Empirical Likelihood (EL).

Art Owen (1988, 1990) introduced ↑

This is a general nonparametric methodology for creating likelihood-type inference without specifying a joint distributional form for the data

↳ we won't guess wrong!

→ no likelihood!

EL is going to use the fact that the empirical cdf is a nonparametric MLE to assess how plausible a value of a parameter is to perform inference.

likelihood methods

(motivating).

1 Mean Case

Suppose $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ ($\in \mathbb{R}^q$) are iid with mean $\boldsymbol{\mu}$ ($\in \mathbb{R}^q$) and covariance-variance Σ ($q \times q$). For simplicity, say we are interested in estimating $\boldsymbol{\mu}$.

Imagine assigning probabilities ($p_i \mapsto Y_i$) $p_1, \ldots, p_n$ to the data $Y_1, \ldots, Y_n$ where $0 \le p_i \le 1$ and $\sum_{i=1}^n p_i = 1$. (*)

Unlike parametric likelihood, where we assume a functional form for p_i's, only constraints (*).

Define a multinomial likelihood: $\prod_{i=1}^n p_i$ (likelihood for $Y_1, \ldots, Y_n$ using $p_1, \ldots, p_n$).

Recall from class (Likelihood Notes pg. 10) if you maximize $\prod_{i=1}^n p_i$, the maximizer is $p_1 = p_2 = \ldots = p_n = \frac{1}{n}$ ($\leftarrow$ 1 obs in each class).

We have also seen, that the empirical cdf

$$F_n(y) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(Y_i \le y), \quad y \in \mathbb{R}^q$$

is the MLE (pg. 23 on likelihood notes).

In other words, given the data the empirical cdf maximizes $\prod_{i=1}^n p_i$.

To perform *nonparametric* likelihood inference on $\boldsymbol{\mu}$, we can consider a constrained multinomial likelihood, known as the **Empirical Likelihood function of $\boldsymbol{\mu}$:**

mean of a dsn $(p_1, \dots, p_n)$ on $(Y_1, \dots, Y_n)$

$$L_n(\underline{\mu}) = L_n(\boldsymbol{\mu}|\boldsymbol{Y}) = \sup\left\{\prod_{i=1}^{n} p_i : p_i \mapsto \boldsymbol{Y}_i, \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} \boldsymbol{Y}_i p_i = \boldsymbol{\mu}\right\}.$$

function of μ; $p_i \geq 0$; EL function; multinomial likelihood; mean constraint on $(p_1, \dots, p_n)$

Given a parameter value $\underline{\mu}$ and data $\underline{Y}$, $L_n(\underline{\mu})$ assesses how plausible the value of $\underline{\mu}$ is.

$L_n(\underline{\mu})$ is the largest multinomial likelihood possible for a probability assignment to the data having mean $\underline{\mu}$.

The largest possible value of $L_n(\boldsymbol{\mu}|\boldsymbol{Y})$ is

$$\prod_{i=1}^{n} \frac{1}{n} = L_n(\bar{\underline{Y}}).$$

So $\bar{\underline{Y}} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is a nonparametric ML estimator of $\underline{\mu}$, i.e. the EL estimator $\hat{\underline{\mu}} = \bar{\underline{Y}}$ of $\underline{\mu}$.

2 Statistical Inference

We can form an EL ratio for $\boldsymbol{\mu}$

$$R_n(\boldsymbol{\mu}) = \frac{L_n(\boldsymbol{\mu}|\boldsymbol{Y})}{L_n(\hat{\boldsymbol{\mu}}|\boldsymbol{Y})}$$

$$= \frac{L_n(\underline{\mu}|\underline{Y})}{\prod_{i=1}^{n} \frac{1}{n}} \leftarrow \hat{\mu} = \bar{Y}_n \Rightarrow p_i = \frac{1}{n}$$

$$= n^n L_n(\underline{\mu}|\underline{Y})$$

$$= \sup\left\{ \prod_{i=1}^{n} n p_i : p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} y_i p_i = \underline{\mu} \right\}$$

$$\sum_{i=1}^{n} (y_i - \mu) p_i = \underline{0}_p.$$

Theorem (Wilk's Theorem): If $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n \in \mathbb{R}^q$ are iid with mean $\boldsymbol{\mu}_0$ and covariance-variance Σ where $\text{rank}(\Sigma) = q$, then

$$-2 \log R_n(\boldsymbol{\mu}_0) \xrightarrow{d} \chi^2_q \text{ as } n \to \infty.$$

material for exam ends here.

3 EL with Estimating Equations

For EL inference on $\boldsymbol{\theta} \in \mathbb{R}^b$, we make an EL function

Then we can get a point estimate, EL ratio, and corresponding CIs, as well as “profile” EL:

Theorem: Suppose $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \cdots \in \mathbb{R}^q$ are iid with $\mathrm{E}\boldsymbol{\psi}(\boldsymbol{Y}_1, \boldsymbol{\theta}_0) = \mathbf{0}_r$ and $\mathrm{Var}[\boldsymbol{\psi}(\boldsymbol{Y}_1, \boldsymbol{\theta}_0)] = \mathrm{E}\boldsymbol{\psi}(\boldsymbol{Y}_1, \boldsymbol{\theta}_0)\boldsymbol{\psi}(\boldsymbol{Y}_1, \boldsymbol{\theta}_0)^\top$ is positive definite, where $\boldsymbol{\theta}_0$ denotes the true parameter value.

Suppose also that $\partial\boldsymbol{\psi}(\boldsymbol{y}, \boldsymbol{\theta})/\partial\boldsymbol{\theta}$ and $\partial^2\boldsymbol{\psi}(\boldsymbol{y}, \boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top$ are continuous in a neighborhood of $\boldsymbol{\theta}_0$ and that, in this neighborhood, $||\boldsymbol{\psi}(\boldsymbol{Y}_1, \boldsymbol{\theta})||^3$, $||\partial\boldsymbol{\psi}(\boldsymbol{y}, \boldsymbol{\theta})/\partial\boldsymbol{\theta}||$ and $||\partial^2\boldsymbol{\psi}(\boldsymbol{y}, \boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top||$ are bounded by an integrable function $\Psi(\boldsymbol{Y}_1)$.

Finally, suppose the $r \times b$ matrix $D_\psi \equiv \mathrm{E}\partial\boldsymbol{\psi}(\boldsymbol{y}, \boldsymbol{\theta})/\partial\boldsymbol{\theta}$ has full column rank b.

Then, as $n \to \infty$,

i. $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}_b, V)$, where $V = (D_\psi^\top \mathrm{Var}[\boldsymbol{\psi}(\boldsymbol{Y}_1, \boldsymbol{\theta}_0)] D_\psi)^{-1}$.

ii. If $r > b$, the asymptotic variance V cannot increase if an estimating function is added.

iii. To test $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, we may use $-2\log R_n(\boldsymbol{\theta}_0)$ and when H_0 is true,

$$-2\log R_n(\boldsymbol{\theta}_0) \xrightarrow{d} \chi^2_b$$

iv. If $r > b$, to test $H_0 : \mathrm{E}\psi(\boldsymbol{Y}_1, \boldsymbol{\theta}) = \mathbf{0}_r$ holds for some $\boldsymbol{\theta}$, we may use

$$-2\log \frac{L_n(\hat{\boldsymbol{\theta}})}{\prod_{i=1}^{n}(1/n)} =$$

and when H_0 is true this quantity converges in distribution to χ^2_{r-b}.

v. To test the profile assumption $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^0 \in \mathbb{R}^q$, we can use the profile EL ratio $-2\log R_n(\boldsymbol{\theta}_1^0)$ and , when H_0 is true, $-2\log R_n(\boldsymbol{\theta}_1^0) \xrightarrow{d} \chi^2_q$.

4 Computation

Technically, for a given value of $\boldsymbol{\theta}$, define $L_n(\boldsymbol{\theta}|\boldsymbol{Y}) = 0$ if

$$\mathcal{A}_n(\boldsymbol{\theta}) = \left\{ \prod_{i=1}^{n} p_i : p_i \mapsto \boldsymbol{Y}_i, \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} \boldsymbol{p}_i \boldsymbol{\psi}(\boldsymbol{Y}_i, \boldsymbol{\theta}) = \mathbf{0}_r \right\}$$

is empty.

If $\mathbf{0}_r$ is in the interior convex hull of $\{\psi(\boldsymbol{Y}_i, \boldsymbol{\theta})\}_{i=1}^{n}$, then $\mathcal{A}_n(\boldsymbol{\theta})$ will not be empty.

The supremum in the definition of $L_n(\boldsymbol{\theta}|\boldsymbol{Y})$ looks nasty, but the form simplifies if $L_n(\boldsymbol{\theta}|\boldsymbol{Y}) > 0$ for a given $\boldsymbol{\theta} \in \mathbb{R}^b$. To see this, fix $\boldsymbol{\theta}$ and let

$$\mathcal{B}_n(\boldsymbol{\theta}) = \left\{ (p_1, \ldots, p_n) : p_i \geq 0, \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} \boldsymbol{p}_i \boldsymbol{\psi}(\mathbf{Y}_i, \boldsymbol{\theta}) = \mathbf{0}_r \right\} \subset [0,1]^n$$

To maximize $\prod_{i=1}^{n} p_i$ on $\mathcal{B}_n(\boldsymbol{\theta})$ and find $(p_1^*, \ldots, p_n^*)$, use Lagrange multipliers $a \in \mathbb{R}$ and $\boldsymbol{\lambda} \in \mathbb{R}^r$ and maximize

$$f(p_1, \ldots, p_n, a, \boldsymbol{\lambda}) = \log \prod_{i=1}^{n} p_i + a \left(1 - \sum_{i=1}^{n} p_i \right) - n\boldsymbol{\lambda}^\top \left(\sum_{i=1}^{n} \boldsymbol{p}_i \boldsymbol{\psi}(\mathbf{Y}_i, \boldsymbol{\theta}) \right)$$

over $p_i \in [0,1], a \in \mathbb{R}$, and $\boldsymbol{\lambda} \in \mathbb{R}^r$.

Take derivatives & set to zero: