Empirical likelihood based survival function estimation for current status data

(work in progress)

Ian McKeague
Department of Statistics
Florida State University

Coauthors:

Nils Hjort, University of Oslo
Ingrid Van Keilegom, Université catholique de Louvain

Outline

- Review of standard empirical likelihood (EL)
- EL with plug-in
- EL under slower-than-root-n asymptotics (with plug-in)
- Survival function estimation for current status data

Standard empirical likelihood

Nonparametric likelihood

For $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$ in $\mathbb{R}^d$

$$L(F) = \prod_{i=1}^{n} F\{X_i\}$$

EL ratio

$$\tilde{R}(F) = \frac{L(F)}{L(F_n)} = \prod_{i=1}^{n} np_i$$

where (part of) the mass on X_i is $p_i \geq 0$, $\sum_{i=1}^{n} p_i = 1$.

EL function

$$\begin{aligned} R(\theta_0) &= \sup\{\tilde{R}(F) : \theta(F) = \theta_0\} \\ &= \frac{\sup\{L(F) : \theta(F) = \theta_0\}}{\sup\{L(F)\}} \end{aligned}$$

EL hypothesis tests

Accept $\theta(F) = \theta_0$ when $R(\theta_0) \geq r_0$ for some threshold r_0.

EL confidence regions

$$\{\theta : R(\theta) \geq r_0\}$$

with r_0 chosen via an EL analogue of Wilks's theorem.

EL for means

$\mu = E(X) \in \mathbb{R}^d$

$$R(\mu) = \max\left\{\prod_{i=1}^{n} np_i : \sum_{i=1}^{n} p_i X_i = \mu, p_i \geq 0, \sum_{i=1}^{n} p_i = 1\right\}$$

Maximize

$$\sum_{i=1}^{n} \log(np_i)$$

under the constraints:

$$n\sum_{i=1}^{n} p_i(X_i - \mu) = 0, \quad 1 - \sum_{i=1}^{n} p_i = 0$$

Write

$$G = \sum_{i=1}^{n} \log(np_i) - n\lambda \sum_{i=1}^{n} p_i(X_i - \mu) - \gamma \left(1 - \sum_{i=1}^{n} p_i\right)$$

λ and γ are Lagrange multipliers.

$$\frac{\partial G}{\partial p_i} = \frac{1}{p_i} - n\lambda(X_i - \mu) + \gamma = 0$$

so

$$0 = \sum_{i=1}^{n} p_i \frac{\partial G}{\partial p_i} = n + \gamma$$

giving $\gamma = -n$. Thus

$$p_i = \frac{1}{n} \frac{1}{1 + \lambda(X_i - \mu)}$$

Plugging this back into the constraint:

$$g(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \frac{X_i - \mu}{1 + \lambda(X_i - \mu)} = 0$$

This equation has a unique solution for $\lambda = \lambda(\mu)$.

Theorem (ELT, Owen 1990) If X has finite mean μ_0 and finite covariance matrix of rank $q > 0$, then

$$-2 \log R(\mu_0) \xrightarrow{\mathcal{D}} \chi^2_q.$$

Sketch of proof Case $d = q = 1$. Note that $g(0) = \bar{X} - \mu_0$. Let $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(X_i - \mu_0)^2$. Taylor expanding g:

$$\begin{aligned} 0 &= g(\lambda) = g(0) + \lambda g'(0) + o_P(n^{-1/2}) \\ &= \bar{X} - \mu_0 - \lambda\hat{\sigma}^2 + o_P(n^{-1/2}) \end{aligned}$$

Thus $\lambda = (\bar{X} - \mu_0)/\hat{\sigma}^2 + o_P(n^{-1/2}) = O_P(n^{-1/2})$. Recall

$$p_i = \frac{1}{n}\frac{1}{1 + \lambda(X_i - \mu_0)}$$

so, using the Taylor expansion $\log(1+x) = x - x^2/2 + O(x^3)$,

$$\begin{aligned}
-2\log R(\mu_0) &= -2\sum_{i=1}^{n}\log(np_i) = 2\sum_{i=1}^{n}\log(1 + \lambda(X_i - \mu_0)) \\
&= 2n\lambda(\bar{X} - \mu_0) - n\lambda^2\hat{\sigma}^2 + o_P(1) \\
&= 2n(\bar{X} - \mu_0)^2/\hat{\sigma}^2 - n(\bar{X} - \mu_0)^2/\hat{\sigma}^2 + o_P(1) \\
&= n(\bar{X} - \mu_0)^2/\hat{\sigma}^2 + o_P(1) \\
&\xrightarrow{\mathcal{D}} \chi^2_1
\end{aligned}$$

EL with estimating equations

Estimating function: $m(x, \theta)$ with $E(m(X, \theta_0)) = 0$; e.g., median, $m(X, \theta) = 1\{X \leq \theta\} - .5$.

$$R(\theta) = \max \left\{ \prod_{i=1}^{n} np_i : \sum_{i=1}^{n} p_i m(X_i, \theta) = 0, p_i \geq 0, \sum_{i=1}^{n} p_i = 1 \right\}$$

Theorem (ELT) If $m(X, \theta_0)$ has a finite covariance matrix of rank $q > 0$, then

$$-2 \log R(\theta_0) \xrightarrow{\mathcal{D}} \chi^2_q.$$

Proof Immediate from basic ELT upon replacing X by $m(X, \theta_0)$.

EL with plug-in

Estimating function $m(x, \theta, h)$ (having q components):

$$E(m(X, \theta_0, h)) = 0$$

now involves a nuisance parameter h estimated by $\hat{h}$.

EL function:

$$R(\theta, \hat{h}) = \max \left\{ \prod_{i=1}^{n} np_i : \sum_{i=1}^{n} p_i m(X_i, \theta, \hat{h}) = 0, p_i \geq 0, \sum_{i=1}^{n} p_i = 1 \right\}$$

Assumptions ($q = 1$)

- $n^{-1/2} \sum_{i=1}^{n} m(X_i, \theta_0, \hat{h}) \xrightarrow{\mathcal{D}} N(0, V_1)$
- $\frac{1}{n} \sum_{i=1}^{n} m^2(X_i, \theta_0, \hat{h}) \xrightarrow{P} V_2$
- $\max_{1 \le i \le n} |m(X_i, \theta_0, \hat{h})| = o_P(n^{1/2})$

Theorem

$$-2 \log R(\theta_0, \hat{h}) \xrightarrow{\mathcal{D}} r\chi_1^2$$

where $r = V_1/V_2$.

For $q \geq 1$ components:

$$-2 \log R(\theta_0, \hat{h}) \xrightarrow{\mathcal{D}} r_1 \chi^2_{1,1} + \ldots + r_q \chi^2_{1,q}$$

$\chi^2_{1,1}, \ldots, \chi^2_{1,q}$ independent χ^2_1 r.v.'s

$r_1, \ldots, r_q$ are the eigenvalues of $V_2^{-1} V_1$

Example: estimating a symmetric cdf F

F assumed symmetric about an unknown location a:

$$\theta_0 = F(x) = 1 - F(2a - x) \quad \text{for } x < a$$

Estimating function (having $q = 2$ components):

$$m(X, \theta, \hat{a}) = \begin{pmatrix} 1(X \leq x) - \theta \\ 1(X > 2\hat{a} - x) - \theta \end{pmatrix}$$

where $\hat{a}$ is the sample median; x is fixed.

Then, provided F is continuous,

$$-2 \log R(\theta_0, \hat{a}) \xrightarrow{\mathcal{D}} \chi^2_2$$

Example: integral of a squared density

Of interest for various problems related to nonparametric density estimation:

$$\theta_0 = \int f_0^2 \, dx$$

Estimating function:

$$m(X, \theta, \hat{f}) = \hat{f}(X) - \theta$$

where $\hat{f}$ is a kernel density estimator having a symmetric kernel and bandwidth $b = b_n$.

Usual estimator of θ_0:

$$\hat{\theta} = n^{-1} \sum_{i=1}^{n} \hat{f}(X_i) = \int \hat{f}\, dF_n$$

EL function:

$$R(\theta, \hat{f}) = \max \left\{ \prod_{i=1}^{n} np_i : \sum_{i=1}^{n} p_i \hat{f}(X_i) = \theta, p_i \geq 0, \sum_{i=1}^{n} p_i = 1 \right\}$$

Our result yields:

$$-2 \log R(\theta_0, \hat{f}) \xrightarrow{\mathcal{D}} 4\chi_1^2$$

Sketch of proof

Checking first assumption: by asymptotic theory for U-statistics,

$$n^{-1/2}\sum_{i=1}^{n} m(X_i, \theta_0, \hat{f}) = n^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{D}} N(0, 4V)$$

provided $nb \to \infty$. Here

$$V = \int (f_0 - \theta_0)^2 f_0 \, dx = \int f_0^3 \, dx - \left(\int f_0^2 \, dx \right)^2$$

is the variance of the limit distribution of

$$n^{-1/2}\sum_{i=1}^{n} m(X_i, \theta_0, f_0)$$

Second assumption:

$$\begin{aligned} n^{-1}\sum_{i=1}^{n} m^2(X_i, \theta_0, \hat{f}) &= n^{-1}\sum_{i=1}^{n}(\hat{f}(X_i) - \theta_0)^2 \\ &= \int \hat{f}^2\, dF_n - 2\theta_0\hat{\theta} + \theta_0^2 \\ &\xrightarrow{P} \int f_0^3\, dx - 2\theta_0^2 + \theta_0^2 \\ &= V \end{aligned}$$

Third assumption:

$$\max_{i \le n} |\hat{f}(X_i) - \theta_0| = O(b^{-1}) = o_P(n^{1/2})$$

provided $n^{1/2}b \to \infty$.

EL under slower-than-root-n asymptotics

Sequence of estimating functions $m_n(x, \theta, \hat{h})$

Assumptions (for rate n^α, $0 < \alpha < 1/2$)

- $n^{\alpha-1} \sum_{i=1}^n m_n(X_i, \theta_0, \hat{h}) \xrightarrow{\mathcal{D}} N(0, V_1)$
- $n^{2(\alpha-1)} \sum_{i=1}^n m_n^2(X_i, \theta_0, \hat{h}) \xrightarrow{P} V_2$
- $\max_{1 \le i \le n} |m_n(X_i, \theta_0, \hat{h})| = O_P(n^\alpha)$

Theorem

$$-2 \log R(\theta_0, \hat{h}) \xrightarrow{\mathcal{D}} r\chi_1^2$$

where $r = V_1/V_2$.

EL for current status data

$T \sim F$ failure time, $S = 1 - F$, pdf f, $\theta_0 = S(t)$
$C \sim G$ check-up time, assumed independent of T, pdf g

Only get to observe $X = (C, \Delta)$ where $\Delta = 1\{T \leq C\}$.

Nonparametric likelihood

$$
\begin{aligned}
L(S) &= \prod_{i=1}^{n} (1 - S(C_i))^{\Delta_i} S(C_i)^{1-\Delta_i} \\
S_n(t) &= \arg\max_S L(S)
\end{aligned}
$$

Groeneboom (1987)

$$n^{1/3}(S_n(t) - S(t)) \xrightarrow{\mathcal{D}} 2c\mathbb{Z}$$

$c = \{F(t)(1 - F(t)f(t)/(2g(t))\}^{1/3}$
$\mathbb{Z} = \operatorname{argmin}(W(s) + s^2)$
W is a two-sided Brownian motion

Banerjee and Wellner (2002)

Found a universal limit law for $-2 \log R(\theta_0)$, where

$$R(\theta) = \frac{\sup\{L(S) : S(t) = \theta\}}{\sup\{L(S)\}}$$

Limit is an integral involving greatest convex minorants of $W(s) + s^2$.

Is there an estimating function version of EL in this setting?

van der Laan and Robins (1998):

Efficient influence curve for the functional

$$\theta_0 = \int_0^\infty k(u)S(u)\,du$$

(estimable at root-n rate) given by

$$\begin{aligned} m(X,\theta,F,g,k) &= \frac{k(C)(1-\Delta)}{g(C)} - \theta - \frac{k(C)(F(C)-1)}{g(C)} \\ &\quad + \int_0^\infty k(u)(1-F(u))\,du \end{aligned}$$

Provides an efficient (plug-in) estimating function $m(X, \theta, \hat{F}, \hat{g}, k)$ when $\hat{F}$ or $\hat{g}$ is consistent

Covariates Z: approach extends to the marginal survival function of T, assuming $T \perp C|Z$.

ELT: Our result for estimating functions with plug-in gives

$$-2 \log R(\theta_0, \hat{F}, \hat{g}, k) \xrightarrow{\mathcal{D}} \chi^2_1$$

van der Laan and van der Vaart (2000)

Asymptotically normal estimator for $\theta_0 = S(t)$:

$$n^{1/3}(\hat{S}(t) - S(t)) \xrightarrow{\mathcal{D}} N(0, \sigma^2)$$

where σ^2 depends on $F(t)$, $g(t)$ and the limits of $\hat{g}(t)$, $\hat{F}(t)$.

Replace k by $k_n = k_{b,t}$, a kernel function of bandwidth $b = b_n = b_1 n^{-1/3}$ centered at t.

Sequence of (plug-in) estimating functions $m(X, \theta, \hat{F}, \hat{g}, k_n)$.

Main assumptions

- $\hat{g}$, $\hat{F}$ belong to classes of functions having uniform entropy of order $(1/\epsilon)^V$, $V < 2$, w.p. tending to 1
- $\hat{g}$ or $\hat{F}$ locally consistent at t.

Note: If $\hat{g}' \to g'$ uniformly in probability, then $\hat{g}$ belongs to the class of Lipschitz functions, w.p. tending to 1.

ELT: Our result for estimating functions (with plug-in) under cube-root asymptotics ($\alpha = 1/3$) gives

$$-2 \log R(S(t), \hat{F}, \hat{g}, k_n) \xrightarrow{\mathcal{D}} \chi^2_1$$

Concluding remarks

- Banerjee–Wellner approach avoids smoothing, but won't allow adjustment for covariate effects.
- Critical values of the limit distribution of Banerjee and Wellner need to be found by via Monte Carlo.
- Simulation studies needed . . .