M-Estimation

Statistics · Module 7 — Asymptotic Theory

Node 7.5

Position in Knowledge Graph

System: Statistics. Structural Domain: Module 7 — Asymptotic Theory. Node 7.5 — M-Estimation.

M-estimation provides a unified asymptotic framework that encompasses MLE, method of moments, robust estimators, and regression estimators as special cases. This node depends on consistency (Node 7.2), asymptotic normality (Node 7.3), the delta method (Node 7.4), and the likelihood framework (Node 3.2), and enables the likelihood asymptotics of Node 7.6.

Formal Definition

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P_0$.

M-estimator. An M-estimator maximizes a sample objective:

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} m(X_i, \theta).$$

Z-estimator. A Z-estimator (estimating equations estimator) $\hat{\theta}_n$ solves

$$\frac{1}{n} \sum_{i=1}^{n} \psi(X_i, \hat{\theta}_n) = 0,$$

where $\psi(x, \theta) : \mathcal{X} \times \Theta \to \mathbb{R}^k$ is the estimating function.

If $m(x, \theta)$ is differentiable in θ, the M-estimator solves the Z-estimating equations with $\psi(x, \theta) = \nabla_\theta m(x, \theta)$. The MLE is the M-estimator with $m(x, \theta) = \log p_\theta(x)$.

Intuition

M-estimation abstracts away the specific form of the objective function and asks: under what conditions does a sample optimizer converge to a population optimizer, and what is its limiting distribution? The answer separates into two independent components — consistency (the optimizer of the sample criterion converges to the optimizer of the population criterion) and asymptotic normality (the fluctuations around the limit are Gaussian with a specific covariance). The sandwich covariance formula captures the mismatch between the curvature of the objective and the variability of the estimating function, which coincide under correct model specification but diverge under misspecification.

Structural Role

M-estimation unifies MLE, method of moments, robust estimators, and regression estimators under a single asymptotic framework. It decouples the consistency argument (uniform law of large numbers applied to the criterion) from the normality argument (CLT applied to the estimating function at the solution), making both components verifiable independently. The sandwich covariance provides valid inference under misspecification and is the theoretical basis for robust standard errors in econometrics and biostatistics.

Mathematical Development

Consistency of M-estimators.

Let $M(\theta) = \mathbb{E}_{P_0}[m(X, \theta)]$ and $M_n(\theta) = n^{-1} \sum_{i=1}^n m(X_i, \theta)$.

Conditions: 1. $M(\theta)$ has a unique maximizer θ_0. 2. $\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{p} 0$ (uniform convergence). 3. Θ is compact (or the maximizer is well-separated).

Conclusion: $\hat{\theta}_n \xrightarrow{p} \theta_0$.

The proof is the same argmax continuous mapping theorem argument as in Node 7.2: uniform convergence of the criterion plus unique maximization of the limit implies convergence of the maximizers.

Consistency of Z-estimators.

Let $\Psi(\theta) = \mathbb{E}_{P_0}[\psi(X, \theta)]$. If $\Psi(\theta_0) = 0$ has a unique solution, $n^{-1} \sum_i \psi(X_i, \theta) \xrightarrow{p} \Psi(\theta)$ uniformly, and Ψ is continuous, then $\hat{\theta}_n \xrightarrow{p} \theta_0$.

Asymptotic normality of Z-estimators.

Define the Jacobian and variance matrices:

$$A(\theta_0) = \mathbb{E}[\nabla_\theta \psi(X, \theta_0)], \qquad B(\theta_0) = \mathbb{E}[\psi(X, \theta_0)\psi(X, \theta_0)^\top].$$

Regularity conditions: (i) $A(\theta_0)$ is nonsingular; (ii) ψ is differentiable in θ with the derivative satisfying a ULLN; (iii) $B(\theta_0)$ is finite and positive definite.

Theorem. Under these conditions,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}\big(0,\ A(\theta_0)^{-1} B(\theta_0) [A(\theta_0)^{-1}]^\top\big)\,.$$

Proof. Taylor-expand the sample estimating equation around θ_0:

$$0 = \frac{1}{n} \sum_{i=1}^n \psi(X_i, \hat{\theta}_n) = \frac{1}{n} \sum_{i=1}^n \psi(X_i, \theta_0) + \left[\frac{1}{n} \sum_{i=1}^n \nabla_\theta \psi(X_i, \theta_n^*) \right] (\hat{\theta}_n - \theta_0),$$

where θ_n^* lies between $\hat{\theta}_n$ and θ_0. Multiply through by $\sqrt{n}$:

$$0 = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(X_i, \theta_0) + \left[\frac{1}{n} \sum_{i=1}^n \nabla_\theta \psi(X_i, \theta_n^*) \right] \sqrt{n}(\hat{\theta}_n - \theta_0).$$

By the CLT, $n^{-1/2} \sum_i \psi(X_i, \theta_0) \xrightarrow{d} \mathcal{N}(0, B(\theta_0))$.

By the WLLN and consistency, $n^{-1} \sum_i \nabla_\theta \psi(X_i, \theta_n^*) \xrightarrow{p} A(\theta_0)$.

Solving:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -A(\theta_0)^{-1} \cdot \frac{1}{\sqrt{n}} \sum_i \psi(X_i, \theta_0) + o_p(1).$$

By Slutsky's theorem, the limit is $-A(\theta_0)^{-1} \cdot \mathcal{N}(0, B) = \mathcal{N}(0, A^{-1}BA^{-\top})$. $\square$

Sandwich variance estimator. A consistent estimator of the asymptotic covariance is

$$\widehat{V}_n = \hat{A}_n^{-1}\hat{B}_n[\hat{A}_n^{-1}]^\top,$$

where $\hat{A}_n = n^{-1}\sum_i \nabla_\theta \psi(X_i, \hat{\theta}_n)$ and $\hat{B}_n = n^{-1}\sum_i \psi(X_i, \hat{\theta}_n)\psi(X_i, \hat{\theta}_n)^\top$.

Information matrix equality. Under correct model specification with $\psi = \nabla_\theta \log p_\theta$, the Bartlett identities give $A(\theta_0) = -I(\theta_0)$ and $B(\theta_0) = I(\theta_0)$, so $A^{-1}BA^{-\top} = I(\theta_0)^{-1}$. The sandwich reduces to the inverse Fisher information.

Connection to robust statistics. Huber's M-estimator uses $\psi(x, \theta) = \psi_c(x - \theta)$ where

$$\psi_c(u) = \begin{cases} u & |u| \le c, \\ c \cdot \operatorname{sign}(u) & |u| > c, \end{cases}$$

which corresponds to $m(x, \theta) = \rho_c(x - \theta)$ with $\rho_c(u) = u^2/2$ for $|u| \le c$ and $c|u| - c^2/2$ for $|u| > c$. This estimator is consistent for the location parameter under symmetric distributions and has bounded influence function, providing robustness against outliers at the cost of efficiency: the ARE relative to the sample mean under normality is $[2\Phi(c) - 1 - 2c\phi(c) + 2c^2(1 - \Phi(c))]^{-1} \cdot (2\Phi(c) - 1)^2$.

Connection to empirical risk minimization. In machine learning, the empirical risk $n^{-1}\sum_i L(Y_i, f_\theta(X_i))$ is an M-estimation objective with $m(x, \theta) = -L(y, f_\theta(x))$. Consistency requires the function class to be Glivenko-Cantelli; asymptotic normality requires finite VC dimension or analogous complexity control. The sandwich variance appears in the asymptotic distribution of the ERM solution.

Worked Examples

Example 1 (Huber estimator). Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F$ with F symmetric about μ. The Huber M-estimator with $c = 1.345$ solves $\sum_i \psi_{1.345}(X_i - \hat{\mu}) = 0$. Under $F = \mathcal{N}(\mu, 1)$: $A = -\mathbb{E}[\psi'_{1.345}(Z)] = -(2\Phi(1.345) - 1) \approx -0.8227$ and $B = \mathbb{E}[\psi_{1.345}(Z)^2] \approx 0.7101$. The asymptotic variance is $B/A^2 \approx 1.049$, giving ARE ≈ 0.953 relative to the sample mean — 95.3% efficiency at the normal with much greater robustness to outliers.

Example 2 (Heteroskedasticity-robust inference). In the linear model $Y_i = x_i^\top\beta + \varepsilon_i$ with $\mathbb{E}[\varepsilon_i|x_i] = 0$ but $\mathrm{Var}(\varepsilon_i|x_i) = \sigma^2(x_i)$ (heteroskedastic), OLS is an M-estimator with $\psi(X_i, \beta) = x_i(Y_i - x_i^\top\beta)$. Here $A = -\mathbb{E}[x_i x_i^\top]$ and $B = \mathbb{E}[\varepsilon_i^2 x_i x_i^\top]$. The sandwich variance $A^{-1}BA^{-\top}$ is the Huber-White heteroskedasticity-consistent covariance estimator, estimated by $(X^\top X)^{-1}(\sum_i \hat{e}_i^2 x_i x_i^\top)(X^\top X)^{-1}$ where $\hat{e}_i$ are OLS residuals.

Cross-Links

AI. Empirical risk minimization is M-estimation; the training loss is the M-objective, and the sandwich variance appears in the asymptotic distribution of learned parameters. Stochastic gradient descent can be analyzed as an approximate Z-estimator solving the sample score equation.

Economics. Generalized Method of Moments (GMM) is a Z-estimation framework with $\psi = g(\theta)^\top W \cdot h(X_i, \theta)$; the sandwich variance is the Eicker-Huber-White robust standard error, standard in applied econometrics.

Linear Algebra. The sandwich form $A^{-1}BA^{-\top}$ is a congruence transformation of B by A^{-1}, the same algebraic operation that governs covariance transformation under linear maps (Node 8.1).

Quantum Mechanics. Quantum state estimation via maximum likelihood on measurement outcomes is an M-estimation problem; the sandwich variance accounts for measurement-model mismatch when the assumed measurement model differs from the true POVM.

Structural Compression

Invariant. Z-estimators solving $\mathbb{E}[\psi(X, \theta_0)] = 0$ are $\sqrt{n}$-consistent and asymptotically normal with sandwich covariance $A^{-1}BA^{-\top}$.

Mechanism. Taylor linearization of the estimating equation at the solution maps the CLT for ψ into a Gaussian limit for $\hat{\theta}_n$ via the inverse Jacobian A^{-1}. The sandwich structure emerges because the variability (B) and the curvature (A) of the estimating equation are distinct objects that coincide only under correct specification.

Cross-System Echo. M-estimation is the frequentist analogue of variational inference (both minimize an objective over parameters). The sandwich variance is the standard covariance correction in generalized estimating equations (GEE) in biostatistics, in heteroskedasticity-robust standard errors in econometrics, and in quasi-likelihood theory.

Exercises

1. Show that the MLE is a special case of M-estimation with $m(x, \theta) = \log p_\theta(x)$, and verify that the sandwich reduces to $I(\theta)^{-1}$ under correct specification.
2. Derive the influence function $\mathrm{IF}(x; T, F) = -A^{-1}\psi(x, \theta_0)$ for a Z-estimator and verify that $\int \mathrm{IF}(x; T, F)\, dF(x) = 0$.
3. For the Huber M-estimator with tuning constant c, compute the breakdown point and the asymptotic efficiency at the normal as functions of c.
4. In the linear regression model with heteroskedastic errors, derive the Huber-White sandwich estimator from the general Z-estimator formula with $\psi_i = x_i(Y_i - x_i^\top \beta)$.
5. Show that the sample median is a Z-estimator with $\psi(x, \theta) = \mathrm{sign}(x - \theta)$. Compute A and B when $X \sim \mathcal{N}(\mu, \sigma^2)$ and derive the asymptotic variance $\pi\sigma^2/(2n)$.
6. Consider empirical risk minimization with the hinge loss $m(x, \theta) = -\max(0, 1 - y \cdot x^\top \theta)$. Explain why the standard M-estimation asymptotics do not apply directly (non-differentiability) and describe modifications needed.
7. Argue, structurally, why the sandwich variance is the correct asymptotic covariance even under model misspecification, while the model-based variance A^{-1} alone is invalid, connecting this to the failure of the information matrix equality.

Statistics · Module 7 — Asymptotic Theory · Node 7.5
Concepts: maximum-likelihood-estimation, convergence, cost-function
Prerequisites: 7.2, 7.3, 7.4, 3.2
Enables: 7.6

hbar.university — own.your.cognition