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 $\overset{\in \mathbb{R}^q}{\boldsymbol{\mu}}$ and covariance-variance $\overset{q \times q}{\Sigma}$. For simplicity, say we are interested in estimating $\boldsymbol{\mu}$.

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

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

$$F_n(y) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(Y_i \le y), \quad y \in \mathbb{R}^q \text{ is the MLE}$$ (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 $(Y_1, \dots, Y_n)$

$$\underbrace{L_n(\underline{\mu}) = L_n(\overset{\text{function of } \mu}{\boldsymbol{\mu}}|\boldsymbol{Y})}_{\text{EL function}} = \sup\left\{\underbrace{\prod_{i=1}^n p_i}_{\text{multinomial likelihood}} : p_i \mapsto \boldsymbol{Y}_i, \quad \sum_{i=1}^n \overset{p_i \geq 0}{p_i} = 1, \quad \underbrace{\sum_{i=1}^n \boldsymbol{Y}_i p_i = \boldsymbol{\mu}}_{\text{mean constraint on } (p_1, \dots, p_n)}\right\}.$$

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 $\bar{\underline{Y}} = \frac{1}{n}\sum_{i=1}^n \underline{Y}_i$ 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 $\boldsymbol{\mu}$

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

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

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

$$= \sup\left\{ \prod_{i=1}^{n} n p_i : p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \underbrace{\sum_{i=1}^{n} Y_i p_i = \underline{\mu}}_{\sum_{i=1}^{n}(Y_i - \mu)p_i = \underline{0}_p} \right\}$$

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.

In other words, for $H_0: \underline{\mu} = \underline{\mu}_0 \in \mathbb{R}^q$, if H_0 is true, $-2\log R_n(\underline{\mu}_0) \xrightarrow{d} \chi^2_q$ as $n \to \infty$.

* EL behave like parametric likelihoods for log ratios! *

So, if $\chi^2_{1-\alpha,q}$ denotes the $1-\alpha$ quantile of χ^2_q, an approximate $100(1-\alpha)\%$ confidence region for $\underline{\mu}$ is

$$CR = \left\{ \underline{\mu} \in \mathbb{R}^q : -2\log R_n(\underline{\mu}) \leq \chi^2_{1-\alpha,q} \right\}$$

By inverting the EL test

$$P(\underline{\mu}_0 \in CR) = P\left(-2\log R_n(\mu_0) \leq \chi^2_{1-\alpha,q}\right) \xrightarrow{\text{as } n\to\infty} P\left(\chi^2_q \leq \chi^2_{1-\alpha,q}\right) = 1-\alpha \,//$$

For proof of this theorem, see Owen (1988).

3 EL with Estimating Equations

(Qin and Lawless, 1994).

Recall:

For $\underline{Y}_1, \dots, \underline{Y}_n$ iid and $\underline{\theta} \in \mathbb{R}^b$ a parameter of interest.

Estimating equations link a data point $\underline{Y}_i$ to parameters through $r \geq p$ functions

$$\underline{\Psi}(\underline{Y}_i, \underline{\theta}) \text{ which satisfies } E\underline{\Psi}(\underline{Y}_i, \underline{\theta}) = \underline{0}_r.$$

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

extends mean example to any other estimating function!

$$L_n(\underline{\theta}) = \sup\left\{ \prod_{i=1}^n p_i : p_i \geq 0, \sum_{i=1}^n p_i = 1, \underbrace{\sum_{i=1}^n p_i \underline{\Psi}(\underline{Y}_i, \underline{\theta}) = \underline{0}_r}_{p_i\text{'s are placed on } \underline{\Psi}(\underline{Y}_i, \underline{\theta}) \text{ to have expectation zero}} \right\}$$

The EL function judges the plausibility of a given value of $\underline{\theta}$ based on data.

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

point estimate: maximize $L_n(\underline{\theta})$ to obtain maximum EL estimator $\hat{\underline{\theta}}$

EL ratio: $R_n(\underline{\theta}) = \dfrac{L_n(\underline{\theta})}{L_n(\hat{\underline{\theta}})}$ (just like parametric likelihood)

confidence region: $CR = \left\{ \underline{\theta} \in \mathbb{R}^b : -2\log R_n(\underline{\theta}) \leq \chi^2_{1-\alpha, q} \right\}$ (invert EL ratio).

Profile EL: Suppose $\underline{\theta} = (\underline{\theta}_1, \underline{\theta}_2)$, $\underline{\theta}_1 \in \mathbb{R}^q$, $\underline{\theta}_2 \in \mathbb{R}^{b-q}$. Given $\underline{\theta}_1$, define $\hat{\underline{\theta}}_{2,\underline{\theta}_1}$ where

$$L_n(\underline{\theta}_1, \hat{\underline{\theta}}_{2,\theta_1}) = \sup_{\underline{\theta}_2} L_n(\underline{\theta}_1, \underline{\theta}_2).$$

Then the profile EL ratio for $\underline{\theta}_1$ is $R_n(\underline{\theta}_1) = \dfrac{L_n(\theta_1, \hat{\theta}_{2,\theta_1})}{L_n(\hat{\theta})}$.

main EL result

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 $V = (D_\psi^\top \mathrm{Var}[\boldsymbol{\psi}(\boldsymbol{Y}_1, \boldsymbol{\theta}_0)] D_\psi)^{-1}$. EL point estimates are asymptotically normal.

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

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,

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

$$R_n(\underline{\theta}_0) = \frac{L_n(\underline{\theta}_0)}{L_n(\hat{\underline{\theta}})}$$

$\Rightarrow$ confidence regions: $CR = \{\underline{\theta} \in \mathbb{R}^b : -2\log R_n(\underline{\theta}) \le \chi^2_{b,1-\alpha}\}$.

iv. $\underbrace{\text{If } r > b}_{\text{more functions than parameters}}$, to test $H_0 : \underbrace{\mathrm{E}\psi(\boldsymbol{Y}_1, \boldsymbol{\theta}) = \mathbf{0}_r}_{\text{moment condition}}$ holds for some $\boldsymbol{\theta}$, we may use

"How compatible is the moment condition?"

$$-2\log \frac{L_n(\hat{\boldsymbol{\theta}})}{\underbrace{\prod_{i=1}^n (1/n)}_{\text{biggest possible value could ever have for an EL function without moment constraints.}}} = -2\log\left(n^n L_n(\hat{\underline{\theta}})\right).$$

and when H_0 is true this quantity converges in distribution to $\chi^2_{\underbrace{r-b}_{\#\text{ excess estimating functions}}}$.

Asymptotically, $-2\log R_n(\underline{\theta}_0)$ and $-2\log(n^n L_n(\hat{\underline{\theta}}))$ are independent.

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, $-2\log R_n(\boldsymbol{\theta}_1^0) \xrightarrow{d} \chi^2_q$.

\# parameters after profiling.

4 Computation

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

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

is empty. (EL function might not be computible over all possible parameter values).

If $\mathcal{A}_n(\underline{\theta})$ is empty, $\mathcal{A}_n(\underline{\theta}) = \phi$, then $L_n(\underline{\theta})$ is not defined.

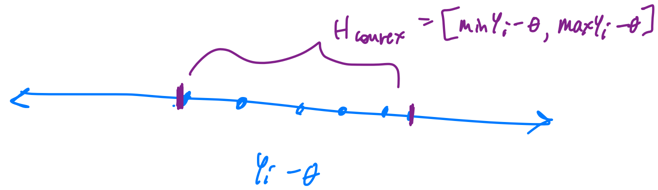
$\Rightarrow$ Define $L_n(\underline{\theta}) = 0$. (this is the smallest it can be anyways).

If $\boldsymbol{0}_r$ 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.

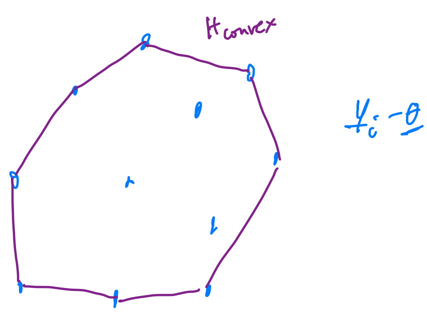
H_{convex} is the smallest convex set containing $\underline{\Psi}(\underline{Y}_1, \underline{\theta}), \ldots, \underline{\Psi}(\underline{Y}_n, \underline{\theta})$.

E.g.: If $\Psi(\underline{Y}_i, \underline{\theta}) = \underline{Y}_i - \underline{\theta}$ (mean case) and

$q = 1$:



$q = 2$:



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

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

Closed and bounded (compact) set in $\mathbb{R}^n$.

$\Rightarrow$ The supremum $L_n(\underline{\theta}) = \sup A_n(\underline{\theta}) = \sup\{\prod_{i=1}^n p_i : (p_1, \ldots, p_n) \in \mathcal{B}_n(\underline{\theta})\}$ is attained as a maximum of $\prod_{i=1}^n p_i$ on $\mathcal{B}(\underline{\theta})$. $\exists\ p_1^*, \ldots, p_n^* \in \mathcal{B}_n(\underline{\theta})$ where $L_n(\underline{\theta}) = \prod_{i=1}^n p_i^*$.

(uniqueness). Suppose $q_1^*, \ldots, q_n^* > 0$ & $p_1^*, \ldots, p_n^* \geq 0$ lie in $\mathcal{B}_n(\underline{\theta})$ and $\prod_{i=1}^n p_i^* = \prod_{i=1}^n q_i^*$.

Now let $r_i^* = \alpha p_i^* + (1-\alpha) q_i^*$ for $\alpha \in [0,1]$.

If $\alpha \in (0,1)$, it holds that $\sum_{i=1}^n \log r_i^* > \alpha \sum_{i=1}^n \log p_i^* + (1-\alpha) \sum_{i=1}^n \log q_i^*$ (Jensen's).

If $\prod_{i=1}^n q_i^* = \prod_{i=1}^n p_i^* = M$ holds, where $M > 0$ is the maximum of $\prod_{i=1}^n p_i$ on $\mathcal{B}_n(\underline{\theta})$.

(contradiction!)

Then $\sum_{i=1}^n \log r_i^* > \sum_{i=1}^n \log p_i^*$ which cannot be true because $\prod_{i=1}^n p_i^*$ is the max of $\prod_{i=1}^n p_i$ on $\mathcal{B}_n(\underline{\theta})$.

$\Rightarrow$ $(p_1^*, \ldots, p_n^*)$ which maximize $\prod_{i=1}^n p_i$ on $\mathcal{B}_n(\underline{\theta})$ must be unique. //

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

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

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

Take derivatives & set to zero:

$$\frac{\partial}{\partial p_i} f(p_1,\dots,p_n,a,\boldsymbol{\lambda}) = \frac{1}{p_i} - a - n\boldsymbol{\lambda}^T\psi(y_i,\boldsymbol{\theta}) \overset{\text{set}}{=} 0 \implies a p_i = 1 - n p_i \boldsymbol{\lambda}^T\psi(y_i,\boldsymbol{\theta})$$

$$\sum_{i=1}^n a p_i = \sum_{i=1}^n \{1 - n p_i \boldsymbol{\lambda}^T \boldsymbol{\psi}(y_i,\boldsymbol{\theta})\}$$

$$\frac{\partial}{\partial a} f(p_1,\dots,p_n,a,\boldsymbol{\lambda}) = \underbrace{1 - \sum_{i=1}^n p_i \overset{\text{set}}{=} 0}$$

$$a = n - n\boldsymbol{\lambda}^T \sum_{i=1}^n p_i \boldsymbol{\psi}(y_i,\boldsymbol{\theta})$$

$$\frac{\partial}{\partial \boldsymbol{\lambda}} f(p_1,\dots,p_n,a,\boldsymbol{\lambda}) = \underbrace{-n\sum_{i=1}^n p_i \boldsymbol{\psi}(y_i,\boldsymbol{\theta}) = 0}$$

$$a = n$$

$$p_i = \frac{1}{a} - \frac{1}{a} n p_i \boldsymbol{\lambda}^T\psi(y_i,\boldsymbol{\theta})$$

$$= \frac{1}{n} - p_i\boldsymbol{\lambda}^T\psi(y_i,\boldsymbol{\theta})$$

$$\Rightarrow p_i\left(1+\boldsymbol{\lambda}^T\boldsymbol{\psi}(y_i,\boldsymbol{\theta})\right) = \frac{1}{n}$$

$$\Rightarrow p_i = \frac{1}{n}\left(\frac{1}{1+\boldsymbol{\lambda}^T\psi(y_i,\boldsymbol{\theta})}\right)$$

convex hull of $\{\psi(y_i,\boldsymbol{\theta})\}_{i=1}^n$

If $\boldsymbol{0}_r \in H_{\text{convex}}$, then

$$L_n(\boldsymbol{\theta}) = \prod_{i=1}^n \frac{1}{n}\left(\frac{1}{1+\boldsymbol{\lambda}^T\boldsymbol{\psi}(y_i,\boldsymbol{\theta})}\right)$$ where

lagrange multiplier.

$\boldsymbol{\lambda}$ is determined by solving

$$\boldsymbol{0} = \sum_{i=1}^n p_i\psi(y_i,\boldsymbol{\theta})$$

$$= \frac{1}{n}\sum_{i=1}^n \frac{\psi(y_i,\theta)}{1+\lambda^T\boldsymbol{\psi}(y_i,\theta)}$$

See Art Owen's website for code/R package.