1 Empirical Likelihood

Empirical likelihood a nonparametric method without having to assume the form of the underlying distribution. It retains some of the advantages of likelihood based inference.
Example: (Somites of Earthworms) Earthworms have segmented bodies. The segments are known as somites. As a worm grows, both the number and the length of the somites increases. The dataset contains the number of somites on each of 487 worms gathered near Ann Arbor in 1902. The histogram shows that the distribution is skewed to the left, and has a heavier tail to the left.

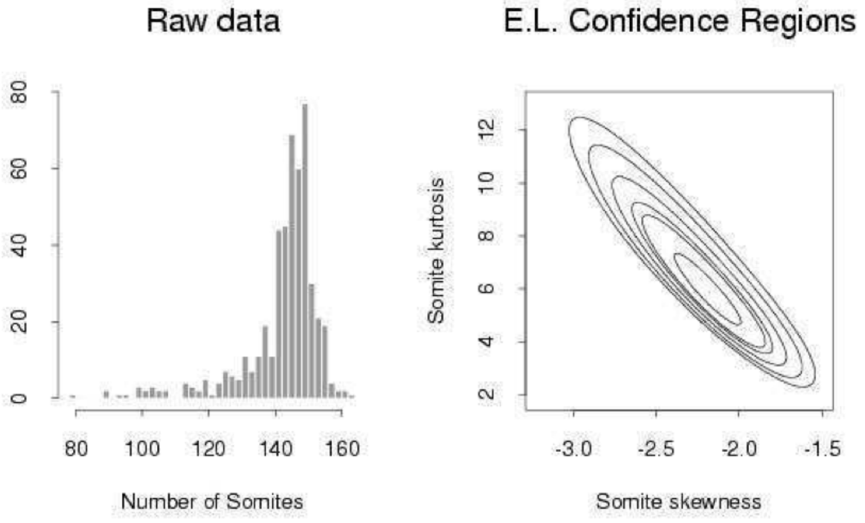


Figure 1: In the second panel, the empirical likelihood confidence regions (i.e. contours) correspond to confidence levels of 50%, 90%, 95%, 99%, 99.9% and 99.99%. Note: $(\gamma, \kappa) = (0, 0)$ is not contained in the confidence regions.

1.1 Why do conventional methods not apply?

Here are the existing methods:

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

2. **Bootstrap:** Difficult in picking out the confidence region from a point cloud consisting of a large number of bootstrap estimates for (γ, κ). For example, given 1000 bootstrap estimates for (γ, κ), ideally 95% confidence region should contain 950 central points. In practice, we restrict to rectangle or ellipse regions in order to facilitate the estimation.

Recall the measures of skewness (symmetry) and kurtosis (tail-heaviness):

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

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

Remark 1. *1. For $N(\mu, \sigma^2), \gamma = 0$ and $\kappa = 0$.*

2. For symmetric distributions, $\gamma = 0$.

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

1.2 Estimation of γ and κ

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

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

How to find confidence sets for (γ, κ)? In this section, we will define $l(\gamma, \kappa)$ as 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., $P(l(\gamma, \kappa) > C\} = 1 - \alpha$.

1.3 Introducing empirical likelihood

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

n known numbers.

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

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

where

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

What is the likelihood function of $\{p_i\}$ and 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 that

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

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

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

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

Example: Find the MELE for $\mu = EX_1$.

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

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

Remark 2. *1. MELEs, without further constraints, are simply the method of moment estimators, which is not new.*

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

2 Empirical likelihood inference for means

Let $X_1, \ldots, X_n$ be a random sample from an unknown distribution.
Goal: test hypothesis on $\mu = EX_1$, or find confidence intervals for μ.

2.1 Empirical likelihood ratio (ELR)

Consider the hypothesis

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

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

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

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

1. How do we find $\{\tilde{p}_i\}$?
2. What is the distribution of T under H_0?

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

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

subject to the conditions

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

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

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

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

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

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

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 $\big(n/(x_{(1)} - \mu), n/(x_{(n)} - \mu)\big)$.

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

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

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

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

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

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

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) implies (1). By (1) and (5),

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

It follows from (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

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

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

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

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

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

Note that for any λ in this interval,

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

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

Remark 3. a. When $\mu = \bar{x}$, $\lambda = 0$, and

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

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

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

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

b. We may view

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

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

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

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

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

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

the ELR statistic is

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

where λ is the unique solution of

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

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

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

Proof. (Sketch) Under H_0, $E(X_i) = \mu_0$. 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 $0 = f(\lambda_n) \approx f(0) + \dot{f}(0)\lambda_n$, implying

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

Now,

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

By the LLN, $n^{-1}\sum_i (X_i - \mu_0)^2 \to \mathrm{Var}(X_1)$. By the CLT, $\sqrt{n}(\bar{X} - \mu_0) \to \mathrm{N}(0, \mathrm{Var}(X_1))$ 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 $2 \log T < \xi^2_{1,1-\alpha}$, 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

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

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

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

The sample mean $\bar{X} = 2.61$ and the standard error $s = 4.71$.
Is the gain significant?
Intuitively: YES, if the negative observations -8.4 and -6.0 do not exist.
Let $\mu = EX_i$ and set up the hypotheses as

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

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

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

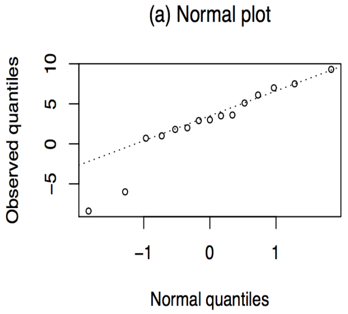
   Since $T = t(14)$ under H_0, the p-value is 0.06 - significant but not overwhelming. Is $\mathrm{N}(\mu, \sigma^2)$ an appropriate assumption? as the data do not appear to be normal (with a heavy left tail); see Figure 2.

2. Consider a generalized normal family

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

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

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



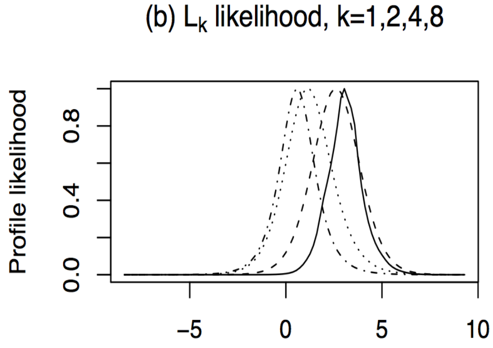
QQ-plot: Quantile of $N(0,1)$ vs Quantile of the empirical distribution.

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

Hence

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

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



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

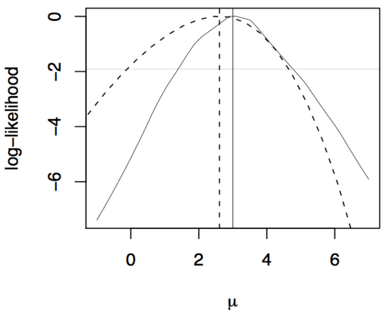
Figure 3: Profile likelihood

3. The empirical likelihood ratio test statistic $2 \log T = 3.56$, which rejects H_0 with the

p-value 0.04. The 95% credible interval is

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

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



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.)

Figure 4: Profile likelihood

3 Empirical likelihood for random vectors

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

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

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

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

Without any further constraints, the MELEs are

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

3.1 EL for multivariate means

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

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

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

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

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

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

in distribution.

Remark 4.

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

2. The null hypothesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ 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 $P\{\chi^2_d \leq \chi^2_{d,1-\alpha}\} = 1 - \alpha$.

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

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

4. Bootstrap calibration: Since (i) and (ii) are based on an asymptotic result, when n is small and d large, $\chi^2_{d,1-\alpha}$ may be replaced by the $\lceil B\alpha \rceil$-th value among $2\log T_1^*, \ldots, 2\log T_B^*$ which are computed as follows:

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

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

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

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

c. 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.,

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

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

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

ii. Transform the constrained problem to an unconstrained problem.

iii. Apply a Newton-Raphson algorithm.

Let

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

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

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

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

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

Hence

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

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

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

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

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

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

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

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

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

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

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

It is easy to see that $\log_*(z)$ has two continuous derivatives on $\mathbb{R}$. Set $M_*(\boldsymbol{\lambda}) = -\sum_{i=1}^{n} \log_*\{n - \boldsymbol{\lambda}^{\mathrm{T}}(\mathbf{X}_i - \boldsymbol{\mu})\}$. Then $M_*(\boldsymbol{\lambda}) = M(\boldsymbol{\lambda})$ on $\mathcal{H}_{\boldsymbol{\lambda}}$ and $M_*(\boldsymbol{\lambda})$ is a convex function on whole of $\mathbb{R}^d$. Hence $M_*(\boldsymbol{\lambda})$ and $M(\boldsymbol{\lambda})$ share the same minimizer which is the solution of (8).

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

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

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

Remark 5. *S-code "el.S", available from www-stat.stanford.edu/∼owen/empirical calculates the empirical likelihood ratio*

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

and other related quantities.

3.2 EL for smooth functions of means

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

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

where $\boldsymbol{\mu} = E\mathbf{X}_i$, 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 $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be $d \times 1$ i.i.d random variables with mean $\boldsymbol{\mu}_0$ and $|\text{Var}(\mathbf{X}_1)| \neq 0$. Let $\boldsymbol{\theta} = h(\boldsymbol{\mu})$ be a smooth function from $\mathbb{R}^d \to \mathbb{R}^q$ where $q \leq d$, and $\boldsymbol{\theta}_0 = h(\boldsymbol{\mu}_0)$. We assume that

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

For any $r > 0$, let

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

and

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

Then as $n \to \infty$,

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

Remark 6. 1. *The idea of bootstrap calibration may be appropriate here too.*

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

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

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

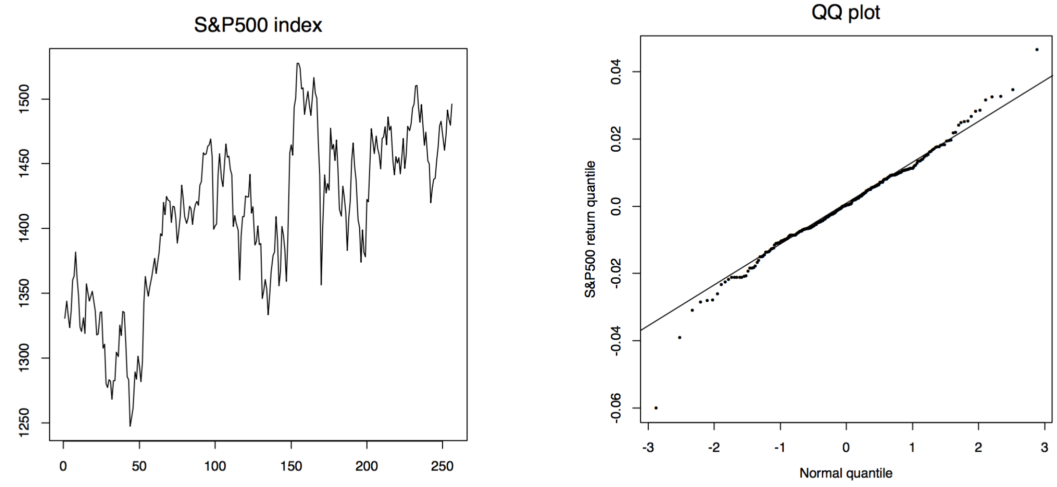
4. *In general, $P(\boldsymbol{\mu} \in \mathcal{C}_{1,r} \leq P(\boldsymbol{\theta} \in \mathcal{C}_{2,r})$.*

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

6. *The profile empirical likelihood function of $\boldsymbol{\theta}$ is*

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

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



(a) S&P Stocks

(b) QQ plot for S&P Stocks

Figure 5: S&P Stocks

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 ith day. By trating X_i i.i.d, we construct confidence intervals for the annual volatility

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

The simple point-estimator is

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

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

4 Estimating Equations

4.1 Estimation via estimating equations

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ 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 a $q \times 1$ vector, m is a $s \times 1$ vector-valued function. For example:

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

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

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

Obviously, in case F is in a parametric family and m is the score function, $\hat{\boldsymbol{\theta}}$ is the ordinary MLE.
Determined case $q = s$: $\hat{\theta}$ may be uniquely determined by (10).
Determined case $q > s$: The solutions of (10) may form a $(q - s)$-dimensional set.
Overdetermined case $q < s$: (10) may not have an exact solution, approximating solutions are sought. One such an example is so-called the generalised method of moments estimation which is very popular in Econometrics.
Example: Let $\{(X_i, Y_i), i = 1, \ldots, n\}$ 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 $\sigma_y^2 = \text{Var}(Y_1)$. Putting $\boldsymbol{\theta}^{\mathrm{T}} = (\mu_x, \mu_y, \sigma_y^2, \gamma)$, and

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

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

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

Remark 7. *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 $\boldsymbol{\theta}$:

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

The following Theorem follows from Theorem 2 immidiately.

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

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

Then as $n \to \infty$,

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

in distribution.

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

Remark 8.

1. *In general* $L(\boldsymbol{\theta})$ *can be calculated using the method for EL for multivariate means, treating* $m(\mathbf{X}_i, \boldsymbol{\theta})$ *as a random vector.*
2. *For* $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ *which is the solution of*

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

   $L(\hat{\boldsymbol{\theta}}) = (1/n)^n$.
3. *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*

   $$\log\{L(\boldsymbol{\theta}_0)\} + n\log n \leq -0.5\chi^2_{s,1-\alpha}.$$
4. *Any* $(1-\alpha)$ *confidence set for* $\boldsymbol{\theta}$ *determined by* $E\{m(\mathbf{X}_1, \boldsymbol{\theta})\} = 0$ *is*

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

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

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

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

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

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

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

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

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

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

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

Thus,

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

Hence

$$\begin{aligned} \Theta_\alpha = \{\theta_\alpha : \log\{L(\theta_\alpha)\} \geq -n \log n - 0.5\chi^2_{1,1-\alpha}\} \\ = \left\{ \theta_\alpha : r \log \frac{n\alpha}{r} + (n-r) \log \frac{n(1-\alpha)}{n-r} > -0.5\chi^2_{1,1-\alpha} \right\} \end{aligned}$$

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

5 Empirical likelihood for estimating conditional distribution

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.
- 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

$$Y = X_1\beta_1 + \cdots + X_d\beta_d + \epsilon, \tag{11}$$

where ϵ has mean 0 and variance σ^2. This model is linear with respect to unknown coefficients $\beta_1, \ldots, \beta_d$ as the variable $X_1, \ldots, X_d$ may be

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

Put $\beta = (\beta_1, \ldots, \beta_d)^{\mathrm{T}}$. With observation $\{Y_i, \mathbf{X}_i, 1 \le i \le n\}$, where $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})^{\mathrm{T}}$, the LSE minimises

$$\hat{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y},$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, and $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})^{\mathrm{T}}$, the LSE minimises

$$\sum_{i=1}^{n}(Y_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})^2, \tag{12}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, and $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^{\mathrm{T}}$ is an $n \times d$ matrix. The fitted model is $\hat{Y} = \mathbf{X}\hat{\boldsymbol{\beta}}$. This is a global fitting, since the model is assumed to be true everywhere in the sample space and the estimator $\hat{\boldsymbol{\beta}}$ is obtained all the available data. Such a global fitting is efficient if the assumed form of the regression function (11) is correct. In general (11) 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 (12) as follows. Suppose we fit a local linear model in a small neighborhood of the observation $\mathbf{X}_k$, with the coefficient $\boldsymbol{\beta} = \boldsymbol{\beta}_k$, the LSE minimizes

$$\sum_{i=1}^{n} (Y_i - \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_k)^2 w(\mathbf{X}_i, \mathbf{X}_k) \tag{13}$$

where the weight function may be taken as

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

where $p \geq 1$ is a prescribed small integer. Although the sum in (13) only has p non-zero terms, the LSE can be expressed formally as

$$\hat{\boldsymbol{\beta}}_k = (\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{Y},$$

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

Remark 9. *1. The local estimator $\hat{\boldsymbol{\beta}}_k$ only makes use of the p (out of n) observations around $\mathbf{X}_k$, may depend on the choice of p sensitively.*

2. Intuitively, the local estimator $\hat{\boldsymbol{\beta}}_k$ may catch some local structure better that the global estimator $\hat{\beta}$. But the variance of $\hat{\boldsymbol{\beta}}_k$ is larger than that of $\hat{\boldsymbol{\beta}}$.

5.2 Kernel methods

5.2.1 Introduction

We observe $\{(Y_i, X_i), i = 1, \ldots, n\}$ from

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

where $f(\cdot)$ is an unknown and smooth function. We may use the idea of local smoothing to estimate f. Let $\hat{f}(x)$ is the average of all those Y_i for which X_i is among the k nearest neighbors of x. Hence

$$\frac{1}{k}\sum_{i=1}^{k} Y_i w(x, X_i) = \frac{\sum_{i=1}^{n} y_i w(x, X_i)}{\sum_{i=1}^{n} w(x, X_i)}$$

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

5.2.2 Nadaraya-Watson estimator

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

Instead of specifying k-the number of neighbors used in estimation, we may determine the number by choosing

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

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 $k(x) = 0.5I(|x| \leq 1)$, only those X_i within h distance from x are used in estimating $f(x)$. The number of points may vary with respect to x. The resulting estimator

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

is called the Nadaraya-Watson estimator. In fact $\hat{f}(\cdot)$ is a local LSE, since

$$\hat{f}(x) = \text{argmin}_a \sum_{i=1}^{n} \{Y_i - a\}^2 K\left(\frac{X_i - x}{h}\right).$$

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

Remark 10. 1. *Commonly used kernel functions:*

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

2. *The bandwidth h controls the amount of data used in local estimation, determines the smoothness of the estimated curve $\hat{f}(\cdot)$ For example, with $K(x) = 0.5I(|x| \leq 1)$, $\hat{f}(x) \to \bar{Y}$ as $h \to \infty$ global constant fitting; $\hat{f}(X_i) \to Y_i$ as $h \to 0$ interpolating the observations. h is also called a smooth parameter.*

3. *The goodness of the estimator $\hat{f}(\cdot)$ 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.*

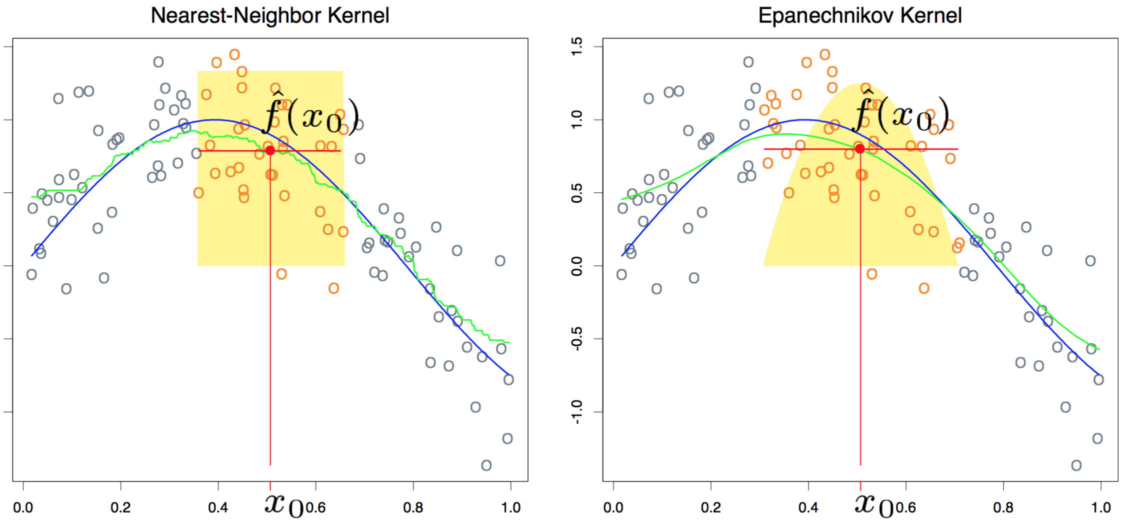


Figure 6: In each panel, 100 pairs x_i, y_i are generated at random from the blue curve with Gaussian errors: $Y = \sin(4X) + \epsilon, X \sim \text{Unif}(0,1), \epsilon \sim N(0, 1/3)$. In the left panel, the green curve is the result of 30-nearest-neighbor running-mean smoother. The red point is the 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 the observations. In the right panel, the green curve is the kernel-weighted average, using an Epanechnikov kernel with (half) window width $\lambda = 0.2$

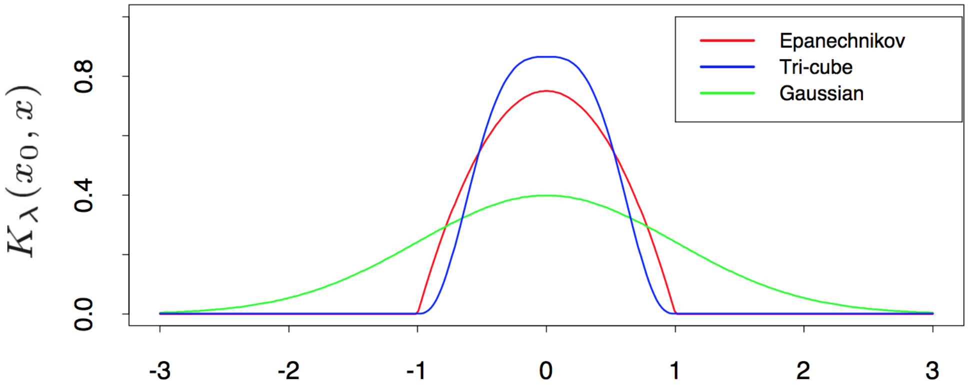


Figure 7: A comparison of three popular kernels for local smoothing. Each has been calibrated to integrate to 1. the tri-cube kernel is compact and has two continuous derivatives at the boundary of its support, while the Epanechnikoc kernel has none. The Gaussian kernel is continuously differentiable, but has infinite support.

5.3 Bias and variance calculations

Regularity conditions:

1. $\{Y_i, X_i\}$ are i.i.d and
$$f(x) = E(Y_i \mid X_i = x), \quad \epsilon_i = Y_i - f(X_i).$$
Further both $f(\cdot)$ and $p(\cdot)$ have two continuous derivatives, where $p(\cdot)$ denotes the pdf of X_i.

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

Let $\sigma_0^2 = \int u^2 K(u)du$, $\mathbf{X} = (X_1, \ldots, X_n)^{\mathrm{T}}$ and
$$\sigma^2(x) = \mathrm{Var}(Y_i \mid X_i = x) = E(Y_i^2 \mid X_i = x) - f^2(x).$$

Theorem 6. *Under conditions (i) and (ii) above, it holds that for x with $p(x) > 0$,*
$$\begin{aligned} E\{\hat{f}(x) - f(x) \mid \mathbf{X}\} &\asymp \frac{h^2\sigma_0^2}{2}\left\{\ddot{f}(x) + \frac{2\dot{f}(x)\dot{p}(x)}{p(x)}\right\}, \\ Var\{\hat{f}(x) \mid \mathbf{X}\} &\asymp \frac{1}{nh}\frac{\sigma(x)^2}{p(x)}\int K^2(u)du. \end{aligned}$$

Proof. We only provide a sketch of the proof for the bias. Let
$$K_i = h^{-1}K\left(\frac{X_i - x}{h}\right).$$

Then
$$\hat{f}(x) = \sum_i Y_i K_i / \sum_i K_i.$$

Note that (i) implies $E\{\epsilon_i \mid \mathbf{X}\} = E\{\epsilon_i \mid X_i\} = 0$.
$$\begin{aligned} E\{\hat{f}(x) - f(x) \mid \mathbf{X}\} &= \sum_i E\{Y_i - f(x) \mid \mathbf{X}\}K_i / \sum_i K_i \\ &= \sum_i \{f(X_i) - f(x)\}K_i / \sum_i K_i. \end{aligned}$$

It follows from the Law of Large numbers that
$$\frac{1}{n}\sum_{i=1}^n K_i \asymp E(K_1) = \int \frac{1}{h}K\left(\frac{X - x}{h}\right)p(X)dX = \int K(u)p(x + hu)du \to p(x), \qquad (14)$$

and

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\{f(X_i)-f(x)\}K_i &\asymp \int\{f(X)-f(x)\}\frac{1}{h}K\biggl(\frac{X-x}{h}\biggr)p(X)dX \\
&= \int\{f(x+hu)-f(x)\}K(u)p(x+hu)du \qquad (15)\\
&= \int\{hu\dot{f}(x)+\frac{h^2u^2}{2}\ddot{f}(x)\}\{p(x)+hu\dot{p}(x)\}K(u)du+O(h^3) \\
&= h^2\sigma_0^2\{\dot{f}(x)\dot{p}(x)+0.5\ddot{f}(x)p(x)\}+O(h^3). \qquad (16)
\end{aligned}$$

Combining (14) and (16), we obtain the required asymptotic formula for the bias. □

Remark 11. *1. An approximate MSE:*

$$\begin{aligned}
E[\{\hat{f}(x)-f(x)\}^2 \mid \mathbf{X}] &= \mathit{Bias}^2+\mathit{Variance} \\
&\approx \frac{h^4\sigma_0^4}{4}\biggl\{\ddot{f}(x)+\frac{2\dot{f}(x)\dot{p}(x)}{p(x)}\biggr\}^2+\frac{1}{nh}\frac{\sigma(x)^2}{p(x)}\int K^2(u)du.
\end{aligned}$$

Increasing h, variance decreases and bias increases. A good choice of h is a trade-off between the variance and the bias. Minimizing 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.

2. It can be shown that

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

converges in distribution to

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

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.

5.3.1 Kernel density estimation

From (15), a natural estimator for the density function of X_i is

$$\hat{p}(x)=\frac{1}{nh}\sum_{i=1}^{n}K\biggl(\frac{X_i-x}{h}\biggr)$$

which is called a kernel density estimator. (15) *also implies that* $\hat{p}(x)$ *is a consistent estimator. Further,*

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

5.3.2 Local linear regression estimation

The Nadaraya-Watson estimation is a local constant estimation, i.e., for y is a small neighborhood of x, we approximate $f(y) \approx f(x)$, and minimize

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

Intuitively, the estimation can be improved by using a local-linear approximation:

$$f(y) \approx f(x) + \hat{f}(x)(y - x).$$

This leads to the local-linear regression estimator: $\hat{f}(x) \equiv \hat{a}$, where $(\hat{a}, \hat{b})$ minimizes

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

Obviously, a natural estimator for $\dot{f}$ is $\hat{\dot{f}}(x) \equiv \hat{b}$. Let $\mathcal{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, $\boldsymbol{\theta} = (a, b)^{\mathrm{T}}$, $\mathcal{X}$ be a $n \times 2$ matrix with $(1, (X_i - x))$ as its ith row, and $\mathcal{K}$ is a $n \times n$ diagonal matrix with $K\{(X_i - x)/h\}$ as its (i, i)-th element. Then (17) can be written as

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

Therefore the LSE method leads to

$$\begin{pmatrix} \hat{f}(x) \\ \hat{\dot{f}}(x) \end{pmatrix} = \hat{\boldsymbol{\theta}} = (\mathcal{X}^{\mathrm{T}} \mathcal{K} \mathcal{X})^{-1} \mathcal{X}^{\mathrm{T}} \mathcal{K} \mathcal{Y}$$

Hence like the Nadaraya-Watson estimator, the local linear estimator for $f(x)$ is a linear combination of $Y_1, \ldots, Y_n$ (given $\mathbf{X} = (X_1, \ldots, X_n)^{\mathrm{T}}$). 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 'ls.s'. Splus and R function 'loess' offers more flexibilities for local regression fitting.

5.3.3 Why is a local linear estimator better

1. Simpler (and often smaller) bias formula
2. 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).

Table 1: default

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

5.4 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 $F(\cdot \mid x)$ nonparametrically.
Motivation: quantile regression, prediction and etc.

5.4.1 Nadaraya-Watson and local linear estimators

Note: $E\{I(Y_i \le y) \mid X_i = x\} = F(y \mid x)$. Hence $G(y \mid x)$ is a regression of $Z_i \equiv I(Y_i \le y)$ on X_i as $E(Z_i \mid X_i) = F(y \mid X_i)$.
Nadaraya-Watson estimator:

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

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

$$w_i(x) = K\left(\frac{X_i - x}{h}\right) / \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$$

In the above expression, $K(\cdot)$ is a pdf and $h > 0$ is a bandwidth. $\hat{F}_{nw}(y \mid x)$ itself is a proper distribution function ! In fact, $\hat{F}_{nw}(y \mid x)$ 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 glocal estimator $\bar{Z}$.
Local-linear estimator: $\hat{F}_{ll}(y \mid x) \equiv \hat{a}$, where $(\hat{a}, \hat{b})$ 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: $\hat{Z}_i = \hat{a} + \hat{b}(X_i - x)$. $\hat{F}_{ll}(y \mid x)$ has superior bias properties (and other types of efficiency). But $\hat{F}_{ll}(y \mid x)$ is not necessarily a distribution function, as it may take value outside the interval $[0, 1]$, and is not necesarily monotonically increases in y.
An ideal estimator: Combine the advantages of both $\hat{F}_{nw}(y \mid x)$ and $\hat{F}_{ll}(y \mid x)$ together. Write $Z_i = I(Y_i \leq y) = F(y \mid X_i) + \epsilon_i$ and $K_h(x) = h^{-1}K(x/h)$. Let $g(\cdot)$ be the pdf of X_i. Then as $n \to \infty$, $(1/n)\sum_{i=1}^{n} K_h(X_i - x) \to g(x)$. Hence

$$\begin{aligned}
\hat{F}_{nw}(y \mid x) &\approx \frac{1}{ng(x)}\sum_{i=1}^{n} \epsilon_i K_h(X_i - x) + \frac{1}{g(x)}\sum_{i=1}^{n} F(y \mid X_i)K_h(X_i - x), \text{and} \\
&\approx \frac{1}{n}\sum_{i=1}^{n} F(y \mid X_i)K_h(X_i - x) \\
&\approx \frac{1}{n}\sum_{i=1}^{n} F(y \mid x)K_h(X_i - x) + \dot{F}(y \mid x)\frac{1}{n}\sum_{i=1}^{n}(X_i - x)K_h(X_i - x) + \cdots
\end{aligned}$$

The extra bias term is due to the fact that $\dot{F}(y \mid x)\frac{1}{n}\sum_{i=1}^{n}(X_i - x)K_h(X_i - x) \neq 0$. Idea: Change the weights $(1/n)$ to force the sum equal to 0.

5.4.2 Empirical Likelihood estimator

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

where $p_i(x), i = 1, \ldots, n$ are the maximum empirical likelihood estimators defined as maximize

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

subject to

$$p_i(x) \geq 0, \sum_{i=1}^{n} p_i(x) = 1, \sum_{i=1}^{n} p_i(x) U_i(x) = 0,$$

where $U_i(x) = (X_i - x) K_h(X_i - x)$. By Lagrangian techniques

$$p_i(x) = \frac{1}{n - \lambda U_i(x)}, \quad \lambda \equiv \lambda(x)$$

where $\lambda(x)$ is the unique solution of

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

The empirical likelihood estimator $\hat{F}_{el}(| \, x)$

a. is a distribution function, and

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