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

Kurtosis: $\kappa = \frac{E\{(X-EX)^4\}}{\{\text{Var}(X)\}^2} - 3$, — 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 $\hat{\sigma}^2 = (n-1)^{-1} \sum_{1 \le i \le n} (X_i - \bar{X})^2$.

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

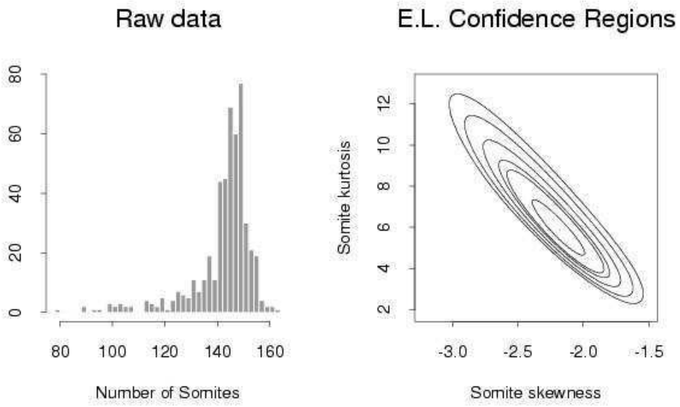
How to find the confidence sets for (γ, κ)?

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

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

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

Two problems:

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

(i) The constrained MELEs $\tilde{p}_i = p_i(\mu_0)$, 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

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

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

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

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

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

(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

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

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,

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

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, $n^{-1}\sum_i (X_i - \mu_0)^2 \to \text{Var}(X_1)$. By the CLT, $\sqrt{n}(\bar{X} - \mu_0) \to N(0, \text{Var}(X_1))$ 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:

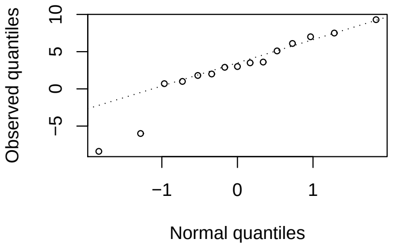
6.1, -8.4, 1.0, 2.0, 0.7, 2.9, 3.5, 5.1, 1.8, 3.6, 7.0, 3.0, 9.3, 7.5, -6.0

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

Is the gain significant?

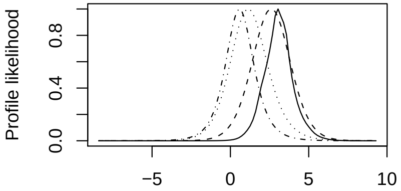
Intuitively: YES, if no two negative observations -8.4 and -6.0.

(a) Normal plot



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

(b) L$_k$ likelihood, k=1,2,4,8



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 $N(\mu, \sigma^2)$.

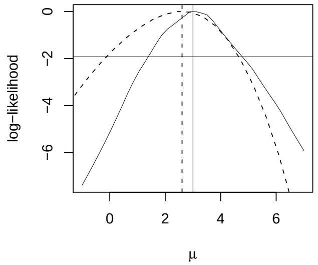
To find the profile likelihood of μ, the 'MLE' for σ is

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

Hence

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

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

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

3.1 EL for multivariate means

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

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

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 Σ and $|\Sigma| \neq 0$. 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 $|\Sigma| = 0$, 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 $|\text{Var}(\mathbf{Y}_i)| \neq 0$, 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

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

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

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

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

(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: $|\text{Var}(\mathbf{X}_i)| \neq 0$, 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

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

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

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

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 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

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

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

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

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 $\boldsymbol{\mu}$.

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 $|\text{Var}(\mathbf{X}_1)| \neq 0$. 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

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

and

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

Then as $n \to \infty$,

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

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

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

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

$\mathcal{C}_{2,r}$ 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 $\boldsymbol{\mu}_0$,

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

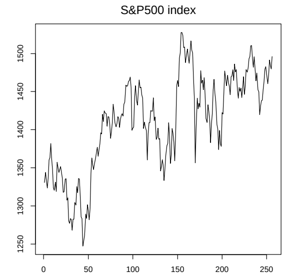
(iii) In general, $P\{\boldsymbol{\mu} \in \mathcal{C}_{1,r}\} \leq P\{\boldsymbol{\theta} \in \mathcal{C}_{2,r}\}$.

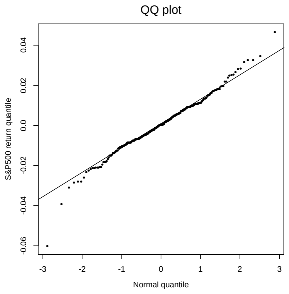
(By Theorem 3, $P\{\boldsymbol{\mu} \in \mathcal{C}_{1,r}\} \to P(\chi_d^2 \leq r)$)

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

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

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

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.

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

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

4.2 EL for estimating equations

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

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

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

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

$$E\{m(\mathbf{X}_1, \boldsymbol{\theta}_0)\} = 0, \qquad \Big|\text{Var}\{m(\mathbf{X}_1, \boldsymbol{\theta}_0)\}\Big| \ne 0.$$

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

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

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.

Let

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

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 $\widehat{\theta}_\alpha \in \Theta_\alpha$.

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

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

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.

Such a global fitting is efficient **if** the assumed form of the regression function (12) is correct.

In general (12) may be incorrect globally. But it may provide a reasonable approximation at any small area in the sample space. We fit for each given small area a different linear model — This is the basic idea of **local** fitting.

Technically, a local fitting may be achieved by adding a weight function in (13) as follows.

Suppose we fit a local linear model in a small neighbourhood of the observation $\mathbf{X}_k$, with the coefficient $\boldsymbol{\beta} = \boldsymbol{\beta}_k$. the LSE minimises

$$\sum_{i=1}^{n} \left(Y_i - \mathbf{X}_i^\tau \boldsymbol{\beta}_k\right)^2 w(\mathbf{X}_i, \mathbf{X}_k) \tag{14}$$

where the weight function may be taken as

$$w(\mathbf{X}_i, \mathbf{X}_k) = \begin{cases} 1 & \text{if } \mathbf{X}_i \text{ is among the } p \text{ nearest neighbours of } \mathbf{X}_k, \\ 0 & \text{otherwise,} \end{cases}$$

where $p \geq 1$ is a prescribed small integer.

Although the sum in (14) only has p non-zero terms, the local LSE can be expressed formally as

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

where $\mathbf{W} = \text{diag}\{w(\mathbf{X}_1, \mathbf{X}_k), \cdots, w(\mathbf{X}_n, \mathbf{X}_k)\}$.

Remark. (i) The local estimator $\widehat{\boldsymbol{\beta}}_k$ only makes use of the p (out of n) observations around $\mathbf{X}_k$, *may depend on the choice of p sensitively*.

(ii) Intuitively the local estimator $\widehat{\boldsymbol{\beta}}_k$ may catch some local structure better than the global estimator $\widehat{\boldsymbol{\beta}}$. But the variance of $\widehat{\boldsymbol{\beta}}_k$ is larger than that of $\widehat{\boldsymbol{\beta}}$.

Example. (*Linear model for classification*)

We have two sets of pair observation (X_1, X_2), each of size 100. The two sets were generated from two different distributions; Red and Green. Define

$$Y = \begin{cases} 1 & (X_1, X_2) \sim \text{Red}, \\ 0 & (X_1, X_2) \sim \text{Green}. \end{cases}$$

Putting the two sets of data together, we fit a linear model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon,$$

the LSE leads to the predictor

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 X_2.$$

We may classify a new observation (X_1, X_2) into the class Red if $\widehat{Y} > 0.5$, into Green if $\widehat{Y} \leq 0.5$. *This effectively divides the whole (x_1, x_2)-plane into two half-planes by the straight line $\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 = 0.5$.*

Elements of Statistical Learning ©Hastie, Tibshirani & Friedman 2001 Chapter 2

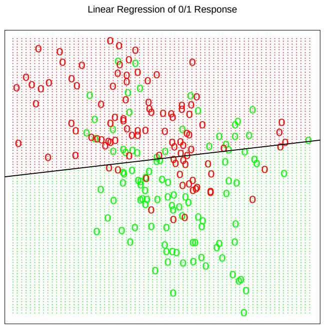


Figure 2.1: *A classification example in two dimensions. The classes are coded as a binary variable—* GREEN $= 0$, RED $= 1$*—and then fit by linear regression. The line is the decision boundary defined by $x^T\hat{\beta} = 0.5$. The red shaded region denotes that part of input space classified as* RED*, while the green region is classified as* GREEN.

There are couple of misspecifications on both sides of the linear boundary; *indicating that the linear model is too rigid*. We employ a local fitting as follows.

For given $\mathbf{x} = (x_1, x_2)$, we fit a local model

$$\widehat{y} \equiv \widehat{y}(\mathbf{x}) = \widehat{\beta}_0(\mathbf{x}) + \widehat{\beta}_1(\mathbf{x})x_1 + \widehat{\beta}_2(\mathbf{x})x_2,$$

where $(\widehat{\beta}_0(\mathbf{x}), \widehat{\beta}_1(\mathbf{x}), \widehat{\beta}_2(\mathbf{x}))$ minimises

$$\sum_i \{Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2}\}^2 w_i(x),$$

and $w_i(x) = 1$ if (X_{i1}, X_{i2}) is among *the d nearest neighbours* of (x_1, x_2), and 0 otherwise.

Remark. The estimation depends on the choice of d sensitively.

Elements of Statistical Learning ©Hastie, Tibshirani & Friedman 2001 Chapter 2

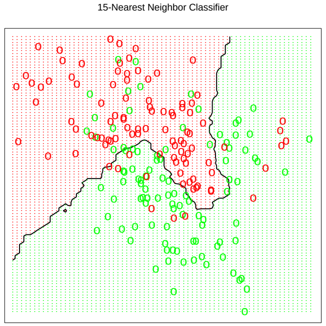


Figure 2.2: *The same classification example in two dimensions as in Figure 2.1. The classes are coded as a binary variable* (GREEN $= 0$, RED $= 1$) *and then fit by* 15*-nearest-neighbor averaging as in (2.8). The predicted class is hence chosen by majority vote amongst the* 15*-nearest neighbors.*

Elements of Statistical Learning ©Hastie, Tibshirani & Friedman 2001 Chapter 2

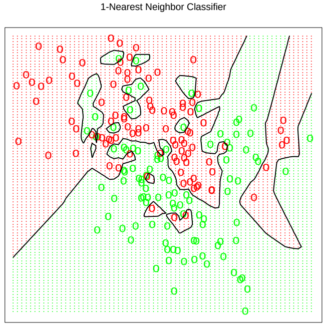


Figure 2.3: *The same classification example in two dimensions as in Figure 2.1. The classes are coded as a binary variable* (GREEN $= 0$, RED $= 1$)*, and then predicted by* 1*-nearest-neighbor classification.*

5.2 Kernel Methods

5.2.1 Introduction

We observe $\{(Y_i, X_i), i = 1, \cdots, 100\}$ from

$$Y_i = f(X_i) + \varepsilon_i, \quad \varepsilon_i \sim (0, \sigma^2)$$

where $f(\cdot)$ is a unknown and *smooth* function.

We may use the idea of **local smoothing** to estimate f:

$$\begin{aligned}\widehat{f}(x) &= \text{the average of those } Y_i \text{ for which } X_i \text{ is} \\ &\quad\ \text{among the } k \text{ nearest neighbours of } x \\ &= \frac{1}{k}\sum_{i=1}^n Y_i w(x, X_i) \;=\; \sum_{i=1}^n Y_i w(x, X_i) \Big/ \sum_{i=1}^n w(x, X_i),\end{aligned}$$

where $w(x, X_i) = 1$ if X_i is among the k nearest neighbours of x, and 0 otherwise.

We may give more weights to X_i closer to x, i.e. let $w(x, X_i) = w(|x - X_i|)$ be a monotonically decreasing function.

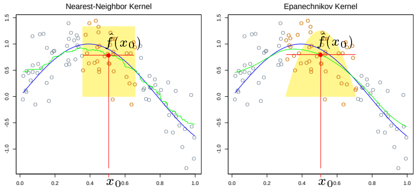


Figure 6.1: *In each panel 100 pairs x_i, y_i are generated at random from the blue curve with Gaussian errors: $Y = \sin(4X) + \varepsilon$, $X \sim U[0,1]$, $\varepsilon \sim N(0, 1/3)$. In the left panel the green curve is the result of a 30-nearest-neighbor running-mean smoother. The red point is the fitted constant $\hat{f}(x_0)$, and the orange shaded circles indicate those observations contributing to the fit at x_0. The solid orange region indicates the weights assigned to observations. In the right panel, the green curve is the kernel-weighted average, using an Epanechnikov kernel with (half) window width $\lambda = 0.2$.*

5.2.2 Nadaraya-Watson estimator.

$$Y_i = f(X_i) + \varepsilon_i.$$

Instead of specifying k – the number of neighbours used in estimation, we may determine the number by choosing

$$w(x, X_i) = K\Big(\frac{X_i - x}{h}\Big),$$

where $K(\cdot) \geq 0$ is a *kernel function*, and $h > 0$ is a *bandwidth*. Conventionally, we use K such that $\int K(u)du = 1$.

When, for example, $K(x) = \frac{1}{2}I(|x| \leq 1)$, only those X_i within h distance from x are used in estimating $f(x)$. The number of those points may vary wrt x.

The resulting estimator

$$\hat{f}(x) = \sum_{i=1}^{n} Y_i K\Big(\frac{X_i - x}{h}\Big) \Big/ \sum_{i=1}^{n} K\Big(\frac{X_i - x}{h}\Big)$$

is called a *Nadaraya-Watson estimator*.

In fact, $\hat{f}(\cdot)$ is a *local LSE*, since

$$\hat{f}(x) = \arg\min_a \sum_{i=1}^{n} \{Y_i - a\}^2 K\Big(\frac{X_i - x}{h}\Big).$$

Therefore, $\hat{f}(\cdot)$ is also called *local constant regression estimator*.

Remarks. (i) Commonly used kernel functions:

- Gaussian kernel $K(x) = (2\pi)^{-1/2} \exp(-x^2/2)$
- Epanechnikov kernel $K(x) = \frac{3}{4}(1 - x^2)I(|x| \leq 1)$
- Tri-cube kernel $K(x) = (1 - |x|^3)^3 I(|x| \leq 1)$

Both Epanechnikov and tri-cube kernels have compact support $[-1, 1]$ while Gaussian kernel has infinite support.

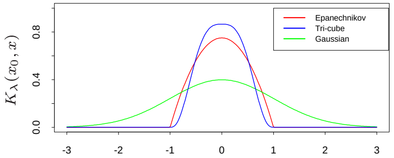


Figure 6.2: *A comparison of three popular kernels for local smoothing. Each has been calibrated to integrate to 1. The tri-cube kernel is compact and has two continuous derivatives at the boundary of its support, while the Epanechnikov kernel has none. The Gaussian kernel is continuously differentiable, but has infinite support.*

(ii) The bandwidth h controls the amount of data used in local estimation, determines the smoothness of the estimated curve $\widehat{f}(\cdot)$. For example, with $K(x) = 0.5I(|x| \le 1)$, $\widehat{f}(x) \to \bar{Y}$ as $h \to \infty$ — global constant fitting; $\widehat{f}(X_i) \to Y_i$ as $h \to 0$ — interpolating the observations.

h is also called a *smooth parameter*.

(iii) The goodness of the estimator $\widehat{f}(\cdot)$ *depends on the bandwidth h sensitively*, while the difference from using different kernel functions may be absorbed to a large extent by adjusting the value of h accordingly.

Bias and variance of $\widehat{f}(x)$

Regularity conditions: (i) $\{(Y_i, X_i)\}$ are i.i.d, and

$$f(x) = E(Y_i|X_i = x), \quad \varepsilon_i = Y_i - f(X_i).$$

Further both $f(\cdot)$ and $p(\cdot)$ have two continuous derivatives, where $p(\cdot)$ denotes the pdf of X_i.

(ii) $K(\cdot)$ is a symmetric density function with a bounded support, and $n \to \infty$, $h \to 0$ and $nh \to \infty$.

Put $\sigma_0^2 = \int u^2 K(u)du$, $\mathbf{X} = (X_1, \cdots, X_n)^\tau$ and

$$\sigma(x)^2 = \text{Var}(Y_i|X_i = x) = E(Y_i^2|X_i = x) - f(x)^2.$$

We write $\xi_n \sim \eta_n$ iff $\xi_n/\eta_n \to 1$ in probability.

Theorem. Under conditions (i) and (ii) above, it holds that for x with $p(x) > 0$

$$\begin{aligned} E\{\widehat{f}(x) - f(x)|\mathbf{X}\} &\sim \frac{h^2\sigma_0^2}{2}\Big\{\ddot{f}(x) + \frac{2\dot{f}(x)\dot{p}(x)}{p(x)}\Big\}, \\ \text{Var}\{\widehat{f}(x)|\mathbf{X}\} &\sim \frac{1}{nh}\frac{\sigma(x)^2}{p(x)}\int K(u)^2 du. \end{aligned}$$

Proof. We only give a sketchy proof for the bias.

Put $K_i = h^{-1}K(\frac{X_i - x}{h})$. Then

$$\widehat{f}(x) = \sum_i Y_i K_i \Big/ \sum_i K_i.$$

Note (i) implies $E\{\varepsilon_i|\mathbf{X}\} = E\{\varepsilon_i|X_i\} = 0$. Hence

$$\begin{aligned} E\{\widehat{f}(x) - f(x)|\mathbf{X}\} &= \sum_i E\{Y_i - f(x)|\mathbf{X}\}K_i \Big/ \sum_i K_i \\ &= \sum_i \{f(X_i) - f(x)\}K_i \Big/ \sum_i K_i. \end{aligned}$$

It follows the LLN that

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n K_i \sim E(K_1) &= \int \frac{1}{h}K\Big(\frac{X - x}{h}\Big)p(X)dX \\ &= \int K(u)p(x + hu)du \ \to \ p(x), \end{aligned} \tag{15}$$

and

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^n \{f(X_i) - f(x)\}K_i \\ \sim\ & \int \{f(X) - f(x)\}\frac{1}{h}K\Big(\frac{X - x}{h}\Big)p(X)dX \\ =\ & \int \{f(x + hu) - f(x)\}K(u)p(x + hu)du \\ =\ & \int \{hu\dot{f}(x) + \frac{h^2u^2}{2}\ddot{f}(x)\}\{p(x) + hu\dot{p}(x)\}K(u)du + O(h^3) \\ =\ & h^2\sigma_0^2\{\dot{f}(x)\dot{p}(x) + \frac{1}{2}\ddot{f}(x)p(x)\} + O(h^3). \end{aligned} \tag{16}$$

Combining (15) and (16), we obtain the required asymptotic formula for the bias.

Remarks. (i) An approximate MSE:

$$E[\{\widehat{f}(x)-f(x)\}^2|\mathbf{X}] \;=\; \text{Bias}^2 + \text{Variance}$$

$$\approx \quad \frac{h^4\sigma_0^4}{4}\Big\{\ddot{f}(x)+\frac{2\dot{f}(x)\dot{p}(x)}{p(x)}\Big\}^2+\frac{1}{nh}\frac{\sigma(x)^2}{p(x)}\int K(u)^2du$$

Increasing h, Variance decreases and Bias increases. A good choice of h is a trade-off between the variance and the bias. Minimising the RHS of the above over h, we obtain

$$h_{op}=n^{-1/5}C(x),$$

where $C(x)$ is a function of x, depending on p, f and K. Note that $C(x)$ is unknown in practice.

(ii) It can be shown that

$$\sqrt{nh}\Big[\widehat{f}(x)-f(x)-\frac{h^2\sigma_0^2}{2}\Big\{\ddot{f}(x)+\frac{2\dot{f}(x)\dot{p}(x)}{p(x)}\Big\}\Big]$$

converges in distribution to

$$N\Big(0,\ \frac{\sigma(x)^2}{p(x)}\int K(u)^2du\Big).$$

Note that the convergence rate is $\sqrt{nh}$ (instead of the standard $\sqrt{n}$). This reflects the nature of local estimation; effectively only the date lying within h-distance from given x are used in estimation, and the number of those data is of the size nh.

5.2.3 Kernel density estimation

From (5), a natural estimator for the density function of X_i is

$$\widehat{p}(x)=\frac{1}{nh}\sum_{i=1}^n K\Big(\frac{X_i-x}{h}\Big),$$

which is called a *kernel density estimator*.

(5) implies $\widehat{p}(x)$ is a consistent estimator. Further,

$$E\{\widehat{p}(x)\}=p(x)+O(h^2).$$

5.2.4 Local linear regression estimation

The Nadaraya-Watson estimation is a local constant estimation, i.e. for y in a small neighbourhood of x, we approximate

$$f(y)\approx f(x),$$

and minimise

$$\sum_{i=1}^n\{Y_i-a\}^2K\Big(\frac{X_i-x}{h}\Big).$$

Intuitively, the estimation may be improved by using a *local linear approximation*:

$$f(y)\approx f(x)+\dot{f}(x)(y-x),$$

this leads to the **local linear regression estimator**: $\widehat{f}(x)\equiv\widehat{a}$, where $(\widehat{a},\widehat{b})$ minimises

$$\sum_{i=1}^n\{Y_i-a-b(X_i-x)\}^2K\Big(\frac{X_i-x}{h}\Big). \tag{17}$$

Obviously a natural estimator for $\dot{f}$ is $\widehat{\dot{f}}(x)\equiv\widehat{b}$.

Let $\mathcal{Y} = (Y_1, \cdots, Y_n)^\tau$, $\boldsymbol{\theta} = (a, b)^\tau$, $\mathcal{X}$ be an $n \times 2$ matrix with $(1, (X_i - x))$ as its i-row, and $\mathcal{K}$ is an $n \times n$ diagonal matrix with $K(\frac{X_i - x}{h})$ as its (i,i)-th element. Then (17) can be written as

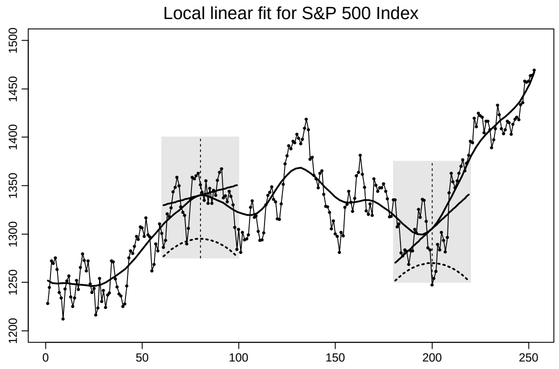
$$(\mathcal{Y} - \mathcal{X}\boldsymbol{\theta})^\tau \mathcal{K} (\mathcal{Y} - \mathcal{X}\boldsymbol{\theta}).$$

Therefore, the LSE method leads to

$$\begin{pmatrix} \widehat{f}(x) \\ \widehat{\dot{f}}(x) \end{pmatrix} \equiv \widehat{\boldsymbol{\theta}} = (\mathcal{X}^\tau \mathcal{K} \mathcal{X})^{-1} \mathcal{X}^\tau \mathcal{K} \mathcal{Y}. \qquad (18)$$

Hence like the Nadaraya-Watson estimator, the local linear estimator for $f(x)$ is *a linear combination of* $Y_1, \cdots, Y_n$ (given $\mathbf{X} = (X_1, \cdots, X_n)^\tau$). Such an estimator is called a **linear estimator**.

Note. Both Nadaraya-Watson estimator and local linear estimator with prescribed bandwidth h can be computed using S-function *'lls.s'*. Splus and R function 'loess' offers more flexibilities for local regression fitting.



Local linear fit for the S&P 500 Index from January 4, 1999 to December 31, 1999, using the Epanechnikov kernel and bandwidth $h = 20$. *The dashed parabola in each window indicates the weight that each local data point receives.*

Why is a local linear estimator better?

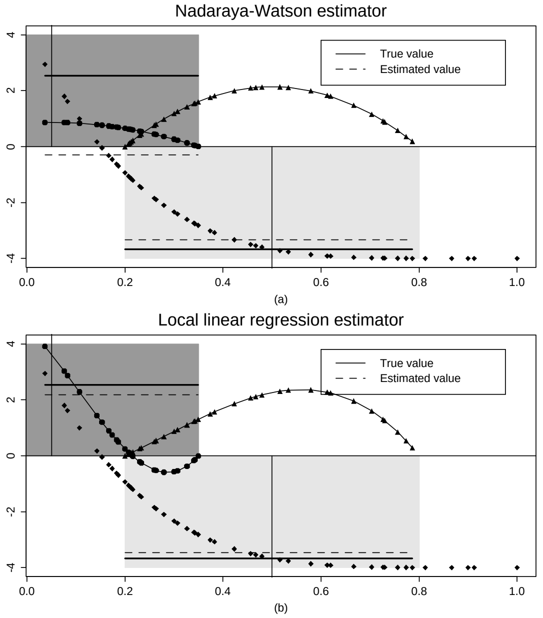
- Simpler (and often smaller) bias formula
- Automatic boundary carpentry

The table below lists the (first order) biases and variances of the Nadaraya-Watson estimator (N-W) and the local linear estimator (LL).

	Bias	Variance
N-W	$\frac{h^2\sigma_0^2}{2}\left\{\ddot{f}(x) + \frac{2\dot{f}(x)\dot{p}(x)}{p(x)}\right\}$	$\frac{1}{nh}\frac{\sigma(x)^2}{p(x)}\int K(u)^2 du$
LL	$\frac{h^2\sigma_0^2}{2}\ddot{f}(x)$	$\frac{1}{nh}\frac{\sigma(x)^2}{p(x)}\int K(u)^2 du$

The asymptotic MSE of the local linear estimator:

$$E[\{\widehat{f}(x) - f(x)\}^2|\mathbf{X}] \approx \frac{h^4\sigma_0^4}{4}\{\ddot{f}(x)\}^2 + \frac{1}{nh}\frac{\sigma(x)^2}{p(x)}\int K(u)^2 du.$$



Effective weights assigned to local data points at an interior point $x_0 = 0.5$ (weights denoted by ◀) and a boundary point $x_0 = 0.05$ (weights denoted by ∘) for the local constant fit (Nadaraya-Watson method) and the local linear fit, with K being the Epanechnikov kernel. The horizontal solid and dashed lines are the heights of true and estimated functions at $x_0 = 0.05$ and $x_0 = 0.5$, respectively. Their differences are biases at these two points. (a) The Nadaraya–Watson estimator; (b) the local linear fit. For clarity, the data (⧫) contain no noise.

A local linear estimator is *automatically adaptive at the boundary* in the sense that the biases at the boundary points are smaller than those of a Nadaraya-Watson estimator.

Nadaraya-Watson estimator:

$$\widehat{F}_{nw}(y|x) = \sum_{i=1}^n I(Y_i \le y) K\Big(\frac{X_i - x}{h}\Big) \Big/ \sum_{i=1}^n K\Big(\frac{X_i - x}{h}\Big) = \sum_{i=1}^n Z_i w_i(x),$$

where $Z_i = I(Y_i \le y)$, and

$$w_i(x) = K\Big(\frac{X_i - x}{h}\Big) \Big/ \sum_{j=1}^n K\Big(\frac{X_j - x}{h}\Big) \ge 1, \qquad \sum_{i=1}^n w_i(x) = 1.$$

In the above expression, $K(\cdot)$ is a pdf and $h > 0$ is a bandwidth.

$\widehat{F}_{nw}(y|x)$ itself is a proper distribution function!

In fact, $\widehat{F}_{nw}(y|x)$ is a local constant estimator in the sense that it minimizes

$$L(a) = \sum_{i=1}^n w_i(x)(Z_i - a)^2.$$

If we replace $w_i(x)$ by $1/n$, we obtain the global estimator $\bar{Z}$.

5.3 Estimation for conditional distributions

Observations: $\{(X_1, Y_1), \cdots, (X_n, Y_n)\}$ i.i.d.

Let $F(\cdot|x)$ denote the conditional distribution of Y_i given $X_i = x$.

Goal: estimate $F(y|x)$ nonparametrically.

Motivation: quantile regression, prediction and etc.

5.3.1 Nadaraya-Watson and local linear estimators

Note: $E\{I(Y_i \le y)|X_i = x\} = F(y|x)$

Hence $G(y|x)$ is a regression of $Z_i \equiv I(Y_i \le y)$ on X_i as $E(Z_i|X_i) = F(y|X_i)$.

Local linear estimator: $\widehat{F}_{ll}(y|x) \equiv \widehat{a}$, where $(\widehat{a}, \widehat{b})$ minimizes

$$\sum_{i=1}^n w_i(x)\{Z_i - a - b(X_i - x)\}^2.$$

Note. If we replace $w_i(x)$ by $1/n$, this is the standard linear regression estimation: $\widehat{Z}_i = \widehat{a} + \widehat{b}(X_i - x)$.

$\widehat{F}_{ll}(y|x)$ has superior bias properties (and other types of efficiency)

But $\widehat{F}_{ll}(y|x)$ is not necessarily a distribution function, as it may take value outside the interval $[0, 1]$, and is not necessarily monotonically increasing in y.

An ideal estimator: combine the advantages of both $\widehat{F}_{nw}(y|x)$ and $\widehat{F}_{ll}(y|x)$ together.

Write $Z_i \equiv I_{\{Y_i \le y\}} = F(y|X_i) + \epsilon_i$, and $K_h(x) = h^{-1}(x/h)$.

Let $g(\cdot)$ be the pdf of X_i. Then as $n \to \infty$, $\frac{1}{n}\sum_{i=1}^n K_n(X_i - x) \to g(x)$. Hence

$$\widehat{F}_{nw}(y|x) \approx \frac{1}{ng(x)}\sum_{i=1}^n \epsilon_i K_h(X_i - x) + \frac{1}{ng(x)}\sum_{t=i}^n F(y|X_i)K_h(X_i - x),$$

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n F(y|X_i)K_h(X_i - x) &= \frac{1}{n}\sum_{i=1}^n F(y|x)K_h(X_i - x)\\ &+ \dot{F}(y|x)\sum_{i=1}^n \frac{1}{n}(X_i - x)K_h(X_i - x) + \cdots\end{aligned}$$

The extra bias term is due to the fact that

$$\sum_{i=1}^n \frac{1}{n}(X_i - x)K_h(X_i - x) \neq 0.$$

Idea: change the weights $\frac{1}{n}$ to force the sum equal to 0!

The empirical likelihood estimator $\widehat{F}_{el}(\cdot|x)$

(a) is a distribution function, and

(b) shares the same (the first order) asymptotic bias and variance as the local linear estimator $\widehat{F}_{ll}(\cdot|x)$.

5.3.2 Empirical likelihood estimator:

$$\widehat{F}_{el}(y|x) = \sum_{i=1}^n p_i(x)Z_iK_h(X_i - x)\Big/\sum_{j=1}^n p_j(x)K_h(X_j - x),$$

where $p_i(x)$ are the maximum empirical likelihood estimators defined as

$$\prod_{i=1}^n p_i(x) = \text{Max!}$$

subject to

$$p_i(x) \ge 0, \quad \sum_{i=1}^n p_i(x) = 1 \quad \text{and} \quad \sum_{i=1}^n p_i(x)U_i(x) = 0,$$

where $U_i(x) = (X_i - x)K_h(X_i - x)$.

By Theorem 1 in §2.1, $p_i(x) = \frac{1}{n - \lambda U_i(x)}$ and $\lambda \equiv \lambda(x)$ is the unique solution of

$$\sum_{i=1}^n \frac{U_i(x)}{n - \lambda U_i(x)} = 0.$$

6. Tests for Lyapunov exponents in deterministic systems

Reference:

Wolff, R.C., Yao, Q. and Tong H. (2004). Statistical tests for Lyapunov exponents of deterministic systems. *Studies in Nonlinear Dynamics and Econometrics*, **8**. Available at
http://stats.lse.ac.uk/q.yao/qyao.links/paper/wyt.pdf

6.1 Chaos

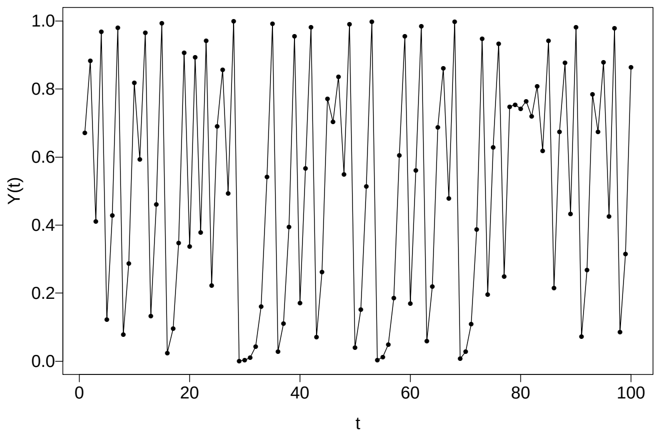
What is chaos?

- Nonlinear and deterministic system such as *Logistic map*:

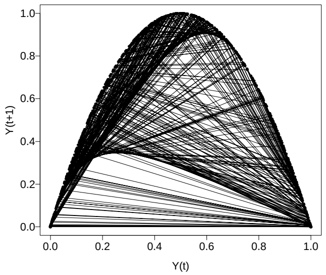
$$Y_{t+1} = 4Y_t(1 - Y_t)$$

- Random-like features

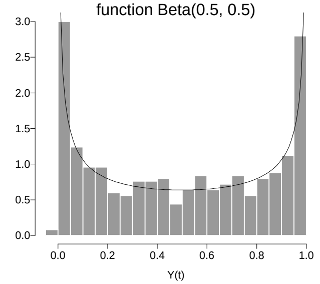
Time series generated by Logistic map



Scatter plot



Histogram, and the true density function Beta(0.5, 0.5)

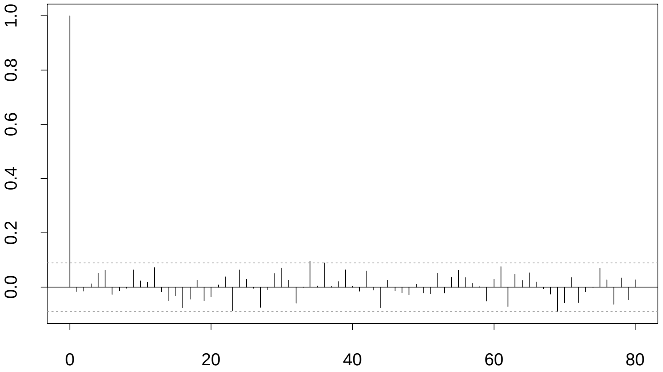


No linear dependence!

Autocorrelation function (ACF):

$$\rho^2(\tau) = \text{Corr}(Y_t, Y_{t+\tau}), \quad \tau = 1, 2, \cdots .$$

Estimated ACF of Logistic map

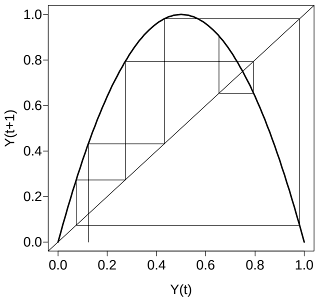


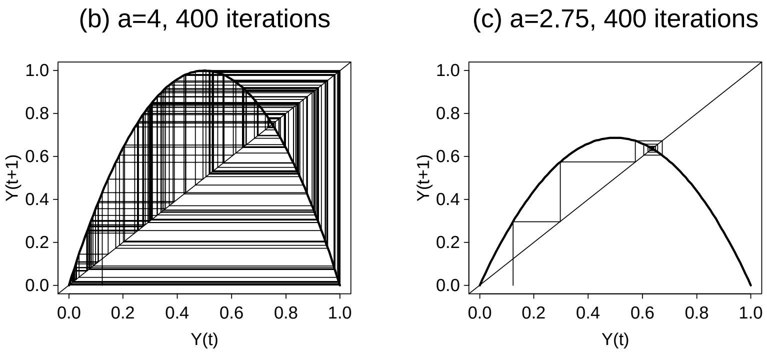
An important feature of chaos: Sensitivity to initial conditions

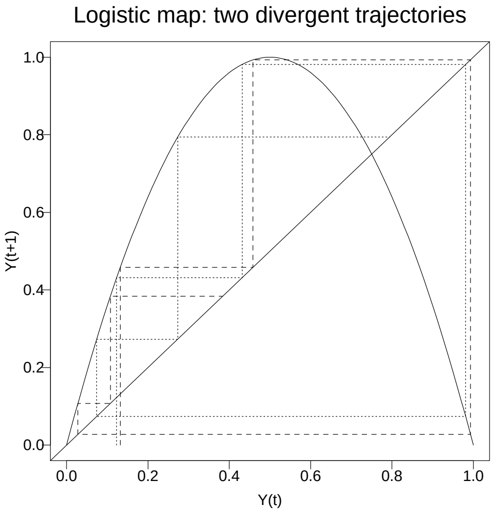
Two trajectories starting at nearby initial positions diverge fast!!!

Graphical Iteration: $Y_{t+1} = aY_t(1 - Y_t)$.

(a) a=4, 6 iterations







Five iterations of map $Y_{t+1} = 4Y_t(1 - Y_t)$ from initial values $Y_0 = 0.123$ (dotted lines) and $Y_0 = 0.132$ (dashed line):

$\Delta Y_0 = 0.009$,
$\Delta Y_1 = 0.027$,
$\Delta Y_2 = 0.012$,
$\Delta Y_3 = 0.046$,
$\Delta Y_4 = 0.163$,
$\Delta Y_5 = 0.410$

6.2 Lyapunov exponents

For model $Y_{t+1} = f(Y_t)$, $Y_1 = f(Y_0)$ and

$$Y_2 = f(Y_1) = f(f(Y_0)) = f^{(2)}(Y_0).$$

In general,

$$Y_m = f(Y_{m-1}) = f^{(m)}(Y_0).$$

Suppose two trajectories start at $Y_0 = x$ and $Y_0 = x + \delta$. After time m, they differ by distance

$$|f^{(m)}(x+\delta) - f^{(m)}(x)| \approx |\frac{d}{dx}f^{(m)}(x) \cdot \delta|.$$

By the chain rule,

$$\begin{aligned} &|\frac{d}{dx}f^{(m)}(x)| = |\dot{f}(x)\dot{f}[f(x)]\cdots\dot{f}[f^{(m-1)}(x)]| \\ &= \exp\{\sum_{k=0}^{m-1}\log|\dot{f}[f^{(k)}(x)]|\} \approx e^{m\lambda}, \end{aligned}$$

provided the limit

$$\lambda \equiv \lim_{m\to\infty}\frac{1}{m}\sum_{k=0}^{m-1}\log|\dot{f}[f^{(k)}(x)]| \quad \text{exists.}$$

λ **is called the Lyapunov exponent**, is a measure for the sensitivity to the initial conditions.

For chaotic systems, λ exists and is positive. Therefore,

$$|Y_m(x+\delta) - Y_m(x)| \approx \delta e^{m\lambda},$$

which diverges *exponentially* fast.

For Logistic model, $\lambda = \log 2$.

Remark. The existence of a positive Lyapunov exponent is often taken as a working definition of chaos.

Sensitivity to initial condition could be an issue wherever nonlinearity occurs.

Question: For any positive definite matrix A, find its square root $A^{\frac{1}{2}}$ for which $A^{\frac{1}{2}}A^{\frac{1}{2}} = A$.

Theorem. For any positive definite matrix B_0 for which $AB_0 = B_0A$, define

$$B_n = \frac{1}{2}(B_{n-1} + AB_{n-1}^{-1}).$$

Then $B_n \to A^{\frac{1}{2}}$ as $n \to \infty$.

The above algorithm seldom works in practice!

Example:

$$A = \begin{pmatrix} 9 & 0.5 \\ 0.5 & 4 \end{pmatrix}, \quad B_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

$$B_{n+1} = \frac{1}{2}(B_n + AB_n^{-1}), \quad \text{MSE} = ||A - B_n^2||^2/4.$$

n	MSE	n	MSE	n	MSE
1	66.78	2	2.33	3	0.91
4	1.77	5	3.10	6	4.90
7	7.28	8	10.35	9	12.35
10	19.14	11	25.25	12	32.81
13	42.14	14	53.59	15	67.60
16	84.66	17	105.41	18	130.55
19	160.96	20	197.66		

In numerical calculation, rounding errors keep changing ‘initial’ values in each iterations!

6.3 Ergodic measures

If the percentage of the points in the orbit

$$\{Y_0, f(Y_0), f^{(2)}(Y_0), \cdots, f^{(Y)}(X_0)\}$$

falling into an arbitrary set B stabilises as $n \to \infty$, we may define a probability measure

$$P(B) = \lim_{n\to\infty} \frac{1}{n+1} \sum_{t=0}^{n} I\{f^{(t)}(Y_0) \in B\}, \qquad a.s.\ P.$$

P is an ergodic measure.

It is easy to see that P is *invariant* in the sense that

$$P(A) = P\{f^{-1}(A)\}.$$

A dynamic system may admit many *mutually singular* ergodic measures, and those sit on sets with positive Lebesgue measures are of interest.

For logistic may $Y_{t+1} = 4Y_t(1 - Y_t)$, the distribution Beta$(0.5, 0.5)$ is an interesting ergodic measure with $(0, 1)$ as its support; the degenerate distribution at 0 is an uninteresting ergodic measure.

From now on, we assume P is an ergodic measure with the support of positive Lebesgue measures. Let $g(\cdot)$ be its density function.

Then the Lyapunov exponent can be expressed as

$$\begin{aligned} \lambda &= \int \log|\dot{f}(x)|P(dx) = \int \log|\dot{f}(x)|g(x)dx \\ &= E\{\log|\dot{f}(Y_t)|\} = E\{\log|\dot{f}\{f^{(t)}(Y_0)\}|\} \\ &= \lim_{n\to\infty} \sum_{t=0}^{n-1} \log|\dot{f}\{f^{(t)}(Y_0)\}| \qquad a.s.\ P. \end{aligned}$$

Furthermore, $\{Y_t\}$ can be treated as a *strictly stationary stochastic process* with the marginal density g.

6.4 Tests for Lyapunov exponents

Goal: detect if a Lyapunov exponent is positive; indicative for Chaos.

Let $X_t = \log|\dot{f}(Y_t)|$. Under an appropriate ergodic measure, the Lyapunov exponent is simply $\lambda = E(X_t)$.

A natural estimator: $\widehat{\lambda} = n^{-1}\sum_{i=1}^n X_i$

A test for the mean:

$$H_0: \ \lambda = 0 \qquad \text{vs} \qquad H_1: \lambda > 0.$$

The standard statistical tests (such as the t-test) do not apply to deterministic system.

Note. For one-dimensional deterministic process, the estimation for $f(\cdot)$ (therefore also for $\dot{f}(\cdot)$) is a trivial matter: plot Y_{t+1} against Y_t.

Bootstrap test

1. Construct a distribution $\widehat{g}(\cdot)$ satisfying the following two conditions

   (a) null hypothesis $H_0: \lambda = 0$, i.e. $\int x\widehat{g}(x)dx = 0$

   (b) the distribution is the closest to the observed 'sample' $\{X_1, \cdots, X_n\}$

2. Draw independent sample $X_1^*, \ldots, X_n^*$ from distribution $\widehat{g}(.)$, and compute $\lambda^* = n^{-1}\sum_i X_i^*$.

3. Repeat Step 2 B times, and the P-value of the test is $\frac{\#\{\widehat{\lambda} \le \lambda^*\}}{B}$.

Construction of $\widehat{g}(\cdot)$ — Empirical kernel density estimator:

$$\widehat{g}(x) = \frac{1}{h}\sum_{i=1}^n \widehat{p}_i K\left(\frac{X_i - x}{h}\right),$$

where $K(\cdot)$ is a density function, $h > 0$ is a bandwidth, and

$$(\widehat{p}_1, \ldots, \widehat{p}_n) = \arg\max \prod_{i=1}^n p_i$$

subject to $p_i \ge 0$, $\sum_{1 \le i \le n} p_i = 1$ and

$$\frac{1}{h}\sum_{i=1}^n p_i \int xK\left(\frac{X_i - x}{h}\right) dx = 0.$$

Remark. (i) For $K(x) = (2\pi)^{-1/2}e^{-x^2/2}$, $\int \frac{x}{h}K\left(\frac{X_i - x}{h}\right) dx = X_i$. Hence $\widehat{p}_i$ are the solution of "empirical likelihood for mean":

$$\max \prod_{i=1}^n p_i \ \text{ subject to } \ p_i \ge 0, \ \sum_{i=1}^n p_i = 1 \ \text{ and } \ \sum_{i=1}^n p_i X_i = 0.$$

Hence

$$\widehat{p}_i = \frac{1}{n - \omega X_i}, \qquad \sum_{j=1}^n \frac{X_j}{n - \omega X_j} = 0.$$

(ii) A random sample $\{X_1^*, \cdots, X_n^*\}$ from $\widehat{g}(x) = \frac{1}{h}\sum_{i=1}^n \widehat{p}_i K\left(\frac{X_i - x}{h}\right)$ may be drawn as follows:

$$X_i^* = Z_i + h\varepsilon_i, \qquad i = 1, \cdots, n,$$

where ε_i are independent $N(0,1)$, and Z_i are independently drawn from the discrete distribution

	X_1	X_2	$\ldots$	X_n
probability	$\widehat{p}_1$	$\widehat{p}_2$	$\ldots$	$\widehat{p}_n$

(iii) For deterministic data, we use very small h or even $h = 0$.