Emanuele Pardini

Mathematical Statistics

Notes from the lectures of the course "Mathematical Statistics" held by prof. A. Agazzi and prof. K. Papagiannouli at University of Pisa during the academic year 2023/24.

Contents

1

Estimation Theory

1.1 General theory of estimators

Definition 1.1.1 (Statistical model): A *statistical model* is a triplet $(\mathbb{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ s.t.

- $(\mathbb{X}, \mathcal{F})$ is a measurable space;
- $\mathbb{P}_\theta$ is a probability measure for every $\theta \in \Theta$.

The set Θ is called *parameter space*.

Remark 1.1.2: Usually $\Theta = \mathbb{R}^p$.

Remark 1.1.3: Fix a statistical model $(\mathbb{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$. In the following, given a measurable function $T : \mathbb{X} \to \mathbb{R}^n$ we will use the following notations

$$\mathbb{E}_\theta [T] = \int_{\mathbb{X}} T \, \mathrm{d}\mathbb{P}_\theta,$$
$$\mathrm{Var}_\theta(T) = \mathbb{E}_\theta \left[\|T - \mathbb{E}_\theta [T]\|^2 \right].$$

Moreover given a σ-algebra $\mathcal{G} \subset \mathcal{F}$ the r.v. $\mathbb{E}_\theta [T \,|\, \mathcal{G}]$ will be the conditional expectation of T given $\mathcal{G}$ w.r.t. $\mathbb{P}_\theta$ and

$$\mathrm{Var}_\theta(T \,|\, \mathcal{G}) = \mathbb{E}_\theta \left[\|T - \mathbb{E}_\theta [T \,|\, \mathcal{G}]\|^2 \,|\, \mathcal{G} \right].$$

Definition 1.1.4: A statistical model $(\mathbb{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ is *dominated* if there exists a σ-finite measure μ s.t.

$$\mathbb{P}_\theta \ll \mu \quad \forall \theta \in \Theta.$$

Definition 1.1.5: Let $(\mathbb{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ be a statistical model dominated by the measure μ. A *likelihood function* of the model is a function $L : \Theta \times \mathbb{X} \to [0, 1]$ s.t. for every $\theta \in \Theta$ the function $L_\theta = L_\cdot(\theta) = L(\theta, \cdot)$ is a density of $\mathbb{P}_\theta$ w.r.t. μ.

Definition 1.1.6: Fix a statistical model $(\mathbb{S}, \mathcal{S}, (P_\theta)_{\theta\in\Theta})$. A *sample* of size $n \in \mathbb{N}_+$ is a list of n i.i.d. r.v.'s $X = (X_1, \dots, X_n)$, $X_j : \Omega \to \mathbb{S}$ for $i = 1, \dots, n$, with common distribution P_θ for some $\theta \in \Theta$. An observation of the sample X is just $X(\omega)$ for an $\omega \in \Omega$.

Remark 1.1.7: In the context of the previous Definition, the law of the sample X is $\mathbb{P}_\theta^{\otimes n}$. In particular if the model is dominated X has density p_θ^n.

In the following when we will consider the *statistical model induced by a sample* $X = (X_1, \dots, X_n)$ where $X_j : \Omega \to \mathbb{S}$ are r.v.'s with values in some measurable space $(\mathbb{S}, \mathcal{S})$ with some given law P_θ dependent of a parameter $\theta \in \Theta$, we will always refer to the statistical model $(\mathbb{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ where $\mathbb{X} = \mathbb{S}^n$, $\mathcal{F} = \mathcal{S}^{\otimes n}$ and $\mathbb{P}_\theta = P_\theta^{\otimes n}$ for every $\theta \in \Theta$ and given a measurable function $f : \mathbb{X} \to E$ for some measurable space $(E, \mathcal{E})$, sometimes we would write with a little abuse of notation $\mathbb{E}_\theta[f(X)], \mathrm{Var}_\theta(f(X))$ to indicate the expectation and th variance w.r.t the r.v. $X \sim \mathbb{P}_\theta$, that is

$$\mathbb{E}_\theta[f(X)] = \mathbb{E}_\theta[f] \quad \text{and} \quad \mathrm{Var}_\theta(f(X)) = \mathrm{Var}_\theta(f).$$

Definition 1.1.8 (Statistic): Fix a statistical model $(\mathbb{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ and a measurable space $(E, \mathcal{E})$. A *statistic* is a measurable map $T : \mathbb{X} \to E$ that does not depend on θ.

Example 1.1.9: For a fixed $m \in \mathbb{N}$ consider $\mathbb{S} = \{1, \dots, m\}$, ν the counting measure and the probabilities

$$P_\theta(\{h\}) = \binom{m}{h}\theta^h(1-\theta)^{n-h}, \quad \text{for } \theta \in \Theta = [0, 1].$$

Then if $\mathcal{S} = \mathcal{P}(\mathbb{S})$, we have that the statistical model $(\mathbb{S}, \mathcal{S}, (P_\theta)_{\theta\in\Theta})$ is dominated by ν with densities $q_{\theta(h)} = P_\theta(\{h\})$.

Consider the statistical model induced by a sample X of size n with $X_j \overset{\text{i.i.d.}}{\sim} P_\theta$, then $T_1(x) = \sum_{i=1}^n x_j$ is a statistic while $T_2(x) = x_1 + \theta$ is not.

Example 1.1.10 (Pareto distribution): Take $\mathbb{S} = [0, \infty)$, $\nu = \mathcal{L}^1_{|[0,\infty)}$ and densities $q_\theta(x) = \sigma(1-\theta)^{-1-\theta}$ for $x > 0$ and $\Theta = (0, \infty)$. Consider the statistical model induced by a sample X of size n with $X_j \overset{\text{i.i.d.}}{\sim} P_\theta$, hence

$$T(x) = \frac{\frac{1}{n}\sum_{j=1}^n x_j}{1 + \frac{1}{n}\sum_{j=1}^n x_j}$$

is a statistic.

Take a (general) statitical model $(\mathbb{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$, the problem is that for a function $g : \Theta \to \Gamma$ we aim to infer $\gamma = g(\theta^*)$ from an observation of a r.v. X defined on $\mathbb{X}$.

Definition 1.1.11 (Estimator): Fix a statitical model $(\mathbb{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ and a function $g : \Theta \to \Gamma$. An *estimator* is a r.v. $\hat{\gamma} : \mathbb{X} \to \Gamma$ that does not depend on θ. The image of an estimator $\hat{\gamma}(X)$ for some r.v. X with values in $\mathbb{X}$ is called an *estimate*.

Remark 1.1.12: When we find an estimator $\hat{\gamma}$ we obtain a function $x \mapsto \hat{\gamma}(x)$ but in practice our statistical model is induced by a sample $X = (X_1, \dots, X_n)$ and we are

interested in the estimate $\hat{\gamma}(X)$ with our sample X. So we are mostly interested in the r.v. $\hat{\gamma}(X)$ rather than just the function $\hat{\gamma}$. Hence, for convenience, when we deal with a practical statistical model induced by a sample X, sometimes we will present an estimator $\hat{\gamma}$ using its estimate $\hat{\gamma}(X)$. From $\hat{\gamma}(X)$ it is easy to recover the actual estimator computing $\hat{\gamma} : x \mapsto \hat{\gamma}(x)$.

Remark 1.1.13: Given r.v.'s $X_1, \dots, X_n$ we will indicate with $(X_{(1)}, \dots, X_{(n)})$ the random vector s.t.

$$X_{(1)}(\omega) \leq \cdots \leq X_{(n)}(\omega) \quad \text{for every } \omega.$$

Example 1.1.14 (Location model): Assume $X_j = \mu + \varepsilon_j$ with $\varepsilon_j \sim P$ where P ha cumulative distribution function F in the family

$$\mathscr{F} = \{F : \mathbb{R} \to [0,1] \mid F \text{ is a cumulative distribution function and } F(-y) = 1 - F(y) \ \ \forall y \in \mathbb{R}\}.$$

Then consider the statistical model $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), (\mathbb{P}_\theta)_{\theta\in\Theta})$ with $\theta = (\mu, F)$ induced by the sample X. We want to estimate the *location* of the mean μ of the law of the $X_j's$, so

$$\mu = \gamma = g(\theta).$$

Possible choices for $\hat{\gamma}$ are:

- the *sample mean* $\hat{\mu}_1(x) = \frac{1}{n}\sum_{j=1}^n x_j$ that minimizes

$$R_1(\mu, x) = \sum_{j=1}^n (x_j - \mu)^2.$$

- the *sample median* $\hat{\mu}_2(x) = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ \frac{x_{(n/2)} + x_{(n/2+1)}}{2} & \text{if } n \text{ is even} \end{cases}$, that minimizes

$$R_2(\mu, x) = \sum_{j=1}^n |x_j - \mu|.$$

- the *Huber estimator* $\hat{\mu}_3(x) = \arg\min_{\mu\in\mathbb{R}} \sum_{j=1}^n \rho_h(\mu - x_j)$, with $\rho_h(y) = \begin{cases} y^2 & \text{if } |y| \leq h \\ h(2|y| + 1) & \text{if } |y| > h \end{cases}$.

1.2 Methods to construct estimators

In this section we will see some methods to construct estimators. Let us fix a statistical model $(\mathbb{S}, \mathcal{S}, (P_\theta)_{\theta\in\Theta})$ with a sample $X = (X_1, \dots, X_n)$ and let $(\mathbb{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ be the statistical model induced by the sample, that is $\mathbb{X} = \mathbb{S}^n$, $\mathcal{F} = \mathcal{S}^{\otimes n}$, $\mathbb{P}_\theta = P_\theta^{\otimes n}$ for every $\theta \in \Theta$.

1.2.1 Maximum likelihood method

Assume that the statistical model $(\mathbb{S}, \mathcal{S}, (P_\theta)_{\theta\in\Theta})$ is dominated with densities $(q_\theta)_{\theta\in\Theta}$, hence a likelihood is given by

$$L_x(\theta) = \prod_{j=1}^n q_\theta(x_j) \quad \forall \theta \in \Theta.$$

Definition 1.2.1 (Maximum likelihood estimator): The *maximum likelihood estimator (MLE)* of θ is

$$\hat{\theta}(x) = \arg\max_{\theta\in\Theta} L_x(\theta) = \arg\max_{\theta\in\Theta} \log L_x(\theta).$$

Example 1.2.2 (Location model):

- Assume $F \in \mathscr{F}_1 = \{y \mapsto \Phi(y/\sigma) \mid \sigma \in (0,\infty)\}$, where Φ is the cumulative distribution function of the standard normal measure. Then $X_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu,\sigma^2)$ and the MLE of μ is

$$\begin{aligned}\hat{\mu}(x) &= \pi_1\left(\arg\max_{(\mu,\sigma)\in\mathbb{R}\times(0,\infty)} \log L_x(\mu,\sigma)\right)\\ &= \pi_1\left(\arg\max_{(\mu,\sigma)\in\mathbb{R}\times(0,\infty)} -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma}\sum_{j=1}^n (x_j-\mu)^2\right)\\ &= \arg\min_{\mu\in\mathbb{R}} \sum_{j=1}^n (\mu - x_j)^2 = \frac{1}{n}\sum_{j=1}^n x_j.\end{aligned}$$

  where π_1 is the projection in the first component. In particular the estimate of μ is the r.v. $\hat{\mu}(X) = \frac{1}{n}\sum_{j=1}^n X_j$.

- Assume $F \in \mathscr{F}_2 = \left\{F(y) = \int_{-\infty}^y \frac{1}{2\sigma}\exp\left(-\frac{|z|}{\sigma}\right) \mathrm{d}z \mid \sigma \in (0,\infty)\right\}$ (distribution functions of the Laplace laws). Then $X_j \overset{\text{i.i.d.}}{\sim} \text{Laplace}(\mu,\sigma)$ and the MLE estimator for μ is

$$\begin{aligned}\hat{\mu}(x) &= \pi_1\left(\arg\max_{(\mu,\sigma)\in\mathbb{R}\times(0,\infty)} \log L_x(\mu,\sigma)\right)\\ &= \pi_1\left(\arg\max_{(\mu,\sigma)\in\mathbb{R}\times(0,\infty)} -n\log\left(\frac{1}{2\sigma}\right) - \frac{1}{\sigma}\sum_{j=1}^n |x_j-\mu|\right)\\ &= \arg\min_{\mu\in\mathbb{R}} \sum_{j=1}^n |\mu - x_j| = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd}\\ \frac{x_{(n/2)}+x_{(n/2+1)}}{2} & \text{if } n \text{ is even}\end{cases}.\end{aligned}$$

  In particular the estimate of μ is the r.v.

$$\hat{\mu}(X) = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd}\\ \frac{X_{(n/2)}+X_{(n/2+1)}}{2} & \text{if } n \text{ is even}\end{cases}.$$

1.2.2 Method of moments

Definition 1.2.3: Let $x = (x_1,\ldots,x_n) \in \mathbb{R}^n$, the *empirical distribution* associated to this numbers is the probability measure with cumulative distribution function

$$\hat{\mathbb{P}}^x = \frac{1}{n}\sum_{j=1}^n \delta_{x_j}$$

where δ_y for $y \in \mathbb{S}$ is the Dirac's delta measure located in y.

Remark 1.2.4: Of course given a sample $X = (X_1,\ldots,X_n)$ we can consider the empirical distribution associated to X that is the random measure

$$\hat{\mathbb{P}}^X = \frac{1}{n}\sum_{j=1}^n \delta_{X_j}.$$

If X is a sample from the distribution ν, the empirical distribution associated to X can be seen as an approximator of it.

The idea of the method is to match the moments of the empirical distribution associated to our sample with $\mathbb{P}_\theta$ varying $\theta \in \Theta$..

Definition 1.2.5: Take $j \in \mathbb{N}_+$. The *j-th moment* of a probability measure ν is

$$m_j(\nu) = \mathbb{E}\left[Y^j\right] \quad \text{where } Y \sim \nu.$$

In our case were we have a statistical model we will indicate $m_j(\theta) = m_j(\mathbb{P}_\theta)$ for every $\theta \in \Theta$.

Definition 1.2.6 (Moment estimator): Let $\theta \in \Theta \subset \mathbb{R}^p$ for some $p \in \mathbb{N}_+$ and let $\int_{\mathbb{X}} |x|^p \, \mathrm{d}\mathbb{P}_\theta(x) < +\infty$ for every $\theta \in \Theta$. We define $m : \Theta \to \mathbb{R}^p$ s.t.

$$m(\theta) = (m_j(\theta))_{j=1}^p.$$

Moreover if m has a continuous inverse we define the *moment estimator*

$$\hat{\theta}(x) = m^{-1}\left(\hat{m}_1(x), \dots, \hat{m}_p(x)\right)$$

where $\hat{m}_j(x) = m_j(\hat{\mathbb{P}}^x)$ for every $j = 1, \dots, p$.

Example 1.2.7 (Normal): $X_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, then considering the statistical model induced by the sample

$$\begin{cases} m_1(\mu, \sigma^2) = \mu \\ m_2(\mu, \sigma^2) = \sigma^2 + \mu^2 \end{cases} \Rightarrow \begin{cases} \hat{\mu} = \hat{m}_1 = \frac{1}{n}\sum_{j=1}^n x_j \\ \hat{\sigma}^2 + \hat{\mu}^2 = \hat{m}_2 = \frac{1}{n}\sum_{j=1}^n x_j^2 \end{cases} \Rightarrow \begin{cases} \hat{\mu} = \frac{1}{n}\sum_{j=1}^n x_j \\ \hat{\sigma}^2 = \hat{m}_2 - \hat{m}_1^2 = \frac{1}{n}\sum_{j=1}^n (x_j - \hat{m}_1)^2 \end{cases}$$

Example 1.2.8 (Negative binomial): $X_j \overset{\text{i.i.d.}}{\sim} \mathrm{NegBin}(k, \theta)$ with $\theta \in (0, 1)$ and k fixed. We have to match only the first moment:

$$m(\theta) = m_1(\theta) = \mathbb{E}_\theta[X] = k\frac{1-\theta}{\theta} \Rightarrow \hat{\theta} = m^{-1}(\hat{m}_1) = \frac{k}{\hat{m}_1 + k}$$

Example 1.2.9 (Pareto distribution): $p_\theta(x) = \theta(1+x)^{-(1+\theta)}$ for $x > 0$, $\theta \in (0, \infty) = \Theta$. Again we have to match only the first moment. If $\theta > 1$ then

$$m(\theta) = m_1(\theta) = \frac{1}{\theta - 1} \Rightarrow \hat{\theta} = m^{-1}(\hat{m}_1) = 1 + \frac{1}{\hat{m}_1}.$$

1.2.3 Plug-in estimators

In moments estimators we:

(1) found the map $m : \Theta \to \mathbb{R}^p$;

(2) substituted the empirical moments (the evaluation on the empirical measure of an extension of m on a $\mathcal{M} \subset \mathcal{M}_1(\mathbb{X})$ that contains $(\mathbb{P}_\theta)_{\theta\in\Theta}$ and the empirical measures) in the inverse of the map m.

However, this procedure can be applied with applied with different maps $Q : \mathcal{M} \to \Theta$ with $\mathcal{M} \subset \mathcal{M}_1(\mathbb{X})$ that contains $(\mathbb{P}_\theta)_{\theta\in\Theta}$ and the empirical measures.

Definition 1.2.10: Let $Q : \mathcal{M} \to \Theta$ with $\mathcal{M} \subset \mathcal{M}_1(\mathbb{X})$ that contains $(\mathbb{P}_\theta)_{\theta\in\Theta}$ and the empirical measures, then

$$T(x) = Q(\hat{\mathbb{P}}^x)$$

is called *plug-in estimator* for θ associated to Q.

Example 1.2.11:

- Suppose $\Theta \subset \mathbb{R}^p$, the map

$$Q(\nu) = m^{-1}\left(\int_{\mathbb{X}} x \,\mathrm{d}\nu(x), \ldots, \int_{\mathbb{X}} x^p \,\mathrm{d}\nu(x)\right)$$

gives the method of moments.

- The map

$$Q(\nu) = \arg\max_{\theta\in\Theta} \int_{\mathbb{X}} \log(L_\theta(x)) \,\mathrm{d}\nu(x)$$

gives the MLE. Indeed

$$Q(\mathbb{P}_n^x) = \arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{j=1}^{n} \log(q_\theta(x_j)) = \arg\max_{\theta\in\Theta} L_x(\theta).$$

Moreover the following holds.

Lemma 1.2.12: *For every $\theta_0 \in \Theta$ holds $Q(\mathbb{P}_{\theta_0}) = \arg\max_{\theta\in\Theta} \mathbb{E}_{\theta_0}[\log(p_{\theta_0})] = \theta_0$.*

Proof. Remember that for $\theta > 0$ holds $\log(x) \le x - 1$, hence

$$\begin{aligned}
\mathbb{E}_{\theta_0}[\log(L_\theta)] - \mathbb{E}_{\theta_0}[\log(L_{\theta_0})] &= \mathbb{E}_{\theta_0}\left[\log\left(\frac{L_\theta}{L_{\theta_0}}\right)\right] \\
&\le \mathbb{E}_{\theta_0}\left[\frac{L_\theta}{L_{\theta_0}} - 1\right] \\
&= \int_{\mathbb{X}} L_\theta(x) \,\mathrm{d}\mu(x) - \int_{\mathbb{X}} L_{\theta_0}(x) \,\mathrm{d}\mu(x) = 0
\end{aligned}$$

for every $\theta > 0$, so the thesis follows. □

1.3 Sufficiency and Neyman-Fisher factorization

Definition 1.3.1: Let $(\mathbb{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ be a statistical model dominated by the measure μ and $(E, \mathcal{E})$ a measurable space. A statistic $S : \Omega \to E$ is *sufficient* if $\mathbb{P}_\theta(\cdot \,|\, S)$ does not depend on θ.

Remark 1.3.2: If $(\mathbb{X}, \mathcal{F})$ is a Borel space (any polish space is Borel) then regular conditional densities exists. So in this case, in the context of the previous Definition, the statistic S is sufficient if and only if the regular conditional probability $k_\theta(s, \cdot)$ of $\mathbb{P}_\theta$ w.r.t. S does not depend on θ.

For example when $\mathbb{X}$ is discrete we have $k_\theta(s, \{x\}) = \mathbb{P}_\theta(\{x\} \,|\, S = s)$, where $S = s$ means as always $\{x \in \mathbb{X} \,|\, S(x) = s\}$.

Example 1.3.3: Consider the statistical model induced by the sample $X = (X_1, \dots, X_n)$, with $X_j \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\theta)$ and take $S(x) = \sum_{j=1}^n x_j$. Hence taking $\mathbb{P}_\theta$ the law of $X = (X_1, \dots, X_n)$

$$\mathbb{P}_\theta(\{(x_1, \dots, x_n)\} \,|\, S = s) = \begin{cases} 0 & \text{if } s \neq \sum_{j=1}^n x_j \\ \frac{\prod_{j=1}^n e^{-\theta} \frac{\theta^{x_j}}{x_j!}}{e^{-n\theta} \frac{(n\theta)^s}{s!}} & \text{if } s = \sum_{j=1}^n x_j \end{cases}$$

and

$$\frac{\prod_{j=1}^n e^{-\theta} \frac{\theta^{x_j}}{x_j!}}{e^{-n\theta} \frac{(n\theta)^s}{s!}} = \frac{s! e^{-n\theta} \theta^{x_1 + \dots + x_n} \prod_{j=1}^n \frac{1}{x_j!}}{e^{-n\theta} n^s \theta^t} = \frac{s!}{n^s \prod_{j=1}^n x_j!}$$

that does not depend on θ, so S is a sufficient statistic.

Theorem 1.3.4 (Neyman-Fisher factorization): *Consider a statistical model $(\mathbb{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ dominated by the measure μ, a measurable space $(E, \mathcal{E})$ and a statistic $S : \mathbb{X} \to E$. The following are equivalent:*

(1) S is a sufficient statistic;

(2) given $L(\theta, x)$ a likelihood function, for every $\theta \in \Theta$ and μ-a.e. $x \in \mathbb{X}$ holds that the likelihood is of the form:

$$L(\theta, x) = g_\theta(S(x)) h(x).$$

Proof of the discrete case. Assume that $\mathbb{X}$ is discrete and μ is the counting measure. Take Q_θ the law of S, that is $Q_\theta(\{s\}) = \sum_{\substack{x \in \mathbb{X} \\ S(x) = s}} \mathbb{P}_\theta(\{x\})$. For every $x \in \mathbb{X}$ s.t. $S(x) = s$ we have

$$\mathbb{P}_\theta(\{x\} \,|\, S = s) = \begin{cases} 0 & \text{if } S(x) \neq s \\ \frac{\mathbb{P}_\theta(\{x\})}{Q_\theta(\{s\})} & \text{if } S(x) = s \end{cases}.$$

If (1) holds, then S is sufficient we have that $\mathbb{P}_\theta(\{x\} \,|\, S = s) = h(x)$ does not depend on θ and for every $x \in \mathbb{X}$ taking $s = S(x)$ we can write

$$L(\theta, x) = \mathbb{P}_\theta(\{x\}) = \mathbb{P}_\theta(\{x\} \,|\, S = s) Q_\theta(\{t\}) = h(x) g_\theta(S(x)).$$

Viceversa if (2) holds, then

$$Q_\theta(\{s\}) = \sum_{\substack{x \in \mathbb{X} \\ S(x) = s}} \mathbb{P}_\theta(\{x\}) = g_\theta(t) \sum_{\substack{x \in \mathbb{X} \\ S(x) = s}} h(x)$$

hence for every $x \in \mathbb{X}$ s.t. $S(x) = s$ holds

$$\mathbb{P}_\theta(\{x\} \,|\, S = s) = \frac{g_\theta(s) h(x)}{g_\theta(s) \sum_{\substack{x \in \mathbb{X} \\ S(x) = s}} h(x)} = \frac{h(x)}{\sum_{\substack{x \in \mathbb{X} \\ S(x) = s}} h(x)}$$

that does not depend on θ. □

1.4 Bias and the Blackwell-Rao theorem

Fix a statistical model $(\mathbb{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$, a function of the parameter $g : \Theta \to \mathbb{R}^n$ and an estimator $T : \mathbb{X} \to \mathbb{R}^n$ of $g(\theta)$.

Definition 1.4.1: The *bias* of the estimator T is

$$\text{bias}_\theta(T) = \mathbb{E}_\theta[T] - g(\theta).$$

The estimator T is called *unbiased* if $\text{bias}_\theta(T) = 0$ for every $\theta \in \Theta$.

Example 1.4.2:

- Let $X \sim \text{Bin}(n, \theta)$ with $\theta \in (0, 1)$. Take $\text{Bin}(n, \theta)$ with $\theta \in (0, 1)$. Take $T(x) = \frac{x}{n}$ where x is the number of successes, then
$$\text{bias}_{\theta\in\Theta}(T) = \mathbb{E}_\theta[T] - \theta = \frac{1}{n}\mathbb{E}_\theta[x] - \theta = \frac{1}{n}n\theta - \theta = 0$$
so T is an unbiased estimator for θ.
- Let $X = (X_1 \ldots, X_n)$ be a sample with $X_j \sim \text{Poisson}(\theta)$. Consider $T(x) = \frac{1}{n}\sum_{j=1}^n x_j$ then
$$\text{bias}_{\theta\in\Theta}(T) = \mathbb{E}_\theta[T] - \theta = \frac{1}{n}\sum_{j=1}^n \mathbb{E}_\theta[x_j] - \theta = \frac{1}{n}n\theta - \theta = 0$$
hence T is an unbiased estimator for θ.
- Let $X = (X_1 \ldots, X_n)$ be a sample with $X_j \sim \mathcal{N}(\mu, \sigma^2)$ with $\theta = (\mu, \sigma^2) \in \mathbb{R} \times [0, \infty)$. One can check that $\overline{x} = \frac{1}{n}\sum_{j=1}^n x_j$ is an unbiased estimator for μ. Moreover $S^2(x) = \frac{1}{n-1}\sum_{j=1}^n (x_n - \overline{x})^2$ is an unbiased estimator for σ^2 while $\overline{\sigma}^2 = \frac{1}{n}\sum_{j=1}^n (x_n - \overline{x})^2$ is a biased estimator for σ^2.

Definition 1.4.3: If $\mathbb{E}_\theta\left[\|T\|^2\right] < +\infty$ for every $\theta \in \Theta$, the *mean squared error (MSE)* of the estimator T is

$$\text{MSE}_\theta(T) = \mathbb{E}_\theta\left[\|T - g(\theta)\|^2\right].$$

Given a second estimator $S : \Theta \to \mathbb{R}^n$ of $g(\theta)$ s.t. $\mathbb{E}_\theta\left[\|S\|^2\right] < +\infty$ for every $\theta \in \Theta$ it is called *preferable* if $\text{MSE}_\theta(S) \le \text{MSE}_\theta(T)$ for every $\theta \in \Theta$ and *strictly preferable* if $\text{MSE}_\theta(S) < \text{MSE}_\theta(T)$ for every $\theta \in \Theta$.

Consider $\mathcal{D}$ a family of estimators of $g(\theta)$, an estimator $T \in \mathcal{D}$ is called *optimal* among the estimators in $\mathcal{D}$ if it is preferable to any other estimator in $\mathcal{D}$.

Lemma 1.4.4 (Bias-variance decomposition): *Suppose that* $\mathbb{E}_\theta\left[\|T\|^2\right] < +\infty$ *for every* $\theta \in \Theta$, *then*

$$\text{MSE}_\theta(T) = \|\text{bias}_\theta(T)\|^2 + \text{Var}_\theta(T) \quad \textit{for every } \theta \in \Theta.$$

Proof. Holds that

$$\begin{aligned}
\mathbb{E}_\theta\left[\|T - g(\theta)\|^2\right] &= \mathbb{E}_\theta\left[\|T\|^2\right] + \|g(\theta)\|^2 - 2\langle g(\theta), \mathbb{E}_\theta[T]\rangle \\
&= \mathbb{E}_\theta\left[\|T\|^2\right] - \|\mathbb{E}_\theta[T]\|^2 + \|\mathbb{E}_\theta[T]\|^2 + \|g(\theta)\|^2 - 2\langle g(\theta), \mathbb{E}_\theta[T]\rangle \\
&= \text{Var}_\theta(T) + \|\text{bias}_\theta(T)\|^2
\end{aligned}$$

□

Example 1.4.5: Consider the statistical model induced by the sample $X = (X_1, \dots, X_n)$ with $X_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ defined on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and $S^2(x) = \frac{1}{n-1}\sum_{j=1}^n (x_j - \overline{x})^2$, $\hat{\sigma}^2(x) = \frac{1}{n}\sum_{j=1}^n (x_j - \overline{x})^2$ with $\overline{x} = \frac{1}{n}\sum_{j=1}^n x_j$. One can check that S^2 is an unbiased estimator for σ^2 while $\hat{\sigma}^2$ is not, but we are going to see that at the same time the MSE of S^2 is much bigger that the MSE of $\hat{\sigma}^2$. So be unbiased is not "enough" to obtain a low MSE.

Recall that $\frac{1}{\sigma^2}\sum_{j=1}^n (X_j - \overline{X})^2 \sim \chi^2_{n-1} = \Gamma(\frac{n-1}{2}, \frac{1}{2})$ and that if $Y \sim \Gamma(k, \lambda)$ then the mean of Y is $\frac{k}{\lambda}$ and its variance is $\frac{k}{\lambda^2}$. Then

$$\mathbb{E}\left[\frac{1}{\sigma^2}\sum_{j=1}^n (X_j - \overline{X})^2\right] = \frac{n-1}{2}2 = n-1$$

$$\mathrm{Var}\left(\frac{1}{\sigma^2}\sum_{j=1}^n (X_j - \overline{X})^2\right) = \frac{n-1}{2}4 = 2(n-1)$$

hence

$$\mathrm{MSE}_\theta(S^2) = \mathrm{Var}_\theta(S^2) = \mathrm{Var}_\theta\left(\frac{\sigma^2}{n-1}\frac{1}{\sigma^2}\sum_{j=1}^n (X_j - \overline{X})^2\right) = \frac{\sigma^4}{(n-1)^2}2(n-1) = \frac{2\sigma^4}{n-1}$$

and

$$\begin{aligned}
\mathrm{MSE}_\theta(\hat{\sigma}^2) &= \mathrm{bias}_\theta(\hat{\sigma}^2) + \mathrm{Var}_\theta(\hat{\sigma}^2) \\
&= \left(\mathbb{E}_\theta\left[\frac{n-1}{n}S^2\right] - \sigma^2\right)^2 + \mathrm{Var}_\theta(\hat{\sigma}^2) \\
&= \left(\frac{n-1}{n}\sigma^2 - \sigma^2\right)^2 + \left(\frac{n-1}{n}\right)^2 \mathrm{Var}_\theta(S^2) \\
&= \frac{1}{n^2}\sigma^4 + \frac{2(n-1)\sigma^4}{n^2} = \frac{(2n-1)\sigma^4}{n^2}
\end{aligned}$$

and actually $\mathrm{MSE}_\theta(\hat{\sigma}^2) \leq \mathrm{MSE}_\theta(S^2)$.

Now we fix a measurable space $(E, \mathcal{E})$. We want to use a sufficient statistic $S : \mathbb{X} \to E$ to reduce the variance of an estimator. From now on we suppose that for every $\theta \in \Theta$ the probability $\mathbb{P}_\theta$ admits a regular conditional probability $k_\theta(s, \cdot)$ w.r.t. S (for example if $(\mathbb{X}, \mathcal{F})$ is a Borel space).

Remark 1.4.6: If $S : \mathbb{X} \to E$ is a sufficient statistic and T is an estimator for $g(\theta)$, then $\mathbb{E}_\theta[T \mid S]$ actually does not depend on θ since

$$\mathbb{E}_\theta[T \mid S = s] = \int_{\mathbb{X}} T(y) k_\theta(s, \mathrm{d}y)$$

and k_θ does not depend on θ since S is sufficient. In this cases we will write $\mathbb{E}_*[T \mid S]$ to point out the independence from θ.

Theorem 1.4.7 (Blackwell-Rao): *Let $S : \mathbb{X} \to E$ be a sufficient statistic and let U be an estimator of $g(\theta)$ with $\mathbb{E}_\theta[\|U\|^2] < +\infty$ for every $\theta \in \Theta$. Define $V = \mathbb{E}_*[U \mid S]$, then*

(1) if U is unbiased, so is V;

(2) $MSE_\theta(V) \leq \mathrm{MSE}_\theta(U)$ *for every* $\theta \in \Theta$.

Furthermore, either $U = h(S)$ $\mathbb{P}_\theta$*-a.s for every* $\theta \in \Theta$ *for some measurable function* h *or there exists* $\theta_0 \in \Theta$ *s.t. the inequality in (2) is strict.*

Proof. (1) We have $\mathbb{E}_\theta[V] = \mathbb{E}_\theta[U]$ by definition of conditional expectation.

(2) We have

$$\mathrm{MSE}_\theta(V) = \|\mathbb{E}_\theta[V] - g(\theta)\|^2 + \mathrm{Var}_\theta(V)$$

and

$$\mathrm{MSE}_\theta(U) = \|\mathbb{E}_\theta[U] - g(\theta)\|^2 + \mathrm{Var}_\theta(U)$$

but $\mathbb{E}_\theta[V] = \mathbb{E}_\theta[U]$ so since

$$\mathrm{Var}_\theta(\cdot) = \mathbb{E}_\theta[\mathrm{Var}_*(\cdot \mid S)] + \mathrm{Var}_\theta(\mathbb{E}_*[\cdot \mid S])$$

we can write

$$\begin{aligned}\mathrm{MSE}_\theta(U) - \mathrm{MSE}_\theta(V) &= \mathrm{Var}_\theta(U) - \mathrm{Var}_\theta(V)\\ &= \mathrm{Var}_\theta(\mathbb{E}_*[U \mid S]) + \mathbb{E}_\theta[\mathrm{Var}_*(U \mid S)] - \mathrm{Var}_\theta(\mathbb{E}_*[U \mid S])\\ &= \mathbb{E}_\theta[\mathrm{Var}_*(U \mid S)] \geq 0.\end{aligned}$$

Assume now that for all measurable functions h there exist $\theta(h) \in \Theta$ s.t. $\mathbb{P}_\theta(h(S) \neq U) > 0$. By Doob's measurability criterion there exists a measurable function h_0 s.t. $V = h_0(S)$, hence

$$\begin{aligned}\mathbb{E}_\theta[\mathrm{Var}_*(U \mid S)] &= \mathbb{E}_\theta\left[\mathbb{E}_*\left[(U-V)^2 \mid S\right]\right]\\ &= \mathbb{E}_\theta\left[\mathbb{E}_*\left[(U-h_0(S))^2 \mid S\right]\right]\\ &= \mathbb{E}_\theta\left[(U-h_0(S))^2\right] > 0\end{aligned}$$

choosing $\theta = \theta_0 = \theta(h_0)$. □

Example 1.4.8: We consider the statistical model induced by the sample $X = (X_1, \ldots, X_n)$, where $X_j \overset{\text{i.i.d.}}{\sim} \mathrm{Poisson}(\theta)$. Take $g(\theta) = e^{-\theta}$. A possible esimator for $g(\theta)$ is $T(x) = \mathbb{1}(x_1 = 0)$, that is unbiased since

$$\mathbb{E}_\theta[T] = \mathbb{P}_\theta(\{x_1 = 0\}) = e^{-\theta} = g(\theta).$$

Probably this estimator has high variance, so we want to use Blackwell-Rao Theorem 1.4.7 to improve it. Remember that in this case $S(x) = \sum_{j=1}^n x_j$ is a sufficient statistic, so we have

$$\begin{aligned}\mathbb{E}_*[T \mid S = s] &= \mathbb{P}_\theta(\{x_1 = 0\} \mid S = s)\\ &= \frac{\mathbb{P}_\theta(S = s \mid \{x_1 = 0\})\mathbb{P}_\theta(\{x_1 = 0\})}{\mathbb{P}_\theta(S = s)}\\ &= \frac{\mathbb{P}_\theta\left(\left\{\sum_{j=2}^n x_j = s\right\}\right)\mathbb{P}_\theta(\{x_1 = 0\})}{\mathbb{P}_\theta(S = s)}\\ &= \frac{e^{-(n-1)\theta}\frac{((n-1)\theta)^s}{s!}e^{-\theta}}{e^{-n\theta}\frac{(n\theta)^s}{s!}e^{-\theta}} = \left(\frac{n-1}{n}\right)^s\end{aligned}$$

so a better estimator for $g(\theta)$ is $V = \mathbb{E}_*[T \mid S] = \left(\frac{n-1}{n}\right)^S$ and it remains unbiased (by the mentioned theorem).

1.5 Completeness and UMVU estimators

Fix a statistical model $(\mathbb{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$, a measurable space $(E, \mathcal{E})$, a function of the parameter $g : \Theta \to \mathbb{R}^n$ and an estimator $T : \mathbb{X} \to \mathbb{R}^n$ of $g(\theta)$.

Definition 1.5.1: An unbiased estimator T^* is *unifom minimum variance unbiased (UMVU)* if for any other estimator T we have $\mathrm{Var}_\theta(T^*) \leq \mathrm{Var}_\theta(T)$ for every $\theta \in \Theta$.

Definition 1.5.2: A statistic $S : \mathbb{X} \to E$ is *complete* if for every $h : E \to \mathbb{R}$ measurable and indipendent of θ holds that for every $\theta \in \Theta$

$$\mathbb{E}_\theta [h(S)] = 0 \Rightarrow h(S) = 0 \ \ \mathbb{P}_\theta\text{-a.s. .}$$

Lemma 1.5.3 (Lehmann-Scheffè): *Let T be an unbiased estimator for $g(\theta)$ with $\mathbb{E}_\theta [\|T\|^2] < +\infty$ for every $\theta \in \Theta$. Let S be a sufficient and complete statistic, then:*

(1) $T^* = \mathbb{E}_* [T \mid S]$ is UMVU;

(2) for every unbiased estimator $\tilde{T}$ of $g(\theta)$ with $\mathbb{E}_\theta \left[\|\tilde{T}\|^2\right] < +\infty$ for every $\theta \in \Theta$ holds that $\mathbb{E}_* \left[\tilde{T} \mid S\right] = T^*$ $\mathbb{P}_\theta$-a.s. for every $\theta \in \Theta$.

Proof. Using Blackwell-Rao Theorem 1.4.7 we get

$$\mathrm{Var}_\theta(\mathbb{E}_* \left[\tilde{T} \mid S\right]) = \mathrm{MSE}_\theta(\mathbb{E}_* \left[\tilde{T} \mid S\right]) \leq \mathrm{MSE}_\theta(\tilde{T}) = \mathrm{Var}_\theta(\tilde{T})$$

for every unbiased estimator $\tilde{T}$ of $g(\theta)$ with $\mathbb{E}_\theta \left[\|\tilde{T}\|^2\right] < +\infty$ for every $\theta \in \Theta$. Hence we can search for the unbiased estimator with minimal variance inside the family of estimators of the form $\mathbb{E}_* \left[\tilde{T} \mid S\right]$ with $\tilde{T}$ like before. Now if $\tilde{T}$ and $\tilde{T}'$ are two estimators like before we get

$$\mathbb{E}_\theta \left[\mathbb{E}_* \left[\tilde{T} \mid S\right] - \mathbb{E}_* \left[\tilde{T}' \mid S\right]\right] = g(\theta) - g(\theta) = 0 \ \ \forall \theta \in \Theta$$

so using the completeness of S we get $\mathbb{E}_* \left[\tilde{T} \mid S\right] = \mathbb{E}_* \left[\tilde{T}' \mid S\right]$ $\mathbb{P}_\theta$-a.s. for every $\theta \in \Theta$. □

Example 1.5.4: (1) Consider the statistical model induced by the sample $X = (X_1, \ldots, X_n)$ with $X_j \overset{\text{i.i.d.}}{\sim} \mathrm{Poisson}(\theta)$ and $g(\theta) = e^{-\theta}$, $\theta \in \Theta = (0, \infty)$. Let us see that the sufficient statistic $S(x) = \sum_{j=1}^n x_j$ (that is sufficient) is also complete. Let h be s.t.

$$\mathbb{E}_\theta [h(S)] = \sum_{s\in\mathbb{N}} h(s) e^{-n\theta} \frac{(n\theta)^s}{s!} = 0$$

we are going to see that $h = 0$ $\mathbb{P}_\theta$-a.s. for every $\theta > 0$. The sum $\sum_{s\in\mathbb{N}} h(s) e^{-n\theta} \frac{(n\theta)^s}{s!}$ can be interpreted as a Taylor expansion of the costantly 0 function (because the sum is 0 by choice of h), then every coefficient is 0, that implies $h(s) = 0$ for every $s \in \mathbb{N}$, that is $h = 0$ and this holds for very θ. Hence S is complete and from that follows that $T^* = \mathbb{E}_* [T \mid S] = \left(\frac{n-1}{n}\right)^S$ is UMVU.

(2) Consider the statistical model induced by the sample $X = (X_1, \ldots, X_n)$ with $X_j \overset{\text{i.i.d.}}{\sim}$ Uniform(θ) and $g(\theta) = \theta$, $\theta \in \Theta = (0, \infty)$. The statistic $S(x) = \max_{1\leq j\leq n} x_j$ is sufficient and has cumulative distribution function $F_\theta(s) = \left(\frac{s}{\theta}\right)^n$ for $s \in [0, \theta]$. Hence

S has density $f_\theta(s) = \frac{ns^{n-1}}{\theta^n}\mathbb{1}(s \in [0,\theta])$. Let us see that S is complete. If h is a measurable function s.t.

$$\mathbb{E}_\theta[h(S)] = \int_0^\theta h(s) f_\theta(s\,\mathrm{d}s) = 0 \quad \forall \theta \in \Theta$$

then $\int_0^\theta h(s)s^{n-1}\,\mathrm{d}s = 0$for every $\theta \in \Theta = (0,\infty)$ and then $h(s) = 0$ for every $s > 0$. We have

$$\mathbb{E}_\theta[S] = \int_0^\theta s\frac{ns^{n-1}}{\theta^n}\,\mathrm{d}s = \frac{n}{n+1}\theta$$

hence $T = \frac{n+1}{n}S$ is an unbiased estimator of θ and $\mathbb{E}_*[T \mid S] = T = \frac{n+1}{n}S$ is UMVU.

1.6 Fisher information and Cramèr-Rao lower bound

Let us fix a statistical model $(\mathbb{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ dominated by a measure μ with densities $(p_\theta)_{\theta\in\Theta}$. Consider the following assumptions:

(1) assume $\Theta \subset \mathbb{R}^k$ to be open and that $p_\theta(x) > 0$ for every $x \in \mathbb{X}$ and $\theta \in \Theta$;

(2) for all measurable function $f \in L^2(\mathbb{X}, \mu)$ holds

$$\nabla_\theta \mathbb{E}_\theta[f] = \nabla_\theta \mathbb{E}_\mu[p_\theta\, f] = \mathbb{E}_\mu[\nabla_\theta p_\theta\, f];$$

(3) for every $\theta \in \Theta$ holds $\mathbb{E}_\theta[\|\nabla_\theta \log(p_\theta)\|^2] < +\infty$.

In the following, given a two r.v.'s $Y_1, Y_2 : \mathbb{X} \to \mathbb{R}$, we will use the notation

$$\mathrm{Cov}_\theta(Y_1, Y_2) = \mathbb{E}_\theta[(Y_1 - \mathbb{E}_\theta[Y_1])(Y_2 - \mathbb{E}_\theta[Y_2])]$$

for the covariance between Y_1 and Y_2 under $\mathbb{P}_\theta$. Furthermore given a r.v. $Y = (Y_1, \dots, Y_k) : \mathbb{X} \to \mathbb{R}^k$ we will call $\mathrm{Cov}_\theta(Y)$ the covariance matrix of the r.v. Y under $\mathbb{P}_\theta$.

Remark 1.6.1: Assumption (1) does holds for example for gaussians distributions but not for the uniform ones on $\mathbb{R}$.

Remark 1.6.2: Assumption (2) holds for example if $\nabla_\theta p_\theta$ exists and $\nabla_\theta p_\theta \in L^2(\mathbb{X}, \mu)$. for every $\theta \in \Theta$.

Lemma 1.6.3: *Under the assumptions (1),(2) we have* $\mathbb{E}_\theta[\nabla_\theta \log(p_\theta)] = 0$ *for every* $\theta \in \Theta$.

Proof.

$$\begin{aligned}
\mathbb{E}_\theta[\nabla_\theta \log(p_\theta)] &= \mathbb{E}_\theta\left[\frac{1}{p_\theta}\nabla_\theta p_\theta\right] \\
&= \int_{\mathbb{X}} \frac{1}{p_\theta}(\nabla_\theta p_\theta) p_\theta\,\mathrm{d}\mu \\
&= \int_{\mathbb{X}} \nabla_\theta p_\theta\,\mathrm{d}\mu \\
&= \nabla_\theta \int_{\mathbb{X}} p_\theta\,\mathrm{d}\mu = \nabla_\theta(1) = 0.
\end{aligned}$$

□

Definition 1.6.4 (Fisher information matrix): Let assumption (1)+(2)+(3) hold. We call the matrix-valued function $\mathcal{I} : \Theta \to \mathbb{R}^{k\times k}$ with entries

$$\mathcal{I}(\theta)_{ij} = \mathbb{E}_\theta \left[(\partial_{\theta_i} \log(p_\theta))(\partial_{\theta_j} \log(p_\theta)) \right] \quad \forall i,j \in \{1,\dots,k\}$$

the *Fisher information matrix* of the dominated model $(p_\theta)_{\theta\in\Theta}$. Further we call the function $s : \Theta \times \mathbb{X} \to \mathbb{R}$ s.t.

$$s(\theta, x) = s_\theta(x) = \nabla_\theta \log(p_\theta(x))$$

the *score function* of the dominated model $(p_\theta)_{\theta\in\Theta}$.

Remark 1.6.5: By the previous Lemma the Fisher information matrix is the covariance matrix for $Y_j = \partial_{\theta_j} \log(p_\theta) : \mathbb{X} \to \mathbb{R}$ with $j \in \{1,\dots,k\}$ since

$$\mathrm{Cov}_\theta(Y_i, Y_j) = \mathbb{E}_\theta \left[(Y_i - \mathbb{E}_\theta [Y_i])(Y_j - \mathbb{E}_\theta [Y_j]) \right] = \mathbb{E}_\theta [Y_i Y_j] = \mathcal{I}(\theta)_{ij}$$

for every $i, j \in \{1,\dots,k\}$.

Remark 1.6.6: Alternatively holds

$$\mathcal{I}(\theta)_{ij} = \mathbb{E}_\theta \left[\frac{1}{p_\theta} (\partial_{\theta_i} p_\theta)(\partial_{\theta_j} p_\theta) \right] \quad \forall i,j \in \{1,\dots,k\}.$$

Remark 1.6.7: The Fisher information matrix is symmetric and positive semidefinite (since it is a covariance matrix).

Lemma 1.6.8: *Let the assumptions (1),(2),(3) hold. For every $i, j \in \{1,\dots,k\}$ holds*

$$\mathbb{E}_\theta \left[\frac{1}{p_\theta} \partial^2_{\theta_i,\theta_j} p_\theta \right] = 0 \quad \forall \theta \in \Theta.$$

Proof. We have

$$\begin{aligned} \mathbb{E}_\theta \left[\frac{1}{p_\theta} \partial^2_{\theta_i,\theta_j} p_\theta \right] &= \int_{\mathbb{X}} \frac{1}{p_\theta} (\partial^2_{\theta_i,\theta_j} p_\theta) p_\theta \, \mathrm{d}\mu \\ &= \partial^2_{\theta_i,\theta_j} \int_{\mathbb{X}} p_\theta \, \mathrm{d}\mu = \partial^2_{\theta_i,\theta_j}(1) = 0. \end{aligned}$$

□

Proposition 1.6.9: *Let the assumptions (1),(2),(3) hold. For all $\theta \in \Theta$ we have*

$$\mathcal{I}(\theta)_{ij} = -\mathbb{E}_\theta \left[\partial^2_{\theta_i,\theta_j} \log(p_\theta) \right] \quad \forall i,j \in \{1,\dots,k\}.$$

Proof. We have

$$\begin{aligned} \partial^2_{\theta_i,\theta_j} \log(p_\theta) &= \partial_{\theta_i} \partial_{\theta_j} \log(p_\theta) \\ &= \partial_{\theta_i} \left(\frac{1}{p_\theta} \partial_{\theta_j} p_\theta \right) \\ &= \frac{1}{p_\theta} \partial^2_{\theta_i,\theta_j} p_\theta - \frac{1}{p_\theta^2} (\partial_{\theta_i} p_\theta)(\partial_{\theta_j} p_\theta) \end{aligned}$$

that, taking the expectations and using the previous Lemma, gives

$$\mathbb{E}_\theta\left[\partial^2_{\theta_i,\theta_j}\log(p_\theta)\right] = \mathbb{E}_\theta\left[\frac{1}{p_\theta}\partial^2_{\theta_i,\theta_j}p_\theta\right] - \mathcal{I}(\theta)_{ij} = -\mathcal{I}(\theta)_{ij}.$$

□

Example 1.6.10 (Gaussians): Consider the statistical model induced by a sample $X = (X_1, \dots, X_n)$ with $X_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ and $\theta = \mu$. It holds

$$\log(p_\mu(x)) = -\sum_{j=1}^{n} -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x-m)^2}{2\sigma^2}$$

hence

$$\partial_\mu \log(p_\mu(x)) = \sum_{j=1}^{n} \frac{x-\mu}{2\sigma^2}$$

that holds

$$-\partial^2_\mu \log(p_\mu(x)) = \frac{n}{\sigma^2} = \mathcal{I}(\mu).$$

Theorem 1.6.11: *Consider two statistical models $(\mathbb{S}_1, \mathcal{S}_1, (\mathbb{P}^{(1)}_\theta)_{\theta\in\Theta})$, $(\mathbb{S}_2, \mathcal{S}_2, (\mathbb{P}^{(2)}_\theta)_{\theta\in\Theta})$ both satisfying assumptions (1),(2),(3) with dominating measures μ_1 and μ_2 respectively. Then the product statistical model $(\mathbb{S}_1\times\mathbb{S}_2, \mathcal{S}_1\otimes\mathcal{S}_2, (\mathbb{P}^{(1)}_\theta\otimes\mathbb{P}^{(2)}_\theta)_{\theta\in\Theta})$ satisfies assumptions (1),(2),(3) with dominating measure $\mu_1\otimes\mu_2$ and*

$$\mathcal{I}(\theta) = \mathcal{I}_1(\theta) + \mathcal{I}_2(\theta)$$

where $\mathcal{I}_1$ and $\mathcal{I}_2$ are the Fisher information matrices of the two considered statistica models and $\mathcal{I}$ is the Fisher information matrix of the product statistical model.

Proof. If $p^{(1)}_\theta$ and $p^{(2)}_\theta$ are the densities of the two considered statistical models and p_θ is the density of the product statistical model, we have

$$\begin{aligned}
\mathcal{I}(\theta) &= \mathrm{Cov}_\theta(\nabla_\theta \log(p_\theta)) \\
&= \mathrm{Cov}_\theta(\nabla_\theta(\log(p^{(1)}_\theta) + \log(p^{(2)}_\theta))) \\
\text{(previous Lemma)} \quad &= \mathrm{Cov}_\theta(\nabla_\theta \log(p^{(1)}_\theta)) + \mathrm{Cov}_\theta(\nabla_\theta \log(p^{(2)}_\theta)) \\
&= \mathcal{I}_1(\theta) + \mathcal{I}_2(\theta).
\end{aligned}$$

□

Corollary 1.6.12: *Consider the statistical model $(\mathbb{S}^n, \mathcal{S}^{\otimes n}, (P^{\otimes n}_\theta)_{\theta\in\Theta})$ induced by a sample $X = (X_1, \dots, X_n)$ with the X_j's with values in $\mathbb{S}$. We indicate its Fisher information matrix $\mathcal{I}_n(\theta)$. Then if $\mathcal{I}_1(\theta)$ is the Fisher information matrix of the statistical model $(\mathbb{S}, \mathcal{S}, (P_\theta)_{\theta\in\Theta})$, we have*

$$\mathcal{I}_n(\theta) = n\mathcal{I}_1(\theta) \quad \forall\theta\in\Theta.$$

Lemma 1.6.13: *Let $b \in \mathbb{R}^k$ and $A \in \mathbb{R}^{k\times k}$ symmetric and positive definite, then*

$$\sup_{a\in\mathbb{R}^k} \frac{\langle a, b\rangle^2}{a^T A a} = b^T A^{-1} b.$$

Proof. There exists a symmetric invertible matrix $B \in \mathbb{R}^{k\times k}$ s.t. $A = B^2$, then

$$\frac{\langle a,b\rangle^2}{a^TAa} = \frac{\langle B^{-1}y,b\rangle^2}{\langle AB^{-1}y,B^{-1}y\rangle} = \frac{\langle y,B^{-1}b\rangle^2}{\|y\|^2}$$

and

$$\begin{aligned}\sup_{a\in\mathbb{R}^k}\frac{\langle a,b\rangle^2}{a^TAa} &= \sup_{y\in\mathbb{R}^k}\frac{\langle y,B^{-1}b\rangle^2}{\|y\|^2}\\ &= \|B^{-1}b\|^2 = \langle B^{-1}b,B^{-1}b\rangle\\ &= \langle (B^{-1})^2,b\rangle = \langle A^{-1}b,b\rangle.\end{aligned}$$

□

Theorem 1.6.14 (Cramer-Rao lower bound (CRLB)): *Let the assumptions (1),(2),(3) hold. Let $g:\Theta\subset\mathbb{R}^k\to\Gamma\subset\mathbb{R}$ be a measurable function and T be an unbiased estimator of $g(\theta)$ s.t. $T\in L^2(\mathbb{X},\mathbb{P}_\theta)$ for every $\theta\in\Theta$. Then assuming the Fisher information matrix $\mathcal{I}(\theta)$ to be invertible for every $\theta\in\Theta$, we have*

$$\mathrm{Var}_\theta(T) \geq \nabla_\theta g(\theta)^T\mathcal{I}(\theta)^{-1}\nabla_\theta g(\theta) \quad \forall\theta\in\Theta.$$

Proof. The estimator T is unbiased, hence

$$\begin{aligned}\nabla_\theta g(\theta) &= \nabla_\theta\mathbb{E}_\theta[T] = \nabla_\theta\int_{\mathbb{X}} Tp_\theta\,\mathrm{d}\mu\\ &= \int_{\mathbb{X}} T(\nabla_\theta p_\theta)\frac{p_\theta}{p_\theta}\,\mathrm{d}\mu\\ &= \int_{\mathbb{X}} T(\nabla_\theta\log(p_\theta))p_\theta\,\mathrm{d}\mu \qquad (1.1)\\ &= \mathbb{E}_\theta[T(\nabla_\theta\log(p_\theta))] - \underbrace{\mathbb{E}_\theta[\nabla_\theta\log(p_\theta)]\,\mathbb{E}_\theta[T]}_{=0 \text{ (Lemma 1.6.3)}}\\ &= \mathbb{E}_\theta[\nabla_\theta\log(p_\theta) - (T-\mathbb{E}_\theta[T])].\end{aligned}$$

If $\theta\in\mathbb{R}$ (that is $k=1$), using (1.1) we get

$$\partial_\theta g(\theta) = \mathbb{E}_\theta[\partial_\theta\log(p_\theta) - (T-\mathbb{E}_\theta[T])]$$

and using the Cauchy-Schwarz inequality we get (using Lemma 1.6.3)

$$\partial_\theta g(\theta)^2 \leq \mathbb{E}_\theta\left[(\partial_\theta\log(p_\theta))^2\right]\mathrm{Var}_\theta(T).$$

Now we do the case $k>1$. For every $a\in\mathbb{R}^k$, using (1.1), we can write

$$\langle a,\nabla_\theta g(\theta)\rangle = \mathbb{E}_\theta[\langle a,\nabla_\theta\log(p_\theta)\rangle(T-\mathbb{E}_\theta[T])]$$

and using the Cauchy-Schwarz inequality we get

$$\begin{aligned}\langle a,\nabla_\theta g(\theta)\rangle^2 &\leq \mathbb{E}_\theta\left[\langle a,\nabla_\theta\log(p_\theta)\rangle^2\right]\mathrm{Var}_\theta(T)\\ &= a^T\mathbb{E}_\theta\left[\nabla_\theta\log(p_\theta)\nabla_\theta\log(p_\theta)^T\right]a\mathrm{Var}_\theta(T)\end{aligned}$$

so it follows

$$\mathrm{Var}_\theta(T) \geq \frac{\langle a, \nabla_\theta g(\theta)\rangle^2}{a^T \mathcal{I}(\theta) a} \quad \forall a \in \mathbb{R}^k$$

and if we take the $\sup_{a\in\mathbb{R}^k}$, using the previous Lemma we get

$$\mathrm{Var}_\theta(T) \geq \nabla_\theta g(\theta)^T \mathcal{I}(\theta)^{-1} \nabla_\theta g(\theta).$$

□

Corollary 1.6.15: *Let the assumptions (1),(2),(3) hold. Let $g : \Theta \subset \mathbb{R}^k \to \Gamma \subset \mathbb{R}^m$ be a measurable function and T be an unbiased estimator of $g(\theta)$ s.t. $T \in L^2(\mathbb{X}, \mathbb{P}_\theta; \mathbb{R}^m)$ for every $\theta \in \Theta$. Then assuming the Fisher information matrix $\mathcal{I}(\theta)$ to be invertible for every $\theta \in \Theta$, we have*

$$\mathrm{Cov}_\theta(T) \succeq D_\theta g(\theta)^T \mathcal{I}(\theta)^{-1} D_\theta g(\theta) \quad \forall \theta \in \Theta$$

Proof. Take $T = (T_1, \ldots, T_m)$ and $g = (g_1, \ldots, g_m)$. Since T is unbiased we have $\mathrm{bias}_\theta(T_j) = 0$ for every $\theta \in \Theta$, hence

$$\mathrm{bias}_\theta\left(\sum_{j=1}^m a_j T_j\right) = 0 \quad \forall a_1, \ldots, a_m \in \mathbb{R} \quad \forall \theta \in \Theta$$

where the bias is calculated w.r.t. $\sum_{j=1}^m a_j g_j(\theta)$. It holds that for every $a = (a_1, \ldots, a_m) \in \mathbb{R}^m$

$$\begin{aligned} a^T \mathrm{Cov}_\theta(T) a &= \mathrm{Var}_\theta(a^T T) \\ \text{(By CRLB)} \quad &\geq \nabla_\theta(a^T g(\theta))^T \mathcal{I}(\theta)^{-1} \nabla_\theta(a^T g(\theta)) \\ &= a^T D_\theta g(\theta) \mathcal{I}(\theta)^{-1} D_\theta g(\theta)^T a \end{aligned}$$

where the first equality is due to the following calculations

$$\begin{aligned} \mathrm{Var}_\theta(a^T T) &= \mathbb{E}_\theta\left[\left(\sum_{j=1}^m a_j T_j\right)\left(\sum_{i=1}^m a_i T_i\right)\right] \\ &= \sum_{i,j=1}^m a_j \mathbb{E}_\theta\left[T_j T_i\right] a_i = a^T \mathrm{Cov}_\theta(T) a. \end{aligned}$$

□

Definition 1.6.16: An unbiased estimator $T : \mathbb{X} \to \Gamma \subset \mathbb{R}^m$ of $g(\theta)$ s.t. $T \in L^2(\mathbb{X}, \mathbb{P}_\theta; \mathbb{R}^m)$ for every $\theta \in \Theta$ is called *efficient* if

$$\mathrm{Cov}_\theta(T) = D_\theta g(\theta)^T \mathcal{I}(\theta)^{-1} D_\theta g(\theta) \quad \forall \theta \in \Theta.$$

Remark 1.6.17: In the one-dimensional case $k = m = 1$, if $g(\theta) = \theta$ the CRLB gives

$$\mathrm{Var}_\theta(T) \geq \frac{1}{\mathcal{I}(\theta)} \quad \forall \theta \in \Theta.$$

Example 1.6.18: Consider the statistical model induced by a sample $X = (X_1, \ldots, X_n)$ with $X_j \overset{\text{i.i.d.}}{\sim} \mathrm{Poisson}(\theta)$, $\theta \in (0, \infty)$. We have

$$\log(p_\theta(x)) = -\theta + x \log(\theta) - \log(x!),$$

hence $s_\theta(x) = \nabla_\theta \log(p_\theta(x)) = -1 + \frac{x}{\theta}$ and

$$\mathcal{I}_1(\theta) = \mathrm{Var}_\theta(s_\theta) = \mathrm{Var}_\theta(\frac{X_1}{\theta}) = \frac{1}{\theta^2}\mathrm{Var}_\theta(X_1) = \frac{1}{\theta}.$$

Take $g(\theta) = \theta$ and the estimator $\hat{\theta}(x) = \sum_{j=1}^n x_j$, then

$$\mathrm{Var}_\theta(\hat{\theta}) = \frac{1}{n}\theta = \frac{1}{n - \frac{1}{\theta}} = \frac{(\partial_\theta g(\theta)^2)}{\mathcal{I}_n(\theta)}$$

so $\hat{\theta}$ is efficient.

Taking $g(\theta) = e^{-\theta}$ and the estimator $T(x) = \left(\frac{n-1}{n}\right)^{\sum_{j=1}^n x_j}$ we have that this is UMVU, but

$$\begin{aligned}
\mathbb{E}_\theta\left[T^2\right] &= \sum_{j\in\mathbb{N}} \left(1 + \frac{1}{n}\right)^{2j} \frac{(n\theta)^j}{j!} e^{-n\theta} \\
&= e^{-n\theta} \sum_{j\in\mathbb{N}} \frac{1}{j!} \left(\frac{(n-1)^2}{n}\theta\right)^j \\
&= e^{-n\theta} e^{-\frac{(n-1)^2}{n}\theta} = \exp\left(\frac{(1-2n)\theta}{n}\right)
\end{aligned}$$

hence

$$\begin{aligned}
\mathrm{Var}_\theta(T) &= \mathbb{E}_\theta\left[T^2\right] - \mathbb{E}_\theta\left[T\right]^2 \\
&= \exp\left(\frac{(1-2n)\theta}{n}\right) - \exp(-2\theta) \\
&= e^{-2\theta}(e^{\frac{\theta}{n}} - 1) \\
&> e^{-2\theta}\frac{\theta}{n} \\
&= \frac{(\partial_\theta g(\theta))^2}{n\mathcal{I}_1(\theta)} = \frac{(\partial_\theta g(\theta))^2}{\mathcal{I}_n(\theta)}
\end{aligned}$$

since $\partial_\theta g(\theta) = -e^{-\theta}$, so T is not efficient.

Remark 1.6.19: It is true that efficient $\Rightarrow$ UMVU, but from the above example follows that UMVU $\not\Rightarrow$ efficient.

1.7 Kullback-Leibler divergence

In this section we give a method to quantify how easy it is to distinguish between two choices of parameters $\theta_1, \theta_2 \in \Theta$ in some model.

Fix a statistical model $(\mathbb{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ dominated by a measure μ with densities $(p_\theta)_{\theta\in\Theta}$.

Definition 1.7.1: The *discriminatory power* between $\theta_1, \theta_2 \in \Theta$ at $x \in \mathbb{X}$ is

$$h(\theta_1 \,|\, \theta_2)(x) = \log\left(\frac{p_{\theta_1}(x)}{p_{\theta_2}(x)}\right) \in \mathbb{R} \cup \{\pm\infty\}$$

defined on $\{p_{\theta_1} > 0\} \cup \{p_{\theta_2} > 0\}$. We use the conventions $\frac{a}{0} = +\infty$ for $a > 0$ and $\log(0) = -\infty$.

Remark 1.7.2: THe set $A_{\theta_1,\theta_2} = \{p_{\theta_1} > 0\} \cup \{p_{\theta_2} > 0\}$ is s.t. $\mathbb{P}_{\theta_1}(A_{\theta_1,\theta_2}) = \mathbb{P}_{\theta_2}(A_{\theta_1,\theta_2}) = 1$, indeed

$$\mathbb{P}_{\theta_1}(A_{\theta_1,\theta_2}) \geq \mathbb{P}_{\theta_1}(\{p_{\theta_1} > 0\}) = \mathbb{P}_{\theta_1}(\mathbb{X}) = 1$$

and the same for θ_2.

Definition 1.7.3: Take $\theta_1, \theta_2 \in \Theta$. The *Kullback-Leibler divergence (KL-divergence)* between θ_1 and θ_2 is

$$\mathrm{D}_{\mathrm{KL}}(\theta_1 \,|\, \theta_2) = \mathbb{E}_\theta\left[h(\theta_1 \,|\, \theta_2)\right] = \int_{\mathbb{X}} \log\left(\frac{p_{\theta_1}}{p_{\theta_2}}\right) p_{\theta_1} \,\mathrm{d}\mu.$$

Remark 1.7.4: For $\theta_1, \theta_2 \in \Theta$ hold $\mathrm{D}_{\mathrm{KL}}(\theta_1 \,|\, \theta_2) \neq \mathrm{D}_{\mathrm{KL}}(\theta_2 \,|\, \theta_1)$, so the KL-divergence is not a metric.

Recall the following classical result.

Theorem 1.7.5 (Jensen's inequality): *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $Y \in L^1(\Omega, \mathbb{P})$ and $\phi : \mathbb{R} \to \mathbb{R}$ be a convex function. Then $\mathbb{E}[\phi(Y)_-] < +\infty$ and the expectation $\mathbb{E}[\phi(Y)]$ is well defined and in $\mathbb{R} \cup \{+\infty\}$. Moreover the following inequality holds*

$$\phi(\mathbb{E}[Y]) \leq \mathbb{E}[\phi(Y)]$$

with equality if and only if Y is $\mathbb{P}$-a.s. constant.

Proposition 1.7.6: *For every $\theta_1, \theta_2 \in \Theta$ we have*

(1) the KL-divergence $\mathrm{D}_{\mathrm{KL}}(\theta_1 \,|\, \theta_2)$ is well defined in $\mathbb{R} \cup \{+\infty\}$;

(2) $\mathrm{D}_{\mathrm{KL}}(\theta_1 \,|\, \theta_2) \geq 0$;

(3) $\mathrm{D}_{\mathrm{KL}}(\theta_1 \,|\, \theta_2) = 0$ if and only if $\mathbb{P}_{\theta_1} = \mathbb{P}_{\theta_2}$.

Proof. (1) Take $Y = \frac{p_{\theta_2}}{p_{\theta_1}}$, then

$$\mathbb{E}_{\theta_1}[|Y|] = \mathbb{E}_{\theta_1}[Y] = \int_{\mathbb{X}} \frac{p_{\theta_2}}{p_{\theta_1}} p_{\theta_1} \,\mathrm{d}\mu = \mathbb{P}_{\theta_2}(\mathbb{X}) = 1$$

hence by the previous Theorem we obtain (1).

(2) Observe that

$$\begin{aligned} \mathrm{D}_{\mathrm{KL}}(\theta_1 \,|\, \theta_2) &= \mathbb{E}_{\theta_1}\left[-\log\left(\frac{p_{\theta_1}}{p_{\theta_2}}\right)\right] \\ \text{(Jensen's ineq.)} &\geq -\log\left(\int_{\mathbb{X}} \frac{p_{\theta_1}}{p_{\theta_2}} p_{\theta_2} \,\mathrm{d}\mu\right) = 0. \end{aligned}$$

(3) Obviously if $\mathbb{P}_{\theta_1} = \mathbb{P}_{\theta_2}$, then $p_{\theta_1} = p_{\theta_2}$ μ-a.e. on $\mathbb{X}$ and $\mathrm{D}_{\mathrm{KL}}(\theta_1 \,|\, \theta_2) = 0$. Viceversa if $\mathrm{D}_{\mathrm{KL}}(\theta_1 \,|\, \theta_2) = 0$, then the Jensen's inequality in point (2) is an equality and as a consequence $Y = \frac{p_{\theta_2}}{p_{\theta_1}} = c$ $\mathbb{P}_{\theta_1}$-a.s. for some constant $c \geq 0$. But $c = \mathbb{E}_{\theta_1}[Y] = 1$ necessarily, hence for every $A \in \mathcal{F}$ holds

$$\mathbb{P}_{\theta_2}(A) = \int_A p_{\theta_2} \,\mathrm{d}\mu = \int_A \frac{p_{\theta_2}}{p_{\theta_1}} p_{\theta_1} d\mu = \mathbb{P}_{\theta_1}(A)$$

since $\frac{p_{\theta_2}}{p_{\theta_1}} = 1$ μ-a.e.

□

Theorem 1.7.7: *Let the assumptions (1),(2),(3) hold for the considered statistical model and assume that* $\partial^2_{\theta'_j,\theta'_i} \int_{\mathbb{X}} \log\left(\frac{p_\theta}{p_{\theta'}}\right) p_\theta \, \mathrm{d}\mu = \int_{\mathbb{X}} \partial^2_{\theta'_j,\theta'_i} \log\left(\frac{p_\theta}{p_{\theta'}}\right) p_\theta \, \mathrm{d}\mu$ *for every* $\theta, \theta' \in \Theta$. *Then*

$$\partial^2_{\theta'_j,\theta'_i} \mathrm{D}_{\mathrm{KL}}\left(\theta \,|\, \theta'\right)\Big|_{\theta'=\theta} = \mathcal{I}(\theta)_{ij} \quad \forall \theta \in \Theta.$$

Proof. Deriving w.r.t θ'_j we get

$$\partial_{\theta'_j} \mathrm{D}_{\mathrm{KL}}\left(\theta \,|\, \theta'\right) = -\int_{\mathbb{X}} (\partial_{\theta'_j} \log(p_{\theta'})) p_\theta \, \mathrm{d}\mu = -\int_{\mathbb{X}} \frac{1}{p_{\theta'}} (\partial_{\theta'_j} p_{\theta'}) p_\theta \, \mathrm{d}\mu$$

then deriving w,r,t, θ'_i we obtain

$$\partial^2_{\theta'_j,\theta'_i} \mathrm{D}_{\mathrm{KL}}\left(\theta \,|\, \theta'\right) = \int_{\mathbb{X}} \left(\frac{(\partial_{\theta'_j} p_{\theta'})(\partial_{\theta'_i} p_{\theta'})}{p^2_{\theta'}} - \frac{\partial^2_{\theta'_j,\theta'_i} p_{\theta'}}{p_{\theta'}} \right) p_\theta \, \mathrm{d}\mu.$$

Calculating for $\theta' = \theta$ we get

$$\partial^2_{\theta'_j,\theta'_i} \mathrm{D}_{\mathrm{KL}}\left(\theta \,|\, \theta'\right)\Big|_{\theta'=\theta} = \int_{\mathbb{X}} \left(\frac{(\partial_{\theta'_j} p_{\theta'})(\partial_{\theta'_i} p_{\theta'})}{p^2_{\theta'}} \right) p_\theta \, \mathrm{d}\mu \Big|_{\theta'=\theta}$$

since

$$\begin{aligned}
\int_{\mathbb{X}} \frac{\partial^2_{\theta'_j,\theta'_i} p_{\theta'}}{p_{\theta'}} p_\theta \, \mathrm{d}\mu \Big|_{\theta'=\theta} &= \int_{\mathbb{X}} \frac{\partial^2_{\theta'_j,\theta'_i} p_\theta}{p_\theta} p_\theta \, \mathrm{d}\mu \\
&= \int_{\mathbb{X}} \partial^2_{\theta_j,\theta_i} p_\theta \, \mathrm{d}\mu \\
&= \partial^2_{\theta_j,\theta_i} \int_{\mathbb{X}} p_\theta \, \mathrm{d}\mu = \partial^2_{\theta_j,\theta_i}(1) = 0.
\end{aligned}$$

Finally

$$\begin{aligned}
\partial^2_{\theta'_j,\theta'_i} \mathrm{D}_{\mathrm{KL}}\left(\theta \,|\, \theta'\right)\Big|_{\theta'=\theta} &= \int_{\mathbb{X}} \left(\frac{(\partial_{\theta'_j} p_{\theta'})(\partial_{\theta'_i} p_{\theta'})}{p^2_{\theta'}} \right) p_\theta \, \mathrm{d}\mu \Big|_{\theta'=\theta} \\
&= \int_{\mathbb{X}} \frac{(\partial_{\theta'_j} p_{\theta'})(\partial_{\theta'_i} p_{\theta'})}{p^2_{\theta'}} \, \mathrm{d}\mathbb{P}_\theta \Big|_{\theta'=\theta} \\
&= \mathbb{E}_\theta \left[\frac{(\partial_{\theta'_j} p_{\theta'})(\partial_{\theta'_i} p_{\theta'})}{p^2_{\theta'}} \right] \Big|_{\theta'=\theta} \\
&= \mathbb{E}_\theta \left[\frac{(\partial_{\theta_j} p_\theta)(\partial_{\theta_i} p_{\theta'})}{p^2_\theta} \right] = \mathcal{I}(\theta)_{ij}.
\end{aligned}$$

□

1.8 Exponential families

Fix a statistical model $(\mathbb{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ dominated by a measure μ with densities $(p_\theta)_{\theta\in\Theta}$.

Definition 1.8.1: The family of probability measures $(\mathbb{P}_\theta)_{\theta\in\Theta}$ is a *k-dimensional exponential family*, with $k \in \mathbb{N}_+$, if

$$p_\theta(x) = \exp\left(\sum_{j=1}^{k} c_j(\theta)T_j(x) - d(\theta)\right) h(x)$$

where $T = (T_1, \dots, T_k)$ is a statistic with values in $\mathbb{R}^k$ and $h : \mathbb{X} \to [0, +\infty)$ is a measurable function.

Remark 1.8.2: In the context of the previous Definition:

(1) $T_1, \dots, T_k$ and $c_1, \dots, c_k$ are not unique, for example we can substitute T_j with aT_j and c_j with $\frac{1}{a}c_j$ where $a \neq 0$;

(2) the statistic T is sufficient (obvious from Theorem 1.3.4);

(3) necessarily $h : \mathbb{X} \to [0, \infty)$.

Remark 1.8.3: Consider the statistical model $(\mathbb{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ induced by a sample $X = (X_1, \dots, X_n)$ with $X_j \overset{\text{i.i.d.}}{\sim} P_\theta$ where $(P_\theta)_{\theta\in\Theta}$ an exponential family w.r.t. the dominating measure ν, then $(\mathbb{P}_\theta)_{\theta\in\Theta}$ is an exponential family w.r.t. $\mu = \nu^{\otimes n}$.

Indeed, if q_θ is the density of P_θ w.r.t. ν, then $p_\theta(x) = \prod_{j=1}^n q_\theta(x_j)$ is the density of $\mathbb{P}_\theta$ w.r.t. μ and we have

$$\begin{aligned} p_\theta(x) &= \prod_{j=1}^{n} q_j(x) \\ &= \exp\left(\sum_{j=1}^{n}\left(c_j(\theta) n \frac{1}{n}\sum_{i=1}^{n} T_i(x)\right) - nd(\theta)\right)\prod_{l=1}^{n} h(x_l) \\ &= \exp\left(\sum_{j=1}^{n}\left(c_j(\theta) \sum_{i=1}^{n} T_i(x)\right) - nd(\theta)\right)\prod_{l=1}^{n} h(x_l). \end{aligned}$$

Example 1.8.4: It is not difficult to see that the distributions Poisson(θ), $\mathcal{N}(\mu, \sigma^2)$ with $\theta = (\mu, \sigma^2)$ and Bin(n, θ) are all exponential families (with θ that varies in the respective Θ).

Lemma 1.8.5: *Suppose that $(\mathbb{P}_\theta)_{\theta\in\Theta}$ is an exponential family with densities*

$$p_\theta(x) = \exp\left(\sum_{j=1}^{k} c_j(\theta)T_j(x) - d(\theta)\right) h(x)$$

with $h : X \to [0, +\infty)$ measurable s.t. $h > 0$ μ-a.e. on $\mathbb{X}$. If the set

$$\mathcal{C} = \{(c_1(\theta), \dots, c_k(\theta)) \mid \theta \in \Theta\}$$

contains an open set of $\mathbb{R}^k$, then the statistic $T = (T_1, \dots, T_k)$ is complete.

Proof. Consider $f : E \to \mathbb{R}$ s.t. $\int_\theta (f \circ T) p_\theta \, \mathrm{d}\mu = 0$ for every $\theta \in \Theta$ Note that $\mathcal{C}$ contains an open set $\mathcal{C}'$ of $\mathbb{R}^k$. For every $\theta \in \Theta$ s.t. $c(\theta) \in \mathcal{C}'$ we write

$$\begin{aligned} 0 &= \int_{\mathbb{X}} f(T(x)) \exp\left(\sum_{j=1}^k c_j(\theta) T_j(x) - d(\theta) \right) h(x) \, \mathrm{d}\mu(x) \\ &= \int_{\mathbb{R}^k} f(t) \exp\left(\sum_{j=1}^k c_j(\theta) t_j - d(\theta) \right) \mathrm{d}\nu_T(t) \end{aligned}$$

where $\nu_T = T_*(h \, \mathrm{d}\mu)$ (observe that since h is non-negative $h \, \mathrm{d}\mu$ is a non-negative measure). Hence

$$\int_{\mathbb{R}^k} f_+(t) \exp(\langle c, t \rangle) \, \mathrm{d}\nu_T(t) = \int_{\mathbb{R}^k} f_-(t) \exp(\langle c, t \rangle) \, \mathrm{d}\nu_T(t) \quad \forall c \in \mathcal{C}'$$

that implies $f_+ = f_-$ on $\mathrm{supp}(\nu_T)$ (since the Laplace transform is injective: we obtain the equality of the two non-negative measures $f_+ \, \mathrm{d}\nu_T, f_- \, \mathrm{d}\nu_T$ on the open set $\mathcal{C}'$), so $f = 0$ on $\mathrm{supp}(\nu_T)$. □

Theorem 1.8.6: *Suppose that $(\mathbb{P}_\theta)_{\theta \in \Theta}$ is an exponential family with $k = 1$ and $\Theta \subset \mathbb{R}$, so that*

$$p_\theta(x) = \exp(\langle c(\theta), T(x) \rangle - d(\theta)) h(x) \quad \forall x \in \mathbb{X} \quad \forall \theta \in \Theta.$$

Then the sufficient statistic T is an efficient estimator for $g(\theta) = \mathbb{E}_\theta[T]$.

Proof. Observe that the score function is

$$\begin{aligned} s_\theta(x) &= \partial_\theta \log(p_\theta) = -\partial_\theta d(\theta) + T(x) \partial_\theta c(\theta) \\ &= -\partial_{\theta d(\theta) + \mathbb{E}_\theta[T] \partial_\theta c(\theta) + (T - \mathbb{E}_\theta[T]) \partial_\theta c(\theta)} \cdot \end{aligned}$$

As seen in the proof of CRLB (Theorem 1.6.14)

$$\partial_\theta g(\theta) = \mathbb{E}_\theta \left[s_\theta (T - \mathbb{E}_\theta[T]) \right]$$

but s_θ is affine in $T - \mathbb{E}_\theta[T]$, so by Cauchy-Schwarz inequality in the case of equality, follows

$$\partial_\theta g(\theta)^2 = \mathbb{E}_\theta \left[s_\theta^2 \right] \mathrm{Var}_\theta(T)$$

that implies $\mathrm{Var}_\theta(T) = \frac{\partial_\theta g(\theta)^2}{\mathcal{I}(\theta)}$. □

2

Asymptotics

2.1 Generalities and M-estimators

In this chapter we study the following situation: consider a statistical model $(\mathbb{S}, \mathcal{S}, (P_\theta)_{\theta\in\Theta})$ and an infinite sample, that is a sequence of r.v.'s $X = (X_n)_{n\in\mathbb{N}_+}$, $X_j : \Omega \to \mathbb{S}$ for every $j \in \mathbb{N}_+$. We want to study what happens in this limit, that is "for $n \to +\infty$". To do so we take the statistical model induced by the sample X, that is the statistical model $(\mathbb{X}.\mathcal{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ with $\mathbb{X} = \mathbb{S}^{\mathbb{N}_+}$,, $\mathcal{F} = \mathcal{S}^{\otimes\mathbb{N}_+}$ and $\mathbb{P}_\theta = P_\theta^{\otimes\mathbb{N}_+}$ for every $\theta \in \Theta$. Where the probabilities $\mathbb{P}_\theta$ are defined using the Ionescu-Tulcea theorem.

In the following sections given a r.v. $f : \mathbb{S} \to \mathbb{R}$ and any $\theta \in \Theta$, we will use the followng notation:

$$\mathbb{E}_{P_\theta}[f] = \int_{\mathbb{S}} f \, \mathrm{d}P_\theta.$$

Lastly we fix a measurable function $g : \Theta \to \Gamma \subset \mathbb{R}^p$.

Definition 2.1.1: A *sequence of estimators* for $g(\theta)$ is a sequence of estimators $(T_n)_{n\in\mathbb{N}}$ for $g(\theta)$ s.t. $T_n : \mathbb{X} \to \Gamma$ for every $n \in \mathbb{N}_+$.

Definition 2.1.2: An estimator $T_n : \mathbb{X} \to \Gamma$ of $g(\theta)$ is called *M-estimator* if there exists a function $\rho : \mathbb{S} \times \mathbb{R}^p \to \mathbb{R}$, called *cost* or *loss*, s.t.

$$T_n(x) = \arg\min_{c\in\Gamma} \frac{1}{n} \sum_{j=1}^{n} \rho(x_j, c) \quad \forall x \in \mathbb{X}.$$

Example 2.1.3: Suppose that $(\mathbb{S}, \mathcal{S}, (P_\theta)_{\theta\in\Theta})$ is and take dominated $L_\theta(x)$ a likelihood function. If $g(\theta) = \theta$ (so $c = \theta$ and $\Gamma = \Theta$), taking $\rho(x, \theta) = -\log(L_\theta(x))$ we get the the MLE is an M-estimator.

Remark 2.1.4: In the context of the previous Definition, if $\rho(x, \cdot)$ is C^1 for every $x \in \mathbb{S}$,

then the M-estimator can be obtained setting

$$\sum_{j=1}^{n} \nabla_\theta \rho(x_j, \theta) = 0.$$

Definition 2.1.5: An estimator $T_n : \mathbb{X} \to \Gamma$ of $g(\theta)$ is called *Z-estimator* if there exists a measurble function $\psi : \mathbb{S} \times \Gamma \to \mathbb{R}$ s.t.

$$\sum_{j=1}^{n} \psi(x_j, T_n(x)) = 0 \quad \forall x \in \mathbb{X}.$$

Remark 2.1.6: So if the cost function is C^1 in θ, then a M-estimator is also a Z-estimator.

We would like that the "accuracy" of the estimate $T_n(X)$ of $g(\theta)$ to increase an n grows. Mathematically we want to study when $T_n(X) \to g(\theta)$ for $n \to +\infty$ for some type of convergence.

Definition 2.1.7: A sequence of estimators $(T_n)_{n\in\mathbb{N}_+}$ for $g(\theta)$ is called *consistent* if $T_n \overset{\mathbb{P}_\theta}{\to} g(\theta)$ for all $\theta \in \Theta$. Meanwhile it is called *strongly consistent* if $T_n \to g(\theta)$ $\mathbb{P}_\theta$-q.c. for every $\theta \in \Theta$.

Definition 2.1.8: A sequence of estimators $(T_n)_{n\in\mathbb{N}_+}$ for $g(\theta)$ is called *asymptotically normal* with *asymptotic covariance matrix* $\Sigma(\theta)$ if

$$\sqrt{n}(T_n - g(\theta)) \overset{\mathcal{L}}{\to} \mathcal{N}(0, \Sigma(\theta)) \quad \text{w.r.t. } \mathbb{P}_\theta \text{ for all } \theta \in \Theta.$$

Definition 2.1.9: Consider $c \in \Gamma$ and $\rho : \mathbb{S} \times \Gamma \to \mathbb{R}$ a loss function. We call *risk* the function $\mathcal{R} : \Theta \times \Gamma \to \mathbb{R}$ s.t.

$$\mathcal{R}(\theta, c) = \mathcal{R}_\theta(c) = \mathbb{E}_{P_\theta}[\rho_c] \quad \forall c \in \Gamma \quad \forall \theta \in \Theta.$$

where $\rho_c = \rho(\cdot, c)$. Given an $n \in \mathbb{N}_+$, we call *empirical risk (of size n)* the function $\hat{\mathcal{R}}_n : \mathbb{X} \times \Gamma \to \mathbb{R}$ s.t.

$$\hat{\mathcal{R}}_n(x, c) = \frac{1}{n} \sum_{j=1}^{n} \rho_c(x_j) \quad \forall x \in \mathbb{X} \quad \forall c \in \Gamma.$$

Lastly, we call

$$\gamma(\theta) = \arg\min_{c\in\Gamma} \mathcal{R}_\theta(c).$$

Remark 2.1.10: The r.v. $\hat{\gamma}_n : \mathbb{X} \to \Gamma$ s.t. $\hat{\gamma}_n(x) = \arg\min_{c\in\Gamma} \hat{\mathcal{R}}_n(x, c)$ is an unbiased M-estimator of $g(\theta)$, so $\hat{\gamma}_n(X) = \arg\min_{c\in\Gamma} \hat{\mathcal{R}}_n(X, c)$ is an M-estimate of $\gamma(\theta)$.

Recall the following two theorems from probability theory that will be useful in the following.

Theorem 2.1.11 (Cramer-Wold): *Let $(Z_n)_{n\in\mathbb{N}}$ and Z be of $\mathbb{R}^p$-valued r.v.'s. Then*

$$Z_n \overset{\mathcal{L}}{\to} Z \iff \langle a, Z_n \rangle \overset{\mathcal{L}}{\to} \langle a, Z \rangle \quad \forall a \in \mathbb{R}^p.$$

Proof. For a r.v. Y we will indicate its characteristic function as ϕ_Y.

The thesis is a simple consequence of the Levy's continuity theorem. It states that $Z_n \xrightarrow{\mathcal{L}} Z$ if and only if $\phi_{Z_n}(a) \to \phi_Z(a)$ for every $a \in \mathbb{R}^p$, that means

$$\mathbb{E}\left[\exp(i\langle a, Z_n\rangle)\right] \to \mathbb{E}\left[\exp(i\langle a, Z\rangle)\right] \quad \forall a \in \mathbb{R}^p$$

and this happens if and only if $\langle a, Z_n\rangle \to \langle a, Z\rangle$ for every $a \in \mathbb{R}^p$. □

Theorem 2.1.12 (Portmanteau): *Let $(Z_n)_{n\in\mathbb{N}}$ and Z be $\mathbb{R}^p$-valued r.v.'s. Denote Q the cumulative distribution function of Z*

$$Q(z) = \mathbb{P}(Z \leq z) = \mathbb{P}(Z_1 \leq z_1, \ldots, Z_p \leq z_p) \quad \forall z \in \mathbb{R}^p.$$

The following are equivalent:

(1) $Z_n \xrightarrow{\mathcal{L}} Z$, that is $\mathbb{E}\left[f(Z_n)\right] \to \mathbb{E}\left[f(Z)\right]$ for every $f \in C_b(\mathbb{R}^p)$;

(2) $\mathbb{E}\left[f(Z_n)\right] \to \mathbb{E}\left[f(Z)\right]$ for every $f : \mathbb{R}^p \to \mathbb{R}$ bounded and Lipschitz-continuous;

(3) $\mathbb{E}\left[f(Z_n)\right] \to \mathbb{E}\left[f(Z)\right]$ for every $f : \mathbb{R}^p \to \mathbb{R}$ bounded and $(Z_*\mathbb{P})$-a.s. continuous;

(4) $\mathbb{P}(Z_n \leq z) \to G(z)$ for every continuity point $z \in \mathbb{R}^p$ of G.

Proof. Not covered. □

Theorem 2.1.13 (Slutsky): *Let $(Z_n)_{n\in\mathbb{N}}$, $(A_n)_{n\in\mathbb{N}}$ and Z be $\mathbb{R}^p$-valued r.v.'s and $a \in \mathbb{R}^p$. If $Z_n \xrightarrow{\mathcal{L}} Z$ and $A_n \xrightarrow{\mathbb{P}} a$, then*

$$\langle A_n, Z_n\rangle \xrightarrow{\mathcal{L}} \langle a, Z\rangle.$$

Proof. Not covered. □

Now let us fix some notations.

Definition 2.1.14: Let $(Z_n)_{n\in\mathbb{N}}$ be a sequence of $\mathbb{R}^p$-valued r.v.'s, we will say that it is *bounded in probability* or *uniformly tight* and we will write $Z_n = \mathcal{O}_{\mathbb{P}}(1)$, if

$$\lim_{M\to+\infty} \limsup_{n\to+\infty} \mathbb{P}(\|Z_n\| > M) = 0.$$

While if $Z_n \xrightarrow{\mathbb{P}} 0$ we will write $Z_n = o_{\mathbb{P}}(1)$.

Moreover if $(r_n)_{n\in\mathbb{N}}$ is a sequence of $(0,\infty)$-valued r.v.'s we will write $Z_n = \mathcal{O}_{\mathbb{P}}(r_n)$ if $\frac{Z_n}{r_n} = \mathcal{O}_{\mathbb{P}}(1)$ and $Z_n = o_{\mathbb{P}}(1)$ if $\frac{Z_n}{r_n} = o_{\mathbb{P}}(1)$.

Lemma 2.1.15: *Let $(Z_n)_{n\in\mathbb{N}}$ and Z be of $\mathbb{R}^p$-valued r.v.'s. If $Z_n \xrightarrow{\mathcal{L}} Z_n$ then $Z_n = \mathcal{O}_{\mathbb{P}}(1)$.*

Proof. Using the Cramer-Wold Theorem 2.1.11 we can restrict to study the 1-dimensional case ($p = 1$). By Portmanteau Theorem 2.1.12, if G is the cumulative distribution function of Z we have that

$$\mathbb{P}(Z_n > M) \to 1 - G(M)$$

for every $M \in \mathbb{R}$ continuity point of G and analogously

$$\mathbb{P}(Z_n \leq -M) \to G(-M)$$

for every $M \in \mathbb{R}$ s.t. $-M$ is a continuity point of G. Since $1 - G(M), G(-M) \to 0$ for $M \to +\infty$ the thesis follows. Indeed

$$\lim_{M\to+\infty} \limsup_{n\to+\infty} \mathbb{P}(|Z_n| > M) = \lim_{M\to+\infty} (1 - G(M) + G(-M)) = 0.$$

□

Definition 2.1.16: A sequence of estimators $(T_n)_{n\in\mathbb{N}_+}$ for $g(\theta)$ is called *asymptotically linear* with *influence function* $l_\theta : \mathbb{S} \to \mathbb{R}^p$ if $\mathbb{E}_\theta[l_\theta] = 0$,

$$\mathbb{E}_\theta\left[\|l_\theta\|^2\right] < +\infty$$

and

$$T_n - g(\theta) = \frac{1}{n}\sum_{j=1}^{n} l_\theta(X_j) + o_{\mathbb{P}_\theta}\left(\frac{1}{\sqrt{n}}\right)$$

for $n \to +\infty$ for every $\theta \in \Theta$.

2.2 Consistency and asymptotic normality of M-estimators

Consider the context of the previous section. We fix also a cost function $\rho : \mathbb{S} \times \Gamma \to \mathbb{R}$ and consider the associated risks $\mathcal{R}$ and $\hat{\mathcal{R}}_n$ for $n \in \mathbb{N}_+$. We also consider the minimizer $\gamma(\theta) \in \Gamma$ of $\mathcal{R}_\theta$ and the M-estimators $\hat{\gamma}_n(x) = \arg\min_{c\in\Gamma} \hat{\mathcal{R}}_n(x, c)$.

Definition 2.2.1: The minimizer γ is called *well separeted* if for all $\varepsilon > 0$

$$\inf\{\mathcal{R}_\theta(c) \mid \|c - \gamma\| > \varepsilon\} > \mathcal{R}_\theta(\gamma(\theta)).$$

Theorem 2.2.2: *Suppose that*

(1) Γ *is compact;*

(2) $c \mapsto \rho_c(x)$ *is continuous for every* $x \in \mathbb{X}$*;*

(3) $\mathbb{E}_\theta[\sup_{c\in\Gamma} |\rho_c|] < +\infty$.

Then we have

$$\max_{c\in\Gamma} |\hat{\mathcal{R}}_n(\cdot, c) - \mathcal{R}_\theta(c)| \to 0 \quad \mathbb{P}_\theta\text{-a.s. for every } \theta \in \Theta.$$

In particular

$$\mathcal{R}_\theta(\hat{\gamma}_n) \to \mathcal{R}_\theta(\gamma(\theta)) \quad \mathbb{P}_\theta\text{-a.s. for every } \theta \in \Theta$$

and if γ *is well separeted this implies*

$$\hat{\gamma}_n \to \gamma(\theta) \quad \mathbb{P}_\theta\text{-a.s. for every } \theta \in \Theta$$

or, in other words, the sequence of estimators $(\hat{\gamma}_n)_{n\in\mathbb{N}_+}$ *for* $\gamma(\theta)$ *is strongly consistent.*

Proof. (Optional) For every $x \in \mathbb{S}$, $\delta > 0$ and $c \in \Gamma$ define

$$\omega(x, \delta, c) = \sup\{|\rho_{c'}(x) - \rho_c(x)| \mid c' \in \Gamma, \|c' - c\| < \delta\}.$$

Then

$$\omega(x,\delta,c)\to 0 \ \text{ when } \delta\searrow 0 \text{ for every } x\in\mathbb{S}$$

so by dominated convergence

$$\mathbb{E}_{P_\theta}\left[\omega(\cdot,\delta,c)\right]\to 0 \ \text{ when } \delta\searrow 0.$$

Hence for every $\varepsilon>0$ and $c\in\Gamma$ there exists a $\delta_c>0$ s.t.

$$\mathbb{E}_{P_\theta}\left[\omega(\cdot,\delta_c,c)\right]\le\varepsilon.$$

Take $B_c=\{c'\in\Gamma \mid \|c'-c\|<\delta_c\}$, then $\{B_c\}_{c\in\Gamma}$ is an open covering of Γ that is compact, so there exists a finite sub-covering $\{B_{c_1},\dots,B_{c_N}\}$. For any $j\in\{1,\dots,N\}$ and any $x\in B_{c_j}$ holds

$$|\rho_c(x)-\rho_{c_j}(c)|\le\omega(x,\delta_{c_j},c_j)\ \ \forall x\in\mathbb{S},$$

as a consequence for $c\in B_{c_j}$

$$\begin{aligned}
&\max_{c\in\Gamma}|\hat{\mathcal{R}}_n(\cdot,c)-\mathcal{R}_\theta(c)|\le\\
&\underbrace{\max_{1\le j\le N}|\hat{\mathcal{R}}_n(\cdot,c_j)-\mathcal{R}_\theta(c_j)|}_{\to 0\ \mathbb{P}_\theta\text{-a.s.}}+\max_{1\le j\le N}\frac{1}{n}\sum_{j=1}^{n}\omega(\cdot,\delta_{c_j},c_j)+\max_{1\le j\le N}\mathbb{E}_{P_\theta}\left[\omega(\cdot,\delta_{c_j},c_j)\right]\\
&\to 2\max_{1\le j\le N}\mathbb{E}_{P_\theta}\left[\omega(\cdot,\delta_{c_j},c_j)\right]\le 2\varepsilon\ \ \mathbb{P}_\theta\text{-a.s.}\ \ \forall\theta\in\Theta
\end{aligned}$$

that gives the first part of the thesis, where we used the strong law of large numbers to conclude that $\max_{1\le j\le N}|\hat{\mathcal{R}}_n(\cdot,c_j)-\mathcal{R}_\theta(c_j)|\to 0$ and $\max_{1\le j\le N}\frac{1}{n}\sum_{j=1}^{n}\omega(\cdot,\delta_{c_j},c_j)\to\max_{1\le j\le N}\mathbb{E}_{P_\theta}\left[\omega(\cdot,\delta_{c_j},c_j)\right]$ $\mathbb{P}_\theta$-a.s. for every $\theta\in\Theta$

Now the second convergence result of the theorem follows:

$$\begin{aligned}
0&\le\mathcal{R}_\theta(\hat{\gamma}_n)-\mathcal{R}_\theta(\gamma(\theta))\\
&=-[(\hat{\mathcal{R}}_n(\cdot,\hat{\gamma}_n)-\mathcal{R}_\theta(\hat{\gamma}_n))-(\hat{\mathcal{R}}_n(\cdot,\gamma(\theta))-\mathcal{R}_\theta(\gamma(\theta)))]+[\hat{\mathcal{R}}_n(\cdot,\hat{\gamma}_n)-\hat{\mathcal{R}}_n(\cdot,\gamma(\theta))]\\
&\le-[(\hat{\mathcal{R}}_n(\cdot,\hat{\gamma}_n)-\mathcal{R}_\theta(\hat{\gamma}_n))-(\hat{\mathcal{R}}_n(\cdot,\gamma(\theta))-\mathcal{R}_\theta(\gamma(\theta)))]\\
&\le|\hat{\mathcal{R}}_n(\cdot,\hat{\gamma}_n)-\mathcal{R}_\theta(\hat{\gamma}_n)|+|\hat{\mathcal{R}}_n(\cdot,\gamma(\theta))-\mathcal{R}_\theta(\gamma(\theta))|\\
&\le 2\max_{c\in\Gamma}|\hat{\mathcal{R}}_n(\cdot,c)-\mathcal{R}_\theta(c)|\to 0\ \ \mathbb{P}_\theta\text{-a.s.}\ \ \forall\theta\in\Theta.
\end{aligned}$$

□

Theorem 2.2.3: *Take $g(\theta)=\theta$, n particular $\Gamma=\Theta$ and we call $\hat{\gamma}_n=\hat{\theta}_n$ and the M-estimator $\gamma(\theta)=\theta_M(\theta)=\arg\min_{\theta'\in\Theta}\mathcal{R}_\theta(\theta')$. Suppose that*

(1) Θ is compact;

(2) $\theta\mapsto\rho_\theta(x)$ is twice differentiable on $\mathrm{Int}(\Theta)$ for every $x\in\mathbb{X}$;

(3) For every $\theta\in\Theta$ called $H(y,\theta)=\nabla^2_{\theta\theta}\rho(y,\theta)$ and $s(y,\theta)=\nabla_\theta\rho(y,\theta)$ for $y\in\mathbb{S}$ and $\theta\in\Theta$, assume $H(y,\cdot)$ to be bounded for every $y\in\mathbb{S}$, $A_{\theta'}(\theta)=\mathbb{E}_{P_\theta}\left[H(\cdot,\theta')\right]$ to be positive definite for every $\theta,\theta'\in\Theta$, $\mathbb{E}_{P_\theta}\left[s(\cdot,\theta')\right]=0$, $\mathrm{Var}_{P_\theta}(s(\cdot,\theta'))<+\infty$ for every $\theta,\theta'\in\Theta$.

Let us indicate $B_{\theta'}(\theta) = \mathbb{E}_{P_\theta}\left[s(\cdot,\theta')^T S(\cdot,\theta')\right]$. *Then*

$$\sqrt{n}(\hat{\theta}_n - \theta_M(\theta)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_M(\theta)) \quad \forall \theta \in \Theta,$$

where $\Sigma_M(\theta) = A^{-1}_{\theta_M}(\theta) B_{\theta_M}(\theta) A^{-1}_{\theta_M}(\theta)$ *forevery* $\theta \in \Theta$.

Proof. Not covered. □

2.3 The δ-method

The delta method is a method of deriving the asymptotic distribution of a sequence of random variables.

Theorem 2.3.1 (δ-method): *Let* $(T_n)_{n\in\mathbb{N}}$ *and* Z *be of* $\mathbb{R}^p$*-valued r.v.'s,* $c \in \mathbb{R}^p$, $(r_n)_{n\in\mathbb{N}} \subset (0,\infty)$ *s.t.* $r_n \searrow 0$ *and* $h : \mathbb{R}^p \to \mathbb{R}$ *continuous and differentiable at* c *with gradient* $\nabla h(c)$. *Suppose that*

$$\frac{T_n - c}{r_n} \xrightarrow{\mathcal{L}} Z$$

then

$$\frac{h(T_n) - h(c)}{r_n} \xrightarrow{\mathcal{L}} \langle \nabla h(c), Z \rangle.$$

Proof. By Slutsky's Theorem 2.1.13 we have that

$$\frac{\langle \nabla h(c), T_n - c \rangle}{r_n} \xrightarrow{\mathcal{L}} \langle \nabla h(c), Z \rangle.$$

Since $\frac{T_n - c}{r_n}$ converges in law we have that $\frac{\|T_n - c\|}{r_n} = \mathcal{O}_{\mathbb{P}}(1)$, hence $\|T_n - c\| = \mathcal{O}_{\mathbb{P}}(r_n)$, but by Taylor's theorem

$$\begin{aligned} h(T_n) - h(c) &= \langle \nabla(h)(c), T_n - c \rangle + o(\|T_n - c\|) \\ &= \langle \nabla(h)(c), T_n - c \rangle + o_{\mathbb{P}}(r_n) \end{aligned}$$

hence

$$\frac{h(T_n) - h(c)}{r_n} = \frac{\langle \nabla(h)(c), T_n - c \rangle}{r_n} + o_{\mathbb{P}}(1)$$

and the result follows (a sequence of r.v.'s that is $o_{\mathbb{P}}(1)$ converges in law to 0). □

Corollary 2.3.2: *Let* $(T_n)_{n\in\mathbb{N}}$ *be an asymptotically normal sequence of estimators for* $g(\theta) \in \Theta \subset \mathbb{R}^p$ *with limiting distributions* $\mathcal{N}(0, \Sigma(\theta))$, $\theta \in \Theta$, *and let* $h : \mathbb{R}^p \to \mathbb{R}$ *be a continuous function differentiable at* $\theta_0 \in \Theta$ *with gradient* $\nabla h(\theta_0)$. *Then*

$$\sqrt{n}(h(T_n) - h(g(\theta_0))) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \langle \nabla h(\theta_0)\Sigma, \nabla h(\theta_0) \rangle) \quad w.r.t.\ \mathbb{P}_{\theta_0}.$$

Corollary 2.3.3: *Let* $(T_n)_{n\in\mathbb{N}}$ *be an asymptotically linear sequence of estimators for* $g(\theta) \in \Theta \subset \mathbb{R}^p$ *with influence function* l_θ *and let* $h : \mathbb{R}^p \to \mathbb{R}$ *be a continuous function differentiable at* $\theta_0 \in \Theta$ *with gradient* $\nabla h(\theta_0)$. *Then* $(h(T_n))_{n\in\mathbb{N}_+}$ *is an asymptotivally linear sequence of estimators for* $h(g(\theta))$ *with influence function* $\langle \nabla h(\theta_0), l_\theta(\cdot) \rangle$.

3

Gaussian Random Variables and Linear Models

3.1 Distributions related to gaussians

Recall that given $X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ then

$$Y = \sum_{j=1}^{n} X_j^2 \sim \chi^2(n)$$

where $\chi^2(n)$ is the *chi-squared distribution with n degrees of freedom*, in particular

$$f_Y(y) = \frac{1}{\Gamma(n/2)2^{n/2}} y^{n/2-1} e^{-y/2}.$$

Theorem 3.1.1 (Cochran): *Fix $X_1, \dots, N_n \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and denote $X = (X_1, \dots, X_n)$. Let $E_1 \oplus \cdots \oplus E_k$ be an orthogonal decomposition of $\mathbb{R}^n$ with respective dimensions $r_1, \dots r_k$ (so $\sum_{j=1}^k r_j = n$). Further denote by $\pi_{E_j} : \mathbb{R}^n \to E_j$ the projection to E_j. Then the r.v.'s $Y_j = \pi_{E_j} X$ are mutually independent and $\|Y_j\|^2 \sim \chi^2(r_j)$.*

Proof. Let $\{\eta_l\}_{l=1,\dots,n}$ be an orthonormal basis of $\mathbb{R}^n$ s.t.

$$E_j = \operatorname{span}\{\eta_l \mid l = \sum_{h=1}^{j-1} r_h + 1, \dots, \sum_{h=1}^{j} r_h\},$$

where we set $r_0 = 1$ and we consider the orthogonal matrix $A^T = (\eta_1 \mid \dots \mid \eta_n) \in \mathbb{R}^{n \times n}$. Observe that for every $j = 1, \dots, n$

$$Z_j = (AX)_j = \langle X, \eta_j \rangle \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$$

then if $J_i = \{\sum_{h=1}^{i-1} r_h + 1, \dots, \sum_{h=1}^{i} r_h\}$, $i = 1, \dots, k$, we have

$$\pi_{E_i} X = \sum_{j \in J_i} \langle X, \eta_j \rangle \eta_j = \sum_{j \in J_i} Z_j \eta_j$$

hence $\pi_{E_j}X$ is independent of $\pi_{E_k}X$ when $j \neq k$ since they are functions of disjoint subsets of independent random variables. Furthermore

$$\|\pi_{E_i}X\|^2 = \sum_{j\in J_i} Z_j^2 \sim \chi^2(r_i).$$

□

In the following given a sample $(X_1, \ldots, X_n)$ we will denote

$$\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$$

$$S^2 = S^2(X) = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2.$$

Theorem 3.1.2: *Let $X = (X_1, \ldots, X_n)$ be a sample with $X_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(m, \sigma^2)$.*

(1) $\frac{\overline{X}-m}{\sigma^2}\sqrt{n} \sim \mathcal{N}(0,1)$*;*

(2) $\frac{1}{\sigma^2}\sum_{j=1}^{n}(X_j - \overline{X})^2 = \frac{n-1}{\sigma^2}S^2 \sim \chi^2(n-1)$*;*

(3) $\overline{X}$ *is independent of* S^2 *and* $\frac{\overline{X}-m}{S}\sqrt{n} \sim t(n-1)$*.*

Proof. (1) Follows from the well known properties of gaussians.

(2) Let $E_1 = \text{span}\{\eta\}$, where $\eta = (1/\sqrt{n}, \ldots, 1/\sqrt{n})^T$ and $E_2 = E^{\perp}$. For $j = 1, \ldots, n$ define $Y_j = \frac{X_j - m}{\sigma}$ and observe that $\overline{Y} = \frac{\overline{X}-m}{\sigma}$ and that $\pi_{E_1}Y = \sqrt{n}\overline{Y}$. Moreover defining $Y = (Y_1, \ldots, Y_n)^T$

$$\begin{aligned}
\frac{\sum_{j=1}^{n}(X_j - \overline{X})^2}{\sigma^2} &= \sum_{j=1}^{n}\left(\frac{(X_j - m) - (\overline{X} - m)}{\sigma}\right)^2 \\
&= \sum_{j=1}^{n}(Y_j - \overline{Y})^2 \\
&= \sum_{j=1}^{n} Y_j^2 + n\overline{Y}^2 - 2\overline{Y}\sum_{j=1}^{n} Y_j \\
&= \sum_{j=1}^{n} Y_j^2 + n\overline{Y}^2 - 2n\overline{Y} \\
&= \sum_{j=1}^{n} Y_j^2 - n\overline{Y}^2 \\
&= \|Y\|^2 - \|\pi_{E_1}Y\|^2 = \|\pi_{E_2}Y\|^2
\end{aligned}$$

and using Cochan's Theorem 3.1.1 we get

$$\frac{\sum_{j=1}^{n}(X_j - \overline{X})^2}{\sigma^2} = \|\pi_{E_2}Y\|^2 \sim \chi^2(\dim(E_2)) = \chi^2(n-1).$$

(3) Follows again by Cochran's Theorem 3.1.1 since $\frac{\overline{X}-m}{\sigma}\sqrt{n} = \overline{Y}\sqrt{n} = \langle \pi_{E_1}Y, \xi\rangle$ (with $\xi = (\sqrt{n}, \ldots, \sqrt{n})$) and $S^2 = \frac{\sigma^2}{n-1}\|\pi_{E_2}Y\|^2$ are functions of independent r.v.'s.

□

3.2 Linear models and least squares linear regression

We collect data $\{(X_j, Y_j)\}_{j=1}^n$, where (X_j, Y_j) are i.i.d., and we suppose:

$$Y_j \,|\, X_j \sim \langle \theta, \phi(X_j) \rangle + \sigma Z_j$$

were we can interpret θ as the influence of the quantity represented by the X_j's on the quantity represented by the Y_j's and the σZ_j's as the influence of other external random factors, where Z_j is supposed to be a centered r.v. with unitary variance. We want to find the conditional law of Y_j w.r.t. X_j, that is characrterized by the parameters (θ, σ).

The least squares linear regression is a technique to estimate θ with the slope of the line minimizing the MSE with the given data, that is:

$$\hat{\theta} \text{ minimizes } \sum_{j=1}^{n} (Y_j - \theta\phi(X_j))^2.$$

Example 3.2.1 (Offsets): Consider the affine model

$$Y_j = \theta_0 + \theta_1 X_j + Z_j$$

with Z_j are centered i.i.d. r.v.'s. Then we can fit

$$\hat{\theta} = \text{argmin}_{(\theta_0,\theta_1)\in\mathbb{R}^2} \sum_{j=1}^{n} (Y_j - (\theta_0 + \theta_1 X_j))^2 = \text{argmin}_{(\theta_0,\theta_1)\in\mathbb{R}^2} \sum_{j=1}^{n} \left(Y_j - \left\langle \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix}, \begin{pmatrix} 1 \\ X_j \end{pmatrix} \right\rangle \right)^2$$

.

Example 3.2.2 (Quadratic): Consider now the model

$$Y_j = \theta_0 + \theta_1 X_j + \theta_2 X_j^2 + Z_j$$

where the Z_j's are as before. We sill find a model fitting the dataset minimizing

$$\hat{\theta} = \text{argmin}_{(\theta_0,\theta_1)\in\mathbb{R}^2} \sum_{j=1}^{n} \left(Y_j - \left\langle \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{pmatrix}, \begin{pmatrix} 1 \\ X_j \\ X_j^2 \end{pmatrix} \right\rangle \right)^2 .$$

Observe that *the model is still linear in* $\theta = (\theta_0, \theta_1, \theta_2)^T$.

Definition 3.2.3: A *linear model* is a statistical model induced by a sample $Y = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$ where we suppose that for some $\theta \in \mathbb{R}^p$ and some $\sigma \in [0, \infty)$ holds

$$Y_j = \sum_{l=1}^{p} \theta_l \Phi_{jl} + \sigma Z_j \quad , \; j = 1, \ldots n$$

where $\Phi \in \mathbb{R}^{n\times p}$, $\mathbb{E}[Z_j] = 0$, $\mathbb{E}[Z_i Z_j] = \delta_{i,j}$ and the unknown parameters are $(\theta_1, \ldots, \theta_p)^T \in \mathbb{R}^p$ and $\sigma^2 \in [0, \infty)$. The above identities can be reformulated in vector notation

$$Y = \Phi\theta + \sigma Z$$

where $Y = (Y_1, \ldots, Y_n)^T$, $Z = (Z_1, \ldots, Z_n)^T$, $\Phi \in \mathbb{R}^{n\times p}$ is called the *desing matrix* and $\theta = (\theta_1, \ldots, \theta_p)^T \in \mathbb{R}^p$ with $\sigma^2 \in [0, \infty)$ are the *parameters*. We say that the above linear model is *underparametrized* if $\text{rank}(\Phi) = p$.

Remark 3.2.4: If we want to represent the situation using an usual statistical model, then we can take

$$(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \{\mathbb{P}_{\theta,\sigma^2}\}_{(\theta,\sigma^2)\in\mathbb{R}^p\times[0,\infty)})$$

where $\mathbb{P}_{\theta,\sigma^2}$ is the law of $\Phi\theta + \sigma Z$.

Remark 3.2.5: Observe that the previous Definition include also the situation described at the beginning of this section since, as we said, we try to figure out the conditional law of the Y_j's w.r.t the X_j's, hence we can consider the latter as fixed so that the random matrix $\Phi(X)_{jl} = \phi(X_j)_l$ becomes a deterministic design matrix.

Definition 3.2.6: The *least squares estimator (LSE)* $\hat{\theta} : \mathbb{R}^n \to \mathbb{R}^p$ of θ is given by

$$\hat{\theta}(y) = \text{argmin}_{\theta\in\mathbb{R}^p}\|y - \Phi\theta\|^2.$$

Lemma 3.2.7: *In the underparametrized setting the LSE $\hat{\theta}$ is unique and it can be written as*

$$\hat{\theta} = (\Phi^T\Phi)^{-1}\Phi^T y.$$

Proof. We want to minimize $f(\theta) = \|y-\Phi\theta\|^2 = (y-\Phi\theta)^T(y-\Phi\theta)$, si imposing $\nabla_\theta f(\theta) = 0$ we easily obtain

$$\Phi^T\Phi\theta = \Phi^T y.$$

Since $\Phi^T\Phi \in \mathbb{R}^{p\times p}$ has full rank, it is invertible. Hence the thesis follows from the above equation. □

Definition 3.2.8: The equations summarized in $\Phi^T\Phi\theta = \Phi^T y$ are called *normal equations*.

Remark 3.2.9: The LSE $\hat{\theta} = \hat{\theta}(y)$ is a linear function.

Remark 3.2.10 (Interpretations of the LSE): We can interpet the LSE in two ways:

- $\hat{\theta} = \text{argmin}_{\theta\in\mathbb{R}^p}\sum_{j=1}^n (y_j - (\Phi\theta)_j)^2$: the LSE minimizes the total y-distance between n points in $p+1$ dimensions (points of the form $(x_1, \dots, x_p, y)^T$) and a p-dimensional plane;

- $\hat{\theta} = \text{argmin}_{\theta\in\mathbb{R}^p}\|y - \Phi\theta\|^2$: so the LSE contains the coefficients of the linear combination of the column vectors of Φ minimizing the euclidean distance with y, that is

  $$\Phi\hat{\theta} = \Phi(\Phi^T\Phi)^{-1}\Phi^T y = \pi_\Phi y$$

  where π_Φ is a projection on the image of Φ. Note in particular that $\pi_\Phi^T = \pi_\Phi$ (if $u, v \in \mathbb{R}^n$, then we can write $u = \pi_\Phi u + a$ with a orthogonal to the image of Φ, so we easily obtain $\langle u, \pi_\Phi v\rangle = \langle \pi_\Phi u, \pi_\Phi v\rangle$ and analogously $\langle \pi_\Phi u, v\rangle = \langle \pi_\Phi u, \pi_\Phi v\rangle$, hence $\langle \pi_\Phi u, v\rangle = \langle u, \pi_\Phi v\rangle$).

Definition 3.2.11: We say that the linear model $Y = \Phi\theta + \sigma Z$ is *gaussian* if $Z \sim \mathcal{N}(0, \mathbb{1}_n)$, where $\mathbb{1}_n \in \mathbb{R}^{n\times n}$ is the identity matrix.

Remark 3.2.12: Observe that if $Y = \Phi\theta + \sigma Z$ is gaussian then $Y \sim \mathcal{N}(\Phi\theta, \sigma^2 \mathbb{1}_n)$, that has the density

$$\begin{aligned} p_{\theta,\sigma^2}(y) &= (2\pi)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\|y - \Phi\theta\|^2\right) \\ &= (2\pi)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\|y\|^2 + \frac{1}{\sigma^2}\langle y, \Phi\theta\rangle - \frac{1}{2\sigma^2}\|\Phi\theta\|^2 - n\log(\sigma)\right) \\ &= (2\pi)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\|y\|^2 + \left\langle \frac{\Phi\theta}{\sigma^2}, \pi_\Phi y\right\rangle - \frac{1}{2\sigma^2}\|\Phi\theta\|^2 - n\log(\sigma)\right) \end{aligned}$$

is a 2-dimensional exponential family with statistics

$$T_1(y) = \pi_\Phi y \quad , \quad c_1(\theta) = \frac{\Phi\theta}{\sigma^2}$$

$$T_2(y) = \|y\|^2 \quad , \quad c_2(\theta) = -\frac{1}{2\sigma^2}.$$

Note also that if we are in the underparametrized case, that is Φ has full rank, then $T(y) = (T_1(y), T_2(y))^T$ is a sufficient and complete statistic.

Lemma 3.2.13: *Let $A \in \mathbb{R}^{n\times n}$ be an matrix and Z be a random vector in $\mathbb{R}^n$ with $\mathbb{E}[Z] = 0$ and $\mathrm{Cov}(Z_i, Z_j) = \Sigma_{ij}$ for every $i, j = 1, \ldots, n$. Then for $W = AZ$ we have*

$$\mathrm{Cov}(W_i, W_j) = (A\Sigma A^T)_{ij}$$

for every $i, j = 1, \ldots, n$.

Proof. Fix $i, j = 1, \ldots, n$, then

$$\begin{aligned} \mathrm{Cov}(W_i, W_j) = \mathbb{E}[W_i W_j] &= \mathbb{E}\left[\left(\sum_{l=1}^n A_{il} Z_l\right)\left(\sum_{k=1}^n A_{jk} Z_k\right)\right] \\ &= \sum_{l,k=1}^n A_{il} A_{jk} \mathbb{E}[Z_l, Z_k] = \sum_{l,k=1}^n A_{il} A_{jk} \Sigma_{lk} = (A\Sigma A^T)_{ij}. \end{aligned}$$

□

We mantain the notations expectations used in the previous sections.

Theorem 3.2.14 (Gauss-Markov): *Suppose that we are in the underparametrized case. The estimator $y \mapsto (\Phi^T\Phi)^{-1}\Phi^T y$ is an unbiased estimator of θ, optimal among all the linear, unbiased estimators and $y \mapsto \|\Phi(\Phi^T\Phi)^{-1}\Phi y - y\|^2$ is an unbiased estimator of $(n-p)\sigma^2$. Moreover in the gaussian case, these estimators are optimal among all the unbiased estimators (linear and nonlinear w.r.t. θ) of $\theta, (n-p)\sigma^2$.*

Proof. Let $T_V : y \mapsto Vy$ be a linear estimator of θ (V is a matrix). Then

$$\mathbb{E}_{\theta,\sigma^2}[T_V] = \mathbb{E}_{\theta,\sigma^2}[VY] = \mathbb{E}_{\theta,\sigma^2}[V(\Phi\theta + \sigma Z)] = V\Phi\theta + V\sigma\mathbb{E}[Z] = V\Phi\theta$$

so T_V is unbiased if and only if $\theta = V\Phi\theta$ for every θ, that is if and only if $V\Phi = \mathbb{1}_p$. Now if we choose $V = V_\Phi = (\Phi^T\Phi)^{-1}\Phi$, then $V_\Phi\Phi = (\Phi^T\Phi)^{-1}\Phi^T\Phi = \mathbb{1}_p$, hence $y \mapsto (\Phi^T\Phi)^{-1}\Phi y$

is unbiased. Now for the risk

$$\begin{aligned}\mathbb{E}_{\theta,\sigma^2}\left[\|VY-\theta\|^2\right] &= \mathbb{E}\left[\|V\Phi\theta+\sigma Z-\theta\|^2\right] = \sigma^2\mathbb{E}\left[\|VZ\|^2\right] \\ &= \sigma^2\mathbb{E}\left[\sum_{i=1}^{p}\left(\sum_{l=1}^{n}V_{il}Z_l\right)^2\right] = \sigma^2\sum_{i=1}^{p}\mathbb{E}\left[\sum_{j,l=1}^{n}V_{ij}V_{il}Z_jZ_l\right] \\ &= \sigma^2\sum_{i=1}^{p}V_{ij}V_{il}\delta_{jl} = \sigma^2\sum_{i=1}^{p}\sum_{l=1}^{n}V_{il}^2 = \sigma^2\|V\|_F^2.\end{aligned}$$

So to minimize $\mathbb{E}\left[\|VY-\theta\|^2\right]$ subject to $V\Phi=\mathbb{1}_p$ we minimize $\|V\|_F^2$ subject to $V\Phi=\mathbb{1}_p$. We have

$$V_\Phi = \mathbb{1}_pV_\Phi = V\Phi V_\Phi = V\pi_\Phi$$

hence

$$\|V_\Phi\|_F^2 = \|V_\Phi^T\|_F^2 = \|\pi_\Phi^TV^T\|_F^2 = \|\pi_\Phi V^T\|_F^2 \leq \|V^T\|_F^2 = \|V\|_F^2$$

since the projection π_Φ does not increase the norm. We have proved the first part of the theorem.

Consider now the estimator $y\mapsto\|\pi_\Phi y-y\|^2$, we have

$$\begin{aligned}\|\pi_\Phi Y-Y\|^2 &= \|\pi_\Phi(\Phi\theta+\sigma Z)-(\Phi\theta+\sigma Z)\|^2 \\ &= \|(\pi_\Phi\Phi-\Phi)\theta+\sigma(\pi_\Phi Z-Z)\|^2 \\ &= \sigma^2\|\pi_\Phi Z-Z\|^2.\end{aligned}$$

In general there exists a matrix $A\in O(n)$ s.t. $\pi_\Phi = A^T\pi_pA$, where $\pi_p=\mathrm{diag}(1,\dots,1,0\dots,0)$ with p 1's. So we can write

$$\begin{aligned}\|\pi_\Phi Z-Z\|^2 &= \|A^T\pi_pAZ-A^TAZ\|^2 \\ &= \|A^T(\pi_pAZ-AZ)\|^2 = \sum_{j=p+1}^{n}(AZ)_j^2\end{aligned}$$

that holds

$$\begin{aligned}\mathbb{E}\left[\|\pi_\Phi Z-Z\|^2\right] &= \sum_{j=p+1}^{n}\mathbb{E}\left[(AZ)_j^2\right] \\ &= \sum_{j=p+1}^{n}\mathrm{Cov}((AZ)_j,(AZ)_j) \\ &= \sum_{j=p+1}^{n}(A^T\mathbb{1}_nA)_{jj} = \sum_{j=p+1}^{n}(\mathbb{1}_n)_{jj} = n-p.\end{aligned}$$

This gives that $y\mapsto\|\pi_\Phi y-y\|^2$ is an unbiased estimator of $(n-p)\sigma^2$.

Assume now $Z_j\overset{i.i.d.}{\sim}\mathcal{N}(0,1)$, $j=1,\dots,n$. From what we said on the previous remark we have that $S(y)=(\pi_\Phi y,\|y\|^2)^T$ is a suffiecient and complete statistic, so we only have to prove that $y\mapsto(\Phi^T\Phi)^{-1}\Phi^Ty$ and $y\mapsto\|\pi_\Phi y-y\|^2$ are functions of $S(y)$ (so that their conditional expectations under S are themselves). Observe that

$$\begin{aligned}(\Phi^T\Phi)^{-1}\Phi^Ty &= (\Phi^T\Phi)^{-1}(\pi_\Phi\Phi)^Ty \\ &= (\Phi^T\Phi)^{-1}\Phi^T\pi_\Phi^Ty \\ &= (\Phi^T\Phi)^{-1}\Phi^T(\pi_\Phi y)\end{aligned}$$

Furthermore we can write $\|y\|^2 = \|\pi_\Phi y\|^2 + \|y - \pi_\Phi y\|^2$, that gives

$$\|y - \pi_\Phi y\|^2 = \|y\|^2 - \|\pi_\Phi y\|^2.$$

□

Remark 3.2.15: In the context of the previous theorem, in the gaussian case, we have

$$\begin{aligned}\frac{1}{\sigma^2}\|Y - \pi_\Phi Y\|^2 &= \frac{1}{\sigma^2}\|\Phi\theta + \sigma Z - \pi_\Phi(\Phi\theta + \sigma Z)\|^2 \\ (\pi_\Phi \Phi\theta = \Phi\theta) \ &= \|(\mathbb{1} - \pi_\Phi)Z\|^2 \sim \chi^2(n-p).\end{aligned}$$

4

Theory of Tests

4.1 Confidence intervals

A *cumulative density function (CDF)* F is a function $F : \mathbb{R} \to [0,1]$ s.t. it is nondecreasing and continuous on the right. For a real r.v. X its cumulative distribution function F_X is a CDF as

$$F_X(x = \mathbb{P}(X \leq x)) \quad \forall x \in \mathbb{R}.$$

Definition 4.1.1: For a given CDF F, define $F^{\leftarrow} : (0,1) \to \mathbb{R}$ the *quantile function* (or *generalized inverse*) of F as

$$F^{\leftarrow}(\alpha) = \inf\{x \in \mathbb{R} \mid F(x) \geq \alpha\} \quad \forall \alpha \in (0,1).$$

For a fixed $\alpha \in (0,1)$ the number $F^{\leftarrow}(\alpha)$ is the *quantile of order* α.

Remark 4.1.2: If F is continuous and strictly increasing (in particular bijective) then

$$F^{\leftarrow}(\alpha) = F^{-1}(\alpha) \quad \forall \alpha \in (0,1).$$

If F is also symmetric (i.e. $F(-x) = 1 - F(x)$), then

$$F^{-1}(x) = -F^{-1(1-\alpha)}.$$

Indeed

$$F(-F^{-1(1-\alpha)}) = 1 - F(F^{-1}(1-\alpha)) = 1 - 1 + \alpha = \alpha.$$

Let us fix some notation. For a fixed $\alpha \in (0,1)$ we denote $\phi_\alpha, t_\alpha(n), \chi^2_\alpha(n)$ the quantiles of order α for the standard normal, t-student and χ^2 (with n degrees of freedom) disributions respectively.

Fix now a statistical model $(\mathbb{X}, \mathcal{F}, \{\mathbb{P}_\theta\}_{\theta \in \Theta})$.

Definition 4.1.3: Let $\alpha \in (0,1)$, $g : \Theta \to \mathbb{R}$ and $S : \mathbb{X} \to \mathcal{B}(\mathbb{R})$ be functions s.t. for all $\theta \in \Theta$ holds

$$\{x \in \mathbb{X} \mid g(\theta) \notin S(x)\} \in \mathcal{F}.$$

We say that S is a *confidence interval (CI)* of level $1-\alpha$ for $g(\theta)$ if

$$\mathbb{P}_\theta(\{x \in \mathbb{X} \mid g(\theta) \in S(x)\}) \geq 1-\alpha \quad \forall \theta \in \Theta.$$

Remark 4.1.4: Tipically $S(x)$ is a (random) interval of $\mathbb{R}$ if $\Theta \subset \mathbb{R}$. There are two special cases:

- if $S(x) = (S_-(x), S_+(x))$ the CI is two-sided;
- if $S(x) = (-\infty, S_+(x))$ the CI is one-sided (left).

Example 4.1.5: Let $X \sim \text{Exp}(\theta)$. We want to find T_1, T_2 s.t. $\mathbb{P}_\theta(\theta \in (T_1(X), T_2(X))) \geq 0.95$.

Note that $Q = \theta X \sim \text{Exp}(1)$, hence

$$1 - e^{-t} = \mathbb{P}_\theta(Q \leq t) = \mathbb{P}_\theta(X\theta \leq t) = \mathbb{P}_\theta(\theta \leq t/X)$$

from which follows

$$(1-e^{-b}) - (1-e^{-a}) = \mathbb{P}_\theta(Q \in (a,b)) = \mathbb{P}_\theta(\theta \in (a/X, b/X))$$

so $(a/X, b/X)$ is a CI at level $1-\alpha$ if $e^{-a} - e^{-b} = 1-\alpha$.

In what follows we use the method of *pivotal quantity* to identify CIs. This method relies on finding a *pivot* $Q : \mathbb{X} \times \Theta \to \mathbb{R}$ that is invertible w.r.t. $\theta \in \Theta$ and s.t. its law does not depend on $\theta \in \Theta$.

Given an estimator $T : \mathbb{X} \to \mathbb{R}$ of $g(\theta) \in \mathbb{R}$ we try to find a CI by

(1) finding a function $\tilde{Q} : \Gamma \times \Theta \to \mathbb{R}$ s.t. $Q(x;\theta) = \tilde{Q}(T(x);\theta)$ is a pivot;

(2) identify the CI: e.g. for two sided CI

$$S(x) = Q^{-1}(x, (F^{\leftarrow}(\alpha/2), F^{\leftarrow}(1-\alpha/2))).$$

Example 4.1.6 (Normal distribution, known variance): Suppose $X_j \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$ and take $g(\theta) = \mu$. Consider the estimator $\overline{x} = \frac{1}{n}\sum_{j=1}^n x_j$. Standardize the estimation $\overline{X}$ to find a pivot

$$Q = \frac{\overline{X} - \mu}{\sigma}\sqrt{n} \sim \mathcal{N}(0,1).$$

So if Φ is the CDF of $\mathcal{N}(0,1)$ we have

$$\begin{aligned}
\Phi(b) - \Phi(a) &= \mathbb{P}_\theta(Q \in (a,b)) \\
&= \mathbb{P}_\theta\left(\frac{\overline{X} - \mu}{\sigma}\sqrt{n} \in (a,b)\right) \\
&= \mathbb{P}_\theta\left(\mu \in \left(\overline{X} - \frac{b}{\sqrt{n}}\sigma, \overline{X} + \frac{a}{\sqrt{n}}\sigma\right)\right)
\end{aligned}$$

So we found a CI for μ at level $\Phi(b) - \Phi(a)$. If a level $1-\alpha$ is fixed choose e.g. $b = \Phi^{-1}(1-\alpha/2)$ and $a = \Phi^{-1}(\alpha/2)$.

4.2 General theory of tests

Performing a *statistical test* aims to verify if a certain *hypothesis* concerning the parameter θ that, generated the data, is *plausible*.

Definition 4.2.1: Given a statistical model $(\mathbb{X}, \mathcal{F}, \{\mathbb{P}_\theta\}_{\theta\in\Theta})$, consider a partition $\Theta = \Theta_0 \sqcup \Theta_1$ in nonempty sets Θ_0, Θ_1. The *null hypothesis* is

$$H_0 : \theta \in \Theta_0$$

the *alternative hypothesis* is

$$H_1 : \theta \in \Theta_1.$$

Example 4.2.2 (Quality control): Sample at random products from an assembly line

$$X_j = \begin{cases} 1 & \text{if the } j\text{-th sample has a defect} \\ 0 & \text{otherwise} \end{cases} \sim \mathrm{Ber}(\theta).$$

Take $\Theta = (0, 1)$, then one can choose e.g. $\Theta_0 = (0, 0.1)$ and $\Theta_1 = [0.1, 1)$, so that the null and alternative hypothesis are $H_0 : \theta \in (0, 0.1)$ and $H_1 : \theta \in [0.1, 1)$.

Definition 4.2.3: We call *deterministic test* a measurable function $\Phi : \mathbb{X} \to \{0, 1\}$ of the form

$$\Phi(x) = \mathbb{1}_C(x) \quad \text{for a } C \in \mathcal{F}.$$

The set C is called *critical region* of the test.

Remark 4.2.4: The critical region is the set of experiment results that are highly incompatible with the null hypothesis H_0 (give evidence that H_0 does not hold, so if $\Phi(\mathbb{X}) = \{1\}$ then we accept the alternative hypothesis H_1). The set C^c can be called *acceptence region* of the test.

Example 4.2.5 (Quality control): Assume $n = 100$ sampled items and we reject H_0 if the number of samples with defect is strictly larger than 11. Then we consider

$$T(x) = \sum_{j=1}^{100} x_j \quad , \quad \Phi(x) = \begin{cases} 0 & \text{if } T(x) \leq 11 \\ 1 & \text{if } T(x) > 11 \end{cases}$$

that is our test Φ has critical region $C = \{x = (x_1, \ldots, x_{100})^T \mid T(x) > 11\}$.

Tests can also be random.

Definition 4.2.6: A *(randomized) test* is a measurable function $\Phi : \mathbb{X} \to [0, 1]$.

Remark 4.2.7: Here, for a given $x \in \mathbb{X}$, $\Phi(x)$ can be interpreted as the probability $\mathbb{P}(H_0 \text{ rej.} \,|\, x)$ that the randomized test will reject H_0 given x.

For example $\Phi(x) = 0.05$ for all $x \in \mathbb{X}$ is rejecting H_0 with $\mathbb{P} = 5\%$ for all $x \in \mathbb{X}$.

Remark 4.2.8: Setting $g(\theta) = \mathbb{1}_{\Theta_1}(\theta)$, then we can interpret Φ as an estimator of $g(\theta)$.

Analogously to what was done in estimation theory, we "evaluate" the prediction of a test through a "cost function", this cost function should be 0 for correct prediction and 1 for wrong prediction, e.g.

$$\ell(\theta, \mathbb{1}_C(x)) = |\mathbb{1}_{\Theta_1}(\theta) - \mathbb{1}_C(x)|$$

so the loss/cost function can be chosen as

$$\ell(\theta, a) = |\mathbb{1}_{\Theta_1}(\theta) - a| = \begin{cases} a & \text{if } \theta \in \Theta_0 \\ 1 - a & \text{if } \theta \in \Theta_1 \end{cases}.$$

This can be extended to randomized tests in the natural way

$$\ell(\theta, \Phi(x)) = |\mathbb{1}_{\Theta_1}(\theta) - \Phi(x)| = \begin{cases} \mathbb{P}(H_0 \text{ rej.} \,|\, x) & \text{if } \theta \in \Theta_0 \\ 1 - \mathbb{P}(H_0 \text{ rej.} \,|\, x) & \text{if } \theta \in \Theta_1 \end{cases}.$$

To compare two estimators we need to comparetheir average error.

Definition 4.2.9: For a test Φ the *risk* is given by

$$\mathcal{R}_\Phi(\theta) = \mathbb{E}_\theta\left[\ell(\theta, \Phi)\right] = \begin{cases} \mathbb{E}_\theta\left[\Phi\right] & \text{if } \theta \in \Theta_0 \\ 1 - \mathbb{E}_\theta\left[\Phi\right] & \text{if } \theta \in \Theta_1 \end{cases}.$$

Remark 4.2.10: In both cases ($\theta \in \Theta_0$ and $\theta \in \Theta_1$) the risk $\mathcal{R}_\Phi$ can be interpreted as probabilities:

$$\mathbb{E}_\theta\left[\Phi\right] = \int_{\mathbb{X}} \Phi(x)\, \mathrm{d}\mathbb{P}_\theta(x) = \int_{\mathbb{X}} \mathbb{P}(H_0 \text{ rej.} \,|\, x)\, \mathrm{d}\mathbb{P}_\theta(x) = \mathbb{P}_\theta(\text{reject } H_0).$$

So we can say

$$\mathcal{R}_\Phi(\theta) = \begin{cases} \mathbb{P}_\theta(\text{reject } H_0) & \text{if } \theta \in \Theta_0 \\ \mathbb{P}_\theta(\text{accept } H_0) & \text{if } \theta \in \Theta_1. \end{cases}$$

Each of the two quantities on the RHS of the above equation corresponds to the probabilities of one of the two types of errors that one can make:

(1) *type 1 error*: rejecting the null hypothesis H_0 when H_0 is true;

(2) *type 2 error*: accepting the null hypothesis H_0 when it is false.

Like what we did with estimators we will rank tests based on their risk.

Definition 4.2.11: A test Φ is *preferable* to a test Φ' if

$$\mathcal{R}_\Phi(\theta) \leq \mathcal{R}_{\Phi'}(\theta) \quad \forall \theta \in \Theta$$

and *strictly preferable* if

$$\mathcal{R}_\Phi(\theta) < \mathcal{R}_{\Phi'}(\theta) \quad \forall \theta \in \Theta.$$

A test is called *optimal* if it is preferable to all other tests an *admissible* if no other test is strictly preferable to it.

Remark 4.2.12: This defines only a partial order (one can make examples).

Remark 4.2.13: Like in the case for estimators, where we can decide that the two components (variance and bias) play an asymmetric role for the choice of the estimator, here we can decide to believe (and this is done quite often) that is better to commit a type 2 error then a type 1 (for example in the case of DNA for homicide with null hypothesis H_0: innocence).

Definition 4.2.14: We say that a test Φ has level α for $\alpha \in (0,1)$ if

$$\sup_{\theta \in \Theta_0} \mathbb{E}_\theta [\Phi] \leq \alpha.$$

In particular if a test is of level α then the probability of a type 1 error is uniformly bounded by α. On the contrary considering the probability of not type 2 error:

Definition 4.2.15: The *power* of a test Φ on Θ_1 is the function $\Theta_1 \ni \theta \mapsto \mathbb{E}_\theta [\Phi]$.

Usually one fixes the level of the tests and ranks them according to their power (that is similar to fixing the bias and ranking estimators according to the variance).

Definition 4.2.16: Fix the hypothesis H_0 and H_1. The test Φ is *uniformly most powerful (UMP)* among tests of level α if $\sup_{\theta \in \Theta_0} \mathbb{E}_\theta [\Phi] = \alpha$ and for all tests Φ' of level α

$$\mathbb{E}_\theta [\Phi] \geq \mathbb{E}_\theta [\Phi'] \quad \forall \theta \in \Theta_1.$$

Definition 4.2.17: We say that a test is *unbiased* if

$$\mathbb{E}_{\theta_0} [\Phi] \leq \mathbb{E}_{\theta_1} [\Phi]$$

for all $\theta_0 \in \Theta_0$ and $\theta_1 \in \Theta_1$. A test that is unbiased and UMP will be called *uniformly most powerful unbiased (UMPU).*

Definition 4.2.18: A test is called *one-sided* if it has the form

$$H_0 : \theta \leq \theta_0 \quad , \quad H_1 : \theta > \theta_0 \quad (\textit{right-sided alternative})$$

or

$$H_0 : \theta \geq \theta_0 \quad , \quad H_1 : \theta < \theta_0 \quad (\textit{left-sided alternative})$$

for a $\theta_0 \in \mathbb{R}$ and *two-sided* if

$$H_0 : \theta \in [\theta_1, \theta_2] \quad , \quad H_1 : \theta \notin [\theta_1, \theta_2]$$

for some $\theta_1, \theta_2 \in \mathbb{R}$, $\theta_1 \leq \theta_2$ (possily/often $\theta_1 = \theta_2$).

Example 4.2.19 (Quality control): If $\theta \in (0,1)$ the test

$$H_0 : \theta \in (0, 1/10] \quad , \quad H_1 : \theta \in (1/10, 1)$$

is a one-sided (right-sided alternative) test.

4.3 Neyman-Pearson tests

In this section (that is until we will specify otherwise) will be $\Theta = \{\theta_0, \theta_1\}$, $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$. In this setting the "best" tests can be identified by Neyman-Pearson lemma.

Consider the dominated statistical model $(\mathbb{X}, \mathcal{F}, \{\mathbb{P}_i\}_{i=1,2})$ with dominating measure μ (that can be e.g. $\mu = \mathbb{P}_1 + \mathbb{P}_2$). Let p_i be the density of $\mathbb{P}_i$ w.r.t. μ for $i = 1, 2$.

Definition 4.3.1: A *Neyman-Pearson test* is a randomized test Φ s.t.

$$\Phi(x) = \begin{cases} 1 & \text{if } p_1(x) > cp_0(x) \\ 0 & \text{if } p_1(x) < cp_0(x) \end{cases}$$

for a non-negative constant $c > 0$ (note that it can b anything in the case $p_1(x) = cp_0(x)$).

Lemma 4.3.2 (Neyman-Pearson): *The Neyman-Pearson test for the hypothesis H_0: $\theta = \theta_0$ and H_1: $\theta = \theta_1$ given by*

$$\Phi_{\mathrm{NP}}(x) = \begin{cases} 1 & \textit{if } p_1(x) > cp_0(x) \\ \gamma & \textit{if } p_1(x) = cp_0(x) \\ 0 & \textit{if } p_1(x) < cp_0(x) \end{cases}$$

for constants $c > 0$ and $\gamma \in [0, 1]$ at level $\alpha = \mathbb{E}_{\theta_0}[\Phi_{\mathrm{NP}}]$ is admissible and UMP among all the tests at level α.

Furthermore, in this setting all admissible tests are Neyman-Pearsons tests.

Proof. Let Φ be another test for the same H_0 and H_1. Let also

$$g(x) = (\Phi(x) - \Phi_{\mathrm{NP}}(x))(p_1(x) - cp_0(x)) \quad \forall x \in \mathbb{X}.$$

Note that:

- if $p_1(x) < cp_0(x) \Rightarrow \Phi_{\mathrm{NP}}(x) = 0 \Rightarrow \Phi(x) - \Phi_{\mathrm{NP}}(x) \geq 0 \Rightarrow g(x) \leq 0$;
- if $p_1(x) = cp_0(x) \Rightarrow g(x) = 0$;
- if $p_1(x) > cp_0(x) \Rightarrow \Phi_{\mathrm{NP}}(x) = 1 \Rightarrow \Phi(x) - \Phi_{\mathrm{NP}}(x) \leq 0 \Rightarrow g(x) \leq 0$.

So $g(x) \leq 0$ for every $x \in \mathbb{X}$ and it holds

$$\begin{aligned} 0 &\geq \int_{\mathbb{X}} g(x)\,\mathrm{d}\mu(x) \\ &= \left[\int_{\mathbb{X}} p_1(x)\Phi(x)\,\mathrm{d}\mu(x) - \int_{\mathbb{X}} p_1(x)\Phi_{\mathrm{NP}}(x)\,\mathrm{d}\mu(x)\right] - c\left[\int_{\mathbb{X}} p_0(x)\Phi(x)\,\mathrm{d}\mu(x) - \int_{\mathbb{X}} p_0(x)\Phi_{\mathrm{NP}}(x)\,\mathrm{d}\mu(x)\right] \end{aligned}$$

that is

$$\mathbb{E}_{\theta_1}[\Phi] - \mathbb{E}_{\theta_1}[\Phi_{\mathrm{NP}}] \leq c\,[\mathbb{E}_{\theta_0}[\Phi] - \mathbb{E}_{\theta_0}[\Phi_{\mathrm{NP}}]] \tag{4.1}$$

Now we see that Φ_{NP} is admissible. Assume that $\mathcal{R}_{\Phi_{\mathrm{NP}}} > \mathcal{R}_{\Phi}$ for every $\theta \in \Theta$, then

$$\mathbb{E}_{\theta_0}[\Phi] < \mathbb{E}_{\theta_0}[\Phi_{\mathrm{NP}}]$$

$$\mathbb{E}_{\theta_1}[\Phi] > \mathbb{E}_{\theta_1}[\Phi_{\mathrm{NP}}]$$

and this contradicts (4.1) since $c > 0$ (the quantity in the RHS of (4.1) would be both > 0 and < 0).

Then we see that Φ_{NP} is UMP. Fix the level $\alpha \in (0,1)$ of Φ_{NP} and take Φ another test of level α, then by (4.1) we have

$$\mathbb{E}_{\theta_1}\left[\Phi_{\mathrm{NP}}\right] \geq \mathbb{E}_{\theta_1}\left[\Phi\right].$$

We do not prove that all admissible tests in this setting are NP. □

Remark 4.3.3: The test Φ_{NP} rejects H_0 when the so-called *likelihood ratio* $\frac{p_1}{p_0}$ is larger than a threshold $c = c(\alpha)$, i.e. if H_1 is more likely than H_0 by a certain margin.

Corollary 4.3.4: *Suppose we are in the setting of the Neyman-Pearson Lemma 4.3.2. The Neyman-Pearson test* Φ_{NP} *is UMPU.*

Proof. Set $\alpha = \mathbb{E}_{\theta_0}\left[\Phi_{\mathrm{NP}}\right]$ and define the constant test $\Phi = \alpha$. Then since by the previous Theorem Φ_{NP} is UMP we have

$$\mathbb{E}_{\theta_1}\left[\Phi_{\mathrm{NP}}\right] \geq \mathbb{E}_{\theta_1}\left[\Phi\right] = \alpha = \mathbb{E}_{\theta_0}\left[\Phi_{\mathrm{NP}}\right].$$

□

4.4 Increasing likelihood ratio models

How can we construct "efficient" tests for non-simple hypothesis classes?

Consider a general statistical model $(\mathbb{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ dominated by a measure μ with densities $(p_\theta)_{\theta\in\Theta}$.

Definition 4.4.1: If there exists a function $T : \mathbb{X} \to \mathbb{R}$ s.t. for every $\theta_1, \theta_2 \in \Theta$ with $\theta_1 < \theta_2$ there exists a strictly increasing function $f_{\theta_1,\theta_2} : T(\mathbb{X}) \subset \mathbb{R} \to \mathbb{R}$ s.t.

$$f_{\theta_1,\theta_2}(T(x)) = \frac{p_{\theta_2}(x)}{p_{\theta_1}(x)}$$

then we say that the model is an *increasing likelihood ratio model* w.r.t. T.

Example 4.4.2: Suppose that $(\mathbb{P}_\theta)_{\theta\in\Theta}$ is an exponential family with densities $(p_\theta)_{\theta\in\Theta}$ s.t.

$$p_\theta(x) = \exp\left(T(x)c(\theta) - d(\theta)\right) h(x)$$

with $\theta \mapsto c(\theta)$ increasing. Then

$$\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)} = \exp\left(T(x)(c(\theta_2) - c(\theta_1)) - (d(\theta_2) - d(\theta_1))\right)$$

is increasing as a function of T, that is this is an increasing likelihood ratio model.

Example 4.4.3: Consider the statistical model induced by a sample $X = (X_1, \ldots, X_n)$ with $X_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with $\theta = \mu$. Then we are in the contect of the previous Example

since this model has densities

$$p_\theta(x) = \exp\left(-\frac{n}{2\sigma^2}\left(\mu^2 - \frac{2\mu}{n}\sum_{j=1}^n x_j + \left(\frac{1}{n}\sum_{j=1}^n x_j\right)^2\right)\right)$$
$$= \exp\left(\left(\frac{1}{n}\sum_{j=1}^n x_j\right)\left(\frac{n\mu}{2\sigma^2}\right) - \frac{n\mu}{2\sigma^2}\right)\exp\left(-\frac{1}{2\sigma^2}\left(\sum_{j=1}^n x_j\right)^2\right).$$

so it is an increasing likelihood ratio model w.r.t. $T(x) = \frac{1}{n}\sum_{j=1}^n x_j$.

Theorem 4.4.4 (Neyman-Pearson): *Suppose that the considered statistical model is an increasing likelihood ratio model w.r.t. T. Fix $C \in \mathbb{R}$, $\gamma \in (0,1)$ and consider*

$$\Phi_{\text{NP}}(x) = \mathbb{1}(T(x) > C) + \gamma \mathbb{1}(T(x) = C).$$

Then Φ_{NP} is admissible and UMPU among all tests for

$$H_0 : \theta \leq \theta_0 \quad , \quad H_1 : \theta > \theta_0$$

at level $\mathbb{E}_{\theta_0}[\Phi_{\text{NP}}]$.

Proof. For every $\tilde{\theta}_1, \tilde{\theta}_2 \in \Theta$ s.t. $\tilde{\theta}_1 < \tilde{\theta}_2$ consider the test

$$\tilde{H}_0 : \theta = \tilde{\theta}_1 \quad , \quad \tilde{H}_1 : \theta = \tilde{\theta}_2.$$

The critical region of the test is

$$\{x \in \mathbb{X} \mid T(x) > C\} = \left\{x \in \mathbb{X} \mid f_{\tilde{\theta}_1,\tilde{\theta}_2}(T(x)) > f_{\tilde{\theta}_1,\tilde{\theta}_2}(C)\right\}$$
$$= \left\{x \in \mathbb{X} \mid \frac{p_{\tilde{\theta}_2}(x)}{p_{\tilde{\theta}_1}(x)} > f_{\tilde{\theta}_1,\tilde{\theta}_2}(C)\right\}$$
$$= \left\{x \in \mathbb{X} \mid p_{\tilde{\theta}_2}(x) > f_{\tilde{\theta}_1,\tilde{\theta}_2}(C) p_{\tilde{\theta}_1}(x)\right\}$$

so calling $\tilde{C} = f_{\tilde{\theta}_1,\tilde{\theta}_2}(C)$ we have that

$$\Phi_{\text{NP}}(x) = \mathbb{1}\left(p_{\tilde{\theta}_2}(x) > \tilde{C} p_{\tilde{\theta}_1}(x)\right) + \gamma \mathbb{1}\left(p_{\tilde{\theta}_2}(x) = \tilde{C} p_{\tilde{\theta}_1}(x)\right)$$

is a Neyman-Pearson test for every $\tilde{\theta}_1, \tilde{\theta}_2 \in \Theta$ s.t. $\tilde{\theta}_1 < \tilde{\theta}_2$. In Corollary 4.3.4 we proved that a Neyman-Pearson test is unbiased and in our case this means that for every $\tilde{\theta}_1, \tilde{\theta}_2 \in \Theta$ s.t. $\tilde{\theta}_1 < \tilde{\theta}_2$ holds

$$\mathbb{E}_{\tilde{\theta}_1}[\Phi_{\text{NP}}] < \mathbb{E}_{\tilde{\theta}_2}[\Phi_{\text{NP}}]. \tag{4.2}$$

In particular Φ_{NP} is unbiased and

$$\sup_{\theta \leq \theta_0} \mathbb{E}_\theta[\Phi_{\text{NP}}] = \mathbb{E}_{\theta_0}[\Phi_{\text{NP}}]$$

so the level of Φ_{NP} is $\mathbb{E}_{\theta_0}[\Phi_{\text{NP}}]$. Now we show that the test is UMP (that implies also that it is admissible). Take Φ another test at level $\mathbb{E}_{\theta_0}[\Phi_{\text{NP}}]$, we want to show that for every $\theta_1 > \theta_0$ holds $\mathbb{E}_{\theta_1}[\Phi_{\text{NP}}] \geq \mathbb{E}_{\theta_1}[\Phi]$. Consider any $\theta_1 > \theta_0$, then Φ_{NP} is a Neyman-Pearson test for the hypothesis

$$\tilde{H}_0 : \theta = \theta_0 \quad , \quad \tilde{H}_1 : \theta = \theta_1$$

so, by Lemma 4.3.2 we have that Φ_{NP} is UMP on the hypothesis $\tilde{H}_0, \tilde{H}_1$. Hence

$$\mathbb{E}_{\theta_1}[\Phi_{\text{NP}}] \geq \mathbb{E}_{\theta_1}[\Phi]$$

for every $\theta_1 > \theta_0$, so that Φ_{NP} is UMP on H_0, H_1. □

Example 4.4.5 (One-sided test for the mean of a gaussian): Suppose we are in the context of Example 4.4.3. The test

$$\Phi_{\text{NP}}(x) = \mathbb{1}(T(x) > C) + \gamma \mathbb{1}(T(x) = C)$$

for some $C \in \mathbb{R}$ and $\gamma \in [0, 1]$ is UMPU for the hypothesis

$$H_0 : \theta \leq \theta_0 \quad , \quad H_1 : \theta > \theta_0$$

and it is at level

$$\begin{aligned}\mathbb{E}_{\theta_0}[\Phi_{\text{NP}}] &= \mathbb{P}_{\theta_0}(T > C) \\ &= 1 - \mathbb{P}_{\theta_0}(T \leq C) = 1 - F\left(\frac{C - \theta_0}{\sigma}\right)\end{aligned}$$

so if $\mathbb{E}_{\theta_0}[\Phi_{\text{NP}}] = \alpha$ we can choose $C = C_\alpha = \theta_0 + \sigma F^{\leftarrow}(1-\alpha)$ to obtain the level α. Where F is the cumulative density function of the standard gaussian distribution.

Theorem 4.4.6: *Under the hypothesis of the Neyman-Pearson Theorem 4.4.4, for a given value of $\alpha \in (0, 1)$ assume that there exists an open interval C_α s.t. one of the following holds*

(1) $\mathbb{P}_{\theta_0}(C_\alpha^c) = \alpha$, $\partial_\theta \mathbb{P}_{\theta_0}(C_\alpha) = 0$ with

$$H_0 : \theta = \theta_0 \quad , \quad H_1 : \theta \neq \theta_0;$$

(2) $\mathbb{P}_{\theta_1}(C_\alpha^c) = \mathbb{P}_{\theta_2}(C_\alpha) = \alpha$ with

$$H_0 : \theta \in [\theta_1, \theta_2] \quad , \quad H_1 : \theta \notin [\theta_1, \theta_2].$$

Then the (deterministic) test $\Phi(x) = \mathbb{1}(x \in C_\alpha)$ for H_0, H_1 is at level α and is UMPU among all test at that level.

Proof. Not covered. □

Remark 4.4.7: Consider again the context ot the Example 4.4.3 and take the hypothesis

$$H_0 : \theta = \theta_0 \quad , \quad H_1 : \theta \neq \theta_0$$

then we see that the test induced by the symmetric confidence interval of μ with critical region

$$C_\alpha = (\theta_0 + \sigma F^{\leftarrow}(\alpha/2), \theta_0 + \sigma F^{\leftarrow}(1 - \alpha/2))$$

satisfies (1) of the previous Theorem.

5

Introduction to Bayesian Statistics

5.1 Introduction to bayesian statistic

Definition 5.1.1: A *bayesian statistical model* is a statistical model $(\mathbb{X}, \mathcal{F}, \{\mathbb{P}_\theta\}_{\theta\in\Theta}, \mathcal{T})$ where $(\Theta, \mathcal{T})$ is a measurable space and fixed $A \in \mathcal{F}$ the function $\Theta \ni \theta \mapsto \mathbb{P}_\theta(A)$ is measurable. We call a probability measure ν over $(\Theta, \mathcal{T})$ a *prior distribution* (or *a priori distribution*) of the parameter of the model.

Remark 5.1.2: In the framework of bayesian inference, in the context of the previous Definition, we consider the parameter θ also as a r.v. with values in the measurable space $(\Theta, \mathcal{T})$ with law given by the prior distribution ν.

Remark 5.1.3: The prior distribution need to be intended as the knowleadge that one has on the present situation before doing any statistical investigation.

Remark 5.1.4: Observe that the map $\Theta \times \mathcal{F} \ni (\theta, A) \mapsto \mathbb{P}(\theta, A) = \mathbb{P}_\theta(A)$ is a transition probability from $(\Theta, \mathcal{T})$ to $(\Omega, \mathcal{F})$

Remark 5.1.5: Suppose that we have a sample $X = (X_1, \dots, X_n)$ with values in a measurable space $(\mathbb{S}, \mathcal{S})$ with law $X_i \overset{\text{i.i.d.}}{\sim} P_\theta$ dependent on a parameter $\theta \in \Theta$, then we consider the statistical model $(\mathbb{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ induced by the sample X as usual, but in the context of bayesian inference we also fix a prior distribution ν on $(\Theta, \mathcal{T})$

Let us fix a bayesian statistical model $(\mathbb{X}, \mathcal{F}, \{\mathbb{P}_\theta\}_{\theta\in\Theta}, \mathcal{T})$ with prior distribution ν, where $\Theta \subset \mathbb{R}^d$ and $\mathcal{T}$ the Borel σ-algebra on Θ. Furthermore we suppose the statistical model $(\mathbb{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ to be dominated by a measure μ with densities $(p(\cdot \,|\, \theta))_{\theta\in\Theta}$ s.t. the likelihood function $L(x, \theta) = p(x \,|\, \theta)$ is $(\mathcal{F} \otimes \mathcal{T})$-measurable (assumption that guarantees the functions $\theta \mapsto p(x \,|\, \theta)$ and $\theta \mapsto \mathbb{P}(\theta, A)$ to be $\mathcal{T}$-measurables for every $x \in \mathbb{X}$ and every $A \in \mathcal{F}$ respectively) and ν to be absolutely continuous w.r.t. the Lebesgue measure on $\mathbb{R}^d$ with density π. Lastly we consider the probability measure Q on the measurable space

$(\Theta \times \mathbb{X}, \mathcal{T} \otimes \mathcal{F})$ s.t.

$$Q(T \times A) = \int_T \mathbb{P}(\theta, A)\, \mathrm{d}\nu(\theta) \quad \forall T \in \mathcal{T} \quad \forall A \in \mathcal{F}$$

that is well defined.

Definition 5.1.6: On $\Theta \times \mathbb{X}$ define the σ-algebra

$$\mathcal{F}_\Theta = \{\Theta \times A \mid A \in \mathcal{F}\}.$$

The *posterior distribution* (or *a posteriori distribution*) is the transition probability $N : \mathbb{X} \times \mathcal{T} \to [0,1]$ s.t. for every $(\mathcal{T} \otimes \mathcal{F})$-measurable and bounded r.v. $X : \Theta \times \mathbb{X} \to \mathbb{R}$ holds

$$\mathbb{E}_Q\left[X \mid \mathcal{F}_\Theta\right](\theta, x) = \int_\Theta X(\tau, x)\, N(x, \mathrm{d}\tau) \quad Q\text{-a.s. .}$$

Then the function $T \mapsto \nu^x(T) = N(x, T)$ is called *posterior distribution given* x.

Remark 5.1.7: Observe that a r.v. $X : \Theta \times \mathbb{X} \to \mathbb{R}$ is $\mathcal{F}_\Theta$-measurable if and only if $X(\theta, x) = W(x)$ for some r.v. $W : \mathbb{X} \to \mathbb{R}$ $\mathcal{F}$-measurable.

Theorem 5.1.8: *Define*

$$\begin{aligned}
&p(x) = \int_\Theta p(x \mid \theta)\, \mathrm{d}\nu(\theta) \in [0, +\infty] \quad \forall x \in \mathbb{X} \\
&G = \{x \in \mathbb{X} \mid p(x) = 0\} \\
&M = \{x \in \mathbb{X} \mid p(x) = +\infty\} \\
&p(\theta \mid x) = \begin{cases} \frac{p(x \mid \theta)\pi(\theta)}{p(x)} & \text{if } p(x) \notin G \\ 1 & \text{if } p(x) \in G \end{cases}
\end{aligned}$$

where we use the convention $\frac{a}{+\infty} = 0$. *Then*

(1) $Q(\Theta \times G) = 0$;

(2) $\mu(M) = 0$;

(3) for every $x \in M^c$ *the function* $\theta \mapsto p(\theta \mid x)$ *is a probability density function on* $(\Theta, \mathcal{T})$ *w.r.t* ν;

(4) for every $x \in M^c$ *the transition probability defined by*

$$(x, T) \mapsto N(x, T) = \int_T p(\theta \mid x)\, \mathrm{d}\nu(\theta)$$

is the posterior distribution.

Proof. (Optional)

(1) Using Fubini-Tonelli theorem we obtain

$$\begin{aligned}
Q(\Theta \times G) &= \int_\theta \mathrm{d}\nu(\theta) \int_G \mathbb{P}(\theta, \mathrm{d}x) \\
&= \int_\theta \mathrm{d}\nu(\theta) \int_G p(x \mid \theta)\, \mathrm{d}\mu(x) \\
&= \int_G \mathrm{d}\mu(x) \int_\Theta p(x \mid \theta)\, \mathrm{d}\nu(\theta) = 0
\end{aligned}$$

where the last equality follows by the definition of G.

(2) We have

$$\begin{aligned}\int_{\mathbb{X}} \mathrm{d}\mu(x)p(x) &= \int_{\mathbb{X}} \mathrm{d}\mu(x) \int_{\Theta} p(x \,|\, \theta)\, \mathrm{d}\nu(\theta) \\ &= \int_{\Theta} \mathrm{d}\nu(\theta) \int_{\mathbb{X}} p(x \,|\, \theta)\, \mathrm{d}\mu(x) \\ &= \int_{\Theta} \mathrm{d}\nu(\theta)\mathbb{P}(\theta, \mathbb{X}) = \int_{\Theta} \mathrm{d}\nu(\theta) = 1 < +\infty\end{aligned}$$

so, since $p(x)$ is non-negative follows that $p(x) < +\infty$ μ-a.s., that is $\mu(M) = 0$.

(3) We need to prove that for every $x \in M^c$ we have $p(\theta \,|\, x) \geq 0$ ν-a.s. and that $\int_{\Theta} p(\theta \,|\, x)\, \mathrm{d}\nu(\theta) = 1$. Let us fix a $x \in M^c$. The fact $p(\theta \,|\, x) \geq 0$ ν-a.s. is obvious, for the other part observe that if $x \in G^c$

$$\int_{\Theta} p(\theta \,|\, x)\, \mathrm{d}\nu(\theta) = \frac{1}{\int_{\Theta} p(x \,|\, \theta)\, \mathrm{d}\nu(\theta)} \int_{\Theta} p(x \,|\, \theta)\, \mathrm{d}\nu(\theta) = 1$$

while if $x \in G$

$$\int_{\Theta} p(\theta \,|\, x)\, \mathrm{d}\nu(\theta) = \int_{\Theta} 1\, \mathrm{d}\nu(\theta) = 1.$$

(4) Observe that (by Fubini-Tonelli theorem) the function $x \mapsto N(x, T)$ is $\mathcal{F}$-measurable for every $T \in \mathcal{T}$ and that $T \mapsto N(x, T)$ is a probability measure for every $x \in \mathbb{X}$ by point (3). So N is in fact a transition probability. Now we have to prove that, given a r.v. $X : \Theta \times \mathbb{X} \to \mathbb{R}$ $(\mathcal{T} \otimes \mathcal{F})$-measurable, setting

$$\begin{aligned}V(\theta, x) = V(x) &= \int_{\Theta} X(\theta, x) p(\theta \,|\, x)\, \mathrm{d}\nu(\theta) \\ &= \begin{cases} \int_{\Theta} X(\theta, x) \frac{1}{\int_{\Theta} p(x \,|\, \tau)\, \mathrm{d}\nu(\tau)} p(x \,|\, \theta)\, \mathrm{d}\nu(\theta) & \text{if } x \in G^c \\ \int_{\Theta} X(\theta, x)\, \mathrm{d}\nu(\theta) & \text{if } x \in G \end{cases}\end{aligned}$$

for every $B \in \mathcal{T}$ holds

$$\int_{\Theta \times B} X\, \mathrm{d}Q = \int_{\Theta \times B} V\, \mathrm{d}Q.$$

To see this observe that, by definition of Q and by (1), we have

$$\begin{aligned}\int_{\Theta \times B} X\, \mathrm{d}Q &= \int_{\Theta \times (B \cap G^c)} X\, \mathrm{d}Q \\ &= \int_{\Theta} \mathrm{d}\nu(\theta) \int_{B \cap G^c} X(\theta, x)\mathbb{P}(\theta, \mathrm{d}x) \\ &= \int_{\Theta} \mathrm{d}\nu(\theta) \int_{B \cap G^c} X(\theta, x) p(x \,|\, \theta)\, \mathrm{d}\mu(x)\end{aligned}$$

and at the same time

$$\begin{aligned}\int_{\Theta\times B} V\,\mathrm{d}Q &= \int_{\Theta\times(B\cap G^c)} V\,\mathrm{d}Q\\ &= \int_\Theta \mathrm{d}\nu(\theta)\int_{B\cap G^c} V(\theta,x)\mathbb{P}(\theta,\,\mathrm{d}x)\\ &= \int_\Theta \mathrm{d}\nu(\theta)\int_{B\cap G^c}\left[\int_\Theta X(s,x)\frac{1}{\int_\Theta p(x\,|\,\tau)\,\mathrm{d}\nu(\tau)}p(x\,|\,s)\,\mathrm{d}\nu(s)\right]p(x\,|\,\theta)\,\mathrm{d}\mu(x)\\ &= \int_\Theta \mathrm{d}\nu(s)\int_{B\cap G^c} X(s,x)p(x\,|\,s)\,\mathrm{d}\mu(x)\frac{\int_\Theta p(x\,|\,\theta)\,\mathrm{d}\nu(\theta)}{\int_\Theta p(x\,|\,\tau)\,\mathrm{d}\nu(\tau)}\\ &= \int_\Theta \mathrm{d}\nu(s)\int_{B\cap G^c} X(s,x)p(x\,|\,s)\,\mathrm{d}\mu(x).\end{aligned}$$

□

Remark 5.1.9: The posterior distribution of the parameter needs to be interpreted as our subjective belief about θ after we observe some data.

Remark 5.1.10 (Gaussian approximations of the posterior distribution): Suppose that the posterior distribution has densities $p(\theta\,|\,x)$ that are C^1 in θ and have a unique maximum in the variable θ. Observe that if we fix a $\theta'\in\Theta$, taking Taylor's expansion in θ with Lagrange's remainder we get

$$\log(p(\theta\,|\,x)) = \log(p(\theta'\,|\,x)) + (\theta-\theta')\partial_\theta\log(p(\theta'\,|\,x)) + \frac{(\xi(\theta)-\theta')^2}{2}\partial_\theta^2\log(p(\theta'\,|\,x))$$

for some $\xi(\theta)\in\Theta$ between θ and θ'. Now taking θ' the unique maximum of $p(\cdot\,|x)$ we get

$$\log(p(\theta\,|\,x)) = \log(p(\theta'\,|\,x)) - \frac{\mathcal{I}(\xi(\theta)\,|\,x)^2}{2}(\theta-\theta')^2$$

where $\mathcal{I}(\theta\,|\,x) = -\partial_\theta^2\log(p(\theta\,|\,x))$. So we have

$$p(\theta\,|\,x)\propto\exp\left(-\frac{\mathcal{I}(\xi(\theta))^2}{2}(\theta-\theta')\right)$$

that says to us that the posterior is "proportional" to a gaussian distribution with mean θ' and variance $\mathcal{I}(\theta'\,|\,x)^{-1}$ (supposing $\mathcal{I}(\theta'\,|\,x)\neq 0$).

5.2 (Bayesian) Decision theory

Fix a bayesian statistical model $(\mathbb{X},\mathcal{F},(\mathbb{P}_\theta)_{\theta\in\Theta},\mathcal{T})$, dominated by a measure μ with densities $(p(x\,|\,\theta))_{\theta\in\Theta}$, and a prior distribution ν.

Fix a measurable space $(A,\mathcal{A})$ called *set of actions* and a measurable function $\ell:\Theta\times A\to[0,\infty)$ called *loss* (or *cost*) *function*. The set A that represents the possible actions that one can do after observing a random phenomenon with law that depends on a parameter $\theta\in\Theta$, then the taken action $a\in A$ lead to a loss (or cost) $\ell(\theta,a)\geq 0$ that depends itself by $\theta\in\Theta$.

Definition 5.2.1: A *decision rule* is a measurable function $\delta:\mathbb{X}\to A$.

Remark 5.2.2: Observe that the considered family of decision rules is completely determined by the set of actions A (more specifically by the measurable space $(A, \mathcal{A})$).

Example 5.2.3: A classical estimator is a decision rule with $A = \Gamma$ (see Definition 1.1.11) and, since we observed that a test is an estimator itself, also tests can be seen as part of decision theory.

Definition 5.2.4: Given a decision rule δ we define its *risk* as

$$\mathcal{R}_\delta(\theta) = \mathbb{E}_\theta \left[\ell(\theta, \delta)\right].$$

Furthermore, in the context of bayesian inference, we define also its *bayesian risk* as

$$\mathcal{R}_\delta^b = \int_\Theta \mathcal{R}_\delta(\theta)\, \mathrm{d}\nu(\theta).$$

Remark 5.2.5: Like we did for estimators (and then for tests) we have defined the notion of risk $\mathcal{R}_\delta(\theta)$ for a decision rule δ and, in the context of classical statistics, we would try to minimize it w.r.t. δ for every $\theta \in \Theta$ (like we did for estimators and test).

In the context of bayesian statistics we have a (prior) probability distribution ν over Θ, so we do not minimize $\mathcal{R}_\delta(\theta)$ w.r.t. δ with θ fixed but instead we try to minimize its integral w.r.t. the prior ν, that is we try to minimize its bayesian risk. Mathematically means that we want to find a decision rule δ_0 s.t.

$$\mathcal{R}_{\delta_0}^b = \int_\Theta \mathcal{R}_{\delta_0}(\theta)\, \mathrm{d}\nu(\theta) \leq \int_\Theta \mathcal{R}_\delta(\theta)\, \mathrm{d}\nu(\theta) = \mathcal{R}_\delta^b \quad \text{for every decision rule } \delta.$$

Definition 5.2.6: Fix ρ a probability measure over $(\Theta, \mathcal{T})$. We call *bayesian decision* relative to ρ any $a_0 \in A$ s.t.

$$\int_\Theta \ell(\theta, a_0)\, \mathrm{d}\rho(\theta) = \inf_{a \in A} \int_\Theta \ell(\theta, a)\, \mathrm{d}\rho(\theta).$$

When a bayes decision relative to ρ exists we will indicate one of them with $d(\rho)$.

Theorem 5.2.7: *For every $x \in \mathbb{X}$ we denote ν^x the posterior distribution w.r.t x. Suppose that for every $x \in \mathbb{X}$ there is a bayes decision $d(\nu^x)$ and that the function $x \mapsto d_\nu(x) = d(\nu^x)$ is $\mathcal{F}$-measurable. Then d_ν is a decision rule and for any decision rule δ holds*

$$\int_\Theta \mathcal{R}_{d_\nu}(\theta)\, \mathrm{d}\nu(\theta) \leq \int_\Theta \mathcal{R}_\delta(\theta)\, \mathrm{d}\nu(\theta).$$

Proof. (Optional) Recall the objects and notations used in Theorem 5.1.8. Recall also that

$Q(\Theta \times G) = 0$ (Theorem 5.1.8 (1)). By the definition of Q we have

$$\begin{aligned}
\int_\Theta \mathcal{R}_\delta(\theta)\,\mathrm{d}\nu(\theta) &= \int_\theta \mathrm{d}\nu(\theta) \int_{\mathbb{X}} \mathbb{P}(\theta,\, \mathrm{d}x)\ell(\theta, \delta(x)) \\
&= \int_{\Theta\times\mathbb{X}} \ell(\theta, \delta(x))\,\mathrm{d}Q(\theta, x) \\
&= \int_{\Theta\times(\mathbb{X}\cap G^c)} \ell(\theta, \delta(x))\,\mathrm{d}Q(\theta, x) \\
&= \int_\Theta \mathrm{d}\nu(\theta) \int_{\mathbb{X}\cap G^c} \mathbb{P}(\theta,\, \mathrm{d}x)\ell(\theta, \delta(x)) \\
&= \int_\Theta \mathrm{d}\nu(\theta) \int_{\mathbb{X}\cap G^c} \mathrm{d}\mu(x) p(x\,|\,\theta)\ell(\theta, \delta(x)) \\
&= \int_{\mathbb{X}\cap G^c} \mathrm{d}\mu(x) \int_\Theta \mathrm{d}\nu(\theta) p(x\,|\,\theta)\ell(\theta, \delta(x)) \\
&= \int_{\mathbb{X}\cap G^c} \mathrm{d}\mu(x) \left(\int_\Theta p(x\,|\,\tau)\,\mathrm{d}\nu(\tau)\right) \int_\Theta \mathrm{d}\nu(\theta) \frac{p(x\,|\,\theta)}{\int_\Theta p(x\,|\,\tau)\,\mathrm{d}\nu(\tau)} \ell(\theta, \delta(x)) \\
&= \int_{\mathbb{X}\cap G^c} \mathrm{d}\mu(x) \left(\int_\Theta p(x\,|\,\tau)\,\mathrm{d}\nu(\tau)\right) \int_\Theta \mathrm{d}\nu^x(\theta)\ell(\theta, \delta(x)) \\
&\geq \int_{\mathbb{X}\cap G^c} \mathrm{d}\mu(x) \left(\int_\Theta p(x\,|\,\tau)\,\mathrm{d}\nu(\tau)\right) \int_\Theta \mathrm{d}\nu^x(\theta)\ell(\theta, d(\nu^x)) \\
&= \int_\Theta \mathcal{R}_{d_\nu}(\theta)\,\mathrm{d}\nu(\theta)
\end{aligned}$$

where the last equality can be seen doing in the opposite sense the calculations done for the previous equalities. □

Definition 5.2.8: Let $g : \Theta \to \Gamma \subset \mathbb{R}^p$ be a measurable function and take as set of actions $A = \Gamma$. An estimator $T : \mathbb{X} \to \Gamma$ of $g(\theta)$ is called *bayesian estimator* if $T(x)$ is a bayesian decision relative to ν^x for every $x \in \mathbb{X}$.

Example 5.2.9 (Posterior mean): Take $A = \mathbb{R}$, a bounded measurable function $g : \Theta \to \mathbb{R}$ and $\ell(\theta, a) = |g(\theta, a)|^2$. We want to find a bayesian estimator for $g(\theta)$.

Firstly we take a generic probability measure ρ on $(\Theta, \mathcal{T})$ and we minimize the integral

$$\int_\Theta |g(\theta) - a|^2\,\mathrm{d}\rho(\theta)$$

w.r.t. α. Denote $a_0 = \int_\Theta g(\theta)\,\mathrm{d}\rho(\theta)$ and observe that

$$\begin{aligned}
\int_\Theta |g(\theta) - a|^2\,\mathrm{d}\rho(\theta) &= \int_\Theta |(g(\theta) - a_0) + (a_0 - a)|^2\,\mathrm{d}\rho(\theta) \\
&= \int_\Theta (g(\theta) - a_0)^2\,\mathrm{d}\rho(\theta) + (a_0 - a)^2 + 2(a_0 - a)\underbrace{\int_\theta (g(\theta) - a_0)\,\mathrm{d}\rho(\theta)}_{=0} \\
&= \int_\Theta (g(\theta) - a_0)^2\,\mathrm{d}\rho(\theta) + (a_0 - a)^2 \geq \int_\Theta (g(\theta) - a_0)^2\,\mathrm{d}\rho(\theta) \quad \forall a \in A = \mathbb{R}.
\end{aligned}$$

with equality if and only if $a = a_0$, so the bayesian decision relative to ρ is $d(\rho) = a_0 =$

$\int_\Theta g(\theta)\,\mathrm{d}\rho(\theta)$ and it is unique. Hence the bayesian estimator for $g(\theta)$ is

$$T(x) = d(\nu^x) = \int_\Theta g(\theta)\,\mathrm{d}\nu^x(\theta) = \begin{cases} \int_\Theta g(\theta)\frac{p(x\,|\,\theta)}{p(x)}\,\mathrm{d}\nu(\theta) & \text{if } p(x) \neq 0 \\ \int_\Theta g(\theta)\,\mathrm{d}\nu(\theta) & \text{if } p(x) = 0. \end{cases}$$

Note that by Theorem 5.1.8 holds

$$T(x) = \mathbb{E}_Q\left[g\,|\,\mathcal{F}_\Theta\right](\theta, x) \quad Q\text{-a.s.}$$

5.3 MAP estimator

Fix a bayesian statistical model $(\mathbb{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta\in\Theta}, \mathcal{T})$, dominated by a measure μ with densities $(p(x\,|\,\theta))_{\theta\in\Theta}$, and a prior distribution ν. In the previous section we discussed bayesian decisions, concluding with the definition of bayesian estimator associated to a given loss. Now we want to analyze another particular estimator.

Definition 5.3.1: The *maximum a posteriori estimator (MAP estimator)* is the one given by

$$\hat{\theta}_{\mathrm{MAP}}(x) = \arg\max_{\theta\in\Theta} p(\theta\,|\,x) \quad \forall x \in \mathbb{X}.$$

Remark 5.3.2: Using Bayes' theorem we can write

$$\begin{aligned} \hat{\theta}_{\mathrm{MAP}}(x) &= \arg\max_{\theta\in\Theta} \frac{p(x\,|\,\theta)\pi(\theta)}{p(x)} \\ &= \arg\max_{\theta\in\Theta} p(x\,|\,\theta)\pi(\theta) \\ &= \arg\max_{\theta\in\Theta} \log(p(x\,|\,\theta)) + \log(\pi(\theta)) \end{aligned}$$

so the MAP estimator is the MLE plus the term $\log(\pi(\theta))$ that depends only on the prior distribution. In particular if we choose a prior distribution that has a very large support, then the MAP estimator will be similar to the MLE.

Example 5.3.3 (MAP, gaussian case): Take the statistical model induced by a sample $X = (X_1, \ldots, X_n)$ with $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, \sigma^2)$. So $\Theta = \mathbb{R}$ and we make the model bayesian considering $\theta \sim \mathcal{N}(\mu, \sigma_0^2)$. Let us compute the MAP estimator. We have

$$\begin{aligned} p(x\,|\,\theta) &= \prod_{j=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\theta)^2}{2\sigma^2}\right) \\ \pi(\theta) &= \frac{1}{\sqrt{2\theta\sigma_0^2}} \exp\left(-\frac{(\theta-\mu)^2}{2\sigma_0^2}\right) \end{aligned}$$

hence

$$\hat{\theta}_{\mathrm{MAP}}(x) = \arg\max_{\theta\in\Theta} \left(-\sum_{j=1}^n \frac{(x_i-\theta)^2}{2\sigma^2} - \frac{(\theta-\mu)^2}{2\sigma_0^2}\right).$$

Deriving we obtain

$$\partial_\theta \left(-\sum_{j=1}^n \frac{(x_i-\theta)^2}{2\sigma^2} - \frac{(\theta-\mu)^2}{2\sigma_0^2}\right) = \sum_{j=1}^n \frac{x_i-\theta}{\sigma^2} - \frac{\theta-\mu}{\sigma_0^2}$$

that imposed equals to 0 gives

$$\hat{\theta}_{\mathrm{MAP}}(x) = \frac{\sigma_0^2 \left(\frac{1}{n} \sum_{j=1}^n x_j \right) + \frac{\sigma^2 \mu}{n}}{\sigma_0^2 + \frac{\sigma^2}{n}}$$

but the remember that the MLE is $\hat{\theta}_{\mathrm{MLE}}(x) = \frac{1}{n} \sum_{j=1}^n x_j$, hence

$$\hat{\theta}_{\mathrm{MAP}} \overset{n \to \infty}{\to} \hat{\theta}_{\mathrm{MLE}}$$

so when n is large one can consider the MLE directly, while for small n can be useful to use the prior distribution and then the MAP estimator. Furthermore when σ_0^2 is low we are confident about the prior distribution and then we can use the MAP estimator, while for large values of σ_0^2 it is important to use the informations given by data, so in this case it can be better to use the MLE.

Example 5.3.4 (Posterior distribution, gaussian case): Take the byesian statistical model of the previous Example. We have

$$-\log(p(x \,|\, \theta)\pi(\theta)) = \sum_{j=1}^n \frac{(x_i - \theta)^2}{2\sigma^2} + \frac{(\theta - \mu)^2}{2\sigma_0^2}.$$

Imposing

$$\sum_{j=1}^n \frac{(x_i - \theta)^2}{2\sigma^2} + \frac{(\theta - \mu)^2}{2\sigma_0^2} = \frac{(\theta - \hat{\theta}_{\mathrm{MAP}})^2}{2\hat{\sigma}^2_{\mathrm{MAP}}}$$

and using the expression for $\hat{\theta}_{\mathrm{MAP}}$ found in the previous Example, we find that

$$\hat{\theta}_{\mathrm{MAP}}(x) = \frac{\sigma_0^2 \left(\frac{1}{n} \sum_{j=1}^n x_j \right) + \frac{\sigma^2 \mu}{n}}{\sigma_0^2 + \frac{\sigma^2}{n}}$$
$$\frac{1}{\hat{\sigma}^2_{\mathrm{MAP}}(x)} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}$$

so that since $p(x) = 1$ for every $x \in \mathbb{X}$ we have that the posterior distribution is $\nu^x = \mathcal{N}(\hat{\theta}_{\mathrm{MAP}}(x), \hat{\sigma}^2_{\mathrm{MAP}}(x))$.

5.4 Non-informative prior distributions

In some cases, like when we do not have any prior distribution that suit the situation or when we want to give to each parameter the same importance, a so called *non-informative* (or *flat*) *prior distribution* is used. The idea is that with such prior distribution will be only the data to determine the posterior distrbution and not a given a priori distribution.

Consider a bayesian statistical model $(\mathbb{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta \in \Theta}, \mathcal{T})$ dominated bt a measure μ with densities $(p(x \,|\, \theta))_{\theta \in \Theta}$ and fix a reference measure τ on $(\Theta, \mathcal{T})$ (it needs to be tought as the counting measure or the Lebesgue measure in the discrete and continuous cases respectively).

Definition 5.4.1: An *improper prior density* is a positive measurable function $\pi : \Theta \to [0, \infty]$ s.t. $\int_\Theta \pi(\theta) \, \mathrm{d}\tau(\theta) = +\infty$.

Improper prior densities arises when we try to create a non-informative prior. For example a natural thing that comes to mind when thinking about the idea of non-informative prior is a uniform density function, that for example on $\mathbb{R}$ with the Lebesgue measure is an improper prior density.

Definition 5.4.2: Consider $\mathcal{I}(\theta)$ the Fisher information matrix of the considered statistical model. A *Jeffrey's prior density* is a measurable function $\pi : \Theta \to [0, \infty]$ s.t.

$$\pi(\theta) \propto \sqrt{\det(\mathcal{I}(\theta))} \quad \forall \theta \in \Theta.$$

Remark 5.4.3: A Jeffrey's prior density can be an improper prior density.

Definition 5.4.4: Consider $\mathcal{I}(\theta)$ the Fisher information matrix of the considered statistical model. When $\sqrt{\det(\mathcal{I}(\theta))} \in L^1(\Theta, \tau)$, then a Jeffrey's prior can be normalized defining a probability measure on $(\Theta, \mathcal{T})$. A prior distribution defined in this way is called *Jeffrey's prior distribution.*

Remark 5.4.5: Anyway any Jeffrey's prior density defines a measure on $(\Theta, \mathcal{T})$ (not necessarily finite).

Example 5.4.6: Consider the statistical model induced by a sample $X = (X_1, \dots, X_n)$ where $X_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with $\theta = \mu \in \mathbb{R}$. We have

$$\mathcal{I}(\theta) = \frac{n}{\sigma^2}$$

so a Jeffrey's prior density is constant given by $\pi(\theta) = \sqrt{\frac{n}{\sigma^2}}$ (in particular it is improper w.r.t. the Lebesgue measure).

Remark 5.4.7 (Transformations of the Jeffrey's prior density): We want to study how a Jeffrey's prior density change w.r.t. a transformation, in particular we want to show that it gives again a Jeffrey's prior density. Suppose to have a C^1-diffeomorphism $g : \Theta \subset \mathbb{R} \to g(\Theta) \subset \mathbb{R}$ and we call $\phi = g(\theta)$. We have

$$\tilde{\pi}(\phi) = |\partial_\phi g^{-1}(\phi)| \pi(g^{-1}(\phi)) \propto |\partial_\phi g^{-1}(\phi)| \sqrt{\mathcal{I}(g^{-1}(\phi))}.$$