Empirical likelihood method in statistical inference 2

Hiroaki Ogata

at Université Libre de Bruxelles

March 9-12, 2010

Random vectors

- Here, We consider independent random vectors $\boldsymbol{X}_i \in \mathbb{R}^d$ assuming that they have a common distribution F_0.
- Since it is no longer convenient to describe F_0 by a cumulative distribution function, we describe distributions by the probabilities that they attach to sets, i.e.,

$$F(A) = \Pr(\boldsymbol{X} \in A) \text{ for } \boldsymbol{X} \sim F \text{ and } A \subset \mathbb{R}^r.$$

- We let $\delta_{\boldsymbol{x}}$ denote the distribution under which $\boldsymbol{X} = \boldsymbol{x}$ with probability 1, i.e.,

$$\delta_{\boldsymbol{x}}(A) = \begin{cases} 1 & (\boldsymbol{x} \in A) \\ 0 & (\boldsymbol{x} \notin A) \end{cases}$$

Empirical distribution function

As well as univariate case, we define the empirical distribution function as follows.

Definition

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$. The *empirical distribution function (EDF)* of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ is

$$F_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{\boldsymbol{X}_i}$$

Nonparametric likelihood

As well as univariate case, we define the nonparametric likelihood as follows.

Definition

Given $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$, assumed independent with common distribution function (DF) F_0, the *nonparametric likelihood* of the DF F is

$$L(F) = \prod_{i=1}^{n} F(\{\boldsymbol{X}_i\})$$

Here, $F(\{\boldsymbol{X}_i\})$ is the probability of getting the value $\boldsymbol{X}_i$ in a sample form F.

Nonparametric maximum likelihood

As well as univariate case, the nonparametric likelihood $L(\cdot)$ is maximized by the empirical distribution function F_n.

Theorem

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$ be independent random variables with a common DF F_0. Let F_n be their EDF and let F be any DF. If $F \neq F_n$, then

$$L(F) < L(F_n).$$

EL for a multivariate mean

- EL ratio for a multivariate mean $\boldsymbol{\mu} = \int_{\mathbb{R}^d} \boldsymbol{x} dF(\boldsymbol{x})$ is

$$\mathcal{R}(\boldsymbol{\mu}) = \max_{\boldsymbol{p}} \left\{ \prod_{i=1}^{n} np_i \,\middle|\, \sum_{i=1}^{n} p_i \boldsymbol{X}_i = \boldsymbol{\mu}, p_i \geq 0, \sum_{i=1}^{n} p_i = 1 \right\}$$

- The confidence region may still be written

$$\begin{aligned} C_{r,n} &= \{\boldsymbol{\mu} \mid \mathcal{R}(\boldsymbol{\mu}) \geq r\} \\ &= \left\{ \sum_{i=1}^{n} p_i \boldsymbol{X}_i \,\middle|\, \prod_{i=1}^{n} np_i \geq r, p_i \geq 0, \sum_{i=1}^{n} p_i = 1 \right\} \end{aligned}$$

where r is some threshold.

EL for a multivariate mean

Theorem (Multivariate ELT)

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be independent random vectors in $\mathbb{R}^d$ with common distribution F_0 having mean $\boldsymbol{\mu}_0$ and finite variance covariance matrix V_0 of rank $q > 0$. Then $C_{r,n}$ is a convex set and

$$-2\log(\mathcal{R}(\boldsymbol{\mu}_0)) \xrightarrow{d} \chi^2_{(q)} \qquad (n \to \infty).$$

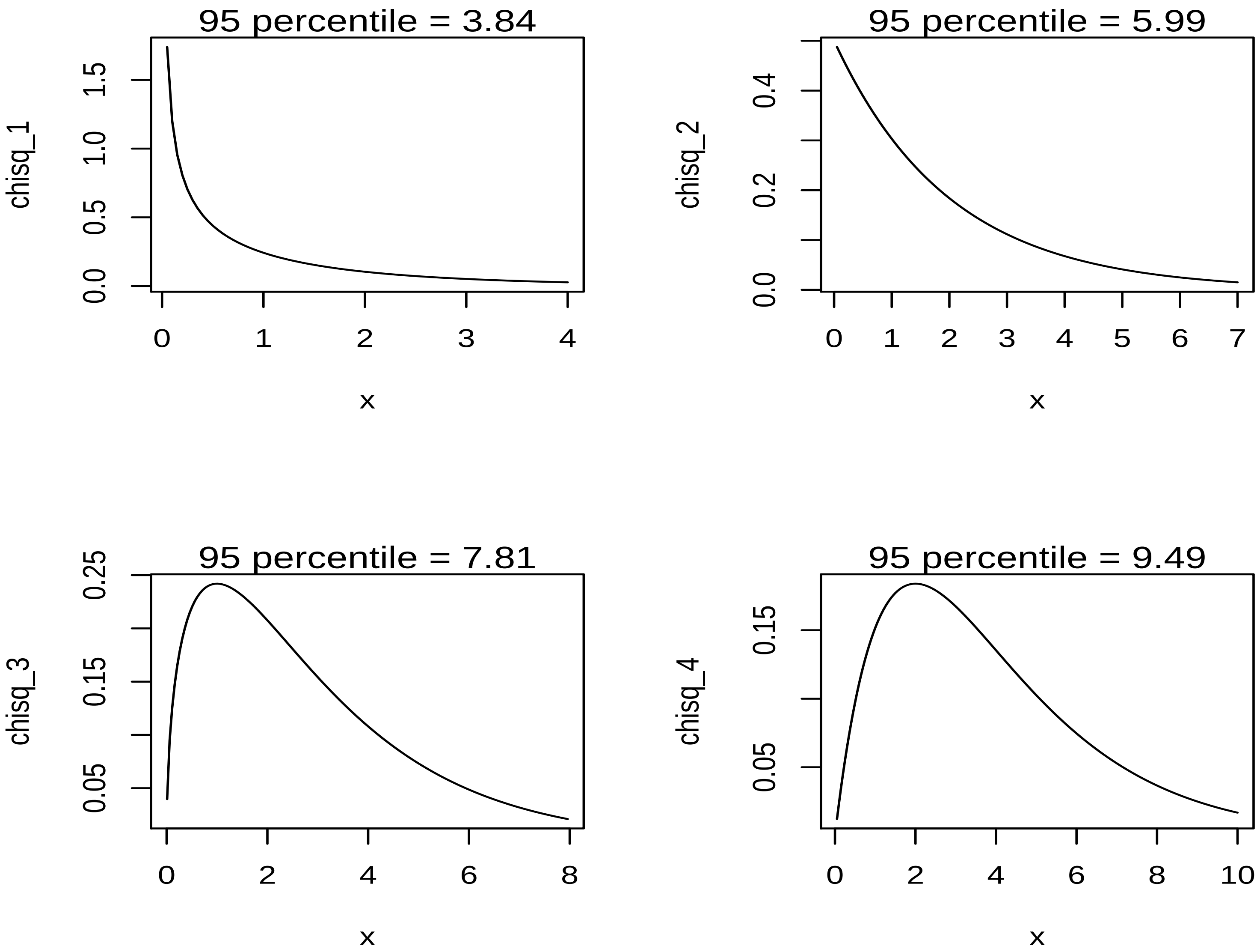
- We reject the value $\boldsymbol{\mu}_0$ at the α level when

$$-2\log(\mathcal{R}(\boldsymbol{\mu}_0)) > \chi^{2,1-\alpha}_{(q)}.$$

- We construct $100(1-\alpha)\%$ confidence region as

$$\{\boldsymbol{\mu} \mid -2\log(\mathcal{R}(\boldsymbol{\mu})) < \chi^{2,1-\alpha}_{(q)}\}.$$

Density functions of chi-square



Estimating functions

- Estimating equations provide an extremely flexible way to describe parameters.
- For a random vector $\boldsymbol{X} \in \mathbb{R}^d$, a parameter $\boldsymbol{\theta} \in \mathbb{R}^p$ is specified through the following equation

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

where $\boldsymbol{m}(\boldsymbol{X}, \boldsymbol{\theta}) \in \mathbb{R}^s$ is a proper vector-valued function of $\boldsymbol{X}$ and $\boldsymbol{\theta}$, which is called the estimating function.

Estimating equations

- The usual setting has $p = s$ (the dimension of parameter $\boldsymbol{\theta}$ is equal to that of estimating function $\boldsymbol{m}(\boldsymbol{X}, \boldsymbol{\theta})$). Then, under conditions on $\boldsymbol{m}(\boldsymbol{X}, \boldsymbol{\theta})$, there is a unique solution of

$$E[\boldsymbol{m}(\boldsymbol{X}, \boldsymbol{\theta})] = \mathbf{0},$$

  with respect to $\boldsymbol{\theta}$.

- In this just determined case, the true value $\boldsymbol{\theta}_0$ may be estimated by solving

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

  which is called estimating equation.

Examples of estimating functions

- Mean

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

- Probability $\Pr(\boldsymbol{X} \in A)$

$$\boldsymbol{m}(\boldsymbol{X}, \boldsymbol{\theta}) = \mathbb{I}_{\boldsymbol{X} \in A} - \boldsymbol{\theta}$$

  - α-quantile (scalar case)

$$m(X, \theta) = \mathbb{I}_{X \leq \alpha} - \theta$$

- Maximum empirical likelihood (MLE)

$$\boldsymbol{m}(\boldsymbol{X}, \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{X}, \boldsymbol{\theta}) \qquad (f(\boldsymbol{X}, \boldsymbol{\theta}) : \text{density})$$

Empirical likelihood with estimating function

Empirical likelihood and estimating equations are well suited to each other. The empirical likelihood ratio function for $\boldsymbol{\theta}$ is defined by

$$\mathcal{R}(\boldsymbol{\theta}) = \max_{\boldsymbol{p}} \left\{ \prod_{i=1}^{n} np_i \,\middle|\, \sum_{i=1}^{n} p_i \boldsymbol{m}(\boldsymbol{X}_i, \boldsymbol{\theta}) = \boldsymbol{0}, p_i \geq 0, \sum_{i=1}^{n} p_i = 1 \right\}$$

Theorem 3 (ELT with estimating function)

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$ be independent random vectors with common distribution F_0. For $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$, and $\boldsymbol{X} \in \mathbb{R}^d$, let $\boldsymbol{m}(\boldsymbol{X}_i, \boldsymbol{\theta}) \in \mathbb{R}^s$. Let $\boldsymbol{\theta}_0 \in \Theta$ be such that $\mathrm{Var}(\boldsymbol{m}(\boldsymbol{X}_i, \boldsymbol{\theta}_0))$ is finite and has rank $q > 0$. If $\boldsymbol{\theta}_0$ satisfies $E(\boldsymbol{m}(\boldsymbol{X}_i, \boldsymbol{\theta}_0)) = \boldsymbol{0}$, then $-2\log(\mathcal{R}(\boldsymbol{\theta}_0)) \xrightarrow{d} \chi^2_{(q)}$ as $n \to \infty$.

Underdetermined cases

- If $s < p$ (the dimension of estimating equation is less than that of parameter), it is called as underdetermined case.
- In this case, there may exist more than one solution $\boldsymbol{\theta}$ which satisfies

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

- Each solution $\boldsymbol{\theta}$ has an asymptotic probability $1 - \alpha$ of being in the confidence region which is based on Theorem 3 under condition that the rank q is common to all solutions.
- That region might not shrink down to a single point as $n \to \infty$.

Overdetermined cases

- If $s > p$ (the dimension of estimating equation is greater than that of parameter), it is called as overdetermined case.
- In this case, there may be no $\boldsymbol{\theta}$ which satisfies

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

- In this case, there is the possibility that confidence region constructed using Theorem 3 will be empty.

Estimating equation with nuisance parameters

- To handle nuisance parameters, write the estimating function as $\boldsymbol{m}(\boldsymbol{X},\boldsymbol{\theta},\boldsymbol{\nu}) \in \mathbb{R}^s$ where
  - $\boldsymbol{\theta} \in \mathbb{R}^p \cdots$ Vector of parameters of interest
  - $\boldsymbol{\nu} \in \mathbb{R}^q \cdots$ Vector of nuisance parameters
- Assume that the parameters $(\boldsymbol{\theta},\boldsymbol{\nu})$ satisfy the equation

$$E[\boldsymbol{m}(\boldsymbol{X},\boldsymbol{\theta},\boldsymbol{\nu})] = \boldsymbol{0}.$$

Example of nuisance parameters

- Suppose that we are interested only in variance. Then, we formulate two estimating equations:

$$\begin{aligned} 0 &= E(X-\mu) \\ 0 &= E((X-\mu)^2-\sigma^2) \end{aligned}$$

  for the parameter $\boldsymbol{\theta} = (\mu, \sigma)$.

- In this case, the mean μ is nuisance parameter.

Profile empirical likelihood ratio function

- Now we define

$$\mathcal{R}(\boldsymbol{\theta}, \boldsymbol{\nu}) = \max_{\boldsymbol{p}} \left\{ \prod_{i=1}^{n} np_i \;\middle|\; \sum_{i=1}^{n} p_i \boldsymbol{m}(\boldsymbol{X}_i, \boldsymbol{\theta}, \boldsymbol{\nu}) = \boldsymbol{0}, p_i \geq 0, \sum_{i=1}^{n} p_i = 1 \right\}$$

  and

$$\mathcal{R}(\boldsymbol{\theta}) = \max_{\boldsymbol{\nu}} \mathcal{R}(\boldsymbol{\theta}, \boldsymbol{\nu})$$

- The function $\mathcal{R}(\boldsymbol{\theta})$ is called the profile empirical likelihood ratio function.
- Under mild conditions, $-2 \log \mathcal{R}(\boldsymbol{\theta}_0) \xrightarrow{d} \chi^2_{(p)}$.

α quantiles

- For $0 < \alpha < 1$, any value Q^α which satisfies

$$\Pr(X \leq Q^\alpha) \geq \alpha, \quad \text{and} \quad \Pr(X \geq Q^\alpha) \geq 1 - \alpha$$

  is called the α quantile of the distribution of X.

- If X has a continuous distribution, the α quantile Q^α is unique and the definition above is equivalent to

$$E(\mathbb{I}_{X \leq Q^\alpha} - \alpha) = 0$$

- It is an important problem to estimate Q^α in many fields including finance. (Ex. Value at Risk)

ELR for α quantiles

- For given $0 < \alpha < 1$, let us take the estimating function as

$$m_\alpha(X_i, q) = \mathbb{I}_{X_i \le \alpha} - q \qquad q \in (-\infty, \infty)$$

  for α quantile. Then define empirical likelihood ratio as

$$\mathcal{R}_\alpha(q) = \max_{\boldsymbol{p}} \left\{ \prod_{i=1}^n n p_i \,\middle|\, \sum_{i=1}^n p_i m_\alpha(X_i, q) = 0, p_i \ge 0, \sum_{i=1}^n p_i = 1 \right\}.$$

- If we set $X_{(k)} \le q \le X_{(k+1)}$, then the first condition becomes

$$\frac{1-\alpha}{\alpha} = \frac{\sum_{i=k+1}^n \; p_i}{\sum_{i=1}^k \; p_i}$$

ELR for α quantiles

- Then, the maximizer $\boldsymbol{p}$ under those constrains is

$$p_i = \begin{cases} \alpha/(n\hat{\alpha}) & (1 \le i \le k) \\ (1-\alpha)/(n(1-\hat{\alpha})) & (k+1 \le i \le n) \end{cases}$$

  where $\hat{\alpha} = \sharp\{X_i \le q\}/n$. So, we can explicitly write the empirical likelihood ratio as

$$\mathcal{R}_\alpha(q) = \left(\frac{\alpha}{\hat{\alpha}}\right)^{n\hat{\alpha}} \left(\frac{1-\alpha}{1-\hat{\alpha}}\right)^{n(1-\hat{\alpha})},$$

  and therefore

$$-\log(\mathcal{R}_\alpha(q)) = n\left(\hat{\alpha}\log\frac{\hat{\alpha}}{\alpha} + (1-\hat{\alpha})\log\frac{1-\hat{\alpha}}{1-\alpha}\right).$$

ELR for α quantiles and tail probabilities

- The empirical likelihood ratio function for α quantile Q^α is piecewise constant, which is clear from the defenition of the estimating function $m_\alpha(X_i, q)$.

- If we want to estimate the probability below the q, then we only have to take α as a parameter. In this case, the empirical likelihood ratio function for α is smooth because the estimating function $m_\alpha(X_i, q)$ is continuous with respect to α.

Computation of ELR for a vector mean

- Denote the convex hull of $\boldsymbol{X}_i \in \mathbb{R}^d$ by

$$\mathcal{H} = \mathcal{H}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = \left\{ \sum_{i=1}^{n} p_i \boldsymbol{X}_i \,\middle|\, p_i \geq 0, \sum_{i=1}^{n} p_i = 1 \right\}$$

  It generalizes the interval $[X_{(1)}, X_{(n)}]$ from the one dimensional case.

- If $\boldsymbol{\mu} \notin \mathcal{H}$, then we take $\mathcal{R}(\boldsymbol{\mu}) = \mathbf{0}$.

- If $\boldsymbol{\mu}$ is on the boundary of $\mathcal{H}$, then we also take $\mathcal{R}(\boldsymbol{\mu}) = \mathbf{0}$ unless $\boldsymbol{X}_i$ all lie in a q-dimensional hyperplane with $1 \leq q < d$.

Nontrivial case

- EL ratio for a multivariate mean $\boldsymbol{\mu} = \int_{\mathbb{R}^d} \boldsymbol{x} dF(\boldsymbol{x})$ is

$$\mathcal{R}(\boldsymbol{\mu}) = \max_{\boldsymbol{p}} \left\{ \prod_{i=1}^n np_i \,\middle|\, \sum_{i=1}^n p_i \boldsymbol{X}_i = \boldsymbol{\mu}, p_i \geq 0, \sum_{i=1}^n p_i = 1 \right\}$$

- If $\boldsymbol{\mu} \in \mathcal{H}$, then we introduce the Lagrangian

$$G = \sum_{i=1}^n \log(np_i) - n\boldsymbol{\lambda}' \left(\sum_{i=1}^n p_i (\boldsymbol{X}_i - \boldsymbol{\mu}) \right) + \gamma \left(\sum_{i=1}^n p_i - 1 \right)$$

where $\boldsymbol{\lambda} \in \mathbb{R}^d$ and $\gamma \in \mathbb{R}$ are Lagrange multipliers.

Nontrivial case

- As in the univariate case we can find that $\gamma = -n$ and maximizing weights

$$p_i = \frac{1}{n} \frac{1}{1 + \boldsymbol{\lambda}'(\boldsymbol{X}_i - \boldsymbol{\mu})}$$

where $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\boldsymbol{\mu})$ satisfies d equations given by

$$\mathbf{0} = \frac{1}{n} \sum_{i=1}^{n} \frac{\boldsymbol{X}_i - \boldsymbol{\mu}}{1 + \boldsymbol{\lambda}'(\boldsymbol{X}_i - \boldsymbol{\mu})} \qquad (*)$$

Nontrivial case

- So, log empirical likelihood ratio for $\boldsymbol{\mu}$ is

$$\begin{aligned}\log \mathcal{R}(\boldsymbol{\mu}) &= \log\left\{\prod_{i=1}^{n} np_i\right\} \\ &= -\sum_{i=1}^{n} \log(1+\boldsymbol{\lambda}'(\boldsymbol{X}_i-\boldsymbol{\mu})) \equiv \mathbb{L}(\boldsymbol{\lambda})\end{aligned}$$

where $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\boldsymbol{\mu})$ satisfies d equations given by

$$\mathbf{0} = \frac{1}{n}\sum_{i=1}^{n} \frac{\boldsymbol{X}_i-\boldsymbol{\mu}}{1+\boldsymbol{\lambda}'(\boldsymbol{X}_i-\boldsymbol{\mu})} \qquad (*)$$

Convex duality

- Note that

$$\text{soving the } d \text{ equations } (*)$$

$$\Updownarrow \text{ equivalent}$$

$$\text{setting the gradient of } \mathbb{L} \text{ with respect to } \lambda \text{ to zero.}$$

$$\frac{\partial \mathbb{L}(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = \mathbf{0}$$

- The Hessian of $\mathbb{L}$ with respect to λ

$$\frac{\partial^2 \mathbb{L}(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\lambda}'} = \sum_{i=1}^{n} \frac{(\boldsymbol{X}_i - \boldsymbol{\mu})(\boldsymbol{X}_i - \boldsymbol{\mu})'}{[1 + \boldsymbol{\lambda}'(\boldsymbol{X}_i - \boldsymbol{\mu})]^2}.$$

  This is a positive definite function of $\boldsymbol{\lambda}$.

Convex duality

- The domain of $\mathbb{L}$ must exclude any $\boldsymbol{\lambda}$ for which some $p_i \geq 0$. Therefore $\boldsymbol{\lambda}$ must be in the region

$$D = \{\boldsymbol{\lambda} \mid 1 + \boldsymbol{\lambda}'(\boldsymbol{X}_i - \boldsymbol{\mu}) > 0\}, \qquad \forall i = 1, \ldots, n.$$

  This region D is an intersection of n half spaces and contains origin $\boldsymbol{\lambda} = \mathbf{0}$. Furthermore, D is a convex subset of $\mathbb{R}^d$.

- After all, maximization problem over n variables p_i subject to $d + 1$ equality constraints

$$\Downarrow$$

  minimization problem of a strictly convex function $\mathbb{L}(\boldsymbol{\lambda})$ over a convex domain D.

Slight modification

- For practical use, we may modify the objective function $\mathbb{L}$ into

$$\mathbb{L}_*(\boldsymbol{\lambda}) = -\sum_{i=1}^{n} \log_*(1 + \boldsymbol{\lambda}(\boldsymbol{X}_i - \boldsymbol{\mu}))$$

where

$$\log_*(x) = \begin{cases} \log(x), & (x \geq 1/n) \\ \log(1/n) - 1.5 + 2nx - (nx)^2/2, & (x \leq 1/n) \end{cases}$$

- $\mathbb{L}_*$ is defined on all of $\mathbb{R}^d$ and twice differentiable. This modification does not affect the value near the solution.

Search for λ

- The derivatives of $\mathbb{L}_*$ can be expressed as

$$\frac{\partial}{\partial \boldsymbol{\lambda}} \mathbb{L}_* = -J'\boldsymbol{y}, \quad \text{and} \quad \frac{\partial^2}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\lambda}'} \mathbb{L}_* = J'J,$$

where J is the $n \times d$ matrix with i'th row

$$J_i = \left[-\log''_*(1 + \boldsymbol{\lambda}'(\boldsymbol{X}_i - \boldsymbol{\mu})) \right]^{1/2} (\boldsymbol{X}_i - \boldsymbol{\mu})',$$

and $\boldsymbol{y}$ is the column vector of n components with i'th component

$$y_i = \frac{\log'_*(1 + \boldsymbol{\lambda}'(\boldsymbol{X}_i - \boldsymbol{\mu}))}{\left[-\log''_*(1 + \boldsymbol{\lambda}'(\boldsymbol{X}_i - \boldsymbol{\mu})) \right]^{1/2}}.$$

Search for λ

- The Newton step is

$$\boldsymbol{\lambda} \to \boldsymbol{\lambda} + (J'J)^{-1}J'\boldsymbol{y}$$

- We repeat the above step until the increment $(J'J)^{-1}J'\boldsymbol{y}$ becomes sufficiently small.

- If $\boldsymbol{\mu}$ is not an interior point of $\mathcal{H}$, then iterative algorithms base on Newton's method produce a sequence of vectors $\boldsymbol{\lambda}$ with length $||\boldsymbol{\lambda}||$ diverging to infinity.

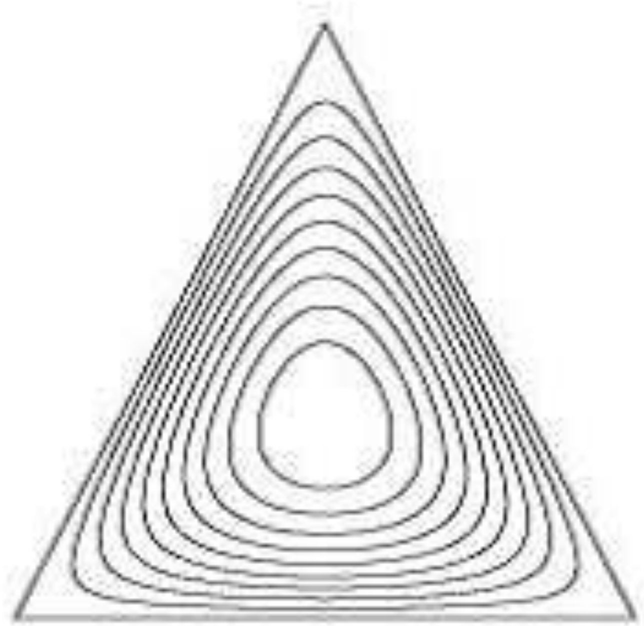
ELR as a measure of the distance

- The log empirical likelihood statistisc $-2\sum_{i=1}^{n}\log(np_i)$ can be interpreted as a measure of the distance of $(p_1,\ldots,p_n)$ from the center (best weight) $(n^{-1},\ldots,n^{-1})$ of the simplex

$$\mathbb{S}_{n-1} = \left\{ (p_1,\ldots,p_n) \,\middle|\, p_i \geq 0, \sum_{i=1}^{n} p_i = 1 \right\}.$$

- Of course, the shorter "distance" is better. So far, we have looked for the point $(p_1,\ldots,p_n)$ which gives the shortest distance from center.

Contours of the log ELR distance



- Shown are contours of $-2\sum_{i=1}^{3}\log(3p_i)$. The minimum value is 0 at the center of the triangle. The maximum value is ∞ at the bounding triangle.
(`http://www-stat.stanford.edu/~owen/empirical/`)

Other measures of the distance

- In addition to the empirical log likelihood distance, there are many other measures of the distance. For example
  - Kullback-Leibler distance : $\sum_{i=1}^{n} p_i \log(np_i)$
  - Hillinger distance : $\sum_{i=1}^{n} \left(p_i^{1/2} - n^{-1/2}\right)^2$
  - Euclidian log likelihood distance : $\frac{1}{2}\sum_{i=1}^{n}(np_i - 1)^2$
- We can replace the empirical log likelihood with the above measures of the distance.

Cressie-Read power divergence statistic

- Cressie-Read power divergence stataistic

$$\mathrm{CR}_\nu(\boldsymbol{p}) = \frac{2}{\nu(\nu+1)} \sum_{i=1}^{n} \left\{ (np_i)^{-\nu} - 1 \right\}$$

  where $-\infty < \nu < \infty$. The degenerate cases $\nu \in \{-1, 0\}$ are handled by takeing limits.

- CR statistic is very general form which includes the empirical log likelihood and previous three examples.

Correspondence to other statistics

$\mathrm{CR}_{-2}(\boldsymbol{p})$	Euclidian log likelihood	$\sum_{i=1}^{n}(np_i - 1)^2$
$\mathrm{CR}_{-1}(\boldsymbol{p})$	Kullback-Liebler statistic	$2\sum_{i=1}^{n} np_i \log(np_i)$
$\mathrm{CR}_{-\frac{1}{2}}(\boldsymbol{p})$	Freeman-Tukey statistic	$-8\sum_{i=1}^{n}(\sqrt{np_i} - 1)$
$\mathrm{CR}_{0}(\boldsymbol{p})$	Empirical log likelihood	$-2\sum_{i=1}^{n}\log(np_i)$
$\mathrm{CR}_{1}(\boldsymbol{p})$	Peason’s chisquare statistic	$\sum_{i=1}^{n}\left(\frac{1}{np_i} - 1\right)$

Asymptotic distribution of CR statistic

Define

$$\mathcal{CR}_\nu(\boldsymbol{\mu}) = \min_{\boldsymbol{p}} \left\{ \mathrm{CR}_\nu(\boldsymbol{p}) \,\middle|\, \sum_{i=1}^{n} p_i \boldsymbol{X}_i = \boldsymbol{\mu}, p_i \geq 0, \sum_{i=1}^{n} p_i = 1 \right\}$$

Theorem 4 (Baggerly (1998))

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be independent random vectors in $\mathbb{R}^d$ with common distribution F_0 having mean $\boldsymbol{\mu}_0$ and finite variance covariance matrix V_0 of rank $q > 0$. Then

$$\mathcal{CR}(\boldsymbol{\mu}_0) \xrightarrow{d} \chi^2_{(q)} \qquad (n \to \infty).$$

for any $-\infty < \nu < \infty$.

Euclidean log likelihood ($\nu = -2$)

- Let us consider the Euclidean log likelihood

$$\mathrm{CR}_{-2}(\boldsymbol{p}) = \sum_{i=1}^{n} (np_i - 1)^2$$

- Unlike the case of log ELR $-2\sum_{i=1}^{n} \log(np_i)$, it is well defined even if some $p_i < 0$. (We do not have to impose the constrain $\forall i, p_i \geq 0$.)
- This negative p_i can yield confidence regions for the mean that extend beyond the convex hull of the data. This can be an advantage when d is large or when n is small.

Computation of Euclidean log likelihood

- Euclidean log likelihood for a multivariate mean $\boldsymbol{\mu} = \int_{\mathbb{R}^d} \boldsymbol{x} dF(\boldsymbol{x})$ is

$$\mathcal{CR}_{-2}(\boldsymbol{\mu}) = \min_{\boldsymbol{p}} \left\{ \sum_{i=1}^{n} (np_i - 1)^2 \,\middle|\, \sum_{i=1}^{n} p_i \boldsymbol{X}_i = \boldsymbol{\mu}, \sum_{i=1}^{n} p_i = 1 \right\}$$

- Introduce the Lagrangian

$$G = \sum_{i=1}^{n} (np_i - 1)^2 - n\boldsymbol{\lambda}' \left(\sum_{i=1}^{n} p_i (\boldsymbol{X}_i - \boldsymbol{\mu}) \right) + \gamma \left(\sum_{i=1}^{n} p_i - 1 \right)$$

where $\boldsymbol{\lambda} \in \mathbb{R}^d$ and $\gamma \in \mathbb{R}$ are Lagrange multipliers.

Computation of Euclidean log likelihood

- Setting

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

we obtain

$$\gamma = n\boldsymbol{\lambda}'(\bar{\boldsymbol{X}} - \boldsymbol{\mu})$$

and minimizing weights

$$p_i = \frac{1}{n}\{1 - \boldsymbol{\lambda}'(\boldsymbol{X}_i - \bar{\boldsymbol{X}})\}.$$

Computation of Euclidean log likelihood

- By some calculations, we obtain

$$\mathbf{0} = (\bar{\boldsymbol{X}} - \boldsymbol{\mu}) - S\boldsymbol{\lambda} \quad \Leftrightarrow \quad \boldsymbol{\lambda} = S^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})$$

  where

$$S = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})'$$

- Therefore, we can eliminate not only γ but also $\boldsymbol{\lambda}$ in the case of Euclidean log likelihood. Then we can explicitely write

$$\mathcal{CR}_{-2}(\boldsymbol{\mu}) = n(\bar{\boldsymbol{X}} - \bar{\mu})'S^{-1}(\bar{\boldsymbol{X}} - \bar{\mu})$$