Method of Moments and Maximum Likelihood Estimation (MLE)

Statistics · Module 3 — Point Estimation

Node 3.2

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Point Estimation

Node:
3.2

Formal Definition

Let $X_1, \dots, X_n$ be data generated under a model $\{f_\theta : \theta \in \Theta\}$.

Method of Moments (MoM)

Choose k population moments $m_j(\theta) = \mathbb{E}_\theta[X^j]$ (or other moment-type quantities), and set them equal to the corresponding sample moments:

$$\hat{m}_j = \frac{1}{n} \sum_{i=1}^{n} X_i^j.$$

Solve the system:

$$m_j(\theta) = \hat{m}_j, \quad j = 1, \dots, k$$

for θ. The solution (if it exists) is the MoM estimator $\hat{\theta}_{\text{MoM}}$.

Maximum Likelihood Estimation (MLE)

Define the likelihood:

$$L(\theta; x_1, \dots, x_n) = \prod_{i=1}^{n} f_\theta(x_i),$$

and log-likelihood:

$$\ell(\theta) = \sum_{i=1}^{n} \log f_\theta(x_i).$$

An MLE is any maximizer:

$$\hat{\theta}_{\text{MLE}} \in \arg\max_{\theta \in \Theta} \ell(\theta).$$

Equivalently (when differentiable), solve the score equation:

$$\nabla_\theta \ell(\theta) = 0$$

and select maxima.

Intuition

MoM: "Match what the model predicts on average to what the data shows on average."
Fast, direct, often algebraic.

MLE: "Choose the parameter under which the observed data would be most plausible."
Optimization-based, often statistically efficient, and the central object of likelihood theory.

Structural Role

These are the two canonical construction principles for estimators:

- MoM provides quick, often closed-form estimators and a gateway to estimating equations.
- MLE is the backbone of classical inference (Wald/Score/LR tests, asymptotics, efficiency, Fisher information).

Understanding their contrast sets up: - consistency and asymptotic normality (Domain 7), - Fisher-information bounds (Node 3.4/3.5), - computational optimization (Domain 9).

Mathematical Development

MoM Example Pattern

If the model implies $\mathbb{E}_\theta[X] = g(\theta)$, then MoM sets:

$$g(\hat{\theta}_{\text{MoM}}) = \bar{X}_n.$$

If g is invertible:

$$\hat{\theta}_{\text{MoM}} = g^{-1}(\bar{X}_n).$$

MLE Example Pattern

MLE uses curvature of $\ell(\theta)$: - gradient (score) gives stationary points, - Hessian indicates maxima/minima, - Fisher information predicts asymptotic variance.

Practical Comparison (Canonical)

- **MoM**: simple, may be less efficient, depends on which moments you pick.
- **MLE**: often efficient under regularity, invariant under reparameterization, but may require numerical optimization and can be sensitive to model misspecification.

Worked Examples

Example 1: Bernoulli(p)

MoM using first moment:

$$\mathbb{E}[X] = p, \quad \hat{m}_1 = \bar{X}_n \Rightarrow \hat{p}_{\text{MoM}} = \bar{X}_n.$$

MLE:

$$\ell(p) = \sum_{i=1}^{n} \left(x_i \log p + (1 - x_i) \log(1 - p)\right).$$

Setting derivative to zero yields:

$$\hat{p}_{\text{MLE}} = \bar{X}_n.$$

Here MoM = MLE.

Example 2: Exponential(λ)

Density: $f_\lambda(x) = \lambda e^{-\lambda x}$ for $x \geq 0$.

MoM:

$$\mathbb{E}[X] = \frac{1}{\lambda} \Rightarrow \hat{\lambda}_{\text{MoM}} = \frac{1}{\bar{X}_n}.$$

MLE:

$$\ell(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i \Rightarrow \hat{\lambda}_{\text{MLE}} = \frac{n}{\sum x_i} = \frac{1}{\bar{X}_n}.$$

Again MoM = MLE.

(They diverge in many multi-parameter or constrained models.)

Cross-Links

- **AI:** MLE corresponds to minimizing negative log-likelihood; MoM resembles matching feature statistics (e.g., method-of-moments in latent variable models).
- **Economics:** Structural estimation often uses MLE; MoM generalizes to GMM (bridge to econometrics).
- **Linear Algebra:** MoM often reduces to solving linear systems; MLE uses gradients/Hessians (quadratic forms).
- **Quantum:** Tomography can be posed as MLE; moment-matching appears via expectation-value constraints.

Structural Compression

Invariant:
Both methods define estimators by enforcing model–data agreement.

Mechanism:
MoM matches expectations; MLE maximizes plausibility via likelihood curvature.

Cross-System Echo:
In control systems: match moments (constraints) vs optimize a global objective.

Exercises

1. Compute MoM and MLE for Poisson(λ) and compare.
2. Give an example where MoM and MLE differ (hint: use a multi-parameter model or a constrained parameter space).
3. Explain why maximizing $\ell(\theta)$ is equivalent to maximizing $L(\theta)$.
4. For Normal(μ, σ^2), derive the MLEs for μ and σ^2.

Statistics · Module 3 — Point Estimation · Node 3.2
Concepts: maximum-likelihood-estimation, likelihood
Prerequisites: 2.3, 2.4, 3.1, 1.5
Enables: 3.3, 3.4

hbar.university — own.your.cognition