Asymptotic Theory

Lecture Notes

Overview

- We now develop the statistical theory governing extremum estimators
- Two key properties to establish:
  1. **Consistency**: Does $\hat{\theta} \to \theta_0$ as the sample grows?
  2. **Inference**: What is the sampling distribution of $\hat{\theta}$ around θ_0?
- Results apply broadly to all extremum estimators, then we specialize to MLE, minimum distance, and GMM
- Treatment follows Newey & McFadden (1994) closely

Definitions

Extremum Estimator

- $\hat{\theta}$ is an **extremum estimator** if:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} Q_N(\theta)$$

  where $\Theta \subset \mathbb{R}^p$ and $Q_N(\cdot)$ depends on the data
- All of our estimators fall into this class; what distinguishes them is the structure of Q_N

M-Estimators

- An **M-estimator** has an objective that is a sample average:

$$Q_N(\theta) = \frac{1}{N} \sum_{n=1}^{N} m(\mathbf{w}_n, \theta)$$

- Two key examples:
  - **Maximum Likelihood**: $m(\mathbf{w}_n, \theta) = \log f(y_n | \mathbf{x}_n, \theta)$
  - **Nonlinear Least Squares**: $m(\mathbf{w}_n, \theta) = -(y_n - \varphi(\mathbf{x}_n, \theta))^2$

GMM Estimator

- Defined by moment conditions $\mathbb{E}[g(\mathbf{w}, \theta_0)] = \mathbf{0}$:

$$Q_N(\theta) = -\frac{1}{2} \mathbf{g}_N(\theta)' \hat{\mathbf{W}} \mathbf{g}_N(\theta), \qquad \mathbf{g}_N(\theta) = \frac{1}{N} \sum_n g(\mathbf{w}_n, \theta)$$

- GMM is itself an M-estimator (expand the quadratic form)

Minimum Distance Estimator

- Works with a first-stage reduced-form estimate $\hat{\pi}$ and model restrictions $\psi(\pi, \theta)$:

$$Q_N(\theta) = -\frac{1}{2}\psi(\hat{\pi}_N, \theta)'\hat{\mathbf{W}}\psi(\hat{\pi}_N, \theta)$$

  where $\psi(\pi_0, \theta_0) = \mathbf{0}$ and $\sqrt{N}(\hat{\pi}_N - \pi_0) \to_d \mathcal{N}(\mathbf{0}, \Omega)$
- Differs from GMM: objective depends on data only through the first-stage statistic $\hat{\pi}$, not through individual observations

---

Consistency

- An extremum estimator solves $\hat{\theta} = \arg\max_{\theta \in \Theta} Q_N(\theta)$
- Let $Q_0(\theta)$ denote the population analogue (probability limit of $Q_N(\theta)$)
- Two conditions needed intuitively:
  1. **Identification**: $Q_0(\theta)$ is uniquely maximized at θ_0
  2. **Convergence**: $Q_N(\theta) \to Q_0(\theta)$ in a sufficiently strong sense

Consistency with Compactness

- **Theorem**: Under the following conditions, $\hat{\theta} \to_p \theta_0$:
  1. Θ is compact
  2. $Q_N(\theta)$ is continuous in θ
  3. $Q_N(\theta)$ is measurable
  4. **Identification**: $Q_0(\theta)$ is uniquely maximized at θ_0
  5. **Uniform convergence**: $\sup_{\theta \in \Theta} |Q_N(\theta) - Q_0(\theta)| \to_p 0$
- **Proof intuition**: Pick any neighborhood $\mathcal{N}$ around θ_0. Since θ_0 uniquely maximizes Q_0 and $\Theta \setminus \mathcal{N}$ is compact, there is a gap:

$$\varepsilon = Q_0(\theta_0) - \sup_{\theta \in \Theta \setminus \mathcal{N}} Q_0(\theta) > 0$$

  Uniform convergence ensures $Q_N(\hat{\theta}) \geq Q_0(\theta_0) - \varepsilon/2$ while $Q_N(\theta) \leq Q_0(\theta_0) - \varepsilon/2$ outside $\mathcal{N}$, so $\hat{\theta}$ must be in $\mathcal{N}$

Consistency without Compactness

- Compactness is a strong assumption—many parameter spaces are unbounded
- **Theorem**: Replace compactness with:
  1. $\theta_0 \in \text{int}(\Theta)$
  2. $Q_N(\theta)$ is **concave** in θ
  3. **Pointwise convergence** (not uniform): $Q_N(\theta) \to_p Q_0(\theta)$ for all θ
- Key insight: concavity turns pointwise convergence into uniform convergence on compact subsets (Rockafellar, 1970)

Uniform Law of Large Numbers for M-Estimators

- For M-estimators, uniform convergence reduces to a uniform LLN
- **Sufficient conditions** (with $\{\mathbf{w}_n\}$ ergodic stationary):
  1. Θ compact
  2. $m(\mathbf{w}, \theta)$ continuous in θ
  3. $m(\mathbf{w}, \theta)$ measurable in $\mathbf{w}$
  4. Dominance: $|m(\mathbf{w}, \theta)| \leq d(\mathbf{w})$ for all θ with $\mathbb{E}[d(\mathbf{w})] < \infty$
- Then $\sup_{\theta \in \Theta} |Q_N(\theta) - Q_0(\theta)| \to_p 0$
- In practice, verify $\mathbb{E}[\sup_{\theta \in \Theta} |m(\mathbf{w}, \theta)|] < \infty$

Consistency of Maximum Likelihood

- For MLE: $Q_N(\theta) = \frac{1}{N} \sum_n \log f(\mathbf{w}_n; \theta)$ and $Q_0(\theta) = \mathbb{E}_{\theta_0}[\log f(\mathbf{w}; \theta)]$
- **Identification via the Kullback-Leibler inequality**: for any two densities g and h:

$$\mathbb{E}_g \left[\log \frac{g(\mathbf{w})}{h(\mathbf{w})} \right] \geq 0$$

  with equality iff $g = h$ a.e.
- Applying with $g = f(\cdot; \theta_0)$ and $h = f(\cdot; \theta)$:

$$\mathbb{E}_{\theta_0}[\log f(\mathbf{w}; \theta_0)] \geq \mathbb{E}_{\theta_0}[\log f(\mathbf{w}; \theta)]$$

  with equality iff $f(\cdot; \theta) = f(\cdot; \theta_0)$ a.e.
- So: as long as different θ imply different densities, the population log-likelihood is uniquely maximized at θ_0
- **Theorem** (MLE Consistency): Under compactness, continuity, identification ($f(\mathbf{w}; \theta_0) \neq f(\mathbf{w}; \theta)$ w.p.p. for $\theta \neq \theta_0$), and dominance, $\hat{\theta}_{ML} \to_p \theta_0$
- Identification here is model-specific: different parameters must imply different distributions
- Analogous result holds without compactness when log-likelihood is concave (e.g. exponential family)

Asymptotic Normality for M-Estimators

- Having established consistency, now characterize the rate and distribution of $\hat{\theta}$ around θ_0
- Define the **score** and **Hessian** of m:

$$\mathbf{s}(\mathbf{w}, \theta) = \frac{\partial m(\mathbf{w}, \theta)}{\partial \theta} \qquad (p \times 1)$$

$$\mathbf{H}(\mathbf{w}, \theta) = \frac{\partial^2 m(\mathbf{w}, \theta)}{\partial \theta \partial \theta'} \qquad (p \times p)$$

Derivation via the Mean Value Theorem

- FOC: $\frac{1}{N}\sum_n \mathbf{s}(\mathbf{w}_n, \hat{\theta}) = \mathbf{0}$
- Mean value expansion around θ_0:

$$\mathbf{0} = \frac{1}{N}\sum_n \mathbf{s}(\mathbf{w}_n, \theta_0) + \left[\frac{1}{N}\sum_n \mathbf{H}(\mathbf{w}_n, \bar{\theta})\right](\hat{\theta} - \theta_0)$$

- Rearranging:

$$\sqrt{N}(\hat{\theta} - \theta_0) = -\left[\frac{1}{N}\sum_n \mathbf{H}(\mathbf{w}_n, \bar{\theta})\right]^{-1} \frac{1}{\sqrt{N}}\sum_n \mathbf{s}(\mathbf{w}_n, \theta_0)$$

- Two standard arguments:
  1. **CLT**: $\frac{1}{\sqrt{N}}\sum_n \mathbf{s}(\mathbf{w}_n, \theta_0) \to_d \mathcal{N}(\mathbf{0}, \Sigma)$ where $\Sigma = \mathbb{E}[\mathbf{s}\mathbf{s}']$
  2. **LLN + continuity**: $\frac{1}{N}\sum_n \mathbf{H}(\mathbf{w}_n, \bar{\theta}) \to_p \mathbb{E}[\mathbf{H}(\mathbf{w}, \theta_0)]$
- Combine via Slutsky's theorem

Asymptotic Normality Theorem

- **Theorem**: Under consistency conditions plus interiority, twice-differentiability, CLT for score, bounded Hessian, and nonsingular expected Hessian:

$$\sqrt{N}(\hat{\theta} - \theta_0) \to_d \mathcal{N}\left(\mathbf{0},\ \mathbb{E}[\mathbf{H}]^{-1}\Sigma\mathbb{E}[\mathbf{H}]^{-1}\right)$$

  where $\mathbb{E}[\mathbf{H}] = \mathbb{E}[\mathbf{H}(\mathbf{w}, \theta_0)]$ and $\Sigma = \mathbb{E}[\mathbf{s}(\mathbf{w}, \theta_0)\mathbf{s}(\mathbf{w}, \theta_0)']$
- This is the **sandwich formula**
- In practice, replace population expectations with sample analogues:

$$\hat{\mathbb{V}}[\hat{\theta}] = \hat{H}^{-1}\hat{\Sigma}\hat{H}^{-1}/N$$

The Information Matrix Equality

- For MLE, $m(\mathbf{w}, \theta) = \log f(\mathbf{w}; \theta)$, and a remarkable simplification occurs
- The **information matrix equality**:

$$\mathcal{I}(\theta_0) \equiv \mathbb{E}[\mathbf{s}\mathbf{s}'] = -\mathbb{E}[\mathbf{H}]$$

- **Why it holds**: Since $\int f(\mathbf{w}; \theta) d\mathbf{w} = 1$, differentiate under the integral:
  - First derivative: $\mathbb{E}_\theta[\mathbf{s}(\mathbf{w}, \theta)] = 0$
  - Second derivative: $\mathbb{E}[\mathbf{H}] + \mathbb{E}[\mathbf{s}\mathbf{s}'] = 0$
- Consequence: for MLE, the sandwich collapses to:

$$\sqrt{N}(\hat{\theta}_{ML} - \theta_0) \to_d \mathcal{N}\left(\mathbf{0},\ \mathcal{I}(\theta_0)^{-1}\right)$$

- $\mathcal{I}(\theta)$ is the **Fisher information matrix**
- MLE variance can be estimated three ways: Hessian, outer product of scores, or sandwich (robust to misspecification)

Probit Standard Errors (Key Results)

- Illustrating the three variance estimators (Hessian, OPG, sandwich) on the probit model from the Roy Model example
- All three should agree when model is correctly specified
- Monte Carlo verification: asymptotic SEs match the standard deviation of MC estimates across replications
- The asymptotic normal approximation matches the MC sampling distribution well

---

The Delta Method

- Often interested in $g(\theta)$ rather than θ itself (e.g. λ and b in the search model)
- **Theorem**: If $\sqrt{N}(\hat{\theta}-\theta_0) \rightarrow_d \mathcal{N}(\mathbf{0}, V)$ and $g : \mathbb{R}^p \rightarrow \mathbb{R}^k$ is continuously differentiable with full-rank Jacobian $\nabla g(\theta_0)$, then:

$$\sqrt{N}(g(\hat{\theta}) - g(\theta_0)) \rightarrow_d \mathcal{N}\left(\mathbf{0},\ \nabla g(\theta_0) V \nabla g(\theta_0)'\right)$$

- Proof: continuous mapping theorem applied to first-order Taylor expansion
- In practice, the Jacobian ∇g can be computed by automatic differentiation

---

Minimum Distance Estimators

- Different approach: two stages
  1. Estimate reduced-form object $\hat{\pi}$ with $\sqrt{N}(\hat{\pi}-\pi_0) \rightarrow_d \mathcal{N}(\mathbf{0}, \Omega)$
  2. Find structural θ that best fits model restrictions $\psi(\pi, \theta) = \mathbf{0}$

The Estimator

$$\hat{\theta} = \arg\min_{\theta} \psi(\hat{\pi}, \theta)' \mathbf{W}_N \psi(\hat{\pi}, \theta)$$

Asymptotic Distribution

- **Theorem**: Under identification ($\psi(\pi_0, \theta) \neq \mathbf{0}$ for $\theta \neq \theta_0$), asymptotic normality of $\hat{\pi}$, $\mathbf{W}_N \rightarrow_p \mathbf{W}$, and full-rank $\nabla_\theta \psi_0$:

$$\sqrt{N}(\hat{\theta}-\theta_0) \rightarrow_d \mathcal{N}(\mathbf{0}, V_{MD})$$

  where:

$$V_{MD} = (\nabla_\theta \psi_0 \mathbf{W} \nabla_\theta \psi_0')^{-1} \nabla_\theta \psi_0 \mathbf{W} \nabla_\pi \psi_0' \Omega \nabla_\pi \psi_0 \mathbf{W} \nabla_\theta \psi_0' (\nabla_\theta \psi_0 \mathbf{W} \nabla_\theta \psi_0')^{-1}$$

- **Derivation intuition**: FOC + expand $\psi(\hat{\pi}, \hat{\theta})$ around (π_0, θ_0), solve for $\sqrt{N}(\hat{\theta}-\theta_0)$ as a linear function of $\sqrt{N}(\hat{\pi}-\pi_0)$

The Optimal Weighting Matrix

- Optimal $\mathbf{W}$ minimizes V_{MD} (in the PSD sense):

$$\mathbf{W}^* = \left(\nabla_\pi \psi_0' \Omega \nabla_\pi \psi_0\right)^{-1}$$

- Variance simplifies to:

$$V_{MD}^* = \left(\nabla_\theta \psi_0 \left(\nabla_\pi \psi_0' \Omega \nabla_\pi \psi_0\right)^{-1} \nabla_\theta \psi_0'\right)^{-1}$$

- **Common case** $\psi(\pi, \theta) = \pi - h(\theta)$: then $\nabla_\pi \psi = I$, $\nabla_\theta \psi = -\nabla_\theta h$, and $\mathbf{W}^* = \Omega^{-1}$:

$$V_{MD}^* = \left(\nabla_\theta h_0 \Omega^{-1} \nabla_\theta h_0'\right)^{-1}$$

Efficiency of Optimal Minimum Distance

- When model is **just-identified** ($\dim(\psi) = \dim(\theta)$) and $\hat{\pi}$ is MLE: the optimally weighted MD estimator achieves the Cramér-Rao lower bound
- Uses the implicit function theorem to show $V_{MD}^* = \mathcal{I}_\theta^{-1}$
- Over-identification introduces some loss relative to MLE but gains robustness

Income Process Standard Errors (Key Results)

- Matching variance of log income at each age to model-implied variances
- With $\psi(\pi, \theta) = \pi - \mathbf{v}(\theta)$: Jacobian $\nabla_\theta \mathbf{v}$ computed via ForwardDiff
- Variance of sample moments Ω estimated using $\text{Var}(\hat{\sigma}^2) \approx 2\sigma^4/(n-1)$
- Optimally weighted estimator is more precise, especially when moments have different scales/variances
- **Important caveat**: variance estimate assumes normality and zero off-diagonal covariances; the latter is explicitly wrong since same individuals appear at multiple ages; the bootstrap (next chapter) will fix this

---

The Generalized Method of Moments

- GMM objective:

$$Q_N(\theta) = -\frac{1}{2}\mathbf{g}_N(\theta)'\mathbf{W}_N \mathbf{g}_N(\theta)$$

Asymptotic Distribution

- **Theorem**: Let $G = \mathbb{E}[\nabla_\theta g(\mathbf{w}, \theta_0)']$ and $S = \mathbb{E}[g(\mathbf{w}, \theta_0) g(\mathbf{w}, \theta_0)']$. Then:

$$\sqrt{N}(\hat{\theta}_{GMM} - \theta_0) \rightarrow_d \mathcal{N}\left(\mathbf{0},\ (G'\mathbf{W}G)^{-1}G'\mathbf{W}S\mathbf{W}G(G'\mathbf{W}G)^{-1}\right)$$

Optimal Weighting Matrix

- Optimal: $\mathbf{W}^* = S^{-1}$
- Variance simplifies to: $V_{GMM}^* = (G'S^{-1}G)^{-1}$
- When **just-identified**: GMM does not depend on $\mathbf{W}$ at all (sample moments set exactly to zero)

Feasible Efficient GMM

- S depends on θ_0 and must be estimated; use **two-step GMM**:
  1. Estimate $\hat{\theta}_1$ with initial $\mathbf{W}$ (e.g. I)
  2. Compute $\hat{S} = \frac{1}{N}\sum_n g(\mathbf{w}_n, \hat{\theta}_1) g(\mathbf{w}_n, \hat{\theta}_1)'$
  3. Re-estimate: $\hat{\theta}_2 = \arg\min_\theta \mathbf{g}_N(\theta)'\hat{S}^{-1}\mathbf{g}_N(\theta)$
- $\hat{\theta}_2$ is asymptotically efficient; first-stage estimation of $\hat{S}$ does not affect the asymptotic variance

---

Efficiency

- $\hat{\theta}_1$ is **asymptotically efficient** relative to $\hat{\theta}_2$ if $V_2 - V_1 \geq 0$ (PSD)

Efficiency of Maximum Likelihood

- MLE achieves the **Cramér-Rao lower bound**: $V[\hat{\theta}] - \mathcal{I}(\theta_0)^{-1} \geq 0$ for any consistent, asymptotically normal estimator
- MLE is efficient relative to any GMM estimator using moments implied by the model
- The efficiency gap: $V_{GMM} - V_{MLE} = \mathbb{E}[\mathbf{ms}']^{-1}\mathbb{E}[\mathbf{UU}']\mathbb{E}[\mathbf{sm}']^{-1} \geq 0$
  - Where $\mathbf{U}$ is the projection residual of the GMM influence function on the MLE score
  - Equals zero only when the GMM estimator fully exploits the likelihood

STOP FOR DISCUSSION
Efficiency vs. Robustness:

- MLE requires the **entire parametric model** to be correctly specified. If the density is wrong, MLE converges to a pseudo-true value and the information matrix equality fails.
- GMM only requires the **moment conditions** to be correct—more robust to partial misspecification.
- The sandwich variance estimator remains valid for MLE even under misspecification.

STOP FOR DISCUSSION
Pros and Cons of MLE:

- Greatest strength: MLE **uses every piece of information in the data**. If the model is identified by the population distribution, you don't have to choose which features to match. Particularly useful for unobserved heterogeneity with panel data.
- Greatest weakness: MLE **uses every piece of information in the data**. You don't control which features the model fits vs. misses. The most credible identification strategy ("whether vs. how") may not coincide with MLE.

Two-Step Estimators

- Many structural estimators proceed in stages (e.g. probit then selection-corrected OLS in the Roy Model)
- Need to account for first-stage estimation uncertainty in second-stage inference

Setup

- **First step**: Estimate $\hat{\gamma}$ via $\frac{1}{N}\sum_n g_1(\mathbf{w}_n, \hat{\gamma}) = \mathbf{0}$
- **Second step**: Estimate $\hat{\beta}$ via $\frac{1}{N}\sum_n g_2(\mathbf{w}_n, \hat{\gamma}, \hat{\beta}) = \mathbf{0}$

Asymptotic Distribution

- Stack moment conditions with $\alpha = (\gamma', \beta')'$:

$$\frac{1}{N}\sum_n \begin{bmatrix} g_1(\mathbf{w}_n, \gamma) \\ g_2(\mathbf{w}_n, \gamma, \beta) \end{bmatrix} = \mathbf{0}$$

- The Jacobian has a **block-triangular** structure:

$$\Gamma = \begin{bmatrix} \Gamma_{1\gamma} & 0 \\ \Gamma_{2\gamma} & \Gamma_{2\beta} \end{bmatrix}$$

  - Zero in the upper-right reflects that the first step does not depend on β

Corrected Variance Formula

- Asymptotic variance of $\hat{\beta}$:

$$V_\beta = \Gamma_{2\beta}^{-1}\mathbb{E}\left[(g_2 - \Gamma_{2\gamma}\Gamma_{1\gamma}^{-1}g_1)(g_2 - \Gamma_{2\gamma}\Gamma_{1\gamma}^{-1}g_1)'\right]\Gamma_{2\beta}^{-1\prime}$$

- The term $\Gamma_{2\gamma}\Gamma_{1\gamma}^{-1}g_1$ is the **correction** for first-stage estimation error
- Ignoring this correction gives incorrect inference in general

When Does the First Stage Not Matter?

- If $\Gamma_{2\gamma} = \mathbb{E}[\nabla_\gamma g_2'] = 0$, the correction vanishes
- This means the second-step moments are **locally insensitive** to first-step parameters at the true values
- Classic example: two-stage IV with a probit first stage
  - $\mathbb{E}[\nabla_\gamma g_2'] = 0$ because the projection of the error on functions of instruments is zero by construction

STOP FOR DISCUSSION

Bootstrap vs. Analytical Standard Errors:

- Computing analytical SEs for two-step estimators can be tedious
- Alternative: **bootstrap**—resample data, re-run the entire two-step procedure, compute SEs from the distribution of bootstrap estimates
- Automatically accounts for first-stage estimation uncertainty without explicit correction terms
- We will discuss the bootstrap formally in the simulation methods chapter