Methods Review of *Maximum Likelihood Estimation of Misspecified Models* by Halbert White

Jim Harmon

University of Washington

May 08, 2012
This slide intentionally left blank.
Vizzini: ...and you’re no match for my brain.
Man in Black: You’re that smart?
Vizzini: Let me put it this way. Have you ever heard of Plato, Aristotle, Socrates?
Man in Black: Yes.
Vizzini: Morons.
Remind me...

Answers questions about the following in a unified framework:

- does MLE converge? (interpretation?)
- if yes, is MLE asymptotically normal?
- can properties of MLE determine model truth?

(now...fasten your seatbelts!)
Underlying Assumptions

A1: true density function $g(u)$ for data $U_t$, with distribution function $G$

A2: family of distributions $F(u, \theta)$, with density $f(u, \theta)$, measurable in $u$
for all $\theta \in \Theta$, and continuous in $\theta$ for all $u \in \Omega$

Define $L_n(U, \theta) = \frac{1}{n} \sum_{t=1}^{n} \log f(U_t, \theta)$.

Define QMLE $= \arg \max_{\theta} L_n(U, \theta)$ (quasi-MLE)

Theorem

Given A1 and A2, for all $n$ there exists a measurable QMLE, $\hat{\theta}_n$.

Note: there is an underlying dominating measure $\nu$
Further Assumptions

A3a: $E(\log g(U_t))$ exists and $|\log f(u, \theta)|$ is bounded by an integrable function of $u$

A3b: KLIC $I(g : f, \theta)$ has a unique minimum at $\theta_* \in \Theta$.

**Theorem**

*Given A1-A3, $\hat{\theta}_n \rightarrow_{a.s.} \theta_*$.***

Note: All expectations are taken w.r.t. the truth, $g$. 
Sandwich Time!!!

Need a consistent estimate of the covariance matrix:

\[ A(\theta) = E \left[ \frac{\partial^2 \log(f(U_t, \theta))}{\partial \theta_i \partial \theta_j} \right] \]

\[ B(\theta) = E \left[ \frac{\partial \log(f(U_t, \theta))}{\partial \theta_i} \cdot \frac{\partial \log(f(U_t, \theta))}{\partial \theta_j} \right] \]

\[ C(\theta) = A(\theta)^{-1} B(\theta) A(\theta)^{-1} \]
A4: \( \frac{\partial \log f(u, \theta)}{\partial \theta_i} \) with \( i = 1, \ldots, p \) are measurable functions of \( u \) for each \( \theta \) and continuously differentiable functions of \( \theta \) for each \( u \).

A5: \( | \frac{\partial^2 \log f(u, \theta)}{\partial \theta_i \partial \theta_j} | \) and \( | \frac{\partial \log f(u, \theta)}{\partial \theta_i} \cdot \frac{\partial \log f(u, \theta)}{\partial \theta_j} | \) with \( i, j = 1, \ldots, p \) are dominated by functions integrable w.r.t. \( G \) for \( u \) and \( \theta \).

A6a: \( \theta_* \) is interior to \( \Theta \)

A6b: \( B(\theta_*) \) is nonsingular

A6c: \( A(\theta) \) has constant rank in some open neighborhood of \( \theta_* \) (regular point)
Two Theorems

Theorem (Identification)

i: Given A1-A3a, A4-A6a, if $\theta_*$ is a unique minimum for $I(g : f, \theta)$ in an open neighborhood of $\Theta$, and if $\theta_*$ is a regular point of $A(\theta)$, then $A(\theta_*)$ is negative definite.

ii: Given A1-A3a, A4-A6a, if $A(\theta_*)$ is negative definite and if $\theta_*$ minimizes $I(g : f, \theta)$ in an open neighborhood of $\Theta$, then there is an open neighborhood of $\Theta$ where $\theta_*$ is a unique minimum of $I(g : f, \theta)$.

Theorem (Asymptotic Normality)

Given A1-A6, $\sqrt{n}(\hat{\theta}_n - \theta_*) \to_d N(0, C(\theta_*))$. Moreover, $C_n(\hat{\theta}_n) \to_{a.s.} C(\theta_*)$ element by element.
A7: \[ \partial \left[ \partial f(u, \theta) / \partial \theta_i \cdot f(u, \theta) \right] / \partial \theta_j \] with \( i, j = 1, \ldots, p \) are dominated by functions integrable with respect to \( \nu \) for all \( \theta \) in \( \Theta \) and the minimal support of \( f(u, \theta) \) does not depend on \( \theta \).

**Theorem (Information Matrix Equivalence)**

Given A1-A7, if \( g(u) = f(u, \theta_0) \) for \( \theta_0 \in \Theta \), then \( \theta_* = \theta_0 \) and

\[ A(\theta_0) = -B(\theta_0), \text{ so that } C(\theta_0) = -A(\theta_0)^{-1} = B(\theta_0)^{-1} \text{ where } -A(\theta_0)^{-1} \]

is Fisher’s Information Matrix.

Note: A1-A7 and \( g(u) = f(u, \theta_0) \) are ”usual MLE regularity conditions”
Wald Test under misspecification

Suppose we wish to test $H_0 : s(\theta_0) = 0$ vs. $H_1 : s(\theta_0) \neq 0$ where $s : \mathbb{R}^p \rightarrow \mathbb{R}^r$ is a continuous vector function of $\theta$ s.t. its Jacobian at $\theta_\ast$, $J_s(\theta_\ast)$ is finite with full row rank $r$.

Theorem (Wald Test)

$$W_n = n \cdot s(\hat{\theta}_n)'[J_s(\hat{\theta}_n)C_n(\hat{\theta}_n)J_s(\hat{\theta}_n)']^{-1}s(\hat{\theta}_n) \rightarrow_d \chi_r^2$$
Let $\tilde{\theta}_n$ solve the constrained maximization problem $\max_{\theta \in \Theta} L_n(U, \theta)$ subject to $s(\theta) = 0$

**Theorem (Lagrange Multiplier Test)**

Given A1-A6 and $H_0$,

$$\mathcal{LM}_n = \nabla L_n(U, \tilde{\theta}_n)' A_n(\tilde{\theta}_n)^{-1} J_s(\tilde{\theta}_n)' \times [J_s(\tilde{\theta}_n) C_n(\tilde{\theta}_n) J_s(\tilde{\theta}_n)']^{-1} \times J_s(\tilde{\theta}_n) A_n(\tilde{\theta}_n)^{-1} \nabla L_n(U, \tilde{\theta}_n) \rightarrow_d \chi_r^2$$

Moreover $\mathcal{W}_n - \mathcal{LM}_n \rightarrow_p 0$
More Notation

\( \theta \) is a \( p \)-dimensional vector.

\[
d_{l}(U_t, \theta) = \partial \log(f(U_t, \theta))/\partial \theta_i \cdot \partial \log(f(U_t, \theta))/\partial \theta_j
\]

\[
+ \partial^2 \log(f(U_t, \theta))/\partial \theta_i \partial \theta_j
\]

\( \text{dim}(d) = q \times 1 \text{ with } q \leq p(p + 1)/2 \)

\[
D_{ln}(\hat{\theta}_n) = n^{-1} \sum_{t=1}^{n} d_{l}(u_t, \hat{\theta}_n)
\]

\[
J_{D}(\theta) = n^{-1} \sum_{t=1}^{n} \partial d(U_t, \theta)/\partial \theta_k
\]

\[
W_{n}(\hat{\theta}_n) = d(U_t, \hat{\theta}_n) - J_{D}(\hat{\theta}_n)A(\hat{\theta}_n)^{-1}\nabla \log(f(U_t, \hat{\theta}_n))
\]

\[
V(\theta) = n^{-1} \sum_{t=1}^{n} W_{n}(\hat{\theta}_n) \cdot W_{n}(\hat{\theta}_n)'
\]
A8: \( \partial d_l(u, \theta) / \partial \theta_k \) for \( l = 1, \ldots, q, \ k = 1, \ldots, p \) exist and are continuous functions of \( \theta \) for each \( u \).

A9: |\( d_l(u, \theta)d_m(u, \theta) \)|, |\( \partial d_l(u, \theta) / \partial \theta_k \)|, and |\( d_l(u, \theta) \partial \log f(u, \theta) / \partial \theta_k \)|, for \( l, m = 1, \ldots, q, \ k = 1, \ldots, p \) are dominated by functions integrable w.r.t. \( G \) for all \( u \) and \( \theta \) in \( \Theta \).

A10: \( V(\theta_\ast) \) is nonsingular

**Theorem (Information Matrix Test)**

*Given A1-A10, if \( g(u) = f(u, \theta_0) \) for some \( \theta_0 \in \Theta \), i)*

\( \sqrt{n}D_n(\hat{\theta}_n) \xrightarrow{d} N(0, V(\theta_0)) \)

*ii)* \( V_n(\hat{\theta}_n) \xrightarrow{a.s.} V(\theta_0) \)

*iii)* \( \mathcal{I}_n = nD_n(\hat{\theta}_n)'V_n(\hat{\theta}_n)^{-1}D_n(\hat{\theta}_n) \xrightarrow{d} \chi^2_q \)
Alternative Consistent QMLEs

Let $\Theta$ and $\Gamma$ be $p$- and $q$- dimensional compact subsets of Euclidean spaces with

$$\Theta = B \times \Psi \quad \text{and} \quad \Gamma = B \times A, \ B \subset \mathbb{R}^k \ (\text{compact})$$

$$\hat{\theta}'_n = (\hat{\beta}'_n, \hat{\psi}'_n) \quad \text{maximizes} \quad n^{-1} \sum \log f(U_t, \theta) \quad \text{over} \quad \Theta$$

$$\tilde{\gamma}'_n = (\tilde{\beta}'_n, \tilde{\alpha}'_n) \quad \text{maximizes} \quad n^{-1} \sum \log h(U_t, \gamma) \quad \text{over} \quad \Gamma$$

$h$ is a density function satisfying

A11: $h$ satisfies A2-A6, and if $g(u) = f(u, \theta_0)$ for any $\theta'_0 = (\beta'_0, \psi'_0) \in \Theta$, then $\gamma'_* = (\beta'_0, \alpha'_*) \in \Gamma$

Note: $\tilde{\beta}_n$ is a consistent estimator of $\beta_0$ and $\sqrt{n}(\tilde{\beta}_n - \beta_0)$ is asymptotically normal, consider $\sqrt{n}(\tilde{\beta}_n - \hat{\beta}_n)$
More Definitions

\( A^f(\theta) = \left( E \left( \partial^2 \log f(U_t, \theta) / \partial \theta_i \partial \theta_j \right) \right), \text{ dimension } p \times p \)

\( B^f(\theta) = \left( E \left( \partial \log f(U_t, \theta) / \partial \theta_i \cdot \partial \log f(U_t, \theta) / \partial \theta_j \right) \right), \text{ dimension } p \times p \)

\( A^h(\gamma) = \left( E \left( \partial^2 \log h(U_t, \gamma) / \partial \gamma_i \partial \gamma_j \right) \right), \text{ dimension } q \times q \)

\( B^h(\gamma) = \left( E \left( \partial \log h(U_t, \gamma) / \partial \gamma_i \cdot \partial \log h(U_t, \gamma) / \partial \gamma_j \right) \right), \text{ dimension } q \times q \)

\( A^{f, \beta \theta}(\theta)^{-1} \) is the matrix obtained by deleting the last \( p - k \) rows from the inverse of \( A^f(\theta) \) above.

\( A^{h, \beta \gamma}(\gamma)^{-1} \) is the matrix obtained by deleting the last \( q - k \) rows from the inverse of \( A^h(\gamma) \) above.
Even more notation

\[ R(\theta, \gamma) = (E(\partial \log f(U_t, \theta)/\partial \theta_i \cdot \partial \log h(U_t, \gamma)/\partial \gamma_j)) \]

\[ S(\theta, \gamma) = A^{h, \beta \gamma(\gamma)^{-1}} B^h(\gamma) A^{h, \beta \gamma(\gamma)^{-1'}} \]

\[ - A^{h, \beta \gamma(\gamma)^{-1}} R(\theta, \gamma)^{'} A^{f, \beta \theta(\theta)^{-1'}} \]

\[ - A^{f, \beta \theta(\theta)^{-1}} R(\theta, \gamma) A^{h, \beta \gamma(\gamma)^{-1'}} \]

\[ + A^{f, \beta \theta(\theta)^{-1}} B^f(\theta) A^{f, \beta \theta(\theta)^{-1'}} \]

A12: \( S(\theta_*, \gamma_*) \) is nonsingular.
First of the second round of tests

Theorem (Hausman Test)

Given A1-A6, A11, and A12, if \( g(u) = f(u, \theta_0) \) for \( \theta_0 \in \Theta \), then

\[ \mathcal{H}_n = n(\tilde{\beta}_n - \hat{\beta}_n)'S_n(\hat{\theta}_n, \tilde{\gamma}_n)^{-1}(\tilde{\beta}_n - \hat{\beta}_n) \to_d \chi^2_k \]
Gradient Test setup

\[ \tilde{\gamma}'_n = (\tilde{\beta}'_n, \tilde{\alpha}'_n) \text{ maximizes } n^{-1} \sum \log h(U_t, \gamma) \text{ over } \Gamma \]

\[ \tilde{\psi}_n \text{ maximizes } \nabla L_n(U, \tilde{\beta}_n, \psi) \text{ over } \Psi. \]

\[ \tilde{\theta}'_n = (\tilde{\beta}'_n, \tilde{\psi}'_n) \]

\[ \nabla_\beta L_n(U, \tilde{\theta}_n) \text{ is an indicator of model misspecification} \]

investigate asymptotic behavior of \[ \sqrt{n} \nabla_\beta L_n(U, \tilde{\theta}_n) \]
The Last One (I Promise!!)

$A_n^{f,\beta\beta}(\theta)^{-1}$ is the $k \times k$ submatrix of $A_n^f(\theta)^{-1}$ obtained by deleting the last $p - k$ columns from $A_n^{f,\beta\theta}(\theta)^{-1}$ (i.e., keep the upper left block)

**Theorem (Gradient Test)**

Given $A1$-$A6$, $A11$, and $A12$, if $g(u) = f(u, \theta_0)$ for some $\theta_0 \in \Theta$, then

$$G_n = \nabla_\beta L_n(U, \tilde{\theta}_n)'A_n^{f,\beta\beta}(\tilde{\theta}_n)^{-1}S_n(\tilde{\theta}_n, \tilde{\gamma}_n)^{-1}A_n^{f,\beta\beta}(\tilde{\theta}_n)^{-1}\nabla_\beta L_n(U, \tilde{\theta}_n) \to_d \chi^2_k$$

Moreover $H_n - \mathcal{G}_n \to_p 0$
THE METHOD (according to White)

- Step 1: Perform first test.
- Step 2a: If you "do not reject", MLE away!
- Step 2b: If you "reject", perform one of the other two tests
- Step 3a: If you "do not reject", use sandwich inference
- Step 3b: If you "reject", reconsider your model choice.
Run simulations to verify that the tests work asymptotically

- Wald test
- Information Matrix Test (1st misspecification test)
- Hausman Test (2nd misspecification test)

Check confidence interval coverage using

- "the method" vs.
- just using sandwich vs.
- just using MLE inference