Maximum likelihood:

counterexamples, examples, and open problems

Jon A. Wellner

University of Washington
visiting Vrije Universiteit, Amsterdam

Talk at BeNeLuxFra

Mathematics Meeting

21 May, 2005

Email: jaw@stat.washington.edu

http: //www.stat.washington.edu/ jaw/jaw.research.html

Outline

1. **Introduction:**
   Maximum Likelihood Estimation

2. **Counterexamples**

3. **Beyond consistency: rates and distributions**

4. **Positive Examples**

5. **Problems and Challenges**

1. Introduction: maximum likelihood estimation

Setting 1: dominated families Suppose that $X_1, \ldots, X_n$ are i.i.d. with density p_{θ_0} with respect to some dominating measure μ where $p_{\theta_0} \in \mathcal{P} = \{p_\theta : \ \theta \in \Theta\}$ for $\Theta \subset \mathbb{R}^d$.

The likelihood is

$$L_n(\theta) = \prod_{i=1}^{n} p_\theta(X_i)\,.$$

Definition: A Maximum Likelihood Estimator (or MLE) of θ_0 is any value $\theta \in \Theta$ satisfying

$$L_n(\theta) = \sup_{\theta \in \Theta} L_n(\theta)\,.$$

Equivalently, the MLE θ maximizes the log-likelihood

$$\log L_n(\theta) = \sum_{i=1}^{n} \log p_\theta(X_i)\,.$$

Example 1. Exponential (θ). If $X_1, \ldots, X_n$ are i.i.d. p_{θ_0} where

$$p_\theta(x) = \theta \exp(-\theta x) 1_{[0,\infty)}(x)$$

Then

$$L_n(\theta) = \theta^n \exp(-\theta \sum_1^n X_i)$$

so

$$\log L_n(\theta) = n \log(\theta) - \theta \sum_1^n X_i$$

and $\theta_n = 1/\overline{X}_n$.

Example 2. Monotone decreasing densities on $[0, \infty)$. If $X_1, \ldots, X_n$ are i.i.d. $p_0 \in \mathcal{P}$ where

$$\mathcal{P} = \text{ all nonincreasing densities on } [0, \infty)$$

Then

$$L_n(p) = \prod_{i=1}^{n} p(X_i)$$

is maximized by the Grenander estimator:

$$\begin{aligned} p_n(x) \;\; = \;\; & \text{left derivative at x of the} \\ & \text{Least Concave Majorant} \\ & \mathbb{C}_n \;\text{ of }\; \mathbb{F}_n \end{aligned}$$

where $\mathbb{F}_n(x) = n^{-1} \sum_{i=1}^{n} 1\{X_i \le x\}$.

(contributions by Birgé!)

Setting 2: non-dominated families
Suppose that $X_1, \ldots, X_n$ are i.i.d. $P_0 \in \mathcal{P}$ where $\mathcal{P}$ is some collection of probability measures on a measurable space $(\mathcal{X}, \mathcal{A})$. If $P\{x\}$ denotes the measure under P of the one-point set $\{x\}$, the empirical likelihood of $X_1, \ldots, X_n$ is defined to be

$$L_n(P) = \prod_{i=1}^{n} P\{X_i\}\,.$$

Then a Maximum Likelihood Estimator (or MLE) of P_0 can be defined as a measure $P_n \in \mathcal{P}$ that maximizes $L_n(P)$; thus

$$L_n(P) = \sup_{P \in \mathcal{P}} L_n(P)$$

if it exists.

Example 3. If $\mathcal{P}$ =all probability measures on $(\mathcal{X}, \mathcal{A})$, then

$$P_n = \mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

where $\delta_x(A) = 1_A(x)$.

Consistency of the MLE:

Wald (1949)
Kiefer and Wolfowitz (1956)
Huber (1967)
Perlman (1972)
Wang (1985)
van de Geer (1993)

Counterexamples:

- Neyman and Scott (1948)
- Bahadur (1958)
- Ferguson (1982)
- LeCam (1975), (1990)
- Barlow et al. (4B s) (1972)
- Boyles, Marshall, and Proschan (1985)

- bivariate right censoring
  Tsai, van der Laan, Pruitt
- left truncation and interval censoring
  Chappell and Pan (1999)

2. Counterexamples: MLE s are not always consistent

Counterexample 1. (Ferguson, 1982).
Suppose that $X_1, \ldots, X_n$ are i.i.d. with density p_{θ_0} where

$$p_\theta(x) = (1-\theta)\frac{1}{\delta(\theta)} f_0\left(\frac{x-\theta}{\delta(\theta)}\right) + \theta f_1(x)$$

for $\theta \in [0, 1]$ where

$$f_1(x) = \frac{1}{2} 1_{[-1,1]}(x) \qquad \text{Uniform}[-1, 1],$$

$$f_0(x) = (1 - |x|) 1_{[-1,1]}(x) \qquad \text{Triangular}[-1, 1]$$

and $\delta(\theta)$ satisfies:

- $\delta(0) = 1$
- $0 < \delta(\theta) \leq 1 - \theta$
- $\delta(\theta) \to 0$ as $\theta \to 1$.

Ferguson (1982) shows that $\theta_n \rightarrow_{a.s.} 1$ no matter what θ_0 is true if $\delta(\theta) \rightarrow 0$ ``fast enough . In fact, the assertion is true if

$$\delta(\theta) = (1-\theta)\exp(-(1-\theta)^{-c}+1)$$

with $c > 2$. (Ferguson shows that $c = 4$ works.) If $c = 2$, Ferguson s argument shows that

$$\begin{aligned} \sup_{0\le\theta\le 1} & \; n^{-1}\log L_n(\theta) \\ \geq & \;\; \frac{n-1}{n}\log(M_n/2) + \frac{1}{n}\log\frac{1-M_n}{\delta(M_n)} \\ \rightarrow_d & \;\; \mathbb{D} \end{aligned}$$

where

$$P(\mathbb{D} \le y) = \exp\left(-\frac{1}{2(y-\log 2)}\right), \quad y \ge \log(2)\,.$$

That is

$$\mathbb{D} \stackrel{d}{=} \log 2 + \frac{1}{2E}$$

where E is an Exponential(1) random variable.

Counterexample 2. (4 B s, 1972). A distribution F on $[0, b)$ is star-shaped if $F(x)/x$ is non-decreasing on $[0, b)$. Thus if F has a density f which is increasing on $[0, b)$ then F is star-shaped. Let $\mathcal{F}_{star}$ be the class of all star-shaped distributions on $[0, b)$ for some b. Suppose that $X_1, \ldots, X_n$ are i.i.d. $F \in \mathcal{F}_{star}$. It is shown by Barlow, Bartholomew, Bremner, and Brunk (1972) that the MLE of a star-shaped distribution function F is

$$F_n(x) = \begin{cases} 0, & x < X_{(1)} \\ \frac{ix}{nX_{(n)}}, & X_{(i)} \le x < X_{(i+1)}, \quad i = 1, \ldots, n-1, \\ 1, & x \ge X_{(n)}. \end{cases}$$

Moreover, BBBB (1972) show that if $F(x) = x$ for $0 \le x \le 1$, then

$$F_n(x) \rightarrow_{a.s.} x^2 \neq x$$

for $0 \le x \le 1$.

Note 1. Since $X_{(i)} \stackrel{d}{=} S_j/S_{n+1}$ where $S_i = \sum_{j=1}^{i} E_j$ with E_j i.i.d. Exponential(1) rv s, the total mass at order statistics equals

$$
\begin{aligned}
\frac{1}{nX_{(n)}} \sum_{i=1}^{n} X_{(i)} &\stackrel{d}{=} \frac{1}{S_n} \sum_{i=1}^{n} S_i, \\
&= \frac{n}{S_n} \frac{1}{n} \sum_{j=1}^{n} \left(1 - \frac{j-1}{n}\right) E_j \\
&\to_p 1 \cdot \int_0^1 (1-t)dt = 1/2 \,.
\end{aligned}
$$

Note 2. BBBB (1972) present consistent estimators of F star-shaped via isotonization due to Barlow and Scheurer (1971) and van Zwet.

Counterexample 3. (Boyles, Marshall, Proschan (1985). A distribution F on $[0, \infty)$ is Increasing Failure Rate Average if

$$\frac{1}{x}\{-\log(1 - F(x))\} \equiv \frac{1}{x}\Lambda(x)$$

is non-decreasing; that is, if Λ is star-shaped. Let $\mathcal{F}_{IFRA}$ be the class of all IFRA-distributions on $[0, \infty)$. Suppose that $X_1, \ldots, X_n$ are i.i.d. $F \in \mathcal{F}_{IFRA}$.
It is shown by Boyles, Marshall, and Proschan (1985) that the MLE F_n of a IFRA-distribution function F is given by

$$-\log(1 - F_n(x)) = \begin{cases} \lambda_j, & X_{(j)} \le x < X_{(j+1)}, \\ & \quad j = 0, \ldots, n-1 \\ \infty, & x > X_{(n)} \end{cases}$$

where

$$\lambda_j = \sum_{i=1}^{j} X_{(i)}^{-1} \log\left(\frac{\sum_{k=i}^{n} X_{(k)}}{\sum_{k=i+1}^{n} X_{(k)}}\right).$$

Moreover, BMP (1985) show that if F is exponential(1), then

$$1 - F_n(x) \to_{a.s.} (1 + x)^{-x} \ne \exp(-x), \quad \text{so}$$
$$\frac{1}{x}\Lambda_n(x) \to_{a.s.} \log(1 + x) \ne 1\,.$$

More counterexamples:

- bivariate right censoring
  Tsai, van der Laan, Pruitt
- left truncation and interval censoring
  Chappell and Pan (1999)
- Possible counterexample?
  bivariate interval censoring
  with a continuous mark
  Hudgens, Maathuis, and Gilbert (2005)

3. Beyond consistency: rates and distributions

Le Cam (1973); Birgé (1983): optimal rate of convergence $r_n = r_n^{opt}$ determined by

$$n r_n^{-2} = \log N_{[\,]}(1/r_n, \mathcal{P}) \tag{1}$$

If

$$\log N_{[\,]}(\epsilon, \mathcal{P}) \asymp \frac{K}{\epsilon^{1/\gamma}} \tag{2}$$

(1) leads to the optimal rate of convergence

$$r_n^{opt} = n^{\gamma/(2\gamma+1)} .$$

On the other hand, the bounds (from Birgé and Massart (1993)), yield achieved rates of convergence for maximum likelihood estimators (and other minimum contrast estimators) $r_n = r_n^{ach}$ determined by

$$\sqrt{n} r_n^{-2} = \int_{c r_n^{-2}}^{r_n^{-1}} \sqrt{\log N_{[\,]}(\epsilon, \mathcal{P})} d\epsilon$$

and if (2) holds, this leads to the rate

$$\begin{cases} n^{\gamma/(2\gamma+1)} & \text{if } \gamma > 1/2 \\ n^{\gamma/2} & \text{if } \gamma < 1/2\,. \end{cases}$$

Thus there is the possibility that maximum likelihood is not (rate-)optimal when $\gamma < 1/2$. Since typically

$$\frac{1}{\gamma} = \frac{d}{\alpha}$$

where d is the dimension of the underlying sample space and α is a measure of the ``smoothness of the functions in $\mathcal{P}$,

$$\alpha < \frac{d}{2}$$

leads to $\gamma < 1/2$.

Many examples with $\gamma > 1/2$!

4. Positive Examples (some still in progress!)

Further Examples:

- Interval censoring (Groeneboom)
  case 1, current status data
  case 2 (Groeneboom)
- panel count data
  (Wellner and Zhang, 2000)
- $k-$monotone densities
  (Balabdaoui and Wellner, 2004)
- competing risks current status data
  (Jewell and van der Laan; Maathuis)
- monotone densities in $\mathbb{R}^d$
  (Polonik; Biau and Devroye)

Example 1. (interval censoring)

Case 1: (van de Geer, 1993).

$Y \sim F$, $T \sim G$ independent

Observe $X = (1\{Y \le T\}, T) \equiv (\Delta, T)$.

Goal: estimate F.   MLE F_n exists

Global rate: $d = 1$, $\alpha = 1$, $\gamma = \alpha/d = 1$.

$\gamma/(2\gamma + 1) = 1/3$, so $r_n = n^{1/3}$:

$$n^{1/3} h(p_{F_n}, p_0) = O_p(1)$$

and this yields

$$n^{1/3} \int |F_n - F_0| dG = O_p(1)\,.$$

Local rate: (Groeneboom, 1987)

$$n^{1/3}(F_n(t_0) - F(t_0))$$
$$\to_d \left\{ \frac{F(t_0)(1 - F(t_0)) f_0(t_0)}{2g(t_0)} \right\}^{1/3} 2\mathbb{Z}$$

where $\mathbb{Z} = \text{argmin}\{W(t) + t^2\}$

Case 2: $Y \sim F$, $(U, V) \sim H$, $U \leq V$ independent of Y

Observe i.i.d. copies of $X = (\Delta, U, V)$ where

$$\begin{aligned} \Delta &= (\Delta_1, \Delta_2, \Delta_3) \\ &= (1\{Y \leq U\}, 1\{U < Y \leq V\}, 1\{V < Y\}) \end{aligned}$$

Goal: estimate F. MLE F_n exists.

Global rate (separated case): If $P(V - U \geq \epsilon) = 1$ $d = 1$, $\alpha = 1$, $\gamma = \alpha/d = 1$ $\gamma/(2\gamma + 1) = 1/3$, so $r_n = n^{1/3}$

$$n^{1/3} h(p_{F_n}, p_0) = O_p(1)$$

and this yields

$$n^{1/3} \int |F_n - F_0| d\mu = O_p(1)$$

where

$$\mu(A) = P(U \in A) + P(V \in A), \quad A \in \mathcal{B}_1$$

Global rate (nonseparated case): (van de Geer, 1993).

$$\frac{n^{1/3}}{(\log n)^{1/6}} h(p_{F_n}, p_0) = O_p(1)\,.$$

Although this looks ``worse in terms of the rate, it is actually better because the Hellinger metric is much stronger in this case.

Local rate (separated case):
(Groeneboom, 1996)

$$n^{1/3}(F_n(t_0) - F_0(t_0)) \to_d \left\{ \frac{f_0(t_0)}{2a(t_0)} \right\}^{1/3} 2\mathbb{Z}$$

where $\mathbb{Z} = \text{argmin}\{W(t) + t^2\}$ and

$$\begin{aligned} a(t_0) &= \frac{h_1(t_0)}{F_0(t_0)} + k_1(t_0) \\ &\quad + k_2(t_0) + \frac{h_2(t_0)}{1 - F_0(t_0)} \end{aligned}$$

with

$$k_1(u) = \int_u^M \frac{h(u,v)}{F_0(v) - F_0(u)} dv$$

$$k_2(v) = \int_0^v \frac{h(u,v)}{F_0(v) - F_0(u)} du$$

Local rate (non-separated case):
(conjectured, G&W, 1992)

$$(n \log n)^{1/3}(F_n(t_0) - F_0(t_0)) \to_d \left\{ \frac{3}{4} \frac{f_0(t_0)^2}{h(t_0, t_0)} \right\}^{1/3} 2\mathbb{Z}$$

where $\mathbb{Z} = \mathrm{argmin}\{W(t) + t^2\}$

Monte-Carlo evidence in support:
Groeneboom and Ketelaars (2005)

Example 2. (k-monotone densities)
A density p on $(0, \infty)$ is $k-$monontone if it is non-negative and nonincreasing when $k = 1$; and if $(-1)^j p^{(j)}(x) \geq 0$ for $j = 0, \ldots, k-2$ and $(-1)p^{(k-2)}$ is convex for $k \geq 2$. Let $\mathcal{D}_k$ the collection of all $k-$monotone densities.

Mixture representation: $p \in \mathcal{D}_k$ iff

$$p(x) = \int_0^\infty \frac{k}{y^k}(y-x)_+^{k-1} dF(y)$$

for some distribution function F on $(0, \infty)$.

$k = 1$: monotone decreasing densities on $\mathbb{R}^+$
$k = 2$: convex decreasing densities on $\mathbb{R}^+$
$k \geq 3$: $\ldots$
$k = \infty$: completely monotone densities
$=$ scale mixtures of exponential

The MLE p_n of $p_0 \in \mathcal{D}_k$ exists and is characterized by

$$\int_0^\infty \frac{k}{y^k} \frac{(y-x)_+^k}{p_n(x)} d\mathbb{P}_n(x) \begin{cases} \le 1, & \text{for all } \ y \ge 0 \\ = 1, & \text{if } \ (-1)^k p_n^{(k-1)}(y-) > p_n^{(k-1)}(y+) \, . \end{cases}$$

$k = 1$; **Grenander estimator:**

$$r_n = n^{1/3}$$

- Global rates and finite n minimax bounds:
  Birgé (1986), (1987), (1989)
- Local rates:
  Prakasa Rao (1969)
  Groeneboom (1985), (1989)
  Kim and Pollard (1990)

$$n^{1/3}(p_n(t_0) - p_0(t_0)) \to_d \left\{ \frac{p_0(t_0)|p_0'(t_0)|}{2} \right\}^{1/3} 2\mathbb{Z}$$

$k = 2$**; convex decreasing density**

$d = 1$, $\alpha = 2$, $\gamma = 2$, $\gamma/(2\gamma + 1) = 2/5$, so $r_n = n^{2/5}$ (forward problem)

- Global rates: nothing yet
- Local rates and distributions:
  Groeneboom, Jongbloed, Wellner (2001)

$k \geq 3$**; k-monotone density**

$d = 1$, $\alpha = k$, $\gamma = k$, $\gamma/(2\gamma + 1) = k/(2k + 1)$, so $r_n = n^{k/(2k+1)}$ (forward problem)?

- Global rates: nothing yet
- Local rates: should be $r_n = n^{k/(2k+1)}$
  progress: Balabdaoui and Wellner (2004)
  local rate is true if a certain conjecture about Hermite interpolation holds

Example 3. (Competing risks with current status data)

Variables of interest (X, Y);

$X =$ failure time; $Y =$ failure cause

$X \in \mathbb{R}^+$, $Y \in \{1, \ldots, K\}$

$T =$ an *observation time*, independent of (X, Y)

Observe: (Δ, T), $\Delta = (\Delta_1, \ldots, \Delta_K, \Delta_{K+1})$ where

$$\Delta_j = 1\{X \le T, Y = j\}, \quad j = 1, \ldots, K$$
$$\Delta_{K+1} = 1\{X > T\}.$$

Goal: estimate $F_j(t) = P(X \le t, Y = j)$

MLE $F_n = (F_{n,1}, \ldots, F_{n,K})$ exists!

Characterization of F_n involves an *interacting system* of slopes of convex minorants

• Global rates. Easy with present methods.

$$n^{1/3} \sum_{k=1}^{K} \int |F_{n,k}(t) - F_{0,k}(t)| dG(t) = O_p(1)$$

• Local rates? Conjecture $r_n = n^{1/3}$
Tricky. Maathuis (2006?)

• Limit distribution theory: will involve slopes of an interacting system of greatest convex minorants Defined in terms of a vector of dependent two-sided Brownian motions

Example 4. (Monotone densities in $\mathbb{R}^d$)
$\alpha = 1$, d, $\gamma = 1/d$, so $\gamma/(2\gamma+1) = 1/(d+2)$
Proofs for entropy results?
Biau and Devroye (2003) using Assouad and direct calculations:

$$r_n^{opt} = n^{1/(2+d)}$$

plus optimal constant of order $S^{d/(d+2)}$ with $S \equiv \log(1+B)$ where $\mathcal{P}_B$ is the family of all coordinate-wise decreasing densities with uniform bound B.

Rate achieved by the MLE:

Natural conjecture:

$$r_n^{ach} = n^{1/2d}, \quad d > 2$$

Biau and Devroye (2003) construct generalizations of Birgé s (1987) histogram estimators that achieve the optimal rate for all $d \geq 2$.

5. Problems and Challenges

- More tools for local rates and distribution theory? Comparison methods?

- Under what additional hypotheses are MLE s globally rate optimal in the case $\gamma > 1/2$?

- More counterexamples to clarify when MLE s do not work?

- What is the limit distribution for interval censoring, case 2? (Does the G&W (1992) conjecture hold?)

- When the MLE is not rate optimal, is it still preferable from some other perspectives? For example, does the MLE provide efficient estimators of smooth functionals (while alternative rate -optimal estimators fail to have this property)? Compare with Bickel and Ritov (2003).

- More rate and optimality theory for Maximum Likelihood Estimators of mixing distributions in mixture models with smooth kernels: e.g. completely monotone densities (scale mixtures of exponential), normal location mixtures (deconvolution problems)

- Stable and efficient algorithms for computing MLE s in models where they exist (e.g. mixture models, missing data).

Selected References

- Bahadur, R. R. (1958). Examples of inconsistency of maximum likelihood estimates. *Sankhya, Ser. A* **20**, 207 - 210.

- Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972). *Statistical Inference under Order Restrictions.* Wiley, New York.

- Barlow, R. E. and Scheurer, E. M (1971). Estimation from accelerated life tests. *Technometrics* **13**, 145 - 159.

- Biau, G. and Devroye, L. (2003). On the risk of estimates for block decreasing densities. *J. Mult. Anal.* **86**, 143 - 165.

- Bickel, P. J. and Ritov, Y. (2003). Nonparametric estimators which can be ``plugged-in .*Ann. Statist.* **31**, 1033 - 1053.

- Birgé, L. (1987). On the risk of histograms for estimating decreasing densities. *Ann. Statist.* **15**, 1013 - 1022.

- Birgé, L. (1989). The Grenander estimator: a nonasymptotic approach. *Ann. Statist.* **17**, 1532-1549.

- Birgé, L. and Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Relat. Fields* **97**, 113 - 150.

- Boyles, R. A., Marshall, A. W., Proschan, F. (1985). Inconsistency of the maximum likelihood estimator of a distribution having increasing failure rate average. *Ann. Statist.* **13**, 413-417.

- Ferguson, T. S. (1982). An inconsistent maximum likelihood estimate. *J. Amer. Statist. Assoc.* **77**, 831--834.

- Ghosh, M. (1995). Inconsistent maximum likelihood estimators for the Rasch model. *Statist. Probab. Lett.* **23**, 165-170.

- Hudgens, M., Maathuis, M., and Gilbert, P. (2005). Nonparametric estimation of the joint distribution of a survival time subject to interval censoring and a continuous mark variable. Manuscript in progress.

- Le Cam, L. (1990). Maximum likelihood: an introduction. *Internat. Statist. Rev.* **58**, 153 - 171.

- Pan, W. and Chappell, R. (1999). A note on inconsistency of NPMLE of the distribution function from left truncated and case I interval censored data. *Lifetime Data Anal.* **5**, 281-291.

- Pan, Wei; Chappell, Rick; Kosorok, Michael R. On consistency of the monotone MLE of survival for left truncated and interval-censored data. *Statist. Probab. Lett.* **38**, 49-57.