

ExpEcon Methods: MLE, Finite Mixture Models, & Model Selection

ECON 8877

P.J. Healy

First version thanks to Hyoeun Park

Updated 2026-01-28

Contents

1. Model estimation via MLE: how to code it Finite mixture models
2. Model selection: Cross-validation vs. BIC vs. AIC

Maximum Likelihood Estimation

Likelihood function

- y : random variable, θ : set of parameters
- $f(\mathbf{y}|\theta)$: pdf, θ identifies possible DGPs (true models)
- The joint density of n i.i.d. observations from this process

$$f(y_1 \dots y_n | \theta) = \prod_{i=1}^n f(y_i | \theta)$$

Maximum Likelihood Estimation

Likelihood function

- y : random variable, θ : set of parameters
- $f(\mathbf{y}|\theta)$: pdf, θ identifies possible DGPs (true models)
- The joint density of n i.i.d. observations from this process

$$f(y_1 \dots y_n | \theta) = \prod_{i=1}^n f(y_i | \theta) = L(\theta | \mathbf{y})$$

- $L(\theta | \mathbf{y})$: function of the unknown parameter vector, θ ,
given observed data \mathbf{y}

Maximum Likelihood Estimation

Example: Behavioral game theory models

- 2-player, 3×3 game
- S_i set of strategies, $s_i \in S_i$
- $\sigma_i(s_i)$: i 's probability of playing s_i in mixed strategy $\sigma_i(\cdot)$
- $u_i(\sigma_1, \sigma_2) = \sum_{(s_1, s_2)} \sigma_1(s_1) \sigma_2(s_2) u_i(s_1, s_2)$
- Some models use deterministic best response:

$$BR_i(\sigma_j) = \arg \max_{s_i} u_i(s_i, \sigma_j)$$

(assume unique for simplicity here)

- Others assume noisy behavior, like logistic response:

$$LR_i(\sigma_j | \lambda_i)(s_i) = \frac{\exp(\lambda u_i(s_i, \sigma_j))}{\sum_{s'_i \in S_i} \exp(\lambda u_i(s'_i, \sigma_j))}$$

Maximum Likelihood Estimation

Model 1: Level- k with logistic trembles

- Observed data: a played strategy
- 3 parameters:
 k (hierarchy level), ϵ (prob. tremble), λ (precision parameter)
- Base model (k): $\sigma_i^{LK}(\cdot | k = 0) = U[S_i]$ is uniform, then $\forall k > 0$

$$\sigma_i^{LK}(s_i | k) = 1 \text{ iff } s_i = BR_i \left(\sigma_j^{LK}(\cdot | k-1) \right)$$

- Problem: deterministic model. Zero likelihood possible.
- Solution (ϵ, λ): modify the model with logistic trembles:

$$f^{LK}(s_i | \epsilon, \lambda, k) = (1-\epsilon) \mathbb{1}_{\{s_i = BR_i(\sigma_j^{LK}(\cdot | k-1))\}} + \epsilon \cdot LR_i(\sigma_j^{LK}(\cdot | (k-1)) | \lambda)(s_i)$$

Maximum Likelihood Estimation

Example

QRE

- One parameter: λ
- Model: defined by fixed point. For each i ,

$$\sigma_i^{QRE}(s_i|\lambda) = LR_i(\sigma_j^{QRE}(\cdot|\lambda)|\lambda, S_i)(s_i)$$

$$f^{QRE}(s_i|\lambda) = \sigma_i^{QRE}(s_i|\lambda)$$

Data: s_i^g for games $g \in \mathbf{G} = \{1, 2, 3 \dots, G\}$.

$$f^{LK}(\mathbf{s}|\epsilon, \lambda, k) = \prod_{g \in \mathbf{G}} f^{LK}(s_i^g|\epsilon, \lambda, k) = L^{LK}(\epsilon, \lambda, k|\mathbf{s})$$

and

$$f^{QRE}(\mathbf{s}|\lambda) = \prod_{g \in \mathbf{G}} f^{QRE}(s_i^g|\lambda) = L^{QRE}(\lambda|\mathbf{s})$$

Maximum Likelihood Estimation

Maximum Likelihood Estimation of a single model (eg, noisy LK)

1. Let's allow different parameters for each subject. So, fix i
2. Set a 3D grid of (ϵ, λ, k) values
3. For each point on that grid calculate

$$\begin{aligned} f^{LK}(\mathbf{s}|\epsilon, \lambda, k) &= \prod_{g \in \mathbf{G}} f^{LK}(s_i^g|\epsilon, \lambda, k) \\ &= \sum_{g \in \mathbf{G}} \log(f^{LK}(s_i^g|\epsilon, \lambda, k)) \end{aligned}$$

4. MLE estimate: $(\epsilon^*, \lambda^*, k^*)$ with highest value

Model Selection:

- Which model has higher likelihood at its MLE parameter?
- Problem: unfair advantage having more parameters
- Solutions: penalize MLE by subtracting param. penalty
 - Akaike Info Criterion (AIC) vs. Bayesian Info Criterion (BIC)

Maximum Likelihood Estimation

Example

Game 1	T	M	B	Game 2	T	M	B	Game 3	T	M	B
T	25	30	100	T	30	50	100	T	10	100	40
M	40	45	65	M	40	45	10	M	0	70	50
B	31	0	40	B	35	60	0	B	20	50	60

- Suppose that a subject plays M, T, B
- Let $\lambda \in \{0.01, 0.05, 1\}$
- $\log(L^{QRE}(0.01|(M, T, B))) = \log(0.3628) + \log(0.3914) + \log(0.3311)$
- $\log(L^{QRE}(0.05|(M, T, B))) = \log(0.5391) + \log(0.5355) + \log(0.3518)$
- $\log(L^{QRE}(1|(M, T, B))) = \log(1) + \log(0.7024) + \log(0.9999)$
- Thus, in this example $\hat{\lambda} = 1$

Finite Mixture Model

- So far, only one model for one likelihood function

Finite Mixture Model

- So far, only one model for one likelihood function
- Is it a valid approach?

Finite Mixture Model

- So far, only one model for one likelihood function
- Is it a valid approach?
 - For example, Georganas et al., (2015) show that a cognitive hierarchy is not persistent across classes of games
 - Suggests that estimating with *one* hierarchy cannot be valid.

Finite Mixture Model

- So far, only one model for one likelihood function
- Is it a valid approach?
 - For example, Georganas et al., (2015) show that a cognitive hierarchy is not persistent across classes of games
 - Suggests that estimating with *one* hierarchy cannot be valid.
 - Another example at the population level, people might have different risk preferences, etc.,

Finite Mixture Model

"Mix" models

- $m = 1, 2, \dots, M$ denotes model
- $f(\mathbf{y}|\psi) = \sum_{m=1}^M \pi_m f_m(\mathbf{y}|\theta_m)$, where $\psi = (\{\theta_m\}_{m=1}^M, \pi_1, \pi_2, \dots, \pi_M)$
- Usually, $f_m(\mathbf{y}, \theta_m)$ (called component density) are taken to belong to the same parametric family.
 - There are special cases where component densities are taken to be different (nonstandard mixture)
- posterior probability that data is drawn from model m , given observed data \mathbf{y} is $\pi_m \cdot \frac{f_m(\mathbf{y}|\theta_m)}{f(\mathbf{y}|\psi)}$
- A parametric family of densities is primitive. Each component has distinct values
- Using MLE to fitting mixture distributions π (Most commonly used way)

Model Selection

Going back to Level- k and QRE Example

- Suppose that an experimenter wants to compare which model better explains data
- They can horse-race models
- Using MLE?
 - For level- k , find the ML estimates $(\hat{\epsilon}, \hat{\lambda}, \hat{k})$
plug in those values to the level- k model's likelihood function
 - For QRE, find the ML estimate $\hat{\lambda}$
plug in $\hat{\lambda}$ to the QRE's likelihood function
 - Compare the likelihood values of two models and pick the model that gives the higher value

Model Selection

What is the problem with ML approach?

- level- k has three parameters, while QRE has only one parameter
- level- k has more “flexibility”
- Consider a weird model with ∞ numbers of parameters
 - ∞ flexibility
 - Can explain any behavior in the data
 - Then, this model “wins” just because it has more flexibility, not because it is true DGP.
- Need for fixing the problem of over-fitting due to large # of parameters

How to penalize over-fitting due to large numbers of parameters?

- AIC (Akaike information criterion)
- BIC (Bayesian information criterion)
- Cross-Validation

Model Selection: AIC, BIC

- $AIC = 2k \ln(n) - 2 \ln(\hat{L})$
- $BIC = k \ln(n) - 2 \ln(\hat{L})$

where

- \hat{L} = the maximized value of the likelihood function of the model
 - From observed data, get ML estimates, and plug into the ML function
- n = number of observed data point
 - In our example, the number of games subjects played
- k = number of parameters
 - In level- k model, $k=3$
In QRE, $k = 1$

Model Selection: AIC, BIC

- $AIC = 2k \ln(n) - 2 \ln(\hat{L})$
- $BIC = k \ln(n) - 2 \ln(\hat{L})$
- The preferred model is the one with the minimum AIC/BIC value
- The second term gives benefits to the model with goodness-of-fit
- The first term gives a penalty to the number of parameters
- Those can be only used for linear-models
- Can be used only when $n \gg k$

Model Selection: Cross-Validation

- k -fold cross-validations
- Divide data into k sub-samples
 - For example, 12 data points, 4 sub-samples that include 3 data points each
- $k - 1$ sub-samples = training data
 - Fit the data to a model (MLE, MSE ...)
- one sub-sample = testing data
 - Using the fitted parameters from training data, test the model i.e., Plug in the estimated parameter to the goodness-of-fit function used for training (MLE, MSE, ...)
- Repeat this for K times
- Extreme case of K fold cross-validation is leave-one-out cross-validation that $K = n$, where n = number of data points

Model Selection: Cross-Validation

How does Cross-Validation penalize the number of parameters?

Consider the following example..

- Suppose that subjects played four games
- For three games, a subject's choices coincide with the level- k level-1's predictions
- Then $\hat{\epsilon} = 0$ and log-likelihood function value is 0.
- Suppose that the subject did not play level-1's predicted strategy.
- Then for the testing data (fourth game), $\hat{\epsilon} = 0$ results in $-\infty$
- For QRE, less likely to over-fit (since it has only one parameter).
Less likely to have $-\infty$ for testing data

Model Selection: Cross-Validation

- Cross-Validation penalizes the number of parameters internally.
- Over-fitting due to a higher number of parameters penalizes deviation from the prediction a lot in the testing data
- No restriction for models being tested; does not have to be linear

Model Selection

- SO, is Cross-Validation a perfect solution for model selection?

Model Selection

- SO, is Cross-Validation a perfect solution for model selection?
- NO!

Model Selection

- SO, is Cross-Validation a perfect solution for model selection?
- NO!
- The penalization can be "too" severe

Model Selection

- SO, is Cross-Validation a perfect solution for model selection?
- NO!
- The penalization can be "too" severe
- See Healy & Park (2023) for suggestions :)