Advanced statistical methods

Frédéric Pascal

CentraleSupélec, Laboratory of Signals and Systems (L2S), France
frederic.pascal@centralesupelec.fr
http://fredericpascal.blogspot.fr

MSc in Data Sciences & Business Analytics
CentraleSupélec / ESSEC
Oct. 2nd - Dec. 20th, 2017

CentraleSupélec

Part B

Statistical Modelling and Parameter Estimation theory

Part B: Contents

Key references of Part B

From an EE / SP point of view...

- Kay, Steven M. *Fundamentals of Statistical Signal Processing - Estimation Theory*, Vol. 1, Prentice Hall, 1993.
- Poor, Vincent, H. *An Introduction to Signal Detection and Estimation*, 2nd ed, Springer, 1998.

From a statistical point of view...

- Casella, George, and Roger L. Berger. *Statistical inference*, Vol. 2. Pacific Grove, CA: Duxbury, 2002.
- Lehmann, Erich Leo, and Casella, George. *Theory of point estimation*, Springer Science & Business Media, 2006.

+ many many references...

I. Statistical modelling

- Generalities
- Sufficiency
- Exponential family

II. Unbiased estimation

III. Theory of Point Estimation

Statistical modelling

Generalities

- n-sample $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$
- Dominated models $\rightsquigarrow$ Likelihood Function (LF), denoted $L(\mathbf{x}, \theta)$
- Parametric models, i.e. $\theta \in \Theta \subset \mathbb{R}^d$

Definition (Identifiability conditions)

*A model $(\mathscr{X}, \mathscr{A}, \{P_\theta, \theta \in \Theta\})$ is said **identifiable** if the mapping from Θ onto the probabilities space $(\mathscr{X}, \mathscr{A})$, which to θ gives P_θ is injective.*

Definition (Statistic)

*In a statistical model $\{\mathscr{X}, \mathscr{A}, \{P_\theta, \theta \in \Theta\}\}$, one said **statistic**, for any (measurable or σ-finite) mapping S from $(\mathscr{X}, \mathscr{A})$ onto an arbitrary space. Let's say a statistic is a function of r.V. $S(\mathbf{x}_1, \ldots, \mathbf{x}_n)$.*

e.g., $\bar{X}_n, R_{n,X}, \hat{\sigma}^2_n$, or even X ☺,...

Statistical modelling

Sufficient statistics

Very important concept! for high-dimensional data, dimension reduction without reducing the information brought by the data.

Main idea: Where is contain the information of interest (i.e. related to the unknowns) in the data?

Example: Coin toss -> Head and Tails - One want to know the probability of Head or if the coin is biased ... No need to keep the whole dataset...

Definition (Sufficient statistic)

A statistic S is said to be sufficient iff the conditional distribution $\mathscr{L}_\theta(X|S(X))$ does not depend on θ.

Remark (Pros and cons)

- *Difficulty to use the definition*
- *Dimension of S has to be minimal! $(\mathbf{x}_1,\ldots,\mathbf{x}_n)$ is always a sufficient stat.* ☺

Statistical modelling

Sufficient statistics characterization

Theorem (Factorisation Criterion (FC))

A statistic S is sufficient iff the likelihood function can be written as:

$$L(x;\theta) = \psi(S(x);\theta)\lambda(x).$$

This is a sort of separability theorem...

Example: let $(X_1, \ldots, X_n)$ i.i.d following a non-centred exponential dist., i.e. with PDF

$$f(x_i, \theta) = \frac{1}{\theta_2} \exp\left(-\frac{1}{\theta_2}(x_i - \theta_1)\right) \mathbb{1}_{\{x_i \geq \theta_1\}} \quad \text{with} \quad \theta = (\theta_1, \theta_2)^t.$$

$$\Rightarrow S(X) = \left(\min_{i=1,\ldots,n}(X_i), \sum_{i=1}^{n} X_i\right) \text{ is sufficient!}$$

Exponential family

Definition (Complete statistics)

A statistic S is said to be complete if for any measurable real-valued function ϕ, one has

$$\left\{\forall\theta\in\Theta, E_\theta\left[\phi\circ S(X)\right]=0\right\}\Rightarrow\left\{\forall\theta\in\Theta, \phi\circ S(X)=0\ \textit{a.s.}\ [P_\theta]\right\}.$$

Purely theoretical... for optimal unbiased estimation...

Definition (Exponential family)

A model is said to be exponential iff its LF can be written as:

$$L(x;\theta)=h(x)\phi(\theta)\exp\left\{\sum_{i=1}^{r} Q_i(\theta)S_i(x)\right\}. \tag{1}$$

*where $S(.)=(S_1(.),\dots,S_r(.))$ is the **canonical statistic**.*

Discussion: r, large family (discrete and continuous models),...

Exponential family

Some very useful properties in the class of models...

Proposition

The canonical statistic is sufficient.

trivial with FC...

Proposition

For exponential family, if the $S_i(.)$ are linearly independent (affine sense), i.e.,

$$\forall x \in \mathscr{X}, \sum_{i=1}^{r} a_i S_i(x) = a_0 \implies a_0 = a_j = 0 \, \forall j$$

Thus $P_{\theta_1} = P_{\theta_2} \iff Q_j(\theta_1) = Q_j(\theta_2)$.

Corollary

For exponential family, if the $S_i(.)$ are linearly independent,
θ *is identifiable* $\iff \theta \mapsto Q(\theta)$ *is injective.*

Exponential family

Some very useful properties in the class of models...

Theorem

If $Q(\Theta)$ contains a non-empty set of $\mathbb{R}^r$, the canonical statistic is complete.

Proposition

Of course, the canonical statistic follows an exponential model.

Models examples:

- Exponential dist.!
- Gaussian
- Poisson
- Binomial dist.
- ...
- Exhaustive list on Wikipedia ☺

I. Statistical modelling

II. Unbiased estimation

- Generalities
- Fisher information
- Optimality
- Cramer-Rao bound

III. Theory of Point Estimation

Unbiased estimation

Regularity conditions

(A_1) The model is dominated

(A_2) The dist. domain $P_\theta : \Delta = \{x \in \mathscr{X} \,|\, L(x;\theta) > 0\}$ does not depend on $\theta \in \Theta$.

(A_3) $L(x;\theta)$ is twice differentiable: $\dfrac{\partial L}{\partial \theta}(x;\theta)$ and $\dfrac{\partial^2 L}{\partial \theta^2}(x;\theta)$ exist $\forall x \in \Delta$, $\forall \theta$.

(A_4) Functions $\dfrac{\partial L}{\partial \theta}$ and $\dfrac{\partial^2 L}{\partial \theta^2}$ are integrable $\forall \theta$, and $\forall \theta \in \Theta, A \in \mathscr{X}$, one has:

$$\left\{ \begin{array}{l} \dfrac{\partial}{\partial \theta} \displaystyle\int_A L(x;\theta)\, dx = \int_A \dfrac{\partial}{\partial \theta} L(x;\theta)\, dx, \\ \dfrac{\partial^2}{\partial \theta^2} \displaystyle\int_A L(x;\theta)\, dx = \int_A \dfrac{\partial^2}{\partial \theta^2} L(x;\theta)\, dx. \end{array} \right.$$

Definition (Regular model)

If Θ is an open set and if $(A_1), (A_2), (A_3), (A_4)$ are verified, the model is regular.

Fisher Information (FI) Matrix (FIM)

Definition (Score)

*The **score** function is the r.V. $s_\theta(\mathbf{x})$ defined by:*

$$s_\theta(\mathbf{x}) = \frac{\partial}{\partial\theta} l(\mathbf{x};\theta),$$

where $l(x;\theta) = \log(L(x;\theta))$ is the log*-likelihood function.*

Proposition

The score is zero-mean, i.e. $E[s_\theta(\mathbf{x})] = 0$.

Definition (FIM)

If one has (A_5) the score is square-integrable, the FIM is the variance (covariance matrix in multidimensional case) of the score:

$$I(\theta) = var_\theta(s_\theta(\mathbf{x})) = E_\theta\left[s_\theta(\mathbf{x})\, s_\theta(\mathbf{x})^t\right].$$

FIM

Remark

In case of a n-sample, $(\mathbf{x}_1,\ldots,\mathbf{x}_n)$, the score can be written as:

$$s_{n,\theta}(\mathbf{x}) = \frac{\partial}{\partial\theta} l_n(\mathbf{x}_1,\ldots,\mathbf{x}_n;\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} l(\mathbf{x}_i;\theta),$$

where $l_n(x_1,\ldots,x_n;\theta)$ is the $\log$-likelihood function of the n-sample.
In such case, the FIM, $I_n(\theta)$ can be written (by independence) as

$$I_n(\theta) = n\,I(\theta).$$

Proposition

Let's assume a regular model, plus (A_5), then for a real θ, one has:

$$I(\theta) = -E_\theta\left[\frac{\partial^2}{\partial\theta\partial\theta^t} l(\mathbf{x};\theta)\right].$$

FIM

Some examples...

Let us consider a n-sample of r.v. Prove the following results:

1. If $P_\theta \sim B(\theta, 1), \theta \in]0,1[$, thus $I_n(\theta) = \frac{n}{\theta(1-\theta)}$.

2. If $P_\theta \sim \text{Poisson}(\theta), \theta > 0$, thus $I_n(\theta) = \frac{n}{\theta}$.

3. If $P_\theta \sim \mathcal{N}(\mu, \sigma^2), (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$, thus:

$$I_n(\theta) = n \begin{pmatrix} \dfrac{1}{\sigma^2} & 0 \\ 0 & \dfrac{1}{2\sigma^4} \end{pmatrix}.$$

Unbiased estimation - Decision theory

Main idea: give an answer d regarding the data...
Define a loss function $\rho(d,\theta)$ between d and the (true) value of the unknowns θ or $g(\theta)$. Generally,

Definition (quadratic loss)

$$\rho(d,\theta) = (d - g(\theta))^t \mathbf{A}(\theta)(d - g(\theta))$$

where $\mathbf{A}(.)$ is positive-definite

Use $\mathbf{A}(\theta) = \mathbf{I}$ leads to $\rho(d,\theta) = (d - g(\theta))^2$...

Definition (Estimator)

*An **estimator** of $g(\theta)$ is a statistic $\delta(\mathbf{x})$ mapping $\mathscr{X}$ into $\mathscr{D} = g(\Theta)$.*

Definition (Mean Square Error (MSE))

$$R_\delta(\theta) = E_\theta\left[\rho(\theta,\delta(\mathbf{x}))\right] = E_\theta\left[(g(\theta) - \delta(\mathbf{x}))^2\right].$$

Rao-Blackwell (RB) estimator

Goal: minimize the MSE but...

Proposition

$$R_\delta(\theta) = var_\theta(\delta(\mathbf{x})) + b_\delta(\theta)^2,$$

where $b_\delta(\theta)$ is the biais of $\delta(\mathbf{x})$, i.e. $b_\delta(\theta) = E_\theta\left[\delta(\mathbf{x}) - g(\theta)\right]$.

$\Rightarrow$ Unbiased estimation!

Theorem (Rao-Blackwell estimator)

Let δ an estimator and S a sufficient statistic. Let's define

$$\delta_S : \mathbf{x} \to E_\theta\left[\delta(\mathbf{x}) | S(\mathbf{x}) = S(x)\right],$$

Thus

$$\forall \theta \in \Theta, \, R_{\delta_S}(\theta) \le R_\delta(\theta).$$

δ_S is Rao-Blackwell estimator (or the Rao-Blackwellization of δ).
It is unbiased if δ is unbiased.

Optimality: Lehman-Scheffé (LS) theorem

Theorem (Lehmann-Scheffé theorem)

If δ is unbiased and if S is a sufficient and complete statistic, thus the Rao-Blackwell estimator δ_S is optimal in the class of unbiased estimators, i.e. its variance is minimal for all $\theta \in \Theta$.

Proof

Few lines...

Some examples...

Definition (Regular estimator)

Let a regular model, and let an estimator δ of $g(\theta)$ s.t.

$$E_\theta\left[|\delta|^2\right] < \infty, \forall \theta \in \Theta \text{ and } \frac{\partial}{\partial \theta}\int_{\mathscr{X}} \delta(x)\, l(x;\theta)\, dx = \int_{\mathscr{X}} \delta(x)\, \frac{\partial}{\partial \theta} l(x;\theta)\, dx,$$

*Then, δ is **regular estimator** of $g(\theta)$.*

Cramer-Rao lower bound

Theorem (Cramer-Rao lower Bound (CRB) - FDCR inequality)

Let δ an unbiased, regular estimator of $g(\theta) \in \mathbb{R}^k$ where $\theta \in \Theta \subset \mathbb{R}^p$. The function g is of class C^1. Let's also assume that $I(\theta)$ is positive-definite. Thus, for a n-sample, and for all $\theta \in \Theta$, one has:

$$R_\delta(\theta) = var_\theta(\delta) \geq \frac{1}{n} \frac{\partial g}{\partial \theta^t}(\theta)\, I(\theta)^{-1}\, \frac{\partial g^t}{\partial \theta}(\theta),$$

with $\dfrac{\partial g}{\partial \theta^t}(\theta)$ the $p \times k$-matrix defined by $\left(\dfrac{\partial g_i}{\partial \theta_j}(\theta)\right)_{1 \leq i \leq p, 1 \leq j \leq k}$ and $\dfrac{\partial g^t}{\partial \theta}(\theta) = \left(\dfrac{\partial g}{\partial \theta'}(\theta)\right)^t$ its transpose.

Cramer-Rao lower bound

Definition (Efficiency)

An unbiased estimator is said to be efficient iff its variance is the CRB.

Proposition

If T is an efficient estimator of $g(\theta)$, then the affine transform $\mathbf{A}\,T + \mathbf{b}$ is an efficient estimator of $\mathbf{A}\,g(\theta) + \mathbf{b}$ (for $\mathbf{A}$ and $\mathbf{b}$ with appropriate dimensions)

Proposition

An efficient estimator is optimal.
The converse is (obviously) wrong.

Think about the students grades in a given course ☺

Link with exponential family

Consider an exponential model (1), $L(x;\theta) = h(x)\phi(\theta)\exp\left\{\sum_{i=1}^{r} Q_i(\theta) S_i(x)\right\}$ and make the change of variable $\lambda_j = Q_j(\theta)$. Then, one obtains:

Definition (Exponential model under a natural form...)

... when the LR is

$$L(x,\lambda) = K(\lambda)\, h(x) \exp\left[\sum_{j=1}^{r} \lambda_j S_j(x)\right] \tag{2}$$

The new parameters $(\lambda_1, \cdots, \lambda_r) \in \Lambda = Q(\Theta) \subset \mathbb{R}^r$

Theorem (Regularity)

Let an exponentiel model (2). If Λ is a non-empty open set of $\mathbb{R}^r$, then the model is regular and (A_5) is verified, $\Rightarrow I(\lambda)$ exists. Furthermore

$$I(\lambda) = -E_\lambda\left[\frac{\partial^2 \ln L(\mathbf{x},\lambda)}{\partial\lambda\partial\lambda^t}\right]$$

Link with exponential family

Theorem (Identifiability)

Let us consider the exponential model (2) where Λ is a (non-empty) open set of $\mathbb{R}^r$. Then, the model is identifiable, i.e., $(P_{\lambda_1} = P_{\lambda_2} \Longrightarrow \lambda_1 = \lambda_2)$ iff the FIM $I(\lambda)$ is invertible $\forall \lambda \in \Lambda$.

Theorem (Necessary condition)

Let us consider the exponential model (1). Let us assume that the model is regular et let δ an unbiased regular estimator of $g(\theta)$. Moreover, let us assume that g is of class C^1 and that $I(\theta)$ is invertible $\forall \theta \in \Theta$. Thus, if δ is efficient, it is necessary an affine function of $S(\mathbf{x}) = (S_1(\mathbf{x}), \cdots, S_r(\mathbf{x}))^t$.

Remark

Previous theorem is useful for proving the NON efficiency of an estimator...

Theorem (Converse of the CRB - Equality)

*Given a regular model where $\Theta \subset \mathbb{R}^d$ is a non-empty open set, let $g : \Theta \mapsto \mathbb{R}^p$ of class C^1 s.t. $\dfrac{\partial g}{\partial \theta^t}(\theta)$ is a **square** invertible matrix $\forall \theta \in \Theta$* **so that** $p = d$.
Assume that $I(\theta)$ exists and is invertible $\forall \theta \in \Theta$.
Thus $\delta(\mathbf{x})$ is a regular and EFFICIENT (unbiased) estimator of $g(\theta)$ iff $L(x,\theta)$ can be written as:

$$L(x,\theta) = C(\theta)\, h(x) \, exp\left[\sum_{j=1}^{d} Q_j(\theta) S_j(x)\right]$$

where functions Q and C are s.t.

- *Q and C are differentiable $\forall \theta \in \Theta$*
- *$\dfrac{\partial Q}{\partial \theta^t}(\theta)$ is invertible $\forall \theta \in \Theta$*
- $g(\theta) = -\left(\dfrac{\partial Q}{\partial \theta^t}(\theta)\right)^{-1} \dfrac{\partial \ln C}{\partial \theta^t}(\theta).$

CRB equality

Corollary

In an exponential model (2) (in the natural form) where $\Lambda \subset \mathbb{R}^r$ is a non-empty open set and where $I(\lambda)$ is invertible $\forall \lambda \in \Lambda$.
Thus, each statistic $S_j(\mathbf{x})$ is an efficient estimator of $E_\lambda\left[S_j(X)\right]$ which is defined as :

$$g_j(\lambda) = -\frac{\partial \ln K}{\partial \lambda_j}(\lambda)$$

Some applications...

Limitations of unbiased estimation theory: restrictive class, difficult derivation for estimators, limited to exponential family...

I. Statistical modelling

II. Unbiased estimation

III. Theory of Point Estimation

- Basics
- Method of Moment
- Method of Maximum Likelihood

Basics

Let us denote $T_n(\mathbf{x}_1,\ldots,\mathbf{x}_n)$ or $\hat{\theta}_n$ an estimator of θ (or the true value θ_0 if needed).

Definition (Consistancy)

An estimator $\hat{\theta}_n$ of $g(\theta)$ is strongly (resp. weakly) consistant if it P_{θ_0}-almost surely (resp. in proba.) converges towards $g(\theta_0)$, with $g:\Theta \to \mathbb{R}^p$.

Definition (Asymptotically unbiased)

*An estimator $\hat{\theta}_n$ of $g(\theta)$ is **asymptotically unbiased** if its limiting distribution is zero-mean, i.e.,*

$$\exists c_n \to \infty \text{ s.t. } c_n\left(\hat{\theta}_n - g(\theta_0)\right) \xrightarrow[n\to\infty]{dist.} \mathbf{z} \text{ with } E_{\theta_0}[\mathbf{z}] = 0.$$

Remark: Different from "unbiased at the limit": $E_{\theta_0}\left[\hat{\theta}_n\right] \xrightarrow[n\to\infty]{} g(\theta_0)$.

Basics

Definition (Asymptotically normal)

*$\hat{\theta}_n$ is **asymptotically normal** if*

$$\sqrt{n}\left(\hat{\theta}_n - g(\theta_0)\right) \xrightarrow[n\to\infty]{dist.} \mathcal{N}\left(\mathbf{0}, \Sigma(\theta_0)\right)$$

where $\Sigma(\theta_0)$ (PDS) is the asymptotic CM of $\hat{\theta}_n$.
Remark: This implies that $\hat{\theta}_n$ is asymptotically unbiased.

Definition (Asymptotically efficient)

An estimator is **asymptotically efficient** *if it is asymptotically normal and if:*

$$\Sigma(\theta_0) = \frac{\partial g}{\partial \theta^t}(\theta_0)\, I(\theta_0)^{-1}\, \frac{\partial g^t}{\partial \theta}(\theta_0)$$

Method of Moment

Let a n-sample $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ i.i.d. with $\mathbf{x}_1 \sim P_\theta$ where $\theta \in \Theta \subset \mathbb{R}^d$ s.t. $E[\|\mathbf{x}_1\|^d) < \infty$. Let us assume that:

$$m = \begin{pmatrix} m_1 \\ \vdots \\ m_d \end{pmatrix} = \begin{pmatrix} \phi_1(\theta_1, \cdots, \theta_d) \\ \vdots \\ \phi_d(\theta_1, \cdots, \theta_d) \end{pmatrix} = \phi \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_d \end{pmatrix}$$

where $m_k = E_\theta[\mathbf{x}^k]$. If function ϕ is invertible (with inverse ψ), one has:

$$\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_d \end{pmatrix} = \begin{pmatrix} \psi_1(m_1, \cdots, m_d) \\ \vdots \\ \psi_d(m_1, \cdots, m_d) \end{pmatrix} = \psi \begin{pmatrix} m_1 \\ \vdots \\ m_d \end{pmatrix}$$

Theorem

- $U_p \xrightarrow[n\to\infty]{a.s} m_p$ *where* $\forall p,\ U_p = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i^p$
- $\sqrt{n}(\mathbf{U} - \mathbf{m}) \xrightarrow[n\to\infty]{dist.} \mathcal{N}(\mathbf{0}, \mathbf{Z})$ *where* $\mathbf{U} = (U_1, \cdots, U_p)^t$, $\mathbf{m} = (m_1, \cdots, m_p)^t$.

Method of Moment

The estimator of the Method of Moment (MME) is defined as

$$\hat{\theta}_n = \begin{pmatrix} \hat{\theta}_{n1} \\ \vdots \\ \hat{\theta}_{nd} \end{pmatrix} = \begin{pmatrix} \psi_1(U_1, \cdots, U_d) \\ \vdots \\ \psi_d(U_1, \cdots, U_d) \end{pmatrix} = \psi \begin{pmatrix} U_1 \\ \vdots \\ U_d \end{pmatrix}$$

where $\forall p$, $U_p = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i^p$ with $\mathbf{x}_i$ are i.i.d.

Theorem (Asymptotics of the MM estimator)

If function ψ is differentiable, then

- $\hat{\theta}_n \xrightarrow[n\to\infty]{a.s} \theta$
- $\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{dist.} \mathcal{N}\left(\mathbf{0}, \mathbf{A}(\theta)\right)$ *where* $\mathbf{A}(\theta) = \dfrac{\partial \psi}{\partial \theta^t}(m)\Sigma(\theta)\dfrac{\partial \psi^t}{\partial \theta}(m)$ *with* $m = \phi(\theta)$.

MME strongly consistant, asymptotically normal BUT generally NOT asymptotically efficient!

Method of Maximum Likelihood

Assume a regular model + (A_5) +

(A_6) $\forall x \in \Delta$, for θ close to θ_0, $\log(f(x;\theta))$ is $3\times$ differentiable w.r.t. θ and

$$\left| \frac{\partial^3}{\partial\theta_j \partial\theta_k \partial\theta_l} \log\left(f(x;\theta)\right) \right| \leq M(x)$$

with $E_{\theta_0}[M(x)] < +\infty$.

Proposition

Assume the model is identifiable, then $\forall\theta \neq \theta_0$, one has

$$P_{\theta_0}\left(L(\mathbf{x}_1,\ldots,\mathbf{x}_n;\theta_0) > L(\mathbf{x}_1,\ldots,\mathbf{x}_n;\theta)\right) \xrightarrow[n\to=\infty]{} 1$$

where $L(\mathbf{x}_1,\ldots,\mathbf{x}_n;\theta)$ is the LF.

The LF is maximum at the point θ_0...

Method of Maximum Likelihood

Definition (Maximum Likelihood Estimator (MLE))

The MLE is defined by

$$T : (\mathbf{x}_1, \ldots, \mathbf{x}_n) \to \widehat{\theta}_n \in \arg\max_{\theta \in \Theta} L(\mathbf{x}_1, \ldots, \mathbf{x}_n; \theta) \, .$$

The MLE has to verified the following likelihood equations!

$$\left\{ \begin{array}{lcl} \dfrac{\partial}{\partial \theta} l(\mathbf{x}_1, \ldots, \mathbf{x}_n; \theta) & = & 0 \\ \dfrac{\partial^2}{\partial \theta \partial \theta^t} l(\mathbf{x}_1, \ldots, \mathbf{x}_n; \theta) & \leq & 0 \, , \end{array} \right.$$

where $l(\mathbf{x}_1, \ldots, \mathbf{x}_n; \theta) = \log\left(L(\mathbf{x}_1, \ldots, \mathbf{x}_n; \theta)\right)$

Definition

Let $g : \Theta \mapsto \mathbb{R}^p$. If $\hat{\theta}_n$ is a MLE of θ, then $g(\hat{\theta}_n)$ is also a MLE of $g(\theta)$.

⚠ the MLE is not necessary unique...

MLE asymptotics

Theorem

Assume: identifiable model, (A_1), (A_2), $\theta_0 \in \Theta \neq \emptyset$, *compact, and*

- $x_1 \mapsto L(x_1, \theta)$ *is bounded* $\forall \theta \in \Theta$;
- $\theta \mapsto L(x_1, \theta)$ *is continuous* $\forall x_1 \in \Delta$;

Thus, $\hat{\theta}_n^{ML} \xrightarrow[n\to\infty]{a.s} \theta_0$ *(Existence from a given* n_0*)*

Theorem (Classical asymptotics)

Assume: identifiable model, Θ *open set of* $\mathbb{R}^d$ *and* $(A_1) - (A_6)$.
Thus, $\exists \hat{\theta}_n^{ML}$ *(from a given* n_0*) solution to the likelihood equations s.t.*

$$\begin{cases} \hat{\theta}_n^{ML} \xrightarrow[n\to\infty]{a.s} \theta_0 \\ \sqrt{n}\left(\hat{\theta}_n^{ML} - \theta_0\right) \xrightarrow[n\to\infty]{dist.} \mathcal{N}\left(\mathbf{0}, I_1(\theta_0)^{-1}\right) \end{cases}$$

MLE asymptotics

Theorem (Classical asymptotics)

Assume: identifiable model, Θ open set of $\mathbb{R}^d$ and $(A_1)-(A_6)$ AND $g\colon \mathbb{R}^d \to \mathbb{R}^p$ differentiable

Thus, $\exists \hat{\theta}_n^{ML}$ (from a given n_0) solution to the likelihood equations s.t.

$$\begin{cases} g\left(\hat{\theta}_n^{ML}\right) \xrightarrow[n\to\infty]{a.s} g(\theta_0) \\ \sqrt{n}\left(g\left(\hat{\theta}_n^{ML}\right) - g(\theta_0)\right) \xrightarrow[n\to\infty]{dist.} \mathcal{N}\left(\mathbf{0}, \dfrac{\partial g}{\partial \theta^t}(\theta_0)\, I_1(\theta_0)^{-1} \dfrac{\partial g^t}{\partial \theta}(\theta_0)\right) \end{cases}$$

Conclusions

The MLE is strongly consistant, asymptotically normal and asymptotically efficient.

Come back on exponential models

Theorem

Let an exponential model (2) *(under natural form)*

$$L(x,\lambda) = K(\lambda)\,h(x)\exp\left(\sum_{i=1}^{r}\lambda_j S_j(x)\right)$$

where $\lambda \in \Lambda$ *and* Λ *is a non-empty open-setof* $\mathbb{R}^r$*. Moreover, let us assume that* $I(\lambda)$ *is invertible* $\forall \lambda \in \Lambda$ *(identifiable model).*

Thus, the MLE exists (from a given n_0*), is unique, strongly consistant and asymptotically efficient (which includes asymptotically normal).*

Proof

Up to you ...