Maximum Simulated Likelihood

A. Colin Cameron
Univ. of Calif. - Davis

May 2021

1. Introduction

- Maximum simulated likelihood (MSL)
  - for models where the density involves an integral with no closed form solution
  - so replace the integral with a Monte Carlo integral.
- Leading applications
  - random parameter models
    - ⋆ random parameters multinomial logit
  - random utility models
    - ⋆ multinomial probit.

Outline

2. Maximum Simulated Likelihood

- Problem: MLE (with independent data over i) maximizes

$$\ln L(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln f(y_i|\mathbf{x}_i, \boldsymbol{\theta}).$$

  - but $f(y_i|\mathbf{x}_i, \boldsymbol{\theta})$ does not have a closed form solution.
- Example: Random effects where $g(y_i|\mathbf{x}_i, \boldsymbol{\theta}_1, \alpha)$ has a closed form solution but we want to integrate out the random effect α

$$f(y_i|\mathbf{x}_i, \boldsymbol{\theta}) = \int g(y_i|\mathbf{x}_i, \boldsymbol{\theta}_1, \alpha) h(\alpha|\boldsymbol{\theta}_2) d\alpha = ?$$

- Solutions include
  - numerical integration using Gaussian quadrature (see appendix)
    - a good method if only a one-dimensional integral
  - Bayesian MCMC with an uninformative prior
  - maximum simulated likelihood (MSL).

Monte Carlo integration

- Monte Carlo integration is the basis for MSL.
- Suppose X is distributed with density $g(x)$ on (a, b)
- Then

$$\mathrm{E}[h(X)] = \int_a^b h(x)g(x)dx.$$

- If not tractable we could approximate by making draws $x^1, ..., x^s$ from $g(x)$, and average the corresponding values $h(x^1), ..., h(x^s)$, so

$$\widehat{\mathrm{E}}[h(X)] = \tfrac{1}{S}\textstyle\sum_{s=1}^{S} h(x^s).$$

- Provided $\mathrm{E}[h(X)]$ exists, $\widehat{\mathrm{E}}[h(X)] \xrightarrow{p} \mathrm{E}[h(X)]$ by a LLN.

Monte Carlo integration (continued)

- Problems:
  - may require many draws
  - "works" even if $\mathrm{E}[h(X)]$ does not exist!
- Variation: Importance sampling instead transforms so that instead of draws from $g(x)$ we make draws from $p(x)$

$$\begin{aligned} E[h(X)] &= \int h(x)g(x)dx \\ &= \int \left(\frac{h(x)g(x)}{p(x)} \right) p(x)dx \\ &= \int w(x)p(x)dx \end{aligned}$$

where it is easier or better to make draws from $p(x)$.

Maximum Simulated Likelihood

- MLE (with independent data over i) maximizes

$$\ln L(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln f(y_i|\mathbf{x}_i, \boldsymbol{\theta}).$$

- Maximum simulated likelihood (MSL) estimator maximizes

$$\ln \widehat{L}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln \widehat{f}(y_i|\mathbf{x}_i, \boldsymbol{\theta})$$

  - $\widehat{f}(y_i|\mathbf{x}_i, \boldsymbol{\theta})$ is a simulated approximation to $f(\cdot)$ based on S draws
  - the usual gradient methods are used so recompute $\widehat{f}(\cdot)$ at each iteration.

- Example using a frequency simulator

$$f(y_i|\mathbf{x}_i, \boldsymbol{\theta}) = \int g(y_i|\mathbf{x}_i, \boldsymbol{\theta}_1, \alpha) h(\alpha|\boldsymbol{\theta}_2) d\alpha$$

$$\widehat{f}(y_i|\mathbf{x}_i, \boldsymbol{\theta}) = \tfrac{1}{S} \sum_{s=1}^{S} g(y_i|\mathbf{x}_i, \boldsymbol{\theta}, \alpha^{(s)}); \ \alpha^{(s)} \text{ are draws from } h(\alpha|\boldsymbol{\theta}_2).$$

Leading Examples

- **1.** Random parameters multinomial logit ("mixed" logit)
  - Regular multinomial logit except coefficients of alternative-varying regressors are random (joint normally distributed)
    - then the restriction of independence of irrelevant alternatives is relaxed
    - Stata `cmmixlogit` command.
- **2.** Multinomial probit model
  - allow underlying errors for utility of each alternative to be correlated (and normal)
    - integral has dimension the number of alternatives less one
    - Stata `cmmprobit` command.
- **3.** Mixed models (random coefficient models)
  - coefficients of regressors are random and joint normally distributed
    - Stata `meglm` command and more specific commands such as `melogit`.

4. MSL details

- MSLE is **consistent** with the usual MLE asymptotic distribution if
  - $\widehat{f}(\cdot)$ is an unbiased simulator and satisfies other conditions given below
  - $S \to \infty$, $N \to \infty$ and $\sqrt{N}/S \to 0$ where S is the number of simulations.
  - note that many draws S (to compute $\widehat{f}(\cdot)$) are required
  - better to use robust standard errors (sandwich matrix).
- Assumed properties of the simulator:
  - $\widehat{f}(\cdot)$ is an **unbiased simulator** with: $\mathrm{E}[\widehat{f}(y_i|\mathbf{x}_i,\boldsymbol{\theta})] = f(y_i|\mathbf{x}_i,\boldsymbol{\theta})$
  - $\widehat{f}(\cdot)$ is differentiable in $\boldsymbol{\theta}$ (or **smooth simulator**) so gradient methods can be used
  - the underlying draws to compute $\widehat{f}(\cdot)$ are unchanged so no "chatter".
- MSL needs $S \to \infty$ because simulator is nonetheless **biased** for $\ln f(\cdot)$

$$\mathrm{E}[\widehat{f}(\cdot)] = f(\cdot) \quad \nRightarrow \quad \mathrm{E}[\ln \widehat{f}(\cdot)] \neq \ln f(\cdot).$$

MSL further details

- When draws are used to compute $\widehat{f}(\cdot)$ the same underlying draws need to be used at each iteration to avoid "chatter"
  - for multivariate normal draws retain original i.i.d. standard normal draws and use Cholesky decomposition.
- More efficient to use antithetic draws (negatively correlated pairs), rather than independent draws.
- Generate uniform numbers using Halton or Hammersley sequences.
- The (obvious) **frequency simulator** averages
  - e.g. earlier example with $\widehat{f}(\cdot) = \frac{1}{S}\sum_{s=1}^{S} g(y_i|\mathbf{x}_i, \boldsymbol{\theta}, \alpha^{(s)})$.
- But better simulators exist in specific circumstances
  - for multinomial probit use the Geweke-Hajivassiliou-Keane (GHK) simulator.

Method of Simulated Moments (MSM)

- Rather than ML, use moment conditions that allow an unbiased simulator.
- Suppose $\widehat{\boldsymbol{\theta}}$ is a method of moments estimator that solves

$$\sum_{i=1}^{N} \mathbf{m}(y_i|\mathbf{x}_i, \boldsymbol{\theta}) = \mathbf{0}.$$

- Assume there exists an unbiased simulator such that $\mathsf{E}[\widehat{\mathbf{m}}(y_i|\mathbf{x}_i, \boldsymbol{\theta})] = \mathbf{m}(y_i|\mathbf{x}_i, \boldsymbol{\theta})$.
- Then the MSM solves

$$\sum_{i=1}^{N} \widehat{\mathbf{m}}(y_i|\mathbf{x}_i, \boldsymbol{\theta}) = \mathbf{0}.$$

- Computational advantage
  - consistent for $\boldsymbol{\theta}$ even for small number of draws S.
- Disadvantages
  - efficiency loss for low S
    - when $\widehat{\mathbf{m}}(\cdot)$ is the frequency simulator $\mathsf{V}[\widehat{\boldsymbol{\theta}}_{\mathsf{MSM}}] = (1 + \frac{1}{S})\mathsf{V}[\widehat{\boldsymbol{\theta}}_{\mathsf{MSL}}]$.
  - and efficiency loss because not the MLE.

5. Example: Random Parameters Logit (fishing mode choice)

- Explain the multinomial variable y with outcome one of
  - $y = 1$ if fish from beach
  - $y = 2$ if fish from pier
  - $y = 3$ if fish from private boat
  - [$y = 4$ if fish from charter boat is dropped below]
- Regressors are
  - price: varies by alternative and individual
  - catch rate: varies by alternative and individual
  - income: varies by individual but not alternative

Alternative-specific Conditional Logit

- Data on individual i and alternative j for m alternatives.
- Two types of regressors
  - $\mathbf{x}_{ij}$ are alternative-varying regressors (price, catch rate)
  - $\mathbf{z}_i$ are alternative-invariant or case-specific (income)
- Specify

$$p_{ij}(\boldsymbol{\beta},\boldsymbol{\gamma}) = \Pr[y_i = j] = \frac{e^{\mathbf{x}_{ij}'\boldsymbol{\beta}+\mathbf{z}_i'\boldsymbol{\gamma}_j}}{\sum_{k=1}^{m} e^{\mathbf{x}_{ik}'\boldsymbol{\beta}+\mathbf{z}_i'\boldsymbol{\gamma}_k}}, \qquad j = 1, ..., m.$$

  - parameters $\boldsymbol{\gamma}_j$ can vary across alternatives and normalize $\boldsymbol{\gamma}_1 = \mathbf{0}$.
- MLE maximizes

$$\ln L(\boldsymbol{\beta},\boldsymbol{\gamma}) = \sum_{i=1}^{N} \sum_{j=1}^{m} y_{ij} \ln p_{ij}(\boldsymbol{\beta},\boldsymbol{\gamma}).$$

Random Parameters Logit

- Allow coefficient of alternative-varying regressors to differ across individuals.
  - $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{v}_i$ where $\mathbf{v}_i \sim N[\mathbf{0}, \Sigma_{\boldsymbol{\beta}}]$
  - $\mathbf{x}_{ij}'\boldsymbol{\beta}_i = \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{x}_{ij}'\mathbf{v}_i$ where $\mathbf{v}_i \sim N[\mathbf{0}, \Sigma_{\boldsymbol{\beta}}]$
- Given knowledge of $\mathbf{v}_i$

$$p_{ij}(\boldsymbol{\beta}, \boldsymbol{\gamma}|\mathbf{v}_i) = \frac{e^{\mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma}_j + \mathbf{v}_i'\boldsymbol{\beta}}}{\sum_{k=1}^m e^{\mathbf{x}_{ik}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma}_k + \mathbf{v}_i'\boldsymbol{\beta}}}, \qquad j = 1, ..., m.$$

- But we need to integrate out $\mathbf{v}_i$

$$p_{ij}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \Sigma_{\boldsymbol{\beta}}) = \int \frac{e^{\mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma}_j + \mathbf{v}_i'\boldsymbol{\beta}}}{\sum_{k=1}^m e^{\mathbf{x}_{ik}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma}_k + \mathbf{v}_i'\boldsymbol{\beta}}} d\phi(\mathbf{v}_i|\Sigma_{\boldsymbol{\beta}}) d\mathbf{v}_i, \qquad j = 1, ..., m.$$

Application

- Here just β_{price} varies across individuals
  - Hammersley sequence is used with 613 integration points ("draws").

```
. * Alternative-specific mixed logit or random parameters logit estimation
. cmset id fishmode

     Case ID variable: id
Alternatives variable: fishmode

. cmmixlogit d q, casevars(income) random(p) basealternative(pier) ///
>     vce(robust) nolog

Mixed logit choice model                               Number of obs     =      2,190
Case ID variable: id                                   Number of cases   =        730

Alternatives variable: fishmode                        Alts per case: min =          3
                                                                      avg =        3.0
                                                                      max =          3
Integration sequence:          Hammersley
Integration points:                   613                 Wald chi2(4)    =      28.40
Log simulated-pseudolikelihood = -433.92078               Prob > chi2     =     0.0000
```

- Estimates - utility is decreasing in price and increasing in catch rate
  - standard deviation of $\beta_{\text{price},i}$ (0.059) is large relative to mean (-0.107)

(Std. err. adjusted for clustering on id)

d	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
fishmode						
q	.8633073	.8872554	0.97	0.331	-.8756813	2.602296
p	-.107416	.0287078	-3.74	0.000	-.1636823	-.0511497
/Normal						
sd(p)	.0595192	.0187898			.0320582	.1105035
beach						
income	.1203331	.0519823	2.31	0.021	.0184497	.2222165
_cons	-.7802862	.2304865	-3.39	0.001	-1.232031	-.328541
pier	(base alternative)					
private						
income	.1733836	.0773131	2.24	0.025	.0218526	.3249146
_cons	-.2199922	.318053	-0.69	0.489	-.8433647	.4033802

AME of Pr(choose mode j) for change in price of mode k

```
. * Average marginal effects with respect to price
. margins, dydx(p)

Average marginal effects                               Number of obs = 2,190
Model VCE: Robust

Expression: Pr(fishmode), predict()
dy/dx wrt:  p
```

	dy/dx	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
p						
_outcome#fishmode						
beach#beach	-.0122611	.0032902	-3.73	0.000	-.0187099	-.0058123
beach#pier	.0097067	.0028752	3.38	0.001	.0040714	.0153421
beach#private	.0025544	.0004443	5.75	0.000	.0016835	.0034252
pier#beach	.0097067	.0028752	3.38	0.001	.0040714	.0153421
pier#pier	-.0131526	.0033724	-3.90	0.000	-.0197624	-.0065428
pier#private	.0034458	.0005343	6.45	0.000	.0023986	.0044931
private#beach	.0025544	.0004443	5.75	0.000	.0016835	.0034252
private#pier	.0034458	.0005343	6.45	0.000	.0023986	.0044931
private#private	-.0060002	.0009187	-6.53	0.000	-.0078007	-.0041996

6. Appendix: Numerical Integration

- Numerical method for computing integral

$$I = \int_a^b f(x)\, dx$$

- Mid-point rule calculates the Riemann sum at n midpoints

$$\widehat{I}_M = \sum_{j=1}^{n} \frac{b-a}{n} f(\bar{x}_j)$$

- Better variants are trapezoidal rule and Simpson's rule.
- But big problem if range of integration is unbounded
  - $a = -\infty$ or $b = \infty$!
  - so use Gaussian quadrature.
- **Gaussian quadrature is the basis for mixed model estimation in Stata.**

Gaussian quadrature (continued)

- Gaussian quadrature re-expresses the integral as

$$I = \int_a^b f(x)\, dx = \int_c^d w(x) r(x) dx,$$

  - where $w(x)$ is one of the following functions depending on range of x (unbounded from above and below; or unbounded on one side only; or bounded on both sides)
    - ★ $(a, b) = (-\infty, \infty)$: Gauss-Hermite: $w(x) = e^{-x^2}$ & $(c, d) = (-\infty, \infty)$.
    - ★ $[a, b) = [a, \infty)$: Gauss-Laguerre: $w(x) = e^{-x}$ and $(c, d) = (0, \infty)$.
    - ★ $[a, b] = [a, b]$: Gauss-Legendre: $w(x) = 1$ and $(c, d) = [-1, 1]$.
  - In simplest case $r(x) = f(x)/w(x)$, but may need transformation of x.
- Gaussian quadrature approximates the integral by the weighted sum

$$\widehat{I}_G = \sum_{j=1}^{m} w_j r(x_j),$$

  - the researcher chooses m with often $m = 20$ enough
  - given m, the m points of evaluation x_j and associated weights w_j are given in e.g. computer code of Press et al. (1993).

Gaussian quadrature in higher dimensions

- In higher dimensions Gauss-Hermite quadrature does not always provide an adequate approximation.
- Adaptive Gauss-Hermite quadrature may provide better approximation.
- In Stata the quadrature methods for multivariate normal use a Cholesky decomposition to reduce a multidimensional problem to a series of one-dimensional Gauss-Hermite quadratures
  - see `[ME] meglm` for a detailed discussion.
- For normal integrals a faster though less accurate alternative is to use a Laplacian approximation.

7. References

- The general principles of MSL (and simulation) are covered in
  - A. Colin Cameron and Pravin K. Trivedi (2005), Microeconometrics: Methods and Applications, chapter 13, Cambridge University Press.