Information Theory and Statistics
Lecture 6: Fisher information

Łukasz Dębowski
ldebowsk@ipipan.waw.pl

Ph. D. Programme 2013/2014
Suppose we have a parametric family of distributions.

**Definition (maximum likelihood estimator)**

The *maximum likelihood estimator (MLE)* is the function defined for discrete random sample $x^n_1$ as

$$\theta_{ML}(x^n_1) := \arg\max_{\theta \in \Theta} P(X^n_1 = x^n_1 | \theta),$$

whereas for real random sample $x^n_1$ it is defined as

$$\theta_{ML}(x^n_1) := \arg\max_{\theta \in \Theta} \rho(x^n_1 | \theta).$$

Probability $P(X^n_1 = x^n_1 | \theta)$ or density $\rho(x^n_1 | \theta)$ as a function of parameter $\theta$ are called the likelihood function.
Consider a sample of length $n$ drawn from Bernoulli distribution. We have

$$0 = \left. \frac{\partial \ln P(X^n_1 = x^n_1|\theta)}{\partial \theta} \right|_{\theta = \theta_{ML}} = \frac{\sum_{i=1}^{n} x_i}{\theta_{ML}} - \frac{n - \sum_{i=1}^{n} x_i}{1 - \theta_{ML}}.$$

Hence

$$\theta_{ML}(x^n_1) = \frac{1}{n} \sum_{i=1}^{n} x_i.$$
MLE for normal distributions

Consider a sample of length \( n \) drawn from normal distribution:

\[
\ln \rho(x^n_1|\mu, \sigma) = -\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} - \frac{n}{2} \ln(2\pi\sigma^2).
\]

\[
0 = \frac{\partial \ln \rho(x^n_1|\mu, \sigma)}{\partial \mu} \bigg|_{\mu=\mu_{ML}, \sigma=\sigma_{ML}} = \sum_{i=1}^{n} \frac{2(x_i - \mu_{ML})}{2\sigma^2_{ML}},
\]

\[
0 = \frac{\partial \ln \rho(x^n_1|\mu, \sigma)}{\partial \sigma} \bigg|_{\mu=\mu_{ML}, \sigma=\sigma_{ML}} = \sum_{i=1}^{n} \frac{(x_i - \mu_{ML})^2}{\sigma^3_{ML}} - \frac{n}{\sigma_{ML}}.
\]

\[
\mu_{ML}(x^n_1) = \frac{1}{n} \sum_{i=1}^{n} x_i,
\]

\[
\sigma^2_{ML}(x^n_1) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_{ML}(x^n_1))^2.
\]
It seems plausible to accept that the deviation of a good estimator from the true parameter should converge to zero for the sample size tending to infinity.

**Definition (consistent estimator)**

Estimator $T(X^n_1)$ is called **consistent in probability** if $T(X^n_1)$ converges to $\theta$ in probability, i.e., for each $\epsilon > 0$ we have

$$\lim_{n \to \infty} P(|T(X^n_1) - \theta| > \epsilon|\theta) = 0.$$
Let us restrict ourselves to discrete random variables and put

\[ D(\theta||S) = \sum_{x \in X} P(X_i = x|\theta) \inf_{\omega \in S} \ln \frac{P(X_i = x|\theta)}{P(X_i = x|\omega)}. \]

**Theorem (Wald theorem)**

Let \( X^n \) be a random sample drawn from distribution \( P(X_i = x_i|\theta) \), where \( \theta \) takes values in the set of real numbers \( \Theta = \mathbb{R} \). Suppose that for any \( \theta \) the following conditions hold:

1. For every \( \omega \neq \theta \), there is a neighborhood \( S_\omega \) of \( \omega \) such that \( D(\theta||S_\omega) > 0 \).
2. For some constant \( a > 0 \), \( D(\theta||\{\omega : |\omega - \theta| > a\}) > 0 \).

Then the maximum likelihood estimator is consistent.
Bias and MSE

Definition (bias)

The *bias* of an estimator $T(X^n_1)$ is defined as the expectation $E_\theta T(X^n_1) - \theta$. The estimator is called *unbiased* if $E_\theta T(X^n_1) = \theta$.

Definition (mean square error)

The *mean square error* of an estimator $T(X^n_1)$ is defined as $E_\theta [T(X^n_1) - \theta]^2$. We say that estimator $T(X^n_1)$ *dominates* estimator $T'(X^n_1)$ if

$$E_\theta [T(X^n_1) - \theta]^2 \leq E_\theta [T'(X^n_1) - \theta]^2.$$
Suppose that \( \lim_{n \to \infty} E_{\theta} [T(X^n_1) - \theta]^2 = 0 \). Then the estimator \( T(X^n_1) \) is consistent in probability.

The claim follows by Markov inequality

\[
P(|T(X^n_1) - \theta| > \epsilon|\theta) \leq \frac{E_{\theta} [T(X^n_1) - \theta]^2}{\epsilon^2}.
\]
Towards Cramér-Rao theorem

**Example**

Consider a sample of length $n$ drawn from Bernoulli distribution $P(X^n = x^n_1 | \theta) = \prod_{i=1}^{n} \theta^{x_i} (1 - \theta)^{1-x_i}$. Both $X_1$ and $\bar{X}_n = \theta_{ML}(X^n_1) = n^{-1} \sum_{i=1}^{n} X_i$ are unbiased estimators of $\theta$. The variance of $X_i$ equals $\theta(1 - \theta)$, whereas the variance of $\bar{X}_n$ is $\theta(1 - \theta)/n$. Hence $\bar{X}_n$ is consistent whereas it can be checked that $X_i$ is not.

The considerations above raise the question what is the minimal variance of an estimator. The question is answered by the Cramér-Rao theorem.
Expected Fisher information

**Definition (expected Fisher information)**

For a discrete random sample, the *expected Fisher information* is defined as

\[ J_n(\theta) := E_\theta \left[ \frac{\partial}{\partial \theta} \ln P(X_1^n | \theta) \right]^2, \]

whereas, for a real random sample, we put

\[ J_n(\theta) := E_\theta \left[ \frac{\partial}{\partial \theta} \ln \rho(X_1^n | \theta) \right]^2. \]
Expected Fisher information matrix

Definition (expected Fisher information)

If the parameter \( \theta = (\theta_1, \theta_2, ..., \theta_s) \) is a vector then, for a discrete random sample, the expected Fisher information is defined as matrix

\[
(J_n(\theta))_{ij} := E_\theta \left[ \frac{\partial}{\partial \theta_i} \ln P(X_1^n|\theta) \cdot \frac{\partial}{\partial \theta_j} \ln P(X_1^n|\theta) \right].
\]

while, for a real random sample, we put

\[
(J_n(\theta))_{ij} := E_\theta \left[ \frac{\partial}{\partial \theta_i} \ln \rho(X_1^n|\theta) \cdot \frac{\partial}{\partial \theta_j} \ln \rho(X_1^n|\theta) \right].
\]

Project co-financed by the European Union within the framework of the European Social Fund
Example

Consider a sample of length \( n \) drawn from Bernoulli distribution. We have

\[
P(X_1^n = x_1^n | \theta) = \prod_{i=1}^{n} \theta^{x_i} (1 - \theta)^{1-x_i} = \theta^{\sum_{i=1}^{n} x_i} (1 - \theta)^{n-\sum_{i=1}^{n} x_i}.
\]

Hence

\[
J_1(\theta) = \theta \left[ \frac{\partial}{\partial \theta} \ln \theta \right]^2 + (1 - \theta) \left[ \frac{\partial}{\partial \theta} \ln(1 - \theta) \right]^2 = \frac{1}{\theta(1 - \theta)}.
\]
For a discrete random sample, we have

\[ E_\theta \left[ \frac{\partial}{\partial \theta} \ln P(X_1^n|\theta) \right] = 0, \]

\[ E_\theta \left[ -\frac{\partial^2}{\partial \theta^2} \ln P(X_1^n|\theta) \right] = J_n(\theta). \]

Random variable \( V = (\partial / \partial \theta) \ln P(X_1^n|\theta) \) is called score.
Proof

Let us write $L(x_1^n|\theta) = P(X_1^n = x_1^n|\theta)$.

$$
\sum_{x_1^n} L(x_1^n|\theta) \frac{\partial}{\partial \theta} \ln L(x_1^n|\theta) = \sum_{x_1^n} \frac{\partial}{\partial \theta} L(x_1^n|\theta) = \frac{\partial}{\partial \theta} \sum_{x_1^n} L(x_1^n|\theta)
= \frac{\partial}{\partial \theta} 1 = 0.
$$

$$
\sum_{x_1^n} L(x_1^n|\theta) \left[ \frac{\partial}{\partial \theta} \ln L(x_1^n|\theta) \right]^2 + \sum_{x_1^n} L(x_1^n|\theta) \frac{\partial^2}{\partial \theta^2} \ln L(x_1^n|\theta)
= \sum_{x_1^n} L(x_1^n|\theta) \left[ \frac{1}{L(x_1^n|\theta)} \frac{\partial L(x_1^n|\theta)}{\partial \theta} \right]^2 + \frac{\partial}{\partial \theta} \frac{1}{L(x_1^n|\theta)} \frac{\partial L(x_1^n|\theta)}{\partial \theta}
= \sum_{x_1^n} L(x_1^n|\theta) \frac{1}{L(x_1^n|\theta)} \frac{\partial^2 L(x_1^n|\theta)}{\partial \theta^2} = \frac{\partial^2}{\partial \theta^2} \sum_{x_1^n} L(x_1^n|\theta) = 0.
$$
Theorem

For a discrete random sample, we have

\[ J_n(\theta) = nJ_1(\theta). \]

Proof

Observe

\[
J_n(\theta) = E_{\theta} \left[ -\frac{\partial^2}{\partial \theta^2} \sum_{i=1}^{n} \ln P(X_i|\theta) \right]
\]

\[
= \sum_{i=1}^{n} E_{\theta} \left[ -\frac{\partial^2}{\partial \theta^2} \ln P(X_i|\theta) \right] = nJ_1(\theta).
\]
Theorem

For a discrete random sample, we have

\[ \frac{\partial^2}{\partial \omega^2} D(\theta || \omega) \bigg|_{\omega=\theta} = J_1(\theta). \]

Proof

Let us write \( L(x|\theta) = P(X_i = x|\theta) \). Observe that

\[ \frac{\partial^2}{\partial \omega^2} D(\theta || \omega) = \frac{\partial^2}{\partial \omega^2} \sum_{x \in X} L(x|\theta) \ln \frac{L(x|\theta)}{L(x|\omega)} \]

\[ = - \sum_{x \in X} L(x|\theta) \frac{\partial^2}{\partial \omega^2} \ln L(x|\omega), \]

which equals \( J_1(\theta) \) for \( \omega = \theta \).
The mean square error of an estimator is bounded by the inverse of Fisher information multiplied by the squared derivative of the estimator's expectation,

$$E_{\theta} [T(X_1^n) - b(\theta)]^2 \geq \frac{[b'(\theta)]^2}{J_n(\theta)},$$

where $b(\theta) = E_{\theta} T(X_1^n)$. In particular, for an unbiased estimator we have $b'(\theta) = 1$. 
Proof

Let us write the estimator \( T = T(X^n_1) \) and score \( V = (\partial / \partial \theta) \ln P(X^n_1|\theta) \). By Schwarz inequality we have

\[
(E_\theta [(V - E_\theta V)(T - E_\theta T)])^2 \leq E_\theta (V - E_\theta V)^2 E_\theta (T - E_\theta T)^2.
\]

Notice that \( E_\theta V = 0 \). Thus

\[
(E_\theta [VT])^2 \leq J_n(\theta)E_\theta (T - b(\theta))^2.
\]

\[
E_\theta [VT] = \sum_{x^n_1} P(X^n_1 = x^n_1|\theta) \frac{\partial}{\partial \theta} \ln P(X^n_1 = x^n_1|\theta) T(X^n_1)
\]

\[
= \sum_{x^n_1} \frac{\partial}{\partial \theta} P(X^n_1 = x^n_1|\theta) T(X^n_1)
\]

\[
= \frac{\partial}{\partial \theta} \sum_{x^n_1} P(X^n_1 = x^n_1|\theta) T(X^n_1) = b'(\theta).
\]
Efficient estimators

Definition (efficient estimator)

We say is that an unbiased estimator $T(X^n_1)$ is efficient when it satisfies the Cramér-Rao bound with equality, i.e., if

$$E_{\theta} [T(X^n_1) - \theta]^2 = \frac{1}{J_n(\theta)}.$$
Example (efficient estimator)

Let $X_1^n$ be a random sample from the normal distribution with expectation $\mu$ and variance $\sigma^2$. For parameter $\mu$ we obtain

$$J_n(\mu) = \int \left[ \frac{\partial}{\partial \mu} \left[ -\frac{(x - \mu)^2}{2\sigma^2} \right] \right]^2 \frac{1}{\sigma \sqrt{2\pi}} \exp \left[ -\frac{(x - \mu)^2}{2\sigma^2} \right] \, dx$$

$$= \int \frac{(x - \mu)^2}{\sigma^4} \frac{1}{\sigma \sqrt{2\pi}} \exp \left[ -\frac{(x - \mu)^2}{2\sigma^2} \right] = \frac{n}{\sigma^2}.$$ 

On the other hand estimator $\bar{X}_n = \mu_{ML}(X_1^n) = n^{-1} \sum_{i=1}^n X_i$ satisfies $E_\theta \bar{X}_n = \mu$ and

$$E_\theta [\bar{X}_n - \mu]^2 = E_\theta \bar{X}_n^2 - 2\mu E_\theta \bar{X}_n + \mu^2$$

$$= \frac{1}{n} (\sigma^2 + \mu^2) + \frac{n-1}{n} \mu^2 - 2\mu^2 + \mu^2 = \frac{\sigma^2}{n}.$$
An important property of the maximum likelihood estimator is that it is asymptotically unbiased and efficient.
Convergence in distribution

**Definition (convergence in distribution)**

We say that a series of random real variables \((Y_i)_{i=1}^{\infty}\) converges to the distribution of a random variable \(Y\), when \(\lim_{n \to \infty} P(Y_n \leq r) = P(Y \leq r)\) for every real number \(r\) at which the distribution function \(P(Y \leq r)\) is continuous.
MLE is asymptotically efficient

Let $X^n_1$ be a random sample drawn from distribution $P(X_i = x_i | \theta)$, where $\theta$ takes values in a subset $\Theta \subseteq \mathbb{R}$. Suppose:

1. Set $A = \{x \in X : P(X_i = x | \theta) > 0\}$ is independent of $\theta$.
2. For every $x \in A$, $(\partial^2 / \partial \theta^2)P(X_i = x | \theta)$ exists and is continuous in $\theta$.
3. Fisher information $J_1(\theta)$ exists and is finite.
4. For every $\theta$ in the interior of $\Theta$ there exists $\epsilon > 0$ such that
   \[
   E_{\theta} \left[ \sup_{\omega \in [\theta-\epsilon, \theta+\epsilon]} \left| \frac{\partial^2}{\partial \omega^2} \ln P(X_i = x | \omega) \right| \right] < \infty.
   \]
5. The maximum likelihood estimator $\theta_{ML}(X^n_1)$ is consistent.

Then for any $\theta$ in the interior of $\Theta$ statistic $\sqrt{n}(\theta_{ML}(X^n_1) - \theta)$ converges under $P(\cdot | \theta)$ to the normal distribution with expectation $0$ and variance $1/J_1(\theta)$. 
Bayesian statistics

So far we have considered the setting of non-Bayesian statistics, where the parameter is a fixed unknown value.

In Bayesian statistics, it is assumed that the parameter is a random variable with a distribution, called the prior distribution.

The random parameter will be denoted as $\Theta$ and the prior distribution as $\Pi(B) = P(\Theta \in B) = \int_B \pi(\theta) d\theta$, where $\pi(\theta)$ is the prior density.
When we have the prior distribution, the inference about the parameter is done using the Bayes theorem.

Putting \( P(X_1^n \in B | \Theta = \theta) = P(X_1^n \in B | \theta) \), we obtain the posterior distribution

\[
P(\Theta \in B | X_1^n \in A) = \int_B \pi(\theta | X_1^n \in A) d\theta,
\]

where the posterior density \( \pi(\theta | X_1^n \in A) \) is given by

\[
\pi(\theta | X_1^n \in A) = \frac{P(X_1^n \in A | \theta) \pi(\theta)}{\int P(X_1^n \in A | \theta') \pi(\theta') d\theta'}.
\]
Bayesian inference

Differently than in the non-Bayesian setting, in Bayesian statistics we do not need to estimate the parameter to predict the next observation. First, in many cases we can compute the marginal probability

\[ P(X_1^n \in A) = \int P(X_1^n \in A | \theta) \pi(\theta) d\theta \]

and hence we may compute

\[ P(X_{n+1} \in C | X_1^n \in A) = \frac{P(X_{n+1} \in A \times C)}{P(X_1^n \in A)}. \]

In particular, for \( X_i \) conditionally independent given \( \Theta \), we obtain

\[ P(X_{n+1} \in C | X_1^n \in A) = \frac{\int P(X_1^{n+1} \in A \times C | \theta) \pi(\theta) d\theta}{\int P(X_1^n \in A | \theta) \pi(\theta) d\theta} \]

\[ = \int P(X_{n+1} \in C | \theta) \pi(\theta | X_1^n \in A) d\theta. \]
Example

Consider a random sample of length $n$ drawn from Bernoulli distribution with the prior $\pi(\theta) = 1$. We have

$$P(X_1^n = x_1^n) = \int P(X_1^n = x_1^n | \theta) d\theta$$

$$= \int \theta^{\sum_{i=1}^{n} x_i} (1 - \theta)^{n - \sum_{i=1}^{n} x_i} d\theta$$

$$= \frac{\Gamma(\sum_{i=1}^{n} x_i + 1) \Gamma(n - \sum_{i=1}^{n} x_i + 1)}{\Gamma(n + 2)} = \frac{1}{n + 1} \left[ \binom{n}{\sum_{i=1}^{n} x_i} \right]^{-1}.$$

Hence we obtain the so called Laplace rule

$$P(X_{n+1} = 1 | X_1^n = x_1^n) = \frac{(\sum_{i=1}^{n} x_i + 1)! (n - \sum_{i=1}^{n} x_i)! (n + 1)!}{(n + 2)! (\sum_{i=1}^{n} x_i)! (n - \sum_{i=1}^{n} x_i)!}$$

$$= \frac{\sum_{i=1}^{n} x_i + 1}{n + 2}.$$
Jeffreys prior

**Definition (Jeffreys prior)**

The **Jeffreys prior** is defined as

\[
\pi_{\text{Jeffreys}}(\theta) = \frac{\sqrt{\det J_n(\theta)}}{\int \sqrt{\det J_n(\theta)} \, d\theta},
\]

where \( J_n(\theta) \) is the expected Fisher information.

If variables \( X_1, X_2, X_3, \ldots \) are independent given \( \Theta \) and have identical distribution then \( J_n(\theta) = nJ_1(\theta) \) and the Jeffreys prior does not depend on \( n \). As we will see, the inference using this prior is invariant with respect to reparametrization.
Example

Consider a random sample of length $n$ drawn from Bernoulli distribution. We have

$$
\pi_{Jeffreys}(\theta) = \frac{\theta^{-1/2}(1 - \theta)^{-1/2}}{\pi}.
$$

Hence

$$
P(X^n_1 = x^n_1) = \int P(X^n_1 = x^n_1|\theta)\pi_{Jeffreys}(\theta)d\theta
$$

$$
= \frac{1}{n\pi} \left[ \left( \sum_{i=1}^{n} x_i - \frac{1}{2} \right) \right]^{-1}
$$

and we obtain the Krichevski-Trofimov estimator

$$
P(X_{n+1} = 1|X^n_1 = x^n_1) = \frac{\sum_{i=1}^{n} x_i + 1/2}{n + 1}.
$$
The inference using Jeffreys prior does not depend on parametrization, i.e., the marginal distribution

$$P(X_1^n \in A) = \int P(X_1^n \in A | \theta) \pi_{\text{Jeffreys}}(\theta) d\theta$$

is invariant with respect to a one-to-one differentiable reparametrization.
Proof

Let $\phi = \phi(\theta)$ be a one-to-one differentiable function of $\theta$

$$
\int P(X_1^n \in A | \theta) \pi_{\text{Jeffreys}}(\theta) d\theta = \int P(X_1^n \in A | \phi) \pi(\phi) d\phi,
$$

where

$$
\pi(\phi) = \pi_{\text{Jeffreys}}(\theta) \left| \det \frac{\partial \theta_k}{\partial \phi_i} \right|
$$

$$
\propto \sqrt{\det \frac{\partial \theta_k}{\partial \phi_i} \det E_\theta \left[ \frac{\partial \ln P(X_1^n | \theta)}{\partial \theta_k} \frac{\partial \ln P(X_1^n | \theta)}{\partial \theta_l} \right] \det \frac{\partial \theta_l}{\partial \phi_j}}
$$

$$
= \sqrt{\det E_\phi \left[ \frac{\partial \ln P(X_1^n | \phi)}{\partial \phi_i} \frac{\partial \ln P(X_1^n | \phi)}{\partial \phi_j} \right]} \propto \pi_{\text{Jeffreys}}(\phi).
$$

Hence we obtain the claim.
Let \( X^n_1 \) be a random sample drawn from an \( s \)-parameter exponential family. Let \( \Theta_0 \) be a subset of \( \Theta \) such that

1. \( \Theta_0 \) is a compact subset of the interior of \( \Theta \);
2. The interior of \( \Theta_0 \) is nonempty.

Let \( \pi \) be a prior density that is continuous on \( \Theta \) and strictly positive on \( \Theta_0 \), i.e., \( \inf_{\theta \in \Theta_0} \pi(\theta) > 0 \). Finally, let a sequence \((x_i)_{i=1}^{\infty}\) be such that \( \theta_{ML}(x^n_1) \in \Theta_0 \) for sufficiently large \( n \). Then

\[
\ln P(X^n_1 = x^n_1|\theta_{ML}(x^n_1)) - \ln \int P(X^n_1 = x^n_1|\theta)\pi(\theta)d\theta
\]

\[
= \frac{s}{2} \ln \frac{n}{2\pi} + \ln \frac{\sqrt{\det J_1(\theta_{ML}(x^n_1))}}{\pi(\theta_{ML}(x^n_1))} + o(1),
\]

where the convergence is uniform in \( \Theta_0 \).
The Laplace approximation becomes particularly simple for the Jeffreys prior. Namely, we obtain

\[
\ln P(X^n_1 = x^n_1 | \theta_{ML}(X^n_1)) - \ln \int P(X^n_1 = x^n_1 | \theta) \pi_{Jeffreys}(\theta) d\theta = \frac{s}{2} \ln \frac{n}{2\pi} + \ln \int \sqrt{\det J_1(\theta)} d\theta + o(1).
\]
Proof

We give only a sketch of the proof. Write $L(x^n_1|\theta) = P(X^n_1 = x^n_1|\theta)$ and $\theta_n = \theta_{ML}(x^n_1)$. First, we observe that

$$-\frac{1}{n} \frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln L(x^n_1|\theta) = \frac{\partial^2 \ln Z(\theta)}{\partial \theta_i \partial \theta_j} = (J_1(\theta))_{ij}$$

equals the expected Fisher information matrix. Because $(\partial / \partial \theta_i) \ln L(x^n_1|\theta)$ vanishes for $\theta = \theta_{ML}(x^n_1)$, we may approximate

$$\ln L(x^n_1|\theta) \approx \ln L(x^n_1|\theta_n) - \frac{n}{2} (\theta - \theta_n)^T J_1(\theta_n)(\theta - \theta_n)$$

using the Taylor expansion.
Proof (finished)

In the following, it can be shown that the second-order Taylor approximation is so good that we obtain

\[
\int L(x_1^n | \theta) \pi(\theta) d\theta \approx L(x_1^n | \theta_n) \pi(\theta_n)
\]

\[
\times \int \exp \left[ -\frac{n}{2} (\theta - \theta_n)^T J_1(\theta_n)(\theta - \theta_n) \right] d\theta
\]

\[
= L(x_1^n | \theta_n) \pi(\theta_n) \left[ \frac{2\pi}{n} \right]^{s/2} \left[ \det J_1(\theta_n) \right]^{-1/2}.
\]

Hence the claim follows.