Estimation and Testing with Interval-Censored Data

Jon A. Wellner

University of Washington

based on joint work with

Moulinath Banerjee
University of Michigan

Talk at OSU, February 18, 2002

Email: jaw@stat.washington.edu

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

Outline

1. **Introduction: interval censored data**
2. **Estimation of F**
3. **The likelihood ratio test of**
   $H : F(t_0) = \theta_0$
4. **How big is "too big"?**

   The Limit Gaussian problem.
5. **Confidence intervals for $F(t_0)$**
6. **Further Problems**

1. Introduction: interval censored data

Example: Current status data

$$X \sim F, \qquad Y \sim G \qquad X, Y \text{ independent}$$

We observe $(Y, 1\{X \leq Y\}) \equiv (Y, \Delta)$ with density

$$p_F(y, \delta) = F(y)^{\delta}(1 - F(y))^{\delta} g(y)\,.$$

Suppose that (Y_i, Δ_i) are i.i.d. as (Y, Δ).

$$L_n(F) = \prod_{i=1}^{n} F(Y_i)^{\Delta_i}(1 - F(Y_i))^{1-\Delta_i}\,.$$

2. Estimation of F: Nonparametric MLE

$$\widehat{F}_n(t) = \text{argmax}_F L_n(F)\,.$$

Another description: define

$$\mathbb{G}_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{[Y_i \le t]}, \qquad \mathbb{V}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \Delta_i 1_{[Y_i \le t]}\,.$$

Note that

$$\mathbb{G}_n(t) \to_{a.s.} G(t),$$
$$\mathbb{V}_n(t) \to_{a.s.} \int_0^t F(y)dG(y) \equiv V(t)\,.$$

Thus

$$\frac{dV}{dG}(t) = F(t)\,.$$

Let $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$.
The *partial sum diagram* $\mathcal{P} = \{P_i\}$ is given by

$$P_i = (\mathbb{G}_n(Y_{(i)}), \mathbb{V}_n(Y_{(i)})), \quad i = 1, \ldots, n\,.$$

The Nonparametric MLE $\widehat{F}_n$ of F is:

$$\widehat{F}_n(Y_{(i)}) = \text{left derivative of the Greatest Convex Minorant of } \mathcal{P} \text{ at } Y_{(i)}\,.$$

From here on:

Greatest Convex Minorant $=$ GCM

3. The likelihood ratio test of $H : F(t_0) = \theta_0$

- **The likelihood ratio statistic:**

$$\lambda_n = \frac{\sup_F L_n(F)}{\sup_{F:F(t_0)=\theta_0} L_n(F)} = \frac{L_n(\widehat{F}_n)}{L_n(\widehat{F}_n^0)}.$$

- **The constrained MLE $\widehat{F}_n^0$. Recipe:**

A. Break $\mathcal{P}$ into $\mathcal{P}_L$ and $\mathcal{P}_R$ where

$\mathcal{P}_L = \{P_i : \quad Y_{(i)} \le t_0\}, \quad \mathcal{P}_R = \{P_i : \quad Y_{(i)} > t_0\}$.

B. Form the GCM's of $\mathcal{P}_L$ and $\mathcal{P}_R$, say $\tilde{\mathbb{V}}_n^L$ and $\tilde{\mathbb{V}}_n^R$.

C. If the slope of $\tilde{\mathbb{V}}_n^L$ exceeds θ_0, replace it by θ_0; if the slope of $\tilde{\mathbb{V}}_n^R$ drops below θ_0, replace it by θ_0.

D. The resulting (truncated or constrained) slope process yields the constrained MLE $\widehat{F}_n^0$.

$$\lambda_n = \frac{\sup_F L_n(F)}{\sup_{F:F(t_0)=\theta_0} L_n(F)} = \frac{L_n(\widehat{F}_n)}{L_n(\widehat{F}_n^0)}\,.$$

When $H : F(t_0) = \theta_0$ holds, does

$$2 \log \lambda_n \to_d \text{something?}$$

Answer: Yes! Banerjee and Wellner (2001)

4. How big is "too big"?

- **The limiting Gaussian problem**

Suppose that we observe $\{X(t) : t \in R\}$ where

$$X(t) = F(t) + \sigma W(t)$$

where $F(t) = \int_{-\infty}^{t} f(s)ds$, f monotone non-decreasing, and W is standard Browian motion. Suppose that we want to estimate the monotone function f. Equivalently

$$dX(t) = f(t) + \sigma dW(t) \, .$$

The "canonical monotone function" is a linear one, and we can change σ to 1 by virtue of scaling arguments so the "canonical" version of the problem is as follows: if

$$dX(t) = 2t + dW(t) \, ,$$

“estimate” $2t$ when $\{X(t) : \ t \in R\}$, is observed. Thus

$$X(t) = t^2 + W(t) \, .$$

“Estimator”: Slope of GCM of $X(t)$. Call this process of slopes of the GCM $\mathbb{S}$.

What is the "canonical constrained problem"?

Estimate the monontone function $f(t) = 2t$ subject to the constraint that $f(0) = 0$" when $\{X(t) : t \in R\}$ is observed.

What is the "constrained estimator"?
Recipe:

A. Break $\{X(t) : t \in R\}$ into
$X^L \equiv \{X(t) : t < 0\}$ and
$X^R \equiv \{X(t) : t \geq 0\}$.

B. Form the GCM's of X^L and X^R say Y^L and Y^R.

C. If the slope of Y^L exceeds 0, replace it by 0; if the slope of Y^R drops below 0, replace it by 0.

D. The resulting (truncated or constrained) slope process $\mathbb{S}^0$ is the constrained MLE of $f(t) = 2t$ in the Gaussian problem.

• **Limit distributions for $\widehat{F}_n$ and $\widehat{F}_n^0$:** Set

$$\begin{aligned}
\mathbb{G}_n^{loc}(t,h) &= n^{1/3}(\mathbb{G}_n(t+n^{-1/3}h) - \mathbb{G}_n(t)) \\
\mathbb{V}_n^{loc}(t,h) &= n^{1/3}\Big\{n^{1/3}(\mathbb{V}_n(t+n^{-1/3}h) - \mathbb{V}_n(t)) \\
&\qquad - \mathbb{G}_n^{loc}(t,h)F(t)\Big\} .
\end{aligned}$$

Theorem 1. If $g(t_0) = G'(t_0)$ and $f(t_0) = F'(t_0)$ exist, then:
A. $\mathbb{G}_n^{loc}(t_0,h) \to_p g(t_0)h$.
B. $\mathbb{V}_n^{loc}(t_0,h) \Rightarrow aW(h) + bh^2$ where $a = \sqrt{F(t_0)(1-F(t_0))g(t_0)}$, $b = f(t_0)g(t_0)/2$, and W is a two-sided Brownian motion starting from 0.

Now define

$$\begin{aligned}
\mathbb{Z}_n(h) &= n^{1/3}(\widehat{F}_n(t_0 + hn^{-1/3}) - F(t_0)) , \\
\mathbb{Z}_n^0(h) &= n^{1/3}(\widehat{F}_n^0(t_0 + hn^{-1/3}) - F(t_0)) .
\end{aligned}$$

Theorem 2. If the hypotheses of Theorem 1 hold with $f(t_0) > 0$, $g(t_0) > 0$, and $F(t_0) = \theta_0$, then

$$(\mathbb{Z}_n(h), \mathbb{Z}_n^0(h)) \Rightarrow (\mathbb{S}_{a,b}(h), \mathbb{S}_{a,b}^0(h))/g(t_0)$$

where $\mathbb{S}_{a,b}$ and $\mathbb{S}_{a,b}^0$ are the constrained and unconstrained slope processes corresponding to $X_{a,b}(h) = aW(h) + bh^2$.

- **Limit distributions for $2\log\lambda_n$**

Theorem 3. (Banerjee and Wellner, 2001). Suppose that F and G have densities f and g which are strictly positive and continuous in a neighborhood in a neighborhood of t_0. Suppose that $F(t_0) = \theta_0$. Then

$$\begin{aligned} 2\log\lambda_n \ &\to_d \ \frac{1}{g(t_0)a^2}\int ((\mathbb{S}_{a,b}(z))^2 - (\mathbb{S}_{a,b}^0(z))^2)\, dz \\ &=_d \ \int \{(\mathbb{S}(z))^2 - (\mathbb{S}^0(z))^2\} dz \equiv \mathbb{D}, \end{aligned}$$

and the distribution of $\mathbb{D}$ is **universal** (free of parameters).

5. Confidence intervals for $F(t_0)$

- **Wald-type intervals**

$$
\begin{aligned}
\mathbb{Z}_n(0) &= n^{1/3}(\widehat{F}_n(t_0) - F(t_0)) \\
&\rightarrow_d \mathbb{S}_{a,b}(0)/g(t_0) \\
&=_d \left\{ \frac{F(t_0)(1-F(t_0))f(t_0)}{2g(t_0)} \right\}^{1/3} \mathbb{S}(0) \\
&\equiv C(F,f,g)\,\mathbb{S}(0)
\end{aligned}
$$

where

$$\mathbb{S}(0) =_d 2\,\mathbb{Z} \equiv 2\mathrm{argmin}(W(h) + h^2)\,.$$

Wald - interval:

$$\widehat{F}_n(t_0) \pm n^{-1/3} C(\widehat{F}_n, \widehat{f}_n, \widehat{g}_n)\, t_\alpha$$

where $\widehat{f}_n$ and $\widehat{g}_n$ are estimates of f and g (at t_0), and $t_{\alpha/2}$ satisfies

$$P(2\mathbb{Z} > t_{\alpha/2}) = \alpha/2\,.$$

Problem: this involves smoothing to get $\widehat{f}_n$, $\widehat{g}_n$!

- **Confidence Intervals from the LR test**

Invert the test:

$$\{\theta : 2 \log \lambda_n(\theta) \le d_\alpha\}.$$

Advantage: No smoothing needed!
Tradeoff: Need to compute constrained estimator(s) $\widehat{F}_n^0$ of F for many different values of the constraint θ.

6. Further Problems

A. Can we find the distribution of $\mathbb{D}$ analytically?

B. How can we prove that the same limit $\mathbb{D}$ arises as the limit distribution for the likelihood ratio test for a large class of such problems involving monotone functions? (Yes!)

C. What is the appropriate contiguity theory? What is the limit distribution of the likelihood ratio statistic under local alternatives? Done! Banerjee and Wellner (2001, 2002).

D. What happens if we constrain at $k > 1$ points?

E. Can we use a union-intersection test to obtain confidence bands for the whole monotone function F?

F. Related problems for estimating a *convex* function?

G. Related problems for interval censoring in R^2?

Selected References

- Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., and Silverman, E. (1955). An empirical distribution function for sampling with incomplete observations. *Ann. Math. Statist.* **26**, 641 - 647.

- Banerjee, M. and Wellner, J. A. (2001). Likelihood ratio tests for monotone functions. *Ann. Statist.* **29**, 1699 - 1731.

- Groeneboom, P. and Wellner, J. A. (2000). Computing Chernoff's distribution. *J. Computational and Graphical Statistics*, to appear.

- Wellner, J. A. (2000). Gaussian white noise models: some results for monotone functions. *Technical Report*, Department of Statistics, University of Washington.