LARGE SAMPLE THEORY

GIORGIO PICCI

University of Padova, Italy

University of Guangzhou, November 2018

SCOPE OF THESE NOTES

To discuss the main ideas and complement the textbook

ELEMENTS OF LARGE SAMPLE THEORY
by
E.L. Lehmann

You should all have an electronic version of this book. Please **take it to class**. It does not make sense for me to provide notes repeating this material. Another useful reference is

A COURSE IN LARGE SAMPLE THEORY
by
T.S. Ferguson

Outline of the Course

1. Some Probability Background
2. Review of convergence concepts
3. The CLT for i.i.d random sequences
4. Applications
5. Law of Large Numbers and Ergodicity
6. Maximum Likelihood: Asymptotic properties
7. Time series

QUICK REVIEW OF PROBABILITY

- **Probability space**: $\{\Omega, \mathcal{A}, \mathbf{P}\}$ the *elementary event* $\omega \in \Omega$ chosen by "nature". $\mathcal{A}$ contains all subsets of Ω (Events) of which you can compute the probability.

$$\mathbf{P} : \mathcal{A} \to [0, 1], \qquad \text{countably additive set function}$$

- **Random variables** are functions $\mathbf{x} : \Omega \to \mathbb{R}$.
  The *Probability distribution function* of $\mathbf{x}$:

$$F(x) := \mathbf{P}\{\omega \mid \mathbf{x}(\omega) \leq x\}; \qquad x \in \mathbb{R}$$

  right-continuous non-decreasing monotonic function.

- **Expectation** of a random variable

$$\mathbb{E}\mathbf{x} := \int_{\Omega} \mathbf{x}(\omega)\, \mathbf{P}(d\omega) = \int_{\mathbb{R}} x\, dF(x)$$

AN ELEMENTARY EXAMPLE

Assume we are tossing a coin and let p := probability to observe “TAIL”, event which will be denoted by the symbol T and $1-p$:= probability that HEAD will show instead; event which is denoted by the symbol C. Naturally, p is unknown. We want to obtain information on the value of p by tossing the coin N consecutive times, assuming that each toss *does not influence the outcome of the other tosses*.

Let $\Omega = \{$all possible outcomes of N consecutive tosses$\}$.

The set Ω contains all sequences made of N symbols T and C in any possible order say one possible ω being

T T C C C T T C C T T T T$\ldots$C T   N symbols

Let $\mathscr{A}$ be the family of all subsets A of Ω. These are called *Events*. Example

$A = \{\omega \mid$ in ω there are an even number of T’s$\}$

Assume each toss does not influence the outcome of the other tosses. This defines a class of probability measures which describes each toss as being *independent* of the others. In formulas, this means that a class of probability measures

$$\mathscr{P} := \{\mathbf{P}_p\,; 0 < p < 1\}$$

is defined on on $\{\Omega, \mathscr{A}\}$ for each elementary event $\omega \in \Omega$ by the *Bernoulli distribution*

$$\mathbf{P}_p(\{\omega\}) = p^{T(\omega)}\,(1-p)^{N-T(\omega)} \qquad 0 < p < 1 \quad ,$$

where $T(\omega)$ is the number of symbols T in the sequence ω. Clearly the probability measure $\mathbf{P}_p$ is defined as soon as one assigns a value to p in the interval$(0 < p < 1)$.

In this case the family $\mathscr{P}$ is *parametric*; i.e.

$$\mathscr{P} := \{\mathbf{P}_p\,; 0 < p < 1\}\ .$$

Estimating $\mathbf{P}$ is just selecting a plausible value of p based on the observation of the outcomes of N successive coin tosses.

Alternatively, one may want to validate some a priori belief on p for example that $p = 1/2$ (that is, T and C are equiprobable). In this case one deals with an *hypothesis testing* problem: on the basis of some observation $\bar{\omega}$ decide whether $\mathbf{P}_p$ belongs to the class

$$\mathscr{P}_0 := \{\mathbf{P}_{1/2}\} \quad ,$$

or P_p belongs to the complementary family

$$\mathscr{P}_1 := \left\{\mathbf{P}_p \, ; \, p \neq 1/2\right\} .$$

A Bernoulli random variable $\mathbf{x}_k \, ; \, k = 1, 2, \ldots, N$ takes value 1 when at the k-toss T is observed and 0 otherwise. So $\mathbf{x}_k(\omega)$ only depends on the k-th symbol ω_k of the sequence.

A Binomial random variable $\mathbf{s}_N(\omega)$ is the gambler's fortune after N tosses

$$\mathbf{s}_N(\omega) := \sum_{k=0}^{N} \mathbf{x}_k(\omega)$$

The expected value of $\mathbf{x}_k$ is

$$\mathbb{E}\,\mathbf{x}_k = \sum_{\omega_k} \mathbf{x}(\omega_k)\,\mathbf{P}_p(\{\omega_k\})$$

Since ω_k can be T with probability p and C with probability $1-p$, it follows that $\mathbb{E}\,\mathbf{x}_k = p$ for all k.

The probability distribution of a Binomial random variable, assuming the tosses are independent, is

$$\mathbf{P}_p\{\omega\,;\,\mathbf{s}_n(\omega)=n\} = \sum_{k=0}^{n} \left(\begin{array}{c} N \\ k \end{array}\right) p^k (1-p)^{N-k}$$

This is known as the **Binomial distribution** and is denoted $B_{p,N}(n)$.

NOTATIONS

∘ Random variables will be denote by **lower case bold symbols** such as $\mathbf{x}$, $\mathbf{y}, \ldots$ etc. Lehmann notation of using Upper case symbols X, Y .. is bad. Upper case symbols are standarad for MATRICES such as covarinces or loading matrices in linear models. Later we shall need to introduce *multivariate statistics* and Lehmann notation would produce confusion.

∘ The sample size is denoted by N: lower case n is often used for dimension of vectors (either random or non-random) or degrees of freedom. So in general n is fixed while $N \to \infty$.

∘ Names: Pdf instead of cdf; $\mathbf{x} \sim F$ means that the random variable $\mathbf{x}$ has Pdf F. In discrete probability spaces $F(x)$ is a staircase function.

Continuous variables admit a **probability density function** $p(x) := \dfrac{dF(x)}{dx}$ (pdf).

CONVERGENCE OF RANDOM VARIABLES

1. **Almost sure convergence**: is ordinary convergence of functions $\mathbf{x}_N(\omega) \to \mathbf{x}(\omega)$ for all $\omega \in \Omega$, except perhaps a subset of ω's of probability zero. Written $\mathbf{x}_N \overset{a.s.}{\to} \mathbf{x}$

2. **quadratic mean convergence**: $\mathbb{E}\,|\mathbf{x}_N - \mathbf{x}|^2 \to 0$
   This is written $\mathbf{x}_N \overset{q.m.}{\to} \mathbf{x}$

3. **Convergence in probability**:
   $\mathbf{P}\{\omega \mid |\mathbf{x}_N(\omega) - \mathbf{x}(\omega)| > \varepsilon\} \to 0$ for all $\varepsilon > 0$.

   This is written $\mathbf{x}_N \overset{\mathbf{P}}{\to} \mathbf{x}$ or $\mathbf{P} - \lim \mathbf{x}_N = \mathbf{x}$.

IMPLICATIONS:

$\mathbf{1.} \Rightarrow \mathbf{3.}$ $\qquad$ $\mathbf{2.} \Rightarrow \mathbf{3.}$

CHEBYSHEV INEQUALITY

Suppose $\mathbf{x}$ and $\mathbf{y}$ have finite second moment, then for all $\varepsilon > 0$ and constant c

$$\mathbf{P}\{\,|\mathbf{x}-\mathbf{y}\,| \geq \varepsilon\} \leq \frac{1}{\varepsilon^2}\,\mathbb{E}\left[(\mathbf{x}-\mathbf{y})^2\right]$$

Same proof of Lemma 2.1.1 in Lehmann's book for $\mathbf{y} = c$. In fact just call $\mathbf{z} := \mathbf{x}-\mathbf{y}$.

Theorem 1 (Theorem 2.1.1)**.** *Let* $\mathbf{x}_N \overset{q.m.}{\to} \mathbf{x}$ *then* $\mathbf{x}_N \overset{\mathbf{P}}{\to} \mathbf{x}$

Proof: just let $\mathbf{x} \equiv \mathbf{x}_N$ and $\mathbf{y} \equiv \mathbf{x}$.

Examples on p. 49 in Lehmann's book.

THE BINOMIAL DISTRIBUTION

A *Binomial* random variable $\mathbf{s}_N$ is the total number n of Tails (T) in N independent Bernoulli tosses with probability p. Its distribution is

$$B_{p,N}(n) = \sum_{k=0}^{n} \left(\begin{array}{c} N \\ k \end{array} \right) p^k (1-p)^{N-k}$$

Since $\mathbf{s}_N$ is the sum of N independent Bernoulli random variables $\mathbf{x}_k\,; k = 1, 2, \ldots, N$ each taking value 1 when T is observed and 0 otherwise, whose mean is $\mathbb{E}\,\mathbf{x} = p$, we have

$$\mathbb{E}\,\mathbf{s}_N = \mathbb{E}\,\{\mathbf{x}_1\} + \mathbb{E}\,\{\mathbf{x}_2\} + \ldots + \mathbb{E}\,\{\mathbf{x}_N\} = \underbrace{p + \cdots + p}_{N \text{ times}} = Np$$

The variance is computed by summing the variances of each $\mathbf{x}_k$ which can be computed to be $p(1-p)$. Hence the variance of $\mathbf{s}_N$ is $Np(1-p)$ and so $\operatorname{var}\left(\frac{\mathbf{s}_N}{N}\right) = \frac{1}{N^2} Np(1-p) = \frac{1}{N} p(1-p)$. Then

$$\frac{\mathbf{s}_N}{N} \overset{P}{\to} p\,. \qquad \text{(Lehmann Example (2.1.1))}$$

WEAK LAW OF LARGE NUMBERS

Let $\{\mathbf{x}_1,\ \mathbf{x}_2,\ \ldots\ \mathbf{x}_N\}$ be a sequence of random variables (a *sample* of size N)

Definition 1. *The* **sample mean** *of the sequence is the random variable*

$$\hat{\boldsymbol{\mu}}_N := \frac{1}{N}\sum_{k=1}^{N}\mathbf{x}_k$$

(also denoted $\bar{\mathbf{x}}$ or $\bar{\mathbf{x}}_N$). The **sample variance** *of the sequence is the random variable*

$$\hat{\boldsymbol{\sigma}}_N^2 := \frac{1}{N}\sum_{k=1}^{N}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_N)^2$$

Theorem 2. *If the random variables are* **independent identically distributed (i.i.d.)** *then*

$$\hat{\boldsymbol{\mu}}_N \xrightarrow{P} \mu = \mathbb{E}\,\mathbf{x}_k$$

that is the sample mean is a consistent estimator *of the mean.*

PROOF

See p. 49. By Chebyshev inequality

$$\mathbf{P}\{\,|\hat{\boldsymbol{\mu}}_N - \mu\,| \geq \varepsilon\} \leq \frac{1}{\varepsilon^2}\,\mathbb{E}\left[(\hat{\boldsymbol{\mu}}_N - \mu)^2\right] \qquad (\hat{\boldsymbol{\mu}}_N \equiv \bar{\mathbf{x}} \text{ if you like})$$

The quantity $\mathbb{E}\left[(\hat{\boldsymbol{\mu}}_N - \mu)^2\right]$ is the variance of the sample mean:

$$\operatorname{var}\left\{\frac{1}{N}\sum_{k=1}^{N}\mathbf{x}_k\right\} = \frac{1}{N^2}\operatorname{var}\left\{\sum_{k=1}^{N}\mathbf{x}_k\right\} = \frac{1}{N^2}\,N\sigma^2$$

where $\sigma^2 = \operatorname{var}(\mathbf{x}_k)$. Then

$$\mathbf{P}\{\,|\hat{\boldsymbol{\mu}}_N - \mu\,| \geq \varepsilon\} \leq \frac{1}{N\varepsilon^2}\,\sigma^2 \to 0$$

for all $\varepsilon > 0$ as $N \to \infty$. Hence $\hat{\boldsymbol{\mu}}_N \xrightarrow{P} \mu = \mathbb{E}\,\mathbf{x}_k$.

CONSISTENCY OF ESTIMATORS

An **estimator** of a parameter θ (or of a function $g(\theta)$ of the parameter) is a function $\phi(\mathbf{y})$ of the sample $\mathbf{y} := \{\mathbf{y}_1, \mathbf{y}_2 \ldots, \mathbf{y}_N\}$ which does not depend on θ. The estimator is *(uniformly) unbiased* if

$$\mathbb{E}_{\theta}\phi(\mathbf{y}) = \theta\,; \qquad (\text{or } \; g(\theta)\,) \quad \forall\theta$$

Definition 2. *Assume the data are generated by a "true model" corresponding to a "true parameter value" θ_0. Let $\mathbf{P}_0 \equiv \mathbf{P}_{\theta_0}$ be the corresponding "true" probability law of the data. The estimator sequence $\phi_N\,; N = 1, 2, \ldots$ is* ***consistent in probability*** *(or weakly consistent) if*

$$\mathbf{P}_0 - \lim \phi_N = \theta_0$$

That is, the sequence of random variables $\{\phi_N\}$ converges in probability $\mathbf{P}_0$ to a nonrandom constant equal to the true parameter θ_0.

CONSISTENCY AND CHEBYSHEV INEQUALITY

Chebyshev inequality holds also for random vectors (see below):

$$\mathbf{P}_\theta\left(\|\phi_N - \theta\| \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2}\,\mathbb{E}_\theta\left[(\phi_N - \theta)^\top(\phi_N - \theta)\right] = \frac{1}{\varepsilon^2}\,\mathbb{E}_\theta\|\phi_N - \theta\|^2,$$

where $\phi_N := \phi_N(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ and $\|\cdot\|$ is Euclidean norm or absolute value when ϕ_N is scalar. Proof: easy generalization of the proof of Lemma 2.1.1 in Lehmann's book.

If ϕ_N is unbiased $\mathbb{E}_\theta\,\phi_N = \theta$ and the last member is the *scalar variance*, $\sigma_N^2(\theta)$, of $\phi_N(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ divided by ε^2. If

$$\lim_{N\to\infty} \sigma_N^2(\theta) = 0; \forall\theta \in \Theta \quad ,$$

then $\phi_N(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ is (weakly) consistent. (Remember that we do not know the true value θ_0).

Proposition 1. *If $\phi_N(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ is an asymptotically unbiased estimator and if its scalar variance $\sigma_N^2(\theta)$ tends to zero as $N \to \infty$ for all $\theta \in \Theta$, then $\phi_N(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ is consistent.*

APPLICATIONS p. 55

Suppose $\{\mathbf{x}_1,\, \mathbf{x}_2,\, \dots \mathbf{x}_N\}$ is a sequence of **independent** random variables with same mean μ but different variances $\sigma_k^2 = \mathrm{var}\,(\mathbf{x}_k)\,;\, k = 1, 2, \dots N$. Then $\mathbb{E}\left[(\hat{\boldsymbol{\mu}}_N - \mu)^2\right]$ is still the variance of the sample mean:

$$\mathrm{var}\left\{ \frac{1}{N} \sum_{k=1}^{N} \mathbf{x}_k \right\} = \frac{1}{N^2} \mathrm{var}\left\{ \sum_{k=1}^{N} \mathbf{x}_k \right\} = \frac{1}{N^2} \sum_{k=1}^{N} \sigma_k^2$$

If

$$\frac{1}{N} \sum_{k=1}^{N} \frac{\sigma_k^2}{N} \to 0$$

the sample mean $\hat{\boldsymbol{\mu}}_N$ is still a consistent estimator. The variances cannot grow faster than (or as) N^2 otherwise the sum will diverge to ∞. If $\sigma_k^2 < \sigma^2\, k$ the sum has a finite limit when $N \to \infty$ and the variance tends to zero.

SIMPLE LINEAR REGRESSION p. 58-59

Suppose you measure data pairs $\{x_k,\, y_k\,;\, k = 1,2,\dots,N\}$ where the x_k are known exactly but the y_k are uncertain because affected by errors.

You would like to describe approximately these data by a straight line say $y = \alpha + \beta\, x$. What is the best straight line approximating the measured data?

Suppose you model the measurement process by a statistical model

$$\mathbf{y}_k = \alpha + \beta\, x_k + \mathbf{e}_k\,, \qquad k = 1,2,\dots,N$$

where the errors $\mathbf{e}_k$ are **zero-mean independent random variables** with variances σ_k^2. In the given experimental condition ω you have observed the values $\mathbf{y}_k(\omega) = y_k\,;\, k = k = 1,2,\dots,N$ corresponding to errors $\mathbf{e}_k(\omega)$ (which you do not know).

Definition 3. *The* **least squares estimator** *of the parameter $(\alpha,\, \beta)$ is the solution of the minimization problem*

$$\min_{(\alpha,\beta)} \sum_{k=1}^{N} [y_k - (\alpha + \beta\, x_k)]^2$$

SOLUTION OF THE LINEAR REGRESSION PROBLEM 1

The minimizers are (formulas 2.2.14 and 2.2.13 in the book)

$$\hat{\alpha}_N = \bar{y}_N - \hat{\beta}_N \bar{x}_N$$

$$\hat{\beta}_N = \frac{\sum_k (x_k - \bar{x}_N)\, y_k}{\sum_k (x_k - \bar{x}_N)^2}$$

You may imagine that these are sample values of random variables (before collecting the data)

$$\hat{\boldsymbol{\alpha}}_N = \bar{\mathbf{y}}_N - \hat{\boldsymbol{\beta}}_N \bar{x}_N$$

$$\hat{\boldsymbol{\beta}}_N = \frac{\sum_k (x_k - \bar{x}_N)\, \mathbf{y}_k}{\sum_k (x_k - \bar{x}_N)^2}$$

Question: are these consistent estimators of the parameters (α, β)?

The answer depends on how you describe the errors.

SOLUTION OF THE LINEAR REGRESSION PROBLEM 2

The estimators are **unbiased**

$$\mathbb{E}\,\hat{\boldsymbol{\alpha}}_N = \mathbb{E}\,(\bar{\mathbf{y}}_N - \hat{\boldsymbol{\beta}}_N \bar{x}_N)$$

$$\mathbb{E}\,\hat{\boldsymbol{\beta}}_N = \frac{\sum_k (x_k - \bar{x}_N)\,\mathbb{E}\,\mathbf{y}_k}{\sum_k (x_k - \bar{x}_N)^2}$$

First not that

$$\sum_k (x_k - \bar{x}_N)\,\mathbb{E}\,\mathbf{y}_k = \beta \sum_k (x_k - \bar{x}_N)\,x_k$$

but since $\sum_k (x_k - \bar{x}_N) = 0$ then $\sum_k (x_k - \bar{x}_N)\,\bar{x}_N = 0$ as well , and hence

$$\mathbb{E}\,\hat{\boldsymbol{\beta}}_N = \beta\,\frac{\sum_k (x_k - \bar{x}_N)^2}{\sum_k (x_k - \bar{x}_N)^2} = \beta$$

Then can show that $\hat{\boldsymbol{\alpha}}_N$ is also unbiased.

PROOF OF UNBIASEDNESS

Since the errors are zero-mean $\mathbb{E}\,(\bar{\mathbf{y}}_N) = \alpha + \beta\,\bar{x}_N$ and hence

$$\mathbb{E}\,\hat{\boldsymbol{\alpha}}_N = \alpha + \mathbb{E}\,(\beta - \hat{\boldsymbol{\beta}}_N)\,\bar{x}_N$$

On the other hand $\mathbb{E}\,\mathbf{y}_k = \alpha + \beta\,x_k$, and so

$$\mathbb{E}\,(\hat{\boldsymbol{\beta}}_N - \beta) = \frac{\sum_k (x_k - \bar{x}_N)\,(\alpha + \beta\,x_k)}{\sum_k (x_k - \bar{x}_N)^2} - \beta = \beta\frac{\sum_k (x_k - \bar{x}_N)\,x_k}{\sum_k (x_k - \bar{x}_N)^2} - \beta = 0$$

since $\sum_k (x_k - \bar{x}_N)\,\alpha = 0$ and likewise $\sum_k (x_k - \bar{x}_N)\,\bar{x}_N = 0$.

CONSISTENCY OF THE LEAST SQUARES ESTIMATORS

First look at

$$\hat{\boldsymbol{\beta}}_N = \frac{\sum_k (x_k - \bar{x}_N)\,\mathbf{y}_k}{\sum_k (x_k - \bar{x}_N)^2} := \sum_k w_k\,\mathbf{y}_k$$

and since the $\mathbf{y}_k$ are independent (as the $\mathbf{e}_k$ are)

$$\mathrm{var}\,(\hat{\boldsymbol{\beta}}_N) = \sum_k w_k^2\,\mathrm{var}\,(\mathbf{y}_k) = \sum_{k=1}^{N} w_k^2\,\sigma_k^2$$

where $\sigma_k^2 = \mathrm{var}\,(\mathbf{e}_k)$. For convergence in probability of $\hat{\boldsymbol{\beta}}_N$ to β we need

$$\lim_{N\to\infty} \sum_{k=1}^{N} w_k^2\,\sigma_k^2 = 0$$

which, in case $\sigma_k^2 = \sigma^2$ independent of k implies

$$\sum_{k=1}^{N} w_k^2 = \frac{\sum_k (x_k - \bar{x}_N)^2}{[\sum_k (x_k - \bar{x}_N)^2]^2} = \frac{1}{\sum_k (x_k - \bar{x}_N)^2} \;\to\; 0\,.$$

This is the same as

$$\sum_{k=1}^{+\infty} (x_k - \bar{x}_N)^2 = \infty$$

which means that the points x_k should not remain too close to their sample mean. Since

$$\hat{\boldsymbol{\alpha}}_N = \alpha + (\beta - \hat{\boldsymbol{\beta}}_N)\, \bar{x}_N + \frac{1}{N} \sum_{k=1}^{N} \mathbf{e}_k$$

and, under this condition, both the last two terms converge to zero in probability, it is easy to see that the estimator $\hat{\boldsymbol{\alpha}}_N$ is also consistent in probability.

CONVERGENCE IN DISTRIBUTION

Definition 4. *A sequence of Pdf's $\{F_N\}$ (may be multivariable),* converges in law *to a Pdf F; notation: $F_N \stackrel{L}{\to} F$, if the functions $\{F_N(x)\}$, converge to a Pdf $F(x)$ at all points x where F is continuous.*

One also talks abut convergence in distribution (or also *in law*) of **random variables**: a sequence $\{\mathbf{x}_N\}$ (maybe vector valued), converges in distribution: $\mathbf{x}_N \stackrel{L}{\to} \mathbf{x}$ if the Pdf's of $\{\mathbf{x}_N\}$ converge in law to the Pdf of $\mathbf{x}$.
This is a weaker notion than convergence of random variables as defined above.

WARNING: To talk about convergence of random variables $\{\mathbf{x}_N\}$ and $\mathbf{x}$ must be **defined in the same probability space** (the same random experiment). Otherwise $F_N \to F$ does not necessarily mean that $\{\mathbf{x}_N(\omega)\}$ with $\mathbf{x}_N \sim F_N$ converges to a limit random variable $\mathbf{x}(\omega'), \omega' \in \Omega'$ in any reasonable sense. The elementary events ω' must lie in the **same space** Ω.

Theorem 3. *Convergence in probability implies convergence in distribution. (Theorem 2.3.5)*

Convergence in distribution is weaker than (implied by) convergence in probability **except when the limit is a constant (nonrandom) variable**. A *degenerate* PDF is

$$F(x) := \mathbb{1}(x-c) = \begin{cases} 1 & \text{if } x \geq c \\ 0 & \text{if } x < c \end{cases}$$

this is the Pdf of a constant (nonrandom) variable $\mathbf{x}(\omega) = c$ for all $\omega \in \Omega$.
Theorem 4. *Convergence in law to a degenerate Pdf (that is convergence in distribution to a constant) implies and is hence equivalent to convergence in probability to the same constant :*

$$\mathbf{x}_N \overset{L}{\to} c \Leftrightarrow \mathbf{x}_N \overset{P}{\to} c$$

whenever c is a (nonrandom) constant.

This is the setting of Lehmann's book.

Example 1. Let $\{\mathbf{x}_N\}$ be a i.i.d. sequence so that each $\mathbf{x}_N$ has the same distribution say $F_N = F_1 = F$ (a continuous function) for all n. Then obviously $F_N \to F_1$ but $\{\mathbf{x}_N\}$ cannot converge in probability to $\mathbf{x}_1$ since $\mathbf{x}_1$ is independent of all $\{\mathbf{x}_N\,;N > 1\}$.

In fact,

$$\mathbf{P}\{\omega \mid |\mathbf{x}_N(\omega) - \mathbf{x}_1(\omega)| > \varepsilon\} = 1 - \mathbb{E}\,\{F(\mathbf{x}_1 + \varepsilon) - F(\mathbf{x}_1 - \varepsilon)\}$$

which does not depend on n and hence cannot converge to zero.

Theorem 5 (Weak Convergence)**.** *The sequence of random variables $\{\mathbf{x}_N\}$ converges in distribution to $\mathbf{x}$ if and only if*

$$\mathbb{E}\,f(\mathbf{x}_N) \to \mathbb{E}\,f(\mathbf{x})\,; \quad \textit{that is} \quad \int f(x)\,dF_N(x) \to \int f(x)\,dF(x)$$

for all bounded continuous **real valued** *functions f. In fact for all real valued functions f which are bounded and continuous in a set of probability one for the Pdf of $\mathbf{x}$.*

CHARACTERISTIC FUNCTIONS

$$\phi_{\mathbf{x}}(it) := \int e^{itx} dF(x) = \mathbb{E}\, e^{it\,\mathbf{x}}$$

NOTE: The imaginary argument of the exponential here is essential to guarantee boundedness, as $|e^{itx}| = 1$.

Therefore **convergence in distribution implies pointwise convergence of the characteristic functions**

$$\phi_{\mathbf{x}_N}(it) := \mathbb{E}\, e^{it\,\mathbf{x}_N} \to \phi_{\mathbf{x}}(it) := \mathbb{E}\, e^{it\,\mathbf{x}}, \quad \text{for all } \ t \in \mathbb{R}\,.$$

Actually this result can be inverted

Theorem 6 (Levy-Helly Bray)**.** *The convergence of characteristic functions is necessary and sufficient (and hence* equivalent *to) convergence in distribution.*

This is a very useful fact. Used for example in the proof of the CLT.

CONVERGENCE OF MOMENTS

The moments of a Pdf are derivatives of the characteristic function computed at $t = 0$.

$$\phi^{(k)}(it) := i^k \int x^k e^{itx} dF(x) \Rightarrow \phi^{(k)}(0) := i^k \int x^k dF(x) = i^k \mu_k$$

Convergence $\phi_n(it) \to \phi(it)$ does not necessarily imply convergence of the derivatives at $t = 0$. In general **convergence in law does not imply convergence of the moments** . Means, variances etc.., etc., of a sequence $\{\mathbf{x}_N\} \overset{L}{\to} \mathbf{x}$, do not necessarily converge to means, variances etc.., of the limit.

Theorem 7 (Billingsley p.32)**.** *Let* $\mathbf{x}_N \overset{L}{\to} \mathbf{x}$*, and*

$$\sup_N \mathbb{E}\, \mathbf{x}_N^2 < \infty \qquad (1)$$

then all existing moments of $\mathbf{x}_N$ *converge to the respective moments of the limit distribution.*

CONTINUOUS MAPPING THEOREMS

The **Continuous Mapping Theorem** states that **for every continuous function $f(\cdot)$, if $\mathbf{x}_N \overset{P}{\to} \mathbf{x}$, then also $f(\mathbf{x}_N) \overset{P}{\to} f(\mathbf{x})$.**. Does it hold also for convergence in distribution?

We show that for **scalar random variables** this is *true also for convergence in distribution*. If $F_N \overset{L}{\to} F$, for every continuous function g composed with another arbitrary continuous function f we must have

$$\int_{\mathbb{R}} g(f(y))\, dF_N(y) \to \int_{\mathbb{R}} g(f(y))\, dF(y)$$

By a change of variable (suppose f is invertible)

$$\int_{\mathbb{R}} g(x)\, dF_N(f^{-1}(x)) \to \int_{\mathbb{R}} g(x)\, dF(f^{-1}(x))$$

where $F_N(f^{-1}(x))$ and $F((f^{-1}(x)))$ are the Pdf's of $f(\mathbf{x}_N)$ and $f(\mathbf{x})$.
NB: For vector functions this proof does not necessarily work. See Slutsky's Theorem below. See also Billingsley book pp. 29-30.

MULTIVARIATE STATISTICS

In many applications one has to deal with multiple measuremensts taken simultaneously. Also there may be many unknown parameters $\theta \equiv \{\theta_1, \theta_2 \ldots, \theta_p\}$. An estimator must then have the same dimension p of the parameter. Need to work with **multivariate random variables**.

Convenient to introduce **vector notation**. An n-dimensional **random vector** is denoted

$$\mathbf{x} := \begin{bmatrix} \mathbf{x}_1 \\ \ldots \\ \mathbf{x}_n \end{bmatrix}, \qquad \mathbf{x}^\top := \begin{bmatrix} \mathbf{x}_1 & \ldots & \mathbf{x}_n \end{bmatrix}$$

The *mean* $\mathbb{E}\,\mathbf{x} := \mu$ is a vector in $\mathbb{R}^n$. The *Covariance (or simply Variance)* of the vector $\mathbf{x}$ is the $n \times n$ matrix

$$\Sigma := \mathrm{Var}\{\mathbf{x}\} := \mathbb{E}\,(\mathbf{x} - \mu)(\mathbf{x} - \mu)^\top = \begin{bmatrix} \sigma_{1,1} & \ldots & \sigma_{1,n} \\ \ldots & \ldots & \ldots \\ \sigma_{n,1} & \ldots & \sigma_{n,n} \end{bmatrix}$$

Usual convention: $\sigma_{k,k} \equiv \sigma_k^2$ the variance of the k-th component $\mathbf{x}_k$.

BASICS ON LINEAR ALGEBRA

Refer to Lehmann Chapter 5 (with some notations changed) p.277.

A covariance matrix is always *symmetric* $\Sigma = \Sigma^\top$ and *positive semidefinite* that is the quadratic form $x^\top \Sigma x \geq 0$. Excluding pathological cases one has in fact strict positivity except when $x = 0$.

The *scalar variance* of a random vector $\mathbf{y}$ is the trace of the covariance matrix. Notation:

$$\mathrm{var}(\mathbf{y}) := \mathbb{E}\{\mathbf{y}^\top \mathbf{y}\} = \mathrm{Tr}\,\mathbb{E}\{\mathbf{y}\mathbf{y}^\top\} = \mathrm{Tr}\,\mathrm{Var}\,(\mathbf{y}) = \sum_{k=1}^{n} \sigma_k^2$$

Gaussian random vectors have a probability density function depending only on the mean vector μ and the Covariance matrix Σ:

$$p(x_1, \ldots, x_n) = \frac{1}{[2\pi^n \det \Sigma]^{1/2}} \exp -\frac{1}{2} \left\{ \begin{bmatrix} x_1 - \mu_1 & \ldots & x_n - \mu_n \end{bmatrix} \Sigma^{-1} \begin{bmatrix} x_1 - \mu_1 \\ \ldots \\ x_n - \mu_n \end{bmatrix} \right\}$$

Uniquely determined by the parameters μ, Σ.

EIGENVALUES AND EIGENVECTORS

Along some directions a square matrix $A \in \mathbb{R}^{n\times n}$ acts like a multiplication by a scalar

$$Av = \lambda v\,, \qquad v \in \mathbb{R}^n$$

the scalar factor λ is called the **eigenvalue** associated to the *eigenvector* v. Eigenvectors are actually directions in space and are usually normalized to unit norm. In general eigenvalues (and eigenvectors) are complex as they must be roots of the *characteristic polynomial* equation

$$\chi_A(\lambda) := \det(A - \lambda I) = 0$$

which is of degree n in λ and hence has n (not necessarily distinct) complex roots $\{\lambda_1, \ldots, \lambda_n\}$. This set is called *the spectrum of* A and is denoted $\sigma(A)$. The multiplicity of λ_k as a root of the characteristic polynomial is called the *algebraic multiplicity*.

When eigenvectors are linearly independent they form a basis in which the matrix A looks like multiplication by a diagonal matrix whose elements are the eigenvalues. Unfortunately this happens only for special classes of matrices.

SYMMETRIC MATRICES

Theorem 8. *Let $A = A^\top \in \mathbb{R}^{n\times n}$. Then*

1. The eigenvalues of A are real and the eigenvectors can be chosen to be a real orthonormal basis.

2. A is diagonalizable by an orthogonal transformation ($\exists T$ s.t. $T^\top T = I$ and $T^\top AT$ is diagonal).

3. A positive (semi-) definite matrix can always be taken symmetric. Its eigenvalues are real and positive (nonnegative).

MULTIVARIATE CONVERGENCE

Let $\mathbf{x}_N \sim F_N$ and $\mathbf{x} \sim F$ be n-dimensional random vectors; then $\mathbf{x}_N \overset{L}{\to} \mathbf{x}$ means that

$$F_N(x_1, x_2, \ldots, x_n) \to F(x_1, x_2, \ldots, x_n)$$

at all points $x = \begin{bmatrix} x_1 & x_2 & \ldots & x_n \end{bmatrix}^\top \in \mathbb{R}^n$ where F is continuous. This implies that *all marginals* converge, that is

$$F_N(x_1) \to F(x_1); \quad \ldots \quad ; F_N(x_n) \to F(x_n)$$

which can be written

$$\mathbf{x}_{1,N} \overset{L}{\to} \mathbf{x}_1; \quad \ldots \quad ; \mathbf{x}_{n,N} \overset{L}{\to} \mathbf{x}_n$$

But **the converse implication is not necessarily true !** as convergence of the marginals $F_N(x_k); \; k = 1, 2, \ldots, n$ does not imply convergence of the **joint distributions** at all points $x = \begin{bmatrix} x_1 & x_2 & \ldots & x_n \end{bmatrix}^\top \in \mathbb{R}^n$.

Example 2. *Let $\mathbf{x}_N$ and $\mathbf{y}_N$ be two scalar sequences converging separately in distribution to the random variables $\mathbf{x}$ and $\mathbf{y}$. Then $\mathbf{x}_N \overset{L}{\to} \mathbf{x}$ and $\mathbf{y}_N \overset{L}{\to} \mathbf{y}$ does NOT NECESSARILY imply that*

$$\begin{bmatrix} \mathbf{x}_N \\ \mathbf{y}_N \end{bmatrix} \overset{L}{\to} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}, \qquad \textit{(In general not true)}$$

unless **one of the two limits is a degenerate r.v. (a constant)** *c; see Slutsky Theorem.*

SLUTSKY THEOREM

Theorem 9 (Slutsky)**.** *Let* $\mathbf{x}_N \overset{L}{\to} \mathbf{x}$ *scalar random variables. Then:*

1. *For every continuous function* f, $f(\mathbf{x}_N) \overset{L}{\to} f(\mathbf{x})$.

2. *If* $\{\mathbf{y}_N\}$ *is a squence of random v's such that* $(\mathbf{x}_N - \mathbf{y}_N) \to 0$ *in probability, then* $\mathbf{y}_N$ *also converges in law to* $\mathbf{x}$ *(that is:* $\mathbf{y}_N \overset{L}{\to} \mathbf{x}$*).*

3. *Let* $\mathbf{z}_N = \begin{bmatrix} \mathbf{x}_N \\ \mathbf{y}_N \end{bmatrix}$ *and let the sequence* $\{\mathbf{y}_N; N = 1, 2, \ldots\}$ *converge in probability (or in law) to a constant* c. *Then if* $f(z) := f(x, y)$ *is a continuous function of the two arguments* $f(\mathbf{x}_N, \mathbf{y}_N) \overset{L}{\to} f(\mathbf{x}, c)$.

All statements are also valid for random vectors.

SLUTSKY THEOREM CONT'D

In statement (c) one cannot relax the assumption that $\{\mathbf{y}_N; N = 1, 2, \ldots\}$ converges to a constant c to convergence to a non-degenerate random variable.

Example 3 (Ferguson p.40)**.** *Let* $\mathbf{x}_N = \mathbf{x}$ *for all* N *where* $\mathbf{x} \sim U[0,\, 1]$ *and let* $\mathbf{y}_N = \mathbf{x}$ *for* N *odd and* $\mathbf{y}_N = 1 - \mathbf{x}$ *for* N *even.*

Then $\mathbf{y}_N \xrightarrow{L} U[0,\, 1]$ *but the (joint) distribution of the vector* $\mathbf{z}_N = \begin{bmatrix} \mathbf{x}_N \\ \mathbf{y}_N \end{bmatrix}$ *cannot converge in distribution as*

$$F_{\mathbf{z}_N}(x, y) = \begin{cases} \mathbf{P}\{\mathbf{x} \le x,\, \mathbf{x} \le y\} & = \min\{x,\, y\} \quad \text{for } N \text{ odd} \\ \mathbf{P}\{\mathbf{x} \le x,\, \mathbf{x} \le 1 - y\} & = \min\{x,\, 1 - y\} \quad \text{for } N \text{ even} \end{cases}$$

where $x, y \in [0,\, 1]$*.*

FERGUSON EXAMPLE CONT'D

We show that $\mathbf{y}_N \sim U[0,\, 1]$ for all N. Hence $F_{\mathbf{y}_N} \to U[0,\, 1]$ as $N \to \infty$.

Obviously true by definition for N odd. For N even $\mathbf{y} = 1 - \mathbf{x}$ hence

$$\mathbf{P}\{\mathbf{y} \leq y\} = \mathbf{P}\{1 - \mathbf{x} \leq y\} = \mathbf{P}\{\mathbf{x} \geq 1 - y\} = 1 - \mathbf{P}\{\mathbf{x} \leq 1 - y\} = 1 - (1 - y) = y$$

So also for N even $\mathbf{y}_N \sim U[0,\, 1]$.

Similarly you get

$$F_{\mathbf{z}_N}(x,y) = \begin{cases} \mathbf{P}\{\mathbf{x} \leq x,\, \mathbf{x} \leq y\} & = \min\{x,\, y\} \quad \text{for } N \text{ odd} \\ \mathbf{P}\{\mathbf{x} \leq x,\, \mathbf{x} \leq 1 - y\} & = \min\{x,\, 1 - y\} \quad \text{for } N \text{ even} \end{cases}$$

where $x, y \in [0,\, 1]$. Draw a picure of this function on the square $[0,\, 1]x[0,\, 1]$ for N odd and N even and check that it jumps. So $F_{\mathbf{z}_N}(x,y)$ cannot converge.

APPLICATIONS OF SLUTSKY THEOREM

(Lehmann p. 70)

Corollary 1 (Theorem 2.3.3)**.** *If* $\mathbf{x}_N \stackrel{L}{\to} \mathbf{x}$ *and two random variables* $\mathbf{a}_N, \mathbf{b}_N$ *converge in probability to constants* $(a,\, b)$ *(same as also converging in distribution !), then*

$$\mathbf{z}_N := \mathbf{a}_N + \mathbf{b}_N \mathbf{x}_N \stackrel{L}{\to} a + b\mathbf{x}$$

(In Lehmann $\mathbf{x}_N \equiv Y_n$*).*

Proof. Follows from statement (3) of Slutsky Theorem, just let $\mathbf{y}_N = \begin{bmatrix} \mathbf{a}_N \\ \mathbf{b}_N \end{bmatrix}$ and $c := \begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R}^2$ and $f(x,y) = a + bx$. □

Corollary 2 (Corollary 2.3.1)**.** *If* $\mathbf{x}_N \stackrel{L}{\to} \mathbf{x}$ *and* $\mathbf{r}_N \stackrel{L}{\to} 0$ *then*

$$\mathbf{x}_N + \mathbf{r}_N \stackrel{L}{\to} \mathbf{x}\,.$$

ABOUT SLUTSKY THEOREM

(Lehmann p. 70)

Warning: The Pdf of the sum of two random variables is **not the sum of the two Pdf's!!**

You may want to compute the Pdf (easier the pdf) of $\mathbf{x}+c$ given that of $\mathbf{x}$.

Also the Pdf of a random variable equal to zero a.s. $(\mathbf{x}=0)$ is **not the zero Pdf!** which in fact is not even a Pdf function!

Two sequences of random variables $\{\mathbf{x}_N\}$, $\{\mathbf{y}_N\}$ such that $(\mathbf{x}_N-\mathbf{y}_N)\overset{P}{\to}0$ in probability (equiv. in distribution, since 0 is a constant), are said to be **asymptotically equivalent**.

THE CENTRAL LIMIT THEOREM (CLT)

The first version of the *Central Limit Theorem* is due to De Moivre and Laplace for binomial random variables and later to Gauss for sum of continuous i.i.d. variables.

What do we mean by a limit distribution ? Let $\{\mathbf{x}_k\}$ be i.i.d. random variables of mean μ and variance σ^2. By the law of large numbers the sample mean $\bar{\mathbf{x}}_N = \frac{1}{N}\sum_{t=1}^{N} \mathbf{x}_t$ converges to the mean $\mu = \mathbb{E}\,\mathbf{x}_k$ in probability (in fact almost surely) as $N \to \infty$. The variance of $\bar{\mathbf{x}}_N$ must then tend to zero [CHECK THIS !] and the limit distribution must obviously be **degenerate**

$$\bar{\mathbf{x}}_N \overset{L}{\to} \mu$$

This is not very interesting. Let us note that $\operatorname{var}\bar{\mathbf{x}}_N$ **tends to zero as** $\frac{1}{N}$. In fact

$$\operatorname{var}(\bar{\mathbf{x}}_N) = \frac{1}{N}\left(\frac{1}{N}\sum_{k=1}^{N}\operatorname{var}(\mathbf{x}_k)\right) = \frac{1}{N}\sigma^2 .$$

THE CLASSICAL CLT

Hence as $N \to \infty$, the variance of the random variable $\sqrt{N}\,[\,\bar{\mathbf{x}}_N - \mu\,]$ has a finite limit (the variance of each $\mathbf{x}_k$). Then its limit distribution cannot be degenerate. The key is to discover the **convergence rate**.

Theorem 10. *Assume $\{\mathbf{x}_k\}$ are i.i.d. random variables of mean μ and finite variance σ^2. Then $\mathbf{y}_N := \sqrt{N}\,(\bar{\mathbf{x}}_N - \mu)$ converges in distribution to a Gaussian of mean zero and variance σ^2. Letting $\tilde{\mathbf{x}}_k := \mathbf{x}_k - \mu$ this is the same as*

$$\sqrt{N}\frac{1}{N}\sum_{k=1}^{N}\tilde{\mathbf{x}}_k \xrightarrow{L} \mathcal{N}(0,\sigma^2)$$

May be written in short also as $\sqrt{N}\,\bar{\tilde{\mathbf{x}}}_N \xrightarrow{L} \mathcal{N}(0,\sigma^2)\,.$

NB *Theorem is no longer true if you do not subtract the mean*:

$$\sqrt{N}\bar{\mathbf{x}}_N = \sqrt{N}\frac{1}{N}\sum_{k=1}^{N}\mathbf{x}_k = \sqrt{N}\frac{1}{N}\sum_{k=1}^{N}\tilde{\mathbf{x}}_k + \sqrt{N}\,\mu$$

since $\sqrt{N}\,\mu$ does not converge in distribution.

ABOUT THE CONVERGENCE RATE

(Lehmann last paragraph of p.70 gives only an intuitive argument)

When $\bar{\mathbf{x}}_N \overset{L}{\rightarrow} c$ then (under suitable assumptions) the variance σ_N^2 of $\bar{\mathbf{x}}_N$ must tend to zero. (*Show that the variance of a degenerate Pdf is zero !*).

The main question is: Does there exist a deterministic function $k(N) \rightarrow \infty$ such that

$$k(N)\,(\bar{\mathbf{x}}_N - c) \overset{L}{\rightarrow} F$$

where $F(x)$ is a (limit) **non-degenerate** Pdf ?

Example: if $\bar{\mathbf{x}}_N$ is the sample mean of N i.i.d. random variables, $c = \mu$ and $k(N) = \sqrt{N}$, then $F \equiv \mathcal{N}(0, \sigma^2)$. This is the CLT.

PROOF OF THE CLT 1

Need to review properties of the characteristic function, see Lehmann p. 581.

Proposition 2. *The characteristic function of* $\mathcal{N}(0,1)$ *is the function*

$$\phi_o(it) = e^{\frac{|it|^2}{2}} = e^{-\frac{t^2}{2}}.$$

Proposition 3. *Let* $\bar{\mathbf{x}} = \sum_{k=1}^{N} \mathbf{x}_k$ *be the sum of independent random variables (not necessarily having the same distribution).The characteristic function of* $\bar{\mathbf{x}}$ *is*

$$\phi_{\bar{\mathbf{x}}}(it) = \phi_{\mathbf{x}_1}(it)\phi_{\mathbf{x}_2}(it)\cdots\phi_{\mathbf{x}_N}(it).$$

The proof is from the property of the exponential function

$$\begin{aligned}\phi_{\bar{\mathbf{x}}}(it) &= \mathbb{E}\, e^{it\,\bar{\mathbf{x}}} = \mathbb{E}\, e^{it\sum_{k=1}^{N}\mathbf{x}_k} \\ &= \prod_{k=1}^{N} \mathbb{E}\, e^{it\,\mathbf{x}_k} = \prod_{k=1}^{N} \int e^{itx_k}\, dF_k(x_k) = \phi_{\mathbf{x}_1}(it)\phi_{\mathbf{x}_2}(it)\cdots\phi_{\mathbf{x}_N}(it).\end{aligned}$$

If the variables are i.i.d. then $\phi_{\bar{\mathbf{x}}}(it) = \phi_{\mathbf{x}_1}(it)^N$.

Another simple fact used in the proof of the CLT:

if $\bar{\mathbf{y}} = \alpha \bar{\mathbf{x}}$ where $\alpha \neq 0$, then

$$\phi_{\bar{\mathbf{y}}}(it) = \mathbb{E}\, e^{it\,\bar{\mathbf{y}}} = \mathbb{E}\, e^{i\dfrac{t}{\alpha}\bar{\mathbf{x}}} = \phi_{\bar{\mathbf{x}}}(i\frac{t}{\alpha})$$

PROOF OF THE CLT 2

Let $\mathbf{y}_k := \mathbf{x}_k - \mu$; then $\mathbf{y}_1,\ \mathbf{y}_2,\ \ldots \mathbf{y}_N$ are i.i.d. of mean zero and variance σ^2. We shall show that

$$\mathbf{z}_N := \sqrt{N}\,\bar{\mathbf{y}}_N \overset{L}{\to} \mathcal{N}(0, \sigma^2)\,.$$

In fact from

$$\phi_{\mathbf{z}_N}(it) = \prod_{k=1}^{N} \mathbb{E}\left[\exp\{\frac{it}{\sqrt{N}}\mathbf{y}_k\}\right]$$

take Taylor expansion about $t = 0$

$$\exp\{\frac{it}{\sqrt{N}}y\} = 1 + \frac{it}{\sqrt{N}}y + \frac{1}{N}\frac{|it|^2}{2}y^2 + o[(\frac{|t|}{N})^2 y^2]$$

substitute $y = \mathbf{y}_k$ and take expectation. Since $\mathbb{E}\,\mathbf{y}_k = 0\,;\, \mathbb{E}\,\mathbf{y}_k^2 = \sigma^2$ you get

$$\phi_{\mathbf{z}_N}(it) = \left[1 + \frac{1}{N}\frac{-t^2}{2}\sigma^2\right]^N + o[(\frac{|t|^2}{N})^N]$$

PROOF OF THE CLT 3

Then pass to the limit as $N \to +\infty$. As in Lehmann p. 41, Prob.4.8

$$\lim_{N\to\infty}\left[1-\frac{1}{N}\frac{t^2}{2}\sigma^2\right]^N = e^{-\dfrac{t^2}{2}\sigma^2}$$

which is the characteristic function of $\mathcal{N}(0,\sigma^2)$. By Levy-Helly-Bray Theorem :

$$\sqrt{N}\,\bar{\mathbf{y}}_N \xrightarrow{L} \mathcal{N}(0,\sigma^2)\,.$$

NB: **this result is independent of the distribution of the $\mathbf{x}_k$'s !!**

It could be anything provided the mean and variance are finite, e.g. Binomial, $U[a,\,b]$ etc. See Lehmann pp. 73-74.

Example 4 (Lehmann p. 75-76)**.** Let $\mathbf{y} := \{\mathbf{y}_1,\, \mathbf{y}_2 \ldots \mathbf{y}_N\}$ a i.i.d. sequence where each $\mathbf{y}_k$ has mean μ and variance σ^2. Find the asymptotic (limit) distribution of the random variable

$$\varphi_N(\mathbf{y}) := \frac{\sqrt{N}\,[\bar{\mathbf{y}}_N - \mu]}{\sqrt{\hat{\sigma}_N^2(\mathbf{y})}}$$

where $\hat{\sigma}_N^2(\mathbf{y})$ is the sample variance

$$\hat{\sigma}_N^2(\mathbf{y}) = \frac{1}{N}\sum_{k=1}^{N}(\mathbf{y}_k - \bar{\mathbf{y}}_N)^2$$

Solution : We know that

$$\sqrt{N}\,[\bar{\mathbf{y}}_N - \mu] \xrightarrow{L} \mathcal{N}(0, \sigma^2)$$

on the other hand, as $N \to \infty$,

$$\hat{\sigma}_N^2(\mathbf{y}) \xrightarrow{P} \sigma^2$$

(in fact also almost surely). By (3) of Slutsky Theorem

$$\varphi_N(\mathbf{y}) \xrightarrow{L} \mathcal{N}(0, 1)\,.$$

If it is known that $\mu = 0$, then $\varphi_N(\mathbf{y})$ is the so called *Student's* $\mathbf{t}$ *statistics*.

Exercise (2.4.4): Assume $\tau^2 = \operatorname{var} \mathbf{y}_k^2$ is finite. Prove that under the i.i.d. assumption

$$\hat{\sigma}_N^2(\mathbf{y}) \xrightarrow{P} \sigma^2 .$$

Use the CLT to prove that if $\mu = \mathbb{E}\,\mathbf{y}_k$ is known, the asymptotic distribution of the sample second order moment is Gaussian; in fact,

$$\sqrt{N}\frac{1}{N}\sum_{k=1}^{N}(\mathbf{y}_k - \mu)^2 \xrightarrow{L} \mathcal{N}(0, \tau^2)$$

But for finding the the asymptotic distribution of the sample variance $\hat{\sigma}_N^2(\mathbf{y})$ we need a more sophisticated tool. See Lehmann pp. 75-76.

SOLUTION

See Lehmann p. 51-52

$$\hat{\sigma}_N^2(\mathbf{y}) = \frac{1}{N}\sum_{k=1}^N (\mathbf{y}_k - \bar{\mathbf{y}}_N)^2 = \frac{1}{N}\{\sum_{k=1}^N (\mathbf{y}_k - \mu)^2 + 2(\mathbf{y}_k - \mu)(\mu - \bar{\mathbf{y}}_N) + (\mu - \bar{\mathbf{y}}_N)^2\}$$

$$= \frac{1}{N}\sum_{k=1}^N (\mathbf{y}_k - \mu)^2 + (\bar{\mathbf{y}}_N - \mu)^2 := \mathbf{s}_N^2(\mathbf{y}) + (\bar{\mathbf{y}}_N - \mu)^2$$

Now $(\bar{\mathbf{y}}_N - \mu)^2 \overset{P}{\to} 0$ since $\bar{\mathbf{y}}_N - \mu \overset{P}{\to} 0$; and since $\mathbb{E}\mathbf{s}_N^2(\mathbf{y}) = \sigma^2$,

$$\mathrm{var}\{\mathbf{s}_N^2(\mathbf{y})\} = \mathbb{E}\left[\frac{1}{N}\sum_{k=1}^N (\mathbf{y}_k - \mu)^2 - \sigma^2\right]^2 = \frac{1}{N}\left[\tau^2 - 2\sigma^4 + \sigma^4\right]$$

which tends to zero for $N \to \infty$. Therefore $\mathbf{s}_N^2(\mathbf{y}) \overset{P}{\to} \sigma^2$ and so does $\hat{\sigma}_N^2(\mathbf{y})$.

THE χ^2 DISTRIBUTION

One says that a scalar random variable $\mathbf{y}$ has a $\chi^2(n)$ distribution if its pdf is supported on the nonnegative real line and has the following form:

$$P(x \leq \mathbf{y} < x+dx) = \frac{1}{2^{n/2}\,\Gamma\left(\frac{n}{2}\right)}\, x^{\left(\frac{n}{2}\right)-1}\, e^{-x/2}\, dx\,, \qquad x \geq 0\,. \tag{2}$$

In this expression n is a natural number called *the number of degrees of freedom* of the distribution. One sees that the χ^2 is a special case of the Gamma distribution. Its characteristic function is

$$\phi(it) := \mathbb{E}\, e^{it\,\mathbf{y}} = (1-2it)^{-n/2}\,. \tag{3}$$

From this one can derive formulas for the moments of the distribution. The first few *central* moments are

$$\begin{aligned} \mu_1 &= n \\ \mu_2 &= 2n \\ \mu_3 &= 8n \\ \mu_4 &= 48n+12n^2 \qquad \text{ecc...} \end{aligned} \tag{4}$$

Lemma 1. *For large n a $\chi^2(n)$ random variable tends in distribution to a Gussian variable with pdf $\mathscr{N}(n, 2n)$.*

Proof. Let $\mathbf{y} \sim \chi^2(n)$; introduce a standardized random variable

$$\mathbf{z}_n := \frac{\mathbf{y} - n}{\sqrt{2n}} \quad ;$$

which for all n has mean zero and unit variance. Of course $\mathbf{z}_n$ is no longer a χ^2 (as this could happen only for *Gaussian* random variables!). We shall show that the limit in distribution, $L - \lim_{n \to \infty} \mathbf{z}_n$, is a standard $\mathscr{N}(0, 1)$ density. Recall that

Proposition 4. *Let $\mathbf{x}_n$ be a sequence of random variables with characteristic functions $\phi_n(it)$ then*

$$\mathbf{x}_n \overset{L}{\to} \mathbf{x} \qquad \textit{if and only if} \qquad \phi_n(it) \to \phi(it)\,; \forall t\,. \tag{5}$$

Proof: the characteristic function, $\phi_n(it)$, of $\mathbf{z}_n$ can be written as,

$$\begin{aligned}
\phi_n(it) &= \mathbb{E}\, e^{it\frac{\mathbf{y}}{\sqrt{2n}}}\, e^{-it\frac{n}{\sqrt{2n}}} = e^{-it\frac{n}{\sqrt{2n}}}\left(1-\frac{2it}{\sqrt{2n}}\right)^{-n/2} \\
&= \left(e^{-it\sqrt{\frac{2}{n}}}\right)^{n/2}\left(1-it\sqrt{\frac{2}{n}}\right)^{-n/2} \\
&= \left[e^{it\sqrt{\frac{2}{n}}}-it\sqrt{\frac{2}{n}}\,e^{it\sqrt{\frac{2}{n}}}\right]^{-n/2} = \left(1-\frac{t^2}{n}+\frac{\psi(n)}{n}\right)^{-n/2} \quad ,
\end{aligned}$$

where $\lim_{n\to\infty}\psi(n)=0$. By a well known formula in Analysis the limit $\lim_{n\to\infty}\phi_n(t)$ is equal to

$$\phi(t)=\lim_{n\to\infty}(1-t^2/n)^{n/2}=e^{-t^2/2} \quad ,$$

which is the characteristic function of a standard Gaussian distribution. □

The χ^2 distribution plays a role in many questions of statistical inference, especially entering in the pdf of estimators.

Proposition 5. *The sum of N independent random variables $\mathbf{y}_i \sim \chi^2(n_i)$ is distributed as $\chi^2(n)$ where*

$$n = \sum_{i=1}^{N} n_i \quad , \tag{6}$$

that is, when summing i.i.d. χ^2's, the degrees of freedom add up.

Proof. Recall that the pdf of the sum $\sum_1^N \mathbf{y}_i$ of i.i.d. random variables is just the N-fold **convolution of the respective p.d.f's**, so that the characteristic functions $\phi_i(t)$, of the $\mathbf{y}_i$'s get multiplied together. It is then clear that multiplying functions like (3) the exponents at the denominators must add up. □

The following is a partial converse of this statement.

Proposition 6. *Let $\mathbf{y} = \mathbf{y}_1 + \mathbf{y}_2$ be the sum of two independent random variables. Assume that $\mathbf{y} \sim \chi^2(n)$ and $\mathbf{y}_2 \sim \chi^2(n_2)$ where $n > n_2$. Then $\mathbf{y}_1 \sim \chi^2(n - n_2)$.*

Proof. By independence the characteristic function of $\mathbf{y}$ is $\phi = \phi_1\phi_2$ so that

$$\phi_1 = \frac{\phi}{\phi_2}$$

and by substituting the relative expressions (3) one sees that the statement must be true. □

Proposition 7. *The pdf of the random variable*

$$\frac{N\,\bar{\mathbf{s}}_N^2}{\sigma^2} := \frac{1}{\sigma^2}\sum_1^N (\mathbf{y}_k - \mu)^2 \quad ,$$

where $\mathbf{y}_k \sim \mathscr{N}(\mu, \sigma^2)$ *i.i.d. is* $\chi^2(N)$*.*

Proof. Just note that, with $\mathbf{y} \sim \mathscr{N}(\mu, \sigma^2)$, the pdf of $\mathbf{z} := (\mathbf{y} - \mu)^2/\sigma^2$ is $\chi^2(1)$ and then use Proposition 5. □

Note also that

Proposition 8. *The pdf of* $\mathbf{z} = \mathbf{x}^2$ *with* $\mathbf{x} \sim \mathscr{N}(0,1)$ *is* $\chi^2(1)$*.*

Proof. Using the well-known rules for the pdf of a function of random variable, say $z = f(x)$ with $f(x) = x^2$, one obtains

$$\begin{aligned} p_{\mathbf{z}}(z) &= \frac{1}{\left|\frac{d}{dx}\, f(x)\Big|_{x=f^{-1}(z)}\right|} \left[p_{\mathbf{x}}(\sqrt{z}) + p_{\mathbf{x}}(-\sqrt{z})\right] \, \mathbb{1}(z) \\ &= \frac{1}{|2\sqrt{z}|} \frac{1}{\sqrt{2\pi}} [e^{-z/2} + e^{-z/2}] \, \mathbb{1}(z) = \frac{1}{\sqrt{2\pi z}} e^{-z/2} ; z \geq 0 \quad , \end{aligned}$$

which is indeed $\chi^2(1)$. □

Proposition 9. *Let* $\mathbf{y}_k \sim \mathcal{N}(\mu, \sigma^2)$*,* $k = 1, \ldots, N$*, i.i.d. Then the pdf of the normalized sample variance:*

$$\frac{N\hat{\boldsymbol{\sigma}}_N^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_1^N (\mathbf{y}_k - \bar{\mathbf{y}}_N)^2 \quad ,$$

is $\chi^2(N-1)$*.*

Proof. This follows from the following remarkable result:

Lemma 2. *Under the above hypotheses, the statistics $\bar{\mathbf{y}}_N$ and $\hat{\boldsymbol{\sigma}}_N^2$ are independent.*

Proof. We just need to show that $\bar{\mathbf{y}}_N$ and $\mathbf{y}_k - \bar{\mathbf{y}}_N$ are uncorrelated for all k's. By Gaussianity, this will imply independence.

Define $\mathbf{x}_k = \mathbf{y}_k - \mu$ and $\bar{\mathbf{x}}_N := \bar{\mathbf{y}}_N - \mu$, so that $\mathbf{y}_k - \bar{\mathbf{y}}_N = \mathbf{x}_k - \ \bar{\mathbf{x}}_N$ and $\mathbb{E}\,\bar{\mathbf{y}}_N(\mathbf{y}_k - \bar{\mathbf{y}}_N) = \mathbb{E}\,\bar{\mathbf{x}}_N(\mathbf{x}_k - \bar{\mathbf{x}}_N) = \mathbb{E}\,\bar{\mathbf{x}}_N\mathbf{x}_k - \mathbb{E}\,\bar{\mathbf{x}}_N^2$. Independence of the variables $\mathbf{y}_k$ implies

$$\mathbb{E}\,\bar{\mathbf{x}}_N\mathbf{x}_k = \frac{1}{N}\,\mathbb{E}\left(\sum_1^N \mathbf{x}_k\mathbf{x}_i\right) = \frac{1}{N}\,\mathbb{E}\,(\mathbf{x}_i)^2 = \frac{\sigma^2}{N}$$

so that, comparing with $\mathbb{E}\,(\bar{\mathbf{x}}_N)^2 = \sigma^2/N$, one gets the conclusion. □

By the identity

$$\sum_1^N (\mathbf{y}_k - \mu)^2 = \sum_1^N (\mathbf{y}_k - \bar{\mathbf{y}}_N)^2 + N(\bar{\mathbf{y}}_N - \mu)^2 \tag{7}$$

one has

$$\sum_{1}^{N} \frac{(\mathbf{y}_k-\mu)^2}{\sigma^2} = \sum_{1}^{N} \frac{(\mathbf{y}_k-\bar{\mathbf{y}}_N)^2}{\sigma^2} + N\,\frac{(\bar{\mathbf{y}}_N-\mu)^2}{\sigma^2}$$

where the two random variables in the right member are *independent*. We know from Proposition 7 that $N\,\mathbf{s}_N^2/\sigma^2 \sim \chi^2(N)$ and that $(\bar{\mathbf{y}}_N-\mu)^2/(\sigma^2/N) \sim \chi^2(1)$ (which also follows from Proposition 7 with $N=1$). By Proposition 6 the pdf of first summand in the second member must be $\chi^2(N-1)$. □

So far we have been discussing the case of scalar variables. Suppose $\mathbf{y}$ is an m-dimensional random vector. We are interested in finding out when the pdf of quadratic forms like $\mathbf{y}^\top Q\mathbf{y}$ con $Q=Q^\top$, is χ^2. The most obvious situation in which this happens is the following.

Proposition 10. *Let* $\mathbf{y} \sim \mathcal{N}(\mu,\Sigma)$ *with* $\mu \in \mathbb{R}^m$ *and* $\Sigma \in \mathbb{R}^{m\times m}$ *positive definite; then*

$$(\mathbf{y}-\mu)^\top\,\Sigma^{-1}(\mathbf{y}-\mu) \sim \chi^2(m)\,. \tag{8}$$

Proof. One just needs to standardize $\mathbf{y}$, by setting $\mathbf{z} := \Sigma^{-1/2}(\mathbf{y} - \mu) \quad$; so that $\mathbf{z} = [z_1, \ldots, \mathbf{z}_m]^\top$ is $\mathscr{N}(0, I)$, in particular $\mathbf{z}_1, \ldots, \mathbf{z}_m$ are i.i.d. and $\mathscr{N}(0, 1)$. It follows that

$$(\mathbf{y} - \mu)^\top \Sigma^{-1} (\mathbf{y} - \mu) = \mathbf{z}^\top \mathbf{z} = \sum_1^m \mathbf{z}_i^2$$

and the last member is $\chi^2(m)$ by Proposition 5. □

A less obvious characterization which is used frequently is the following.

Proposition 11. *Let $\mathbf{z} \sim \mathscr{N}(0, I_m)$ and $Q \in \mathbb{R}^{m \times m}$. Then the quadratic form $\mathbf{z}^\top Q \mathbf{z}$ is χ^2 distributed if and only if Q is idempotent; i.e. $Q = Q^2$. In this case the number of degrees of freedom is equal to $r = \operatorname{rank} Q$.*

Proof. The proof is based on diagonalization of Q. Indeed since Q is symmetric (and can always be assumed to be such) and idempotent, it is really an orthogonal projection in $\mathbb{R}^m$. Its non-zero eigenvalues are all equal to 1 and there are exactly $r = \operatorname{rank} Q$ of them. The spectral decomposition of Q can therefore be written

$$Q = U \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} U^\top, \qquad UU^\top = U^\top U = I_m$$

that is

$$Q = U_1 U_1^\top,$$

where U_1 is an $m \times r$ matrix formed by the first r (orthonormal) columns of U. Hence

$$\mathbf{z}^\top Q \mathbf{z} = \mathbf{z}_1^\top \mathbf{z}_1$$

where the r-dimensional random vector $\mathbf{z}_1 := U_1^\top \mathbf{z}$ is distributed as $\mathscr{N}(0, I_r)$. Proposition 5 then yields the conclusion. □

THE STUDENT DISTRIBUTION

Let $\mathbf{y} \sim \mathcal{N}(0\,,\,1)$ and $\mathbf{x} \sim \chi^2(n)$ be independent. Then the ratio

$$\mathbf{t} := \frac{\mathbf{y}}{\sqrt{\mathbf{x}/n}} \tag{9}$$

has the pdf

$$p_n(t) = \frac{1}{\sqrt{n}B(1/2\,,\,n/2)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}} \qquad t \in \mathbb{R} \tag{10}$$

called a *Student distribution with* n *degrees of freedom*, which we shall denote by the symbol $\mathscr{S}(n)$. In (10) B Euler Beta function:

$$B(p\,,\,q) := \int_0^1 x^{p-1}(1-x)^{q-1}\,dx = \frac{\Gamma(p)\,\Gamma(q)}{\Gamma(p+q)}$$

where the function Γ is the well-known generalization of the factorial. When n is an integer greater than 1, $\Gamma(n) = (n-1)!$.

The Student pdf has a curious history which is reported in all textbooks of classical Statistics. For $n = 1$ it reduces to the *Cauchy distribution* :

$$\mathscr{S}(1) \equiv \frac{1}{\pi(1+t^2)}\,.$$

It can be shown that $\mathscr{S}(n)$ has finite moments only up to order $n-1$, given by the formulas

$$\begin{aligned}
\mu_r &= \quad 0 \quad \text{when } r \text{ i sodd and } \ r < n \\
\mu_r &= \frac{\Gamma(\frac{1}{2}n - r)\Gamma(r + \frac{1}{2})}{\Gamma(\frac{1}{2}n)\Gamma(\frac{1}{2})} \quad \text{when } r \text{ is even and } \ 2r < n\,.
\end{aligned}$$

It is also not hard to show that for $n \to \infty$ the distribution $\mathscr{S}(n)$ converges to $\mathcal{N}(0\,,\,1)$. (Example 4)

Example 5. It is not difficult to check that he Cauchy distribution centered at $y = \theta$

$$\mathcal{C}(y,\theta) = \frac{1}{\pi}\,\frac{1}{1+(y-\theta)^2}; \qquad \theta \in R \quad .$$

has characteristic function

$$\phi(it) = \frac{1}{\pi}\int_{-\infty}^{+\infty} \frac{1}{1+(y-\theta)^2}\, e^{ity}\, dy = e^{it\theta-|t|}$$

Suppose that $\mathbf{x}_1, \mathbf{x}_1, \ldots \mathbf{x}_N$ is an i.i.d sample from the Cauchy distribution. Its sample mean $\bar{\mathbf{x}}_N$ has the characteristic function

$$\phi_{\bar{\mathbf{x}}_N}(it) = \mathbb{E}\, e^{i\frac{t}{N}\sum \mathbf{x}_k} = e^{i\frac{t}{N}[N\theta - N|t|]} = \; e^{it\theta-|t|} = \phi(it)$$

which is invariant so that $\phi_{\bar{\mathbf{x}}_N}(it)$ converges to the same initial characteristic function. In distribution, $\bar{\mathbf{x}}_N \overset{L}{\to} \mathbf{x}_k$ (k arbitrary).

This implies that $\bar{\mathbf{x}}_N$ as an estimator of θ is not consistent (in fact $\mathbb{E}_{\,\theta}\mathbf{x}_k = \infty$!). Moreover the characteristic function of $\sqrt{N}\,\bar{\mathbf{x}}_N$ diverges with N at all points $t \in \mathbb{R}$. *No CLT for the Cauchy distribution!*

CRAMÈR'S THEOREM

Ferguson p. 45, see also Theorem 2.5.2 in Lehmann

Theorem 11. *Let $g : \mathbb{R}^n \to \mathbb{R}^m$ have continuous partial derivatives and consider the $n \times m$ Jacobian matrix:*

$$G(x) := \left[\frac{\partial g_i}{\partial x_j}\right]_{i=1,\dots n,\, j=1,\dots,m}$$

If $\mathbf{y}_1, \mathbf{y}_2, \dots$ is a sequence of n-dimensional random vectors (not necesarily i.i.d.) such that $\sqrt{N}\,(\mathbf{y}_N - \mu) \xrightarrow{L} \mathbf{y}$ then

$$\sqrt{N}\,(\,g(\mathbf{y}_N) - g(\mu)\,) \xrightarrow{L} G(\mu)\mathbf{y}\,.$$

In particular, if $\sqrt{N}\,(\mathbf{y}_N - \mu) \xrightarrow{L} \mathcal{N}(0, \Sigma)$ where Σ is a covariance matrix, then

$$\sqrt{N}\,(\,g(\mathbf{y}_N) - g(\mu)\,) \xrightarrow{L} \mathcal{N}(0, G(\mu)\Sigma G(\mu)^\top)\,.$$

Can compute the asymptotic distribution of many functions of the sequence $\mathbf{y}_1, \mathbf{y}_2, \dots$.

APPLICATIONS OF CRAMÈR THEOREM

Many examples in Lehmann pp 86-90

Example 6. *Here $n = 1$. Suppose that $\bar{\mathbf{x}}_N = \frac{1}{N}\sum_{k=1}^{N} \mathbf{x}_k$ wher $\mathbf{x}_k$ are not necessarily i.i.d. with common mean μ and variance σ^2 and $\sqrt{N}\,(\bar{\mathbf{x}}_N - \mu) \overset{L}{\to} \mathcal{N}(0,\, \sigma^2)$. What is the asymptotic distribution of $\bar{\mathbf{x}}_N^2$?*

Solution: Let $g(x) = x^2$ then $g'(x) = \dfrac{d\,g}{d\,x} = 2x$. Hence

$$\sqrt{N}\,(\bar{\mathbf{x}}_N^2 - \mu^2) \overset{L}{\to} \mathcal{N}(0, 4\mu^2\sigma^2)\,.$$

Warning: If $\mu = 0$ this just says that $\sqrt{N}\,\bar{\mathbf{x}}_N^2 \overset{L}{\to} 0$ which is a degenerate distribution. Means that $\dfrac{1}{\sqrt{N}}$ is not the correct convergence rate. In fact by Proposition 8 and Slutsky Theorem

$$N\frac{\bar{\mathbf{x}}_N^2}{\sigma^2} \overset{L}{\to} \chi^2(1)\,.$$

Example 7 (Lehmann ppp 75-76)**.** *Show that the asymptotic distribution of the sample variance $\hat{\boldsymbol{\sigma}}_N^2(\mathbf{x})$, of an i.i.d. sample having finite fourth order moment μ_4 is*

$$\sqrt{N}\,(\hat{\boldsymbol{\sigma}}_N^2(\mathbf{x}) - \sigma^2) \xrightarrow{L} \mathcal{N}(0, \mu_4 - \sigma^4)\,.$$

Solution: Using the identity (7) the sample variance can be written

$$\hat{\boldsymbol{\sigma}}_N^2(\mathbf{x}) \;=\; \frac{1}{N}\sum_{k=1}^{N}(\mathbf{x}_k - \bar{\mathbf{x}}_N)^2 = \frac{1}{N}\sum_{k=1}^{N}(\mathbf{x}_k - \mu)^2 - (\bar{\mathbf{x}}_N - \mu)^2 := m_2(\mathbf{x}) - m_1(\mathbf{x})^2\,.$$

To find the asymptotic distribution of $\hat{\boldsymbol{\sigma}}_N^2(\mathbf{x})$ we use Cramèr Theorem. Let us define the function $g(m_1, m_2) = -m_1^2 + m_2$. Then

$$\hat{\boldsymbol{\sigma}}_N^2(\mathbf{x}) = -m_1(\mathbf{x})^2 + m_2(\mathbf{x}) := g(m_1(\mathbf{x}), m_2(\mathbf{x})) = g(\begin{bmatrix} m_1(\mathbf{x}) \\ m_2(\mathbf{x}) \end{bmatrix})$$

Note that the two components are correlated. One cannot “add” the two asymptotic distributions. Need to use the joint distribution of the random vector.

Since the sample $\mathbf{x}$ is i.i.d. one has

$$\sqrt{N}\, m_1(\mathbf{x}) \overset{L}{\to} \mathcal{N}(0,\, \sigma^2)$$

. Similarly, since also the $(\mathbf{x}_k - \mu)^2\,;\, k = 1,2,\ldots$ are i.i.d. one has

$$\sqrt{N}\, m_2(\mathbf{x}) := \sqrt{N}\, \frac{1}{N}\sum_{k=1}^{N} (\mathbf{x}_k - \mu)^2 \overset{L}{\to} \mathcal{N}(\sigma^2,\, \mu_4 - \sigma^4)$$

because $\mathrm{var}\,(\mathbf{x}_k - \mu)^2 = \mathbb{E}\,(\mathbf{x}_k - \mu)^4 - 2\sigma^4 + \sigma^4 = \mu_4 - \sigma^4$ is the central fourth order moment.

By the CLT *for random vectors*

$$\sqrt{N}\left\{ \begin{bmatrix} m_1(\mathbf{x}) \\ m_2(\mathbf{x}) \end{bmatrix} - \begin{bmatrix} 0 \\ \sigma^2 \end{bmatrix} \right\} \overset{L}{\to} \mathcal{N}(0, \Sigma)$$

where the random vector $\begin{bmatrix} m_1(\bar{\mathbf{x}}) \\ m_2(\bar{\mathbf{x}}) \end{bmatrix}$ has variance matrix

$$\Sigma = \begin{bmatrix} \mathrm{var}\, \mathbf{x}_k & \mathrm{Cov}\,(\mathbf{x}_k^2, \mathbf{x}_k) \\ \mathrm{Cov}\,(\mathbf{x}_k^2, \mathbf{x}_k) & \mathrm{var}\; \mathbf{x}_k^2 \end{bmatrix} = \begin{bmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4 \end{bmatrix}$$

We want the asymptotic distribution of $g(m_1(\mathbf{x}), m_2(\mathbf{x}))$. Don't need to compute the third order moment μ_3 since the derivatives of g with respect to the two variables m_1, m_2 is $g'(m_1, m_2) = [-2m_1,\, 1]$ so that $g'(0, \sigma^2) = [0,\, 1]$ and hence

$$g'(0,\sigma^2)\Sigma\left[g'(0,\sigma^2)\right]^{\top} = \operatorname{var}\mathbf{x}_k^2 = \mathbb{E}\,\mathbf{x}_k^4 - (\mathbb{E}\,\mathbf{x}_k^2)^2 = \mu_4 - \sigma^4\,.$$

In conclusion:

$$\sqrt{N}\left[\hat{\boldsymbol{\sigma}}_N^2(\mathbf{x}) - g(0,\sigma^2)\right] = \sqrt{N}\left[\hat{\boldsymbol{\sigma}}_N^2(\mathbf{x}) - \sigma^2\right] \xrightarrow{L} \mathcal{N}(0, \mu_4 - \sigma^4)\,.$$

If we know that the pdf of $\mathbf{x}_k$ is Gaussian, $\mu_4 = 3\sigma^4$ and the limit distribution is $\mathcal{N}(0, 2\sigma^4)$.

GENERALIZATIONS OF THE CLT

Theorem 12 (Ferguson Problem 5, p.34, Lehmann Theorem 2.7.4 p.102)**.** *Let* $\{\mathbf{x}_1, \mathbf{x}_2, \dots \mathbf{x}_N\}$ *be a sequence of i.i.d random variables with mean* μ *and variance* σ^2 *and let* $w(N, k)\,;\, k = 1, 2, \dots, N$ *be a sequence of real numbers. Define*

$$\mathbf{y}_N := \sum_{k=1}^{N} w(N, k)\,\mathbf{x}_k\,, \qquad \sigma_N^2 := \mathrm{var}\,(\mathbf{y}_N) = \sigma^2 \sum_{k=1}^{N} w(N, k)^2\,.$$

Then :

$$\frac{\mathbf{y}_N - \mathbb{E}\,\mathbf{y}_N}{\sqrt{\sigma_N^2}} \xrightarrow{L} \mathcal{N}(0, 1)$$

if the following Lindeberg condition *holds: for* $N \to +\infty$

$$\max_{k \leq N} \frac{w(N, k)^2}{\sum_{k=1}^{N} w(N, k)^2} \to 0\,.$$

Note: you can restate the theorem assuming independent random variables defined as $\mathbf{z}_k = w(N, k)\mathbf{x}_k$ which have arbitrary mean $\mu_k = w(N, k)\mu$ and variance $\sigma_k^2 = w(N,k)^2\sigma^2$ then

$$\mathbf{y}_N := \frac{\sum_{k=1}^{N} (\mathbf{z}_k - \mu_k)}{\sqrt{\sigma_N^2}} \xrightarrow{L} \mathcal{N}(0, 1)\,.$$

Lehmann condition (2.7.3) is more complicated to verify

APPLICATION TO LINEAR REGRESSION

See Lehmann p. 101 and 104

Recall the regrssion model

$$\mathbf{y}_k = \alpha + \beta x_k + \mathbf{e}_k, \qquad k = 1, 2, \ldots, N$$

where the errors $\mathbf{e}_k$ are **zero-mean** i.i.d. with variances σ_k^2. The least squares estimates are

$$\hat{\boldsymbol{\alpha}}_N = (\bar{\mathbf{y}}_N - \hat{\boldsymbol{\beta}}_N \bar{x}_N)$$

$$\hat{\boldsymbol{\beta}}_N = \sum_{k=1}^{N} \frac{(x_k - \bar{x}_N)}{\sum_{k=1}^{N} (x_k - \bar{x}_N)^2} \mathbf{y}_k := \sum_{k=1}^{N} w_k \mathbf{y}_k$$

Note that the weights w_k actually depend on N. We shall rewrite them $w(N,k)$. The expression is similar to that in Theorem 12 (see the note below the statement). The variance σ_N^2 in this case is

$$\sigma_N^2 = \operatorname{var}(\hat{\boldsymbol{\beta}}_N) = \sum_{k=1}^{N} w^2(N,k) \operatorname{var}(\mathbf{y}_k) = \sum_{k=1}^{N} w^2(N,k)\, \sigma_k^2$$

where $\sigma_k^2 = \mathrm{var}\,(\mathbf{e}_k)$. For convergence in probability of $\hat{\boldsymbol{\beta}}_N$ to β we need

$$\lim_{N\to\infty} \sum_{k=1}^{N} w^2(N,k)\,\sigma_k^2 = 0$$

In case $\sigma_k^2 = \sigma^2$ independent of k the Lindeberg condition is equivalent to

$$\frac{w^2(N,k)}{\sum_{k=1}^{N} w^2(N,k)} = \frac{(x_k - \bar{x}_N)^2}{[\sum_{k=1}^{N}(x_k - \bar{x}_N)^2]^2} \sum_{k=1}^{N} (x_k - \bar{x}_N)^2 = \frac{(x_k - \bar{x}_N)^2}{\sum_{k=1}^{N}(x_k - \bar{x}_N)^2} \to 0\,.$$

Hence, under the same condition we also have

$$\sqrt{N}\,\frac{\hat{\boldsymbol{\beta}}_N - \beta}{\sigma_N} \xrightarrow{L} \mathcal{N}(0,1)\,.$$

THE ASYMPTOTIC DISTRIBUTION OF $\hat{\boldsymbol{\alpha}}_N$

See Lehmann Prob 7.11 p. 104

Since $\hat{\boldsymbol{\alpha}}_N = \alpha + \bar{x}_N (\beta - \hat{\boldsymbol{\beta}}_N) + \frac{1}{N}\sum_{k=1}^N \mathbf{e}_k$ we have

$$(\hat{\boldsymbol{\alpha}}_N - \alpha) = (\beta - \hat{\boldsymbol{\beta}}_N)\bar{x}_N + \bar{\mathbf{e}}_N$$

that is

$$\sqrt{N}\begin{bmatrix} 1 & \bar{x}_N \end{bmatrix}\left(\begin{bmatrix} \hat{\boldsymbol{\alpha}}_N \\ \hat{\boldsymbol{\beta}}_N \end{bmatrix} - \begin{bmatrix} \alpha \\ \beta \end{bmatrix}\right) = \sqrt{N}\,\bar{\mathbf{e}}_N$$

so that, the random variable say $\sqrt{N}\,\mathbf{z}_N$ in the first member has mean zero and variance $\tau_N^2 := \frac{1}{N}\sigma_N^2$ where $\sigma_N^2 = \sum_{k=1}^N \sigma_k^2 \to \infty$. Therefore under this condition,

$$\frac{\mathbf{z}_N}{\tau_N} \xrightarrow{L} \mathcal{N}(0,1)\,.$$

Need a careful analysis of the **vector** least squares problem.

THE CRAMÈR - RAO BOUND

Ferguson Chap 19

Let $\mathbf{x}$ be a r-dimensional random vector with $\mathbf{x} \sim \{F_\theta\,;\theta \in \Theta\}$ ($\mathbf{x}$ could in particular be a random sample as $(\mathbf{y}_1, \ldots, \mathbf{y}_N)$, but the Cramèr-Rao inequality does not require independence). Assume the following:

[A.1)] F_θ admits a density $p(\cdot, \theta)$ which is twice differentiable w.r.t. θ.

[A.2)] For every statistics ϕ with $\mathbb{E}_\theta \phi < \infty$,

$$\frac{\partial}{\partial \theta_i} \int_{\mathbb{R}^r} \phi(x)\, p(x,\theta)\, dx = \int_{\mathbb{R}^r} \phi(x)\, \frac{\partial}{\partial \theta_i}\, p(x,\theta)\, dx \quad \text{for } i = 1, \ldots, p\,;\ \forall\ \theta \in \Theta.$$

In particular,

$$\frac{\partial}{\partial \theta_i} \int_{\mathbb{R}^r} p(x,\theta)\, dx = \int_{\mathbb{R}^r} \frac{\partial}{\partial \theta_i}\, p(x,\theta)\, dx.$$

[A.3)] $\dfrac{\partial^2}{\partial \theta_i\, \partial \theta_j} \int_{\mathbb{R}^r} p(x,\theta)\, dx = \int_{\mathbb{R}^r} \dfrac{\partial^2}{\partial \theta_i\, \partial \theta_j}\, p(x,\theta)\, dx\,;\forall\, i,j = 1, \ldots, p\,;\forall\ \ \theta \in \Theta.$

Definition 5. *The* **Fisher Information Matrix** $I(\theta)$, *of the parametric family of densities* $\{p_\theta\}$ *is defined as*

$$I(\theta) := \left[\mathbb{E}_\theta \left(\frac{\partial \log p(\mathbf{x},\theta)}{\partial \theta_i} \cdot \frac{\partial \log p(\mathbf{x},\theta)}{\partial \theta_j}\right)\right]_{i,j=1,\dots,p} \tag{11}$$

$I(\theta)$ *can also be written as*

$$I(\theta) = \left[-\mathbb{E}_\theta \frac{\partial^2 \log p(\mathbf{x},\theta)}{\partial \theta_i \, \partial \theta_j}\right]_{i,j=1,\dots,p} \qquad . \tag{12}$$

That (12) and (11) are equivalent follows by differentiating the identity $\int p(x,\theta)\,dx = 1$ (constant with respect to θ) termwise with respect to θ getting

$$\int_{\mathbb{R}^r} \frac{\partial p(x,\theta)}{\partial \theta_i} dx = 0\,, \qquad \int_{\mathbb{R}^r} \frac{\partial^2 p(x,\theta)}{\partial \theta_i \, \partial \theta_j} dx = 0\,, i,j = 1,\dots,p\,.$$

Equation (12) then follows from

$$-\frac{\partial^2 \log p}{\partial \theta_i \, \partial \theta_j} = \frac{\partial \log p}{\partial \theta_i} \frac{\partial \log p}{\partial \theta_j} - \frac{1}{p} \frac{\partial^2 p}{\partial \theta_i \, \partial \theta_j} \quad ,$$

To understand the meaning of $I(\theta)$ we shall bring in the p-dimensional random vector of the **random sensitivities** of $p(\cdot,\theta)$ with respect to the parameter θ,

$$\mathbf{z}_\theta := \left[\frac{\partial \log p(\mathbf{x},\theta)}{\partial \theta_i}\right]_{i=1,\ldots,p} = \left[\frac{\partial\, p(\mathbf{x},\theta)}{\partial \theta_i} / p(\mathbf{x},\theta)\right]_{i=1,\ldots,p}$$

by which

$$I(\theta) = \mathbb{E}_\theta\, \mathbf{z}_\theta\, \mathbf{z}_\theta^\top\,, \tag{13}$$

where the matrix is (at least) positive semidefinite since it is a variance. In fact, it easily follows from [A.2)] that $\mathbb{E}_\theta \frac{\partial \log p}{\partial \theta_i} = 0$ for all i's and so

$$\mathbb{E}_\theta\, \mathbf{z}_\theta = 0\,.$$

Theorem 13 (The Cramèr-Rao Inequality)**.** *Let g be a differentiable function from Θ to $\mathbb{R}^q$ and ϕ be an unbiased estimator of $g(\theta)$. Let $V(\theta)$ be the variance matrix of ϕ and $G(\theta)$ the Jacobian matrix of g,*

$$G(\theta) = \left[\frac{\partial g_i(\theta)}{\partial \theta_j}\right]_{\substack{i=1,\ldots,q \\ j=1,\ldots,p}} \quad . \tag{14}$$

Then, if the Fisher matrix $I(\theta)$ is invertible, one has

$$V(\theta) - G(\theta)\, I^{-1}(\theta)\, G(\theta)^\top \geq 0 \quad , \tag{15}$$

where ≥ 0 means that the matrix on the left is positive semidefinite.

Proof: The proof is based on the classical formula for the error variance of the linear Bayesian estimator $\hat{\phi}(\mathbf{x}) := \mathbb{E}_\theta[\phi(\mathbf{x}) \mid \mathbf{z}_\theta]$ of the vector $\phi(\mathbf{x})$, given $\mathbf{z}_\theta$, that is

$$\mathrm{Var}_\theta\{\phi(\mathbf{x}) - \hat{\phi}(\mathbf{x})\} = \mathrm{Var}_\theta\{\phi(\mathbf{x})\} - \mathrm{Cov}_\theta\{\phi(\mathbf{x}), \mathbf{z}_\theta\}\mathrm{Var}_\theta\{\mathbf{z}_\theta\}^{-1}\mathrm{Cov}_\theta\{\phi(\mathbf{x}), \mathbf{z}_\theta\}^\top . \tag{16}$$

See for example [**?**, p. 27].

Since $\phi(\mathbf{x})$ is an unbiased estimator of $g(\theta)$; i.e.

$$\int_{\mathbb{R}^r} \phi(x)\, p(x,\theta)\, dx = g(\theta)\,, \forall\theta \in \Theta \quad ,$$

by applying property A.3) one gets

$$\mathbb{E}_{\theta}\, \phi(\mathbf{x})\, \mathbf{z}_{\theta}^{j} \;=\; \int_{\mathbb{R}^r} \phi(x)\, \frac{\partial p(x,\theta)}{\partial \theta_j} \cdot \frac{1}{p(x,\theta)} \cdot p(x,\theta)\, dx = \frac{\partial g(\theta)}{\partial \theta_j} \quad ,$$
$$j = 1,\dots,p \quad ,$$

and hence $\frac{\partial g(\theta)}{\partial \theta_j}$ is the j-th column of the covariance matrix of ϕ and $\mathbf{z}_\theta$,

$$\mathbb{E}_{\theta}\, \phi(\mathbf{x})\, \mathbf{z}_{\theta}^{\top} = \mathbb{E}_{\theta}\, \phi(\mathbf{x})\, [\mathbf{z}_{\theta}^{1},\dots,\mathbf{z}_{\theta}^{p}] \quad ,$$

that is,

$$\mathbb{E}_{\theta}\, \phi\, \mathbf{z}_{\theta}^{\top} = G(\theta) \quad . \tag{17}$$

The inequality follows since the variance of the random vector $\phi(\mathbf{x}) - G(\theta)\, I(\theta)^{-1}\, \mathbf{z}_\theta$ must be (at least) positive semidefinite.

Remarks

When ϕ is an **unbiased estimator** of θ (that is if g is the identity map) one has $G(\theta) = I \; (p \times p)$ and (15) becomes

$$V(\theta) - I(\theta)^{-1} \geq 0 \quad . \tag{18}$$

Since the scalar variance $\text{var}_{\theta}(\phi) = \sum_1^p \mathbb{E}_{\theta}(\phi_i - \theta_i)^2$ is the trace of $V(\theta)$ and

$$\text{Tr}\, V(\theta) - \text{Tr}\, I^{-1}(\theta) = \text{Tr}\left[V(\theta) - I^{-1}(\theta)\right] \geq 0$$

(the trace is the sum of the eigenvalues and the eigenvalues of a positive semidefinite matrix are all non-negative) it follows that *the scalar variance of any unbiased estimator of the parameter θ cannot be less than the positive number* $\text{Tr}\, I(\theta)^{-1}$, that is

$$\text{var}_{\theta}(\phi) \geq \text{Tr}\left[I(\theta)^{-1}\right], \quad \forall \theta \quad . \tag{19}$$

This lower bound only depends on the probabilistic model class $\{p(\cdot, \theta)\,;\, \theta \in \Theta\}$ and is *independent of which estimation criterion is used to construct* ϕ.

Remark:

One should however be aware of the fact that the Cramèr-Rao bound is just *one* possible bound for the variance which is not necessarily the tightest possible bound.There are in fact unbiased estimators whose variance is strictly larger than $\mathrm{Tr}\left[I(\theta)^{-1}\right]$ but nevertheless have minimum variance.

Example 8. Let $\mathbf{y} \sim \mathcal{N}(\theta, \sigma^2)$ be a scalar random variable with a known variance σ^2. Since

$$\log p(y,\theta) = C - \frac{1}{2}\frac{(y-\theta)^2}{\sigma^2} \quad ,$$
$$\frac{d}{d\theta}\log p(y,\theta) = \frac{y-\theta}{\sigma^2}$$

we have

$$i(\theta) = \mathbb{E}_\theta \left(\frac{\mathbf{y}-\theta}{\sigma^2}\right)^2 = \frac{1}{\sigma^4}\cdot\sigma^2 = 1/\sigma^2 \quad .$$

Hence the variance of any unbiased estimator of θ based on a sample of size one, cannot be smaller that the variance of $\mathbf{y}$. Assume now we have a random sample of size N from the same Gaussian distribution. Now we have a random vector $\mathbf{x} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$ of dimension $r = N$ and

$$p(y_1, \ldots, y_N \mid \theta) = \prod_{t=1}^{N} p(y_t, \theta)$$

and hence

$$\begin{aligned} \log p(y_1,\ldots,y_N \mid \theta) &= N \times Const - \frac{1}{2}\sum_{t=1}^{N}\frac{(y_t-\theta)^2}{\sigma^2} \quad , \\ \frac{d\log p}{d\theta} &= \sum_{t=1}^{N}\frac{y_t-\theta}{\sigma^2} \quad . \end{aligned}$$

Since the random variables $\mathbf{y}_1,\ldots,\mathbf{y}_N$ are independent, it follows that,

$$I(\theta) = \mathbb{E}_\theta\left[\frac{d\log p(\mathbf{y},\,\theta)}{d\theta}\right]^2 = \frac{1}{\sigma^4}\cdot N\sigma^2 = \frac{N}{\sigma^2} \quad .$$

Let us consider the sample mean $\bar{\mathbf{y}}_N = \frac{1}{N}\sum_{t=1}^{N}\mathbf{y}_t$ which has distribution $\mathscr{N}(\theta,\sigma^2/N)$. Since $\bar{\mathbf{y}}_N$ is an unbiased estimator of θ with variance σ^2/N, exactly equal to the inverse of the Fisher information, the sample mean is the *best possible estimator of* θ (of course if the sample distribution is Gaussian). One says that an unbiased estimator whose variance is exactly equal to the inverse of the Fisher information matrix, $V(\theta) = I(\theta)^{-1}$ is *efficient*. $\diamondsuit$

Example 9. Let $\mathbf{y} \sim \mathscr{N}(\mu, \theta^2)$, where μ is known and $(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ is a random sample from $\mathscr{N}(\mu, \theta^2)$. Consider the unbiased estimator

$$\tilde{\boldsymbol{\sigma}}_N^2 = \frac{N\hat{\boldsymbol{\sigma}}_N^2}{N-1} = \frac{1}{N-1} \sum_{k=1}^{N} (\mathbf{y}_k - \bar{\mathbf{y}}_N)^2 \quad ;$$

We know that $\frac{N\hat{\boldsymbol{\sigma}}_N^2}{\theta^2}$ has a *chi square* distribution with $N-1$ degrees of freedom, which has expectation $N-1$ and variance $2(N-1)$.

Its variance is then $\frac{2\theta^4}{N-1}$. The Cramèr-Rao bound in this case is $2\theta^4/N$ and hence the variance of $\tilde{\boldsymbol{\sigma}}_N^2$ is *strictly greater than* $I(\theta)^{-1}$. One can however show [**?**] that any unbiased estimator of θ^2 cannot have a smaller variance than that of $\tilde{\boldsymbol{\sigma}}_N^2$.

From this example it follows that $I(\theta)^{-1}$ is not the best possible lower bound. ♢

Note that a **biased** *estimator can have smaller variance than the C-R limit. From Proposition 9 it follows that* $\operatorname{var} \hat{\boldsymbol{\sigma}}_N^2 = 2\sigma^4 \frac{N-1}{N^2}$*.*

Exercises

1-1 Let $I(\theta)$ be the Fisher matrix relative to an arbitrary density $p(\mathbf{y},\theta)$. Show that for a random sample of size N one has $I_N(\theta) = N I(\theta)$.

1-2 Show, without using the χ^2 distribution, that the Cramèr-Rao bound for a random sample from $\mathscr{N}(\mu,\theta^2)$ of size N is $2\theta^4/N$.

1-3 Show that the Cramèr-Rao bound for $\mathscr{N}(\theta_1,\theta_2^2)$ (two dimensional parameter θ) is

$$I(\theta)^{-1} = \begin{bmatrix} \theta_2^2/N & 0 \\ 0 & 2\theta_2^4/N \end{bmatrix} \quad .$$

THE KULLBACK-LEIBLER DISTANCE

See Ferguson p112

In this section we shall define a measure of deviation of two random variables $\mathbf{x}_1 \sim p(\cdot, \theta_1)$ and $\mathbf{x}_2 \sim p(\cdot, \theta_2)$ described by the same parametric family of distributions.

We shall use this measure to quantify in rather precise terms, the ability of observations extracted from the model, to *discriminate* between different values of the parameter θ.

Definition 6. *Let f and p be probability densities such that $p(x) = 0 \Rightarrow f(x) = 0$. The* Kullback-Leibler (pseudo-)distance *between f and p, is*

$$K(f,p) := \int_{\mathbb{R}^r} [\log f - \log p] \, f(x) \, dx = \int_{\mathbb{R}^r} \log f/p \, f(x) \, dx = \mathbb{E}_f \log f/p \, ; \quad (20)$$

It is immediate that $K(f,p) = 0$ if and only if $f = p$.

From Jensen inequality:

$$\int \log g(x)\, d\mu \le \log\{\int g(x)\, d\mu\}$$

which holds for $g(x) > 0$ and an arbitrary positive measure μ, one gets

$$-K(f,p) = \int_{\mathbb{R}^r} \log \frac{p}{f} f\, dx \le \log\{\int_{\mathbb{R}^r} \frac{p}{f} f\, dx\} = \log\{1\} = 0$$

so that $K(f,p) \ge 0$.

For this reason $K(f,p)$ can be interpreted as a measure of deviation of the probability density p from a “reference” density f.
Note in fact that $K(f,p)$ is not symmetric; i.e. $K(p,f) \neq K(f,p)$ and does not satisfy the triangle inequality. In Information Theory $K(f,p)$ is called *divergence* and is denoted by the symbol $D(f\|p)$ (here p is the approximation of f). The article in Wikipedia on *Kullback-Leibler divergence* provides a rather complete overview and a bibliography.

Let us assume that the family $p(\cdot,\theta)$ satisfies the same regularity assumptions listed for the Cramèr-Rao bound and let $f \equiv p(\cdot,\theta_0)$ and $p \equiv p(\cdot,\theta)$, $\theta_0,\theta \in \Theta$. Denoting $K(p(\cdot,\theta_0),p(\cdot,\theta))$ by $K(\theta_0,\theta)$ and letting $\theta = \theta_0 + \Delta\theta$, one has

$$K(\theta_0,\theta) = K(\theta_0,\theta_0) + \left.\frac{\partial K}{\partial \theta}\right|_{\theta_0} \Delta\theta + \frac{1}{2}\Delta\theta^\top \left[\frac{\partial^2 K}{\partial \theta_i\, \partial \theta_j}\right]_{\theta_0} \Delta\theta + o(\|\Delta\theta\|^2)\,.$$

Since $K(\theta_0,\theta_0) = 0$ and

$$\frac{\partial K}{\partial \theta_i} = -\int_{\mathbb{R}^r} p(x,\theta_0)\,\frac{\partial \log p(x,\theta)}{\partial \theta_i}\,dx \quad ,$$

it follows that

$$\left.\frac{\partial K}{\partial \theta_i}\right|_{\theta_0} = -\int_{\mathbb{R}^r} \left[\frac{\partial p(x,\theta)}{\partial \theta_i}\right]_{\theta_0} dx = 0$$

for all $i = 1,\ldots,p$.

In the same way one can verify that

$$\left.\frac{\partial^2 K}{\partial \theta_i\, \partial \theta_j}\right|_{\theta_0} = -\int_{\mathbb{R}^r} p(x,\theta_0)\ \left[\frac{\partial^2 \log p(x,\theta)}{\partial \theta_i\, \partial \theta_j}\right]_{\theta_0} dx = -\mathbb{E}_{\,\theta_0}\ \left[\frac{\partial^2 \log p(x,\theta)}{\partial \theta_i\, \partial \theta_j}\right]_{\theta_0}$$

and hence the first member of this equality is the (i, j)-th element of the Fisher matrix $I(\theta_0)$. Hence, for small variation of the parameter θ, it holds

$$K(\theta_0, \theta) \cong \frac{1}{2} \Delta\theta^\top \, I(\theta_0)\, \Delta\theta \quad ; \tag{21}$$

which says that, for small deviations $\Delta\theta$ of the parameter from the reference value θ_0, **the Kullback-Leibler distance between $p(\cdot, \theta)$ and $p(\cdot, \theta_0)$ is a quadratic form whose weighting matrix is the Fisher matrix $I(\theta_0)$**. In the next section we will see a remarkable consequence of this fact.

Exercise Compute the Kullback-Leibler distance between the two Gaussian densities, $f \equiv \mathscr{N}(\mu\,, \sigma_0^2)$ and $p \equiv \mathscr{N}(\mu\,, \sigma^2)$. Check what happens if you invert the order of the two densities.

IDENTIFIABILITY

See Ferguson p112

The observations may be structurally incapable of providing enough information to uniquely locate the value of the parameter θ which has generated them. A rather trivial example :

Let θ be a two-dimensional parameter $[\theta_1, \theta_2]^\top$, ranging on $\Theta = \mathbb{R}^2$ and let F_θ depend on (θ_1, θ_2) only through their product $\theta_1\theta_2$; for example $F_\theta \sim \mathcal{N}(\theta_1\theta_2, \sigma^2)$.

For any fixed value $\bar{\theta} = (\bar{\theta}_1, \bar{\theta}_2)^\top$, the parameters $\hat{\theta} = \left(\alpha\bar{\theta}_1, \frac{1}{\alpha}\bar{\theta}_2\right)^\top$, $\alpha \neq 0$, define the same PDF; that is $F_{\bar{\theta}}(x) = F_{\hat{\theta}}(x)$, $\forall x$. Hence a sample observation extracted from this family, irrespective of its size N, will never be able to distinguish between $\bar{\theta}$ and $\hat{\theta}$.

Definition 7. *Two parameter values θ' and $\theta^{''}$ in Θ are said to be* indistinguishable *if $F_{\theta_1}(x) = F_{\theta_2}(x)$, $\forall x \in \mathbb{R}^r$. Notation: $\theta' \simeq \theta^{''}$.*

The family of PDF's $\{F_\theta\,;\, \theta \in \Theta\}$ (sometimes one says improperly that the parameter *$\theta \in \Theta$) is* globally identifiable *if $\theta' \simeq \theta''$, or, equivalently, $F_{\theta'} = F_{\theta''}$, implies that $\theta' = \theta''$ for all θ', θ'' in Θ.*

For many applications global identifiability is too restrictive. A weaker condition is the local notion.

Definition 8. *The family of PDF's $\{F_\theta\,;\,\theta \in \Theta\}$ is* locally identifiable about *θ_0 if there exists an open neighborhood of θ_0 which does not contain parameter values θ which are indistinguishable from θ_0 (of course, except θ_0 itself).*

This concept is often overlooked. There is a remarkable relation between (local) identifiability and nonsingularity of the Fisher matrix.

Theorem 14 (Rothenberg)**.** *Let the parametric model $\{p_\theta\,;\,\theta \in \Theta\}$ satisfy the assumptions A.1, A.2, A.3. Then θ_0 is locally identifiable if and only if $I(\theta_0)$ is non-singular.*

Proof. Is based on the properties of the Kullback-Leibler (pseudo)-metrics which guarantees that $K(\theta_0,\theta) = 0 \Leftrightarrow p(\cdot,\theta_0) = p(\cdot,\theta)$. For small deviations $\Delta\theta$ of the parameter θ about the reference value θ_0, the Kullback-Leibler distance between the two densities $p(\cdot,\theta)$ and $p(\cdot,\theta_0)$ is the quadratic form $\frac{1}{2}\Delta\theta^\top\, I(\theta_0)\,\Delta\theta$. It follows that in any small enough neighborhhod of θ_0 one can have parameter values $\theta \neq \theta_0$ for which $p(\cdot,\theta) = p(\cdot,\theta_0)$ if and only if $I(\theta_0)$ is singular. □

In the previous trivial example one has

$$I(\theta) = \mathbb{E}_{\theta} \begin{bmatrix} \frac{(\mathbf{x} - \theta_1\theta_2)^2}{\sigma^4}\theta_2^2 & \frac{(\mathbf{x} - \theta_1\theta_2)^2}{\sigma^4}\theta_1\theta_2 \\ \frac{(\mathbf{x} - \theta_1\theta_2)^2}{\sigma^4}\theta_1\theta_2 & \frac{(\mathbf{x} - \theta_1\theta_2)^2}{\sigma^4}\theta_1^2 \end{bmatrix} = \frac{1}{\sigma^2} \begin{bmatrix} \theta_2^2 & \theta_1\theta_2 \\ \theta_1\theta_2 & \theta_1^2 \end{bmatrix}.$$

one sees that $\det I(\theta) = 0$, $\forall\theta \in \mathbb{R}^2$ and hence the model is never locally identifiable about any arbitrary parameter value θ. In fact, the model is globally unidentifiable as all indistinguishability classes contain infinitely many parameter values.

ERGODICITY AND THE STRONG LAW OF LARGE NUMBERS

Lehmann p. 62

Let us pretend that we have an infinite sequence of observations indexed by time, extending from $t = -\infty$ to the infinite future $t = +\infty$. This is called a **stochastic process** denoted

$$\mathbf{y} = \{\mathbf{y}(t)\}, \qquad t \in \mathbb{Z}$$

the symbol $\mathbb{Z}$ (Zhalen in German) stands for integer numbers.

Definition 9. *A stochastic process $\{\mathbf{y}(t)\}$ is* **stationary** *(in the strict sense) if all Pdf's relative to $\mathbf{y}(t_1), \mathbf{y}(t_2), \ldots \mathbf{y}(t_n)$ say $F_n(x_1, \ldots, x_n, t_1, \ldots, t_n)$ are invariant for temporal translation, that is for every n it must hold that,*

$$F_n(x_1, \ldots, x_n, t_1 + \Delta, \ldots, t_n + \Delta) = F_n(x_1, \ldots, x_n, t_1, \ldots, t_n) \quad ,$$

(same function of $x_1, \ldots, x_n$, $t_1, \ldots, t_n$), whatever the time shift $\Delta \in \mathbb{Z}$.

Consequences

- The Pdf $F(x,t)$ of any variable $\mathbf{y}(t)$ does not depend on t; that is the random variables $\mathbf{y}(t)$, $t \in \mathbb{Z}$, are *identically distributed*;

- The *second order* joint Pdf $F_2(x_1,x_2,t_1,t_2)$ of the variables$\mathbf{y}(t_1)$, $\mathbf{y}(t_2)$, only depends on $\tau = t_1 - t_2$ and not on the date.
  In particulao, $\mu(t) := E\,\mathbf{y}(t)$,ia constant equal to $\mu \in \mathbb{R}^m$ and the *Covariance*

  $$\Sigma(t_1,t_2) := E\,[\mathbf{y}(t_1) - \mu(t_1)]\;[\mathbf{y}(t_2) - \mu(t_2)]^\top$$

  depends only on $\tau = t_1 - t_2$.

THE ERGODIC THEOREM

Let $f(\mathbf{y})$ denote a statistic, function of any number of random variables of the process, *which does not depend on time*. Denote by $f_k(\mathbf{y})$ the same function in which all time indices of these variables are shifted by k units.

Theorem 15 (Birkhoff Ergodic Theorem)**.** *Let $\{\mathbf{y}(t)\}$ be a strictly stationary process. The limit*

$$\bar{\mathbf{z}} := \lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} f_k(\mathbf{y}) \tag{22}$$

exists with probability one for all functions f such that $\mathbb{E}\,|f(\mathbf{y})| < \infty$

The limit can either be random or constant. If it is random it must be a “very special” random variable. These are called *invariant random variables*. We shall not investigate them.

If the limit is a constant then the process is called **Ergodic.**

Note now that

$$\mathbb{E}\{\frac{1}{T}\sum_{k=1}^{T} f_k(\mathbf{y})\} = \frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\, f_k(\mathbf{y}) = \mathbb{E}\, f(\mathbf{y})$$

since $\mathbf{z}(k) = f_k(\mathbf{y})$ is itself a strictly stationary process. Take expectation in both members of (22). For $T \to \infty$ one finds

$$\mathbb{E}\,\bar{\mathbf{z}} = \mathbb{E}\, f(\mathbf{y})\,.$$

Corollary 3. *If $\{\mathbf{y}(t)\}$ is ergodic*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} f_k(\mathbf{y}) = \mathbb{E}\, f(\mathbf{y}) \tag{23}$$

with probability one whatever may be $f(\mathbf{y})$ having finite expectation.

Proof: In fact $\bar{\mathbf{z}}$ must be a constant and hence coincides with its own expectation $\bar{\mathbf{z}} = \mathbb{E}\,\bar{\mathbf{z}} = \mathbb{E}\, f(\mathbf{y})$. □

THE STRONG LAW OF LARGE NUMBERS

Theorem 16 (Kolmogorov)**.** *Every i.i.d. process having finite expectation is ergodic.*

The following is an important consequence.
Corollary 4. *Let* $\mathbf{y}$ *be an ergodic process and* $\mathbf{z}(t) := f_t(\mathbf{y})$ *the sequence of translates of an arbitrary function of the process, having finite expectation. The* $\{\mathbf{z}(t)\}$ *is stationary and ergodic. In particular the translates of every time-invariant function of an i.i.d. process form an ergodic process.*

For example if $\mathbf{e}$ is i.i.d. of finite variance and $\sum_{-\infty}^{+\infty} |c_k| < \infty$, the translated random variables

$$\mathbf{z}(t) := \sum_{-\infty}^{+\infty} c_k \, \mathbf{e}(t+k) \; , = \sum_{-\infty}^{+\infty} c_{-k} \, \mathbf{e}(t-k) \, ; \qquad \text{CONV}$$

form an ergodic process. In fact by Cauchy-Schwartz inequality

$$\mathbb{E} \, | \sum_{-N}^{+N} c_k \, \mathbf{e}(t+k) | \leq \sum_{-N}^{+N} |c_k|^2 \, \mathbb{E} \, |\mathbf{e}(t+k)|^2 = \sum_{-N}^{+N} |c_k|^2 \, \sigma_{\mathbf{e}}^2$$

Proposition 12. *An ergodic process cannot admit limit for* $t \to \pm\infty$ *unless it reduces to a deterministic sequence (with probability 1).*

In fact such a limit should be a constant random variable.

STRONG CONSISTENCY OF ML

Assume $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ is an i.i.d. sample from a family of pdf's $\{\, p_\theta(x)\,;\, \theta \in \Theta\}$ with $\Theta \subset \mathbb{R}^p$. After performing an experiment you observe a sequence of sample values $x := (x_1, x_2, \ldots, x_N\,)$. The *likelihood function* of θ corresponding to these sample values is the function

$$L_N(\theta) = L_N(\theta \mid x) = \prod_{k=1}^{N} p_\theta(x_k)\,.$$

The function $l_N(\theta) := \log L_N(\theta)$ is called the *log-likelihood*. A *maximum likelihood estimate (MLE)* of θ is any function $\hat{\theta}_N(x)$ such that

$$L_N(\hat{\theta}_N(x)) = \sup_{\theta \in \Theta} L_N(\theta \mid x)$$

or equivalently $l_N(\hat{\theta}_N(x)) = \sup_{\theta \in \Theta} l_N(\theta \mid x)$. This supremum (which by definition always exists) may be $+\infty$ for all x and the (MLE) as a function of x may not exist. It certainly exists if Θ is a compact set and $L_N(\theta)$ is continuous (minimum requirement: upper semicontinuous).

Suppose the sample is generated by an unknown *true value* θ_0 of the parameter and assume the model is locally identifiable about θ_0 (in practice need to check this condition for all θ) then the KL distance

$$K(\theta_0, \theta) := \int_{\mathbb{R}} \log \frac{p_{\theta_0}(x_k)}{p_\theta(x_k)} p_{\theta_0}(x_k)\, dx_k = \mathbb{E}_{\theta_0} \log \frac{p_{\theta_0}(\mathbf{x}_k)}{p_\theta(\mathbf{x}_k)}$$

is positive (independent of k) and can be zero only if $\theta = \theta_0$.
Denote the log-likelihood by $\mathbf{l}_N(\theta) = \mathbf{l}_N(\theta \mid \mathbf{x})$ then

$$\mathbf{l}_N(\theta) - \mathbf{l}_N(\theta_0) = \sum_{k=1}^{N} (\log p_\theta(\mathbf{x}_k) - \log p_{\theta_0}(\mathbf{x}_k)) = \sum_{k=1}^{N} \log \frac{p_\theta(\mathbf{x}_k)}{p_{\theta_0}(\mathbf{x}_k)}$$

By the law of large numbers, for every $\theta \in \Theta$,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \log \frac{p_\theta(\mathbf{x}_k)}{p_{\theta_0}(\mathbf{x}_k)} = \mathbb{E}_{\theta_0} \log \frac{p_\theta(\mathbf{x}_k)}{p_{\theta_0}(\mathbf{x}_k)} = -K(\theta_0, \theta) < 0\,,$$

with probability 1.

Therefore for $N \to \infty$, $\frac{1}{N}\left(\mathbf{l}_N(\theta) - \mathbf{l}_N(\theta_0)\right) \overset{a.s.}{\to} -K(\theta_0, \theta) < 0$ this means that for N large, $\mathbf{l}_N(\theta) < \mathbf{l}_N(\theta_0)$ for all $\theta \neq \theta_0$ so, taking exponentials, the likelihood of θ_0 will be larger than that of any other $\theta \in \Theta$. In fact we have asymptotic exponential decay:

$$\frac{L_N(\theta)}{L_N(\theta_0)} = O(e^{-NK(\theta_0, \theta)}); \qquad \forall\, \theta \in \Theta$$

which means that the ratio will certainly be < 1 for N large. When Θ is a finite set this implies that any maximum, $\hat{\theta}_N$, of $L_N(\theta)$ must converge to θ_0.

Theorem 17. *If Θ is compact, $p_\theta(x)$ is continuous in θ for all x and there is $K(x)$ such that*

$$\log \frac{p_\theta(x)}{p_{\theta_0}(x)} \leq K(x); \qquad \mathbb{E}_{\theta_0} K(\mathbf{x}_k) < \infty \tag{24}$$

then any maximizing $\hat{\theta}_N(x)$ converges almost surely to θ_0 as $N \to \infty$.

ASYMPTOTIC NORMALITY OF MLE

Theorem 18 (Cramèr)**.** *Assume again that* $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ *is an i.i.d. sample from a family of pdf's* $\{\, p_\theta(x)\, ;\, \theta \in \Theta\}$ *with* $\Theta \subset \mathbb{R}^p$ *now an open subset, satisfying the three conditions [A.1)][A.2)][A.3)] of the Cramèr-Rao Theorem. Let* θ_0 *be the true value of the parameter and assume that the information matrix* $I(\theta)$ *is non singular at* $\theta = \theta_0$*. In other words*

$$I(\theta_0) > 0$$

(which implies local identifiability). Assume further that condition (24) *is satisfied. Then there exists a strongly consistent sequence* $\hat{\boldsymbol{\theta}}_N$ *of roots of the likelihood equation*

$$\frac{\partial \mathbf{l}_N(\theta)}{\partial \theta_i} = 0\,; \qquad i = 1, 2, \ldots, p$$

such that

$$\sqrt{N}\,(\hat{\boldsymbol{\theta}}_N - \theta_0) \overset{L}{\to} \mathcal{N}(0, I(\theta_0)^{-1})$$

ASYMPTOTIC NORMALITY OF MLE 2

Lehmann book pp.459-476 does only the scalar case of 1 parameter

Somewhat confusing. Will see a more general CLT statement later

ASYMPTOTICS OF STATISTICAL TESTS

See Lehmann p. 133-4, and 137

Example 10. Consider an i.i.d sample with finite fourth order moment μ_4. We want to test the hypothesis

$$H_0 : \sigma^2 = \sigma_0^2 \,,$$

against the alternative $\sigma^2 > \sigma_0^2$ based on $N\,(\to \infty)$ independent observations. We reject the hypothesis H_0 if $\hat{\boldsymbol{\sigma}}_N^2$ is large. The asymptotic distribution

$$\sqrt{N}\,(\hat{\boldsymbol{\sigma}}_N^2(\mathbf{x}) - \sigma^2) \xrightarrow{L} \mathcal{N}(0, \mu_4 - \sigma^4)\,.$$

has mean σ^2 and variance $\mu_4 - \sigma^4$ which in case of near normality is $2\sigma^4 - \sigma^4 = \sigma^4$ and the rejection region becomes

$$\sqrt{N}\left(\hat{\boldsymbol{\sigma}}_N^2 - \sigma_0^2\right) \geq u_\alpha \sqrt{2}\,\sigma_0^2$$

does not depend on the parameter μ of the parent distribution.

TIME SERIES

Suppose you have a sequence of scalar regression data

$$y^N := \{y(t)\,;\, t = 1,2,\ldots,N\}\,, \qquad u^N := \{u(t)\,;\, t = 1,2,\ldots,N\}$$

where you measure the $u(t)$'s exactly but the $y(t)$'s are random due to errors of various kinds. We shall imagine that they are extracted from two stochastic processes $\{\mathbf{y}(t),\ \mathbf{u}(t)\,\}$which are jointly stationary zero-mean and have finite second order joint moments.

There is **serial correlation** among successive sample values so that $\mathbf{y}(t)$ is correlated with its past values $\mathbf{y}(t-1), \mathbf{y}(t-2), \mathbf{y}(t-3), \ldots$. We don't care much about $\mathbf{u}$ since it is going to be observed exactly. The simplest generalization of regression models to describe serial correlation is

$$\mathbf{y}(t) + \sum_{k=1}^{n} a_k \mathbf{y}(t-k) = \sum_{k=1}^{m} b_k \mathbf{u}(t-k) + \mathbf{e}(t)\,, \tag{ARX}$$

where $\mathbf{e} := \{\mathbf{e}(t), t \in \mathbb{Z}\}$ is a process of random errors.

In econometric applications the variable $\mathbf{u}$ is an external forcing term called *exogenous variable*. Sometimes you want to describe how $\mathbf{y}$ changes in time as a consequences of time-varying exogenous variables. If there is no $\mathbf{u}$ then the model is called (purely) **Auto-Regressive** and is denoted by the acronym AR.

$$\mathbf{y}(t) + \sum_{k=1}^{n} a_k \mathbf{y}(t-k) = \mathbf{e}(t)\,, \tag{AR}$$

There are also more complicated models: ARMA, ARMAX, GARCH etc..

We shall assume that $\mathbf{e}$ is an i.i.d. process. The model ARX depends on $p := n+m$ unknown parameters written as a **column vector**:

$$\theta := \begin{bmatrix} a_1 & \dots & a_n & b_1 & \dots & b_m \end{bmatrix}^\top$$

and can be written in regression form as

$$\mathbf{y}(t) = \boldsymbol{\varphi}(t)^\top \theta + \mathbf{e}(t)\,.$$

where

$$\boldsymbol{\varphi}(t)^\top = \begin{bmatrix} -\mathbf{y}(t-1) & \dots & -\mathbf{y}(t-n) & \mathbf{u}(t-1) & \dots & \mathbf{u}(t-m) \end{bmatrix}$$

(so $\boldsymbol{\varphi}(t)$ is a column vector). The function of the past data

$$\hat{\mathbf{y}}_\theta(t \mid t-1) = \boldsymbol{\varphi}(t)^\top \theta$$

is the (one step ahead) **predictor function** associated to the model. Note that the predictor function is a linear function of θ and of the previous $n+m$ past samples of the joint process.

PEM IDENTIFICATION OF TIME SERIES

We want to estimate the parameter θ from observed data (y^N, u^N).

Often we do not know the probability distribution of the error process. We may assume it is Gaussian but we shall see that this assumption is not so useful. We try to do by just assuming that $\mathbf{e}$ is an i.i.d. process.

We shall use the **Prediction Error Minimization (PEM)** approach. This is a variant of Empirical Risk Minimization (ERM) in Machine Learning.

Given the model (ARX) one does as follows:

1. For a generic value of θ, construct a **predictor** based on data up to time $t-1$ of the next output, $y(t)$. For each θ the predictor is a deterministic function of the past data denoted $\hat{y}_\theta(t \mid t-1)$. For analysis purpose we may consider $\hat{y}_\theta(t \mid t-1)$ as a function (of θ) and of the past **random** observed data denoted $\hat{\mathbf{y}}_\theta(t \mid t-1)$.

2. Form the *empirical prediction errors* incurred by using θ:

$$\varepsilon_\theta(t) := y(t) - \hat{y}_\theta(t \mid t-1); \qquad t = 1, 2, \ldots, N$$

these are *numbers* but may also be interpreted as sample values of a random variables, written $\boldsymbol{\varepsilon}_\theta(t)$.

3. Minimize with respect to θ the **empirical average prediction error**

$$V_N(\theta) := \frac{1}{N} \sum_{t=1}^{N} \varepsilon_\theta(t)^2$$

More generally may introduce a *discount factor* for past errors: a positive sequence $\beta(N,t)$,

$$V_N(\theta) := \frac{1}{N}\sum_{t=1}^{N} \beta(N,t)\varepsilon_\theta(t)^2 \qquad \beta(t,N) > 0$$

For small N, the function β gives small weight to errors incurred at the beginning. One designs the weighting function so that For $N \to \infty$, $\beta(N,t) \to 1$

.

The parameter estimate

$$\hat{\theta}_N := \text{Arg}\min_{\theta} V_N(\theta)$$

becomes a function of the data (y^N, u^N). For the simple ARX model it can be computed explicitly.

Define as an estimate of $\lambda^2 = \text{var}\{\mathbf{e}(t)\}$, the *residual quadratic error*,

$$\hat{\lambda}_N^2 := V_N(\hat{\theta}_N)$$

where V_N is defined above.

ARX MODEL ESTIMATION

Assume we have data $\{y(t)\,,u(t)\,;t=t_0,t_0-1,\ldots,0,1,2,\ldots,N\}$ to be described by an (ARX) model using the PEM method. Write a N-vector model for all data as

$$\mathbf{y}=\Phi_N\,\theta+\mathbf{e}\qquad\text{column vectors of dimension } N$$

where $\mathbf{y}$ and $\mathbf{e}$ have components $\mathbf{y}(t)$ and $\mathbf{e}(t)$ for $t=1,2,\ldots,N$ and Φ_N is an $N\times p$ matrix of past data:

$$\Phi_N:=\begin{bmatrix}\varphi(1)^\top\\ \vdots\\ \varphi(N)^\top\end{bmatrix},$$

Assuming the initial time t_0 is far enough, we can fill in Φ_N with data from time say $t=1$. The N-dimensional vector of predictors and prediction errors are $\hat{\mathbf{y}}_\theta=\Phi_N\,\theta\,,\qquad \boldsymbol{\varepsilon}_\theta=\mathbf{y}-\Phi_N\,\theta$. Then $V_N(\theta)$ is the squared Euclidean norm of $\boldsymbol{\varepsilon}_\theta$,

$$V_N(\theta)=\frac{1}{N}\sum_{t=1}^{N}[y(t)-\varphi(t)^\top\theta\,]^2=\frac{1}{N}\|\mathbf{y}-\Phi_N\,\theta\|^2\,.$$

DETERMINISTIC VECTOR LEAST SQUARES

Hence the estimation leads to a vector **Least Squares Regression Problem**. More generally to *weighted Least squares*: if the t-th measurement is more reliable weight the t-th prediction errors by larger weight q_t^2. If the error variances are approximately known, the optimal choice is take $q_t^2 = \dfrac{1}{\mathrm{var}(\mathbf{e}(t))}$ the inverse of the error variance.

Linear Regression: Given a deterministic linear model class $\{\hat{y}_\theta = X\theta\,,\ \theta \in \mathbb{R}^p\}$ where X is a known $N \times p$ matrix. Choose the unknown p-dimensional parameter θ in such a way to describe in the best way a vector of observed data $y \in \mathbb{R}^N$.
Let $Q = \mathrm{diag}\{q_1^2, \ldots, q_N^2\}$ be a matrix of weights. Problem is to minimize with respect to θ the quadratic form

$$V_Q(\theta) = [y - X\theta]^\top Q\,[y - X\theta] := \|y - X\theta\|_Q^2\,.$$

(forget $\frac{1}{N}$) where $Q = Q^\top$ is a positive definite weight matrix. The minimization can be done by elementary calculus. However it is more instructive to do this by geometric means.

GEOMETRY OF VECTOR LEAST SQUARES

Make $\mathbb{R}^N$ into an inner product space by introducing the inner product

$$\langle x, y\rangle_Q = x^\top Q y = \sum_{k=1}^{N} x_k q_k^2 y_k$$

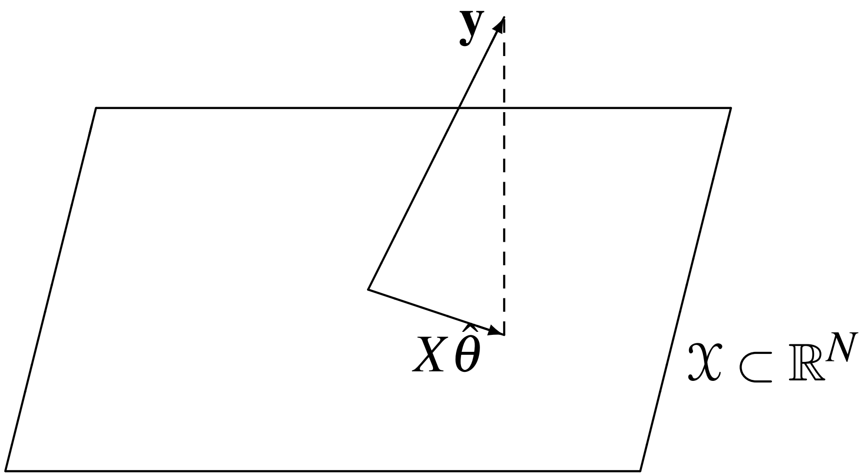
and let the corresponding norm be denoted by $\|x\|_Q^2 := x^\top Q x$. Note that for $Q = I$ the norm $\|x\|_Q^2$ is just $\sum x_k^2$.

Let $\mathcal{X}$ be the linear subspace of $\mathbb{R}^N$ linearly spanned by the columns of the matrix $X = \begin{bmatrix} x_1 & \dots & x_p \end{bmatrix}$ that is

$$\mathcal{X} := \{\sum_{k=1}^{p} \theta_k x_k,\ \theta_1, \dots, \theta_p \,\text{arbitrary real numbers}\}$$

Then the minimization of $\|y - X\theta\|_Q^2$ is just the minimum distance problem of finding the vector $\hat{y} \in \mathcal{X}$ **of shortest distance from the data vector** y according to the distance defined by the norm $\|\cdot\|_Q$.

THE GEOMETRY OF LEAST SQUARES



$X\hat{\theta}$ is the **Orthogonal Projection** of $y \in \mathbb{R}^N$ onto $\mathcal{X}$.

THE ORTHOGONALITY PRINCIPLE

Want to find the vector $\hat{y} \in \mathcal{X}$ **of shortest distance from the data vector** y.

Theorem 19. *The minimizer of* $V_Q(\theta) = \|y - X\theta\|_Q^2$ *must make the error* $y - X\theta$ **orthogonal (according to the scalar product** $\langle x, y \rangle_Q$**) to the subspace** $\mathcal{X}$, *or, equivalently, to the columns of* X, *that is*

$$X^\top Q(y - X\theta) = 0\,,$$

Equivalently the optimal θ *must solve the* Normal Equations

$$X^\top Q\,X\,\theta = X^\top Q y\,.$$

The **normal equations** of Least-Squares for $Q = I_N$:

$$X^\top X\,\theta = X^\top y\,.$$

To solve need invertibility of $X^\top X$.

Let us now assume that

$$\operatorname{rank} X = p \leq N\,. \tag{25}$$

This is an *identifiability condition* of the model class. Each model corresponds $1:1$ to a unique value of the parameter. Under this condition the Normal have a **unique solution** which we denote $\hat{\boldsymbol{\theta}}(y)$ given by

$$\hat{\boldsymbol{\theta}}(y) = [X^\top Q X]^{-1}\, X^\top Q y \quad , \tag{26}$$

which is a linear function of the observations y. For short we shall denote $\hat{\boldsymbol{\theta}}(y) = Ay$. Then $X\hat{\boldsymbol{\theta}}(y) := XAy$ is the orthogonal projection of y onto the subspace $\mathcal{X} = \operatorname{span}(X)$. In other words the matrix $P \in \mathbb{R}^{N\times N}$, defined as

$$P = XA \quad ,$$

is the orthogonal projector, with respect to the inner product $\langle\cdot,\cdot\rangle_Q$, *from* $\mathbb{R}^N$ *onto* $\mathcal{X}$. In fact P is idempotent ($P = P^2$), since

$$XA \cdot XA = X \cdot I \cdot A = XA$$

STATISTICAL ANALYSIS

Assume that the error $\mathbf{e}$ in the model

$$\mathbf{y} = X\theta + \mathbf{e}. \tag{27}$$

is zero-mean random vector with known variance $R = \mathbb{E}\,\mathbf{e}\mathbf{e}^\top$ positive definite and that the rank condition 25 holds.

Theorem 20. *The estimator $\hat{\boldsymbol{\theta}}(\mathbf{y}) = A\mathbf{y}$ is unbiased and has variance matrix*

$$\mathrm{Var}\,\hat{\boldsymbol{\theta}}(\mathbf{y}) = \left[X^\top R^{-1} X\right]^{-1}$$

If $\{\mathbf{e}_k, k = 1, 2, \ldots, N\}$ are i.i.d. that is $R = \sigma^2 I$, then

$$\mathrm{Var}\,\hat{\boldsymbol{\theta}}(\mathbf{y}) = \sigma^2 \left[X^\top X\right]^{-1}$$

In the ARX case just assume $Q = I_N$ ($N \times N$ identity matrix) then the PEM estimator of θ is

$$\hat{\theta}_N = \left[\Phi_N^\top \Phi_N\right]^{-1} \Phi_N^\top \mathbf{y}$$

which can also be rewritten

$$\hat{\theta}_N = \left[\sum_{t=1}^{N} \varphi(t)\varphi(t)^\top\right]^{-1} \sum_{t=1}^{N} \varphi(t)\mathbf{y}(t)\,.$$

where we assume that the inverse exists for suitably large N.

- What are the statistical properties of this estimator?
- Note that $\hat{\theta}_N$ is a **non linear** function of the observed data. Don't know if it is unbiased; even if $\mathbf{y}$ and $\mathbf{u}$ were Gaussian the Pdf of $\hat{\theta}_N$ is impossible to compute. Can only try to see what happens for $N \to \infty$.

CONSISTENCY OF THE PEM ESTIMATOR

(Preview)

Theorem 21. *Assume there is a true model describing the data having orders n, m as in the candidate ARX model and true parameter θ_0. Assume also that the true model is* **causal** *that is*

$$\mathbb{E}_{\theta_0} \boldsymbol{\varphi}(t)\, \mathbf{e}(t) = 0\,; \qquad \forall t \in \mathbb{Z}\,; \qquad \text{(INNOV)}$$

and that $\mathbb{E}_{\theta_0} \boldsymbol{\varphi}(t) \boldsymbol{\varphi}(t)^\top > 0$; then

$$\lim_{N \to \infty} \hat{\boldsymbol{\theta}}_N = \theta_0$$

with probability one.

STRONG CONSISTENCY OF THE PEM ESTIMATOR

Rewrite $\hat{\boldsymbol{\theta}}_N$ as

$$\hat{\theta}_N = \left[\frac{1}{N}\sum_{t=1}^{N}\varphi(t)\varphi(t)^\top\right]^{-1}\frac{1}{N}\sum_{t=1}^{N}\varphi(t)\mathbf{y}(t)\,.; \qquad \text{(EST)}$$

substitute $\mathbf{y}(t) = \varphi(t)^\top\theta_0 + \mathbf{e}(t)$ (true model) and define the sample covariance matrix of $\boldsymbol{\varphi}(t)$

$$\hat{\boldsymbol{\Sigma}}_N := \frac{1}{N}\sum_{t=1}^{N}\boldsymbol{\varphi}(t)\boldsymbol{\varphi}(t)^\top; \qquad \in \mathbb{R}^{p\times p}$$

For notation simplicity we do the case of no exogenous input $(\mathbf{u} \equiv 0)$.

Lemma 3. *If $\{\mathbf{e}\}$ is an i.i.d. process then $\{\mathbf{y}\}$ is ergodic and $\hat{\boldsymbol{\Sigma}}_N$ converges almost surely for $N \to \infty$ to the positive semidefinite covariance matrix*

$$\Sigma := \mathbb{E}_{\theta_0}\{\begin{bmatrix}\mathbf{y}(t-1)\\ \dots\\ \mathbf{y}(t-n)\end{bmatrix}\begin{bmatrix}\mathbf{y}(t-1) & \dots & \mathbf{y}(t-n)\end{bmatrix}\};$$

PROOF OF STRONG CONSISTENCY OF THE PEM ESTIMATOR

Proof. The ergodicity follows from Corollary 4 but we shall need to understand why $\mathbf{y}(t)$ admits such a representation. We shall do that in the next slides □

Now just go to the limit in formula (EST)

$$\lim_{N\to\infty}\left[\frac{1}{N}\sum_{t=1}^{N}\boldsymbol{\varphi}(t)\boldsymbol{\varphi}(t)^{\top}\right]^{-1}\frac{1}{N}\sum_{t=1}^{N}\boldsymbol{\varphi}(t)\left(\boldsymbol{\varphi}(t)^{\top}\theta_0+\mathbf{e}(t)\right)$$

to get, by ergodicity and in virtue if the two main assumptions

$$\lim_{N\to\infty}\hat{\boldsymbol{\theta}}_N=\Sigma^{-1}\Sigma\,\theta_0=\theta_0$$

here $\Sigma:=\mathbb{E}_{\theta_0}\boldsymbol{\varphi}(t)\boldsymbol{\varphi}(t)^{\top}>0$. For ARMAX models we need to assume that the input process $\mathbf{u}$ is also ergodic and uncorrelated with the noise input. Of course the reason why the two main assumptions should hold needs to be investigated.

SOLVING DIFFERENCE EQUATIONS

Linear difference equations (with constant coefficients) arise as deterministic mathematical models of many physical or economic systems. They may be written as

$$y(t) + \sum_{k=1}^{n} a_k y(t-k) = f(t)\,, \qquad t \in \mathbb{Z} \tag{28}$$

or, equivalently as

$$y(t+n) + \sum_{k=1}^{n} a_k y(t+n-k) = g(t), \qquad t \in \mathbb{Z} \tag{29}$$

where $f(t)$ or $g(t) = f(t+n)$ are exogenous signals. To find a solution first look at the homogeneous case where $f(t) = 0$. Try with a simple exponential $y(t) = \lambda^t$; which leads to

$$\lambda^{t+n} + \sum_{k=1}^{n} a_k \lambda^{t+n-k} = 0$$

since we want $\lambda \neq 0$ we can collect λ^t and end with an algebraic equation od degree n

$$\lambda^n + \sum_{k=1}^{n} a_k \lambda^{n-k} = 0$$

which is the **characteristic equation** of the system. It has n complex solutions $\lambda_k\,;\, k = 1, 2, \ldots, n$ not necessarily distinct. Hence any solution must be a linear combination of these n exponentials. Assuming all roots are distinct, the general solution turns out to be

$$y(t) = \sum_{k=1}^{n} c_k\, \lambda_k^t$$

where all coefficients c_k can be determined from the **initial conditions** say $y(0), y(1), \ldots y(n-1),$. Note that in general the λ_k may be complex numbers.

CONVOLUTION AND DIFFERENCE EQUATIONS

Suppose now that you want to solve

$$y(t+n)+\sum_{k=1}^{n} a_k y(t+n-k) = \delta(t)\,; \qquad\qquad t \in \mathbb{Z}$$

where the input sequence $\delta(t)$ is equal to 1 for t=0 and zero otherwise. This function is called the elementary or *unit impulse function*. You can work out an equivalent **homogeneous equation** to (DE) by solving for successive n initial conditions the system of equations obtained by writing (DE) at times $t = 0, 1, \ldots, n-1$.

Let's denote the solution by $h(t)$; this is called the *impulse response of the system*.

Once we have $h(t)$ we can solve the equation for an arbitrary input $f(t)$. Note that the system represented by the DE operates a **linear transformation on the input $f(t)$**. Since any input function can be expressed as a (possibly infinite) linear combinations of impulse functions located at all times $t = k$,

$$f(t) = \sum_{k=-\infty}^{+\infty} f(k)\,\delta(t-k), \qquad t \in \mathbb{Z}$$

because of linearity the response of the system can be written as a sum of infinitely many impulse responses to the $\delta(t-k)$'s each located at times $t = k$ and weighted by amplitude $f(k)$. This leads to the **convolution representation**

$$y(t) = \sum_{k=-\infty}^{+\infty} f(k)\,h(t-k)\,; \qquad \Leftrightarrow \qquad y(t) = \sum_{k=-\infty}^{+\infty} h(k)\,f(t-k)$$

Let's go back to the AR model. Now the input is an i.i.d. process $\{\mathbf{e}(t)\}$. We can still write the solution as a convolution sum

$$\mathbf{y}(t) = \sum_{k=-\infty}^{+\infty} h(k)\,\mathbf{e}(t-k)$$

which is of the same form of the representation (CONV) on p. 78. We need however to check under what circumstances the convergence condition $\sum_{k=-\infty}^{+\infty} |h(k)| < \infty$ is satisfied.

Lemma 4. *If and only if all roots of the characteristic equation have modulus strictly less than 1; i.e* $|\lambda_k| < 1\,;\, k = 1, 2, \ldots, n$, *one has :*

1. ***Causality*** *i.e.* $h(t) = 0$ *for* $t < 0$

2. ***Stability*** *i.e.* $\sum_{k=0}^{+\infty} |h(k)| < \infty\,$.

Clearly when $|\lambda_k| < 1$ then $\lim_{t\to+\infty} \lambda_k^t = 0$. True also for multiple roots. the impulse response is summable.

If $|\lambda_k| > 1$ then $\lim_{t\to+\infty} \lambda_k^t = \infty$ the impulse response is not causal. May be causal but then *not summable*.

If λ_k imaginary i.e $|\lambda_k| = 1$; since the characteristic polynomial is real they come always as pairs of compl. conjugate roots

$$a_k e^{i\lambda\, t} + \bar{a}_k e^{-i\lambda\, t} = \Re e\, a_k \cos\lambda\, t + \Im m\, a_k \sin\lambda\, t$$

oscillatory behaviour. In this case the process $\mathbf{y}$ is **not ergodic**.

CAUSALITY

When $|\lambda_k| < 1\,; k = 1, 2, \ldots, n$ the process $\mathbf{y}$ is ergodic but more is true. Since $h(t) = 0$ for $t < 0$, one can write

$$\mathbf{y}(t) = \sum_{k=0}^{+\infty} h(k)\,\mathbf{e}(t-k) \;=\; \sum_{k=-\infty}^{t} h(t-k)\,\mathbf{e}(k)$$

so $\mathbf{y}(t)$ **depends only on the past history of** $\{\mathbf{e}(t)\}$. In general infinite, since $h(t)$ is non zero for all $t \geq 0$. Write

$$\mathbf{y}(t-i) = \sum_{k=-\infty}^{t-i} h(t-i-k)\,\mathbf{e}(k)\,; \qquad \text{linear function of past } \mathbf{e}\text{'s at times} \leq t-i$$

Therefore, since the $\mathbf{e}(t)$ are uncorrelated,

$$\mathbb{E}\,\mathbf{e}(t)\mathbf{y}(t-i)\; = 0),; \qquad i = 1, 2, \ldots, n$$

If there is an input process this needs to be assumed.

CAUSALITY AND THE PREDICTION ERROR

For a **causal AR model**

$$\mathbf{y}(t) = \sum_{k=1}^{n} a_k \mathbf{y}(t-k) + \mathbf{e}(t)$$

the two terms in the right side are uncorrelated.
The second term is the **conditional expectation** of $\mathbf{y}(t)$, given its (infinite) past history up to time $t-1$. This is the (optimal) one-step-ahead predictor of $\mathbf{y}(t)$ given its own past. The i.i.d. process $\mathbf{e}(t)$ is then the (random) **prediction error**.

This is true also if there is an external input assumed independent of $\mathbf{e}$. If in the AR model all roots of the characteristic equation are in absolute value less than one. The i.i.d. process $\mathbf{e}$ has the interpretation of one-step-ahead prediction error of $\mathbf{y}(t)$ given the joint past of $\mathbf{y}$ and $\mathbf{u}$ at time $t-1$.

CONDITIONAL EXPECTATION

Denote by $L^2(\mathbf{y})$, the vector space of statistics of the process $\mathbf{y}$ which have finite second order moment; that is $f(\mathbf{y}(s)\,;\, s\in\mathbb{Z})$ with $\mathbb{E}\, f(\mathbf{y})^2 < \infty$.

Let $\mathbf{x}$ be a random variable of the same experiment. We want to find the statistic in $L^2(\mathbf{y})$ which approximates $\mathbf{x}$ in some optimal way; that is such that the expected square error

$$\mathbb{E}\,[\mathbf{x} - f(\mathbf{y})]^2, \qquad\qquad f(\mathbf{y}) \in L^2(\mathbf{y})$$

is minimal. The solution of this problem is the **conditional expectation** denoted $\mathbb{E}\,(\mathbf{x} \mid \mathbf{y})$, of $\mathbf{x}$ given $\mathbf{y}$. We have the following (stochastic) orthogonality principle

Theorem 22. *The conditional expectation is the unique random variable in $L^2(\mathbf{y})$ satisfying the orthogonality condition:*

$$\mathbf{x} - \mathbb{E}\,(\mathbf{x} \mid \mathbf{y}) \perp L^2(\mathbf{y}) \tag{30}$$

where the orthogonality is with respect to the covariance inner product $\langle \mathbf{x}\,,\, \mathbf{z}\rangle := \operatorname{cov}\{\mathbf{x}\,,\, \mathbf{z}\}$.

MARTINGALE-DIFFERENCE PROCESSES

For AR models with i.i.d. noise input, the (one-step-ahead) prediction error turns out to coincide with $\mathbf{e}(t)$ itself and is (statistically) independent of the optimal predictor. This was defined as the conditional expectation of $\mathbf{y}(t)$ given the past history of the process up to time $t-1$:

$$\hat{\mathbf{y}}(t \mid t-1) = \mathbb{E}\left[\mathbf{y}(t) \mid \mathbf{y}^{t-1}\right].$$

Denote the prediction error by

$$\tilde{\mathbf{y}}(t) := \mathbf{y}(t) - \hat{\mathbf{y}}(t \mid t-1)$$

By the orthogonality property of conditional expectation $\tilde{\mathbf{y}}(t)$ is **uncorrelated but not necessarily independent** of all statistics of the past say

$$\mathbb{E}\,\tilde{\mathbf{y}}(t) f(\mathbf{y}^{t-1}) = 0, \qquad \forall f \in L^2(\mathbf{y}^{t-1})$$

This is the **martingale difference property** which is weaker and generalizes the i.i.d. property.

Since $\mathbb{E}\{\tilde{\mathbf{y}}(t) \mid \mathbf{y}^{t-1}\} = 0$,

$$\mathbb{E}\,\tilde{\mathbf{y}}(t) = \mathbb{E}\{\mathbb{E}[\tilde{\mathbf{y}}(t) \mid \mathbf{y}^{t-1}]\} = 0$$

Often the conditional variance $\mathbb{E}\{\tilde{\mathbf{y}}(t)\tilde{\mathbf{y}}(t)^\top \mid \mathbf{y}^{t-1}\}$ does not depend on the conditioning variables, that is

$$\mathbb{E}\{\tilde{\mathbf{y}}(t)\tilde{\mathbf{y}}(t)^\top \mid \mathbf{y}^{t-1}\} = \mathbb{E}\{\tilde{\mathbf{y}}(t)\,\tilde{\mathbf{y}}(t)^\top) = \operatorname{Var}(\tilde{\mathbf{y}}(t))$$

we say that $\tilde{\mathbf{y}}$ has a non random conditional variance.

For ARX and ARMAX models the predictor also depends on the past of some exogenous variable $\mathbf{u}$. The definition is generalized as conditional expectation given the joint past histories $(\mathbf{y}^{t-1}, \mathbf{u}^{t-1})$ and

$$\tilde{\mathbf{y}}(t) = \mathbf{y}(t) - \mathbb{E}\{\mathbf{y}(t) \mid \mathbf{y}^{t-1}, \mathbf{u}^{t-1}\} \qquad t \in \mathbb{Z}$$

In this case one replaces $L^2(\mathbf{y}^{t-1})$ with $L^2(\mathbf{y}^{t-1}\,\mathbf{u}^{t-1})$.

GENERAL D-MARTINGALES

Definition 10. *Let $\{\mathbf{z}_t\,;\, t\in\mathbb{Z}\}$ be a stationary vector process and consider the sequence of subspaces $L^2(\mathbf{z}^t)$, which is non-decreasing i.e. $L^2(\mathbf{z}^t)\subset L^2(\mathbf{z}^{t+1})$. A stochastic process $\{\mathbf{x}(t)\,;\, t\in\mathbb{Z}\}$is a* martingale difference, *or briefly, a* d-martingale *with respect to the family $\{L^2(\mathbf{z}^t)\}$, if,*

○ *For all t, $\mathbf{x}(t)\in L^2(\mathbf{z}^t)$; i.e. $\mathbf{x}(t)$ is itself a statistic of the past history of $\mathbf{z}$ at time t, having finite variance.*

○ *$\mathbf{z}(t+1)$ is uncorrelated with all random variables in the space $L^2(\mathbf{z}^t)$ that is $\mathbb{E}\{\mathbf{x}(t+1)\,f(\mathbf{z}^t)\}=0$ for all $f(\mathbf{z}^t)\in L^2(\mathbf{z}^t)$ which is equivalent to:*

$$\mathbb{E}\{\mathbf{x}(t+1)\mid \mathbf{z}^t\}=0 \qquad t\in\mathbb{Z}. \qquad \textit{(DMART)}$$

Note that by the very notion of a projection,

$$\mathbb{E}\{\mathbf{x}(t)\mid \mathbf{z}^s\}=\mathbb{E}\{\mathbb{E}\,[\mathbf{x}(t)\mid \mathbf{z}^{t-1}]\mid \mathbf{z}^s\}=0 \qquad \forall s<t\,.$$

In particular a d-martingale has mean zero.

Example:
Let $\tilde{\mathbf{y}}(t)$ be the random prediction error defined in a previous slide an consider the process

$$\mathbf{x}(t) := \boldsymbol{\varphi}(t)\,\tilde{\mathbf{y}}(t), \qquad\qquad \boldsymbol{\varphi}(t) \in L^2(\mathbf{y}^{t-1}, \mathbf{u}^{t-1})$$

where $\boldsymbol{\varphi}(t) = \boldsymbol{\varphi}(\mathbf{y}^{t-1}, \mathbf{u}^{t-1})$ is a function of the past history of $\mathbf{y}, \mathbf{u}$ up to time $t-1$. This is also a d-martigale; in fact

$$\mathbb{E}\{\mathbf{x}(t) \mid L^2(\mathbf{y}^{t-1}, \mathbf{u}^{t-1})\} = \boldsymbol{\varphi}(t)\,\mathbb{E}\{\tilde{\mathbf{y}}(t) \mid L^2(\mathbf{y}^{t-1}, \mathbf{u}^{t-1})\} = 0 \qquad t \in \mathbb{Z}\,.$$

A *martingale* is the discrete integral of a d-martingale,

$$\mathbf{m}(t) = \mathbf{m}(0) + \sum_{s=1}^{t} \mathbf{x}(s) \tag{31}$$

It is a non-stationary process, generalization of the well-kown *random walk* process.

The following lemma generalizes the sum-of-variances property, known to hold for sums of i.i.d. processes, to d-martingales.

Lemma 5. *For every d-martingale* $\mathbf{x}$ *one has*

$$\operatorname{Var}\{\sum_{t=1}^{N}\mathbf{x}(t)\} = \sum_{t=1}^{N}\operatorname{Var}\{\mathbf{x}(t)\} \tag{32}$$

When the d-martigale $\mathbf{x}$ *is stationary the second member is just* $N\,\sigma_{\mathbf{x}}^2$*. This property holds unchanged in the vector case.*

Proof.

$$\mathbb{E}\{\sum_{t=1}^{N}\mathbf{x}(t)\}^2 = \mathbb{E}\{\mathbf{x}(1)^2+\mathbf{x}(2)^2+\ldots+\mathbf{x}(N)^2\}+\mathbb{E}\{2\sum_{t>s}\mathbf{x}(t)\mathbf{x}(s)\} =$$

$$= \sum_{t=1}^{N}\operatorname{Var}\{\mathbf{x}(t)\}+2\sum_{t>s}\mathbb{E}\,\mathbf{x}(t)\mathbf{x}(s)$$

But the last term is zero since for $t>s$, $\mathbf{x}(s)\in L^2(\mathbf{z}^{\,s})$, and by the d-martingale property

$$\mathbb{E}\,\mathbf{x}(t)\mathbf{x}(s) = \mathbb{E}\{\mathbb{E}\,[\mathbf{x}(t)\mathbf{x}(s)\mid\mathbf{z}^{\,s}]\} = \mathbb{E}\{\,\mathbf{x}(s)\,\mathbb{E}\,[\,\mathbf{x}(t)\mid\mathbf{z}^{\,s}]\,\} = 0$$

THE CLT FOR D-MARTINGALES

Theorem 23 (Levy, Doob, Billingsley, Ibragimov)**.** *Let* $\{\mathbf{x}(t)\}$ *be a stationary (vector) d-martingale having constant conditional variance, equal to* $\Sigma_{\mathbf{x}} = \mathbb{E}\,\mathbf{x}(t)\mathbf{x}(t)^\top$. *One has*

$$\sqrt{T}\bar{\mathbf{x}}_T \xrightarrow{L} \mathcal{N}(0, \Sigma_{\mathbf{x}}) \tag{33}$$

that is the modified sample mean $\sqrt{T}\bar{\mathbf{x}}_T$ *converges in distribution to a multivariate Gaussian distribution of mean zero and variance* $\Sigma_{\mathbf{x}}$.

Proof (for the scalar case): We shall use the *conditional* characteristic function, substituting the variable it with $i\lambda$. By stationarity the variances of the random variables $\{\mathbf{x}(t)\}$ are uniformly bounded and $\phi_{\mathbf{x}(t)}(i\lambda \mid \mathbf{z}^{t-1})$ admits a second derivative at $\lambda = 0$ equal to the (conditional) variance, σ^2, of $\mathbf{x}(t)$. One can write

$$\mathbb{E}\left[e^{i\lambda\mathbf{x}(t)} \mid \mathbf{z}^{t-1}\right] = \mathbb{E}\left[1 + i\lambda\mathbf{x}(t) - \frac{\lambda^2}{2}\mathbf{x}(t)^2 + \boldsymbol{\eta}(\lambda, \mathbf{x}(t)) \mid \mathbf{z}^{t-1}\right] = 1 - \frac{\sigma^2\lambda^2}{2} + o(\lambda^2)$$

where $o(\lambda^2)$ is a random variable in $L^2(\mathbf{z}^{t-1})$ which tens to zero as $\lambda \to 0$ faster than λ^2. Call $\phi_T(\lambda)$ the conditional characteristic function of the sum

$\bar{\mathbf{x}}(T) := \sum_{k=1}^{T} \mathbf{x}(t)$. Then

$$\begin{aligned}
\phi_T(\lambda) &= \mathbb{E}\left\{\mathbb{E}\left[e^{i\lambda \mathbf{x}(T)} \mid \mathbf{z}^{T-1}\right] e^{i\lambda\,\bar{\mathbf{x}}(T-1)}\right\} = \\
&= \left[1 - \frac{\sigma^2\lambda^2}{2}\right] \mathbb{E}\{e^{i\lambda\bar{\mathbf{x}}(T-1)}\} + \mathbb{E}\{o(\lambda^2)e^{i\lambda\bar{\mathbf{x}}(T-1)}\} = \\
&= \left[1 - \frac{\sigma^2\lambda^2}{2}\right] \phi_{T-1}(\lambda) + \bar{o}(\lambda^2)
\end{aligned}$$

where $\bar{o}(\lambda^2)$ is the expected value of a variable in $L^2(\mathbf{z}^{T-1})$ having the same absolute value of $o(\lambda^2)$ hence tending to zero faster than λ^2. Solving the difference equation one finds

$$\phi_T(\lambda) = \left[1 - \frac{\sigma^2\lambda^2}{2}\right]^T + \bar{o}_T(\lambda^2)$$

where $\bar{o}_T(\lambda^2)$ is still infinitesimal of higher order than λ^2 for $\lambda \to 0$.

Now, the characteristic function of $\mathbf{s}(T) := \frac{1}{\sqrt{T}} \sum_{k=1}^{T} \mathbf{x}(t)$ is the same function ϕ_T computed in $\lambda/\sqrt{T}$,so that

$$\phi_T(\frac{\lambda}{\sqrt{T}}) = \left[1 - \frac{\sigma^2}{2}\frac{\lambda^2}{T}\right]^T + \bar{o}_T(\frac{\lambda^2}{T})$$

where the second term tends to zero for $T \to \infty$, for whatever fixed value of λ while the limit of the first term is the well known function $\exp\{-\frac{\sigma^2\lambda^2}{2}\}$. Hence the characteristic function of $\mathbf{s}(T)$ converges pointwise to that of the Gaussian $\mathcal{N}(0, \sigma^2)$. □

THE CLT FOR THE PEM ESTIMATOR

We shall consider simultaneously models of AR, ARX, ARMA, ARMAX type depending smoothly on a p-dimensional parameter θ. Assume at least local identifiability about the **true value** θ_0 of the **true model** which has generated the data. Assume also that the true model generates data which are stationary and ergodic and that the PEM estimator is consistent.

Each candidate model defines a probability density which belongs to a parametric class $\{p(y \mid \theta)\}$ each member being uniquely defined by just selecting a p-dimensional paremeter value θ. So all models in the class have the same order or complexity. In particular the true model belongs to this class and is uniquely described by assigning the true parameter θ_0.

We assume that the one-step-ahead predictor function $\hat{\mathbf{y}}_\theta(t \mid t-1)$ is stationary and has a known expression. Our goal is to study the asymptotic distribution of minimizer(s) of some generalized average square prediction error criterion which we shall denote $V_N(\theta)$.

Recall that each candidate minimizer must solve the gradient equation,

$$\frac{\partial V_N(\theta)}{\partial \theta} := V_N(\theta)' = 0 \,. \qquad \text{(a p-vector function)}$$

To approximate $V_N(\hat{\boldsymbol{\theta}}_N)'$ use Taylor's formula about the point $\theta = \theta_0$ truncated to the first order. There is always some point $\bar{\theta}$, such that, exactly

$$V_N(\hat{\boldsymbol{\theta}}_N)' = V_N(\theta_0)' + V_N(\bar{\boldsymbol{\theta}})''(\hat{\boldsymbol{\theta}}_N - \theta_0) = 0 \tag{34}$$

wher $V_N(\bar{\boldsymbol{\theta}})''$ is the second derivatives (Hessian) matrix computed at some unknown $\bar{\theta}$ which however must belong to the$p$-dimensional interval having extremes θ_0 and$\hat{\boldsymbol{\theta}}_N$, that is

$$\theta_0^k \leq \bar{\boldsymbol{\theta}}^k \leq \hat{\boldsymbol{\theta}}_N^k, \qquad k = 1, 2, \ldots, p$$

Assuming the Hessian matrix is invertible from (34) one gets

$$\hat{\boldsymbol{\theta}}_N - \theta_0 = -\left[\frac{1}{2} V_N(\bar{\boldsymbol{\theta}})''\right]^{-1} \frac{1}{2} V_N(\theta_0)' \tag{35}$$

where the factor $\frac{1}{2}$ is intruduced for convenience.

Let's compute the gradient and the Hessian matrix using the expression $V_N(\theta) = \frac{1}{N}\sum_{t=1}^{N} \boldsymbol{\varepsilon}_\theta(t)^2$. Define:

$$\boldsymbol{\psi}_\theta(t) := \frac{\partial \boldsymbol{\varepsilon}_\theta(t)}{\partial \theta} = -\frac{\partial \hat{\mathbf{y}}_\theta(t \mid t-1)}{\partial \theta}$$

to get

$$\frac{1}{2}V_N(\theta)' = \frac{1}{N}\sum_{k=1}^{N} \boldsymbol{\psi}_\theta(t)\boldsymbol{\varepsilon}_\theta(t) \tag{36}$$

$$\frac{1}{2}V_N(\theta)'' = \frac{1}{N}\sum_{k=1}^{N} \left\{ \boldsymbol{\psi}_\theta(t)\boldsymbol{\psi}_\theta(t)^\top + \boldsymbol{\varepsilon}_\theta(t) \left[\frac{\partial^2 \boldsymbol{\varepsilon}_\theta(t)}{\partial \theta_i \partial \theta_j}\right] \right\} \tag{37}$$

The asymptotic behaviour of the second derivative is discussed in the following lemma,

Lemma 6. *One has*

$$\lim_{N\to\infty} \frac{1}{2}V_N(\bar{\boldsymbol{\theta}})'' = \mathbb{E}_{\theta_0}\left\{ \boldsymbol{\psi}_{\theta_0}(t)\boldsymbol{\psi}_{\theta_0}(t)^\top \right\} \tag{38}$$

with probability one.

Proof : By consistency, $\hat{\boldsymbol{\theta}}_N \to \theta_0$ and hence also $\bar{\boldsymbol{\theta}} \to \theta_0$ (with probability one). Since the time average in (37) converges to the expectation, we have,

$$\frac{1}{2}V_N(\bar{\boldsymbol{\theta}})'' \to \mathbb{E}_{\theta_0}\left\{\boldsymbol{\psi}_{\theta_0}(t)\boldsymbol{\psi}_{\theta_0}(t)^\top + \boldsymbol{\varepsilon}_{\theta_0}(t)\left[\frac{\partial^2\boldsymbol{\varepsilon}_\theta(t)}{\partial\theta_i\partial\theta_j}\right]_{|\theta=\theta_0}\right\}$$

almost surely. Since the true model belongs to the model class, $\boldsymbol{\varepsilon}_{\theta_0}(t) = \mathbf{e}_0(t)$ where $\mathbf{e}_0(t)$ is the true random prediction error (a d-martingale). Finally, both the gradient $(\boldsymbol{\psi}_\theta(t))$, and the second derivative of $\hat{\mathbf{y}}_\theta(t \mid t-1)$ are functions (often linear functions) only of the past data $(\mathbf{y}^{t-1}, \mathbf{u}^{t-1})$, all entries in the second derivative matrix in the second member are uncorrelated with $\mathbf{e}_0(t)$ and hence the expectation of the last term is zero. □

Note now that the last term in (35), has the expression

$$\frac{1}{2}V_N(\theta_0)' = \frac{1}{N}\sum_{k=1}^{N}\boldsymbol{\psi}_{\theta_0}(t)\mathbf{e}_0(t) \tag{39}$$

AN INTERMEDIATE CLT

Theorem 24. *Assume the prediction error $\mathbf{e}_0$, is a stationary d-martingale with respect to the flow of past data $(\mathbf{y}^t, \mathbf{u}^t)$ having finite variance. Then also the process $\{\boldsymbol{\psi}_{\theta_0}(t)\mathbf{e}_0(t)\}$ is a d-martingale and,*

$$\sqrt{N}\frac{1}{2}V_N(\theta_0)' \xrightarrow{L} \mathcal{N}(0, Q) \tag{40}$$

If the conditional variance of $\mathbf{e}_0(t)$ doesnot depend on the data $(\mathbf{y}^{t-1}, \mathbf{u}^{t-1})$, that is if

$$\mathbb{E}_0\{\mathbf{e}_0(t)^2 \mid \mathbf{y}^{t-1}, \mathbf{u}^{t-1}\} = \mathbb{E}_0\{\mathbf{e}_0(t)^2\} = \sigma_0^2\,, \tag{41}$$

the asymptotic variance Q is given by the formula,

$$Q = \sigma_0^2\,\mathbb{E}_0\{\boldsymbol{\psi}_{\theta_0}(t)\boldsymbol{\psi}_{\theta_0}(t)^\top\}\,. \tag{42}$$

Proof : The first statement follows from the previous observation that $\{\boldsymbol{\psi}_{\theta_0}(t)\mathbf{e}_0(t)\}$ is also a d-martingale with respect to the past data flow $\{\mathbf{y}^t, \mathbf{u}^t\}$.

The second is a corollary of the CLT for d-martingales 23. The xpression for the Variance matrix Q follows from the property (41), which implies,

$$\begin{aligned}\mathrm{Var}\{\boldsymbol{\psi}_{\theta_0}(t)\mathbf{e}_0(t)\} &= \quad \mathbb{E}_0\{\mathbb{E}_0\left[\mathbf{e}_0(t)^2 \mid \mathbf{y}^{t-1}, \mathbf{u}^{t-1}\right]\}\,\boldsymbol{\psi}_{\theta_0}(t)\boldsymbol{\psi}_{\theta_0}(t)^\top\} = \\ &= \quad \mathbb{E}_0\{\mathbf{e}_0(t)^2\}\,\mathbb{E}_0\{\boldsymbol{\psi}_{\theta_0}(t)\boldsymbol{\psi}_{\theta_0}(t)^\top\}\,.\end{aligned}$$

□

Obviously the theorem includes the case where $\mathbf{e}_0$ is an i.i.d. process but applies to a much wider variety of cases.

THE ASYMPTOTIC DISTRITBUTION OF THE PEM ESTIMATOR

The following theorem is a fundamental result in time series analysis. It actually includes asymptotic normality of the Maximum Likelihood estimate, which is asymptotically equivalent to PEM.

Theorem 25. *Under the same assumptions of Theorem 24 the PEM estimator has a Gaussian asymptotic distribution as described by,*

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \theta_0) \xrightarrow{L} \mathcal{N}(0, P)\,, \tag{43}$$

where the asymptotic variance matrix P is given by the expression

$$P = \sigma_0^2 \left[\mathbb{E}_0\{\boldsymbol{\psi}_{\theta_0}(t)\boldsymbol{\psi}_{\theta_0}(t)^\top\}\right]^{-1} \tag{44}$$

and the inverse of the matrix between square brackets exists.

Proof: Follows from (35). For $N \to \infty$

$$\hat{\boldsymbol{\theta}}_N - \theta_0 \simeq - \left[\frac{1}{N} \sum_{k=1}^{N} \boldsymbol{\psi}_{\bar{\boldsymbol{\theta}}}(t) \boldsymbol{\psi}_{\bar{\boldsymbol{\theta}}}(t)^\top \right]^{-1} \frac{1}{N} \sum_{k=1}^{N} \boldsymbol{\psi}_{\theta_0}(t) \boldsymbol{\varepsilon}_{\theta_0}(t)$$

and use the third statement of Slutsky. The expression for P follows from

$$P = \left[\mathbb{E}_{\theta_0} \left\{ \boldsymbol{\psi}_{\theta_0}(t) \boldsymbol{\psi}_{\theta_0}(t)^\top \right\} \right]^{-1} Q \left[\mathbb{E}_{\theta_0} \left\{ \boldsymbol{\psi}_{\theta_0}(t) \boldsymbol{\psi}_{\theta_0}(t)^\top \right\} \right]^{-1}$$

where Q is the asymptotic variance of the limit (40). The invertibility of P is equivalent to that of the Fisher matrix. □

ASYMPTOTIC VARIANCE

Recall that the variance of a consistent estimator must tend to zero as $N \to \infty$. The concept of *asymptotic variance* of a consistent estimator must therefore be defined properly. Here is one possible definition.

Definition 11. *Let $\{\phi_N(\mathbf{y}) \,;\, N = 1, 2, \ldots\}$ be a consistent sequence of estimators of the parameter θ and $d(N)$ a function of N which is increasing to $+\infty$ with N and strictly positive. One say that $\phi_N(\mathbf{y})$ has asymptotic variance Σ if*

$$\sqrt{d(N)}\,[\,\phi_N(\mathbf{y}) - \theta_0\,] \xrightarrow{L} D(0, \Sigma)$$

where $D(0, \Sigma)$ is a pdf having variance Σ, possibly depending on $theta_0$, which is finite and strictly positive definite.

Hence for N large the variance of $\phi_N(\mathbf{y})$ can be approximated by $\frac{1}{d(N)}\Sigma$. In most asymptotically normal examples discussed above $d(N)$ can be taken equal to N.

The condition of strict positivity $\Sigma > 0$ is essential since it excludes the possibility of linear combinations of the components of $\phi_N(\mathbf{y})$ whose variance tends to zero, which just means that the order of infinitesimal of the variance of these combinations will be different from $O(\frac{1}{d(N)})$.

ASYMPTOTIC EFFICIENCY

Discussion in Lehmann p.510

One may compare consistent estimators based on the asymptotic variance, saying that estimator 1 is more efficient than estimator 2 when $\Sigma_1 \leq \Sigma_2$. Unfortunately however in general the asymptotic variance is a function of the true parameter which is unknown so it may be that for different values of θ_0 the two estimators compare in the opposite way or do not compare at all.

For maximum likelihood under the usual identifiability condition, the asymptotic variance is $I(\theta_0)^{-1}$ which in force of the Cramèr-Rao bound is the best possible (for a fixed θ_0). One usually says that **maximum likelihood is an asymptotically efficient estimator** but this sentence must of course be interpreted with a grain of salt.