1. Introduction

1.1 Why empirical likelihood

- nonparametric method: without having to assume the form of the underlying distribution
- likelihood based inference: taking the advantages of likelihood methods
- alternative method when other (more conventional) methods are not applicable

Remark. (i) For $N(\mu, \sigma^2)$, $\gamma = 0$ and $\kappa = 0$.

(ii) For symmetric distributions, $\gamma = 0$.

(iii) When $\kappa > 0$, heavier tails than those of $N(\mu, \sigma^2)$.

Example 1. Somites of earthworms.

Earthworms have segmented bodies. The segments are known as somites. As a worm grows, both the number and the length of its somites increases.

The dataset contains the No. 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.

Skewness: $\gamma = \frac{E\{(X-EX)^3\}}{\{\mathrm{Var}(X)\}^{3/2}}$, — a measure for symmetry

Kurtosis: $\kappa = \frac{E\{(X-EX)^4\}}{\{\mathrm{Var}(X)\}^2} - 3$, — a measure for tail-heaviness

Estimation for γ and κ

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

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

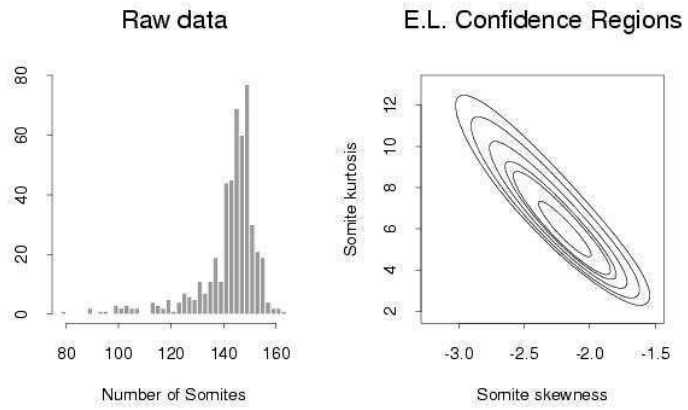
How to find the confidence sets for (γ, κ)?

Answer: Empirical likelihood contours.

Let $l(\gamma, \kappa)$ be the (log-) empirical 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. $P\{l(\gamma, \kappa) > C\} = 1 - \alpha$.



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?

Parametric likelihood. Not normal distribution! Likelihood inference for high moments is typically not robust wrt a misspecified distribution.

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.

1.2 Introducing empirical likelihood

Let $\mathbf{X} = (X_1, \cdots, X_n)^\tau$ be a random sample from an unknown distribution $F(\cdot)$. We *know nothing* about $F(\cdot)$.

In practice we observe $X_i = x_i$ $(i = 1, \cdots, n)$, $x_1, \cdots, x_n$ are n known numbers.

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

$$p_i = F(x_i), \qquad i = 1, \cdots, n,$$

where

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

What is the likelihood function of $\{p_i\}$? What is the MLE?

Since

$$P\{X_1 = x_1, \cdots, X_n = x_n\} = p_1 \cdots p_n,$$

the likelihood is

$$L(p_1, \cdots, p_n) \equiv L(p_1, \cdots, p_n; \mathbf{X}) = \prod_{i=1}^{n} p_i,$$

which is called an *empirical likelihood*.

Remark. The number of parameters is the same as the number of observations.

Note

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

the equality holds iff $p_1 = \cdots = p_n = 1/n$.

Put $\widehat{p}_i = 1/n$, we have

$$L(p_1, \cdots, p_n; \mathbf{X}) \le L(\widehat{p}_1, \cdots, \widehat{p}_n; \mathbf{X})$$

for any $p_i \ge 0$ and $\sum_i p_i = 1$.

Hence the MLE based on the empirical likelihood, which is called **maximum empirical likelihood estimator (MELE)**, puts the equal probability mass $1/n$ on the n observed values $x_1, \cdots, x_n$.

Namely the MELE for F is the uniform distribution on observed data points. The corresponding distribution function

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \le x)$$

is called the **empirical distribution** of the sample $\mathbf{X} = (X_1, \cdots, X_n)^\tau$.

Remarks. (i) MELEs, without further constraints, are simply the method of moments estimators, which is not new.

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

Example 2. Find the MELE for $\mu \equiv EX_1$.

Corresponding to the EL,

$$\mu = \sum_{i=1}^{n} p_i x_i = \mu(p_1, \cdots, p_n).$$

Therefore, the MELE for μ is

$$\widehat{\mu} = \mu(\widehat{p}_1, \cdots, \widehat{p}_n) = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}.$$

Similarly, the MELE for $\mu_k \equiv E(X_1^k)$ is the simply the sample k-th moment:

$$\widehat{\mu}_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k.$$

2. Empirical likelihood for means

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

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

Tool: *empirical likelihood ratios (ELR)*

2.1 Tests Consider the hypotheses

$$H_0 : \mu = \mu_0 \qquad \text{vs} \qquad H_1 : \mu \ne \mu_0.$$

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

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

where $\{\tilde{p}_1\}$ are the constrained MELEs for $\{p_i\}$ under H_0.

Two problems:

(i) $\widetilde{p}_i$ =?
(ii) What is the distribution of T under H_0?

(i) The constrained MELEs $\widetilde{p}_i = p_i(\mu_0)$, where $\{p_i(\mu)\}$ are the solution of the maximisation problem:

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

subject to the conditions

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

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

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

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

Theorem 1. For $\mu \in (x_{(1)},\ x_{(n)})$,

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

where λ is the unique solution of the equation

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

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

Proof. We use the Lagrange multiplier technique to solve this optimisation problem. Put

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

Letting the partial derivatives of Q w.r.t. p_i, ψ and λ equal 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),

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

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

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

It follows (4) that

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

This together with (7) imply (2).

Now let $g(\lambda)$ be the function on the LHS of (2). Then

$$\dot{g}(\lambda) = \sum_i \frac{(x_i - \mu)^2}{\{n - \lambda(x_i - \mu)\}^2} > 0.$$

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

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

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

$$\Big(\frac{n}{x_{(1)} - \mu},\ \frac{n}{x_{(n)} - \mu}\Big).$$

Note for any λ in this interval,

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

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

Remarks. (a) When $\mu = \bar{x} = \bar{X}$, $\lambda = 0$, and

$$p_i(\mu) = 1/n, \qquad i = 1, \cdots, n.$$

It may be shown for μ close to $E(X_i)$, and n large

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

where $S(\mu) = \frac{1}{n}\sum_i (x_i - \mu)^2$.

(b) We may view

$$L(\mu) = L\{p_1(\mu), \cdots, p_n(\mu)\}.$$

as a **profile empirical likelihood** for μ.

Hypothetically consider an 1-1 parameter transformation from $\{p_1, \cdots, p_n\}$ to $\{\mu, \theta_1, \cdots, \theta_{n-1}\}$. Then

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

(c) The likelihood function $L(\mu)$ may be calculated using R-code and Splus-code, downloaded at

http://www-stat.stanford.edu/∼owen/empirical/

(ii) The asymptotic theorem for the classic likelihood ratio tests (i.e. Wilks' Theorem) still holds for the ELR tests.

Let $X_1, \cdots, X_n$ 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

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

where λ is the unique solution of

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

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

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

in distribution as $n \to \infty$.

A sketch proof. Under H_0, $EX_i = \mu_0$. 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 $f(\lambda_n) = 0$, where

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

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

$$\lambda_n \approx -f(0)\big/\dot{f}(0) = -(\bar{X} - \mu_0)\Big/\frac{1}{n}\sum_j (X_j - \mu_0)^2.$$

Now

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

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

2.2 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 $2\log T < \chi^2_{1,1-\alpha}$, where $P\{\chi^2_1 \le \chi^2_{1,1-\alpha}\} = 1-\alpha$. For $\alpha = 0.05$, $\chi^2_{1,1-\alpha} = 3.84$.

Hence a $100(1-\alpha)\%$ confidence interval for μ is

$$\begin{aligned} & \left\{\mu \,\middle|\, -2\log\{L(\mu)n^n\} < \chi^2_{1,1-\alpha}\right\} \\ = & \left\{\mu \,\middle|\, \sum_{i=1}^n \log p_i(\mu) > -0.5\chi^2_{1,1-\alpha} - n\log n\right\} \\ = & \left\{\mu \,\middle|\, \sum_{i=1}^n \log\{np_i(\mu)\} > -0.5\chi^2_{1,1-\alpha}\right\}. \end{aligned}$$

Let $\mu = EX_i$.

$$H_0 : \ \mu = 0 \qquad vs \qquad H_1 : \mu > 0$$

(i) *Standard approach*: Assume $\{X_1, \cdots, X_{15}\}$ is a random sample from $N(\mu, \sigma^2)$

MLE: $\hat{\mu} = \bar{X} = 2.61$

The t-test statistic:

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

Since $T \sim t(14)$ under H_0, the p-value is 0.06 — *significant but not overwhelming*.

Is $N(\mu, \sigma^2)$ an appropriate assumption? as the data do not appear to be normal (with a heavy left tail); see Fig(a).

Example 3. Darwin's data: gains in height of plants from cross-fertilisation

$$X = \text{height(Cross-F) - height(Self-F)}$$

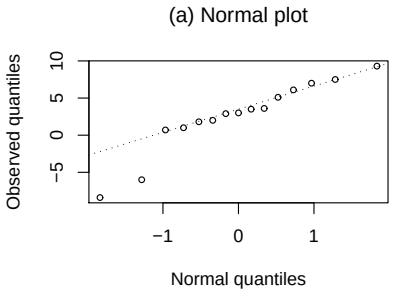
15 observations:

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

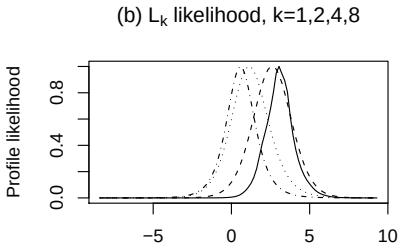
The sample mean $\bar{X} = 2.61$, the standard error $s = 4.71$.

Is the gain significant?

Intuitively: YES, if no two negative observations -8.4 and -6.0.



QQ-plot: Quantile of $N(0,1)$ vs Quantile of the empirical distribution.



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

(ii) Consider a generalised normal family

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

which has the mean μ. When $k=2$, it is $N(\mu,\sigma^2)$.

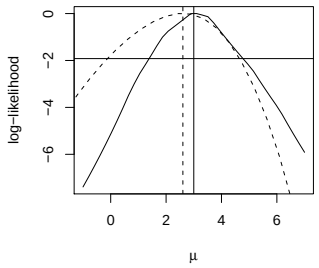
To find the profile likelihood of μ, the 'MLE' for σ is

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

Hence

$$l_k(\mu) = l_k(\mu,\widehat{\sigma}) = -n\log\Gamma(1+\frac{1}{k}) - n(1+\frac{1}{k})\log 2 - n\log\widehat{\sigma} - \frac{n}{k}.$$

Fig.(b) shows the MLE $\widehat{\mu}=\widehat{\mu}(k)$ varies with respect to k. In fact $\widehat{\mu}(k)$ increases as k decreases.



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.)

If we use the distribution functions with $k=1$ to fit the data, the p-value for the test is 0.03 – much more significant than that under the assumption of normal distribution.

(iii) The empirical likelihood ratio test statistic $2\log T = 3.56$, which rejects H_0 with the p-value 0.04.

The 95% confidence interval is

$$\{\mu \,\Big|\, \sum_{i=1}^{15}\log p_i(\mu) > -1.92 - 15\log(15)\} = [0.17, 4.27].$$

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

$$-n\log\sum_i |X_i-\mu|.$$

3. Empirical likelihood for random vectors

Let $\mathbf{X}_1,\cdots,\mathbf{X}_n$ be i.i.d. random vectors from distribution F.

Similar to the univariate case, we assume

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

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

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

Without any further constraints, the MELEs are

$$\widehat{p}_i = 1/n, \quad i=1,\cdots,n.$$

3.1 EL for multivariate means

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

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

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

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

Theorem 3. Let $\mathbf{X}_1, \cdots, \mathbf{X}_n$ be $d \times 1$ i.i.d. with mean $\boldsymbol{\mu}$ and finite covariance matrix $\boldsymbol{\Sigma}$ and $|\boldsymbol{\Sigma}| \neq 0$. Then as $n \to \infty$,

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

in distribution.

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

(ii) The null hypothesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ will be rejected at the significance level α iff

$$\sum_{i=1}^n \log\{np_i(\boldsymbol{\mu}_0)\} \leq -0.5\chi^2_{d,1-\alpha},$$

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

(iii) A $100(1-\alpha)\%$ confidence region for $\boldsymbol{\mu}$ is

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

(iv) *Bootstrap calibration*. Since (ii) and (iii) are based on an asymptotic result. When n is small and d is large, $\chi^2_{d,1-\alpha}$ may be replaced by *the* $[B\alpha]$*-th largest value* among $2\log T^*_1, \cdots,$ $2\log T^*_B$ which are computed as follows.

(a) Draw i.i.d. sample $\mathbf{X}^*_1, \cdots, \mathbf{X}^*_n$ from the uniform distribution on $\{\mathbf{X}_1, \cdots, \mathbf{X}_n\}$. Let

$$T^* = 1 \Big/ \prod_{i=1}^n \{np^*_i(\bar{\mathbf{X}})\},$$

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

(b) Repeat (a) B times, denote the B values of T^* as $T^*_1, \cdots, T^*_B$.

We may draw an $\mathbf{X}^*$ from the uniform distribution on $\{\mathbf{X}_1, \cdots, \mathbf{X}_n\}$ as follows: draw $Z \sim U(0,1)$, define $\mathbf{X}^* = \mathbf{X}_i$ if $Z \in [\frac{i-1}{n}, \frac{i}{n})$.

Since the limiting distribution is free from the original distribution of $\{X_i\}$, we may draw X^*_i from any distribution $\{\pi_1, \cdots, \pi_n\}$ instead of the uniform distribution used above. Of course now $p^*_i(\bar{\mathbf{X}})$ should be replaced by $p^*(\tilde{\mu})$, where $\tilde{\mu} = \sum_i \pi_i \mathbf{X}_i$.

(v) Computing $p_i(\boldsymbol{\mu})$.

Assumptions: $|\mathrm{Var}(\mathbf{X}_i)| \neq 0$, and $\boldsymbol{\mu}$ is an inner point of *the convex hull* spanned by the observations, i.e.

$$\boldsymbol{\mu} \in \Big\{ \sum_{i=1}^n p_i \mathbf{X}_i \,\Big|\, p_i > 0,\ \sum_{i=1}^n p_i = 1 \Big\}.$$

This ensures the solutions $p_i(\boldsymbol{\mu}) > 0$ exist.

We solve the problem in 3 steps:

1. Transform the constrained n-dim problem to a constrained d-dim problem.
2. Transform the constrained problem to an unconstrained problem.
3. Apply a Newton-Raphson algorithm.

Put

$$l(\mu) \equiv \log L(\boldsymbol{\mu}) \;=\; \sum_{i=1}^{n} \log p_i(\boldsymbol{\mu})$$
$$= \max\Big\{ \sum_{i=1}^{n} \log p_i \;\Big|\; p_i \geq 0,\; \sum_{i=1}^{n} p_i = 1,\; \sum_{i=1}^{n} p_i \mathbf{X}_i = \boldsymbol{\mu} \Big\}.$$

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

$$n - \boldsymbol{\lambda}^\tau(\mathbf{X}_i - \boldsymbol{\mu}) > 0, \quad i = 1, \cdots, n. \tag{10}$$

Note. (10) and (8) together imply $\sum_i p_i(\boldsymbol{\mu}) = 1$.

The original n-dimensional optimisation problem is equivalent to a d-dimensional problem of minimising $M(\boldsymbol{\lambda})$ subject to the constraints (10).

Let $\mathcal{H}_\lambda$ be the set consisting all the values of $\boldsymbol{\lambda}$ satisfying

$$n - \boldsymbol{\lambda}^\tau(\mathbf{X}_i - \boldsymbol{\mu}) > 1, \quad i = 1, \cdots, n.$$

Then $\mathcal{H}_\lambda$ a convex set in R^d, which contains the minimiser of the convex function $M(\boldsymbol{\lambda})$. (See 'Note' above)

Unfortunately $M(\boldsymbol{\lambda})$ is not defined on the sets

$$\{\boldsymbol{\lambda} \mid n - \boldsymbol{\lambda}^\tau(\mathbf{X}_i - \boldsymbol{\mu}) = 0\}, \quad i = 1, \cdots, n.$$

Step 1:

Similar to Theorem 1, the LM method entails

$$p_i(\boldsymbol{\mu}) = \frac{1}{n - \boldsymbol{\lambda}^\tau(\mathbf{X}_i - \boldsymbol{\mu})}, \quad i = 1, \cdots, n,$$

where $\boldsymbol{\lambda}$ is the solution of

$$\sum_{j=1}^{n} \frac{\mathbf{X}_j - \boldsymbol{\mu}}{n - \boldsymbol{\lambda}^\tau(\mathbf{X}_j - \boldsymbol{\mu})} = 0. \tag{8}$$

Hence

$$l(\mu) = -\sum_{i=1}^{n} \log\{n - \boldsymbol{\lambda}^\tau(\mathbf{X}_i - \boldsymbol{\mu})\} \equiv M(\boldsymbol{\lambda}). \tag{9}$$

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

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

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

Define

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

It is easy to see that $\log_\star(z)$ has two continuous derivatives on R.

Put $M_\star(\boldsymbol{\lambda}) = -\sum_{i=1}^{n} \log_\star\{n - \boldsymbol{\lambda}^\tau(\mathbf{X}_i - \boldsymbol{\mu})\}$. Then

- $M_\star(\boldsymbol{\lambda}) = M(\boldsymbol{\lambda})$ on $\mathcal{H}_\lambda$.
- $M_\star(\boldsymbol{\lambda})$ is a convex function on whole R^d.

Hence, $M_\star(\boldsymbol{\lambda})$ and $M(\boldsymbol{\lambda})$ share the same minimiser which is the solution of (8).

Step 3: We apply a Newton-Raphson algorithm to compute $\boldsymbol{\lambda}$ iteratively:

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

A convenient initial value would $\boldsymbol{\lambda}_0 = 0$, corresponding to $p_i = 1/n$.

Remarks. (i) S-code "el.S", available from

www-stat.stanford.edu/∼owen/empirical

calculates the empirical likelihood ratio

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

and other related quantities.

Theorem 4. Let $\mathbf{X}_1, \cdots, \mathbf{X}_n$ be $d \times 1$ i.i.d. r.v.s with mean $\boldsymbol{\mu}_0$ and $|\mathrm{Var}(\mathbf{X}_1)| \neq 0$. Let $\boldsymbol{\theta} = h(\boldsymbol{\mu})$ be a smooth function from R^d to R^q $(q \leq d)$, and $\boldsymbol{\theta}_0 = h(\boldsymbol{\mu}_0)$. We assume

$$|GG^\tau| \neq 0, \quad G = \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\mu}^\tau}.$$

For any $r > 0$, let

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

and

$$\mathcal{C}_{3,r} = \left\{\boldsymbol{\theta}_0 + G(\boldsymbol{\mu} - \boldsymbol{\mu}_0) \;\middle|\; \boldsymbol{\mu} \in \mathcal{C}_{1,r}\right\}.$$

Then as $n \to \infty$,

$$P\{\boldsymbol{\theta} \in \mathcal{C}_{3,r}\} \to P(\chi^2_q \leq r).$$

3.2 EL for smooth functions of means

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

$$\sigma^2 = EY_i^2 - (EY_i)^2 = h(\boldsymbol{\mu}),$$

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

Remarks. (i) The idea of bootstrap calibration may be applied here too.

(ii) Under more conditions, $P\{\boldsymbol{\theta} \in \mathcal{C}_{2,r}\} \to P(\chi^2_q \leq r)$, where

$$\mathcal{C}_{2,r} = \left\{h(\boldsymbol{\mu}) \;\middle|\; \boldsymbol{\mu} \in \mathcal{C}_{1,r}\right\}.$$

$\mathcal{C}_{2,r}$ is a practical feasible confidence set, while $\mathcal{C}_{3,r}$ is not since μ_0 and θ_0 are unknown in practice. Note for $\boldsymbol{\mu}$ close to $\boldsymbol{\mu}_0$,

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

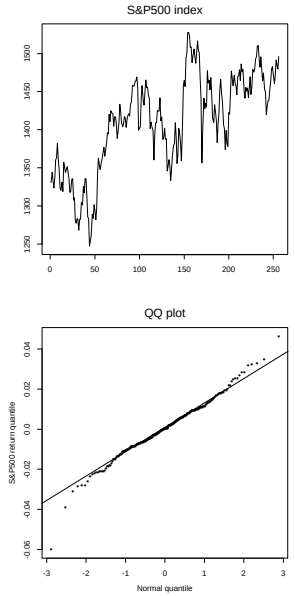
(iii) In general, $P\{\boldsymbol{\mu} \in \mathcal{C}_{1,r}\} \leq P\{\boldsymbol{\theta} \in \mathcal{C}_{2,r}\}$.

(By Theorem 3, $P\{\boldsymbol{\mu} \in \mathcal{C}_{1,r}\} \to P(\chi^2_d \leq r)$)

(iv) The profile empirical likelihood function of $\boldsymbol{\theta}$ is

$$\begin{aligned} L(\boldsymbol{\theta}) &= \max\Big\{ \prod_{i=1}^{n} p_i(\boldsymbol{\mu}) \,\Big|\, h(\boldsymbol{\mu}) = \boldsymbol{\theta} \Big\} \\ &= \max\Big\{ \prod_{i=1}^{n} p_i \,\Big|\, h\Big(\sum_{i=1}^{n} p_i \mathbf{X}_i \Big) = \boldsymbol{\theta},\ p_i \geq 0, \\ &\qquad \sum_{i=1}^{n} p_i = 1 \Big\}, \end{aligned}$$

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



Example 4. 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,

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

which is the return, i.e. the percentage of the change on the i-th day.

By treating 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

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

The 95% confidence intervals are:

Method	C.I.
EL	[0.1895, 0.2422]
Normal	[0.1950, 0.2322]

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.

4. Estimating equations

4.1 Estimation via estimating equations

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

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

where $\boldsymbol{\theta}$ is $q \times 1$ vector, m is a $s \times 1$ vector-valued function.

For example,

- $\theta = EX_1$ if $m(x, \theta) = x - \theta$,
- $\theta = E(X_1^k)$ if $m(x, \theta) = x^k - \theta$,
- $\theta = P(X_1 \in A)$ if $m(x, \theta) = I(x \in A) - \theta$,
- θ is the α-quantile if $m(x, \theta) = I(x \le \theta) - \alpha$.

Example 5. Let $\{(X_i, Y_i),\ i = 1, \cdots, n\}$ be a random sample. Find a set of estimating equations for estimating $\gamma \equiv \mathrm{Var}(X_1)/\mathrm{Var}(Y_1)$.

In order to estimate γ, we need to estimate $\mu_x = E(X_1)$, $\mu_y = E(Y_1)$ and $\sigma_y^2 = \mathrm{Var}(Y_1)$. Put $\boldsymbol{\theta}^\tau = (\mu_x, \mu_y, \sigma_y^2, \gamma)$, and

$$m_1(X, Y, \boldsymbol{\theta}) = X - \mu_x, \quad m_2(X, Y, \boldsymbol{\theta}) = Y - \mu_y,$$

$$m_3(X, Y, \boldsymbol{\theta}) = (Y - \mu_y)^2 - \sigma_y^2,$$

$$m_4(X, Y, \boldsymbol{\theta}) = (X - \mu_x)^2 - \sigma_y^2\gamma,$$

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

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

the solution of the above equation is an estimator $\widehat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$.

Remark. Estimating equation method does not facilitate hypothesis tests and interval estimation for $\boldsymbol{\theta}$.

A natural estimator for $\boldsymbol{\theta}$ is determined by *the estimating equation*

$$\frac{1}{n}\sum_{i=1}^n m(\mathbf{X}_i, \widehat{\boldsymbol{\theta}}) = 0. \tag{11}$$

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

Determined case $q = s$: $\widehat{\boldsymbol{\theta}}$ may be uniquely determined by (11)

Underdetermined case $q > s$: the solutions of (11) may form a $(q - s)$-dimensional set

Overdetermined case $q < s$: (11) 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.

4.2 EL for estimating equations

Aim: construct statistical tests and confidence intervals for $\boldsymbol{\theta}$

The profile empirical likelihood function of $\boldsymbol{\theta}$:

$$L(\boldsymbol{\theta}) = \max\{\prod_{i=1}^n p_i \,\Big|\, \sum_{i=1}^n p_i m(\mathbf{X}_i, \boldsymbol{\theta}) = 0,\ p_i \ge 0,\ \sum_{i=1}^n p_i = 1\}$$

The following theorem follows from Theorem 2 immediately.

Theorem 5. Let $\mathbf{X}_1, \cdots, \mathbf{X}_n$ be i.i.d., $m(\mathbf{x}, \boldsymbol{\theta})$ be an $s \times 1$ vector-valued function. Suppose

$$E\{m(\mathbf{X}_1, \boldsymbol{\theta}_0)\} = 0, \quad \Big|\mathrm{Var}\{m(\mathbf{X}_1, \boldsymbol{\theta}_0)\}\Big| \neq 0.$$

Then as $n \to \infty$,

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

in distribution.

The theorem above applies in all *determined, underdetermined* and *overdetermined* cases.

Remarks (i) In general $L(\boldsymbol{\theta})$ can be calculated using the method for EL for multivariate means in §3.1, treating $m(\mathbf{X}_i, \boldsymbol{\theta})$ as a random vector.

(ii) For $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ which is the solution of

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

$L(\widehat{\boldsymbol{\theta}}) = (1/n)^n$.

(iii) For $\boldsymbol{\theta}$ determined by $E\{m(\mathbf{X}_1, \boldsymbol{\theta})\} = 0$, we will reject the null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ iff

$$\log\{L(\boldsymbol{\theta}_0)\} + n\log n \le -0.5\chi^2_{s,1-\alpha}.$$

(iii) An $(1-\alpha)$ confidence set for $\boldsymbol{\theta}$ determined by $E\{m(\mathbf{X}_1, \boldsymbol{\theta})\} = 0$ is

$$\{\boldsymbol{\theta} \;\Big|\; \log\{L(\boldsymbol{\theta})\} + n\log n > -0.5\chi^2_{s,1-\alpha}\}$$

Let

$$\begin{aligned} L(\theta_\alpha) \;=\; \max\Big\{ \prod_{i=1}^{n} p_i \;\Big|\; \sum_{i=1}^{n} p_i I(X_i \le \theta_\alpha) = \alpha, \\ p_i \ge 0,\ \sum_{i=1}^{n} p_i = 1\Big\}. \end{aligned}$$

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

$$\Theta_\alpha = \{\theta_\alpha \;\Big|\; \log\{L(\theta_\alpha)\} > -n\log n - 0.5\chi^2_{1,1-\beta}\}.$$

Note $L(\widehat{\theta}_\alpha) = (1/n)^n \ge L(\theta_\alpha)$ for any θ_α. It is always true that $\widehat{\theta}_\alpha \in \Theta_\alpha$.

Example 6. (Confidence intervals for quantiles)

Let $X_1, \cdots, X_n$ be i.i.d. For a given $\alpha \in (0,1)$, let

$$m(x, \theta_\alpha) = I(x \le \theta_\alpha) - \alpha.$$

Then $E\{m(X_i, \theta_\alpha)\} = 0$ 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

$$\sum_{i=1}^{n} m(X_i, \widehat{\theta}_\alpha) = \sum_{i=1}^{n} I(X_i \le \theta_\alpha) - n\alpha = 0$$

entails

$$\widehat{\theta}_\alpha = X_{(n\alpha)},$$

where $X_{(i)}$ denotes the i-th smallest value among $X_1, \cdots, X_n$. We assume $n\alpha$ is an integer to avoid insignificant (for large n, e.g. $n = 100$) technical details.

In fact $L(\theta_\alpha)$ can be computed explicitly as follows.

Let $r = r(\theta_\alpha)$ be the integer for which

$X_{(i)} \le \theta_\alpha$ for $i = 1, \cdots, r$, and

$X_{(i)} > \theta_\alpha$ for $i = r+1, \cdots, n$.

Thus

$$\begin{aligned} L(\theta_\alpha) \;&=\; \max\Big\{ \prod_{i=1}^{n} p_i \;\Big|\; p_i \ge 0,\ \sum_{i=1}^{r} p_i = \alpha,\ \sum_{i=r+1}^{n} p_i = 1-\alpha\Big\} \\ &=\; (\alpha/r)^r\{(1-\alpha)/(n-r)\}^{n-r}. \end{aligned}$$

Hence

$$\begin{aligned}\Theta_\alpha &= \{\theta_\alpha \,\big|\, \log\{L(\boldsymbol{\theta}_\alpha)\} > -n\log n - 0.5\chi^2_{1,1-\alpha}\} \\ &= \left\{\theta_\alpha \,\middle|\, r\log\frac{n\alpha}{r} + (n-r)\log\frac{n(1-\alpha)}{n-r} > -0.5\chi^2_{1,1-\alpha}\right\},\end{aligned}$$

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

5. Empirical likelihood for estimating conditional distributions

References on kernel regression:

- Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. Springer, New York.
- Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing*. Chapman and Hall, London.

References on nonparametric estimation for distribution functions:

- Hall, P., Wolff, R.C.L. and Yao, Q. (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association*, **94**, 154-163.
- Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer, New York. Sections 10.3 (also Section 6.5).

5.1 From global fitting to local fitting

Consider linear regression model

$$Y = X_1\beta_1 + \cdots + X_d\beta_d + \varepsilon, \tag{12}$$

where $\varepsilon \sim (0, \sigma^2)$.

This model is *linear wrt unknown coefficients* $\beta_1, \cdots, \beta_d$ as the variable $X_1, \cdots, X_d$ may be

- quantitative inputs
- transformations of quantitative inputs, such as log, square-root etc
- interactions between variables, e.g. $X_3 = X_1X_2$
- basis expansions, such as $X_2 = X_1^2, X_3 = X_1^3, \cdots$
- numeric or "dummy" coding of the levels if qualitative inputs

Put $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_d)^\tau$

With observations $\{(Y_i, \mathbf{X}_i),\ 1 \le i \le n\}$, where $\mathbf{X}_i = (X_{i1}, \cdots, X_{id})^\tau$, the LSE minimises

$$\sum_{i=1}^{n}\left(Y_i - \mathbf{X}_i^\tau\boldsymbol{\beta}\right)^2, \tag{13}$$

resulting to

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau\mathbf{Y},$$

where $\mathbf{Y} = (Y_1, \cdots, Y_n)^\tau$, and $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_n)^\tau$ is an $n \times d$ matrix.

The fitted model is

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}.$$

This is a **global** fitting, since the model is assumed to be true everywhere in the sample space and the estimator $\widehat{\boldsymbol{\beta}}$ is obtained using all the available data.