第四讲　经验似然[1]

张伟平
统计与金融系

---

[1] mainly according to slides by Debdeep Pati

第四讲　经验似然

1 likelihoods 2
2 Introducing empirical likelihood 10
3 Empirical likelihood inference for means . . . 20
4 Empirical likelihood for random vectors . . . 36
  4.1 EL for multivariate means 37
  4.2 EL for smooth functions of means . . 44
5 Estimating Equations 49
  5.1 Estimation via estimating equations . 49
  5.2 EL for estimating equations 52

1 likelihoods

■ Parametric likelihoods

Data $X_1, X_2, \ldots, X_n$ have known distribution f_θ with unknown parameter θ

$$\Pr\left(X_1 = x_1, \ldots, X_n = x_n\right) = f\left(x_1, \ldots, x_n; \theta\right)$$

For continuous data $\cdots$ use probability density function.

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

Likelihood function

$$L(\theta) = L\left(\theta; x_1, \ldots, x_n\right) = f\left(x_1, \ldots, x_n; \theta\right)$$

"Chance, under θ, of getting the data we did get"

■ Likelihood inference Maximum likelihood estimate

$$\hat{\theta} = \arg\max_{\theta} L\left(\theta; x_1, \ldots, x_n\right)$$

Likelihood ratio inferences

$$-2\log\left(L\left(\theta_0\right)/L(\hat{\theta})\right) \rightarrow \chi_q^2 \quad \text{Wilks}$$

1) Reject $H_0 : \theta = \theta_0$ if

$$\frac{L\left(\theta_0\right)}{L(\hat{\theta})} < \exp\left(-\frac{1}{2}\chi_q^2(1-\alpha)\right)$$

2) Confidence set for θ_0 $\left\{\theta \mid \frac{L(\theta)}{L(\hat{\theta})} \geq \exp\left(-\frac{1}{2}\chi_q^2(1-\alpha)\right)\right\}$ e.g. 95% confidence if $\alpha = .05$

Statistical advantages

Typically $\cdots$ **Neyman-Pearson, Cramer-Rao**, ...

- $\hat{\theta}$ asymptotically normal
- $\hat{\theta}$ asymptotically efficient
- Likelihood ratio tests powerful
- Likelihood ratio confidence regions small

Other likelihood advantages: can model/undo 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 $\{\theta \mid L(\theta) \geq rL(\hat{\theta})\}$...

as long as we know correct $f(\cdots;\theta)$!

■ Empirical likelihood (经验似然): a nonparametric method without having to assume the form of the underlying distribution. It retains some of the advantages of likelihood based inference.

↑Example

Example (Somites of Earthworms) Earthworms have segmented bodies. The segments are known as somites. As a worm grows, both the number and the length of the somites increases. The dataset contains the number of somites on each of 487 worms gathered near Ann Arbor in 1902. The histogram shows that the distribution is skewed to the left, and has a heavier tail to the left.

↓Example

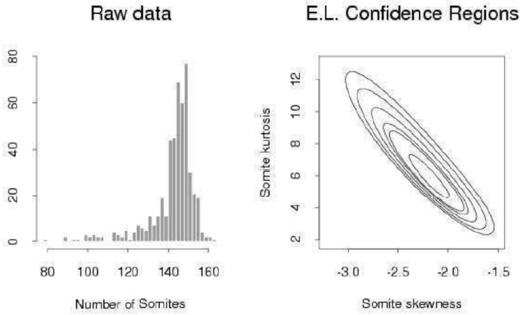


Figure 1. In the second panel, the empirical likelihood confidence regions (i.e. contours) correspond to confidence levels of 50%, 90%, 95%, 99%, 99.9% and 99.99%. Note: $(\gamma, \kappa) = (0, 0)$ is not contained in the confidence regions.

■ Why do conventional methods not apply?

Here are the existing methods:

1. **Parametric likelihood**: Not normal distribution! Likelihood inference for high moments is typically not robust wrt a misspecified distribution.
2. **Bootstrap**: Difficult in picking out the confidence region from a point cloud consisting of a large number of bootstrap estimates for (γ, κ). For example, given 1000 bootstrap estimates for (γ, κ), ideally 95% confidence region should contain 950 central points. In practice, we restrict to rectangle or ellipse regions in order to facilitate the estimation.

Recall the measures of skewness (symmetry) and kurtosis (tailheaviness):

$$\text{Skewness: } \gamma = \frac{E\left\{(X-EX)^3\right\}}{\{\mathrm{Var}(X)\}^{3/2}}$$

$$\text{Kurtosis: } \kappa = \frac{E\left\{(X-EX)^4\right\}}{\{\mathrm{Var}(X)\}^2} - 3$$

Remark 1.

- *For* $N\left(\mu,\sigma^2\right), \gamma = 0$ *and* $\kappa = 0$
- *For symmetric distributions,* $\gamma = 0$
- *When* $\kappa > 0$, *heavier tails than those of* $N\left(\mu,\sigma^2\right)$

■ Estimation of γ and κ

Let $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$ and $\hat{\sigma}^2 = (n-1)^{-1}\sum_{1\le i\le n}\left(X_i - \bar{X}\right)^2$. Then

$$\hat{\gamma} = \frac{1}{n\hat{\sigma}^3}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^3, \quad \hat{\kappa} = \frac{1}{n\hat{\sigma}^4}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^4$$

How to find confidence sets for (γ, κ)? In this section, we will define $l(\gamma, \kappa)$ as the logempirical likelihood function of (γ, κ). The confidence region for (γ, κ) is defined as

$$\{(\gamma, \kappa) : l(\gamma, \kappa) > C\}$$

where $C > 0$ is a constant determined by the confidence level, i.e., $P(l(\gamma, \kappa) > C\} = 1 - \alpha$.

2 Introducing empirical likelihood

Let $\mathbf{X} = (X_1, \ldots, X_n)^{\mathrm{T}}$ be a random sample from an unknown distribution $F(\cdot)$. We know nothing about $F(\cdot)$. In practice, we observe $X_i = x_i, i = 1, \ldots, n$ where $x_1, x_2, \ldots, x_n$ are n known numbers.

Basic idea: Assume F is a discrete distribution on $\{x_1, \cdots, x_n\}$ with

$$p_i = F\left(x_i\right), \quad i = 1, \ldots, n$$

where

$$p_i \geq 0, \sum_{i=1}^{n} p_i = 1$$

which is called an empirical likelihood.

Remark 2. *The number of parameters is the same as the number* *of observations. Note that*

$$\left(\prod_{i=1}^{n} p_i\right)^{1/n} \leq \frac{1}{n}\sum_{i=1}^{n} p_i = \frac{1}{n}$$

the equality holds iff $p_1 = \ldots = p_n = 1/n$. *Putting* $\hat{p}_i = 1/n$, *we have*

$$L\left(p_1, \cdots, p_n; \mathbf{X}\right) \leq L\left(\hat{p}_1, \cdots, \hat{p}_n; \mathbf{X}\right)$$

for any $p_i \geq 0$ *and* $\sum_{i=1}^{n} p_i = 1$. *Hence the MLE based on the empirical likelihood, which is called the maximum empirical likelihood estimator (MELE), puts equal probability mass* $1/n$ *on the* n *observed values* $x_1, x_2, \ldots, x_n$

Given $X_1, \ldots, X_n \in \mathbb{R}$, assumed independent with common CDF F_0, the nonparametric likelihood of the CDFF is

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

Definition

The value $L(F)$ is the probability of getting exactly the observed sample values $X_1, \ldots, X_n$ from the CDF F. One consequence is that $L(F) = 0$ if F is a continuous distribution. To have a positive nonparametric likelihood, a distribution F must place positive probability on every one of the observed data values.

Theorem 1. *Let* $X_1, \ldots, X_n \in \mathbb{R}$ *be independent random variables with a common* $\mathrm{CDF} \mathrm{F}_0$. *Let* F_n *be their* ECDF *and let* F *be any* CDF. *If* $\mathrm{F} \neq \mathrm{F}_n$ *then* $L(F) < L(F_n)$

Proof Let $z_1 < z_2 < \cdots < z_m$ be the distinct values in $\{X_1, \ldots, X_n\}$, and let $n_j \geq 1$ be the number of X_i that are equal to z_j. Let $p_j = F(z_j) - F(z_j-)$ and put $\hat{p}_j = n_j/n$. If $p_j = 0$ for any $j = 1, \ldots, m$, then $L(F) = 0 < L(F_n)$, so we suppose that all $p_j > 0$, and that for at least one $j, p_j \neq \hat{p}_j$. Now $\log(x) \leq x - 1$ for all $x > 0$ with equality only when $x = 1$. Therefore

$$\log\left(\frac{L(F)}{L(F_n)}\right) = \sum_{j=1}^{m} n_j \log\left(\frac{p_j}{\hat{p}_j}\right) = n \sum_{j=1}^{m} \hat{p}_j \log\left(\frac{p_j}{\hat{p}_j}\right)$$

$$< n \sum_{j=1}^{m} \hat{p}_j \left(\frac{p_j}{\hat{p}_j} - 1\right) \leq 0$$

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

↑Example

Example Find the MELE for $\mu = EX_1$. Corresponding to the EL, $\mu = \sum_{i=1}^{n} p_i x_i = \mu\left(p_1, \ldots, p_n\right)$. Therefore, the MELE for μ is

$$\hat{\mu} = \mu\left(\hat{p}_1, \cdots, \hat{p}_n\right) = \bar{X}$$

↓Example

Remark 3. *(1). MELEs, without further constraints, are simply the method of moment estimators, which is not new.*

(2). Empirical likelihood is a powerful tool in dealing with testing hypotheses and interval estimation in a nonparametric matter based on likelihood tradition, which also involves evaluating MELEs under some further constraints.

Inference based on EL

MELE: $T(F)$ by $T(F_n)$, as F_n is the MELE of F.

Testing/CI: Nonparametric Likelihood Ratio

$$R(F) = \frac{L(F)}{L\left(F_n\right)}$$

Assume a parameter of interest: $\theta = T(F), F \in \mathcal{F}$. Define the profile likelihood ratio function:

$$\mathcal{R}(\theta) = \sup\{R(F) \mid T(F) = \theta, F \in \mathcal{F}\}$$

Empirical likelihood hypothesis tests reject $H_0 : T\left(F_0\right) = \theta_0$, when $\mathcal{R}\left(\theta_0\right) < r_0$ for some threshold value r_0. Empirical likelihood confidence regions are of the form

$$\{\theta \mid \mathcal{R}(\theta) \geq r_0\}$$

■ Ties in the data

If there are no ties in the observations and $F(\{x_i\}) = p_i \geq 0$. As

$$\hat{F}(\{x_i\}) = \frac{1}{n}$$

The likelihood ratio is then

$$R(F) = \frac{\prod p_i}{\prod \frac{1}{n}} = \prod np_i$$

and the profile likelihood ratio function for $\theta = T(F)$ is:

$$\mathcal{R}(\theta) = \sup\{\prod np_i \mid T(F) = \theta, p_i \geq 0, \sum p_i \leq 1\}$$

If there are ties, then assuming the distinct values are z_j appearing $n_j \geq 1$ times in the sample, for $F(z_j) = p_j \geq 0$ where $\sum p_j \leq 1$

$$R(F) = \prod_{j=1}^{k} \left(\frac{p_j}{\hat{p}_j}\right)^{n_j} = \prod_{j=1}^{k} \left(\frac{np_j}{n_j}\right)^{n_j}$$

Actually, if we ignore ties, we will ge the same profile likelihood ratio.

To see this, we split atom p_j on z_j into weights w_i on observation x_i, and make it satisfy the constraint:

$$\sum_{i=1}^{n} w_i 1_{x_i=z_j} = p_j, j = 1, \ldots, k$$

Let $\tilde{L}(F) = \prod_{i=1}^{n} w_i$, and maximizing $\tilde{L}(F)$ over w_i, we can get

$$w_i = \frac{p_{j(i)}}{n_{j(i)}}$$

where $x_i = z_j(i)$. So max $\prod_i w_i$ for given F is $\prod_{j=1}^{k} \left(\frac{p_j}{n_j}\right)^{n_j}$. Therefore, the profile likelihood ratio

$$\mathcal{R}(\theta) = \sup\{\prod_{j=1}^{k} \left(\frac{np_j}{n_j}\right)^{n_j} \mid T(F) = \theta, p_i \geq 0, \sum p_i \leq 1\}$$

$$= \sup \left\{ \max\{\prod_{i=1}^{n} nw_i | T(F) = \theta, \sum_{i|x_i=z_j} w_i = p_j, w_i \geq 0, \sum w_i \leq 1\} \right\}$$

$$= \sup \left\{ \prod_{i=1}^{n} nw_i | T(F) = \theta, w_i \geq 0, \sum w_i \leq 1 \right\}.$$

This holds for any family $\mathcal{F}$ of distributions and for whatever function $T(F)$ is used to define θ.

Remark 4. *(1) Intuition for w_i : let*

$$\tilde{X}_i = (X_i, U_i)$$

where $\{U_i\}$ i.i.d. $U(0,1)$, and are independent of all X_i. Then $\tilde{x}_i$ should have no ties. If we define

$$\tilde{F} = F \times U(0,1)$$

$$\tilde{T}(\tilde{F}) = T\,(F)$$

The likelihood ratio for $\left\{\tilde{X}_i\right\}$ *should also be the same as* $R(F)$ *as* U_i *contain no information and we get the same C.I..*

(2) When constructing the profile empirical likelihood function for the mean, we may suppose that $\sum_{i=1}^{n} w_i = 1$.

3 Empirical likelihood inference for means

Let $X_1, \ldots, X_n$ be a random sample from an unknown distribution.

Goal: test hypothesis on $\mu = EX_1$, or find confidence intervals for μ.

Empirical likelihood ratio (ELR)

Consider the hypothesis

$$H_0 : \mu = \mu_0 \quad \text{vs. } H_1 : \mu \neq \mu_0$$

Let $L\left(p_1, \ldots, p_n\right) = \prod_i p_i$. We reject H_0 for large values of the ELR

$$T = \frac{\max L\left(p_1, \ldots, p_n\right)}{\max_{H_0} L\left(p_1, \ldots, p_n\right)} = \frac{L\left(n^{-1}, \ldots, n^{-1}\right)}{L\left(\tilde{p}_1, \ldots, \tilde{p}_n\right)}$$

where $\{\tilde{p}_i\}$ are the constrained MELEs for $\{p_i\}$ under H_0.

Two problems:

1. How do we find $\{\tilde{p}_i\}$?

2. What is the distribution of T under H_0?

The constrained MELEs $\tilde{p}_i = p_i(\mu_0)$, where $\{p_i(\mu)\}$ are the solution of the maximization problem

$$\max_{\{p_i\}} \sum_{i=1}^{n} \log p_i$$

subject to the conditions

$$p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i x_i = \mu$$

The solution for the above problem is given in the Theorem below. Note that

$$x_{(1)} \equiv \min_i x_i \leq \sum_{i=1}^{n} p_i x_i \leq \max_i x_i \equiv x_{(n)}$$

Hence it is natural we require $x_{(1)} \leq \mu \leq x_{(n)}$.

Theorem 2. *For* $\mu \in \left(x_{(1)}, x_{(n)}\right)$,

$$p_i(\mu) = \frac{1}{n - \lambda\left(x_i - \mu\right)} > 0, \quad 1 \leq i \leq n \tag{1}$$

where λ *is the unique solution of the equation*

$$\sum_{j=1}^{n} \frac{x_j - \mu}{n - \lambda\left(x_j - \mu\right)} = 0 \tag{2}$$

in the interval $\left(n/\left(x_{(1)} - \mu\right), n/\left(x_{(n)} - \mu\right)\right)$.

证明. We use the Lagrange multiplier technique to solve this optimization problem. Put

$$Q = \sum_i \log p_i + \psi\left(\sum_i p_i - 1\right) + \lambda\left(\sum_i p_i x_i - \mu\right)$$

Letting the partial derivatives of Q w.r.t. p_i, ψ and λ equal to 0, we have

$$p_i^{-1} + \psi + \lambda x_i = 0 \tag{3}$$

$$\sum_i p_i = 1 \tag{4}$$

$$\sum_i p_i x_i = \mu \tag{5}$$

By (3)

$$p_i = -1/\left(\psi + \lambda x_i\right) \tag{6}$$

Hence, $1 + \psi p_i + \lambda x_i p_i = 0$, which implies $\psi = -(n + \lambda\mu)$. This together with (6) implies (1). By (1) and (5)

$$\sum_i \frac{x_i}{n - \lambda\left(x_i - \mu\right)} = \mu \tag{7}$$

It follows from (4) that

$$\mu = \mu \sum_i p_i = \sum_i \frac{\mu}{n - \lambda\left(x_i - \mu\right)}$$

This together with (7) imply (2). Now, let $g(\lambda)$ be the function on the LHS of (2). Then

$$\frac{d}{d\lambda} g(\lambda) = \sum_i \frac{\left(x_i - \mu\right)^2}{\left\{n - \lambda\left(x_i - \mu\right)\right\}^2} > 0$$

Hence $g(\lambda)$ is a strictly increasing function. Note

$$\lim_{\lambda \uparrow n/(x_{(1)} - \mu)} g(\lambda) = \infty, \lim_{\lambda \downarrow n/(x_{(n)} - \mu)} g(\lambda) = -\infty$$

Hence $g(\lambda) = 0$ has a unique solution in the interval

$$\left(\frac{n}{x_{(n)} - \mu}, \frac{n}{x_{(1)} - \mu}\right)$$

Note that for any λ in this interval,

$$\frac{1}{n - \lambda\left(x_{(1)} - \mu\right)} > 0, \quad \frac{1}{n - \lambda\left(x_{(n)} - \mu\right)} > 0$$

and $1/\{n - \lambda(x - \mu)\}$ is a monotonic function of x. It holds that $p_i(\mu) > 0$ for all $1 \le i \le n$. ☐

Remark 5. *(a). When* $\mu = \bar{x}, \lambda = 0,$ *and*

$$p_i(\mu) = 1/n, \quad i = 1, \ldots, n$$

It may be shown for μ *close* $E\left(X_i\right),$ *and* n *large*

$$p_i(\mu) \approx \frac{1}{n} \frac{1}{1 + \frac{\bar{x} - \mu}{S(\mu)}\left(x_i - \mu\right)}$$

where $S(\mu) = (1/n) \sum_{i=1}^{n} \left(x_i - \mu\right)^2$

(b). We may view

$$L(\mu) = L\left\{p_1(\mu), \ldots, p_n(\mu)\right\}$$

as a profile empirical likelihood for μ. Hypothetically consider an $1-1$ *parameter transformation from* $\{p_1, \ldots, p_n\}$ *to* $\{\mu, \theta_1, \ldots, \theta_n\}$. *Then*

$$L(\mu) = \max_{\{\theta_i\}} L\left(\mu, \theta_1, \ldots, \theta_{n-1}\right) = L\left\{\mu, \hat{\theta}_1(\mu), \ldots, \hat{\theta}_{n-1}(\mu)\right\}$$

(c). The likelihood function $L(\mu)$ may be calculated using R-code and Splus-code, downloaded at http: / /www-stat.stanford.edu/ -owen/empirical.

Testing for μ

The asymptotic theorem for the classic likelihood ratio tests (i.e., Wilk's Theorem) still holds for the ELR tests. Let $X_1, \ldots, X_n$ be i.i.d and $\mu = E(X_1)$. To test

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu \neq \mu_0$$

The ELR statistic is

$$\begin{aligned} T = \frac{\max L(p_1, \ldots, p_n)}{\max_{H_0} L(p_1, \ldots, p_n)} &= \frac{(1/n)^n}{L(\mu_0)} \\ &= \prod_{i=1}^{n} \frac{1}{np_i(\mu_0)} = \prod_{i=1}^{n} \left\{ 1 - \frac{\lambda}{n} (X_i - \mu_0) \right\} \end{aligned}$$

where λ is the unique solution of

$$\sum_{j=1}^{n} \frac{X_j - \mu_0}{n - \lambda (X_j - \mu_0)} = 0$$

Theorem 3. *Let* $E\left(X_1^2\right) < \infty$. *Then under* H_0,

$$2 \log T = 2 \sum_{i=1}^{n} \log \left\{ 1 - \frac{\lambda}{n} \left(X_i - \mu_0 \right) \right\} \rightarrow \chi_1^2$$

Proof. (Sketch) Under $H_0, E\left(X_i\right) = \mu_0$. Therefore μ_0 is close to $\bar{X}$ for large n. Hence the λ, or more precisely, $\lambda_n \equiv \lambda / n$ is small, which is the solution of $f\left(\lambda_n\right) = 0$, where

$$f\left(\lambda_n\right) = \frac{1}{n} \sum_{j=1}^{n} \frac{X_j - \mu_0}{1 - \lambda_n \left(X_j - \mu_0\right)}$$

By a simple Taylor expansion $0 = f\left(\lambda_n\right) \approx f(0) + \dot{f}(0) \lambda_n$, implying

$$\lambda_n \approx -f(0) / \dot{f}(0) = -\left(\bar{X} - \mu_0\right) / \left\{ (1/n) \sum_j \left(X_j - \mu_0\right)^2 \right\}$$

Now,

$$\begin{aligned}2\log T &\approx 2\sum_i \left\{-\lambda_n\left(X_i-\mu_0\right)-\frac{\lambda_n^2}{2}\left(X_i-\mu_0\right)^2\right\}\\ &= -2\lambda_n n\left(\bar{X}-\mu_0\right)-\lambda_n^2\sum_i\left(X_i-\mu_0\right)^2\\ &\approx \frac{n\left(\bar{X}-\mu_0\right)^2}{n^{-1}\sum_i\left(X_i-\mu_0\right)^2}\end{aligned}$$

By the LLN, $n^{-1}\sum_i\left(X_i-\mu_0\right)^2 \to \mathrm{Var}\left(X_1\right)$. By the CLT, $\sqrt{n}\left(\bar{X}-\mu_0\right) \to \mathrm{N}\left(0, \mathrm{Var}\left(X_1\right)\right)$ in distribution. Hence $2\log T \to \chi_1^2$ in distribution.

Confidence intervals for μ

For a given $\alpha \in (0,1)$, since we will not reject the null hypothesis $H_0 : \mu = \mu_0$ iff $2 \log T < \chi_1^2(1-\alpha)$, hence a $100(1-\alpha)$ confidence interval for μ is

$$\begin{aligned}
&\left\{\mu : -2 \log \{L(\mu) n^n\} < \chi_1^2(1-\alpha)\right\} \\
&= \left\{\mu : \sum_{i=1}^n \log p_i(\mu) > -0.5\chi_1^2(1-\alpha) - n \log n\right\} \\
&= \left\{\mu : \sum_{i=1}^n \log \{n p_i(\mu)\} > -0.5\chi_1^2(1-\alpha)\right\}
\end{aligned}$$

↑Example

Example Darwin's data: gains in height of plants from cross-fertilization. X = height (Cross-F) - height(Self-F). There are 15 observations.

6.1,-8.4,1.0,2.0,0.7,2.9,3.5,5.1,1.8,3.6,7.0,3.0,9.3,7.5,-6.0

Is the gain significant?

↓Example

Intuitively: YES, if the negative observations -8.4 and -6.0 do not exist. Let $\mu = EX_i$ and set up the hypotheses as

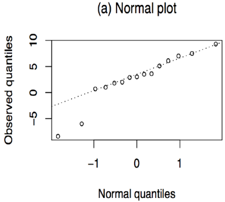
$$H_0 : \mu = 0, \quad \text{vs.} \quad H_1 : \mu > 0$$

The sample mean $\bar{X} = 2.61$ and the standard error $s = 4.71$.

1. Standard approach: Assume $\{X_1, \ldots, X_{15}\}$ is a random sample from $\mathrm{N}\left(\mu, \sigma^2\right)$. The MLE is $\hat{\mu} = \bar{X} = 2.61$. The t-test statistic is

$$T = \sqrt{n}\bar{X}/s = 2.14$$

since $T = t(14)$ under H_0, the p -value is 0.06 - significant but not overwhelming. Is $\mathrm{N}\left(\mu, \sigma^2\right)$ an appropriate assumption? as the data do not appear to be normal (with a heavy left tail ; see Figure 2 .



(a) Normal plot

Figure 2: Quantile of $N(0,1)$ vs Quantile of the empirical distribution

2. Consider a generalized normal family

$$f_k(x \mid \mu, \sigma) = \frac{2^{-1-1/k}}{\Gamma(1+1/k)\sigma} \exp\left\{-\frac{1}{2}\left|\frac{x-\mu}{\sigma}\right|^k\right\}$$

which has the mean μ. When $k = 2$, it is $\mathrm{N}\left(\mu, \sigma^2\right)$. To find the profile likelihood of μ, the 'MLE' for σ is

$$\hat{\sigma}^k \equiv \hat{\sigma}(\mu)^k = \frac{k}{2n} \sum_{i=1}^{n} |X_i - \mu|^k$$

Hence

$$l_k(\mu) = l_k(\mu, \hat{\sigma}) = -n \log \Gamma(1+1/k) - n(1+1/k) \log 2 - n \log \hat{\sigma} - n/k$$

Figure 3 shows that the MLE $\hat{\mu} = \hat{\mu}(k)$ varies with respect to k. In fact $\hat{\mu}(k)$ increases as k decreases. If we use the density with $k = 1$ to fit the data, then the p -value for the test is 0.03 which is much more significant than that under the assumption of normal distribution.

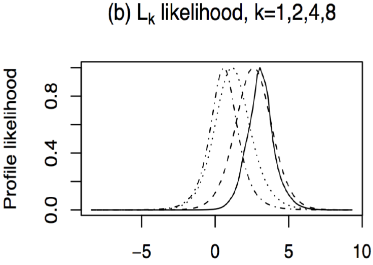


Figure 3: Profile likelihood. The profile likelihood $l_k(\mu)$ is plotted against μ for $k = 1$ (solid), 2 (dashed), 4 (dotted), and 8 (dot-dashed).

3. The empirical likelihood ratio test statistic $2logT = 3.56$, which rejects H_0 with the p-value 0.04. The 95% credible interval is

$$\left\{\mu : \sum_{i=1}^{15} \log p_i(\mu) > -1.92 - 15\log(15)\right\} = [0.17, 4.27]$$

4. The double exponential density is of the form $1/(2\sigma)e^{-|x-\mu|/\sigma}$. With μ fixed, the MLE for σ is $n^{-1}\sum_i |X_i - \mu|$. Hence the parametric log (profile) likelihood is $-n\log\sum_i |X_i - \mu|$. See Figure 4

.

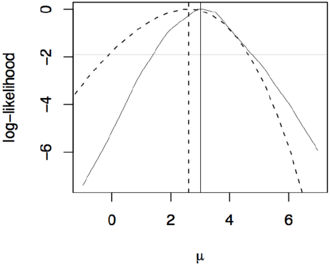


Figure 4: Profile likelihood.Parametric log-likelihood (solid curve) based on the DE distribution, and the empirical log-likelihood (dashed curve). (Both curves were shifted vertically by their own maximum values.)

4 Empirical likelihood for random vectors

Let $\mathrm{X}_1, \ldots, \mathrm{X}_n$ be i.i.d random vectors from distribution F. Similar to the univariate case, we assume

$$p_i = F\left(\mathbf{X}_i\right), \quad i = 1, \ldots, n$$

where $p_i \geq 0$ and $\sum_i p_i = 1$. The empirical likelihood is

$$L\left(p_1, \ldots, p_n\right) = \prod_{i=1}^{n} p_i$$

Without any further constraints, the MELEs are

$$\hat{p}_i = 1/n, i = 1, \ldots, n$$

4.1 EL for multivariate means

The profile empirical likelihood for $\mu = E\mathrm{X}_1$ is

$$L(\boldsymbol{\mu}) = \max \left\{ \prod_{i=1}^{n} p_i : p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i \mathbf{X}_i = \mu \right\}$$

where $p_i(\mu)$ is the MELE of p_i with the additional constraint $E\mathrm{X}_i = \mu$. Define the ELR

$$T \equiv T(\boldsymbol{\mu}) = \frac{L(1/n, \ldots, 1/n)}{L(\boldsymbol{\mu})} = 1/\prod_{i=1}^{n} \{np_i(\boldsymbol{\mu})\}$$

Theorem 4. *Let* $\mathrm{X}_1, \ldots, \mathrm{X}_n$ *be* $d \times 1$ *i.i.d with mean* μ *and finite covariance matrix* Σ *with*$|\Sigma| \neq 0$. *Then as* $n \to \infty$,

$$2 \log\{T(\boldsymbol{\mu})\} = -2 \sum_{i=1}^{n} \log \{np_i(\boldsymbol{\mu})\} \to \chi_d^2$$

in distribution.

Remark 6. *(1). In the case that* $|\Sigma| = 0$*, there exists an integer* $q < d$ *for which,* $\mathbf{X}_i = A\mathbf{Y}_i$ *where* Y_i *is a* $q \times 1$ *random variable such that* $|\mathrm{Var}\,(Y_i)| \neq 0$*, and* A *is a* $d \times q$ *constant matrix. The above theorem still holds with the limit distribution replaced by* χ^2_q

(2). The null hypothesis $H_0 : \mu = \mu_0$ *will be rejected at the significance level* α *iff*

$$\sum_{i=1}^{n} \log\{np_i(\mu_0)\} \leq -0.5\chi^2_{d,1-\alpha}\}$$

where $P\left\{\chi^2_d \leq \chi^2_{d,1-\alpha}\right\} = 1 - \alpha$

(3). A $100(1-\alpha)\%$ *confidence region for* μ *is*

$$\left\{\boldsymbol{\mu} : \sum_{i=1}^{n} \log\left\{np_i(\boldsymbol{\mu})\right\} \geq -0.5\chi^2_{d,1-\alpha}\right\}$$

(4). *Bootstrap calibration:* *since (i) and (ii) are based on an asymptotic result, when* n *is small and* d *large,* $\chi^2_{d,1-\alpha}$ *may be re-*

placed by the $\lceil B\alpha \rceil$ *-th value among* $2\log T_1^*, \ldots, 2\log T_B^*$ *which are computed as follows:*

a. Draw i.i.d sample $\mathrm{X}_1^*, \ldots, \mathrm{X}_n^*$ from the uniform distribution on $\{\mathrm{X}_1, \ldots, \mathrm{X}_n\}$. Let

$$T^* = 1 / \prod_{i=1}^{n} \{ n p_i^* (\bar{X}) \}$$

where $\bar{X} = (1/n) \sum_{i=1}^{n} \mathbf{X}_i$, and $p_i^*(\boldsymbol{\mu})$ is obtained in the same manner as $p_i(\boldsymbol{\mu})$ with $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ replaced by $\{\mathbf{X}_1^*, \ldots, \mathbf{X}_n^*\}$

b. Repeat (a) B times, denote the B values of T^* as $T_1^*, \ldots, T_B^*$

In which, computing $p_i(\mu)$:

by the observations, i.e.,

$$\boldsymbol{\mu} \in \left\{ \sum_{i=1}^{n} p_i \mathbf{X}_i : p_i > 0, \sum_{i=1}^{n} p_i = 1 \right\}$$

This ensures the solutions $p_i(\boldsymbol{\mu})$ exist. We solve the problem in 3 steps.

i. Transform the constrained n -dimensional problem to a constrained d -dimensional problem.

ii. Transform the constrained problem to an unconstrained problem.

iii. Apply a Newton-Raphson algorithm.

Let

$$
\begin{aligned}
l(\boldsymbol{\mu}) = \log L(\boldsymbol{\mu}) &= \sum_{i=1}^{n} \log p_i(\boldsymbol{\mu}) \\
&= \max\{\sum_{i=1}^{n} \log p_i : p_i > 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i \mathbf{X}_i = \boldsymbol{\mu}\}
\end{aligned}
$$

Step 1: Similar to previous Theorem 1, the Lagrangian multiplier method entails:

$$
p_i(\boldsymbol{\mu}) = \frac{1}{n - \boldsymbol{\lambda}^{\mathrm{T}} (\mathbf{X}_i - \boldsymbol{\mu})}, \quad i = 1, 2, \ldots, n
$$

where λ is the solution of

$$\sum_{j=1}^{n} \frac{\mathbf{X}_j - \boldsymbol{\mu}}{n - \boldsymbol{\lambda}^{\mathrm{T}} \left(\mathbf{X}_j - \boldsymbol{\mu}\right)} = 0 \tag{8}$$

Hence

$$l(\boldsymbol{\mu}) = -\sum_{i=1}^{n} \log \left\{ n - \lambda^{\mathrm{T}} \left(\mathbf{X}_i - \boldsymbol{\mu}\right) \right\} \equiv M(\boldsymbol{\lambda})$$

Note $\frac{\partial}{\partial \lambda} M(\boldsymbol{\lambda}) = 0$ leads to (8), and

$$\frac{\partial^2 M(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda} \boldsymbol{\lambda}^{\mathrm{T}}} = \sum_{i=1}^{n} \frac{\left(\mathbf{X}_i - \boldsymbol{\mu}\right) \mathbf{X}_i - \boldsymbol{\mu})^{\mathrm{T}}}{n - \boldsymbol{\lambda}^{\mathrm{T}} \left(\mathbf{X}_i - \boldsymbol{\mu}\right)} > 0$$

Thus $M(\cdot)$ is a convex function on any connected sets satisfying

$$n - \lambda^{\mathrm{T}} \left(\mathbf{X}_i - \boldsymbol{\mu}\right) > 0 \quad i = 1, \ldots, n \tag{9}$$

Note that (9) and (8) together imply $\sum_{i=1}^{n} p_i(\boldsymbol{\mu}) = 1$. The original n -dimensional optimization problem is equivalent to a d -dimensional problem of minimizing $M(\cdot)$ subject to the constraints (9). Let $\mathcal{H}_\lambda$ be the set consisting all the values of λ satisfying

$$n - \lambda^{\mathrm{T}} (\mathrm{X}_i - \mu) > 1, \quad i = 1, \ldots, n$$

Then $\mathcal{H}_\lambda$ is a convex set in $\mathbb{R}^d$, which contains the minimizer of the convex function $M(\boldsymbol{\lambda})$. Unfortunately $M(\boldsymbol{\lambda})$ is not defined on the sets:

$$\left\{\boldsymbol{\lambda} : n - \boldsymbol{\lambda}^{\mathrm{T}} (\mathbf{X}_i - \boldsymbol{\mu}) = 0\right\}, \quad i = 1, 2, \ldots, n$$

Step 2: We extend $M(\boldsymbol{\lambda})$ outside $\mathcal{H}_\lambda$ such that it is still a convex function on the whole $\mathbb{R}^d$. Define

$$\log_*(z) = \begin{cases} \log z, & z \geq 1 \\ -1.5 + 2z - 0.5z^2, & z < 1 \end{cases}$$

It is easy to see that $\log_*(z)$ has two continuous derivatives on $\mathbb{R}$. Set $M_*(\boldsymbol{\lambda}) = -\sum_{i=1}^{n} \log_* \left\{n - \lambda^{\mathrm{T}} (\mathrm{X}_i - \boldsymbol{\mu})\right\}$. Then $M_*(\lambda) = M(\lambda)$ on

$\mathcal{H}_\lambda$ and $M_*(\lambda)$ is a convex function on whole of $\mathbb{R}^d$. Hence $M_*(\boldsymbol{\lambda})$ and $M(\boldsymbol{\lambda})$ share the same minimizer which is the solution of (8)

Step 3 : We apply a Newton-Raphson algorithm to compute λ iteratively:

$$\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k - \left\{\ddot{M}_*\left(\boldsymbol{\lambda}_k\right)\right\}^{-1} \dot{M}_*\left(\boldsymbol{\lambda}_k\right)$$

A convenient initial value would be $\lambda_0 = 0$, corresponding to $p_i = 1/n$

Remark 7. *S-code "el.S", available from `www-stat.stanford.edu/~owen/empirical` calculates the empirical likelihood ratio*

$$\sum_{i=1}^{n} \log\left\{np_i(\boldsymbol{\mu})\right\}$$

and other related quantities.

4.2 EL for smooth functions of means

Basic idea: Let $Y_1, \ldots, Y_n$ be i.i.d random variables with variance σ^2. Note that

$$\sigma^2 = EY_i^2 - E^2(Y_i) = h(\mu)$$

where $\mu = E\mathbf{X}_i$, and $\mathbf{X}_i = \left(Y_i, Y_i^2\right)$. We may deduce a confidence interval for σ^2 from that of μ.

Theorem 5. *Let* $\mathrm{X}_1, \ldots, \mathrm{X}_n$ *be* $d \times 1$ *i.i.d random variables with mean* μ_0 *and* $|\mathrm{Var}(\mathrm{X}_1)| \neq 0$. *Let* $\theta = h(\mu)$ *be a smooth function from* $\mathbb{R}^d \to \mathbb{R}^q$ *where* $q \leq d$, *and* $\theta_0 = h(\mu_0)$. *We assume that*

$$\left|GG^{\mathrm{T}}\right| \neq 0, \quad G = \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\mu}^{\mathrm{T}}}$$

For any $r > 0$, *let*

$$\mathcal{C}_{1,r} = \left\{\mu : \sum_{i=1}^{n} \log\{np_i(\mu\} \geq -0.5r\}\right\}$$

and

$$\mathcal{C}_{3,r} = \{\boldsymbol{\theta}_0 + G(\boldsymbol{\mu} - \boldsymbol{\mu}_0) : \boldsymbol{\mu} \in \mathcal{C}_{1,r}\}$$

Then as $n \to \infty$

$$P(\boldsymbol{\theta} \in \mathcal{C}_{3,r}) \to P\left(\chi_q^2 \leq r\right)$$

Remark 8. *1. The idea of bootstrap calibration may be appropriate here too.*

2. Under more conditions, $P(\theta \in \mathcal{C}_{2,r}) \to P\left(\chi_q^2 \leq r\right)$, *where* $\mathcal{C}_{2,r} = \{h(\mu) : \mu \in \mathcal{C}_{1,r}\}$.

3. $\mathcal{C}_{2,r}$ *is a practical feasible confidence set, while* $\mathcal{C}_{3,r}$ *is not since* $\boldsymbol{\mu}_0$ *and* $\boldsymbol{\theta}_0$ *are unknown in practice. Note that* $\boldsymbol{\mu}$ *close to* $\boldsymbol{\mu}_0$,

$$\boldsymbol{\theta}_0 + G(\boldsymbol{\mu} - \boldsymbol{\mu}_0) \approx h(\boldsymbol{\mu})$$

4. In general, $P(\mu \in \mathcal{C}_{1,r} \leq P(\theta \in \mathcal{C}_{2,r})$.

5. By *Theorem* $4, P(\boldsymbol{\theta} \in \mathcal{C}_{1,r}) \to P\left(\chi_d^2 \leq r\right)$

6. The profile empirical likelihood function of θ is

$$L(\boldsymbol{\theta}) = \max\left\{\prod_{i=1}^{n} p_i(\boldsymbol{\mu}) : h(\boldsymbol{\mu}) = \boldsymbol{\theta}\right\}$$

$$= \max\left\{\prod_{i=1}^{n} p_i : h\left(\sum_{i=1}^{n} p_i \mathbf{X}_i\right) = \boldsymbol{\theta}, p_i \geq 0, \sum_{i=1}^{n} p_i = 1\right\}$$

which may be calculated directly using the Lagrange multiplier method. The computation is more involved for nonlinear $h(\cdot)$.

↑Example

Example S&P500 stock index in 17.8.1999 - 17.8.2000 (256 trading days). Let Y_i be the price on the i -th day

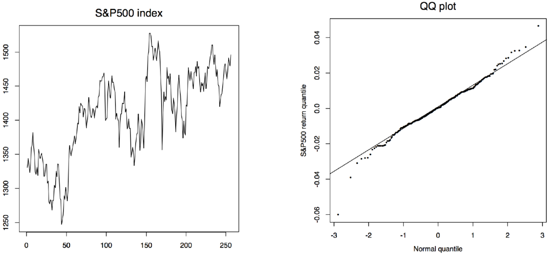
$$X_i = \log\left(Y_i/Y_{i-1}\right) \approx \left(Y_i - Y_{i-1}\right)/Y_{i-1}$$

which is the return, i.e. the percentage of the change on the i th day. By trating X_i i.i.d, we construct confidence intervals for the annual volatility

$$\sigma = \{255\,\mathrm{Var}\,(X_i)\}^{1/2}$$

The simple point-estimator is

$$\hat{\sigma} = \left\{ \frac{255}{255} \sum_{i=1}^{255} (X_i - \bar{X})^2 \right\}^{1/2} = 0.2116$$



(a) S&P Stocks (b) QQ plot for S&P Stocks

Figure 5: S&P Stocks

The 95% confidence intervals for σ the Normal approximation approach is [0.1950,0.2322] and for the EL method is $[0.1895, 0.2422]$. The EL confidence interval is 41.67% wider than the interval based on normal distribution, which reflects the fact that the returns have heavier tails.

↓Example

5 Estimating Equations

5.1 Estimation via estimating equations

Let $\mathrm{X}_1, \ldots, \mathrm{X}_n$ be i.i.d from a distribution F. We are interested in some characteristic $\theta \equiv \theta(F)$, which is determined by equation

$$E\left\{m\left(\mathbf{X}_1, \boldsymbol{\theta}\right)\right\} = 0$$

where θ is a $q \times 1$ vector, m is a $s \times 1$ vector-valued function. For example:

1. $\theta = EX_1$ if $m(x, \theta) = x - \theta$
2. $\theta = EX_1^k$ if $m(x, \theta) = x^k - \theta$
3. $\theta = P\left(X_1 \in A\right)$ if $m(x, \theta) = I(x \in A) - \theta$
4. θ is the α -quantile if $m(x, \theta) = I(x \leq \theta) - \alpha$

A natural estimator for θ is determined by the estimating equation

$$\frac{1}{n}\sum_{i=1}^{n} m\left(\mathbf{X}_1, \hat{\theta}\right) = 0 \tag{10}$$

Obviously, in case F is in a parametric family and m is the score function, $\hat{\theta}$ is the ordinary MLE.

Determined case $q = s$: $\hat{\theta}$ may be uniquely determined by (10)

Determined case $q > s$: The solutions of (10) may form a $(q-s)$ -dimensional set.

Overdetermined case $q < s$: (10) may not have an exact solution, approximating solutions are sought. One such an example is so-called the generalised method of moments estimation which is very popular in Econometrics.

↑Example

Example Let $\{(X_i, Y_i), i = 1, \ldots, n\}$ be a random sample. Find a set of estimating equations for estimating $\gamma \equiv \operatorname{Var}(X_1) / \operatorname{Var}(Y_1)$ In order to estimate γ, we need to estimate $\mu_x = E(X_1), \mu_y = E(Y_1)$ and $\sigma_y^2 = \operatorname{Var}(Y_1)$ Putting $\boldsymbol{\theta}^{\mathrm{T}} = \left(\mu_x, \mu_y, \sigma_y^2, \gamma\right)$, and

$$\begin{gathered} m_1(X, Y, \theta) = X - \mu_x, \quad m_2(X, Y, \theta) = Y - \mu_y \\ m_3(X, Y, \theta) = (Y - \mu_y)^2 - \sigma_y^2 \\ m_4(X, Y, \theta) = (X - \mu_x)^2 - \sigma_y^2 \gamma \end{gathered}$$

and $m = (m_1, m_2, m_3, m_4)^{\mathrm{T}}$. Then $E\{m(X_i, Y_i, \theta)\} = 0$, leading to the estimating equation

$$\frac{1}{n} \sum_{i=1}^{n} m(X_i, Y_i, \boldsymbol{\theta}) = 0$$

↓Example

Remark 9. *Estimating equation method does not facilitate hypothesis tests and interval estimation for θ.*

5.2 EL for estimating equations

Aim: Construct statistical tests and confidence intervals for θ.

The profile empirical likelihood function of θ :

$$L(\boldsymbol{\theta}) = \max\left\{\prod_{i=1}^{n} p_i : \sum_{i=1}^{n} p_i m\left(\mathbf{X}_i, \boldsymbol{\theta}\right) = 0, p_i \geq 0, \sum_{i=1}^{n} p_i = 1\right\}$$

The following Theorem follows from Theorem 2 immidiately.

Theorem 6. *Let* $\mathrm{X}_1, \ldots, \mathrm{X}_n$ *be i.i.d,* $\mathrm{m}(\mathrm{x}, \theta)$ *be an* $s \times 1$ *vector valued function. Suppose*

$$E\left\{m\left(\mathbf{X}_1, \boldsymbol{\theta}_0\right)\right\} = 0, \left|\operatorname{Var}\left\{m\left(\mathbf{X}_1, \boldsymbol{\theta}_0\right)\right\}\right| \neq 0$$

Then as $n \to \infty$

$$-2\log\{L(\boldsymbol{\theta}_0)\} - 2n\log n \to \chi^2_s$$

in distribution.

Remark 10. *1. In general* $L(\theta)$ *can be calculated using the method for EL for multivariate means, treating* $m(\mathbf{X}_i, \boldsymbol{\theta})$ *as a random vector.*

2. For $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ *which is the solution of*

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

3. For $\boldsymbol{\theta}$ *determined by* $E\{\mathrm{m}(\mathrm{X}_1, \boldsymbol{\theta})\} = 0$*, we will reject the null hypothesis* $\mathrm{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ *iff*

$$\log\{L(\theta_0)\} + n\log n \le -0.5\chi^2_{s,1-\alpha}$$

4. Any $(1-\alpha)$ *confidence set for* θ *determined by* $E\left\{m\left(\mathbf{X}_1, \theta\right)\right\}=0$ *is*

$$\left\{\boldsymbol{\theta}: \log \{L(\boldsymbol{\theta})\}+n \log n>-0.5 \chi_{s, 1-\alpha}^2\right\}$$

↑Example

Example (Confidence intervals for quantiles) Let $X_1, \ldots, X_n$ be i.i.d. For a given $\alpha \in(0,1)$, let

$$m\left(x, \theta_\alpha\right)=I\left(x \leq \theta_\alpha\right)-\alpha$$

Then $E\left\{m\left(X_i, \theta_\alpha\right\}=0\right.$ implies θ_α is the α -quantile of the distribution of X_i. We assume the true value of θ_α is between $X_{(1)}$ and $X_{(n)}$. The estimating equation

$$\sum_{i=1}^n m(X_i, \hat{\theta}_\alpha)=\sum_{i=1}^n I\left(X_i \leq \theta_\alpha\right)-n \alpha=0$$

entails $\hat{\theta}_\alpha = X_{(n\alpha)}$, where $X_{(i)}$ denotes the i -th smallest value among $X_1, \ldots, X_n$. Let

$$L(\theta_\alpha) = \max\left\{\prod_{i=1}^n p_i : \sum_{i=1}^n p_i I(X_i \le \theta_\alpha) = \alpha, p_i \ge 0, \sum_{i=1}^n p_i = 1\right\}$$

An $(1-\beta)$ confidence interval for the α -quantile is

$$\Theta_\alpha = \left\{\theta_\alpha : \log\{L(\theta_\alpha)\} > -n\log n - 0.5\chi^2_{1,1-\beta}\right\}$$

Note $L\left(\hat{\theta}_\alpha\right) = (1/n)^n \ge L(\theta_\alpha)$ for any θ_α. It is always true that $\hat{\theta}_\alpha \in \Theta_\alpha$. In fact $L(\theta_\alpha)$ can be computed explicitly as follows. Let $r = r(\theta_\alpha)$ be the integer for which

$$\begin{aligned} X_{(i)} &\le \theta_\alpha, \quad \text{for} \quad i = 1, \ldots, r \\ X_{(i)} &> \theta_\alpha, \quad \text{for } i = r+1, \ldots, n \end{aligned}$$

Thus,

$$L\left(\theta_\alpha\right) = \max\left\{\prod_{i=1}^{n} p_i : p_i \geq 0, \sum_{i=1}^{r} p_i = \alpha, \sum_{i=r+1}^{n} p_i = 1-\alpha\right\}$$
$$= (\alpha/r)^r \{(1-\alpha)/(n-r)\}^{n-r}$$

Hence

$$\Theta_\alpha = \left\{\theta_\alpha : \log\left\{L\left(\theta_\alpha\right)\right\} \geq -n\log n - 0.5\chi_1^2(1-\alpha)\right\}$$
$$= \left\{\theta_\alpha : r\log\frac{n\alpha}{r} + (n-r)\log\frac{n(1-\alpha)}{n-r} > -0.5\chi_1^2(1-\alpha)\right\}$$

which can also be derived directly based on a likelihood ratio test for a binomial distribution.

↓Example

EL with nuisance parameters

For estimating equations with nuisance parameters, we have

$$\mathbb{E}[m(x,\theta,\nu)] = 0$$

where $\theta \in \mathbb{R}^p, \nu \in \mathbb{R}^q$ The profile likelihood ratios are defined as

$$\begin{aligned} \mathcal{R}(\theta,\nu) &= \max \left\{ \prod m_i \mid \sum w_i m\left(X_i,\theta,\nu\right) = 0, w_i \geq 0, \sum w_i = 1 \right\} \\ \mathcal{R}(\theta) &= \max_{\nu} \mathcal{R}(\theta,\nu) = \max_{\nu} \min_{\lambda} L(\nu,\lambda) \end{aligned}$$