Empirical Likelihood

Art B. Owen

Department of Statistics
Stanford University

These lectures

I) Basics of empirical likelihood ✓

II) Estimating equations

III) Research frontier

Today: Basics of EL

1) Parametric likelihood
2) Nonparametric likelihood
3) NPMLEs
4) Nonparametric likelihood ratios
5) EL definition
6) EL computation for the mean
7) Statistical properties of EL for the mean
8) Calibration
9) Euclidean likelihood, Renyi-Cressie-Read
10) Biased sampling

Empirical likelihood provides:

- **likelihood** methods for inference, especially
  - tests, and
  - confidence regions,
- **without** assuming a parametric model for data
- **competitive** power even when parametric model holds

Like the bootstrap, but without resampling.

Some good things about EL

1) (correct) data driven shape for confidence sets Hall
2) power optimality of tests Kitamura
3) allows side constraints O (1991), Qin & Lawless (1993)
4) Bartlett correctable DiCiccio, Hall & Romano (1991)
5) extends for
   (a) censoring
   (b) truncation
   (c) biased sampling,
6) methods for
   (a) time series Kitamura
   (b) survey sampling Qin, Chen, Sitter, . . .

Many more extensions S.-X. Chen; Hjort, McKeague & van Keilegom; Lahiri

Parametric likelihoods

Data have **known** distribution f_θ with **unknown** parameter θ

$$\Pr(X_1 = x_1, \dots, X_n = x_n) = f(x_1, \dots, x_n; \theta)$$
$$\Pr(x_1 \le X_1 \le x_1 + \Delta, \dots, x_n \le X_n \le x_n + \Delta) \propto f(x_1, \dots, x_n; \theta)$$

$f(\cdots;\cdot)$ known, $\theta \in \Theta \subseteq \mathbb{R}^p$ unknown

Likelihood function

$$L(\theta) = L(\theta; x_1, \dots, x_n) = f(x_1, \dots, x_n; \theta)$$

“Chance, under θ, of getting the data we did get”

Likelihood examples

$$X_i \sim \text{Poi}(\theta), \quad \theta \ge 0$$

$$L(\theta) = \prod_{i=1}^n \frac{e^{-\theta}\theta^{x_i}}{x_i!}$$

$$Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2) \quad x_i \text{ fixed}$$

$$L(\beta_0, \beta_1, \sigma) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2}$$

Likelihood inference

Maximum likelihood estimate

$$\hat\theta = \arg\max_\theta L(\theta; x_1, \dots, x_n)$$

Likelihood ratio inferences

$$-2\log(L(\theta_0)/L(\hat\theta)) \to \chi^2_{(q)} \qquad \text{Wilks}$$

Typically . . . Neyman-Pearson, Cramer-Rao, . . .

1) $\hat\theta$ asymptotically normal
2) $\hat\theta$ asymptotically efficient
3) Likelihood ratio tests powerful
4) Likelihood ratio confidence regions small

Other likelihood advantages

- can model data distortion: bias, censoring, truncation
- can combine data from different sources
- can factor in prior information
- obey range constraints: MLE of correlation in $[-1, 1]$
- transformation invariance
- data determined shape for $\{\theta \mid L(\theta) \geq rL(\hat{\theta})\}$
- incorporates nuisance parameters

Unfortunately

We might not know a correct $f(\cdots;\theta)$

No reason to expect that new data belong to one of our favorite families

Wrong models sometimes work (e.g. Normal mean via CLT) and sometimes fail (e.g. Normal variance)

Also,

Usually easy to compute $L(\theta)$, but ...

Sometimes hard to find $\hat{\theta}$

Sometimes hard to compute $\max_{\theta_2} L((\theta_1, \theta_2); x_1, \dots, x_n)$ (Profile likelihood)

Nonparametric methods

Assume only $X_i \sim F$ where

- F is continuous, or,
- F is symmetric, or,
- F has a monotone density, or,
- F has log-concave density, or,
- $\cdots$ other believable, but big, family

Nonparametric usually means infinite dimensional parameter

Sometimes lose power (e.g. sign test), sometimes not

Nonparametric maximum likelihood

For X_i IID from F, $\quad L(F) = \prod_{i=1}^{n} F(\{x_i\})$

The NPMLE is $\quad \widehat{F} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$

where δ_x is a point mass at x

Kiefer and Wolfowitz, 1956

Proof

Distinct values z_j appear n_j times in sample, $j = 1, \dots, m$

Let $F(\{z_j\}) = p_j \ge 0$ and $\widehat{F}(\{z_j\}) = \hat{p}_j = n_j/n$ with some $p_j \ne \hat{p}_j$

$$\begin{aligned}\log\left(\frac{L(F)}{L(\widehat{F})}\right) &= \sum_{j=1}^{m} n_j \log\left(\frac{p_j}{\hat{p}_j}\right)\\ &= n\sum_{j=1}^{m} \hat{p}_j \log\left(\frac{p_j}{\hat{p}_j}\right)\\ &< n\sum_{j=1}^{m} \hat{p}_j \left(\frac{p_j}{\hat{p}_j} - 1\right)\\ &= 0. \quad \square\end{aligned}$$

Other NPMLEs

NPMLEs are useful when we want the analogue of the empirical CDF for nonstandard settings.

Kaplan-Meier	Right censored survival times
Lynden-Bell	Left truncated star brightness
Hartley-Rao	Sample survey data
Grenander	Monotone density for actuarial data

Censoring and Truncation

The likelihood can be used to compensate for sampling distortions.

Censoring

X_i only known to be in set C_i. E.g.: patient survived ≥ 438 days.
If observed exactly, then $C_i = \{X_i\}$ others. Conditional on C_i

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

Truncation

X_i only observed if $X_i \in T_i$. E.g.: star only seen if it is bright enough.

$$L(F) = \prod_{i=1}^{n} \frac{F(\{X_i\})}{F(T_i)} \quad \text{or} \quad \prod_{i=1}^{n} \frac{F(C_i \cap T_i)}{F(T_i)}$$

Kaplan-Meier

$X =$ failure time, or age, or other positive quantity
$Y =$ censoring time. If $X > Y$ we just know $X \in (Y, \infty)$

Let F be the distribution of X.
Let $t_1 < t_2 < \cdots < t_k$ be distinct failure/censoring times.

Discrete case

Represent F via $\lambda_j = \dfrac{F(\{t_j\})}{F([t_j, \infty))}$ (hazard)

$$L(F) = \prod_{j=1}^{k} \lambda_j^{d_j} (1 - \lambda_j)^{r_j - d_j} \qquad d_j \text{ out of } r_j \text{ remaining, fail at time } t_j$$

$$\hat{\lambda}_j = \frac{d_j}{r_j} \qquad \text{MLE}$$

$$\hat{F}(t) = 1 - \prod_{j \mid t_j \le t} \frac{r_j - d_j}{r_j} \qquad \text{product limit}$$

Lynden-Bell

Let X = brightness of a star
and Y = distance from us

Choose units so that observation is possible only when $X \ge Y$

If $X \sim F$ and $Y \sim G$ independently then

$$L = \prod_{i=1}^{n} \frac{F(\{X_i\})}{G([0, X_i])}$$

The MLE for F is also of 'product-limit' form. Lynden-Bell (conditional likelihood) for left truncated data

$$\widehat{F}((-\infty, t]) = 1 - \prod_{i=1}^{n} \Bigl(1 - \frac{1_{x_i \le t}}{\sum_{\ell=1}^{n} 1_{y_\ell < x_i \le x_\ell}}\Bigr)$$

Can have $\widehat{F}((-\infty, x_{(i)}] = 1$ for some $i < n$

Monotone & unimodal

Grenander (1956) $X \in [0, \infty)$ density f non-decreasing NPMLE $\hat{F}$ is 'least concave majorant of the ECDF'

piece-wise linear density

Log concave

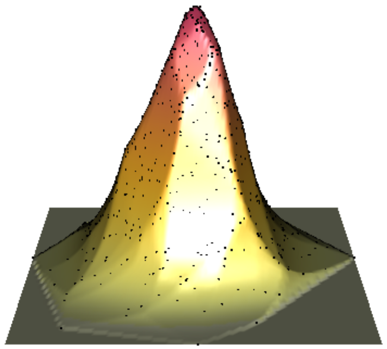
Recent work Samworth, Cule, Walther, Dumbgen $\cdots$

$\log f(\boldsymbol{x})$ concave on $\mathbb{R}^d$

MLE computable for small d

No bandwidth to select

A log concave MLE



Downloaded January 2014 from
http://www.statslab.cam.ac.uk/Statistics/activities/CSI_RS2.png

Nonparametric likelihood ratios

Likelihood ratio:	$R(F) = L(F)/L(\widehat{F})$
Confidence region:	$\{T(F) \mid R(F) \ge r\}$
Profile likelihood:	$\mathcal{R}(\theta) = \sup\{R(F) \mid T(F) = \theta\}$
Confidence region:	$\{\theta \mid \mathcal{R}(\theta) \ge r\}$

Choosing r in a parametric setting,

$$-2\log(r) = \chi^{2,1-\alpha}_{(q)}$$

We seek a nonparametric version

Survival curve

Thomas & Grunkemeier (1975)

$L(F) = \prod_{j=1}^{k} \lambda_j^{d_j} (1-\lambda_j)^{r_j - d_j}$

$\hat{F}$ is Kaplan Meier

$R(F) = L(F)/L(\hat{F})$ Likelihood ratio function

Profile likelihood ratio

$\mathcal{R}(s,t) = \max\{R(F) \mid F([t,\infty)) = s\}$

$s_0 = F_0([t,\infty))$ for true F_0

They find $-2\log(\mathcal{R}(s_0,t)) \to \chi^2_{(1)}$ heuristically

Subsequent empirical likelihood ratios

Data type	Statistic	Reference
Right censoring	Survival prob	Thomas & Grunkemeier, Li, Murphy
Left truncation	Survival prob	Li
Left trunc, right cens	Mean	Murphy & van der Vaart
Right censoring	proportional hazard param	Murphy & van der Vaart
Right censoring	integral vs cumu hazard	Pan & Zhou

General statistic (first with no ties)

Let $w_i = F(\{\boldsymbol{x}_i\})$ $w_i \ge 0$ $\sum_{i=1}^{n} w_i \le 1$ $\boldsymbol{x}_i \in \mathbb{R}^d$

$$L(F) = \prod_{i=1}^{n} w_i \quad L(\widehat{F}) = \prod_{i=1}^{n} 1/n \quad R(F) = \prod_{i=1}^{n} n w_i$$

$$\mathcal{R}(\theta) = \sup\Big\{\prod_{i=1}^{n} n w_i \mid T(F) = \theta\Big\}$$ some parameter $T(F)$

If there are ties . . .

$$L(F) \to L(F) \times \prod_j n_j^{n_j}$$ and, $$L(\widehat{F}) \to L(\widehat{F}) \times \prod_j n_j^{n_j}$$

R and $\mathcal{R}$ unchanged

For the mean of F

$T(F) \equiv \int \boldsymbol{x}\, dF(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{R}^d$

$\widehat{T} \equiv T(\widehat{F}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$

We get $\{T(F) \mid R(F) \ge \epsilon\} = \mathbb{R}^d, \quad \forall r < 1$

Let $F_{\epsilon,\boldsymbol{x}} = (1-\epsilon)\widehat{F} + \epsilon\delta_{\boldsymbol{x}}$

For any $r < 1$,

$R(F_{\epsilon,\boldsymbol{x}}) = \frac{L((1-\epsilon)\widehat{F}+\epsilon\delta_{\boldsymbol{x}})}{L(\widehat{F})} \ge (1-\epsilon)^n \ge r$ for small enough ϵ

Then let $\delta_{\boldsymbol{x}}$ range over $\mathbb{R}^d$

Bounded random variables

If $\Pr(\boldsymbol{X} \in B) = 1$, for known bounded set B, then the confidence region

$$\left\{ \int \boldsymbol{x} dF(\boldsymbol{x}) \mid R(F) \geq c, F(B) = 1 \right\}$$

does not become degenerate.

Which bounded set?

If $\mathbb{E}(\|\boldsymbol{X}\|^2) < \infty$ then it works to take B to be the convex hull of the sample.

(The hull approaches the support fast enough.)

Then maximizing the likelihood for $F(B) = 1$ reduces to maximizing it for $F(\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}) = 1$

Empirical likelihood for the mean

Restrict to $F(\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}) = 1$ i.e. $\sum_{i=1}^n w_i = 1$

Confidence region is

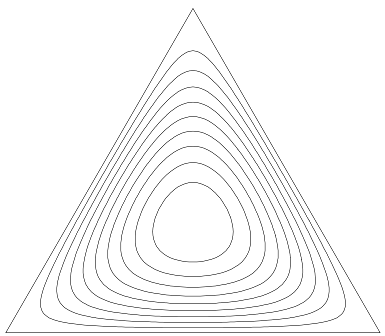
$$C_{r,n} = \left\{ \sum_{i=1}^n w_i \boldsymbol{x}_i \mid w_i \geq 0, \sum_{i=1}^n w_i = 1, \prod_{i=1}^n n w_i > r \right\}$$

Profile likelihood

$$\mathcal{R}(\mu) = \sup\left\{ \prod_{i=1}^n n w_i \mid w_i > 0, \sum_{i=1}^n w_i = 1, \sum_{i=1}^n w_i \boldsymbol{x}_i = \mu \right\}$$

We have a multinomial on the n data points, hence $n - 1$ parameters

Multinomial likelihood for $n = 3$



Contours of $\prod_i n w_i$ MLE at center LR$= i/10, i = 0, \dots, 9$

Empirical likelihood theorem

Suppose that $\boldsymbol{X}_i \sim F_0$ are IID in $\mathbb{R}^d$

$\mu_0 = \int \boldsymbol{x} dF_0(\boldsymbol{x})$

$V_0 = \int (\boldsymbol{x} - \mu_0)(\boldsymbol{x} - \mu_0)^T dF_0(\boldsymbol{x})$ finite

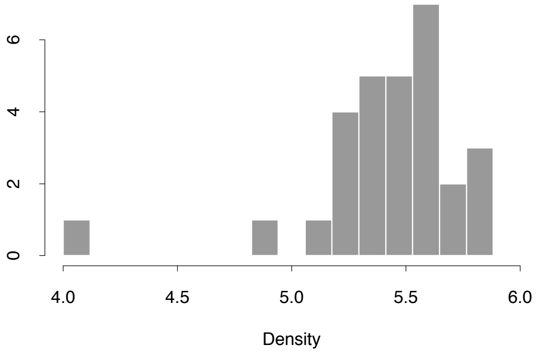
rank$(V_0) = q > 0$

Then as $n \to \infty$

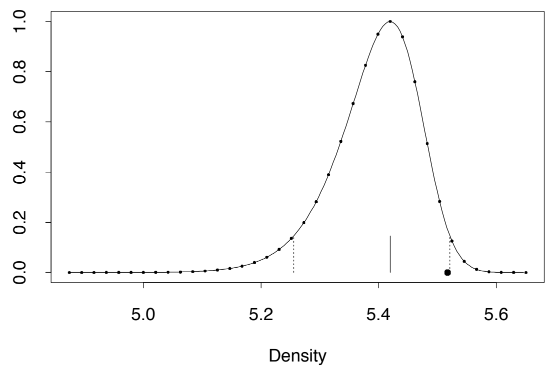
$$-2 \log \mathcal{R}(\mu_0) \to \chi^2_{(q)}$$

same as parametric limit

Cavendish's measurements of Earth's density



Profile empirical likelihood



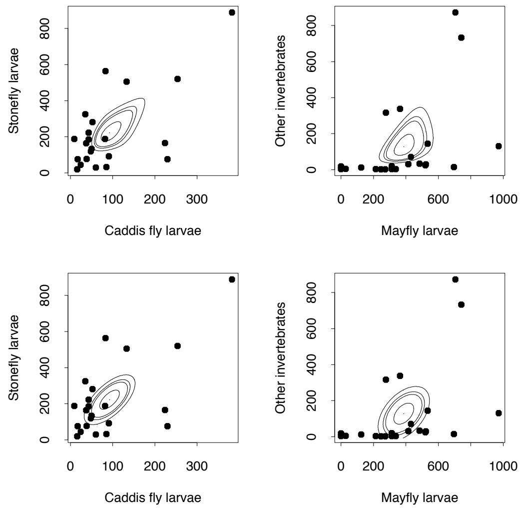
Bars show 95% C.I.

Dot is at presently known value.

Dipper, Cinclus cinclus

Eats larvae of Mayflies, Stoneflies, Caddis flies, other

Dipper diet means



Top row shows EL; bottom Hotelling's T^2 ellipses

Data from Iles

Computing EL for the mean

Start with the convex hull:

$$\mathcal{H} = \mathcal{H}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) = \Big\{ \sum_{i=1}^n w_i \boldsymbol{x}_i \mid w_i \geq 0, \sum_{i=1}^n w_i = 1 \Big\}$$

$$\mu \notin \mathcal{H} \implies \log \mathcal{R}(\mu) = -\infty$$

If $\mu \in \mathcal{H}$ then $\mathcal{R}(\mu) < \infty$

and we will compute it via Lagrange multipliers

Lagrange multipliers

$$G = \sum_{i=1}^n \log(nw_i) - n\lambda^{\mathsf{T}}\Big(\sum_{i=1}^n w_i(\boldsymbol{x}_i - \mu)\Big) + \gamma\Big(\sum_{i=1}^n w_i - 1\Big)$$

$$\frac{\partial}{\partial w_i} G = \frac{1}{w_i} - n\lambda^{\mathsf{T}}(\boldsymbol{x}_i - \mu) + \gamma = 0$$

$$\sum_i w_i \frac{\partial}{\partial w_i} G = n + \gamma = 0 \implies \gamma = -n$$

Solving,

$$w_i = \frac{1}{n} \frac{1}{1 + \lambda^{\mathsf{T}}(\boldsymbol{x}_i - \mu)}$$

Where $\lambda = \lambda(\mu)$ solves

$$0 = \sum_{i=1}^n \frac{\boldsymbol{x}_i - \mu}{1 + \lambda^{\mathsf{T}}(\boldsymbol{x}_i - \mu)}$$

reciprocal tilting

Convex duality

$$\text{Let} \quad \mathbb{L}(\lambda) \equiv -\sum_{i=1}^n \log(1 + \lambda^{\mathsf{T}}(\boldsymbol{x}_i - \mu)) = \log R(F)$$

$$\frac{\partial \mathbb{L}}{\partial \lambda} = -\sum_{i=1}^n \frac{\boldsymbol{x}_i - \mu}{1 + \lambda^{\mathsf{T}}(\boldsymbol{x}_i - \mu)}$$

Minimizing $\mathbb{L}$ sets gradient to 0 and maximizes $\log R$

$$\frac{\partial^2 \mathbb{L}}{\partial \lambda \partial \lambda^{\mathsf{T}}} = \sum_{i=1}^n \frac{(\boldsymbol{x}_i - \mu)(\boldsymbol{x}_i - \mu)^{\mathsf{T}}}{(1 + \lambda^{\mathsf{T}}(\boldsymbol{x}_i - \mu))^2}$$

$\mathbb{L}$ is convex and d dimensional $\implies$ easy optimization

Range extension

Recall

$$\mathbb{L}(\lambda) \equiv -\sum_{i=1}^n \log(1 + \lambda^{\mathsf{T}}(\boldsymbol{x}_i - \mu)) = \log R(F)$$

At the solution

$$w_i = \frac{1}{n} \frac{1}{1 + \lambda^{\mathsf{T}}(\boldsymbol{x}_i - \mu)} \leq 1$$

Therefore

$$1 + \lambda^{\mathsf{T}}(\boldsymbol{x}_i - \mu) \geq 1/n$$

So we may replace $\log$ by

$$\log_\star(z) = \begin{cases} \log(z), & z \geq 1/n \\ Q(z), & z < 1/n. \end{cases}$$

for function $Q(\cdots)$ matching $\log(\cdots)$ and several derivatives at $z = 1/n$

Extended function

Now

$$\mathbb{L}_\star(\lambda) \equiv -\sum_{i=1}^n \log_\star(1+\lambda^\mathsf{T}(\boldsymbol{x}_i-\mu))$$

is well defined for **all** $\lambda \in \mathbb{R}^d$ (no constraints needed)

If $R(F)$ is finite then $\mathbb{L}_\star$ has the same minimizer as $\mathbb{L}$

Optimization

The Newton step for minimizing $\mathbb{L}_\star$ turns out to be least squares.

As a result there are fast and stable algorithms for it.

Recent work O. (2013) shows that we can choose $\mathbb{L}_\star$ to be a self-concordant* convex function. Then global convergence is assured for Newton's method with step reduction Boyd & VandenBerghe.

*$|f'''(x)| \le 2|f''(x)|^{3/2}$, and multidimensional generalizations

Sketch of χ^2 limit proof

WLOG $q=d$, and anticipate a small λ

$$0 = \frac{1}{n}\sum_{i=1}^n \frac{\boldsymbol{x}_i-\mu}{1+(\boldsymbol{x}_i-\mu)^\mathsf{T}\lambda} \qquad 1/(1+\epsilon) = 1-\epsilon+\epsilon^2-\epsilon^3\cdots$$

$$\doteq \frac{1}{n}\sum_{i=1}^n (\boldsymbol{x}_i-\mu)-(\boldsymbol{x}_i-\mu)(\boldsymbol{x}_i-\mu)^\mathsf{T}\lambda, \quad \text{so,}$$

$$\lambda \doteq S^{-1}(\bar{\boldsymbol{x}}-\mu), \quad \text{where,}$$

$$S = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{x}_i-\mu)(\boldsymbol{x}_i-\mu)^\mathsf{T}$$

Sketch continued

$$-2\log\prod_{i=1}^n nw_i = -2\log\prod_{i=1}^n \frac{1}{1+\lambda^\mathsf{T}(\boldsymbol{x}_i-\mu)}$$

$$= 2\sum_{i=1}^n \log(1+\lambda^\mathsf{T}(\boldsymbol{x}_i-\mu)) \qquad \log(1+\epsilon) = \epsilon-(1/2)\epsilon^2+\cdots$$

$$\doteq 2\sum_{i=1}^n \Bigl(\lambda^\mathsf{T}(\boldsymbol{x}_i-\mu)-\frac{1}{2}\lambda^\mathsf{T}(\boldsymbol{x}_i-\mu)(\boldsymbol{x}_i-\mu)^\mathsf{T}\lambda\Bigr)$$

$$= n\Bigl(2\lambda^\mathsf{T}(\bar{\boldsymbol{x}}-\mu)-\lambda^\mathsf{T}S\lambda\Bigr)$$

$$= n\Bigl(2(\bar{\boldsymbol{x}}-\mu)^\mathsf{T}S^{-1}(\bar{\boldsymbol{x}}-\mu)-(\bar{\boldsymbol{x}}-\mu)^\mathsf{T}S^{-1}SS^{-1}(\bar{\boldsymbol{x}}-\mu)\Bigr)$$

$$= n(\bar{\boldsymbol{x}}-\mu)^\mathsf{T}S^{-1}(\bar{\boldsymbol{x}}-\mu)$$

$$\to \chi^2_{(d)}$$

Coverage errors

1) $\Pr(\mu_0 \in C_{r,n}) = 1-\alpha+O\bigl(\frac{1}{n}\bigr)$ as $n\to\infty$ Hall
2) One-sided errors of $O\bigl(\frac{1}{\sqrt{n}}\bigr)$ cancel
3) Bartlett correction DiCiccio, Hall, Romano
   (a) replace $\chi^{2,1-\alpha}$ by $\bigl(1+\frac{a}{n}\bigr)\chi^{2,1-\alpha}$ for carefully chosen a
   (b) get coverage errors $O\bigl(\frac{1}{n^2}\bigr)$
   (c) a does not depend on α
   (d) e.g., $a=(\kappa+3)/2-\gamma^2/3$ for $\mathbb{E}(X)$
   (e) data based $\hat{a}$ gets same rate
   (f) the rate seems to set in slowly

same as for parametric likelihoods

Power

Some nonparametric methods are inefficient

E.g.: sign test for $\#\{X_i > \mu\} \sim \mathrm{Bin}(n/1/2)$ when $X \sim \mathcal{N}(\mu, \sigma^2)$

EL for the mean is efficient

Suppose $X_i \in \mathbb{R}$ with $\mathbb{E}(X) = \mu$ and Var$(X) = \sigma^2 > 0.$

Then

$$-2\log(\mathcal{R}(\mu_0 + \tau\sigma_0 n^{-1/2})) \to \chi^2_{(1)}(\tau^2)$$

noncentral χ^2. Then power $= \Pr(\chi^2_{(1)}(\tau^2) \geq \chi^{2,1-\alpha}_{(1)})$, same as in parametric setting

Finer print

When a parametric model holds, we may use it to generate an MLE of $\hat{\theta}$ EL inferences for that estimate are also as efficient as ones based on parametric likelihood, to a second order analysis in Lazar and Mykland (1998)

Calibrating empirical likelihood

Plain $\chi^{2,1-\alpha}$	undercovers
$F^{1-\alpha}_{d,n-d}$	is a bit better
Bartlett correction	asymptotics slow to take hold
Bootstrap	seems to work best

Bootstrap calibration

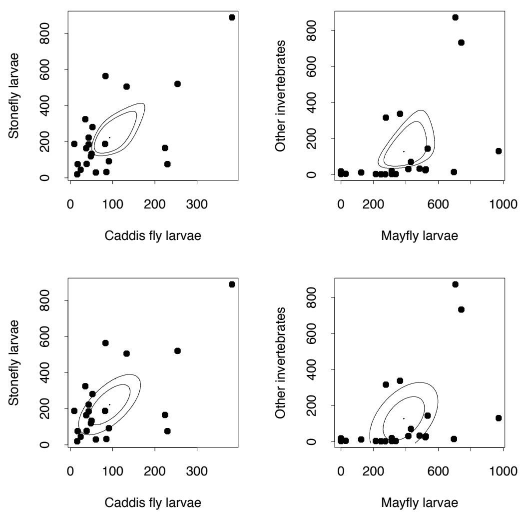
Recipe

Sample $\boldsymbol{X}_i^*$ IID $\widehat{F}$
Get $-2\log\mathcal{R}(\bar{\boldsymbol{x}}; \boldsymbol{x}_1^*, \dots, \boldsymbol{x}_n^*)$
Repeat $B = 1000$ times (or more)
Use $1-\alpha$ bootstrap quantile of $-2\log\mathcal{R}^*$

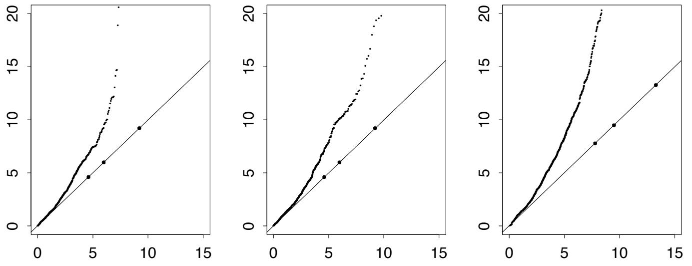
Results

Regions get empirical likelihood shape and bootstrap size
Coverage error $O(n^{-2})$
Same error rate as bootstrapping the bootstrap
Sets in faster than Bartlett correction
Need further adjustments for one-sided inference

Bootstrap (and χ^2) calibrated Dipper regions



Resampled $-2\log\mathcal{R}(\mu)$ values vs χ^2



Caddis vs Stonefly   Mayfly vs other   all four

Euclidean log likelihood

$-\sum_{i=1}^n \log(nw_i)$ is a distance of w from $(1/n,\dots,1/n)$.

Replace loglik by

$$\ell_E = -\frac{1}{2}\sum_{i=1}^n (nw_i - 1)^2$$

Then $-2\ell_E \to \chi^2_{(q)}$ too

Reduces to Hotelling's T^2 for the mean O. (1990)

Reduces to Huber-White covariance for regression

Reduces to continuous updating GMM Kitamura

Quadratic approx to EL, like Wald test is to parametric likelihood

Euclidean likelihood

Minimize $\sum_{i=1}^n (nw_i - 1)^2$

Subject to $\sum_{i=1}^n w_i = 1$, and

$$\sum_{i=1}^n w_i \boldsymbol{x}_i = \mu$$

This is a quadratic programming problem.

Allows $w_i < 0$, and so

Good news confidence regions for means can get out of the convex hull

Bad news but confidence regions no longer obey range restrictions

(e.g. weighted variances can be negative)

Exponential empirical likelihood

Replace $-\sum_{i=1}^n \log(nw_i)$ by

$$\text{KL} = \sum_{i=1}^n w_i \log(nw_i)$$

relates to entropy and exponential tilting

Hellinger distance

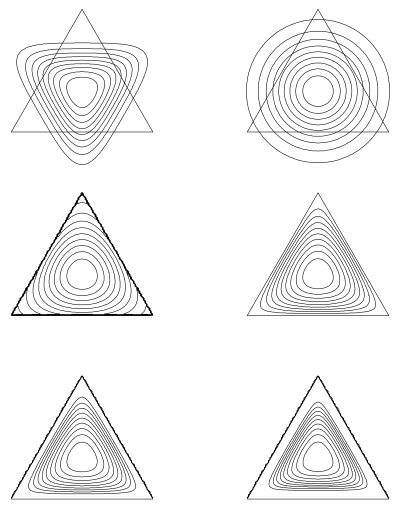
$$\sum_{i=1}^n (w_i^{1/2} - n^{-1/2})^2$$

Renyi, Cressie-Read

$$\frac{2}{\lambda(\lambda+1)}\sum_{i=1}^{n}((nw_i)^{-\lambda}-1)$$

λ	Method
-2	Euclidean log likelihood
$\to -1$	Exponential empirical likelihood
$-1/2$	Freeman-Tukey
$\to 0$	Empirical likelihood
1	Pearson's

Renyi-Cressie-Read contours



Top to bottom, left to right, λ: -5 -2 0 1 2/3 3/2

Alternate artificial likelihoods

All Renyi Cressie-Read familiies have χ^2 calibrations. Baggerly

Only EL is Bartlett correctable Baggerly

$-2\sum_{i=1}^{n}\widetilde{\log}(nw_i)$ Bartlett correctable if

$$\widetilde{\log}(1+z) = z - \frac{1}{2}z^2 + \frac{1}{3}z^3 - \frac{1}{4}z^4 + o(z^4), \quad \text{as } z \to 0$$

Corcoran

$-\widetilde{\log}(\cdot)$ is also convex and self-concordant O (2013)

Biased sampling

Examples

1) Sample children, but record family sizes.
2) Draw blue line over cotton, sample fibers that are partly blue.
3) When $\boldsymbol{Y} = \boldsymbol{y}$ it is recorded as $\boldsymbol{X}$ with prob. $u(\boldsymbol{y})$, lost with prob. $1 - u(\boldsymbol{y})$.

$\boldsymbol{Y} \sim F$, observe $\boldsymbol{X} \sim G$, but we really want F

$$G(A) = \frac{\int_A u(\boldsymbol{y})\,dF(\boldsymbol{y})}{\int u(\boldsymbol{y})\,dF(\boldsymbol{y})}$$

$$L(F) = \prod_{i=1}^{n} G(\{\boldsymbol{x}_i\}) = \prod_{i=1}^{n} \frac{F(\{\boldsymbol{x}_i\})\,u(\boldsymbol{x}_i)}{\int u(\boldsymbol{x})\,dF(\boldsymbol{x})}$$

NPMLE

$$\widehat{G}(\{\boldsymbol{x}_i\}) = \frac{1}{n} \qquad \text{(for simplicity, suppose no ties)}$$

$$\widehat{G}(\{\boldsymbol{x}_i\}) \propto \hat{F}(\{\boldsymbol{x}_i\}) \times u(\boldsymbol{x}_i)$$

$$\widehat{F}(\{\boldsymbol{x}_i\}) = \frac{u_i^{-1}}{\sum_{j=1}^n u_j^{-1}}$$

For the mean

$$\hat{\mu} = \frac{\sum_{i=1}^n \boldsymbol{x}_i/u_i}{\sum_{i=1}^n 1/u_i}$$

Horvitz-Thompson estimator is NPMLE

$$\hat{\mu} = \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i^{-1}\right)^{-1}$$

when $u_i \propto x_i$, so length bias $\implies$ harmonic mean

Biased sampling again

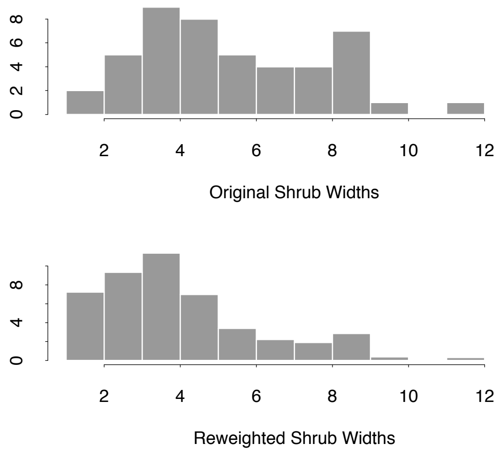
$$0 = \int (\boldsymbol{x} - \mu) dF(\boldsymbol{x}) = \int \frac{\boldsymbol{x} - \mu}{u(\boldsymbol{x})} dG(\boldsymbol{x})$$

$$G(\{\boldsymbol{x}_i\}) = w_i \implies F(\{\boldsymbol{x}_i\}) = \frac{w_i/u_i}{\sum_{j=1}^n 1/u_j}$$

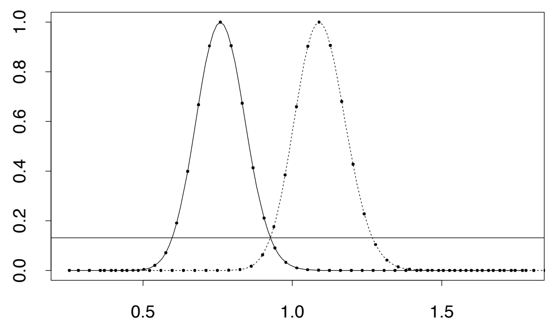
Very simple recipe

$$\mathcal{R}(\theta) = \max\left\{\prod_{i=1}^n nw_i \mid w_i \geq 0, \sum_{i=1}^n w_i = 1, \sum_{i=1}^n w_i \frac{\boldsymbol{x}_i - \mu}{u_i} = 0\right\}$$

Transect sampling of shrubs (Muttlak & McDonald)



Mean shrub width



$$0 = \sum_{i=1}^n w_i \frac{x_i - \mu}{x_i} \qquad \text{Solid, at left}$$

$$0 = \sum_{i=1}^n w_i (x_i - \mu) \qquad \text{Dotted, at right}$$

Next: Estimating equations

The mean is but one of many interesting quantities in statistical problems.

It often happens that a solution for the mean extends readily to other problems.

A key technique is to use estimating equations. Let $\theta \in \mathbb{R}^p$ be defined by

$$\mathbb{E}(m(\boldsymbol{X}, \theta)) = 0$$

where m is usually a function from $\mathbb{R}^{d \times p}$ to $\mathbb{R}^p$.

Then $\hat{\theta}$ is defined by

$$\frac{1}{n}\sum_{i=1}^{n} m(\boldsymbol{x}_i, \hat{\theta}) = 0$$

and we can test $H_0 : \theta = \theta_0$ by testing whether $m(\boldsymbol{X}_i, \theta_0)$ has mean zero.