Chapter 3 :
Likelihood function and inference

The likelihood

Given an usually parametric family of distributions

$$F \in \{F_\theta,\ \theta \in \Theta\}$$

with densities f_θ [wrt a fixed measure ν], the density of the iid sample $x_1, \ldots, x_n$ is

$$\prod_{i=1}^{n} f_\theta(x_i)$$

Note In the special case ν is a counting measure,

$$\prod_{i=1}^{n} f_\theta(x_i)$$

is the probability of observing the sample $x_1, \ldots, x_n$ among all possible realisations of $X_1, \ldots, X_n$

The likelihood

Given an usually parametric family of distributions

$$F \in \{F_\theta,\ \theta \in \Theta\}$$

with densities f_θ [wrt a fixed measure ν], the density of the iid sample $x_1, \ldots, x_n$ is

$$\prod_{i=1}^{n} f_\theta(x_i)$$

Note In the special case ν is a counting measure,

$$\prod_{i=1}^{n} f_\theta(x_i)$$

is the probability of observing the sample $x_1, \ldots, x_n$ among all possible realisations of $X_1, \ldots, X_n$

The likelihood

Definition (likelihood function)

The likelihood function associated with a sample $x_1, \ldots, x_n$ is the function

$$\begin{aligned} L : \Theta &\longrightarrow \mathbb{R}_+ \\ \theta &\longrightarrow \prod_{i=1}^{n} f_\theta(x_i) \end{aligned}$$

same formula as density but different space of variation

The likelihood

Definition (likelihood function)

The likelihood function associated with a sample $x_1, \ldots, x_n$ is the function

$$\begin{aligned} L : \Theta &\longrightarrow \mathbb{R}_+ \\ \theta &\longrightarrow \prod_{i=1}^{n} f_\theta(x_i) \end{aligned}$$

same formula as density but different space of variation

Example: density function versus likelihood function

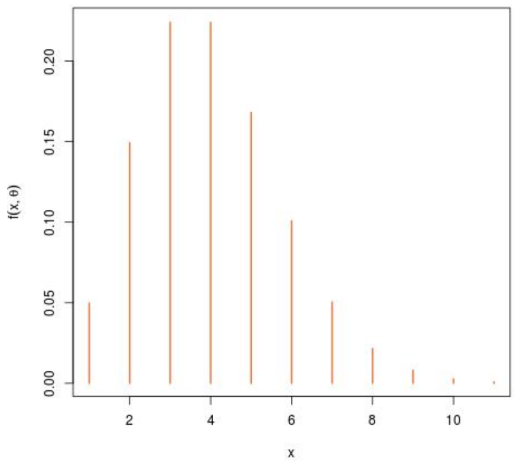
Take the case of a Poisson density
[against the counting measure]

$$f(x;\theta) = \frac{\theta^x}{x!}\, e^{-\theta}\, \mathbb{I}_{\mathbb{N}}(x)$$

which varies in $\mathbb{N}$ as a function of x
versus

$$L(\theta;x) = \frac{\theta^x}{x!}\, e^{-\theta}$$

which varies in $\mathbb{R}_+$ as a function of θ



$\theta = 3$

Example: density function versus likelihood function

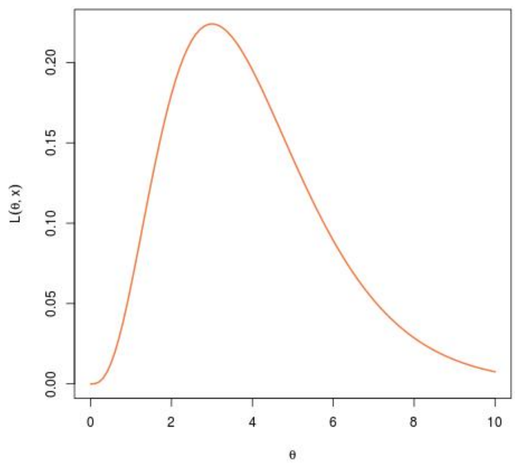
Take the case of a Poisson density
[against the counting measure]

$$f(x;\theta) = \frac{\theta^x}{x!}\, e^{-\theta}\, \mathbb{I}_{\mathbb{N}}(x)$$

which varies in $\mathbb{N}$ as a function of x
versus

$$L(\theta;x) = \frac{\theta^x}{x!}\, e^{-\theta}$$

which varies in $\mathbb{R}_+$ as a function of θ



$x = 3$

Example: density function versus likelihood function

Take the case of a Normal $\mathcal{N}(0, \theta)$ density [against the Lebesgue measure]

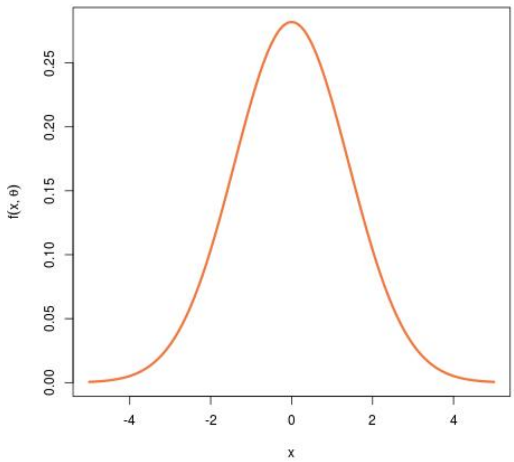
$$f(x;\theta) = \frac{1}{\sqrt{2\pi\theta}}\, e^{-x^2/2\theta}\, \mathbb{I}_{\mathbb{R}}(x)$$

which varies in $\mathbb{R}$ as a function of x

versus

$$L(\theta;x) = \frac{1}{\sqrt{2\pi\theta}}\, e^{-x^2/2\theta}$$

which varies in $\mathbb{R}_+$ as a function of θ



$\theta = 2$

Example: density function versus likelihood function

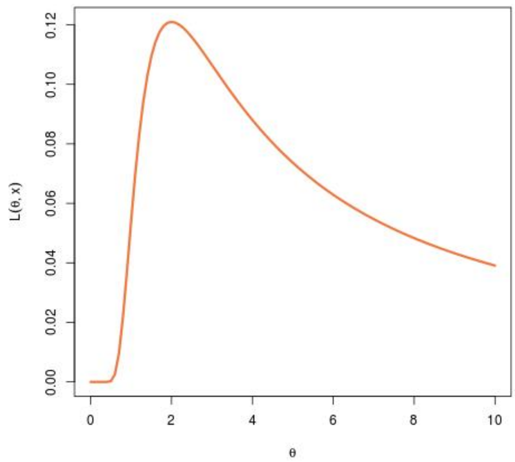
Take the case of a Normal $\mathcal{N}(0, \theta)$ density [against the Lebesgue measure]

$$f(x;\theta) = \frac{1}{\sqrt{2\pi\theta}}\, e^{-x^2/2\theta}\, \mathbb{I}_{\mathbb{R}}(x)$$

which varies in $\mathbb{R}$ as a function of x versus

$$L(\theta;x) = \frac{1}{\sqrt{2\pi\theta}}\, e^{-x^2/2\theta}$$

which varies in $\mathbb{R}_+$ as a function of θ



$x = 2$

Example: density function versus likelihood function

Take the case of a Normal $\mathcal{N}(0, 1/\theta)$ density [against the Lebesgue measure]

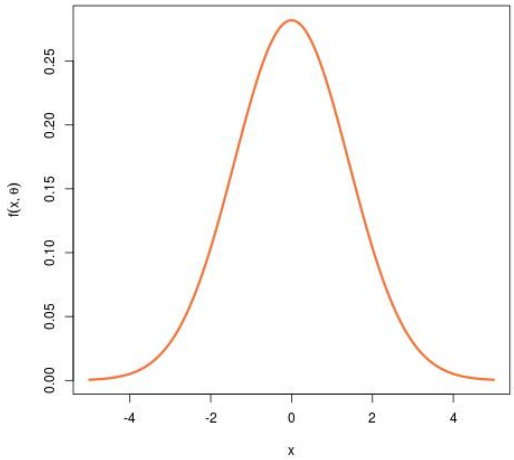
$$f(x;\theta) = \frac{\sqrt{\theta}}{\sqrt{2\pi}}\, e^{-x^2\theta/2}\, \mathbb{I}_{\mathbb{R}}(x)$$

which varies in $\mathbb{R}$ as a function of x

versus

$$L(\theta;x) = \frac{\sqrt{\theta}}{\sqrt{2\pi}}\, e^{-x^2\theta/2}\, \mathbb{I}_{\mathbb{R}}(x)$$

which varies in $\mathbb{R}_+$ as a function of θ



$\theta = 1/2$

Example: density function versus likelihood function

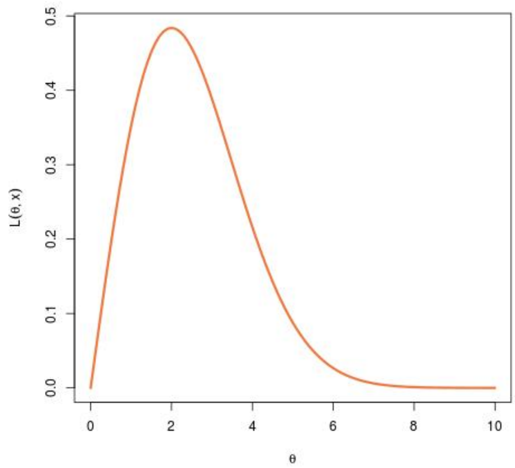
Take the case of a Normal $\mathcal{N}(0, 1/\theta)$ density [against the Lebesgue measure]

$$f(x;\theta) = \frac{\sqrt{\theta}}{\sqrt{2\pi}} e^{-x^2\theta/2} \mathbb{I}_{\mathbb{R}}(x)$$

which varies in $\mathbb{R}$ as a function of x versus

$$L(\theta;x) = \frac{\sqrt{\theta}}{\sqrt{2\pi}} e^{-x^2\theta/2} \mathbb{I}_{\mathbb{R}}(x)$$

which varies in $\mathbb{R}_+$ as a function of θ



$x = 1/2$

Example: Hardy-Weinberg equilibrium

Population genetics:

- Genotypes of biallelic genes AA, Aa, and aa
- sample frequencies n_{AA}, n_{Aa} and n_{aa}
- multinomial model $\mathcal{M}(n; p_{AA}, p_{Aa}, p_{aa})$
- related to population proportion of A alleles, p_A:

$$p_{AA} = p_A^2 \,,\; p_{Aa} = 2p_A(1-p_A)\,,\; p_{aa} = (1-p_A)^2$$

- likelihood

$$L(p_A|n_{AA}, n_{Aa}, n_{aa}) \propto p_A^{2n_{AA}}[2p_A(1-p_A)]^{n_{Aa}}(1-p_A)^{2n_{aa}}$$

[Boos & Stefanski, 2013]

mixed distributions and their likelihood

Special case when a random variable X may take specific values $a_1, \dots, a_k$ and a continum of values $\mathfrak{A}$

Example: Rainfall at a given spot on a given day may be zero with positive probability p_0 [it did not rain!] or an arbitrary number between 0 and 100 [capacity of measurement container] or 100 with positive probability p_{100} [container full]

mixed distributions and their likelihood

Special case when a random variable X may take specific values $a_1, \ldots, a_k$ and a continum of values $\mathfrak{A}$

Example: Tobit model where $y \sim \mathcal{N}(X^T\beta, \sigma^2)$ but $y^* = y \times \mathbb{I}\{y \geqslant 0\}$ observed

mixed distributions and their likelihood

Special case when a random variable X may take specific values $a_1, \ldots, a_k$ and a continum of values $\mathfrak{A}$

Density of X against composition of two measures, counting and Lebesgue:

$$f_X(a) = \begin{cases} \mathbb{P}_\theta(X = a) & \text{if } a \in \{a_1, \ldots, a_k\} \\ f(a|\theta) & \text{otherwise} \end{cases}$$

Results in likelihood

$$L(\theta|x_1, \ldots, x_n) = \prod_{j=1}^{k} \mathbb{P}_\theta(X = a_i)^{n_j} \times \prod_{x_i \notin \{a_1, \ldots, a_k\}} f(x_i|\theta)$$

where n_j # observations equal to a_j

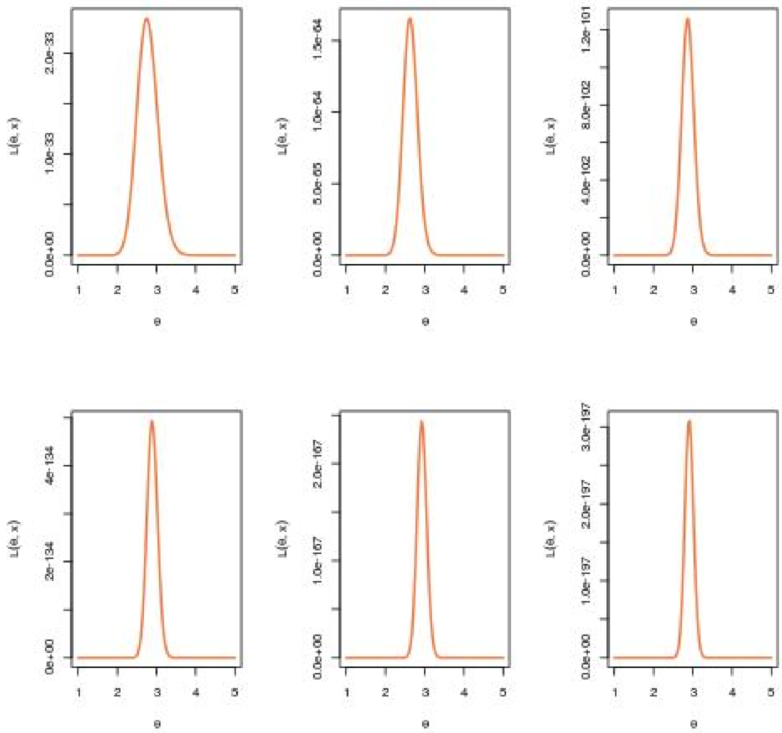
Enters Fisher, Ronald Fisher!

Fisher's intuition in the 20's:

- the likelihood function contains the relevant information about the parameter θ
- the higher the likelihood the more likely the parameter
- the curvature of the likelihood determines the precision of the estimation

Concentration of likelihood mode around “true” parameter

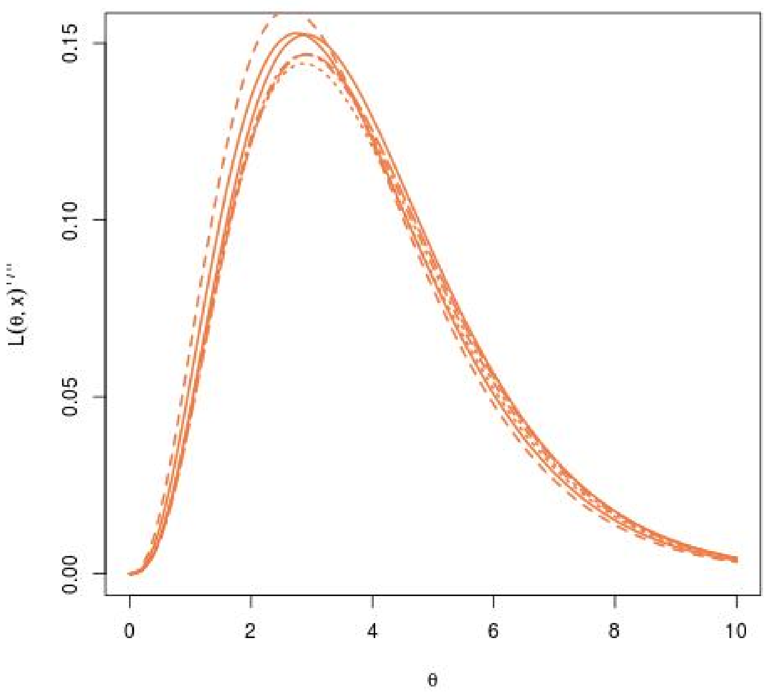
Likelihood functions for $x_1, \ldots, x_n \sim \mathcal{P}(3)$ as n increases



$$n = 40, ..., 240$$

Concentration of likelihood mode around "true" parameter

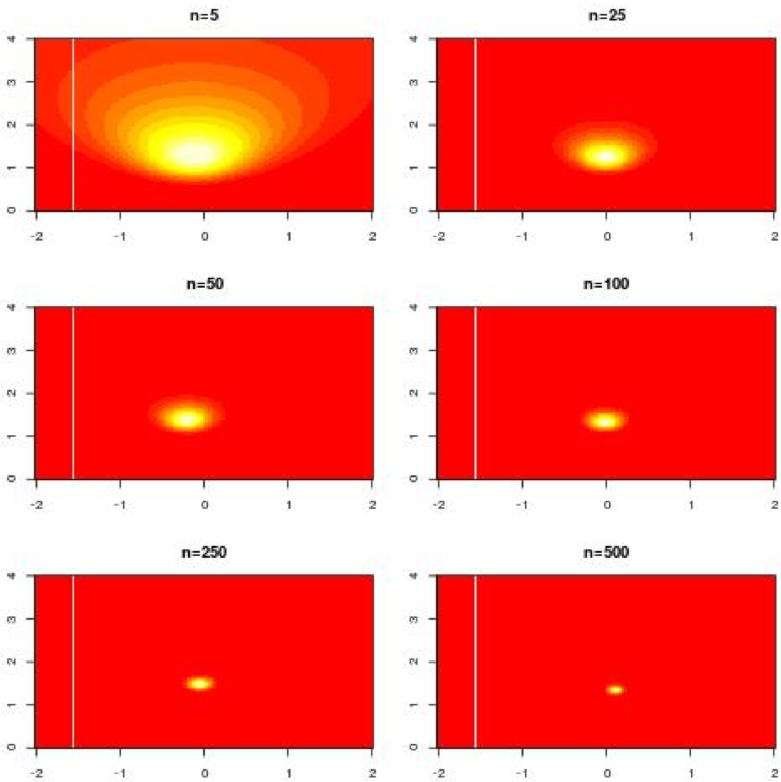
Likelihood functions for $x_1, \ldots, x_n \sim \mathcal{P}(3)$ as n increases



$n = 38, ..., 240$

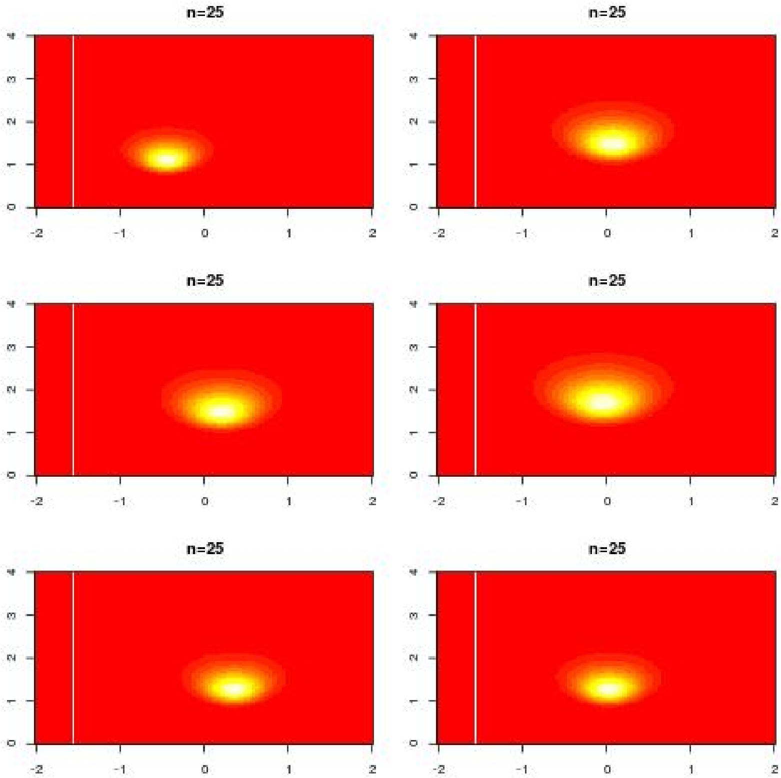
Concentration of likelihood mode around "true" parameter

Likelihood functions for $x_1, \ldots, x_n \sim \mathcal{N}(0,1)$ as n increases



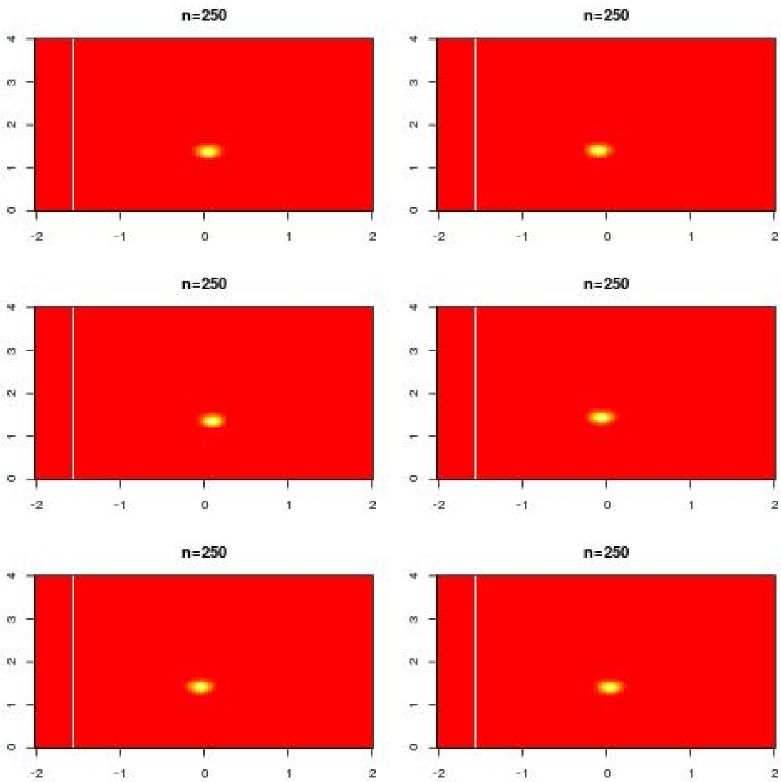
Concentration of likelihood mode around "true" parameter

Likelihood functions for $x_1, \ldots, x_n \sim \mathcal{N}(0, 1)$ as sample varies



Concentration of likelihood mode around “true” parameter

Likelihood functions for $x_1, \ldots, x_n \sim \mathcal{N}(0,1)$ as sample varies



why concentration takes place

Consider

$$x_1, \ldots, x_n \overset{\text{iid}}{\sim} F$$

Then

$$\log \prod_{i=1}^{n} f(x_i|\theta) = \sum_{i=1}^{n} \log f(x_i|\theta)$$

and by LLN

$$^1/_n \sum_{i=1}^{n} \log f(x_i|\theta) \xrightarrow{\mathcal{L}} \int_{\mathcal{X}} \log f(x|\theta)\, dF(x)$$

Lemma

Maximising the likelihood is asymptotically equivalent to minimising the Kullback-Leibler divergence

$$\int_{\mathcal{X}} \log {}^{f(x)}/_{f(x|\theta)}\, dF(x)$$

ⓒ Member of the family closest to true distribution

why concentration takes place

by ◂ LLN

$$1/n \sum_{i=1}^{n} \log f(x_i|\theta) \xrightarrow{\mathcal{L}} \int_{\mathcal{X}} \log f(x|\theta)\, dF(x)$$

Lemma

Maximising the likelihood is asymptotically equivalent to minimising the Kullback-Leibler divergence

$$\int_{\mathcal{X}} \log {}^{f(x)}/_{f(x|\theta)}\, dF(x)$$

ⓒ Member of the family closest to true distribution

Score function

Score function defined by

$$\nabla \log L(\theta|x) = \left(\partial/\partial\theta_1 L(\theta|x), \ldots, \partial/\partial\theta_p L(\theta|x)\right) \Big/ L(\theta|x)$$

Gradient (slope) of likelihood function at point θ

lemma

When $X \sim F_\theta$,

$$\mathbb{E}_\theta[\nabla \log L(\theta|X)] = 0$$

Score function

Score function defined by

$$\nabla \log L(\theta|x) = \left({}^{\partial}/_{\partial\theta_1} L(\theta|x), \ldots, {}^{\partial}/_{\partial\theta_p} L(\theta|x)\right) \Big/ L(\theta|x)$$

Gradient (slope) of likelihood function at point θ

lemma

When $X \sim F_\theta$,

$$\mathbb{E}_\theta[\nabla \log L(\theta|X)] = 0$$

Score function

Score function defined by

$$\nabla \log L(\theta|x) = \left({}^{\partial}\!/\!{}_{\partial\theta_1} L(\theta|x), \ldots, {}^{\partial}\!/\!{}_{\partial\theta_p} L(\theta|x)\right) \big/ L(\theta|x)$$

Gradient (slope) of likelihood function at point θ

lemma

When $X \sim F_\theta$,

$$\mathbb{E}_\theta[\nabla \log L(\theta|X)] = 0$$

Reason:

$$\int_{\mathcal{X}} \nabla \log L(\theta|x)\, dF_\theta(x) = \int_{\mathcal{X}} \nabla L(\theta|x)\, dx = \nabla \int_{\mathcal{X}} dF_\theta(x)$$

Score function

Score function defined by

$$\nabla \log L(\theta|x) = \left({}^{\partial}/_{\partial\theta_1} L(\theta|x), \ldots, {}^{\partial}/_{\partial\theta_p} L(\theta|x) \right) \Big/ L(\theta|x)$$

Gradient (slope) of likelihood function at point θ

lemma

When $X \sim F_\theta$,

$$\mathbb{E}_\theta[\nabla \log L(\theta|X)] = 0$$

Connected with concentration theorem: gradient null on average for true value of parameter

Score function

Score function defined by

$$\nabla \log L(\theta|x) = \left({}^{\partial}/_{\partial\theta_1} L(\theta|x), \ldots, {}^{\partial}/_{\partial\theta_p} L(\theta|x)\right) \Big/ L(\theta|x)$$

Gradient (slope) of likelihood function at point θ

lemma

When $X \sim F_\theta$,

$$\mathbb{E}_\theta[\nabla \log L(\theta|X)] = 0$$

Warning: Not defined for non-differentiable likelihoods, e.g. when support depends on θ

Score function

Score function defined by

$$\nabla \log L(\theta|x) = \left({}^{\partial}/_{\partial\theta_1} L(\theta|x), \ldots, {}^{\partial}/_{\partial\theta_p} L(\theta|x)\right) \big/ L(\theta|x)$$

Gradient (slope) of likelihood function at point θ

lemma

When $X \sim F_\theta$,

$$\mathbb{E}_\theta[\nabla \log L(\theta|X)] = 0$$

Warning (2): Does not imply maximum likelihood estimator is unbiased

Fisher's information matrix

Another notion attributed to Fisher [more likely due to Edgeworth]

Information: covariance matrix of the score vector

$$\mathfrak{I}(\theta) = \mathbb{E}_\theta \left[\nabla \log f(X|\theta) \{\nabla \log f(X|\theta)\}^{\mathrm{T}} \right]$$

Often called Fisher information

Measures curvature of the likelihood surface, which translates as information brought by the data

Sometimes denoted $\mathfrak{I}_X$ to stress dependence on distribution of X

Fisher's information matrix

Another notion attributed to Fisher [more likely due to Edgeworth]

Information: covariance matrix of the score vector

$$\mathfrak{I}(\theta) = \mathbb{E}_\theta \left[\nabla \log f(X|\theta) \left\{ \nabla \log f(X|\theta) \right\}^{\mathrm{T}} \right]$$

Often called Fisher information

Measures curvature of the likelihood surface, which translates as information brought by the data

Sometimes denoted $\mathfrak{I}_X$ to stress dependence on distribution of X

Fisher's information matrix

Second derivative of the log-likelihood as well

lemma

If $L(\theta|x)$ is twice differentiable [as a function of θ]

$$\mathfrak{I}(\theta) = -\mathbb{E}_\theta\left[\nabla^{\mathrm{T}}\nabla \log f(X|\theta)\right]$$

Hence

$$\mathfrak{I}_{ij}(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(X|\theta)\right]$$

Illustrations

Binomial $\mathcal{B}(n, p)$ distribution

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$\partial/\partial p \log f(x|p) = {}^x\!/_p - {}^{n-x}\!/_{1-p}$$

$$\partial^2/\partial p^2 \log f(x|p) = -{}^x\!/_{p^2} - {}^{n-x}\!/_{(1-p)^2}$$

Hence

$$\begin{aligned}\mathfrak{I}(p) &= {}^{np}\!/_{p^2} + {}^{n-np}\!/_{(1-p)^2} \\ &= {}^{n}\!/_{p(1-p)}\end{aligned}$$

Illustrations

Multinomial $\mathcal{M}(n; p_1, \ldots, p_k)$ distribution

$$
\begin{aligned}
f(x|p) &= \binom{n}{x_1 \cdots x_k} p_1^{x_1} \cdots p_k^{x_k} \\
\partial/\partial p_i \log f(x|p) &= x_i/p_i - x_k/p_k \\
\partial^2/\partial p_i \partial p_j \log f(x|p) &= - x_k/p_k^2 \\
\partial^2/\partial p_i^2 \log f(x|p) &= - x_i/p_i^2 - x_k/p_k^2
\end{aligned}
$$

Hence

$$
\mathfrak{I}(p) = n \begin{pmatrix} 1/p_1 + 1/p_k & \cdots & 1/p_k \\ 1/p_k & \cdots & 1/p_k \\ & \ddots & \\ 1/p_k & \cdots & 1/p_{k-1} + 1/p_k \end{pmatrix}
$$

Illustrations

Multinomial $\mathcal{M}(n; p_1, \ldots, p_k)$ distribution

$$\begin{aligned}
f(x|p) &= \binom{n}{x_1 \cdots x_k} p_1^{x_1} \cdots p_k^{x_k} \\
\partial/\partial p_i \log f(x|p) &= x_i/p_i - x_k/p_k \\
\partial^2/\partial p_i \partial p_j \log f(x|p) &= -x_k/p_k^2 \\
\partial^2/\partial p_i^2 \log f(x|p) &= -x_i/p_i^2 - x_k/p_k^2
\end{aligned}$$

and

$$\mathfrak{I}(p)^{-1} = 1/n \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_{k-1} \\ -p_1p_2 & p_2(1-p_2) & \cdots & -p_2p_{k-1} \\ & \ddots & \ddots & \\ -p_1p_{k-1} & -p_2p_{k-1} & \cdots & p_{k-1}(1-p_{k-1}) \end{pmatrix}$$

Illustrations

Normal $\mathcal{N}(\mu, \sigma^2)$ distribution

$$f(x|\theta) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}\exp\left\{-{(x-\mu)^2}/{2\sigma^2}\right\} \quad {\partial}/{\partial\mu}\log f(x|\theta) = {x-\mu}/{\sigma^2}$$
$${\partial}/{\partial\sigma}\log f(x|\theta) = -{1}/{\sigma} + {(x-\mu)^2}/{\sigma^3} \quad {\partial^2}/{\partial\mu^2}\log f(x|\theta) = -{1}/{\sigma^2}$$
$${\partial^2}/{\partial\mu\partial\sigma}\log f(x|\theta) = -2\,{x-\mu}/{\sigma^3} \quad {\partial^2}/{\partial\sigma^2}\log f(x|\theta) = {1}/{\sigma^2} - 3\,{(x-\mu)^2}/{\sigma^4}$$

Hence

$$\mathfrak{I}(\theta) = {1}/{\sigma^2}\begin{pmatrix}1 & 0\\ 0 & 2\end{pmatrix}$$

Properties

Additive features translating as accumulation of information:

- if X and Y are independent, $\mathfrak{I}_X(\theta) + \mathfrak{I}_Y(\theta) = \mathfrak{I}_{(X,Y)}(\theta)$
- $\mathfrak{I}_{X_1,\ldots,X_n}(\theta) = n\mathfrak{I}_{X_1}(\theta)$
- if $X = T(Y)$ and $Y = S(X)$, $\mathfrak{I}_X(\theta) = \mathfrak{I}_Y(\theta)$
- if $X = T(Y)$, $\mathfrak{I}_X(\theta) \leqslant \mathfrak{I}_Y(\theta)$

If $\eta = \Psi(\theta)$ is a bijective transform, change of parameterisation:

$$\mathfrak{I}(\theta) = \left\{\frac{\partial\eta}{\partial\theta}\right\}^{\mathrm{T}} \mathfrak{I}(\eta) \left\{\frac{\partial\eta}{\partial\theta}\right\}$$

"In information geometry, this is seen as a change of coordinates on a Riemannian manifold, and the intrinsic properties of curvature are unchanged under different parametrizations. In general, the Fisher information matrix provides a Riemannian metric (more precisely, the Fisher-Rao metric)." [Wikipedia]

Properties

Additive features translating as accumulation of information:

- if X and Y are independent, $\mathfrak{I}_X(\theta) + \mathfrak{I}_Y(\theta) = \mathfrak{I}_{(X,Y)}(\theta)$
- $\mathfrak{I}_{X_1,\ldots,X_n}(\theta) = n\mathfrak{I}_{X_1}(\theta)$
- if $X = T(Y)$ and $Y = S(X)$, $\mathfrak{I}_X(\theta) = \mathfrak{I}_Y(\theta)$
- if $X = T(Y)$, $\mathfrak{I}_X(\theta) \leqslant \mathfrak{I}_Y(\theta)$

If $\eta = \Psi(\theta)$ is a bijective transform, change of parameterisation:

$$\mathfrak{I}(\theta) = \left\{\frac{\partial\eta}{\partial\theta}\right\}^{\mathrm{T}} \mathfrak{I}(\eta) \left\{\frac{\partial\eta}{\partial\theta}\right\}$$

"In information geometry, this is seen as a change of coordinates on a Riemannian manifold, and the intrinsic properties of curvature are unchanged under different parametrizations. In general, the Fisher information matrix provides a Riemannian metric (more precisely, the Fisher-Rao metric)." [Wikipedia]

Properties

If $\eta = \Psi(\theta)$ is a bijective transform, change of parameterisation:

$$\mathfrak{I}(\theta) = \left\{\frac{\partial\eta}{\partial\theta}\right\}^{\mathrm{T}} \mathfrak{I}(\eta) \left\{\frac{\partial\eta}{\partial\theta}\right\}$$

> *"In information geometry, this is seen as a change of coordinates on a Riemannian manifold, and the intrinsic properties of curvature are unchanged under different parametrizations. In general, the Fisher information matrix provides a Riemannian metric (more precisely, the Fisher-Rao metric)." [Wikipedia]*

Approximations

Back to the Kullback–Leibler divergence

$$\mathfrak{D}(\theta', \theta) = \int_{\mathcal{X}} f(x|\theta') \log {}^{f(x|\theta')}/_{f(x|\theta)} \, dx$$

Using a second degree Taylor expansion

$$\begin{aligned}\log f(x|\theta) = \log f(x|\theta') + (\theta - \theta')^{\mathrm{T}} \nabla \log f(x|\theta') \\ + \frac{1}{2}(\theta - \theta')^{\mathrm{T}} \nabla \nabla^{\mathrm{T}} \log f(x|\theta')(\theta - \theta') + o(||\theta - \theta'||^2)\end{aligned}$$

approximation of divergence:

$$\mathfrak{D}(\theta', \theta) \approx \frac{1}{2}(\theta - \theta')^{\mathrm{T}} \mathfrak{I}(\theta')(\theta - \theta')$$

[Exercise: show this is exact in the normal case]

Approximations

Back to the Kullback–Leibler divergence

$$\mathfrak{D}(\theta', \theta) = \int_{\mathcal{X}} f(x|\theta') \log {}^{f(x|\theta')}/_{f(x|\theta)} \, dx$$

Using a second degree Taylor expansion

$$\log f(x|\theta) = \log f(x|\theta') + (\theta - \theta')^{\mathrm{T}} \nabla \log f(x|\theta') + \frac{1}{2}(\theta - \theta')^{\mathrm{T}} \nabla\nabla^{\mathrm{T}} \log f(x|\theta')(\theta - \theta') + o(||\theta - \theta'||^2)$$

approximation of divergence:

$$\mathfrak{D}(\theta', \theta) \approx \frac{1}{2}(\theta - \theta')^{\mathrm{T}} \mathfrak{I}(\theta')(\theta - \theta')$$

[Exercise: show this is exact in the normal case]

Approximations

Back to the Kullback–Leibler divergence

$$\mathfrak{D}(\theta', \theta) = \int_{\mathcal{X}} f(x|\theta') \log f(x|\theta')/f(x|\theta) \; dx$$

Using a second degree Taylor expansion

$$\begin{aligned}\log f(x|\theta) = \log f(x|\theta') &+ (\theta - \theta')^{\mathrm{T}} \nabla \log f(x|\theta') \\ &+ \frac{1}{2}(\theta - \theta')^{\mathrm{T}} \nabla\nabla^{\mathrm{T}} \log f(x|\theta')(\theta - \theta') + o(\|\theta - \theta'\|^2)\end{aligned}$$

approximation of divergence:

$$\mathfrak{D}(\theta', \theta) \approx \frac{1}{2}(\theta - \theta')^{\mathrm{T}} \mathfrak{I}(\theta')(\theta - \theta')$$

[Exercise: show this is exact in the normal case]

First CLT

Central limit law of the score vector

Given $X_1, \ldots, X_n$ i.i.d. $f(x|\theta)$,

$$1/\sqrt{n}\nabla \log L(\theta|X_1, \ldots, X_n) \approx \mathcal{N}\left(0, \mathfrak{I}_{X_1}(\theta)\right)$$

[at the "true" θ]

Notation $\mathfrak{I}_1(\theta)$ stands for $\mathfrak{I}_{X_1}(\theta)$ and indicates information associated with a single observation

First CLT

Central limit law of the score vector

Given $X_1, \ldots, X_n$ i.i.d. $f(x|\theta)$,

$$^1/\sqrt{n} \nabla \log L(\theta|X_1, \ldots, X_n) \approx \mathcal{N}\left(0, \mathfrak{I}_{X_1}(\theta)\right)$$

[at the "true" θ]

Notation $\mathfrak{I}_1(\theta)$ stands for $\mathfrak{I}_{X_1}(\theta)$ and indicates information associated with a single observation

Sufficiency

What if a transform of the sample

$$S(X_1, \dots, X_n)$$

contains all the information, i.e.

$$\mathfrak{I}_{(X_1,\dots,X_n)}(\theta) = \mathfrak{I}_{S(X_1,\dots,X_n)}(\theta)$$

uniformly in θ?

In this case $S(\cdot)$ is called a sufficient statistic [because it is sufficient to know the value of $S(x_1, \dots, x_n)$ to get complete information]

[A statistic is an arbitrary transform of the data $X_1, \dots, X_n$]

Sufficiency

What if a transform of the sample

$$S(X_1, \ldots, X_n)$$

contains all the information, i.e.

$$\mathfrak{I}_{(X_1,\ldots,X_n)}(\theta) = \mathfrak{I}_{S(X_1,\ldots,X_n)}(\theta)$$

uniformly in θ?

In this case $S(\cdot)$ is called a sufficient statistic [because it is sufficient to know the value of $S(x_1, \ldots, x_n)$ to get complete information]

[A statistic is an arbitrary transform of the data $X_1, \ldots, X_n$]

Sufficiency

What if a transform of the sample

$$S(X_1, \ldots, X_n)$$

contains all the information, i.e.

$$\mathfrak{I}_{(X_1,\ldots,X_n)}(\theta) = \mathfrak{I}_{S(X_1,\ldots,X_n)}(\theta)$$

uniformly in θ?

In this case $S(\cdot)$ is called a sufficient statistic [because it is sufficient to know the value of $S(x_1, \ldots, x_n)$ to get complete information]

[A statistic is an arbitrary transform of the data $X_1, \ldots, X_n$]

Sufficiency (bis)

Alternative definition:

If $(X_1, \ldots, X_n) \sim f(x_1, \ldots, x_n|\theta)$ and if $T = S(X_1, \ldots, X_n)$ is such that the distribution of $(X_1, \ldots, X_n)$ conditional on T does not depend on θ, then $S(\cdot)$ is a sufficient statistic

Factorisation theorem

$S(\cdot)$ is a sufficient statistic if and only if

$$f(x_1, \ldots, x_n|\theta) = g(S(x_1, \ldots, x_n)|\theta) \times h(x_1, \ldots, x_n)$$

another notion due to Fisher

Sufficiency (bis)

Alternative definition:

If $(X_1, \ldots, X_n) \sim f(x_1, \ldots, x_n|\theta)$ and if $T = S(X_1, \ldots, X_n)$ is such that the distribution of $(X_1, \ldots, X_n)$ conditional on T does not depend on θ, then $S(\cdot)$ is a sufficient statistic

Factorisation theorem

$S(\cdot)$ is a sufficient statistic if and only if

$$f(x_1, \ldots, x_n|\theta) = g(S(x_1, \ldots, x_n)|\theta) \times h(x_1, \ldots, x_n)$$

another notion due to Fisher

Sufficiency (bis)

Alternative definition:

If $(X_1, \ldots, X_n) \sim f(x_1, \ldots, x_n|\theta)$ and if $T = S(X_1, \ldots, X_n)$ is such that the distribution of $(X_1, \ldots, X_n)$ conditional on T does not depend on θ, then $S(\cdot)$ is a sufficient statistic

Factorisation theorem

$S(\cdot)$ is a sufficient statistic if and only if

$$f(x_1, \ldots, x_n|\theta) = g(S(x_1, \ldots, x_n)|\theta) \times h(x_1, \ldots, x_n)$$

another notion due to Fisher

Illustrations

Uniform $\mathcal{U}(0, \theta)$ distribution

$$L(\theta|x_1, \dots, x_n) = \theta^{-n} \prod_{i=1}^{n} \mathbb{I}_{(0,\theta)}(x_i) = \theta^{-n} \mathbb{I}_{\theta > \max_i x_i}$$

Hence

$$S(X_1, \dots, X_n) = \max_i X_i = X_{(n)}$$

is sufficient

Illustrations

Bernoulli $\mathcal{B}(p)$ distribution

$$L(p|x_1,\ldots,x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{n-x_i} = \{p/1-p\}^{\sum_i x_i}(1-p)^n$$

Hence

$$S(X_1,\ldots,X_n) = \overline{X}_n$$

is sufficient

Illustrations

Normal $\mathcal{N}(\mu, \sigma^2)$ distribution

$$
\begin{aligned}
L(\mu, \sigma | x_1, \ldots, x_n) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\{-{}^{(x_i-\mu)^2}/_{2\sigma^2}\} \\
&= \frac{1}{\{2\pi\sigma^2\}^{n/2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \bar{x}_n + \bar{x}_n - \mu)^2\right\} \\
&= \frac{1}{\{2\pi\sigma^2\}^{n/2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (\bar{x}_n - \mu)^2\right\}
\end{aligned}
$$

Hence

$$
S(X_1, \ldots, X_n) = \left(\overline{X}_n, \sum_{i=1}^{n} (X_i - \overline{X}_n)^2\right)
$$

is sufficient

Sufficiency and exponential families

Both previous examples belong to exponential families

$$f(x|\theta) = h(x) \exp\left\{T(\theta)^T S(x) - \tau(\theta)\right\}$$

Generic property of exponential families:

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} h(x_i) \exp\left\{T(\theta)^T \sum_{i=1}^{n} S(x_i) - n\tau(\theta)\right\}$$

lemma

For an exponential family with summary statistic $S(\cdot)$, the statistic

$$S(X_1, \ldots, X_n) = \sum_{i=1}^{n} S(X_i)$$

is sufficient

Sufficiency and exponential families

Both previous examples belong to exponential families

$$f(x|\theta) = h(x) \exp\left\{T(\theta)^{\mathrm{T}} S(x) - \tau(\theta)\right\}$$

Generic property of exponential families:

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} h(x_i) \exp\left\{T(\theta)^{\mathrm{T}} \sum_{i=1}^{n} S(x_i) - n\tau(\theta)\right\}$$

lemma

For an exponential family with summary statistic $S(\cdot)$, the statistic

$$S(X_1, \ldots, X_n) = \sum_{i=1}^{n} S(X_i)$$

is sufficient

Sufficiency as a rare feature

Nice property reducing the data to a low dimension transform but...

How frequent is it within the collection of probability distributions?

Very rare as essentially restricted to exponential families

[Pitman-Koopman-Darmois theorem]

with the exception of parameter-dependent families like $\mathcal{U}(0, \theta)$

Sufficiency as a rare feature

Nice property reducing the data to a low dimension transform but...

How frequent is it within the collection of probability distributions?

Very rare as essentially restricted to exponential families

[Pitman-Koopman-Darmois theorem]

with the exception of parameter-dependent families like $\mathcal{U}(0, \theta)$

Sufficiency as a rare feature

Nice property reducing the data to a low dimension transform but...

How frequent is it within the collection of probability distributions?

Very rare as essentially restricted to exponential families [Pitman-Koopman-Darmois theorem]

with the exception of parameter-dependent families like $\mathcal{U}(0, \theta)$

Pitman-Koopman-Darmois characterisation

If $X_1, \ldots, X_n$ *are iid random variables from a density* $f(\cdot|\theta)$ *whose support does not depend on* θ *and verifying the property that there exists an integer* n_0 *such that, for* $n \geqslant n_0$*, there is a sufficient statistic* $S(X_1, \ldots, X_n)$ *with fixed [in* n*] dimension, then* $f(\cdot|\theta)$ *belongs to an exponential family*

[Factorisation theorem]

Note: Darmois published this result in 1935 [in French] and Koopman and Pitman in 1936 [in English] but Darmois is generally omitted from the theorem... Fisher proved it for one-D sufficient statistics in 1934

Pitman-Koopman-Darmois characterisation

If $X_1, \ldots, X_n$ are iid random variables from a density $f(\cdot|\theta)$ whose support does not depend on θ and verifying the property that there exists an integer n_0 such that, for $n \geqslant n_0$, there is a sufficient statistic $S(X_1, \ldots, X_n)$ with fixed [in n] dimension, then $f(\cdot|\theta)$ belongs to an exponential family

[Factorisation theorem]

Note: Darmois published this result in 1935 [in French] and Koopman and Pitman in 1936 [in English] but Darmois is generally omitted from the theorem... Fisher proved it for one-D sufficient statistics in 1934

Minimal sufficiency

Multiplicity of sufficient statistics, e.g., $S'(x) = (S(x), U(x))$ remains sufficient when $S(\cdot)$ is sufficient

Search of a most concentrated summary:

Minimal sufficiency

A sufficient statistic $S(\cdot)$ is minimal sufficient if it is a function of any other sufficient statistic

Lemma

For a minimal exponential family representation

$$f(x|\theta) = h(x) \exp\left\{T(\theta)^{T} S(x) - \tau(\theta)\right\}$$

$S(X_1) + \ldots + S(X_n)$ is minimal sufficient

Minimal sufficiency

Multiplicity of sufficient statistics, e.g., $S'(x) = (S(x), U(x))$ remains sufficient when $S(\cdot)$ is sufficient

Search of a most concentrated summary:

Minimal sufficiency

A sufficient statistic $S(\cdot)$ is minimal sufficient if it is a function of any other sufficient statistic

Lemma

For a minimal exponential family representation

$$f(x|\theta) = h(x) \exp\left\{T(\theta)^{\mathrm{T}} S(x) - \tau(\theta)\right\}$$

$S(X_1) + \ldots + S(X_n)$ is minimal sufficient

Ancillarity

Opposite of sufficiency:

Ancillarity

When $X_1, \ldots, X_n$ are iid random variables from a density $f(\cdot|\theta)$, a statistic $A(\cdot)$ is ancillary if $A(X_1, \ldots, X_n)$ has a distribution that does not depend on θ

Useless?! Not necessarily, as conditioning upon $A(X_1, \ldots, X_n)$ leads to more precision and efficiency:

Use of $F_\theta(x_1, \ldots, x_n|A(x_1, \ldots, x_n))$ instead of $F_\theta(x_1, \ldots, x_n)$

Notion of maximal ancillary statistic

Ancillarity

Opposite of sufficiency:

Ancillarity

When $X_1, \ldots, X_n$ are iid random variables from a density $f(\cdot|\theta)$, a statistic $A(\cdot)$ is ancillary if $A(X_1, \ldots, X_n)$ has a distribution that does not depend on θ

Useless?! Not necessarily, as conditioning upon $A(X_1, \ldots, X_n)$ leads to more precision and efficiency:

Use of $F_\theta(x_1, \ldots, x_n | A(x_1, \ldots, x_n))$ instead of $F_\theta(x_1, \ldots, x_n)$

Notion of maximal ancillary statistic

Ancillarity

Opposite of sufficiency:

Ancillarity

When $X_1, \ldots, X_n$ are iid random variables from a density $f(\cdot|\theta)$, a statistic $A(\cdot)$ is ancillary if $A(X_1, \ldots, X_n)$ has a distribution that does not depend on θ

Useless?! Not necessarily, as conditioning upon $A(X_1, \ldots, X_n)$ leads to more precision and efficiency:

Use of $F_\theta(x_1, \ldots, x_n | A(x_1, \ldots, x_n))$ instead of $F_\theta(x_1, \ldots, x_n)$

Notion of maximal ancillary statistic

Illustrations

1. If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{U}(0, \theta)$, $A(X_1, \ldots, X_n) = (X_1, \ldots, X_n)/X_{(n)}$ is ancillary
2. If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$,
$$A(X_1, \ldots, X_n) = \frac{(X_1 - \overline{X}_n, \ldots, X_n - \overline{X}_n)}{\sum_{i=1}^n (X_i - \overline{X}_n)^2)}$$
is ancillary
3. If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x|\theta)$, $\text{rank}(X_1, \ldots, X_n)$ is ancillary

```
> x=rnorm(10)
> rank(x)
 [1]  7  4  1  5  2  6  8  9 10  3
```

[see, e.g., rank tests]

Basu's theorem

Completeness

When $X_1, \ldots, X_n$ are iid random variables from a density $f(\cdot|\theta)$, a statistic $A(\cdot)$ is complete if the only function Ψ such that $\mathbb{E}_\theta[\Psi(A(X_1, \ldots, X_n))] = 0$ for all θ's is the null function

Let $X = (X_1, \ldots, X_n)$ be a random sample from $f(\cdot|\theta)$ where $\theta \in \Theta$. If V is an ancillary statistic, and T is complete and sufficient for θ then T and V are independent with respect to $f(\cdot|\theta)$ for all $\theta \in \Theta$.

[Basu, 1955]

Basu's theorem

Completeness

When $X_1, \ldots, X_n$ are iid random variables from a density $f(\cdot|\theta)$, a statistic $A(\cdot)$ is complete if the only function Ψ such that $\mathbb{E}_\theta[\Psi(A(X_1, \ldots, X_n))] = 0$ for all θ's is the null function

Let $X = (X_1, \ldots, X_n)$ be a random sample from $f(\cdot|\theta)$ where $\theta \in \Theta$. If V is an ancillary statistic, and T is complete and sufficient for θ then T and V are independent with respect to $f(\cdot|\theta)$ for all $\theta \in \Theta$.

[Basu, 1955]

some examples

Example 1

If $X = (X_1, \ldots, X_n)$ is a random sample from the Normal distribution $\mathcal{N}(\mu, \sigma^2)$ when σ is known, $\bar{X}_n = {}^1\!/\!_n \sum_{i=1}^n X_i$ is sufficient and complete, while $(X_1 - \bar{X}_n, \ldots, X_n - \bar{X}_n)$ is ancillary, hence independent from $\bar{X}_n$.

counter-Example 2

Let N be an integer-valued random variable with known pdf $(\pi_1, \pi_2, \ldots)$. And let $S|N = n \sim \mathcal{B}(n, p)$ with unknown p. Then (N, S) is minimal sufficient and N is ancillary.

some examples

Example 1

If $X = (X_1, \ldots, X_n)$ is a random sample from the Normal distribution $\mathcal{N}(\mu, \sigma^2)$ when σ is known, $\bar{X}_n = {}^1\!/_n \sum_{i=1}^n X_i$ is sufficient and complete, while $(X_1 - \bar{X}_n, \ldots, X_n - \bar{X}_n)$ is ancillary, hence independent from $\bar{X}_n$.

counter-Example 2

Let N be an integer-valued random variable with known pdf $(\pi_1, \pi_2, \ldots)$. And let $S|N = n \sim \mathcal{B}(n, p)$ with unknown p. Then (N, S) is minimal sufficient and N is ancillary.

more counterexamples

counter-Example 3

If $X = (X_1, \ldots, X_n)$ is a random sample from the double exponential distribution $f(x|\theta) = 2\exp\{-|x-\theta|\}$, $(X_{(1)}, \ldots, X_{(n)})$ is minimal sufficient but not complete since $X_{(n)} - X_{(1)}$ is ancillary and with fixed expectation.

counter-Example 4

If X is a random variable from the Uniform $\mathcal{U}(\theta, \theta+1)$ distribution, X and $[X]$ are independent, but while X is complete and sufficient, $[X]$ is not ancillary.

more counterexamples

counter-Example 3

If $X = (X_1, \ldots, X_n)$ is a random sample from the double exponential distribution $f(x|\theta) = 2\exp\{-|x-\theta|\}$, $(X_{(1)}, \ldots, X_{(n)})$ is minimal sufficient but not complete since $X_{(n)} - X_{(1)}$ is ancillary and with fixed expectation.

counter-Example 4

If X is a random variable from the Uniform $\mathcal{U}(\theta, \theta+1)$ distribution, X and $[X]$ are independent, but while X is complete and sufficient, $[X]$ is not ancillary.

last counterexample

Let X be distributed as

x	-5	-4	-3	-2	-1	1	2	3	4	5
p_x	$\alpha' p^2 q$	$\alpha' p q^2$	$p^3/2$	$q^3/2$	$\gamma' p q$	$\gamma' p q$	$q^3/2$	$p^3/2$	$\alpha p q^2$	$\alpha p^2 q$

with

$$\alpha + \alpha' = \gamma + \gamma' = 2/3$$

known and $q = 1 - p$. Then

- $T = |X|$ is minimal sufficient
- $V = \mathbb{I}(X > 0)$ is ancillary
- if $\alpha' \neq \alpha$ T and V are not independent
- T is complete for two-valued functions

[Lehmann, 1981]

Point estimation, estimators and estimates

When given a parametric family $f(\cdot|\theta)$ and a sample supposedly drawn from this family

$$(X_1, \dots, X_N) \stackrel{\text{iid}}{\sim} f(x|\theta)$$

1. an estimator of θ is a statistic $T(X_1, \dots, X_N)$ or $\hat{\theta}_n$ providing a [reasonable] substitute for the unknown value θ.
2. an estimate of θ is the value of the estimator for a given [realised] sample, $T(x_1, \dots, x_n)$

Example: For a Normal $\mathcal{N}(\mu, \sigma^2)$ sample $X_1, \dots, X_N$,

$$T(X_1, \dots, X_N) = \hat{\mu}_n = \overline{X}_N$$

is an estimator of μ and $\hat{\mu}_N = 2.014$ is an estimate

Point estimation, estimators and estimates

When given a parametric family $f(\cdot|\theta)$ and a sample supposedly drawn from this family

$$(X_1, \ldots, X_N) \overset{\text{iid}}{\sim} f(x|\theta)$$

1. an estimator of θ is a statistic $T(X_1, \ldots, X_N)$ or $\hat{\theta}_n$ providing a [reasonable] substitute for the unknown value θ.
2. an estimate of θ is the value of the estimator for a given [realised] sample, $T(x_1, \ldots, x_n)$

Example: For a Normal $\mathcal{N}(\mu, \sigma^2)$ sample $X_1, \ldots, X_N$,

$$T(X_1, \ldots, X_N) = \hat{\mu}_n = \overline{X}_N$$

is an estimator of μ and $\hat{\mu}_N = 2.014$ is an estimate

Rao–Blackwell Theorem

If $\delta(\cdot)$ is an estimator of θ and $T = T(X)$ is a sufficient statistic, then

$$\delta_1(X) = \mathbb{E}_\theta[\delta(X)|T]$$

has a smaller variance than $\delta(\cdot)$

$$\text{var}_\theta(\delta_1(X)) \leqslant \text{var}_\theta(\delta(X))$$

[Rao, 1945; Blackwell, 1947]

mean squared error of Rao–Blackwell estimator does not exceed that of original estimator

Rao–Blackwell Theorem

If $\delta(\cdot)$ is an estimator of θ and $T = T(X)$ is a sufficient statistic, then

$$\delta_1(X) = \mathbb{E}_\theta[\delta(X)|T]$$

has a smaller variance than $\delta(\cdot)$

$$\text{var}_\theta(\delta_1(X)) \leqslant \text{var}_\theta(\delta(X))$$

[Rao, 1945; Blackwell, 1947]

mean squared error of Rao–Blackwell estimator does not exceed that of original estimator

Lehmann–Scheffé Theorem

Estimator δ_0

- unbiased for $\mathbb{E}_\theta[\delta X] = \Psi(\theta)$
- depends on data only through complete, sufficient statistic $S(X)$

is the unique best unbiased estimator of $\Psi(\theta)$

[Lehmann & Scheffé, 1955]

For any unbiased estimator $\delta(\cdot)$ of $\Psi(\theta)$,

$$\delta_0(X) = \mathbb{E}_\theta[\delta(X)|S(X)]$$

Lehmann–Scheffé Theorem

Estimator δ_0

- unbiased for $\mathbb{E}_\theta[\delta X] = \Psi(\theta)$
- depends on data only through complete, sufficient statistic $S(X)$

is the unique best unbiased estimator of $\Psi(\theta)$

[Lehmann & Scheffé, 1955]

For any unbiased estimator $\delta(\cdot)$ of $\Psi(\theta)$,

$$\delta_0(X) = \mathbb{E}_\theta[\delta(X)|S(X)]$$

[Fréchet–Darmois–]Cramér–Rao bound

If $\hat{\theta}$ is an estimator of $\theta \in \mathbb{R}$ with bias

$$b(\theta) = \mathbb{E}_\theta[\hat{\theta}] - \theta$$

then

$$\mathrm{var}_\theta(\hat{\theta}) \geqslant \frac{[1 + b'(\theta)]^2}{\mathfrak{I}(\theta)}$$

[Fréchet, 1943; Darmois, 1945; Rao, 1945; Cramér, 1946]

variance of any unbiased estimator at least as high as inverse Fisher information

[Fréchet–Darmois–]Cramér–Rao bound

If $\hat{\theta}$ is an estimator of $\theta \in \mathbb{R}$ with bias

$$b(\theta) = \mathbb{E}_\theta[\hat{\theta}] - \theta$$

then

$$\mathrm{var}_\theta(\hat{\theta}) \geqslant \frac{[1 + b'(\theta)]^2}{\mathfrak{I}(\theta)}$$

[Fréchet, 1943; Darmois, 1945; Rao, 1945; Cramér, 1946]

variance of any unbiased estimator at least as high as inverse Fisher information

Single parameter proof

If $\delta = \delta(X)$ unbiased estimator of $\Psi(\theta)$, then

$$\mathsf{var}_\theta(\delta) \geqslant \frac{[\Psi'(\theta)]^2}{\mathfrak{I}(\theta)}$$

Take score $Z = \frac{\partial}{\partial\theta} \log f(X|\theta)$. Then

$$\mathsf{cov}_\theta(Z, \delta) = \mathbb{E}_\theta[\delta(X)Z] = \Psi'(\theta)$$

And Cauchy-Schwarz implies

$$\mathsf{cov}_\theta(Z, \delta)^2 \leqslant \mathsf{var}_\theta(\delta)\mathsf{var}_\theta(Z) = \mathsf{var}_\theta(\delta)\mathfrak{I}(\theta)$$

Warning: unbiasedness may be harmful

Unbiasedness is not an ultimate property!

- most transforms $h(\theta)$ do not allow for unbiased estimators
- no bias may imply large variance
- efficient estimators may be biased (MLE)
- existence of UNMVUE restricted to exponential families
- Cramér–Rao bound inaccessible outside exponential families

Maximum likelihood principle

Given the concentration property of the likelihood function, reasonable choice of estimator as mode:

MLE

A maximum likelihood estimator (MLE) $\hat{\theta}_N$ satisfies

$$L(\hat{\theta}_N|X_1,\ldots,X_N) \geqslant L(\theta_N|X_1,\ldots,X_N) \qquad \text{for all } \theta \in \Theta$$

Under regularity of $L(\cdot|X_1,\ldots,X_N)$, MLE also solution of the likelihood equations

$$\nabla \log L(\hat{\theta}_N|X_1,\ldots,X_N) = 0$$

Warning: $\hat{\theta}_N$ is not most likely value of θ but makes observation $(x_1,\ldots,x_N)$ most likely...

Maximum likelihood principle

Given the concentration property of the likelihood function, reasonable choice of estimator as mode:

MLE

A maximum likelihood estimator (MLE) $\hat{\theta}_N$ satisfies

$$L(\hat{\theta}_N|X_1,\ldots,X_N) \geqslant L(\theta_N|X_1,\ldots,X_N) \qquad \text{for all } \theta \in \Theta$$

Under regularity of $L(\cdot|X_1,\ldots,X_N)$, MLE also solution of the likelihood equations

$$\nabla \log L(\hat{\theta}_N|X_1,\ldots,X_N) = 0$$

Warning: $\hat{\theta}_N$ is not most likely value of θ but makes observation $(x_1,\ldots,x_N)$ most likely...

Maximum likelihood principle

Given the concentration property of the likelihood function, reasonable choice of estimator as mode:

MLE

A maximum likelihood estimator (MLE) $\hat{\theta}_N$ satisfies

$$L(\hat{\theta}_N|X_1,\ldots,X_N) \geqslant L(\theta_N|X_1,\ldots,X_N) \qquad \text{for all } \theta \in \Theta$$

Under regularity of $L(\cdot|X_1,\ldots,X_N)$, MLE also solution of the likelihood equations

$$\nabla \log L(\hat{\theta}_N|X_1,\ldots,X_N) = 0$$

Warning: $\hat{\theta}_N$ is not most likely value of θ but makes observation $(x_1,\ldots,x_N)$ most likely...

Maximum likelihood invariance

Principle independent of parameterisation:

If $\xi = h(\theta)$ is a one-to-one transform of θ, then

$$\hat{\xi}_N^{\mathsf{MLE}} = h(\hat{\theta}_N^{\mathsf{MLE}})$$

[estimator of transform = transform of estimator]

By extension, if $\xi = h(\theta)$ is any transform of θ, then

$$\hat{\xi}_N^{\mathsf{MLE}} = h(\hat{\theta}_n^{\mathsf{MLE}})$$

Alternative of *profile likelihoods* distinguishing between parameters of interest and nuisance parameters

Maximum likelihood invariance

Principle independent of parameterisation:

If $\xi = h(\theta)$ is a one-to-one transform of θ, then

$$\hat{\xi}_N^{\text{MLE}} = h(\hat{\theta}_N^{\text{MLE}})$$

[estimator of transform = transform of estimator]

By extension, if $\xi = h(\theta)$ is any transform of θ, then

$$\hat{\xi}_N^{\text{MLE}} = h(\hat{\theta}_n^{\text{MLE}})$$

Alternative of *profile likelihoods* distinguishing between parameters of interest and nuisance parameters

Unicity of maximum likelihood estimate

Depending on regularity of $L(\cdot|x_1, \ldots, x_N)$, there may be

1. an a.s. unique MLE $\hat{\theta}_n^{\mathsf{MLE}}$
2.
3.

1. Case of $x_1, \ldots, x_n \sim \mathcal{N}(\mu, 1)$
2.
3. [with $\tau = +\infty$]

Unicity of maximum likelihood estimate

Depending on regularity of $L(\cdot|x_1, \ldots, x_N)$, there may be

1.
2. several or an infinity of MLE's [or of solutions to likelihood equations]
3.

1.
2. Case of $x_1, \ldots, x_n \sim \mathcal{N}(\mu_1 + \mu_2, 1)$ [and mixtures of normal]
3. [with $\tau = +\infty$]

Unicity of maximum likelihood estimate

Depending on regularity of $L(\cdot|x_1, \ldots, x_N)$, there may be

1.
2.
3. no MLE at all

1.
2.
3. Case of $x_1, \ldots, x_n \sim \mathcal{N}(\mu_i, \tau^{-2})$ [with $\tau = +\infty$]

Unicity of maximum likelihood estimate

Consequence of standard differential calculus results on $\ell(\theta) = \log L(\theta|x_1, \ldots, x_n)$:

lemma

If Θ is connected and open, and if $\ell(\cdot)$ is twice-differentiable with

$$\lim_{\theta \to \partial\Theta} \ell(\theta) < +\infty$$

and if $H(\theta) = \nabla\nabla^T \ell(\theta)$ is positive definite at all solutions of the likelihood equations, then $\ell(\cdot)$ has a unique global maximum

Limited appeal because excluding local maxima

Unicity of maximum likelihood estimate

Consequence of standard differential calculus results on $\ell(\theta) = \log L(\theta|x_1, \ldots, x_n)$:

lemma

If Θ is connected and open, and if $\ell(\cdot)$ is twice-differentiable with

$$\lim_{\theta \to \partial\Theta} \ell(\theta) < +\infty$$

and if $H(\theta) = \nabla\nabla^T \ell(\theta)$ is positive definite at all solutions of the likelihood equations, then $\ell(\cdot)$ has a unique global maximum

Limited appeal because excluding local maxima

Unicity of MLE for exponential families

lemma

If $f(\cdot|\theta)$ is a minimal exponential family

$$f(x|\theta) = h(x) \exp\left\{T(\theta)^{\mathrm{T}} S(x) - \tau(\theta)\right\}$$

with $T(\cdot)$ one-to-one and twice differentiable over Θ, if Θ is open, and if there is at least one solution to the likelihood equations, then it is the unique MLE

Likelihood equation is equivalent to $S(x) = \mathbb{E}_\theta[S(X)]$

Unicity of MLE for exponential families

lemma

If Θ is connected and open, and if $\ell(\cdot)$ is twice-differentiable with

$$\lim_{\theta \to \partial\Theta} \ell(\theta) < +\infty$$

and if $\mathrm{H}(\theta) = \nabla\nabla^{\mathrm{T}}\ell(\theta)$ is positive definite at all solutions of the likelihood equations, then $\ell(\cdot)$ has a unique global maximum

Illustrations

Uniform $\mathcal{U}(0, \theta)$ likelihood

$$L(\theta|x_1, \ldots, x_n) = \theta^{-n} \mathbb{I}_{\theta > \max\limits_i x_i}$$

not differentiable at $X_{(n)}$ but

$$\hat{\theta}_n^{\text{MLE}} = X_{(n)}$$

[Super-efficient estimator]

Illustrations

Bernoulli $\mathcal{B}(p)$ likelihood

$$L(p|x_1,\ldots,x_n) = \{p/1-p\}^{\sum_i x_i} (1-p)^n$$

differentiable over $(0,1)$ and

$$\hat{p}_n^{\text{MLE}} = \overline{X}_n$$

Illustrations

Normal $\mathcal{N}(\mu, \sigma^2)$ likelihood

$$L(\mu, \sigma|x_1, \ldots, x_n) \propto \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\bar{x}_n - \mu)^2\right\}$$

differentiable with

$$(\hat{\mu}_n^{\mathsf{MLE}}, \widehat{\sigma^2}_n^{\mathsf{MLE}}) = \left(\overline{X}_n, \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\right)$$

The fundamental theorem of Statistics

fundamental theorem

Under appropriate conditions, if $(X_1, \ldots, X_n) \overset{\text{iid}}{\sim} f(x|\theta)$, if $\hat{\theta}_n$ is solution of $\nabla \log f(X_1, \ldots, X_n|\theta) = 0$, then

$$\sqrt{n}\{\hat{\theta}_n - \theta\} \xrightarrow{\mathcal{L}} \mathcal{N}_p(0, \mathfrak{I}(\theta)^{-1})$$

Equivalent of CLT for estimation purposes

- $\mathfrak{I}(\theta)$ can be replaced with $\mathfrak{I}(\hat{\theta}_n)$
- or even $\hat{\mathfrak{I}}(\hat{\theta}_n) = -1/n \sum_i \nabla\nabla^T \log f(x_i|\hat{\theta}_n)$

The fundamental theorem of Statistics

fundamental theorem

Under appropriate conditions, if $(X_1, \ldots, X_n) \overset{\text{iid}}{\sim} f(x|\theta)$, if $\hat{\theta}_n$ is solution of $\nabla \log f(X_1, \ldots, X_n|\theta) = 0$, then

$$\sqrt{n}\{\hat{\theta}_n - \theta\} \xrightarrow{\mathcal{L}} \mathcal{N}_p(0, \mathfrak{I}(\theta)^{-1})$$

Equivalent of CLT for estimation purposes

- $\mathfrak{I}(\theta)$ can be replaced with $\mathfrak{I}(\hat{\theta}_n)$
- or even $\hat{\mathfrak{I}}(\hat{\theta}_n) = -1/n \sum_i \nabla\nabla^{\mathrm{T}} \log f(x_i|\hat{\theta}_n)$

Assumptions

- θ identifiable
- support of $f(\cdot|\theta)$ constant in θ
- $\ell(\theta)$ thrice differentiable
- [the killer] there exists $g(x)$ integrable against $f(\cdot|\theta)$ in a neighbourhood of the true parameter such that

$$\left|\frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_k}f(\cdot|\theta)\right| \leqslant g(x)$$

- the following identity stands [mostly superfluous]

$$\mathfrak{I}(\theta) = \mathbb{E}_\theta\left[\nabla \log f(X|\theta)\{\nabla \log f(X|\theta)\}^{\mathrm{T}}\right] = -\mathbb{E}_\theta\left[\nabla^{\mathrm{T}}\nabla \log f(X|\theta)\right]$$

- $\hat{\theta}_n$ converges in probability to θ [similarly superfluous]

[Boos & Stefanski, 2014, p.286; Lehmann & Casella, 1998]

Inefficient MLEs

Example of MLE of $\eta = ||\theta||^2$ when $x \sim \mathcal{N}_p(\theta, I_p)$:

$$\hat{\eta}^{\mathsf{MLE}} = ||x||^2$$

Then $\mathbb{E}_\eta[||x||^2] = \eta + p$ diverges away from η with p

Note: Consistent and efficient behaviour when considering the MLE of η based on

$$Z = ||X||^2 \sim \chi^2_p(\eta)$$

[Robert, 2001]

Inefficient MLEs

Example of MLE of $\eta = ||\theta||^2$ when $x \sim \mathcal{N}_p(\theta, I_p)$:

$$\widehat{\eta}^{\text{MLE}} = ||x||^2$$

Then $\mathbb{E}_\eta[||x||^2] = \eta + p$ diverges away from η with p

Note: Consistent and efficient behaviour when considering the MLE of η based on

$$Z = ||X||^2 \sim \chi^2_p(\eta)$$

[Robert, 2001]

Inefficient MLEs

Example of MLE of $\eta = ||\theta||^2$ when $x \sim \mathcal{N}_p(\theta, I_p)$:

$$\hat{\eta}^{\text{MLE}} = ||x||^2$$

Then $\mathbb{E}_\eta[||x||^2] = \eta + p$ diverges away from η with p

Note: Consistent and efficient behaviour when considering the MLE of η based on

$$Z = ||X||^2 \sim \chi_p^2(\eta)$$

[Robert, 2001]

Inconsistent MLEs

Take $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} f_\theta(x)$ with

$$f_\theta(x) = (1-\theta)\frac{1}{\delta(\theta)} f_0\left({}^{x-\theta}/_{\delta(\theta)}\right) + \theta f_1(x)$$

for $\theta \in [0,1]$,

$$f_1(x) = \mathbb{I}_{[-1,1]}(x) \quad f_0(x) = (1-|x|)\mathbb{I}_{[-1,1]}(x)$$

and

$$\delta(\theta) = (1-\theta)\exp\{-(1-\theta)^{-4} + 1\}$$

Then for any θ

$$\hat{\theta}_n^{\text{MLE}} \xrightarrow{\text{a.s.}} 1$$

[Ferguson, 1982; John Wellner's slides, ca. 2005]

Inconsistent MLEs

Consider X_{ij} $i = 1, \ldots, n$, $j = 1, 2$ with $X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$. Then

$$\hat{\mu}_i^{\mathsf{MLE}} = {}^{X_{i1}+X_{i2}}/_{2} \qquad \widehat{\sigma^2}^{\mathsf{MLE}} = \frac{1}{4n}\sum_{i=1}^{n}(X_{i1} - X_{i2})^2$$

Therefore

$$\widehat{\sigma^2}^{\mathsf{MLE}} \xrightarrow{\text{a.s.}} {}^{\sigma^2}/_{2}$$

[Neyman & Scott, 1948]

Inconsistent MLEs

Consider X_{ij} $i = 1, \ldots, n$, $j = 1, 2$ with $X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$. Then

$$\widehat{\mu}_i^{\text{MLE}} = {}^{X_{i1}+X_{i2}}\!/_2 \qquad \widehat{\sigma^2}^{\text{MLE}} = \frac{1}{4n} \sum_{i=1}^{n} (X_{i1} - X_{i2})^2$$

Therefore

$$\widehat{\sigma^2}^{\text{MLE}} \xrightarrow{\text{a.s.}} {}^{\sigma^2}\!/_2$$

[Neyman & Scott, 1948]

Note: Working solely with $X_{i1} - X_{i2} \sim \mathcal{N}(0, 2\sigma^2)$ produces a consistent MLE

Likelihood optimisation

Practical optimisation of the likelihood function

$$\theta^\star = \arg\max_\theta L(\theta|\mathbf{x}) = \prod_{i=1}^n g(X_i|\theta).$$

assuming $\mathbf{X} = (X_1, \ldots, X_n) \overset{\text{iid}}{\sim} g(x|\theta)$

- analytical resolution feasible for exponential families

$$\nabla T(\theta) \sum_{i=1}^n S(x_i) = n \nabla \tau(\theta)$$

- use of standard numerical techniques like Newton-Raphson

$$\theta^{(t+1)} = \theta^{(t)} + I^{\text{obs}}(\mathbf{X}, \theta^{(t)})^{-1} \nabla \ell(\theta^{(t)})$$

with $\ell(.)$ log-likelihood and I^{obs} observed information matrix

Likelihood optimisation

Practical optimisation of the likelihood function

$$\theta^\star = \arg\max_\theta L(\theta|\mathbf{x}) = \prod_{i=1}^n g(X_i|\theta).$$

assuming $\mathbf{X} = (X_1, \ldots, X_n) \overset{\text{iid}}{\sim} g(x|\theta)$

- analytical resolution feasible for exponential families

$$\nabla T(\theta) \sum_{i=1}^n S(x_i) = n\nabla\tau(\theta)$$

- use of standard numerical techniques like Newton-Raphson

$$\theta^{(t+1)} = \theta^{(t)} + \mathrm{I}^{\text{obs}}(\mathbf{X}, \theta^{(t)})^{-1}\nabla\ell(\theta^{(t)})$$

with $\ell(.)$ log-likelihood and I^{obs} observed information matrix

Likelihood optimisation

Practical optimisation of the likelihood function

$$\theta^\star = \arg\max_\theta L(\theta|\mathbf{x}) = \prod_{i=1}^{n} g(X_i|\theta).$$

assuming $\mathbf{X} = (X_1, \ldots, X_n) \overset{\text{iid}}{\sim} g(x|\theta)$

- analytical resolution feasible for exponential families

$$\nabla T(\theta) \sum_{i=1}^{n} S(x_i) = n \nabla \tau(\theta)$$

- use of standard numerical techniques like Newton-Raphson

$$\theta^{(t+1)} = \theta^{(t)} + I^{\text{obs}}(\mathbf{X}, \theta^{(t)})^{-1} \nabla \ell(\theta^{(t)})$$

  with $\ell(.)$ log-likelihood and I^{obs} observed information matrix

EM algorithm

Cases where g is too complex for the above to work

Special case when g is a marginal

$$g(x|\theta) = \int_{\mathcal{Z}} f(x, z|\theta)\, \mathrm{d}z$$

Z called latent or missing variable

Illustrations

- censored data

$$X = \min(X^*, a) \qquad X^* \sim \mathcal{N}(\theta, 1)$$

- mixture model

$$X \sim .3\,\mathcal{N}_1(\mu_0, 1) + .7\,\mathcal{N}_1(\mu_1, 1),$$

- desequilibrium model

$$X = \min(X^*, Y^*) \qquad X^* \sim f_1(x|\theta) \quad Y^* \sim f_2(x|\theta)$$

Completion

EM algorithm based on completing data x with z, such as

$$(\mathbf{X}, \mathbf{Z}) \sim f(x, z|\theta)$$

$\mathbf{Z}$ missing data vector and pair $(\mathbf{X}, \mathbf{Z})$ complete data vector

Conditional density of $\mathbf{Z}$ given x:

$$k(z|\theta, x) = \frac{f(x, z|\theta)}{g(x|\theta)}$$

Completion

EM algorithm based on completing data x with z, such as

$$(X, Z) \sim f(x, z|\theta)$$

Z missing data vector and pair (X, Z) complete data vector

Conditional density of Z given x:

$$k(z|\theta, x) = \frac{f(x, z|\theta)}{g(x|\theta)}$$

Likelihood decomposition

Likelihood associated with complete data (x, z)

$$L^c(\theta|x, z) = f(x, z|\theta)$$

and likelihood for observed data

$$L(\theta|x)$$

such that

$$\log L(\theta|x) = \mathbb{E}[\log L^c(\theta|x, Z)|\theta_0, x] - \mathbb{E}[\log k(Z|\theta, x)|\theta_0, x] \quad (1)$$

for any θ_0, with integration operated against conditionnal distribution of Z given observables (and parameters), $k(z|\theta_0, x)$

[A tale of] two θ's

There are "two θ's" ! : in (1), θ_0 is a fixed (and arbitrary) value driving integration, while θ both free (and variable)

Maximising observed likelihood

$$L(\theta|x)$$

equivalent to maximise r.h.s. term in (1)

$$\mathbb{E}[\log L^c(\theta|x,Z)|\theta_0,x] - \mathbb{E}[\log k(Z|\theta,x)|\theta_0,x]$$

[A tale of] two θ's

There are "two θ's" ! : in (1), θ_0 is a fixed (and arbitrary) value driving integration, while θ both free (and variable)

Maximising observed likelihood

$$L(\theta|\mathbf{x})$$

equivalent to maximise r.h.s. term in (1)

$$\mathbb{E}[\log L^c(\theta|\mathbf{x}, \mathbf{Z})|\theta_0, \mathbf{x}] - \mathbb{E}[\log k(\mathbf{Z}|\theta, \mathbf{x})|\theta_0, \mathbf{x}]$$

Intuition for EM

Instead of maximising wrt θ r.h.s. term in (1), maximise only

$$\mathbb{E}[\log \mathrm{L}^c(\theta|\boldsymbol{x}, \boldsymbol{Z})|\theta_0, \boldsymbol{x}]$$

Maximisation of complete log-likelihood impossible since z unknown, hence substitute by maximisation of expected complete log-likelihood, with expectation depending on term θ_0

Intuition for EM

Instead of maximising wrt θ r.h.s. term in (1), maximise only

$$\mathbb{E}[\log \mathrm{L}^c(\theta|\boldsymbol{x}, \boldsymbol{Z})|\theta_0, \boldsymbol{x}]$$

Maximisation of complete log-likelihood impossible since $\boldsymbol{z}$ unknown, hence substitute by maximisation of expected complete log-likelihood, with expectation depending on term θ_0

Expectation–Maximisation

Expectation of complete log-likelihood denoted

$$Q(\theta|\theta_0, \mathbf{x}) = \mathbb{E}[\log L^c(\theta|\mathbf{x}, \mathbf{Z})|\theta_0, \mathbf{x}]$$

to stress dependence on θ_0 and sample $\mathbf{x}$

Principle

EM derives sequence of estimators $\hat{\theta}_{(j)}$, $j = 1, 2, \ldots,$ through iteration of **E**xpectation and **M**aximisation steps:

$$Q(\hat{\theta}_{(j)}|\hat{\theta}_{(j-1)}, \mathbf{x}) = \max_{\theta} Q(\theta|\hat{\theta}_{(j-1)}, \mathbf{x}).$$

Expectation–Maximisation

Expectation of complete log-likelihood denoted

$$Q(\theta|\theta_0, \mathbf{x}) = \mathbb{E}[\log L^c(\theta|\mathbf{x}, \mathbf{Z})|\theta_0, \mathbf{x}]$$

to stress dependence on θ_0 and sample $\boldsymbol{x}$

Principle

EM derives sequence of estimators $\hat{\theta}_{(j)}$, $j = 1, 2, \ldots$, through iteration of **E**xpectation and **M**aximisation steps:

$$Q(\hat{\theta}_{(j)}|\hat{\theta}_{(j-1)}, \mathbf{x}) = \max_{\theta} Q(\theta|\hat{\theta}_{(j-1)}, \mathbf{x}).$$

EM Algorithm

Iterate (in m)

1. (*step E*) Compute

$$Q(\theta|\hat{\theta}_{(m)}, \boldsymbol{x}) = \mathbb{E}[\log L^c(\theta|\boldsymbol{x}, \mathbf{Z})|\hat{\theta}_{(m)}, \boldsymbol{x}] \,,$$

2. (*step M*) Maximise $Q(\theta|\hat{\theta}_{(m)}, \boldsymbol{x})$ in θ and set

$$\hat{\theta}_{(m+1)} = \arg\max_{\theta} \; Q(\theta|\hat{\theta}_{(m)}, \boldsymbol{x}).$$

until a fixed point [of Q] is found

[Dempster, Laird, & Rubin, 1978]

Justification

Observed likelihood

$$L(\theta|\boldsymbol{x})$$

increases at every EM step

$$L(\hat{\theta}_{(m+1)}|\boldsymbol{x}) \geqslant L(\hat{\theta}_{(m)}|\boldsymbol{x})$$

[Exercice: use Jensen and (1)]

Censored data

Normal $\mathcal{N}(\theta, 1)$ sample right-censored

$$L(\theta|\mathbf{x}) = \frac{1}{(2\pi)^{m/2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{m}(x_i-\theta)^2\right\}[1-\Phi(a-\theta)]^{n-m}$$

Associated complete log-likelihood:

$$\log L^c(\theta|\mathbf{x}, \mathbf{z}) \propto -\frac{1}{2}\sum_{i=1}^{m}(x_i-\theta)^2 - \frac{1}{2}\sum_{i=m+1}^{n}(z_i-\theta)^2 \,,$$

where z_i's are censored observations, with density

$$k(z|\theta, \mathbf{x}) = \frac{\exp\{-\frac{1}{2}(z-\theta)^2\}}{\sqrt{2\pi}[1-\Phi(a-\theta)]} = \frac{\varphi(z-\theta)}{1-\Phi(a-\theta)}, \qquad a < z.$$

Censored data

Normal $\mathcal{N}(\theta, 1)$ sample right-censored

$$L(\theta|\mathbf{x}) = \frac{1}{(2\pi)^{m/2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{m}(x_i-\theta)^2\right\}[1-\Phi(a-\theta)]^{n-m}$$

Associated complete log-likelihood:

$$\log L^c(\theta|\mathbf{x}, \mathbf{z}) \propto -\frac{1}{2}\sum_{i=1}^{m}(x_i-\theta)^2 - \frac{1}{2}\sum_{i=m+1}^{n}(z_i-\theta)^2 ,$$

where z_i's are censored observations, with density

$$k(z|\theta, \mathbf{x}) = \frac{\exp\{-\frac{1}{2}(z-\theta)^2\}}{\sqrt{2\pi}[1-\Phi(a-\theta)]} = \frac{\varphi(z-\theta)}{1-\Phi(a-\theta)}, \qquad a < z.$$

Censored data (2)

At j-th EM iteration

$$
\begin{aligned}
Q(\theta|\hat{\theta}_{(j)}, \boldsymbol{x}) &\propto -\frac{1}{2}\sum_{i=1}^{m}(x_i-\theta)^2 - \frac{1}{2}\mathbb{E}\left[\sum_{i=m+1}^{n}(Z_i-\theta)^2 \middle| \hat{\theta}_{(j)}, \boldsymbol{x}\right] \\
&\propto -\frac{1}{2}\sum_{i=1}^{m}(x_i-\theta)^2 \\
&\quad -\frac{1}{2}\sum_{i=m+1}^{n}\int_a^{\infty}(z_i-\theta)^2 k(z|\hat{\theta}_{(j)}, \boldsymbol{x})\, dz_i
\end{aligned}
$$

Censored data (3)

Differenciating in θ,

$$n\,\hat{\theta}_{(j+1)} = m\bar{x} + (n-m)\mathbb{E}[Z|\hat{\theta}_{(j)}]\ ,$$

with

$$\mathbb{E}[Z|\hat{\theta}_{(j)}] = \int_a^\infty zk(z|\hat{\theta}_{(j)}, x)\,dz = \hat{\theta}_{(j)} + \frac{\varphi(a-\hat{\theta}_{(j)})}{1-\Phi(a-\hat{\theta}_{(j)})}.$$

Hence, EM sequence provided by

$$\hat{\theta}_{(j+1)} = \frac{m}{n}\bar{x} + \frac{n-m}{n}\left[\hat{\theta}_{(j)} + \frac{\varphi(a-\hat{\theta}_{(j)})}{1-\Phi(a-\hat{\theta}_{(j)})}\right],$$

which converges to likelihood maximum $\hat{\theta}$

Censored data (3)

Differenciating in θ,

$$n\,\hat{\theta}_{(j+1)} = m\bar{x} + (n-m)\mathbb{E}[Z|\hat{\theta}_{(j)}]\ ,$$

with

$$\mathbb{E}[Z|\hat{\theta}_{(j)}] = \int_a^\infty z k(z|\hat{\theta}_{(j)}, \mathbf{x})\,\mathrm{d}z = \hat{\theta}_{(j)} + \frac{\varphi(a-\hat{\theta}_{(j)})}{1-\Phi(a-\hat{\theta}_{(j)})}.$$

Hence, EM sequence provided by

$$\hat{\theta}_{(j+1)} = \frac{m}{n}\bar{x} + \frac{n-m}{n}\left[\hat{\theta}_{(j)} + \frac{\varphi(a-\hat{\theta}_{(j)})}{1-\Phi(a-\hat{\theta}_{(j)})}\right],$$

which converges to likelihood maximum $\hat{\theta}$

Mixtures

Mixture of two normal distributions with unknown means

$$.3\,\mathcal{N}_1(\mu_0, 1) + .7\,\mathcal{N}_1(\mu_1, 1),$$

sample $X_1, \ldots, X_n$ and parameter $\theta = (\mu_0, \mu_1)$
Missing data: $Z_i \in \{0, 1\}$, indicator of component associated with X_i ,

$$X_i|z_i \sim \mathcal{N}(\mu_{z_i}, 1) \qquad Z_i \sim \mathcal{B}(.7)$$

Complete likelihood

$$\begin{aligned}\log L^c(\theta|x, z) &\propto -\frac{1}{2}\sum_{i=1}^n z_i(x_i - \mu_1)^2 - \frac{1}{2}\sum_{i=1}^n (1 - z_i)(x_i - \mu_0)^2 \\ &= -\frac{1}{2}n_1(\hat{\mu}_1 - \mu_1)^2 - \frac{1}{2}(n - n_1)(\hat{\mu}_0 - \mu_0)^2\end{aligned}$$

with

$$n_1 = \sum_{i=1}^n z_i\,, \quad n_1\hat{\mu}_1 = \sum_{i=1}^n z_i x_i\,, \quad (n - n_1)\hat{\mu}_0 = \sum_{i=1}^n (1 - z_i)x_i$$

Mixtures

Mixture of two normal distributions with unknown means

$$.3\,\mathcal{N}_1(\mu_0, 1) + .7\,\mathcal{N}_1(\mu_1, 1),$$

sample $X_1, \ldots, X_n$ and parameter $\theta = (\mu_0, \mu_1)$

Missing data: $Z_i \in \{0, 1\}$, indicator of component associated with X_i ,

$$X_i | z_i \sim \mathcal{N}(\mu_{z_i}, 1) \qquad Z_i \sim \mathcal{B}(.7)$$

Complete likelihood

$$\begin{aligned}\log L^c(\theta|\mathbf{x}, \mathbf{z}) &\propto -\frac{1}{2}\sum_{i=1}^n z_i(x_i - \mu_1)^2 - \frac{1}{2}\sum_{i=1}^n (1 - z_i)(x_i - \mu_0)^2 \\ &= -\frac{1}{2} n_1 (\hat{\mu}_1 - \mu_1)^2 - \frac{1}{2}(n - n_1)(\hat{\mu}_0 - \mu_0)^2\end{aligned}$$

with

$$n_1 = \sum_{i=1}^n z_i\,, \quad n_1\hat{\mu}_1 = \sum_{i=1}^n z_i x_i\,, \quad (n - n_1)\hat{\mu}_0 = \sum_{i=1}^n (1 - z_i) x_i$$

Mixtures (2)

At j-th EM iteration

$$Q(\theta|\hat{\theta}_{(j)}, \mathbf{x}) = \frac{1}{2}\mathbb{E}\left[n_1(\hat{\mu}_1 - \mu_1)^2 + (n - n_1)(\hat{\mu}_0 - \mu_0)^2|\hat{\theta}_{(j)}, \mathbf{x}\right]$$

Differenciating in θ

$$\hat{\theta}_{(j+1)} = \begin{pmatrix} \mathbb{E}\left[n_1\hat{\mu}_1 \left|\hat{\theta}_{(j)}, \mathbf{x}\right.\right] \Big/ \mathbb{E}\left[n_1|\hat{\theta}_{(j)}, \mathbf{x}\right] \\ \\ \mathbb{E}\left[(n - n_1)\hat{\mu}_0 \left|\hat{\theta}_{(j)}, \mathbf{x}\right.\right] \Big/ \mathbb{E}\left[(n - n_1)|\hat{\theta}_{(j)}, \mathbf{x}\right] \end{pmatrix}$$

Mixtures (3)

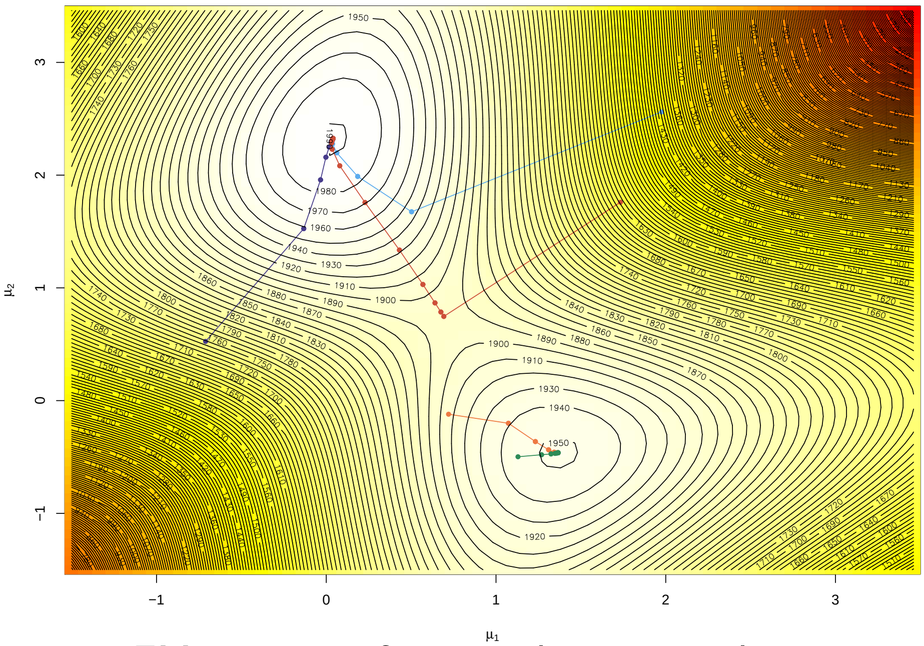
Hence $\widehat{\theta}_{(j+1)}$ given by

$$\left(\begin{array}{c} \sum_{i=1}^{n} \mathbb{E}\left[Z_i \left|\widehat{\theta}_{(j)}, x_i\right.\right] x_i \Big/ \sum_{i=1}^{n} \mathbb{E}\left[Z_i|\widehat{\theta}_{(j)}, x_i\right] \\ \\ \sum_{i=1}^{n} \mathbb{E}\left[(1-Z_i)\left|\widehat{\theta}_{(j)}, x_i\right.\right] x_i \Big/ \sum_{i=1}^{n} \mathbb{E}\left[(1-Z_i)|\widehat{\theta}_{(j)}, x_i\right] \end{array}\right)$$

Conclusion

Step *(E)* in EM replaces missing data Z_i with their conditional expectation, given $\boldsymbol{x}$ (expectation that depend on $\widehat{\theta}_{(m)}$).

Mixtures (3)



EM iterations for several starting values

Properties

EM algorithm such that

- it converges to local maximum or saddle-point
- it depends on the initial condition $\theta_{(0)}$
- it requires several initial values when likelihood multimodal