omega}^2$ is the variance of log-likelihood differences.

Limitation: Both tests assume at least one model is correctly specified.

Robustness

In practice, **all models are misspecified**

- Standard LR / Vuong tests can give misleading inference if both models are wrong or overlapping.

A test that is robust to misspecification

- Weighted log-likelihood ratio:

$$LR = \frac{1}{n}\sum_{i=1}^{n}\left(\omega_i(\varepsilon)LL_{A,i} - \omega_{i+1}(\varepsilon)LL_{B,i}\right) \text{ with}$$

- Where weights alternate:

$$\omega_j(\varepsilon) = \begin{cases} 1, & j \text{ odd} \\ 1+\varepsilon, & j \text{ even} \end{cases}$$

- $\varepsilon = 0$: reduces to Vuong test.
- Small $\varepsilon > 0$: adds robustness to misspecification.

Robust LR Test Statistic

- Test statistic:

$$t = \frac{\sqrt{n}\, LR}{\tilde{\sigma}} \text{ with } \tilde{\sigma}^2 = (1+\varepsilon)\sigma^2 + \frac{\varepsilon^2}{2}(\sigma_A^2 + \sigma_B^2)$$

- Asymptotics:

$$t \overset{H_0}{\sim} \mathcal{N}(0,1)$$

- Valid even if both models are misspecified or overlapping.

Methods for Robust p-values

- In misspecified or clustered models, standard p-values may fail.
- Several robust alternatives are available.

For instance,

1. **Asymptotic LR distribution**
   - Schennach–Wilhelm LR test + Huber sandwich errors.
   - Can appear "misaligned" in small samples.
2. **Normal approximation**
   - Treat estimator $\hat{\lambda}$ as asymptotically normal.
   - Use robust standard errors $\sigma_{\hat{\lambda}}$.
3. **Bootstrap**
   - Resample data, recompute LR many times.
   - Empirical distribution gives p-value.

Caveats and Comments

- Robust methods help with misspecification and clustering, but they do **not** control for
  - Researcher degrees of freedom (model/data selection),
  - Multiple testing or post-hoc inference,
  - Misinterpretation of p-values.

Mixture models

A (quick) Introduction to Structural Modeling

Subject heterogeneity

Subject heterogeneity

Assume the subjects are not homogeneous, but differ in their precision λ

Subject heterogeneity

Assume the subjects are not homogeneous, but differ in their precision λ

Example Distribution of λ and utility parameters in choice under risk (Wilcox, 2008)

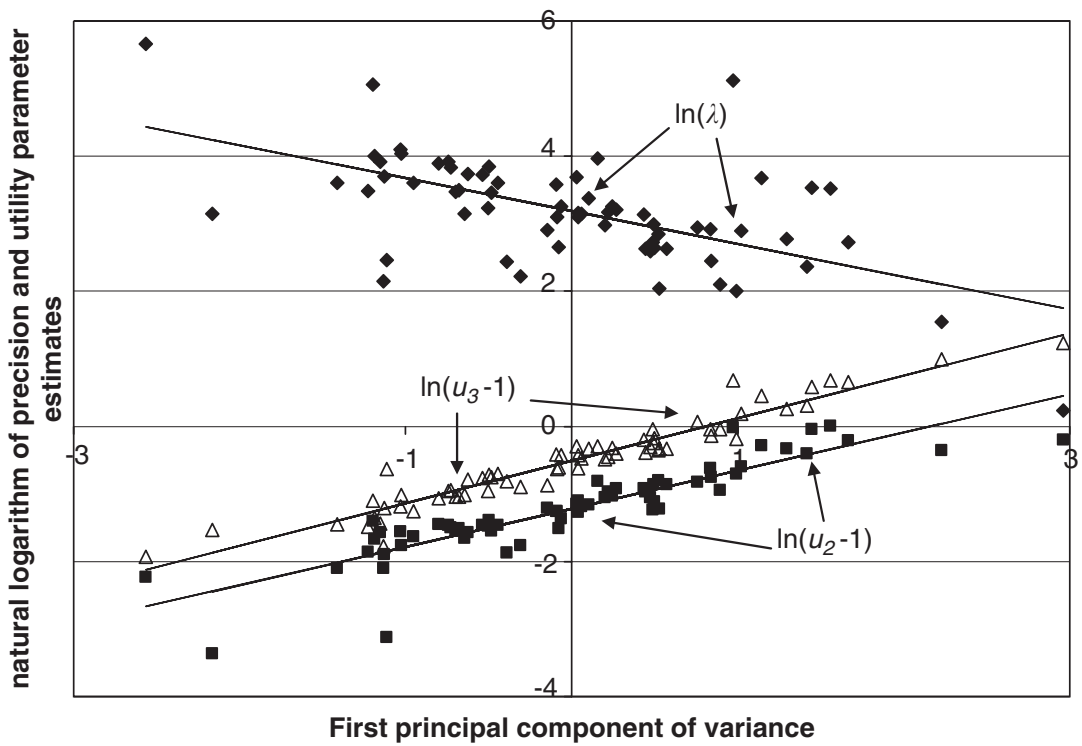


Fig. 3. Shared Variance of Initial Individual Parameter Estimates in the (EU,Strong) Model.

Example: Distribution of social motive weights

Weights of social motives in simple social interactions (Blanco et al., 2011)

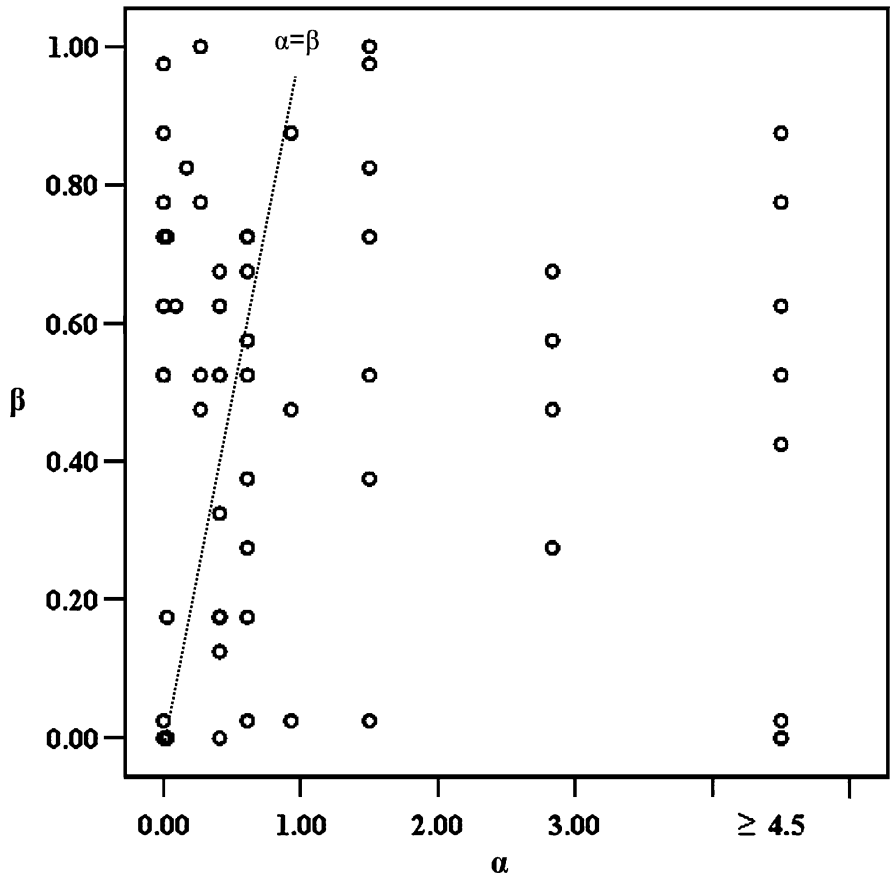


Fig. 1. The joint α–β distribution. Each dot in the figure represents an individual's α and β parameter. Observations to the left of the $\alpha = \beta$ line have $\alpha < \beta$. Observations with the highest level of a cannot be pinned down more narrowly than $\alpha \geqslant 4.5$.

Definition of mixture models

Definition of mixture models

Notation

- Subjects $s \in S$, observations o_s; $O = \{o_s\}_{s \in S}$
- $\Pr(o_s|\lambda)$ is the predicted probability of o_s given $\lambda \geq 0$
- $f_\lambda(\cdot|\alpha)$ is the density of λ in the population, with parameters α

Definition of mixture models

Notation

- Subjects $s \in S$, observations o_s; $O = \{o_s\}_{s \in S}$
- $\Pr(o_s|\lambda)$ is the predicted probability of o_s given $\lambda \geq 0$
- $f_\lambda(\cdot|\alpha)$ is the density of λ in the population, with parameters α

Mixed logit model

$$LL(\alpha \mid O) = \sum_{s \in S} \log \int_0^\infty f_\lambda(\lambda|\alpha) \Pr(o_s|\lambda)\, d\lambda$$

Definition of mixture models

Notation

- Subjects $s \in S$, observations o_s; $O = \{o_s\}_{s \in S}$
- $\Pr(o_s|\lambda)$ is the predicted probability of o_s given $\lambda \geq 0$
- $f_\lambda(\cdot|\alpha)$ is the density of λ in the population, with parameters α

Mixed logit model

$$LL(\alpha \mid O) = \sum_{s \in S} \log \int_0^\infty f_\lambda(\lambda|\alpha) \Pr(o_s|\lambda)\, d\lambda$$

Extension Multiple observations per subject $o_s = \{o_{s,t}\}_{t \leq T}$

$$LL(\alpha \mid O) = \sum_{s \in S} \log \int_0^\infty f_\lambda(\lambda|\alpha) \prod_{t \leq T} \Pr(o_{s,t}|\lambda)\, d\lambda$$

Possible distributions for λ

Possible distributions for λ

Censored normal $\lambda \sim \max\left(0, \mathcal{N}(\mu, \sigma)\right)$

Possible distributions for λ

Censored normal $\lambda \sim \max\left(0, \mathcal{N}(\mu, \sigma)\right)$

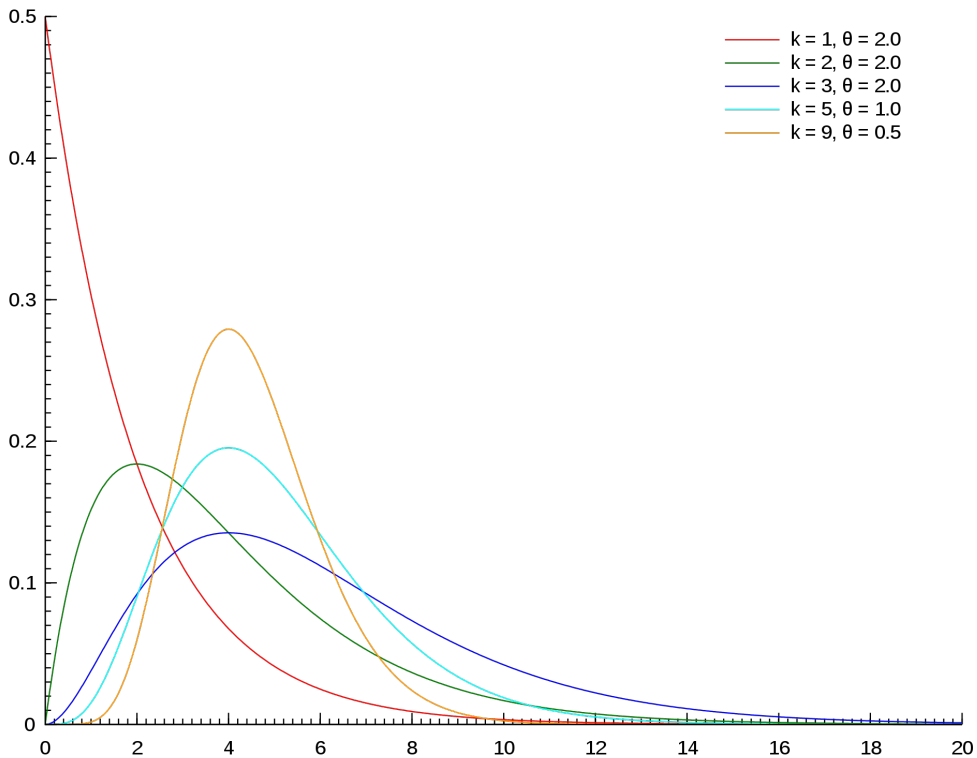
Truncated normal $f(\lambda) = \phi(\lambda)/(1 - \Phi(0))$
with ϕ, Φ as density, cdf of $\mathcal{N}(\mu, \sigma)\big)$

Possible distributions for λ

Censored normal $\lambda \sim \max\left(0, \mathcal{N}(\mu, \sigma)\right)$

Truncated normal $f(\lambda) = \phi(\lambda)/(1 - \Phi(0))$
with ϕ, Φ as density, cdf of $\mathcal{N}(\mu, \sigma)\big)$

Gamma $\lambda \sim \Gamma(\kappa, \theta)\big)$

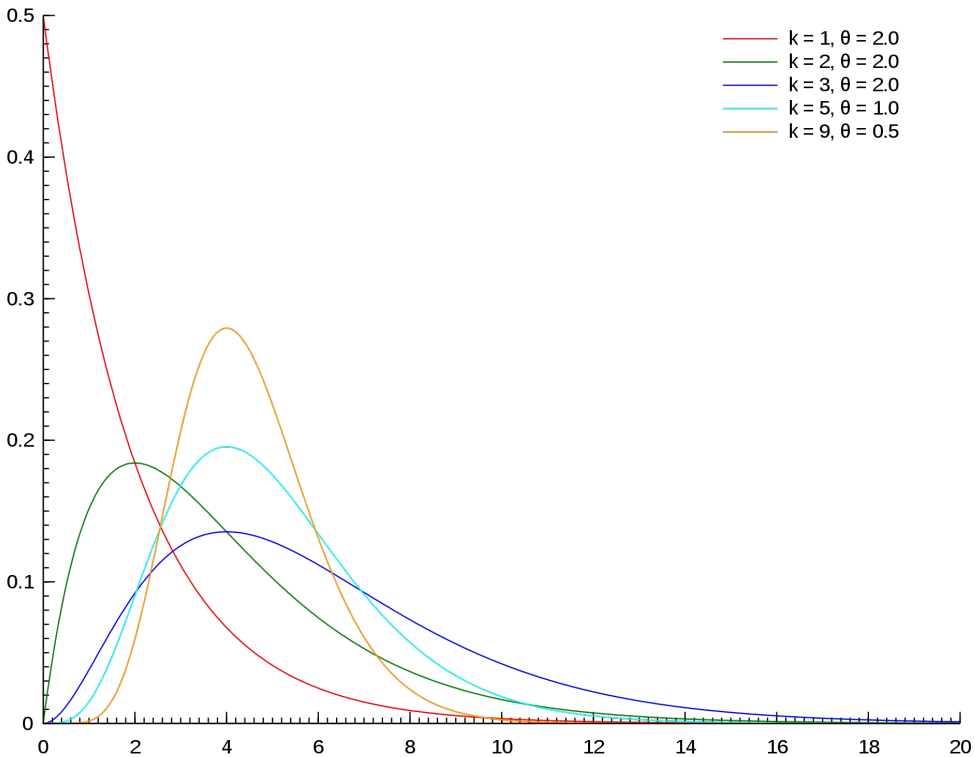


Possible distributions for λ

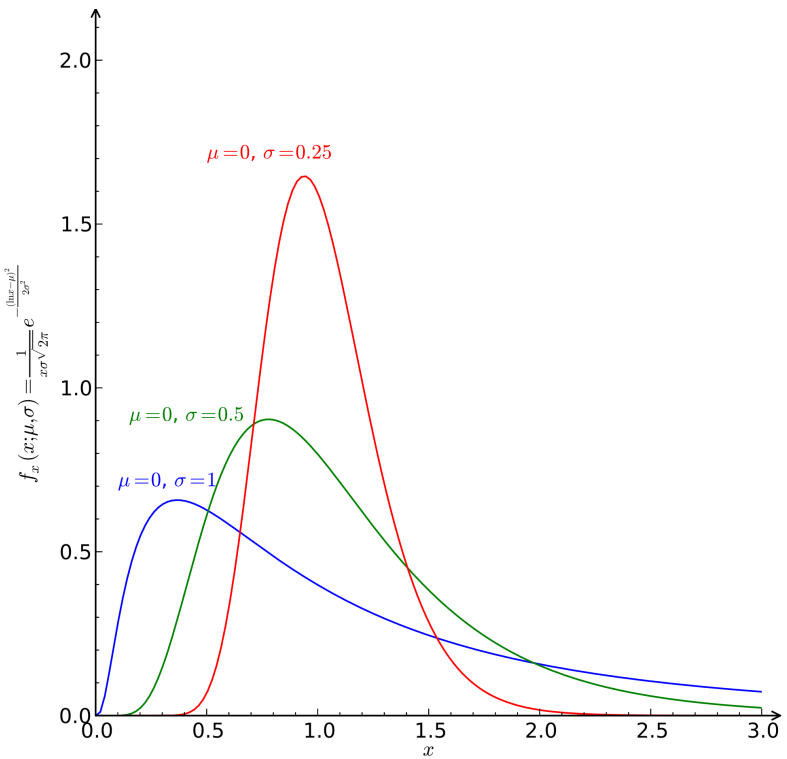
Censored normal $\lambda \sim \max\left(0, \mathcal{N}(\mu, \sigma)\right)$

Truncated normal $f(\lambda) = \phi(\lambda)/(1 - \Phi(0))$
with ϕ, Φ as density, cdf of $\mathcal{N}(\mu, \sigma)$

Gamma $\lambda \sim \Gamma(\kappa, \theta)$



Log-normal $\log(\lambda) \sim \mathcal{N}(\mu, \sigma)$



Pictures: Wikimedia Commons

Maximum simulated likelihood

Maximum simulated likelihood

Let $f_\lambda(\cdot|\alpha)$ denote the density of λ, with parameter α

Maximum simulated likelihood

Let $f_\lambda(\cdot|\alpha)$ denote the density of λ, with parameter α

Let $Q(\alpha) \subset \mathbb{R}$ denote a finite sequence of realizations from $f_\lambda(\cdot|\alpha)$. The simulated log-likelihood is

$$\widetilde{LL}(\alpha \mid O) = \sum_{s \in S} \log \left(\frac{1}{Q(\alpha)} \sum_{q \in Q(\alpha)} \Pr(o_s|q) \right)$$

The α maximizing $\widetilde{LL}$ is called maximum simulated likelihood estimator.

Maximum simulated likelihood

Let $f_\lambda(\cdot|\alpha)$ denote the density of λ, with parameter α

Let $Q(\alpha) \subset \mathbb{R}$ denote a finite sequence of realizations from $f_\lambda(\cdot|\alpha)$. The simulated log-likelihood is

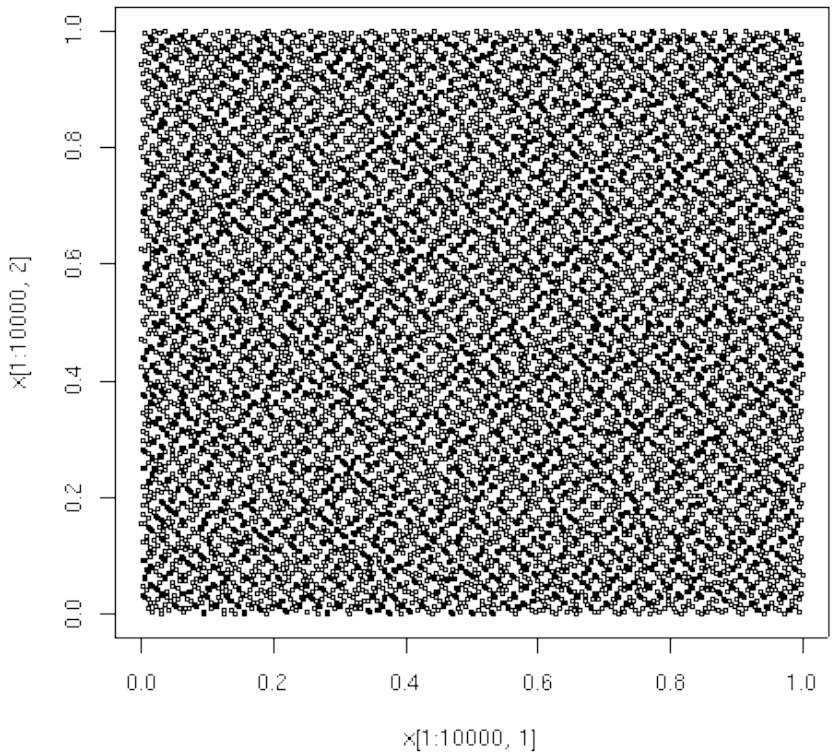
$$\widetilde{LL}(\alpha \mid O) = \sum_{s \in S} \log \left(\frac{1}{Q(\alpha)} \sum_{q \in Q(\alpha)} \Pr(o_s|q) \right)$$

The α maximizing $\widetilde{LL}$ is called maximum simulated likelihood estimator.

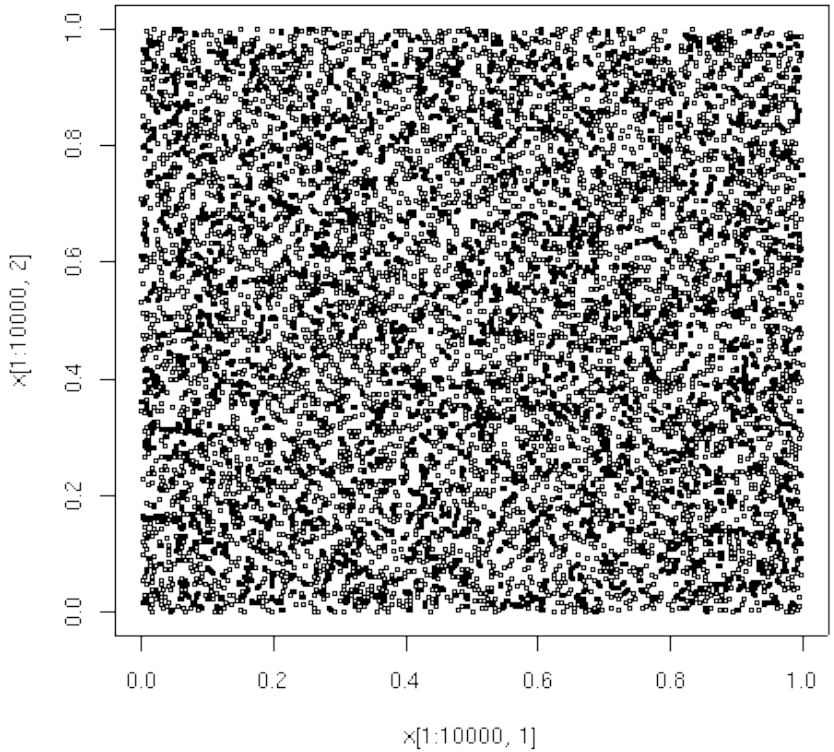
Quasi-random numbers (Low-discrepancy sequences)

- low discrepancy = λ-space is filled in evenly, without gaps
- frequently used in numerical integration/MSL (see e.g. (Train, 2003))

Sobol quasi-random numbers

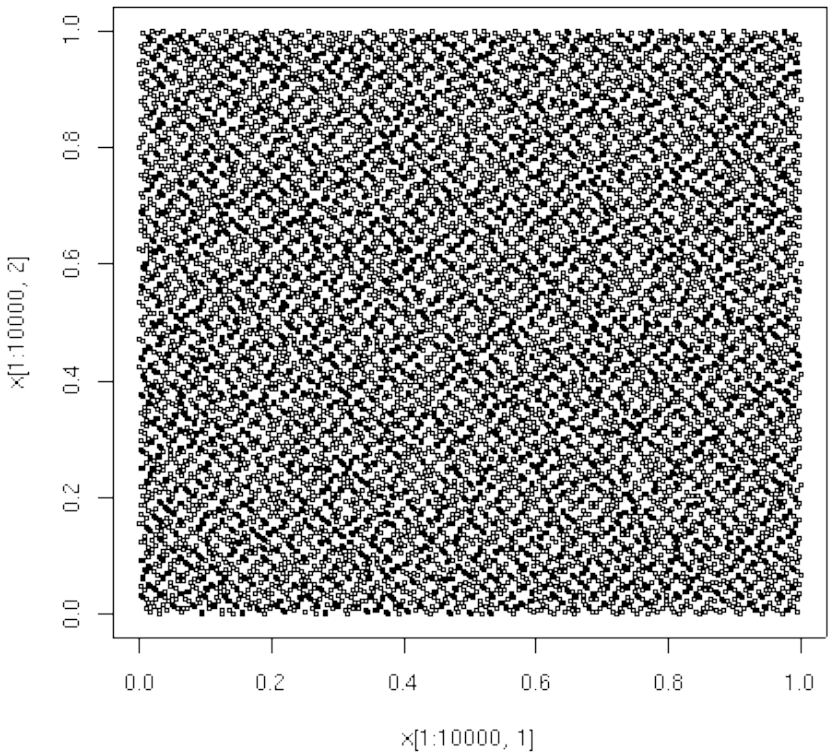


Random numbers

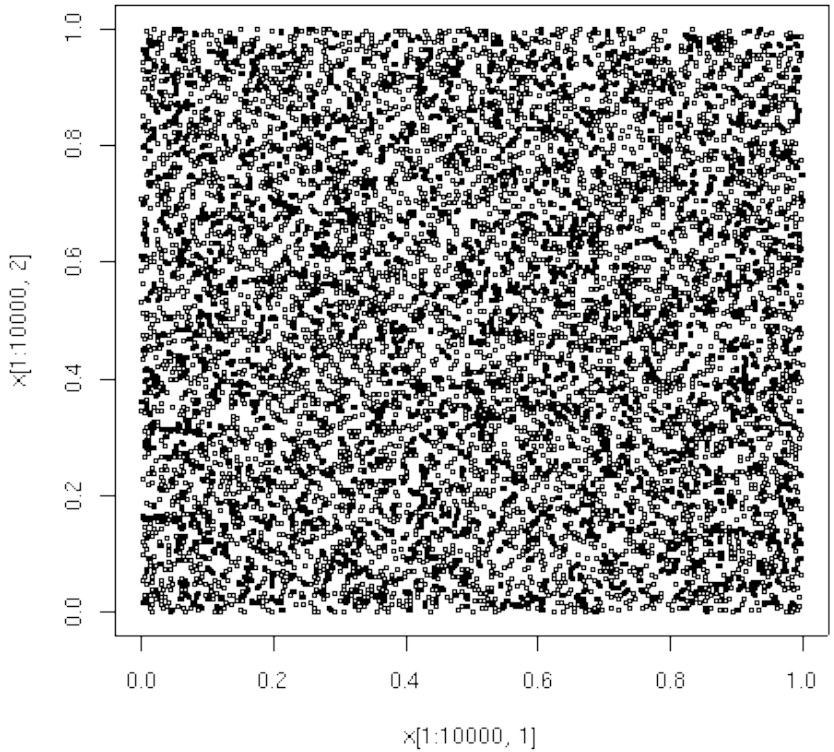


Pictures: Wikimedia Commons

Sobol quasi-random numbers



Random numbers



Pictures: Wikimedia Commons

Generation of Sobol sequences (est-msl.R)

```
library (gsl)

q <- qrng_alloc (type = c("sobol"), dim=2)
qrngSize <- 16384
qrngMat <- qrng_get (q, qrngSize)
```

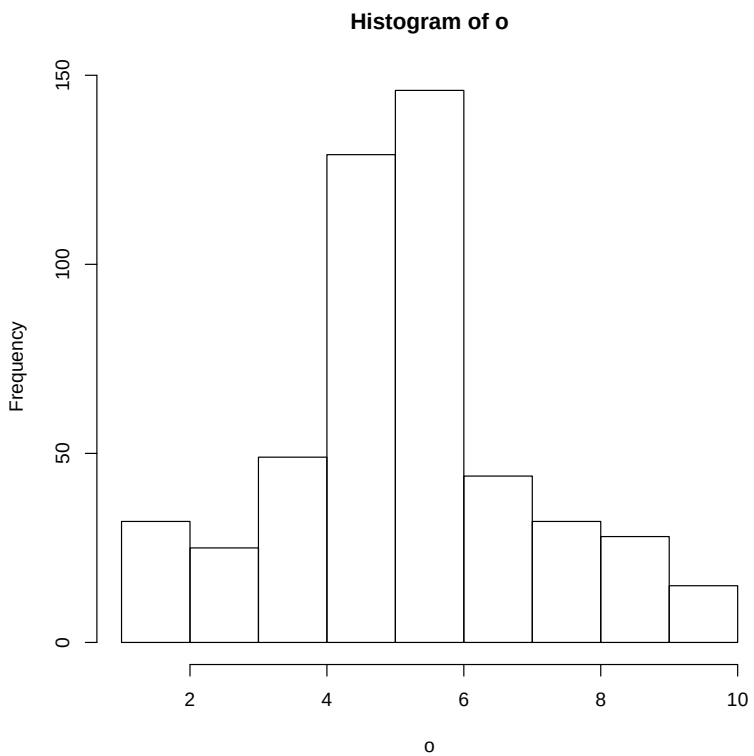
Generating a sample from mixed logit models

```
u <- 5-abs(5.5 - 1:10)
theoLambdaV <- rgamma (n=500, shape=1, scale=1)
o <- rep (0, 500)
for (i in 1:500) {
    LogitProbs <- exp(theoLambdaV[i]*u) / sum(exp(theoLambdaV[i]*u))
    o[i] <- sample (1:10, size=1, replace=TRUE, prob=LogitProbs)
}
```

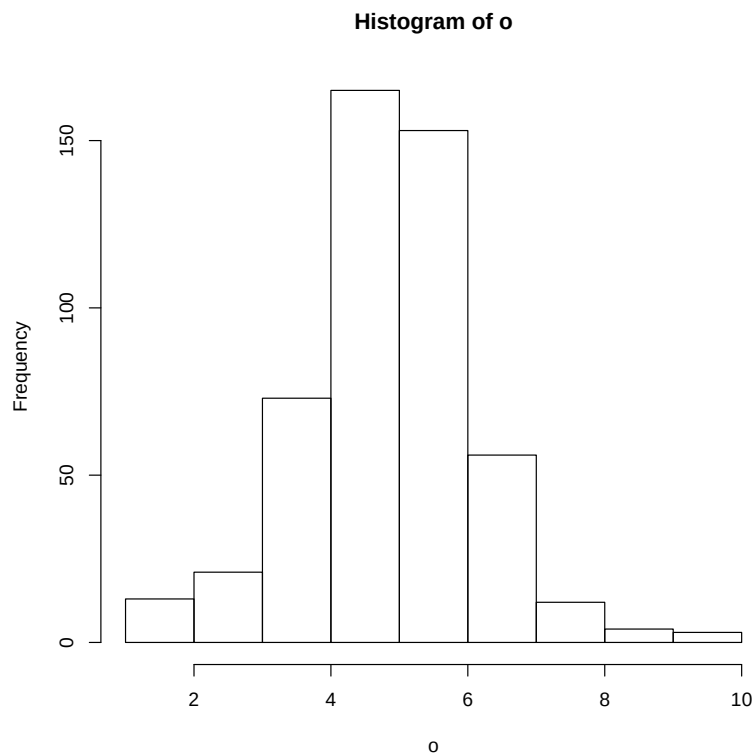
Generating a sample from mixed logit models

```
u <- 5-abs(5.5 - 1:10)
theoLambdaV <- rgamma (n=500, shape=1, scale=1)
o <- rep (0, 500)
for (i in 1:500) {
    LogitProbs <- exp(theoLambdaV[i]*u) / sum(exp(theoLambdaV[i]*u))
    o[i] <- sample (1:10, size=1, replace=TRUE, prob=LogitProbs)
}
```

Mixed logit choices (est-msl.R)



Logit choices (est-lambda.R)



Implementation of MSL (est-msl.R)

Implementation of MSL (est-msl.R)

This is a simple, inefficient implementation. It uses 256 draws to simulate the likelihood, which is not a lot.

Implementation of MSL (est-msl.R)

This is a simple, inefficient implementation. It uses 256 draws to simulate the likelihood, which is not a lot.

Implementation of MSL

```
ll <- function (alpha, obs) {
lambdaV <- qgamma(qrngMat[1:256,1],shape=alpha[1],scale=alpha[2])

ll <- 0
for (os in obs) {
    indL <- 0
    for (lambda in lambdaV) {
       p <- exp(lambda * u) / sum (exp(lambda * u))
         indL <- indL + p[os]
        }
    ll <- ll + log (indL / length(lambdaV))
    }
    cat (alpha, "=>", ll, "\n")
    return (min (1e6, -ll))
}

opt <- nlm (ll, p=c(2,0.5), obs=o)
```

Correlated random effects

Correlated random effects

Consider logit choices of $a \in \{1, 2, \ldots, 10\}$ with Cobb-Douglas utilities $u(a) = (11 - a)^{1-\alpha} (1 + a)^{\alpha}$

Assume subjects are heterogeneous in (λ, α), where $\lambda \geq 0$ and $\alpha \in \mathbb{R}$.
The heterogeneity is unobserved and the effects on λ, α may be correlated

Correlated random effects

Consider logit choices of $a \in \{1, 2, \ldots, 10\}$ with Cobb-Douglas utilities $u(a) = (11 - a)^{1-\alpha} (1 + a)^{\alpha}$

Assume subjects are heterogeneous in (λ, α), where $\lambda \geq 0$ and $\alpha \in \mathbb{R}$. The heterogeneity is unobserved and the effects on λ, α may be correlated

Possible econometric model (Next week shows an alternative approach)
λ is log-normal, α is normal, and subject $s \in S$ has parameters

- Precision $\lambda_s = \exp(\lambda_0 + u_s^{\lambda})$
- Precision $\alpha_s = \alpha_0 + u_s^{\alpha}$

where $(u_s^{\lambda}, u_s^{\alpha})$ are normal with zero mean and covariance matrix Ω

Correlated random effects

Consider logit choices of $a \in \{1, 2, \ldots, 10\}$ with Cobb-Douglas utilities $u(a) = (11 - a)^{1-\alpha} (1 + a)^{\alpha}$

Assume subjects are heterogeneous in (λ, α), where $\lambda \geq 0$ and $\alpha \in \mathbb{R}$. The heterogeneity is unobserved and the effects on λ, α may be correlated

Possible econometric model (Next week shows an alternative approach)
λ is log-normal, α is normal, and subject $s \in S$ has parameters

- Precision $\lambda_s = \exp(\lambda_0 + u_s^{\lambda})$
- Precision $\alpha_s = \alpha_0 + u_s^{\alpha}$

where $(u_s^{\lambda}, u_s^{\alpha})$ are normal with zero mean and covariance matrix Ω

Likelihood function Parameter vector $p = (\lambda_0, \alpha_0, \Omega)$

$$LL(p \mid O) = \sum_{s \in S} \log \int_0^{\infty} \int_{-\infty}^{\infty} f(\lambda, \alpha | p) \Pr(o_s | \lambda, \alpha) \, d\alpha \, d\lambda$$

Simulation of correlated effects

Simulation of correlated effects

Let `qrngMat` denote a $M \times 2$ matrix of quasi-random numbers

Simulation of correlated effects

Let `qrngMat` denote a $M \times 2$ matrix of quasi-random numbers

Transform them into standard normal random numbers

```
uLambda <- qnorm (qrngMat[,1])
uAlpha  <- qnorm (qrngMat[,2])
```

Simulation of correlated effects

Let `qrngMat` denote a $M \times 2$ matrix of quasi-random numbers

Transform them into standard normal random numbers

```
uLambda <- qnorm (qrngMat[,1])
uAlpha  <- qnorm (qrngMat[,2])
```

Compute Cholesky decomposition, $\Omega = R'R$, (R upper triangular)

```
O <- matrix (c(1,0.5,0.5,1), nrow=2)
R <- chol (O)
```

Simulation of correlated effects

Let `qrngMat` denote a $M \times 2$ matrix of quasi-random numbers

Transform them into standard normal random numbers

```
uLambda <- qnorm (qrngMat[,1])
uAlpha  <- qnorm (qrngMat[,2])
```

Compute Cholesky decomposition, $\Omega = R'R$, (R upper triangular)

```
O <- matrix (c(1,0.5,0.5,1), nrow=2)
R <- chol (O)
```

Now, the correlated (but still centralized) draws from $f(\lambda, \alpha)$ with covariance Ω result as $(\lambda, \alpha)' = R' \times (u^\lambda, u^\alpha)'$, or $(\lambda, \alpha) = (u^\lambda, u^\alpha) \times R$

```
corrMat <- cbind (uLambda, uAlpha) %*% R #
```

Verify result `cor(corrMat[,1], corrMat[,2])`, `sd(corrMat[,1])`, ...

Simulation of correlated effects

Let qrngMat denote a $M \times 2$ matrix of quasi-random numbers

Transform them into standard normal random numbers

```
uLambda <- qnorm (qrngMat[,1])
uAlpha  <- qnorm (qrngMat[,2])
```

Compute Cholesky decomposition, $\Omega = R'R$, (R upper triangular)

```
O <- matrix (c(1,0.5,0.5,1), nrow=2)
R <- chol (O)
```

Now, the correlated (but still centralized) draws from $f(\lambda, \alpha)$ with covariance Ω result as $(\lambda, \alpha)' = R' \times (u^\lambda, u^\alpha)'$, or $(\lambda, \alpha) = (u^\lambda, u^\alpha) \times R$

```
corrMat <- cbind (uLambda, uAlpha) %*% R #
```

Verify result `cor(corrMat[,1], corrMat[,2])`, `sd(corrMat[,1])`, ...

Finally, the actual draws from $f(\lambda, \alpha)$ are

```
lambda <- exp (lambda0 + corrMat[,1]); alpha <- alpha0 + corrMat[,2]
```

Finite mixtures

A (quick) Introduction to Structural Modeling

Finite mixtures

Finite mixtures

Sometimes, a subject pool/data set may appear to consist of clearly distinguishable *discrete* components.

How can this be verified/modeled econometrically?

Finite mixtures

Sometimes, a subject pool/data set may appear to consist of clearly distinguishable *discrete* components.

How can this be verified/modeled econometrically?

Example 1 Donations in dictator games (Andreoni and Miller, 2002)

You have a budget of $B > 0$ tokens and player 2 has nothing. You can give him as many tokens as you wish. The value of each token may be different for the two of you. These are histograms of transfers for various price ratios $(2:1, 1:2, 1:1)$

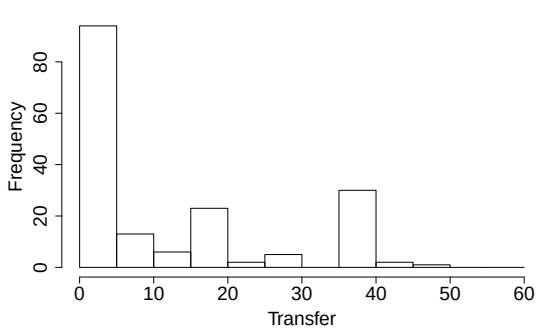
Finite mixtures

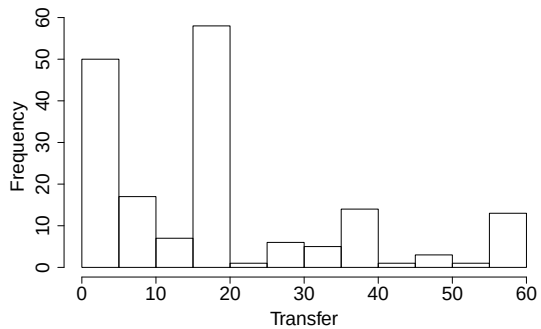
Sometimes, a subject pool/data set may appear to consist of clearly distinguishable *discrete* components.

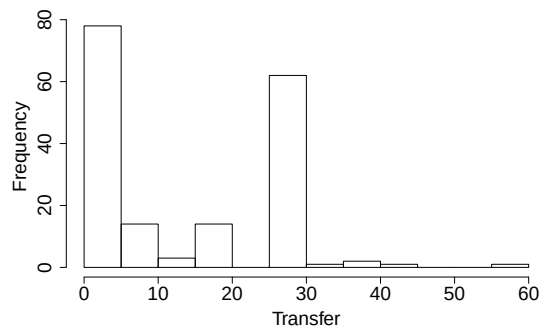
How can this be verified/modeled econometrically?

Example 1 Donations in dictator games (Andreoni and Miller, 2002)

You have a budget of $B > 0$ tokens and player 2 has nothing. You can give him as many tokens as you wish. The value of each token may be different for the two of you. These are histograms of transfers for various price ratios $(2:1, 1:2, 1:1)$





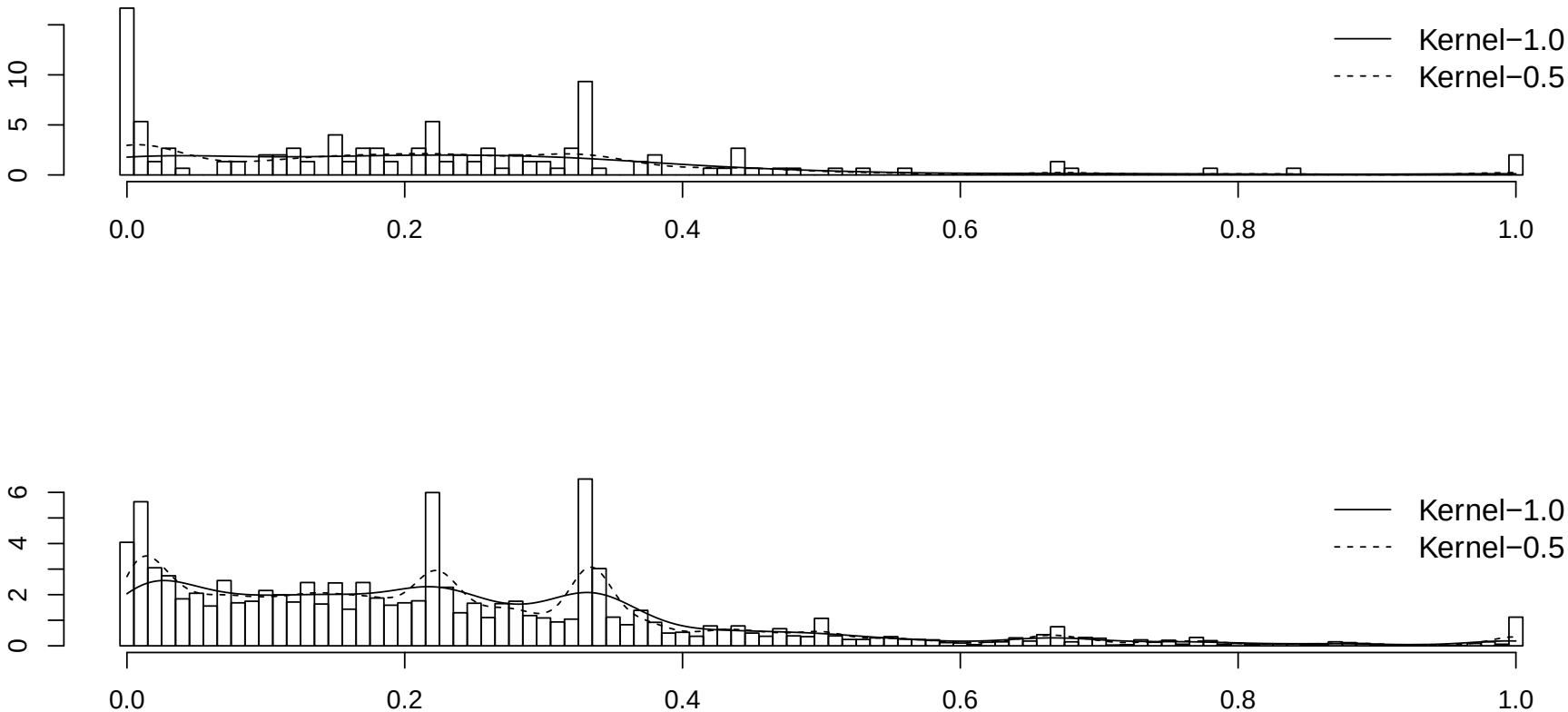


Example 2 Choices in guessing games (Bosch-Domènech et al., 2002)

$n > 2$ players, everybody picks a real number between 0 and 1 (or 100). The guess closest to 2/3 of the mean wins a prize.

Example 2 Choices in guessing games (Bosch-Domènech et al., 2002)

$n > 2$ players, everybody picks a real number between 0 and 1 (or 100). The guess closest to 2/3 of the mean wins a prize.



Finite mixture model

Finite mixture model

Notation

- Components $k \in K$ with weights ρ_k and parameters θ_k
- Subjects $s \in S$ with observations o_s
- Probability of o_s conditional on s belonging to $k \in K$ with θ_k is $\Pr(o_s|\theta_k)$

Finite mixture model

Notation

- Components $k \in K$ with weights ρ_k and parameters θ_k
- Subjects $s \in S$ with observations o_s
- Probability of o_s conditional on s belonging to $k \in K$ with θ_k is $\Pr(o_s|\theta_k)$

Unconditional probability of o_s is $\sum_{k \in K} \rho_k \cdot \Pr(o_s|\theta_k)$

Finite mixture model

Notation

- Components $k \in K$ with weights ρ_k and parameters θ_k
- Subjects $s \in S$ with observations o_s
- Probability of o_s conditional on s belonging to $k \in K$ with θ_k is $\Pr(o_s|\theta_k)$

Unconditional probability of o_s is $\sum_{k \in K} \rho_k \cdot \Pr(o_s|\theta_k)$

Likelihood function

$$LL(\rho, \theta|O) = \sum_{s \in S} \ln \sum_{k \in K} \rho_k \cdot \Pr(o_s|\theta_k)$$

Finite mixture model

Notation

- Components $k \in K$ with weights ρ_k and parameters θ_k
- Subjects $s \in S$ with observations o_s
- Probability of o_s conditional on s belonging to $k \in K$ with θ_k is $\Pr(o_s|\theta_k)$

Unconditional probability of o_s is $\sum_{k \in K} \rho_k \cdot \Pr(o_s|\theta_k)$

Likelihood function

$$LL(\rho, \theta|O) = \sum_{s \in S} \ln \sum_{k \in K} \rho_k \cdot \Pr(o_s|\theta_k)$$

- Heterogenous parameters such as precision/altruism: Continuous mixtures
- Heterogenous depth of reasoning such as level-k: Finite mixture models
- Choice under risk Finite mixture over general aspects (e.g. preference structure) and continuous mixture over actual parameters

Posterior probability that s belongs to k

$$\Pr(k|o_s, \rho, \theta) = \frac{\rho_k \cdot \Pr(o_s|\theta_k)}{\sum_{k' \in K} \rho_{k'} \cdot \Pr(o_s|\theta_{k'})}$$

Define $\hat{\tau}_{sk} = \Pr(k|o_s, \rho, \theta)$

Posterior probability that s belongs to k

$$\Pr(k|o_s, \rho, \theta) = \frac{\rho_k \cdot \Pr(o_s|\theta_k)}{\sum_{k' \in K} \rho_{k'} \cdot \Pr(o_s|\theta_{k'})}$$

Define $\hat{\tau}_{sk} = \Pr(k|o_s, \rho, \theta)$

Information criterion Used to estimate/select number of components
BIC tends to overestimate the number of components – penalty for superfluous components required (further discussion in (McLachlan and Peel, 2000), Chapter 6). Measure for the latter is entropy of the posterior classification

$$\mathrm{En}(\hat{\tau}) = -\sum_{s \in S} \sum_{k \in K} \hat{\tau}_{sk} \ln \hat{\tau}_{sk},$$

Posterior probability that s belongs to k

$$\Pr(k|o_s, \rho, \theta) = \frac{\rho_k \cdot \Pr(o_s|\theta_k)}{\sum_{k' \in K} \rho_{k'} \cdot \Pr(o_s|\theta_{k'})}$$

Define $\hat{\tau}_{sk} = \Pr(k|o_s, \rho, \theta)$

Information criterion Used to estimate/select number of components
BIC tends to overestimate the number of components – penalty for superfluous components required (further discussion in (McLachlan and Peel, 2000), Chapter 6). Measure for the latter is entropy of the posterior classification

$$\mathrm{En}(\hat{\tau}) = -\sum_{s \in S} \sum_{k \in K} \hat{\tau}_{sk} \ln \hat{\tau}_{sk},$$

Integrated classification likelihood

$$\mathit{ICL\text{-}BIC} = -2\,LL + d \cdot \ln |S| + 2\,\mathrm{En}(\hat{\tau})$$

Example: Finite mixture of normal RVs (est-fm-norm.R)

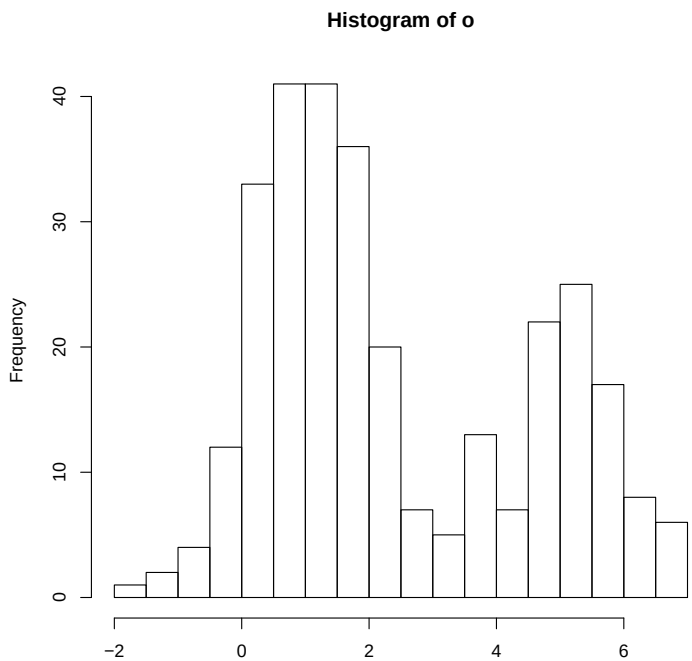
Example: Finite mixture of normal RVs (est-fm-norm.R)

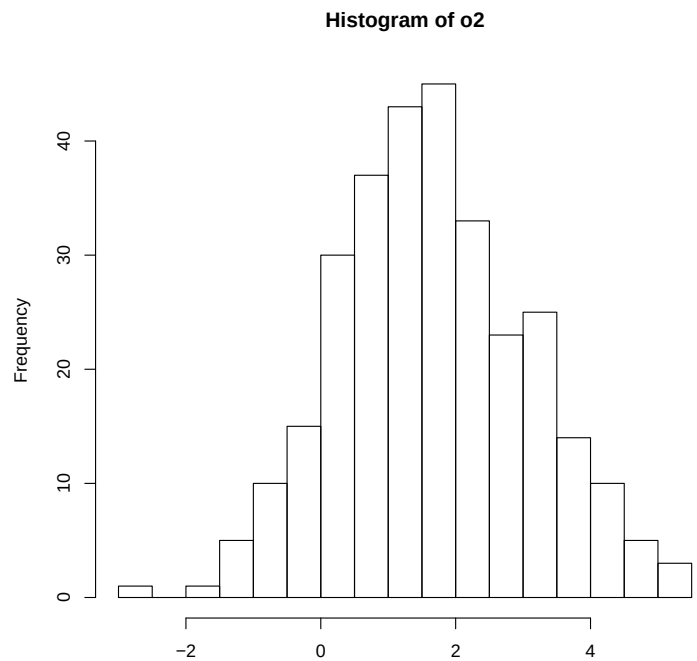
Left 2:1 mixture from $\mathcal{N}(1,1)$ and $\mathcal{N}(5,1)$

```
o <- c(rnorm (200, mean=1, sd=1), rnorm (100, mean=5, sd=1))
```

Right 2:1 mixture from $\mathcal{N}(1,1)$ and $\mathcal{N}(3,1)$

```
o2 <- c(rnorm (200, mean=1, sd=1), rnorm (100, mean=3, sd=1))
```





Estimation by direct likelihood maximization

Estimation by direct likelihood maximization

Estimation assuming $\sigma = 1$.

Estimation by direct likelihood maximization

Estimation assuming $\sigma = 1$.

Log-likelihood

```
ll <- function (p, obs) {
    indL1 <- dnorm (obs, mean=p[2], sd=1)
    indL2 <- dnorm (obs, mean=p[3], sd=1)
    min (1e6, -sum (log (p[1]*indL1 + (1-p[1])*indL2)))
}
```

Estimation by direct likelihood maximization

Estimation assuming $\sigma = 1$.

Log-likelihood

```
ll <- function (p, obs) {
    indL1 <- dnorm (obs, mean=p[2], sd=1)
    indL2 <- dnorm (obs, mean=p[3], sd=1)
    min (1e6, -sum (log (p[1]*indL1 + (1-p[1])*indL2)))
}
```

Maximization by `nlm()`

```
opt <- nlm (ll, p=c(0.5,0,4), obs=o)
```

Estimation by direct likelihood maximization

Estimation assuming $\sigma = 1$.

Log-likelihood

```
ll <- function (p, obs) {
    indL1 <- dnorm (obs, mean=p[2], sd=1)
    indL2 <- dnorm (obs, mean=p[3], sd=1)
    min (1e6, -sum (log (p[1]*indL1 + (1-p[1])*indL2)))
}
```

Maximization by `nlm()`

```
opt <- nlm (ll, p=c(0.5,0,4), obs=o)
```

Alternative Maximize by `newuoa()` or `cmaes()`

Expectation-Maximization (EM) algorithm

[1] With $LL(\rho, \theta|O) = \sum_{s\in S} \ln \sum_{k\in K} \rho_k \cdot \Pr(o_s|\theta_k)$, the FOC $\partial LL/\partial(\rho, \theta) = 0$ yields $\hat{\rho}_k = \frac{1}{n} \sum_{s\in S} \Pr(k|o_s, \hat{\rho}, \hat{\theta})$ as well as $\sum_{s\in S} P(k|o_s, \hat{\rho}, \hat{\theta}) \cdot \frac{\partial \ln \Pr(o_s|\hat{\theta}_k)}{\partial \theta_k} = 0 \quad \forall k$, and thus θ_k^{+1} as provided above maximizes the conditional LL.

Expectation-Maximization (EM) algorithm

Efficient algorithm to estimate FM models (Dempster et al., 1977; Bilmes, 1998)

- yields substantial acceleration in large data sets

[1] With $LL(\rho, \theta|O) = \sum_{s \in S} \ln \sum_{k \in K} \rho_k \cdot \Pr(o_s|\theta_k)$, the FOC $\partial LL / \partial(\rho, \theta) = 0$ yields $\hat{\rho}_k = \frac{1}{n} \sum_{s \in S} \Pr(k|o_s, \hat{\rho}, \hat{\theta})$ as well as $\sum_{s \in S} P(k|o_s, \hat{\rho}, \hat{\theta}) \cdot \frac{\partial \ln \Pr(o_s|\hat{\theta}_k)}{\partial \theta_k} = 0 \quad \forall k$, and thus θ_k^{+1} as provided above maximizes the conditional LL.

Expectation-Maximization (EM) algorithm

Efficient algorithm to estimate FM models (Dempster et al., 1977; Bilmes, 1998)

- yields substantial acceleration in large data sets

Expectation Given (ρ, θ) compute the posterior probability of component inclusion $\Pr(k|o_s, \rho, \theta)$ for all subjects s.

[1] With $LL(\rho, \theta|O) = \sum_{s \in S} \ln \sum_{k \in K} \rho_k \cdot \Pr(o_s|\theta_k)$, the FOC $\partial LL / \partial(\rho, \theta) = 0$ yields $\hat{\rho}_k = \frac{1}{n} \sum_{s \in S} \Pr(k|o_s, \hat{\rho}, \hat{\theta})$ as well as $\sum_{s \in S} P(k|o_s, \hat{\rho}, \hat{\theta}) \cdot \frac{\partial \ln \Pr(o_s|\hat{\theta}_k)}{\partial \theta_k} = 0 \quad \forall k$, and thus θ_k^{+1} as provided above maximizes the conditional LL.

Expectation-Maximization (EM) algorithm

Efficient algorithm to estimate FM models (Dempster et al., 1977; Bilmes, 1998)

- yields substantial acceleration in large data sets

Expectation Given (ρ, θ) compute the posterior probability of component inclusion $\Pr(k|o_s, \rho, \theta)$ for all subjects s.

Maxmization Update (ρ, θ) to be optimal conditional on these posteriors being the correct probabilities of component inclusion.[1] For all $k \in K$,

$$\theta_k^{+1} \in_{\theta'} \sum_{s \in S} \Pr(k|o_s, \rho, \theta) \cdot \ln \Pr(o_s|\theta') \qquad \rho_k^{+1} = \sum_{s \in S} \Pr(k|o_s, \rho, \theta)$$

[1] With $LL(\rho, \theta|O) = \sum_{s \in S} \ln \sum_{k \in K} \rho_k \cdot \Pr(o_s|\theta_k)$, the FOC $\partial LL / \partial(\rho, \theta) = 0$ yields $\hat{\rho}_k = \frac{1}{n} \sum_{s \in S} \Pr(k|o_s, \hat{\rho}, \hat{\theta})$ as well as $\sum_{s \in S} P(k|o_s, \hat{\rho}, \hat{\theta}) \cdot \frac{\partial \ln \Pr(o_s|\hat{\theta}_k)}{\partial \theta_k} = 0 \quad \forall k$, and thus θ_k^{+1} as provided above maximizes the conditional LL.

Expectation-Maximization (EM) algorithm

Efficient algorithm to estimate FM models (Dempster et al., 1977; Bilmes, 1998)

- yields substantial acceleration in large data sets

Expectation Given (ρ, θ) compute the posterior probability of component inclusion $\Pr(k|o_s, \rho, \theta)$ for all subjects s.

Maxmization Update (ρ, θ) to be optimal conditional on these posteriors being the correct probabilities of component inclusion.[1] For all $k \in K$,

$$\theta_k^{+1} \in_{\theta'} \sum_{s \in S} \Pr(k|o_s, \rho, \theta) \cdot \ln \Pr(o_s|\theta') \qquad \rho_k^{+1} = \sum_{s \in S} \Pr(k|o_s, \rho, \theta)$$

Stop if distance between (ρ, θ) and (ρ^{+1}, θ^{+1}) is low.

[1] With $LL(\rho, \theta|O) = \sum_{s \in S} \ln \sum_{k \in K} \rho_k \cdot \Pr(o_s|\theta_k)$, the FOC $\partial LL / \partial(\rho, \theta) = 0$ yields $\hat{\rho}_k = \frac{1}{n} \sum_{s \in S} \Pr(k|o_s, \hat{\rho}, \hat{\theta})$ as well as $\sum_{s \in S} P(k|o_s, \hat{\rho}, \hat{\theta}) \cdot \frac{\partial \ln \Pr(o_s|\hat{\theta}_k)}{\partial \theta_k} = 0 \quad \forall k$, and thus θ_k^{+1} as provided above maximizes the conditional LL.

EM-Algorithm_new.R (thanks to TB)

EM-Algorithm_new.R (thanks to TB)

```
....
Read data
initialize lots of variables


Prost <- function (subjN, gameN, locTheta, k0) {
      # a function computing the likelihood of parameters locTheta
      # for subject subjN in round gameN, given component k0
      # data dependent, skipped here

    prost <- XXX
    return (prost)
}
```

E-Step

E-Step

```
# predM[k,o] contains prediction of model k for task o
# gamma is the noise parameter

sumokrhoL <- rep (0, AnzSubj)
for (s in 1:AnzSubj){
    for (k in 1:AnzStrats) {

        for (o in 1:4)
            coopProbM[k,o] <- min(1-gamma, max(gamma, predM[k,o]))
        L[s,k]<-1

        for (g in 7:13)
            L[s,k] <- L[s,k] * Prost (subjN=s, gameN=g, locTheta=theta

        rhoL[s,k] <- rho[k]*L[s,k]
        sumokrhoL[s] <- sumokrhoL[s]+rhoL[s,k] # sums up over k
    }

    for (k in 1:AnzStrats) tau[s,k] <- rhoL[s,k]/sumokrhoL[s]
}
```

M-Step

M-Step

```
rho_plus<-rep(0,AnzStrats)
for (k in 1:AnzStrats){
    sumtau[1,k]<-sum(tau[,k] ) # adding up columns (over subjects)
    rho_plus[k]<-(1/AnzSubj)*sumtau[1,k] # updated rho
}

calcLLg<-function(par2, SubjV){

    penFac <- 1 + sum(pmax(par2-1, 0))
    par2 <- pmin (par2, 1)

    LL <- rep(0,AnzStrats)
    Lopt <- array (1, c(max(SubjV), AnzStrats))
    for (s in SubjV) {
        for (k2 in 1:AnzStrats) {
            for (g in 7:13)
                Lopt[s,k2] <- Lopt[s,k2] * Prost (subjN=s, gameN=g, l
            LL[k2]<-LL[k2]+tau[s,k2]*log(Lopt[s,k2])
        }
    } # end of s loop

    return (-sum (LL) * penFac)
}
opt2 <- newuoa (par=theta, fn=calcLLg, control=list(npt=2*length(theta
theta_plus <- opt2$par
```

Markov switching models

A (quick) Introduction to Structural Modeling

Markov switching models

Markov switching models

Sometimes exogenous events may affect your mental regime, e.g. your utility function or the way you think strategically (your strategic beliefs).

Markov switching models

Sometimes exogenous events may affect your mental regime, e.g. your utility function or the way you think strategically (your strategic beliefs).

Regime switches may also be triggered in economic experiments.

- Observing opponents' behavior may raise one's level of reasoning
- Observing opponents' generosity may affect one's reference point/asserted social norm

Markov switching models

Sometimes exogenous events may affect your mental regime, e.g. your utility function or the way you think strategically (your strategic beliefs).

Regime switches may also be triggered in economic experiments.

- Observing opponents' behavior may raise one's level of reasoning
- Observing opponents' generosity may affect one's reference point/asserted social norm

One way to econometrically model/control for regime switching is by Markov switching models (for discussion, see (Frühwirth-Schnatter, 2006))

Markov switching models

Sometimes exogenous events may affect your mental regime, e.g. your utility function or the way you think strategically (your strategic beliefs).

Regime switches may also be triggered in economic experiments.

- Observing opponents' behavior may raise one's level of reasoning
- Observing opponents' generosity may affect one's reference point/asserted social norm

One way to econometrically model/control for regime switching is by Markov switching models (for discussion, see (Frühwirth-Schnatter, 2006))

Rarely used in experimental analyses so far

Definition

Definition

Notation

- Components $k \in K$ with initial weights/prior probabilities ρ_k
- Periods $t = 1, \ldots, T$, state at t is an unobserved random variable K_t
- Transition probality $\sigma_{kk'} = \Pr(K_t = k' | K_{t-1} = k)$, independent of t
- Parameters θ_{kt} for component k in period t (possibly time dependent)
- State-paths are $\kappa \in K^T$ and have probability $\Pr(\kappa)$, conditional on ρ, σ
- Subjects s with observations o_{st} in period t
- $\Pr(o_{st} | \theta, k)$ is probability of o_{st} conditional on inclusion in $k \in K$ in period t

Definition

Notation

- Components $k \in K$ with initial weights/prior probabilities ρ_k
- Periods $t = 1, \ldots, T$, state at t is an unobserved random variable K_t
- Transition probality $\sigma_{kk'} = \Pr(K_t = k' | K_{t-1} = k)$, independent of t
- Parameters θ_{kt} for component k in period t (possibly time dependent)
- State-paths are $\kappa \in K^T$ and have probability $\Pr(\kappa)$, conditional on ρ, σ
- Subjects s with observations o_{st} in period t
- $\Pr(o_{st}|\theta, k)$ is probability of o_{st} conditional on inclusion in $k \in K$ in period t

Likelihood function

$$LL(\rho, \sigma, \theta | O) = \sum_{s \in S} \ln \sum_{\kappa \in K^T} \Pr(\kappa) \cdot \prod_{t \leq T} \Pr(o_{st} | \theta, \kappa(t), t).$$

Definition

Notation

- Components $k \in K$ with initial weights/prior probabilities ρ_k
- Periods $t = 1, \ldots, T$, state at t is an unobserved random variable K_t
- Transition probality $\sigma_{kk'} = \Pr(K_t = k' | K_{t-1} = k)$, independent of t
- Parameters θ_{kt} for component k in period t (possibly time dependent)
- State-paths are $\kappa \in K^T$ and have probability $\Pr(\kappa)$, conditional on ρ, σ
- Subjects s with observations o_{st} in period t
- $\Pr(o_{st}|\theta, k)$ is probability of o_{st} conditional on inclusion in $k \in K$ in period t

Likelihood function

$$LL(\rho, \sigma, \theta | O) = \sum_{s \in S} \ln \sum_{\kappa \in K^T} \Pr(\kappa) \cdot \prod_{t \leq T} \Pr(o_{st} | \theta, \kappa(t), t).$$

Estimation The naive approach, to enumerate and explicitly consider all paths, usually takes too much time. This can be accelerated by Monte Carlo simulation, but the following EM algorithm is much quicker.

Backward-Forward algorithm to compute posteriors

Given (ρ, σ, θ), compute the posterior probability of component inclusion for all subjects s and all periods t as follows

Probability of seeing the partial sequence $o_{s1}, \ldots, o_{st}$ ending up in state k at t,

$$\alpha_{sk}(t) = \Pr(O_{s1} = o_{s1}, \ldots, O_{st} = o_{st}, K_t = k),$$

is efficiently defined as

$$\alpha_{sk}(1) = \rho_k \Pr(o_{s1}|\theta, k) \quad \alpha_{sk'}(t+1) = \left[\sum_{k \in K} \alpha_{sk}(t)\sigma_{kk'}\right] \Pr(o_{s,t+1}|\theta, k').$$

Probability of the ending partial sequence $o_{t+1}, \ldots, o_T$ cond. on starting in k at t,

$$\beta_{sk}(t) = \Pr(O_{s,t+1} = o_{s,t+1}, \ldots, O_{sT} = o_{sT}, K_t = k),$$

is efficiently defined as

$$\beta_{sk}(T) = 1 \qquad \beta_{sk}(t) = \sum_{k' \in K} \sigma_{kk'} \Pr(o_{s,t+1}|\theta, k')\beta_{sk'}(t+1).$$

The probability of observing the sequence $(o_{s1}, \ldots, o_{sT})$ unconditional of states is

$$\Pr(o_s) = \sum_{k \in K} \alpha_{sk}(T) = \sum_{k \in K} \beta_{sk}(1) \rho_k \Pr(o_{s1} | \theta, k).$$

The posterior probability of being in state k at time t for observations o_s, $\gamma_{sk}(t) = \Pr(K_t = k | o_s)$, is

$$\gamma_{sk}(t) = \frac{\Pr(O, K_t = k)}{\Pr(O)} = \frac{\Pr(O, K_t = k)}{\sum_{k' \in K} \Pr(O, K_t = k')} = \frac{\alpha_{sk}(t)\beta_{sk}(t)}{\sum_{k' \in K} \alpha_{sk'}(t)\beta_{sk'}(t)}.$$

The posterior probability of s being in state k at t and in k' at $t+1$ is

$$\begin{aligned}
\zeta_{skk'}(t) &= \Pr(K_t = k, K_{t+1} = k' | o_s) \\
&= \frac{\Pr(K_t = k | o_s) \Pr(o_{t+1,\ldots,T}, K_{t+1} = k' | K_t = k)}{\Pr(o_{t+1,\ldots,T} | K_t = k)} \\
&= \frac{\gamma_{sk}(t) \sigma_{kk'} \Pr(o_{s,t+1} | \theta, k') \beta_{sk'}(t+1)}{\beta_{sk}(t)}
\end{aligned}$$

The EM algorithm

Expectation Compute the posteriors using the Backward-Forward algorithm

Maximization For all $k \in K$ and $t \leq T$,

$$\theta_{kt}^{+1} \in_{\theta'} \sum_{s \in S} \gamma_{sk}(t) \cdot \ln \Pr(o_{s,t} | \theta', k)$$

as well as

$$\rho_k^{+1} = \frac{1}{n} \cdot \sum_{s \in S} \gamma_{sk}(1) \qquad \sigma_{kk'}^{+1} = \frac{\sum_{s \in S} \sum_{t < T} \zeta_{skk'}(t)}{\sum_{s \in S} \sum_{t < T} \gamma_{sk}(t)}.$$

Stop if distance between (ρ, θ, σ) and $(\rho^{+1}, \theta^{+1}, \sigma^{+1})$ is low

Extensions

A (quick) Introduction to Structural Modeling

Possible extensions

Besides running and playing with the various R-scripts, here are a few ideas for possible extensions that may be implemented to practice.

Possible extensions

Besides running and playing with the various R-scripts, here are a few ideas for possible extensions that may be implemented to practice.

1. Extending the estimation of λ (`est-lambda.R`), simulate logit choices from $u(a) = 5 - |\chi - a|$ and estimate (λ, χ) by ML.
2. Estimate standard errors for λ (`est-lambda.R`) by bootstrapping. Compute the p-value of $H_0 : \lambda = 1$
3. Generate choices and estimate the mixed-logit model with two correlated effects.
4. Estimate the finite mixture of normals by `newuoa()` (package minqa).
5. Estimate the finite mixture of normals using the EM algorithm.

References

A (quick) Introduction to Structural Modeling

References I

Andreoni, J. and Miller, J. H. (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica*, 70(2):737–753.

Bilmes, J. A. (1998). A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. mimeo.

Blanco, M., Engelmann, D., and Normann, H.-T. (2011). A within-subject analysis of other-regarding preferences. *Games and Economic Behavior*, 72(2):321–338.

Bosch-Domènech, A., Montalvo, J. A., Nagel, R., and Satorra, A. (2002). One, two, (three), infinity, ...: Newspaper and lab beauty-contest experiments. *American Economic Review*, 92(5):1687–1701.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer, New York.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.

Nevo, A. and Whinston, M. D. (2010). Taking the dogma out of econometrics: Structural modeling and credible inference. *Journal of Economic Perspectives*, 24(2):69–82.

References II

Reiss, P. C. and Wolak, F. A. (2007). Structural econometric modeling: Rationales and examples from industrial organization. In Heckman, J. J. and Leamer, E. E., editors, *Handbook of Econometrics*, volume 6A of *Handbook of Econometrics*, pages 4277–4415. Elsevier.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Train, K. E. (2003). *Discrete Choice Methods with Simulation*. Cambridge University Press, Cambridge, 1 edition.

Wilcox, N. T. (2008). Stochastic models for binary discrete choice under risk: A critical primer and review. *Journal of Economic Surveys*, 22(5):808–846.