Chapter 1: A Brief Review of Maximum Likelihood, GMM, and Numerical Tools

Joan Llull

Microeconometrics
IDEA PhD Program

Maximum Likelihood

The Likelihood Principle

Likelihood principle: our estimate of $\boldsymbol{\theta}$ is the one that maximizes the likelihood of our sample $(\boldsymbol{y}, X) = ((y_1, \boldsymbol{x}_1')', ...(y_N, \boldsymbol{x}_N')')'$.

Likelihood is the **"probability" of observing the sample**, i.e. $\Pr[\boldsymbol{y}, X; \boldsymbol{\theta}]$ for discrete data; $f(\boldsymbol{y}, X; \boldsymbol{\theta})$ for continuous.

The **likelihood function** is $L_N^*(\boldsymbol{\theta}) \equiv f(\boldsymbol{y}, X; \boldsymbol{\theta}) = f(\boldsymbol{y}|X; \boldsymbol{\theta}) f(X; \boldsymbol{\theta})$, which for *i.i.d.* data is $f(\boldsymbol{y}, X; \boldsymbol{\theta}) = \prod_{i=1}^N f(y_i, \boldsymbol{x}_i; \boldsymbol{\theta})$.

We assume $f(X; \boldsymbol{\theta}) = f(X)$ so we can focus on $f(\boldsymbol{y}|X; \boldsymbol{\theta})$.

Hence, our object of interest is the (conditional) **log-likelihood function**:

$$\mathcal{L}_{\mathrm{N}}(\boldsymbol{\theta}) \equiv \sum_{\mathrm{i=1}}^{\mathrm{N}} \ln f(y_i | \boldsymbol{x}_i; \boldsymbol{\theta}).$$

The Maximum Likelihood Estimator (MLE)

The MLE is defined by the following optimization problem:

$$\hat{\boldsymbol{\theta}}_{ML} \equiv \arg\max_{\boldsymbol{\theta}\in\Theta} \frac{1}{N}\,\mathcal{L}_{\mathrm{N}}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}\in\Theta} \frac{1}{N}\sum_{\mathrm{i}=1}^{\mathrm{N}} \ln f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}).$$

This estimator is:

- Fully parametric
- An extremum estimator
- An m-estimator

The FOC of the problem is:

$$\frac{1}{N}\sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{\partial \ln f(y_i|\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_{ML})}{\partial \boldsymbol{\theta}} = \mathbf{0}.$$

Identification

The *true* parameter vector $\boldsymbol{\theta}_0$ is identified if there are *no observationally equivalent* parameters.

More formally, $\boldsymbol{\theta}_0$ is identified if the *Kullback-Leibler inequality* is satisfied:

$$\Pr[f(y|\boldsymbol{x};\boldsymbol{\theta}) \neq f(y|\boldsymbol{x};\boldsymbol{\theta}_0)] > 0 \quad \forall \boldsymbol{\theta} \neq \boldsymbol{\theta}_0.$$

Regularity conditions

If the following two assumptions hold,

i. The specified density $f(y|\boldsymbol{x};\boldsymbol{\theta})$ is the **data generating process** (dgp)

ii. The **support** of y does not depend on $\boldsymbol{\theta}$

then the **regularity conditions** are satisfied:

$$\mathbb{E}_f\left[\frac{\partial \ln f(y|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = \mathbf{0}$$

$$-\mathbb{E}_f\left[\frac{\partial^2 \ln f(y|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right] = \mathbb{E}_f\left[\frac{\partial \ln f(y|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\frac{\partial \ln f(y|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right].$$

The latter condition is a.k.a. **information matrix equality**.

Consistency

Using identification and the first regularity condition:

$$\mathbb{E}[\ln f(y|\boldsymbol{x};\boldsymbol{\theta})] < \mathbb{E}[\ln f(y|\boldsymbol{x};\boldsymbol{\theta}_0)] \quad \forall \boldsymbol{\theta} \neq \boldsymbol{\theta}_0.$$

By the Law of Large Numbers (LLN):

$$\frac{1}{N}\sum_{\mathrm{i=1}}^{\mathrm{N}} \ln f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}) \underset{p}{\rightarrow} \mathbb{E}[\ln f(y|\boldsymbol{x};\boldsymbol{\theta})].$$

Then, as $N \to \infty$, $\hat{\boldsymbol{\theta}}_{ML} = \arg\max \mathcal{L}_{\mathrm{N}}(\boldsymbol{\theta}) \underset{p}{\rightarrow} \arg\max \mathcal{L}_0(\boldsymbol{\theta}) = \boldsymbol{\theta}_0,$

whenever:

i. The parameter space Θ is **compact**

ii. $\mathcal{L}_{\mathrm{N}}(\boldsymbol{\theta})$ is **measurable** for all $\boldsymbol{\theta}$

Asymptotic distribution

Using a first order *Taylor Expansion* of the FOC around $\boldsymbol{\theta}_0$:

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = -\left(\frac{1}{N}\sum_{\mathrm{i=1}}^{\mathrm{N}} \frac{\partial^2 \ell_i(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right)^{-1} \frac{1}{\sqrt{N}} \sum_{\mathrm{i=1}}^{\mathrm{N}} \frac{\partial \ell_i(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}},$$

where $\ell_i(\boldsymbol{\theta}) \equiv \ln f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})$, and $\boldsymbol{\theta}^*$ is between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$.

Assuming i.i.d. observations and regularity+consistency conditions, by LLN:

$$-\left(\frac{1}{N}\sum_{\mathrm{i=1}}^{\mathrm{N}} \frac{\partial^2 \ell_i(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right)^{-1} \underset{p}{\rightarrow} -\mathbb{E}\left[\frac{\partial^2 \ell_i(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]^{-1}.$$

By CLT:

$$\frac{1}{\sqrt{N}} \sum_{\mathrm{i=1}}^{\mathrm{N}} \frac{\partial \ell_i(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \underset{d}{\rightarrow} \mathcal{N}\left(0, \mathbb{E}\left[\frac{\partial \ell_i(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \frac{\partial \ell_i(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'}\right]\right).$$

Finally, using Cramer theorem and IM equality:

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \underset{d}{\rightarrow} \mathcal{N}\left(0, \Omega_0\right), \quad \Omega_0 = -\mathbb{E}\left[\frac{\partial^2 \ell_i(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]^{-1} = \mathbb{E}\left[\frac{\partial \ell_i(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \frac{\partial \ell_i(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'}\right]^{-1}.$$

Since $\Omega_0 = IM^{-1}$, it is the **Cramer-Rao** lower bound (efficient estimator).

Generalized Method of Moments

General formulation

Let $\boldsymbol{\theta}$ be the parameter vector of interest, defined by the set of moments (or orthogonality conditions):

$$\mathbb{E}[\psi(\boldsymbol{w};\boldsymbol{\theta})] = \mathbf{0},$$

where

- $\boldsymbol{w}$ is a (vector) random variable,
- and $\psi(\cdot)$ is a vector function such that $\dim(\psi) \geq \dim(\boldsymbol{\theta})$.

Estimation

Consider a random **sample** with N observations $\{\boldsymbol{w}_i\}_{i=1}^N$.

GMM estimation is based on the **sample moment conditions**:

$$\boldsymbol{b}_N(\boldsymbol{\theta}) \equiv \frac{1}{N}\sum_{\mathrm{i=1}}^{\mathrm{N}} \psi(\boldsymbol{w}_i; \boldsymbol{\theta}).$$

The **GMM estimator** minimizes the quadratic distance of $b_N(\boldsymbol{\theta})$ from zero:

$$\hat{\boldsymbol{\theta}}_{GMM} \equiv \arg\min_{\theta\in\Theta} \boldsymbol{b}_N(\boldsymbol{\theta})' W_N \boldsymbol{b}_N(\boldsymbol{\theta}),$$

where W_N is semi-positive definite, and $\mathrm{rank}(W_N) \geq \dim(\boldsymbol{\theta})$.

If the problem is **just-identified** ($\dim(\psi) = \dim(\boldsymbol{\theta})$):

$$\boldsymbol{b}_N(\hat{\boldsymbol{\theta}}_{GMM}) = \mathbf{0}.$$

Consistency

GMM is an **extremum estimator**, so general consistency results hold, as in MLE.

Conditions and intuition are similar to MLE:

- Parameter space $\Theta \in \mathbb{R}^K$ is **compact**.
- $W_N \boldsymbol{b}_N(\boldsymbol{\theta}) \underset{p}{\rightarrow} W_0 \, \mathbb{E}[\psi(\boldsymbol{w}; \boldsymbol{\theta})]$.
- **Identification**: $W_0 \, \mathbb{E}[\psi(\boldsymbol{w}; \boldsymbol{\theta})] = 0 \Leftrightarrow \boldsymbol{\theta} = \boldsymbol{\theta_0}$.

If these conditions hold, $\hat{\boldsymbol{\theta}}_{GMM} \underset{p}{\rightarrow} \boldsymbol{\theta}_0$.

Asymptotic distribution

Following similar steps as in the MLE case, if

- $\hat{\boldsymbol{\theta}}_{GMM}$ is a **consistent estimator** of $\boldsymbol{\theta}_0$,
- $\boldsymbol{\theta}$ is in the **interior** of Θ,
- $\psi(\boldsymbol{w};\boldsymbol{\theta})$ is **once differentiable** with respect to $\boldsymbol{\theta}$,
- $D_N(\boldsymbol{\theta}) \equiv \partial \boldsymbol{b}_N(\boldsymbol{\theta})/\boldsymbol{\theta}' \underset{p}{\rightarrow} D_0(\boldsymbol{\theta})$, $D_0(\boldsymbol{\theta})$ continuous at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$,
- For $D_0 \equiv D_0(\boldsymbol{\theta}_0)$, the matrix $D_0' W_0 D_0$ is **non-singular**,
- $\sqrt{N}\boldsymbol{b}_N(\boldsymbol{\theta}_0) \underset{d}{\rightarrow} \mathcal{N}(0, V_0)$, with $V_0 = \mathbb{E}[\psi(\boldsymbol{w};\boldsymbol{\theta}_0)\psi(\boldsymbol{w};\boldsymbol{\theta}_0)']$,

then $\sqrt{N}(\hat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta}_0) \underset{d}{\rightarrow} \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_0)$, with:

$$\boldsymbol{\Omega}_0 = (D_0' W_0 D_0)^{-1} D_0' W_0 V_0 W_0 D_0 (D_0' W_0 D_0)^{-1}.$$

Optimal weighting matrix

Efficiency is achieved with any W_N that delivers $W_0 = \kappa V_0^{-1}$.

This includes $W_N = V_0^{-1}$ (**unfeasible**), but also $W_N = \hat{V}_N^{-1}$, where $\hat{V}_N$ is any consistent estimator of V_0.

Optimal GMM estimator is implemented in two steps:

1. Obtain $\hat{\boldsymbol{\theta}}_{GMM}(W_N^0)$ for an initial guess W_N^0.
2. Re-estimate using

$$\widehat{W}_{opt} \equiv \left(\sum_{\mathrm{i=1}}^{\mathrm{N}} \psi(\boldsymbol{w}_i; \hat{\boldsymbol{\theta}}_{GMM}(W_N^0)) \psi(\boldsymbol{w}_i; \hat{\boldsymbol{\theta}}_{GMM}(W_N^0))' \right)^{-1}$$

as the new weighting matrix.

Numerical Methods

Differentiation

We use the **definition of a derivative**:

$$f'(x) = \lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x)}{\epsilon} \Rightarrow f'(x) \approx \frac{f(x+h) - f(x)}{h},$$

for a small h (e.g. 10^{-6}).

More accurate (and costly) is the **two-sided** differential:

$$f'(x) \approx \frac{f(x+h) - f(x-h)}{2h}.$$

To compute a **gradient** $\nabla_f(\boldsymbol{x})$, one element is moved at a time.

Newton-Raphson optimization

Originally conceived for **finding roots**.

Approximates the function by the **tangent** line and finds the **intercept** (iterative procedure):

$$\frac{f(x_n) - 0}{x_n - x_{n+1}} = f'(x_n) \Rightarrow x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

Extension to **optimization** is natural:

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}.$$

In the **multivariate** case:

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n - \left[H_f(\boldsymbol{x}_n)\right]^{-1} \nabla_f(\boldsymbol{x}_n).$$

Integration

Numerical integration (quadrature) consists of a **weighted sum** of a finite set of **evaluations** of the integrand.

Integration weights depend on the **method** and on **precision**.

Deterministic methods: midpoint rule, trapezoidal rule, Simpson's rule, Gaussian,...

Alternative: **Monte Carlo integration**.