

# Principles of Statistics

## Lecture 1

### 1 Introduction

Consider a random variable (r.v.)  $X$  defined on some probability space:  $X : (\Omega, \mathcal{A}, \mathbb{P}) \rightarrow \mathbb{R}$ . ( $\Omega$  is the set of outcomes,  $\mathcal{A}$  is the measurable events in  $\Omega$ ,  $\mathbb{P}$  is the probability measure on  $\mathcal{A}$ ), with distribution function  $F(t) = \mathbb{P}(\omega \in \Omega : X(\omega) \leq t), t \in \mathbb{R}$ .

If  $X$  is a discrete r.v. Then

$$F(t) = \sum_{x \leq t} f(x), \text{ where } f \text{ is the probability mass function (pmf).}$$

and, if  $X$  is a continuous r.v., then

$$F(t) = \int_{-\infty}^t f(x)dx, \text{ where } f \text{ is the probability density function (pdf).}$$

We typically only write  $F(t) = \mathbb{P}(X \leq t)$ , where  $P$  is the *law* of  $X$  (i.e. the image measure  $P = \mathbb{P} \cdot X^{-1}$

**Definition 1.1.** A statistical model for the law  $P$  of  $X$  is any collection  $\{f(\cdot, \theta) : \theta \in \Theta\}$ , or  $\{P_\theta : \theta \in \Theta\}$  of pdf/pmf's or probability distributions. The index set  $\Theta$  is the parameter space.

**Example.** 1)  $N(\theta, 1), \theta \in \Theta = \mathbb{R}$ , or  $\Theta = [-1, 1]$

2)  $N(\mu, \sigma^2), (\mu, \sigma^2) = \theta \in \Theta = \mathbb{R} \times (0, \infty)$

3)  $Exp(\theta), \theta \in \Theta = (0, \infty)$

**Definition 1.2.** A statistical model  $\{P_\theta : \theta \in \Theta\}$  is correctly specified (for the law  $P$  of  $X$ ) if  $\exists \theta \in \Theta$  s.t.  $P_\theta = P$ . We often write  $\theta_0$  for this specific 'true' value of  $\theta$ . We say that observations  $X_1, \dots, X_n \stackrel{iid}{\sim}$  arise from the model  $\{P_\theta : \theta \in \Theta\}$  in this case. We refer to  $n$  as the sample size.

The tasks of statistical inference comprise at least:

1) Estimation: Construct an estimator  $\hat{\theta}_n = \hat{\theta}(X_1, \dots, X_n) \in \Theta$  that is close with high probability to  $\theta$  when  $X_1, \dots, X_n \stackrel{iid}{\sim} P_\theta, \forall \theta \in \Theta$ .

- 2) Hypothesis Testing: For  $H_0 : \theta = \theta_0$  vs  $H_1 : \theta \neq \theta_0$ , we want a test (indicator) function  $\psi_n = \psi(X_1, \dots, X_n)$  s.t  $\psi_n = 0$  with high probability when  $H_0$  is true, and  $\psi_n = 1$  otherwise.
- 3) Confidence Regions (Inference): Find regions (intervals)  $C_n = C(X_1, \dots, X_n) \subseteq \Theta$  of confidence in that  $P_\theta(\theta \in C_n) \geq 1 - \alpha, \forall \theta \in \Theta$ . This quantifies the uncertainty in the inference on  $\Theta$  by the size/diameter of  $C_n$ . Here,  $0 < \alpha < 1$  is a pre-described significance level.

## 2 Likelihood Principle

**Example.** Consider a sample  $X_1, \dots, X_n \sim^{iid} Poi(\theta)$  with (unknown)  $\theta > 0$ . If the actual observed values are  $X_1 = x_1, \dots, X_n = x_n$ , then the probability of this particular occurrence of  $x_1, \dots, x_n$  as a function of  $\theta$  is

$$\begin{aligned} f(x_1, \dots, x_n, \theta) &= P_\theta(X_1 = x_1, \dots, X_n = x_n) \\ &= \prod_{i=1}^n P_\theta(X_i = x_i) \\ &= \prod_{i=1}^n e^{-\theta} \frac{\theta^{x_i}}{x_i!} \\ &= L_n(\theta), \text{ a random function of } \theta \end{aligned}$$

**Remark.** (C.F. Gauss, R. Fisher): Maximise  $L_n(\theta)$  over  $\Theta$ , and for continuous variables, replace pmf's by pdf's.

In the example, we can equivalently maximise

$$l_n(\theta) = \log(L_n(\theta)) = -n\theta + \log \theta \sum_{i=1}^n X_i - \sum_{i=1}^n \log(X_i!).$$

over  $(0, \infty)$ . Then  $l'_n(\theta) = -n + \frac{1}{\theta} \sum_{i=1}^n X_i$ , so at maximum  $\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n X_i$

Also,  $l''_n(\theta) = -\frac{1}{\theta^2} \sum_{i=1}^n X_i < 0$  if not all  $X_i = 0$  (in which case  $\hat{\theta} = 0 = \frac{1}{n} \sum_{i=1}^n X_i$ ).

## Lecture 2

**Definition 2.1.** Given a statistical model  $\{f(\cdot, \theta) : \theta \in \Theta\}$  of pdf/pmf's for the law  $P$  of  $X$ , and given 'numerical' observations  $(x_i : i = 1, \dots, n)$  arising as iid copies  $X_i \stackrel{iid}{\sim} P$ , the *likelihood function of the model* is defined as

$$L_n : \Theta \rightarrow \mathbb{R}, L_n(\theta) = \prod_{i=1}^n f(x_i, \theta).$$

Moreover, the *log-likelihood* is

$$l_n : \Theta \rightarrow \mathbb{R} \cup \{-\infty\}, l_n(\theta) = \sum_{i=1}^n \log f(x_i, \theta),$$

and the normalised log-likelihood is

$$\bar{l}_n(\theta) = \frac{1}{n} \sum_{i=1}^n \log f(x_i, \theta).$$

We regard these functions as ('random' via the  $X_i$ 's) maps of  $\theta$ .

**Definition 2.2.** A *maximum likelihood estimator* (MLE) is any  $\hat{\theta} = \hat{\theta}_{MLE}(X_1, \dots, X_n) \in \Theta$  s.t.

$$L_n(\hat{\theta}) = \max_{\theta \in \Theta} L_n(\theta).$$

**Example.** For  $Poi(\theta), \theta \geq 0$ , we have seen  $\hat{\theta}_{MLE} = \frac{1}{n} \sum_{i=1}^n X_i$

**Example.**  $N(\mu, \sigma^2)$ , where  $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty)$ , one can show that the MLE

$$\begin{aligned} \hat{\theta}_{MLE} &= \begin{pmatrix} \hat{\mu}_{MLE} \\ \hat{\sigma}^2 \end{pmatrix} \\ &= \begin{pmatrix} \bar{X}_n \\ \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \end{pmatrix}, \end{aligned}$$

is obtained from solving  $\nabla \ln(\hat{\theta}_{MLE}) = 0$

**Remark.** Calculation of marginal MLEs that optimise only one variable is not sufficient. Typically the MLE for  $\theta \in \Theta \subseteq \mathbb{R}^p$  is found by solving the *score equations*

$$S_n(\hat{\theta}) = 0, \text{ where } S_n : \Theta \rightarrow \mathbb{R}^p \text{ is the score function } S_n(\theta) = \nabla \ln(\theta).$$

Here, we use the implicit notation  $S_n(\hat{\theta}) = \nabla \ln(\theta)|_{\theta=\hat{\theta}}$

**Remark.** The likelihood principle 'works' as soon as a joint pdf/pmf (family  $\{f(\cdot, \theta) : \theta \in \Theta\}$ ) of  $X_1, \dots, X_n$  can be specified, and does not rely on the iid assumption. For instance in the normal linear model,  $N(\chi\beta, \sigma^2 I)$ , where  $\chi$  is an  $n \times p$  matrix,  $(\beta, \sigma^2) = \theta \in \mathbb{R}^p \times (0, \infty)$ , the MLE coincides with the LS-estimator (not iid, but independent).

### 3 Information Geometry

For a r.v.  $X$  of law/distribution  $P_\theta$  on  $\chi \subseteq \mathbb{R}^d$ , and let  $g : \chi \rightarrow \mathbb{R}$  be given. We will write

$$\begin{aligned}\mathbb{E}_\theta g(X) &= \mathbb{E}_{P_\theta} g(X) \\ &= \int_\chi g(x) dP_\theta(x),\end{aligned}$$

which in the continuous case equals  $\int_\chi g(x) f(x, \theta) dx$ , and in the discrete case is  $\sum_{x \in \chi} g(x) f(x, \theta)$ .

**Observation (1).** Consider a model  $\{f(\cdot, \theta) : \theta \in \Theta\}$  for  $X$  of law  $P$  on  $\chi$ , and assume  $\mathbb{E}_P |\log f(x, \theta)| < \infty$ . Then  $\bar{l}_n(\theta) = \frac{1}{n} \sum_{i=1}^n \log f(X_i, \theta)$  as a sample approximation of

$$l(\theta) = \mathbb{E}_P \log f(X, \theta), \theta \in \Theta.$$

If the model is correctly specified with any true value  $\theta_0$  s.t.  $P = P_{\theta_0}$ , then we can rewrite

$$\begin{aligned}l(\theta) &= \mathbb{E}_{P_{\theta_0}} \log f(X, \theta) \\ &= \int_\chi (\log f(x, \theta)) f(x, \theta_0) dx.\end{aligned}$$

Next, we write

$$\begin{aligned}l(\theta) - l(\theta_0) &= \mathbb{E}_{\theta_0} \left[ \log \frac{f(X, \theta)}{f(X, \theta_0)} \right], \text{ which by Jensen's inequality applied to } \log \\ &\leq \log \mathbb{E}_{\theta_0} \left[ \frac{f(X, \theta)}{f(X, \theta_0)} \right] \\ &= \log \int_\chi \frac{f(x, \theta)}{f(x, \theta_0)} f(x, \theta_0) dx = 0, \forall \theta \in \Theta.\end{aligned}$$

Thus  $l(\theta) \leq l(\theta_0) \forall \theta \in \Theta$ , and approximately maximising  $l(\theta)$  appears sensible.

Note next that by the strict version of Jensen's inequality,  $l(\theta) = l(\theta_0)$  can only occur when  $\frac{f(X, \theta)}{f(X, \theta_0)} = \text{const}$  (in  $X$ ), which since  $\int_\chi f(x, \theta) dx = 1$  can only happen when  $f(\cdot, \theta) \stackrel{a.s.}{=} f(\cdot, \theta_0)$ .

The quantity

$$\begin{aligned}0 \leq -(l(\theta) - l(\theta_0)) &= \mathbb{E}_{\theta_0} \log \frac{f(X, \theta_0)}{f(X, \theta)} \\ &\equiv KL(P_{\theta_0}, P_\theta)\end{aligned}$$

is called the Kullback-Leibler divergence (entropy-distance), which builds the basis of statistical information theory. In particular, the differential geometry of the map  $\theta \rightarrow KL(P_{\theta_0}, P_\theta)$  determines what 'optimal' inference in a statistical model could be.

### Lecture 3

Let us say a statistical model  $\{f(\cdot, \theta) : \theta \in \Theta\}$  is regular if  $\frac{d}{d\theta} \frac{d^2}{d\theta d\theta^T} (= \nabla_\theta, \nabla_\theta \nabla_\theta^T)$  of  $f(x, \theta)$  can be interchanged with  $\int(\cdot) dx$  integration.

**Observation (2).** In a regular statistical model  $\{f(\cdot, \theta) : \theta \in \Theta\}$ , we have  $\forall \theta \in \text{int}(\Theta)$  (the interior in  $\mathbb{R}^p$ ) that

$$\begin{aligned} 0 &= \frac{d}{d\theta} 1 \\ &= \frac{d}{d\theta} \int_{\mathcal{X}} f(x, \theta) dx \\ &= \int_{\mathcal{X}} \frac{d}{d\theta} f(x, \theta) dx \\ &= \int_{\mathcal{X}} \frac{d}{d\theta} [\log f(x, \theta)] f(x, \theta) dx \\ &= \mathbb{E}_{\theta} \left[ \frac{d}{d\theta} \log f(X, \theta) \right] \end{aligned}$$

In other words, the score vector will be  $\mathbb{E}_{\theta}$  centred  $\forall \theta \in \text{int}(\Theta)$

**Definition 3.1.** Let  $\Theta \subseteq \mathbb{R}^p, \theta \in \text{int}(\Theta)$ . Then the  $p \times p$  matrix

$$I(\theta) = \mathbb{E}_{\theta} \left[ \frac{d}{d\theta} \log f(X, \theta) \frac{d}{d\theta} \log f(x, \theta)^T \right], \text{ (if it exists)}$$

is called the *Fisher information (matrix)* of the model  $\{f(\cdot, \theta) : \theta \in \Theta\}$  at  $\theta$ .

**Proposition 3.2.** In a regular statistical model  $\{f(\cdot, \theta) : \theta \in \Theta\}$ , we have  $\forall \theta \in \text{int}(\Theta), \Theta \subseteq \mathbb{R}^p, p \geq 1$ ,

$$I(\theta) = -\mathbb{E}_{\theta} \left[ \frac{d^2}{d\theta d\theta^T} \log f(X, \theta) \right]$$

*Proof.* As earlier, we write

$$\begin{aligned} 0 &= \frac{d}{d\theta d\theta^T} 1 \\ &= \frac{d^2}{d\theta d\theta^T} \int_{\mathcal{X}} f(x, \theta) dx \\ &= \int_{\mathcal{X}} \frac{d^2}{d\theta d\theta^T} f(x, \theta) dx \end{aligned} \tag{*}$$

Moreover, using the chain/product rule, we have

$$\begin{aligned} \frac{d^2}{d\theta d\theta^T} \log f(X, \theta) &= \frac{d}{d\theta^T} \left[ \frac{1}{f(X, \theta)} \frac{d}{d\theta} f(X, \theta) \right] \\ &= \frac{1}{f(X, \theta)} \frac{d^2}{d\theta d\theta^T} f(X, \theta) - \frac{1}{f^2(X, \theta)} \frac{d}{d\theta} f(X, \theta) \frac{d}{d\theta} f(X, \theta)^T \end{aligned}$$

Then, taking  $\mathbb{E}_{\theta}$  expectation, we see

$$\mathbb{E}_{\theta} \left[ \frac{d^2}{d\theta d\theta^T} \log f(X, \theta) \right] = \int_{\mathcal{X}} \frac{d^2}{d\theta d\theta^T} f(X, \theta) \frac{f(X, \theta)}{f(X, \theta)} dx - \mathbb{E}_{\theta} \left[ \frac{d}{d\theta} \log f(X, \theta) \frac{d}{d\theta} \log f(X, \theta)^T \right]$$

Hence by (\*), (3.1) holds.  $\square$

**Remark.**

1) When  $p = 1$ , the above expressions simplify and we have

$$\begin{aligned} I(\theta) &= \mathbb{E}_\theta \left( \left[ \frac{d}{d\theta} \log f(X, \theta) \right]^2 \right) \\ &= \text{Var}_\theta \left[ \frac{d}{d\theta} \log f(X, \theta) \right] \\ &= -\mathbb{E}_\theta \left[ \frac{d^2}{d\theta^2} \log f(X, \theta) \right] \end{aligned}$$

2) If  $X = (X_1, \dots, X_n)$  consists of iid copies of  $X$  so that its pdf/pmf equals  $f(x_1, \dots, x_n, \theta) = \prod_{i=1}^n f(x_i, \theta)$ , then the Fisher information *tensorises*, that is

$$\begin{aligned} I_n(\theta) &= \mathbb{E}_\theta \left[ \frac{d}{d\theta} \log f(X_1, \dots, x_n, \theta) \frac{d}{d\theta} \log f(X_1, \dots, X_n, \theta)^T \right] \\ &= \sum_{i,j=1}^n \mathbb{E}_\theta \left[ \frac{d}{d\theta} \log f(X_i, \theta) \frac{d}{d\theta} \log f(X_j, \theta)^T \right] \\ &= \sum_{i=1}^n \mathbb{E}_\theta \left[ \frac{d}{d\theta} \log f(X_i, \theta) \frac{d}{d\theta} \log f(X_i, \theta)^T \right] + \sum_{i \neq j} \mathbb{E}_\theta \left[ \frac{d}{d\theta} \log f(X_i, \theta) \right] \mathbb{E}_\theta \left[ \frac{d}{d\theta} \log f(X_j, \theta) \right] \\ &= nI(\theta) + 0 \end{aligned}$$

Where  $I(\theta)$  is the Fisher information 'per observation', i.e the Fisher info for  $\{f(x, \theta) : \theta \in \Theta, x \in \mathbb{R}\}$ .

**Proposition 3.3.** (Cramer-Rao lower bound/inequality)

Let  $X_1, \dots, X_n \stackrel{iid}{\sim}$  from a regular statistical model  $\{f(\cdot, \theta) : \theta \in \Theta, \Theta \subseteq \mathbb{R}\}$  and suppose  $\tilde{\theta} = \tilde{\theta}(X_1, \dots, X_n)$  is any unbiased estimator of  $\theta$  (i.e  $\mathbb{E}_\theta \tilde{\theta} = \theta \ \forall \theta \in \Theta$ ). Then  $\forall \theta \in \text{int}(\Theta)$ ,

$$\text{Var}_\theta \tilde{\theta} \geq \frac{1}{nI(\theta)} \ \forall n \in \mathbb{N}$$

*Proof.* Assume wlog that  $\text{Var}_\theta \tilde{\theta} < \infty$ , and consider first  $n = 1$ . recall the Cauchy-Schwarz inequality to the affect that  $\text{Cov}^2(Y, Z) \leq \text{Var}Y \ \text{Var}Z$ . For  $Y = \tilde{\theta}$  and for  $Z = \frac{d}{d\theta} \log f(X, \theta)$ . Then  $\mathbb{E}_\theta Z = 0$  by observation 2, and by the preceding remark,  $\mathbb{E}_{\theta} Z = \text{Var}_\theta Z = I(\theta)$ . Thus by Cauchy-Schwarz inequality,

$$\begin{aligned} \text{Var}(\tilde{\theta}) &\geq \frac{\text{Cov}^2(Y, Z)}{I(\theta)}, \text{ since} \\ \text{Cov}(Y, Z) &= \mathbb{E}_\theta(YZ), \text{ as } \mathbb{E}_\theta Z = 0 \\ &= \int_{\chi} \tilde{\theta}(x) \left( \frac{d}{d\theta} \log f(x, \theta) \right) f(x, \theta) dx \\ &= \int_{\chi} \tilde{\theta}(x) \frac{d}{d\theta} f(x, \theta) dx \\ &= \frac{d}{d\theta} \int_{\chi} \tilde{\theta}(x) f(x, \theta) dx \\ &= \frac{d}{d\theta} \mathbb{E}_\theta \tilde{\theta} \\ &= \frac{d}{d\theta} \theta \\ &= 1 \end{aligned}$$

For general n, replace  $Z$  by  $\frac{d}{d\theta} \log \prod_{i=1}^n f(X_i, \theta)$ , and use that

$$\mathbb{E}_\theta g(X_1, \dots, X_n) = \int_{\chi} g(x_1, \dots, x_n) \prod_{i=1}^n f(x_i, \theta) dx_1 \dots dx_n,$$

and use that Fisher information tensorises. □

Let us recall also

**Corollary 3.4.** If  $\tilde{\theta}$  is not necessarily unbiased, the proof still gives

$$\text{Var}_\theta(\tilde{\theta}) \geq \frac{(\frac{d}{d\theta} \mathbb{E}_\theta \tilde{\theta})^2}{nI(\theta)} \ \forall \theta \in \Theta, \Theta \subseteq \mathbb{R}$$

- to be called: CR-inequality for biased estimators.

## Lecture 4

A multi-dimensional version of the CRLB can be obtained from considering estimation of general differentiable functions  $\Phi : \Theta \rightarrow \mathbb{R}, \Theta \in \mathbb{R}^p$ . Then one shows that for any unbiased estimator  $\tilde{\Phi} = \tilde{\Phi}(X_1, \dots, X_n)$ , where  $X_i \stackrel{iid}{\sim} \{f(\cdot, \theta) : \theta \in \Theta\}$ , we have

$$Var_{\theta} \tilde{\Phi} \geq \frac{1}{n} \frac{\partial \Phi^T}{\partial \theta} (\theta) I(\theta)^{-1} \frac{\partial \Phi}{\partial \theta} (\theta), \forall \theta \in \text{int}(\Theta)$$

Indeed, for  $p=1$ , the proof is the same, but replacing  $\frac{d}{d\theta} \mathbb{E}_{\theta} \tilde{\theta} = \frac{d}{d\theta} \theta = 1$  by  $\frac{d}{d\theta} \mathbb{E}_{\theta} \tilde{\Phi} = \frac{d}{d\theta} \Phi(\theta)$ , and for  $p > 1$  it only needs notational adjustment.

In particular, setting  $\Phi(\theta) = \alpha^T \theta$  for any  $\alpha \in \mathbb{R}^p$ , we see that for any unbiased estimator  $\tilde{\theta}$  of  $\theta$  in  $\mathbb{R}^p$ , we also have

$$Var_{\theta}(\alpha^T \tilde{\theta}) \geq \frac{1}{n} \alpha^T I(\theta)^{-1} \alpha \quad \forall \alpha \in \mathbb{R}^p,$$

so that  $Cov_{\theta} \tilde{\theta} - \frac{1}{n} I(\theta)^{-1}$  is positive semi-definite, hence using the order structure on symmetric  $p \times p$  matrices,

$$Cov_{\theta} \tilde{\theta} > \frac{1}{n} I(\theta)^{-1}, \forall \theta \in \text{int}(\Theta)$$

**Note.** Here,  $>$  is in the sense of our order structure: for symmetric  $n \times n$  matrices  $A$  and  $B$ ,  $A > B$  if  $A - B$  is positive semi-definite.

**Example.** Consider  $X \sim N(\theta, \Sigma)$ , where  $\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \in \mathbb{R}^2, \Sigma$  is positive definite [ $n = 1$ ].

Case 1: Suppose one wants to estimate  $\theta_1$ , and  $\theta_2$  is known. Then (see example sheet), one finds the Fisher information  $I_n(\theta)$  of this one-dimensional statistical model  $\{f(\cdot, \theta_1), \theta_1 \in \mathbb{R}\}$ , with CRLB  $I_1(\theta)^{-1}$ .

Case 2: Now suppose  $\theta_2$  is unknown. Then one can compute the  $2 \times 2$  information matrix  $I_2(\theta)$ , and the CRLB for estimating  $\theta_1$  is, with  $\Phi(\theta) = \theta_1$ ,

$$\frac{\partial \Phi^T}{\partial \theta} I(\theta)^{-1} \frac{\partial \Phi}{\partial \theta}$$

One can see  $CRLB(1) < CRLB(2)$  unless  $\Sigma$  is diagonal

## 4 Asymptotic Theory for MLEs

We will investigate the large sample performance of estimators  $\tilde{\theta}(X_1, \dots, X_n)$ , specifically the MLE  $\hat{\theta}_{MLE}$ , as  $n \rightarrow \infty$ . The main goal will be to prove

$$\hat{\theta}_{MLE} \underset{n \rightarrow \infty}{\approx} N\left(\theta, \frac{1}{n} I(\theta)^{-1}\right), \forall \theta \in \Theta,$$

in a sense to be made precise later.



## Stochastic Convergence: Concepts and Facts

**Definition 4.1.** Let  $(X_n : n \in \mathbb{N})$ ,  $X$  be random vectors in  $\mathbb{R}^k$ , defined on some probability space  $(\Omega, \mathcal{A}, \mathbb{P})$

1) We say  $X_n \rightarrow X$  almost surely,  $X_n \xrightarrow{a.s.} X$  as  $n \rightarrow \infty$ , if

$$\begin{aligned} \mathbb{P}(\omega \in \Omega : \|X_n(\omega) - X(\omega)\| \rightarrow 0 \text{ as } n \rightarrow \infty) &= 1 \\ \text{i.e } \mathbb{P}(\|X_n - X\| \rightarrow 0 \text{ as } n \rightarrow \infty) &= 1 \end{aligned}$$

2) We say  $X_n \rightarrow X$  in probability,  $X_n \xrightarrow{P} X$  as  $n \rightarrow \infty$ , if  $\forall \varepsilon > 0$ ,

$$\mathbb{P}(\|X_n - X\| > \varepsilon) \rightarrow 0 \text{ as } n \rightarrow \infty$$

**Remark.** The choice of norm on  $\mathbb{R}^k$  is irrelevant. Also, one shows (see example sheet) that  $X_n \xrightarrow{a.s.P} X$  as  $n \rightarrow \infty$  is equivalent to  $X_{n_j} \xrightarrow{a.s.P} X_j$  as  $n \rightarrow \infty \forall j = 1, \dots, k$ .

**Definition 4.2.** We say  $X_n \rightarrow X$  in distribution (or in law), writing  $X_n \rightarrow^d X$  as  $n \rightarrow \infty$ , if

$$\mathbb{P}(X_n \leq t) \rightarrow \mathbb{P}(X \leq t) \forall t \in \mathbb{R}^k,$$

for which  $t \mapsto \mathbb{P}(X \leq t)$  is continuous.

Recall,  $\mathbb{P}(Z \leq t) := \mathbb{P}(Z_1 \leq t_1, \dots, Z_k \leq t_k)$ .

The following facts on stochastic convergence will be frequently used, and can be proved with measure theory.

**Proposition 4.3.** 1)  $X_n \xrightarrow[n \rightarrow \infty]{a.s.} X \Rightarrow X_n \xrightarrow[n \rightarrow \infty]{P} X \Rightarrow X_n \xrightarrow[n \rightarrow \infty]{d} X$ , but any converse is false in general.

2) (Continuous Mapping Theorem) If  $X_n, X$  take values in  $\chi \subseteq \mathbb{R}^k$ , and  $g : \chi \rightarrow \mathbb{R}^d$  is continuous, then  $X_n \xrightarrow[n \rightarrow \infty]{a.s./P/d} X$  a.s./P/d  $\Rightarrow g(X_n) \xrightarrow[n \rightarrow \infty]{a.s./P/d} g(X)$  a.s./P/d respectively

3) (Slutsky's Lemma) Suppose  $X_n \xrightarrow[n \rightarrow \infty]{d} X, Y_n \xrightarrow[n \rightarrow \infty]{d} c$ ,  $c$  constant (non-stochastic), then

$$\begin{aligned} Y_n &\xrightarrow{P} c, n \rightarrow \infty \\ X_n + Y_n &\xrightarrow{d} X + c, n \rightarrow \infty \\ X_n Y_n &\xrightarrow{d} cX \\ \frac{X_n}{Y_n} &\xrightarrow{d} \frac{X}{c}, \text{ provided } c \neq 0, \text{ as } n \rightarrow \infty \end{aligned}$$

If  $(A_n)_{ij}$  are random matrices s.t.  $(A_n)_{ij} \xrightarrow{P} A_{ij}$ , then

$$A_n X_n \xrightarrow{d} A X \text{ as } n \rightarrow \infty$$

4) If  $X_n \xrightarrow{d} X$  as  $n \rightarrow \infty$ , then  $X_n$  is stochastically bounded  $[= O_P(1)]$ , that is  $\forall \varepsilon > 0, \exists M_\varepsilon$  s.t. for all sufficiently large  $n, \mathbb{P}(\|X_n\| > M_\varepsilon) < \varepsilon$

## Lecture 5

### 5 Law of Large Numbers (LLN) and Central Limit Theorem (CLT)

Consider  $X_1, \dots, X_n \stackrel{iid}{\sim} X \sim P$  on  $\mathbb{R}^k$ . This sequence can be realised as the coordinate projections of the infinite product probability space  $(\Omega, \mathcal{A}, \mathbb{P}) = (\mathbb{R}^{\mathbb{N}}, \mathcal{B}^{\mathbb{N}}, P^{\mathbb{N}})$ , where  $P^{\mathbb{N}} = \otimes_{i=1}^{\infty} \mathbb{P}$ .

Under this product space, we can make some simultaneous statements about the stochastic behaviour of our  $X_i$ .

**Theorem 5.1.** (Weak Law of Large Numbers)

If  $\text{var}(X) < \infty$ , we have that

$$\begin{aligned} \mathbb{P} \left( \left| \frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}(X)) \right| > \varepsilon \right) &\leq \frac{\text{var}(\bar{X} - \mathbb{E}(X))}{n\varepsilon^2} \text{ by Chebyshev} \\ &= \frac{\text{var}(X)}{n\varepsilon^2} \text{ as } X_i \text{ iid} \\ &\rightarrow 0 \text{ as } n \rightarrow \infty \end{aligned}$$

**Theorem 5.2.** (Strong Law of Large Numbers)

Let  $X_1, \dots, X_n \stackrel{iid}{\sim} X \sim P$  on  $\mathbb{R}^k$  such that  $\mathbb{E}(\|X\|) < \infty$ . Then  $\bar{X} \xrightarrow{a.s.P} \mathbb{E}(X)$  as  $n \rightarrow \infty$

This is harder to prove than the Weak version, so we don't.

The stochastic fluctuations of  $\bar{X}$  about  $\mathbb{E}(X)$  are of order  $\frac{1}{\sqrt{n}}$  and as long as  $\text{var}(X) < \infty$ , *always* look normally distributed.

**Theorem 5.3.** (Univariate Central Limit Theorem)

Let  $X_1, \dots, X_n \stackrel{iid}{\sim} X \sim P$  on  $\mathbb{R}$ , with  $\text{var}(X) = \sigma^2 < \infty$ . Then

$$\sqrt{n}(\bar{X} - \mathbb{E}(X)) \xrightarrow[n \rightarrow \infty]{d} N(0, \sigma^2)$$

To give a multivariate version, we recall that  $X \in \mathbb{R}^k$  is multivariate normal if  $\forall t \in \mathbb{R}^k$ ,  $t^T X$  is univariate normal, and write  $X \sim N_{\mu}(\mu, \Sigma)$ , where  $\mu$  is our mean  $\mathbb{E}(X)$ , and  $\Sigma$  is our covariance matrix  $\text{var}(X)$ .

In fact,  $X$  is uniquely characterised as the random variable on  $\mathbb{R}^k$  such that  $t^T X \sim N(t^T \mu, t^T \Sigma t)$  for all  $t \in \mathbb{R}^k$ .

If  $\Sigma$  is invertible, then  $X$  has pdf

$$f(x) = \frac{1}{\sqrt{(2\pi)^n |\Delta \Sigma|}} \exp \left( -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right).$$

If  $A \in \mathbb{R}^{d \times k}$  and  $b \in \mathbb{R}^d$ , then

$$AX + b \sim N_d(A\mu + b, A\Sigma A^T)$$

Furthermore, if  $A_n \xrightarrow{P} A$  are random matrices, and  $X_n \xrightarrow{d} N_k(\mu, \Sigma)$ , then  $A_n X_n \xrightarrow{d} N_d(A\mu, A\Sigma A^T)$ .

**Theorem 5.4.** (Multivariate Central Limit Theorem)

Let  $X_1, \dots, X_n \stackrel{iid}{\sim} X \sim P$  on  $\mathbb{R}^k$  with  $\text{var}(X) = \Sigma$  positive definite. Then

$$\sqrt{n}(\bar{X} - \mathbb{E}(X)) \xrightarrow[n \rightarrow \infty]{d} N_k(0, \Sigma)$$

From this, and the SLLN, we can bound in probability the deviations.

**Definition 5.5.** For a sequences  $Y_1, \dots, Y_n$  and  $c_1, \dots, c_n \in \mathbb{R} \setminus \{0\}$ , we define

$$Y_n = O_P(c_n)$$

if

$$\forall \varepsilon > 0, \exists M, N > 0 \text{ s.t. } \mathbb{P}\left(\left|\frac{Y_n}{c_n}\right| > M\right) < \varepsilon \quad \forall n > N$$

(By Prokhorov's Theorem)

**Corollary 5.6.**

$$\bar{X} - \mathbb{E}(X) = O_P\left(\frac{1}{\sqrt{n}}\right)$$

**Example.** (Confidence Intervals)

Let  $X_1, \dots, X_n \stackrel{iid}{\sim} X \sim P$  on  $\mathbb{R}$  with mean  $\mu_0$  and variance  $\sigma^2$ .

Define  $\mathcal{C}_n = \{\mu \in \mathbb{R} : |\bar{X} - \mu| \leq \frac{\sigma z_\alpha}{\sqrt{n}}\}$ , where  $\mathbb{P}(|Z| \leq z_\alpha) = 1 - \alpha$  for  $Z \sim N(0, 1)$  as our 'confidence region'

$$\begin{aligned} \mathbb{P}(\mu \in \mathcal{C}_n) &= \mathbb{P}\left(|\bar{X} - \mu_0| \leq \frac{\sigma z_\alpha}{\sqrt{n}}\right) \\ &= \mathbb{P}\left(|\bar{X} - \mathbb{E}(X)| < \frac{\sigma z_\alpha}{\sqrt{n}}\right) \\ &= \mathbb{P}\left(\sqrt{n} \left| \frac{1}{n} \sum_{i=1}^n \frac{X_i - \mathbb{E}(X_i)}{\sigma} \right| \leq z_\alpha\right) \\ &\xrightarrow[n \rightarrow \infty]{} \mathbb{P}(|Z| \leq z_\alpha) \text{ by CLT} \\ &= 1 - \alpha \end{aligned}$$

where we have used the continuous mapping theorem, and because  $z_\alpha$  is a continuity point of the distribution of  $Z$ . Therefore  $\mathcal{C}_n$  is an asymptotic confidence region with confidence level (or coverage)  $1 - \alpha$ . (Alternatively of size/significance level  $\alpha$ ).

When  $\sigma$  is unknown, we can replace it in  $\mathcal{C}_n$  by  $S_n$ , where  $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ , and the same conclusion follows, using the asymptotic distribution of the t-distribution,

$$t_n := \frac{\sqrt{n}(\bar{X} - \mathbb{E}(X))}{S_n} \xrightarrow[n \rightarrow \infty]{d} N(0, 1)$$

## Lecture 6

### 6 Consistency of MLEs

**Definition 6.1.** Let  $X_1, \dots, X_n \stackrel{iid}{\sim} X$  form a statistical model  $\{P_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$ . Then we say that an estimator  $\hat{\theta}_n = \hat{\theta}(x_1, \dots, x_n)$  is consistent (for that model) if  $\hat{\theta}_n \xrightarrow[n \rightarrow \infty]{P_\theta^\mathbb{N}} \theta$  in  $(P_\theta^\mathbb{N})$ -probability,  $\forall \theta \in \Theta$ .

**Assumption A.** Suppose a statistical model  $\{f(\cdot, \theta) : \theta \in \Theta\}$ ,  $\Theta \subseteq \mathbb{R}^p$  of pdf/pmf on  $\chi \subseteq \mathbb{R}^d$  satisfies the following conditions:

1.  $f(x, \theta) > 0 \forall (x, \theta) \in (\chi, \Theta)$
2.  $\int_\chi f(x, \theta) dx = 1 \forall \theta \in \Theta$
3. The map  $\theta \mapsto f(x, \theta)$  is continuous  $\forall x \in \chi$
4.  $\Theta \subseteq \mathbb{R}^p$  is compact
5.  $\theta = \theta' \Leftrightarrow f(\cdot, \theta) = f(\cdot, \theta'), \forall \theta, \theta' \in \Theta$
6.  $\mathbb{E}_\theta \sup_{\theta' \in \Theta} |\log f(x, \theta')| < \infty$

**Remark.** 1. The above conditions justify the application of Jensen's inequality in Observation (1) of section 3 (Information Geometry) - in particular the map  $\theta \mapsto l(\theta) = \mathbb{E}_{\theta_0} \log f(X, \theta)$  is maximised uniquely at  $\theta_0 \in \Theta$ .

2. Using the dominated convergence theorem, one can integrate the limit

$$\lim_{\eta \rightarrow 0} |\log f(X, \theta + \eta) - \log f(X, \theta)| = 0$$

wrt  $P_\theta$ , and conclude that also the map  $\theta \mapsto f(\theta)$  is continuous under assumption A.

**Theorem 6.2.** Suppose that the statistical model  $\{f(\cdot, \theta) : \theta \in \Theta\}$  satisfies assumption A. Then an MLE exists, and any MLE is consistent.

*Proof.* The map  $\bar{l}_n(\theta) = \frac{1}{n} \sum_{i=1}^n \log f(X_i, \theta)$  is continuous on the compact set  $\Theta \subseteq \mathbb{R}^p$ , so by the Heine-Borel theorem,  $\bar{l}_n$  obtains a maximum on  $\Theta$ , hence a MLE  $\hat{\theta}_n$  exists.

Now, let  $\hat{\theta}_n$  be any MLE, and for a true (arbitrary) value  $\theta_0 \in \Theta$  we now prove that  $\hat{\theta}_n \xrightarrow[n \rightarrow \infty]{P} \theta_0$  (in  $P_{\theta_0}^\mathbb{N}$  probability). The idea is that maximisers  $\hat{\theta}_n$  of  $\bar{l}_n$  over  $\Theta$  should converge to the unique maximiser  $\theta_0$  of  $l$  over  $\Theta$ , since  $\bar{l}_n(\theta) \xrightarrow[n \rightarrow \infty]{P} l(\theta)$  by the LLN, for all  $\theta \in \Theta$  pointwise.

This is generally false unless one has uniform convergence.

$$(*) \sup_{\theta \in \Theta} |\bar{l}_n(\theta) - l(\theta)| \xrightarrow[n \rightarrow \infty]{P} 0 \text{ [see Ex sheet for counterexample]}$$

We show later that  $(*)$  indeed holds under the maintained hypothesis.

Define for any  $\varepsilon > 0$

$$\Theta_\varepsilon = \{\theta \in \Theta : \|\theta - \theta_0\| \geq \varepsilon\}$$

which is again a compact subset of  $\mathbb{R}^p$  (intersection of closed and compact). Thus the function  $l(\theta)$  attains its bounds on  $\Theta_\varepsilon$ , so

$$c(\varepsilon) = \sup_{\theta \in \Theta_\varepsilon} l(\theta) = l(\bar{\theta}_\varepsilon \in \Theta_\varepsilon) < l(\theta_0) \text{ since } l \text{ is maximised uniquely at } \theta_0$$

Thus we can choose  $\delta(\varepsilon)$  small enough such that

$$c(\varepsilon) + \delta(\varepsilon) < l(\theta_0) - \delta(\varepsilon) \quad (\dagger)$$

$$\text{Now, } \sup_{\theta \in \Theta_\varepsilon} \bar{l}_n = \sup_{\theta \in \Theta_\varepsilon} l(\theta) + (\bar{l}_n(\theta) - l(\theta)) \leq \sup_{\theta \in \Theta_\varepsilon} l(\theta) + \sup_{\theta \in \Theta_\varepsilon} |\bar{l}_n(\theta) - l(\theta)|$$

Next, define events (subsets of  $\mathbb{R}^N$  supporting  $(X_1, X_2, \dots)$ )

$$A_n(\varepsilon) = \{\sup_{\theta \in \Theta_\varepsilon} |\bar{l}_n(\theta) - l(\theta)| \leq \delta(\varepsilon)\}$$

On these events, we have

$$\sup_{\theta \in \Theta_\varepsilon} \bar{l}_n(\theta) \leq c(\varepsilon) + \delta(\varepsilon) < l(\theta_0) - \delta(\varepsilon) \leq \bar{l}_n(\theta_0), \text{ since on } A_n(\varepsilon) \text{ we also have } |l(\theta_0) - \bar{l}_n(\theta_0)| < \delta(\varepsilon)$$

Thus if we assume  $\hat{\theta}_n \in \Theta_\varepsilon$ , then by what precedes

$$\bar{l}_n(\hat{\theta}_n) \leq \sup_{\theta \in \Theta_\varepsilon} \bar{l}_n(\theta) < l(\theta_0) \text{ on } A_n(\varepsilon)$$

This is a contradiction to  $\hat{\theta}_n$  being a maximiser. Therefore on  $A_n(\varepsilon)$  we must have  $\hat{\theta}_n \in \Theta_\varepsilon^c$ , or in other words

$$A_n(\varepsilon) \subseteq \{\|\hat{\theta}_n - \theta_0\| < \varepsilon\}.$$

Now, by  $(*)$ ,  $\mathbb{P}(A_n(\varepsilon)) \rightarrow 1$ . We conclude that

$$\begin{aligned} \mathbb{P}(\|\hat{\theta}_n - \theta_0\| < \varepsilon) &\xrightarrow[n \rightarrow \infty]{} 1 \quad \text{or} \\ \mathbb{P}(\|\hat{\theta}_n - \theta_0\| < \varepsilon) &\xrightarrow[n \rightarrow \infty]{} 0 \end{aligned}$$

Since  $\varepsilon$  is non arbitrary,  $\hat{\theta}_n \xrightarrow[n \rightarrow \infty]{P} \theta_0$ , and the proof is complete modulo the verification of  $(*)$ .  $\square$

**Remark.** The previous proof works as well if  $(\Theta, d)$  is any compact metric space, and if continuity in assumption A is for the metric  $d$ .

## 7 Lecture 7

To verify (\*), we now make the following (non-examinable) digression:

For a (meas.)  $\chi \in \mathbb{R}^d$  and (meas.)  $h : \chi \rightarrow \mathbb{R}$ , and let  $X_1, \dots, X_n \stackrel{iid}{\sim} X$  in  $\chi$  with law  $P$ . Then the  $h(X_i)$ 's are also iid, and if  $\mathbb{E}|h(X)| < \infty$  ( $\mathbb{E} = \mathbb{E}_P$ ), then by the SLLN,

$$\frac{1}{n} \sum_{i=1}^n h(X_i) - \mathbb{E}h(X) \xrightarrow[n \rightarrow \infty]{a.s.} 0$$

Next, let  $h_1, \dots, h_N$  be a finite collection of such functions. Then

$$\mathbb{P} \left( \left| \frac{1}{n} \sum_{i=1}^n h_j(X_i) - \mathbb{E}h_j(X) \right| \xrightarrow[n \rightarrow \infty]{} 0 \right) \equiv \mathbb{P}(A_j) = 1.$$

Moreover,

$$\mathbb{P} \left( \max_{j=1, \dots, N} \left| \frac{1}{n} \sum_{i=1}^n h_j(X_i) - \mathbb{E}h_j(X) \right| \xrightarrow[n \rightarrow \infty]{} 0 \right) = \mathbb{P} \left( \bigcap_{j=1}^N A_j \right) = 1,$$

since

$$\mathbb{P} \left( \left( \bigcap_{j=1}^N A_j \right)^c \right) = \mathbb{P} \left( \bigcup_{j=1}^N A_j^c \right) \leq \sum_{j=1}^N \mathbb{P}(A_j^c) = 0.$$

To transfer to an infinite collection of h's, let us say that a family of brackets  $[\underline{h_j}, \overline{h_j}], \underline{h_j}, \overline{h_j} : \chi \rightarrow \mathbb{R}, j = 1, \dots, N$ , covers a class  $\mathcal{H}$  of maps on  $\chi$  if  $\forall h \in \mathcal{H}, \exists j$  s.t  $\underline{h_j} \leq h(x) \leq \overline{h_j} \forall x \in \chi$ .

**Proposition 7.1.** Suppose that  $\forall \varepsilon > 0$  there exist brackets  $[\underline{h_j}, \overline{h_j}], j = 1, \dots, N(\varepsilon)$  covering  $\mathcal{H}$  and such that

1.  $\mathbb{E}|\underline{h_j}(x)| < \infty, \mathbb{E}|\overline{h_j}(x)| < \infty$
2.  $\mathbb{E}|\overline{h_j}(x) - \underline{h_j}(x)| < \varepsilon$

Then

$$\sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^n h(X_i) - \mathbb{E}h(X) \right| \xrightarrow[n \rightarrow \infty]{a.s.} 0$$

*Non-examinable.* Let  $\varepsilon = \frac{1}{m}$ , where  $m \in \mathbb{N}$  is arbitrary. Then take  $N(\frac{\varepsilon}{3})$ -many brackets covering  $\mathcal{H}$ , and note that by the preceding argument, we have

$$\mathbb{P} \left( \max_{j=1, \dots, N(\frac{\varepsilon}{3})} \left| \frac{1}{n} \sum_{i=1}^n \underline{h_j}(X_i) - \mathbb{E}\underline{h_j}(X) \right| \leq \frac{\varepsilon}{3}, \forall n \geq n_o(\varepsilon) \right) = \mathbb{P}(A_\varepsilon) = 1,$$

and equivalently for  $\overline{h_j}$

Now, pick  $h \in \mathcal{H}$  arbitrary, and write for the respective bracket  $\underline{h_j}, \overline{h_j} \ni h$

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n h(X_i) - \mathbb{E}h(X) &\leq \frac{1}{n} \sum_{i=1}^n \overline{h_j}(X_i) - \mathbb{E}\overline{h_j}(X) + \mathbb{E}\overline{h_j}(X) - \mathbb{E}h(X) \\ &\leq \frac{\varepsilon}{3} + \mathbb{E}|\overline{h_j}(X) - \underline{h_j}(X)| \\ &\leq \frac{2\varepsilon}{3} \end{aligned}$$

Likewise, we get

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n h(X_i) - \mathbb{E}h(X) &\geq \frac{1}{n} \sum_{i=1}^n \underline{h_j}(X_i) - \mathbb{E}\underline{h_j}(X) + \mathbb{E}\underline{h_j}(X) - \mathbb{E}h(X) \\ &\geq -\frac{2\varepsilon}{3} \end{aligned}$$

Therefore on the event  $A = \bigcap_m A_m$  we have

$$\left| \frac{1}{n} \sum_{i=1}^n h(X_i) - \mathbb{E}h(X) \right| < \frac{2\varepsilon}{3} < \varepsilon,$$

and since

$$\mathbb{P}(A^c) \leq \sum_{m=1}^{\infty} \mathbb{P}(A_m^c) = 0,$$

the proof is complete.  $\square$

From this proposition, we deduce

**Proposition 7.2.** Let  $\chi \subseteq \mathbb{R}^d, \Theta \in \mathbb{R}^p$  compact, and suppose  $\theta \mapsto q(x, \theta)$  is continuous  $\forall x$  (and  $x$  measurable  $\forall \theta$ ), and that  $\mathbb{E} \sup_{\theta \in \Theta} |q(X, \theta)| < \infty$ . If  $X_1, \dots, X_n \stackrel{iid}{\sim} X$ , then

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^n q(X_i, \theta) - \mathbb{E}q(X, \theta) \right| \xrightarrow[n \rightarrow \infty]{a.s.}$$

**Remark.** By choosing  $q(X, \theta) = \log f(X, \theta)$ , we verify  $(*)$  in the proof of the last theorem.

**Remark.** The condition that the expectation of the supremum of  $q$  over  $\theta$  is finite can be seen to be necessary as  $\mathbb{E}\|Z\| < \infty$  in the LLN for  $Z_1, \dots, Z_n$  iid in the space  $\mathcal{C}(\Theta)$  of continuous functions on the compact space  $\Theta$ .

## 8 Asymptotic Distribution of MLEs

**Definition 8.1.** We say that an estimator  $\tilde{\theta}_n$  is asymptotically efficient in a regular statistical model  $\{f(\cdot, \theta) : \theta \in \Theta\}$  if  $\lim_{n \rightarrow \infty} n \text{Var}_{\theta}(\tilde{\theta}) = I(\theta)^{-1}, \forall \theta \in \text{int } \Theta$ .

**Assumption B.** Consider a model  $\{f(\cdot, \theta) : \theta \in \Theta\}$ ,  $\Theta \in \mathbb{R}^p$  of pdf/pmfs on  $\chi \subseteq \mathbb{R}^d$  such that  $f(x, \theta) > 0$  for all  $x \in \chi$  and all  $\theta \in \Theta$ , and such that  $\int_{\chi} f(x, \theta) dx = 1$  for every  $\theta \in \Theta$ .

Let  $\theta_0$  be a fixed ('true') value, and assume

1.  $\theta_0 \in \text{int } \Theta$
2. There exists an open set  $U$  satisfying  $\theta_0 \in U \subseteq \Theta$  such that  $f(x, \theta)$  is, for every  $x \in \chi$ , twice continuously differentiable wrt  $\theta$  in  $U$
3. the  $p \times p$  matrix  $\mathbb{E}_{\theta_0} \frac{\partial^2 \log f(X, \theta_0)}{\partial \theta \partial \theta^T}$  is non singular, and

$$\mathbb{E}_{\theta_0} \left\| \frac{\partial \log f(X, \theta_0)}{\partial \theta} \right\|^2 < \infty$$

4. There exists a compact ball  $K \subset U$  (with non-empty interior) centered at  $\theta_0$  such that

$$\mathbb{E}_{\theta_0} \left\| \frac{\partial^2 \log f(X, \theta_0)}{\partial \theta \partial \theta^T} \right\| < \infty,$$

$$\int_{\chi} \sup_{\theta \in K} \left\| \frac{\partial f(x, \theta)}{\partial \theta} \right\| dx < \infty \text{ and } \int_{\chi} \sup_{\theta \in K} \left\| \frac{\partial^2 \log f(X, \theta_0)}{\partial \theta \partial \theta^T} \right\| dx < \infty$$

5. Suppose the MLE  $\hat{\theta}_n$  in the model  $\{f(\cdot, \theta) : \theta \in \Theta\}$  based on the sample  $X_1, \dots, X_n$  exists and is consistent, i.e.  $\hat{\theta}_n \xrightarrow[n \rightarrow \infty]{P_{\theta_0}} \theta_0$ .

**Theorem 8.2.** Suppose a statistical model  $\{f(\cdot, \theta) : \theta \in \Theta\}$  is regular in the sense that it satisfies the conditions B on the handout. Then, if  $\hat{\theta}_n$  is the MLE, based on  $X_1, \dots, X_n \stackrel{iid}{\sim}$  from the model, we have  $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow[n \rightarrow \infty]{d} N(0, I(\theta)^{-1})$

**Idea** (p=1). For  $\hat{\theta}$ , we must have that for  $l_n(\theta) = \sum_{i=1}^n \log f(X_i, \theta)$

$$0 = l'_n(\hat{\theta}) = l'_n(\theta_0) + l''_n(\bar{\theta}_n)(\hat{\theta}_n - \theta_0) \text{ (MVT) },$$

so

$$\sqrt{n}(\hat{\theta} - \theta_0) = \frac{\sqrt{n}l'_n(\theta_0)}{-\frac{1}{n}l''_n(\bar{\theta}_n)} = \frac{\sqrt{n} \sum_{i=1}^n \frac{d}{d\theta} \log f(X_i, \theta) - \mathbb{E}(\cdot, 1)}{-\frac{1}{n} \sum_{i=1}^n \frac{d^2}{d\theta^2} \log f(X_i, \bar{\theta}_n)} \xrightarrow[n \rightarrow \infty]{d} \frac{N(0, I(\theta_0))}{I(\theta_0)} = N(0, I(\theta_0)^{-1})$$

## Lecture 8

**Lemma 8.3.** Observations 2 and 3 from the Information Geometry section are valid.

*Proof.* Apply the dominated convergence theorem and Assumption B. □

*Proof.* (8.2) Here,  $\mathbb{P} \equiv P_{\theta_0}^N$ ,  $\mathbb{E} \equiv \mathbb{E}_{\theta_0}$ .

In proving convergence in distribution (say  $Z_n \rightarrow^d Z$ ), it suffices to restrict to any sequence  $(E_n)$  of events (in  $\mathbb{R}^N$ ) s.t.  $\mathbb{P}(E_n) \rightarrow 1$ . Indeed,

$$|\mathbb{P}(Z_n \leq t) - \mathbb{P}(Z_n \leq t \mid E_n)| \leq \mathbb{P}(E_n^c) \rightarrow 0.$$



By consistency,  $\hat{\theta}_n \xrightarrow{P} \theta_0$ , hence the events  $E_n = \{\hat{\theta}_n \in K\}$  have probability tending to 1, and we restrict to this event in what follows. Therefore we must have

$$0 = \begin{pmatrix} \frac{\partial}{\partial \theta_1} \bar{l}_n(\hat{\theta}_n) \\ \vdots \\ \frac{\partial}{\partial \theta_p} \bar{l}_n(\hat{\theta}_n) \end{pmatrix}$$

Where we recall  $\bar{l}_n(\theta) = \frac{1}{n} \sum_{i=1}^n \log f(X_i, \theta)$ .

For any map  $h : U \rightarrow \mathbb{R}$ , we can apply the mean value theorem along the line segment  $\{t\hat{\theta}_n + (1-t)\theta_0 : 0 < t < 1\}$  connecting  $\hat{\theta}_n$  and  $\theta_0$ , and write

$$h(\hat{\theta}_n) = h(\theta_0) + \frac{\partial h^T}{\partial \theta} \big|_{\theta=\bar{\theta}} (\hat{\theta}_n - \theta_0)$$

here  $\bar{\theta} = \bar{\theta}(h)$  is some mean value on that line segment.

Second derivatives of  $\bar{l}_n(\theta)$  are differentials of the map  $u \mapsto \frac{\partial}{\partial \theta} \ln(\theta) \big|_{\theta=u}$ , and hence applying what precedes  $p$ -times to the vector entries  $\frac{\partial}{\partial \theta_j} \bar{l}_n(\hat{\theta}_k)$ , we obtain

$$\begin{aligned} 0 &= \begin{pmatrix} \frac{\partial}{\partial \theta_j} \\ \vdots \end{pmatrix} \\ &= \begin{pmatrix} \frac{\partial}{\partial \theta_j} \bar{l}_n(\theta_0) \\ \vdots \end{pmatrix} + \begin{pmatrix} \cdots & \frac{\partial^2}{\partial \theta_i \partial \theta_j} \bar{l}_n(\bar{\theta}_{ij}) & \cdots \end{pmatrix} (\hat{\theta}_n - \theta_0) \\ &:= \bar{A}_n(\hat{\theta}_n - \theta_0) \end{aligned}$$

where  $\bar{\theta}_{ij}$  is the  $p \times 1$  vector arising from the  $j^{th}$  application of the MVT.

We will show

$$\bar{A}_n \xrightarrow[n \rightarrow \infty]{P} -I(\theta_0) \tag{\dagger}$$

and in particular this implies convergence of  $\|\bar{A}_n + I(\theta_0)\| \xrightarrow{P} 0$  under the operator norm.

Hence since  $I(\theta_0)$  is non-singular, so is  $\bar{A}_n$  on events of probability converging to 1, and since we can rewrite

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = (-\bar{A}_n)^{-1} \sqrt{n} \frac{\partial}{\partial \theta} \bar{l}_n(\theta_0)$$

and the theorem follows from  $(\dagger)$ , Slutsky's Lemma, and since

$$\begin{aligned} \sqrt{n} \frac{\partial}{\partial \theta} \bar{l}_n(\theta) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left( \frac{\partial}{\partial \theta} \log f(X_i, \theta) - \mathbb{E} \frac{\partial}{\partial \theta} \log f(X_i, \theta) \right) \\ &\xrightarrow[CLT]{d} N(0, I_{\theta_0}) \text{ as } n \rightarrow \infty, \text{ applying Observation 2 to the second term above.} \end{aligned}$$

To verify (†), it suffices (see Ex. sheet) to check convergence in probability of  $\bar{A}_{n_{jk}} \rightarrow (-I(\theta_0))_{jk}$ .

Now, we write

$$\begin{aligned}\bar{A}_{n_{jk}} &= \frac{1}{n} \sum_{i=1}^n \frac{\partial^2}{\partial \theta_j \partial \theta_k} \log f(X_i, \bar{\theta}_{(j)}) - \mathbb{E} \frac{\partial^2}{\partial \theta_j \partial \theta_k} \log f(X_i, \bar{\theta}_{(j)}) \\ &\quad + \mathbb{E} \frac{\partial^2}{\partial \theta_j \partial \theta_k} \log f(X_i, \bar{\theta}_{(j)}) - \mathbb{E} \frac{\partial^2}{\partial \theta_j \partial \theta_k} \log f(X_i, \theta_0) + (-I(\theta_0))_{jk} \quad (\text{by Obs. 3})\end{aligned}$$

We denote this by (1) + (2) -  $I(\theta_0)_{jk}$ . For (1), notice that  $\bar{\theta}_{(j)} \in K$ , and hence with  $q(x, \theta) = \frac{\partial^2}{\partial \theta_j \partial \theta_k} \log f(x, \theta)$ ,

$$|(1)| = \left| \frac{1}{n} \sum q(X_i, \bar{\theta}_j) - \mathbb{E} q(X, \bar{\theta}_{(j)}) \right| \leq \sup_{\theta \in K} \left| \frac{1}{n} \sum q(X_i, \bar{\theta}_j) - \mathbb{E} q(X, \bar{\theta}_{(j)}) \right| \xrightarrow[n \rightarrow \infty]{P} 0 \text{ by the uniform LLN}$$

For (2), we notice that  $\hat{\theta}_n \xrightarrow{P} \theta_0 \Rightarrow \bar{\theta}_{(j)} \xrightarrow{P} \theta_0$  as  $n \rightarrow \infty \forall j$ , and since  $\theta \mapsto \mathbb{E} q(X, \theta)$  is continuous, the continuous mapping theorem implies that

$$(2) = \mathbb{E} q(X, \bar{\theta}_j) - \mathbb{E} q(X, \theta_0) \xrightarrow[n \rightarrow \infty]{P} 0,$$

completing the proof of (†)

□

## Lecture 9

**Remark.** 1. The assumption that  $\theta \mapsto f(x, \theta)$  is  $C^2$  can be relaxed to the existence of first derivatives (weak ones) by more involved proof methods (Le Cam Theory, see van der Vaart (1998)), including in particular the Laplace distribution (where one may show  $I_n(\theta) = n$ ). However, this cannot be weakened further, and for non-smooth parameterisation, the asymptotic theory for MLEs may be different as the example of  $U(0, \theta), \theta \in [0, \infty)$  shows (see Ex. sheet).

2. If the 'true' value  $\theta_0$  lies at the boundary of  $\Theta$ , then the MLE is also not asymptotically normal (Ex. sheet  $N(\theta, 1), \theta \in \Theta = [0, \infty)$ )

3. An asymptotic version of the Cramer-Rao lower bound can also be proved (see Le Cam theory), but it requires a restriction to 'regular' or 'uniformly consistent' (instead of unbiased) estimators to claim asymptotic efficiency. Some restriction on the class of estimators is indeed necessary, as the following example (due to Hodges) shows: Consider a statistical model  $\{P_\theta : \theta \in \Theta\}, \Theta \subseteq \mathbb{R}, 0 \in \Theta$ , s.t.  $\sqrt{n}(\hat{\theta}_{MLE} - \theta) \xrightarrow[n \rightarrow \infty]{d} N(0, I(\theta)^{-1}) \forall \theta \in \text{int } \Theta$ . [Recall that this implies that  $\sqrt{n}(\hat{\theta} - \theta)$  is stochastically bounded, i.e.,  $\forall \varepsilon > 0 \exists M_\varepsilon$  st  $\mathbb{P} \left( |\hat{\theta} - \theta| > \frac{M_\varepsilon}{\sqrt{n}} \right) < \varepsilon$ , in particular  $\hat{\theta} \xrightarrow[n \rightarrow \infty]{P} \theta$ ]

Define

$$\tilde{\theta} = \theta_{Hodges} = \begin{cases} \hat{\theta} & \text{if } |\hat{\theta}| > n^{-\frac{1}{4}} \\ 0 & \text{if } |\hat{\theta}| < n^{-\frac{1}{4}} \end{cases}$$

Now for  $\theta \neq 0$  and under  $P_\theta$ ,

$$\begin{aligned}
\mathbb{P}(\tilde{\theta} \neq \hat{\theta}) &= \mathbb{P}(|\hat{\theta}| \leq n^{-\frac{1}{4}}) \\
&= \mathbb{P}(|\hat{\theta} - \theta + \theta| \leq n^{-\frac{1}{4}}) \\
&\leq \mathbb{P}(|\hat{\theta} - \theta| \geq \theta - n^{-\frac{1}{4}}) \\
&\leq \mathbb{P}\left(|\hat{\theta} - \theta| > \frac{1}{2}|\theta|\right) \text{ for large enough } n \\
&\xrightarrow[n \rightarrow \infty]{} 0 \text{ since } \hat{\theta} \xrightarrow{P} \theta, |\theta| \neq 0
\end{aligned}$$

So for such  $\theta$  we thus have  $\sqrt{n}(\tilde{\theta} - \theta) \xrightarrow[n \rightarrow \infty]{d} N(0, I(\theta)^{-1})$ .

Next, if  $\theta = 0$ , we have under  $P_0$

$$\begin{aligned}
\mathbb{P}(\tilde{\theta} \neq 0) &= \mathbb{P}(|\hat{\theta}| \geq n^{-\frac{1}{4}}) \\
&= \mathbb{P}(|\hat{\theta} - \theta| > n^{-\frac{1}{4}}) \\
&= \mathbb{P}(\sqrt{n}|\hat{\theta} - \theta| > n^{\frac{1}{4}})
\end{aligned}$$

So for any  $\varepsilon > 0$  and  $n$  s.t.  $n^{\frac{1}{4}} > M_\varepsilon$ , we have by stochastic boundedness of  $\sqrt{n}(\hat{\theta} - \theta)$  that the last probability is less than  $\varepsilon$ . Hence we conclude that under  $P_0$ ,  $\sqrt{n}(\tilde{\theta} - \theta) \xrightarrow[n \rightarrow \infty]{d} N(0, 0)$ , and so  $\tilde{\theta}$  'beats' the asymptotic efficiency bound  $I(\theta)^{-1}$  at  $\theta = 0$ .

## 9 Plug-in MLEs and the Delta method

Consider estimating a functional  $\Phi : \Theta \rightarrow \mathbb{R}^k$ ,  $\Theta \subseteq \mathbb{R}^p$  based on  $X_i \stackrel{iid}{\sim} f(\cdot, \theta) : \theta \in \Theta$ , where  $\hat{\theta}$  is the MLE for  $\theta$ . One can show that a MLE in the model  $\{f(\cdot, \phi) : \phi = \Phi(\theta) \text{ for some } \theta \in \Theta\}$

The asymptotic normality and efficiency of  $\hat{\theta}$  then implies the same for  $\Phi(\hat{\theta})$  as long as  $\Phi$  is differentiable.

### **Theorem 9.1.** (Delta-method)

Suppose  $\Phi : \Theta \rightarrow \mathbb{R}^k$  is continuously differentiable at  $\theta \in \Theta$  with gradient vector  $\frac{\partial \Phi}{\partial \theta}(\theta)$ . Suppose further  $\hat{\theta}_n$  are random vectors in  $\Theta$  s.t.  $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n \rightarrow \infty]{d} Z$ , where  $Z$  is some random vector in  $\mathbb{R}^p$ . Then  $\sqrt{n}(\Phi(\hat{\theta}_n) - \Phi(\theta)) \xrightarrow[n \rightarrow \infty]{d} \frac{\partial \Phi}{\partial \theta}(\theta)^T Z$

*Proof.* By the mean value theorem applied to  $\Phi$  on the line segment  $\{t\hat{\theta}_n + (1-t)\theta : 0 < t < 1\}$  we can write for mean value  $\bar{\theta}$

$$\sqrt{n}(\Phi(\hat{\theta}_n) - \Phi(\theta)) = \frac{\partial \Phi}{\partial \theta}(\bar{\theta})^T (\hat{\theta}_n - \theta)$$

Since  $\sqrt{n}(\hat{\theta} - \theta) \rightarrow^d Z$ , we have in particular  $\hat{\theta} \xrightarrow[n \rightarrow \infty]{P} \theta$  (by stochastic boundedness), so also  $\bar{\theta} \xrightarrow[n \rightarrow \infty]{P} \theta$ , and hence by the continuous mapping theorem, we also have  $\frac{\partial \Phi}{\partial \theta}(\bar{\theta}_n) \xrightarrow[n \rightarrow \infty]{P} \frac{\partial \Phi}{\partial \theta}(\theta)$ . Hence by Slutsky's Lemma,  $\sqrt{n}(\Phi(\hat{\theta}) - \Phi(\theta)) \xrightarrow[n \rightarrow \infty]{d} \frac{\partial \Phi}{\partial \theta}(\theta)^T Z$

□

**Remark.** If  $\sqrt{n}(\hat{\theta}_{MLE} - \theta) \rightarrow^d N(0, I(\theta)^{-1})$ , then what precedes implies that the plug-in MLE satisfies  $\sqrt{n}(\Phi(\hat{\theta}_{MLE}) - \Phi(\theta)) \rightarrow^d N(0, \frac{\partial \Phi}{\partial \theta}(\theta)^T I(\theta)^{-1} \frac{\partial \Phi}{\partial \theta}(\theta))$ , in particular the asymptotic covariance attains the CRLB for estimating  $\Phi(\theta)$

## 10 Asymptotic inference with the MLE

Suppose we want to make inference on  $\theta_i$ , the  $i^{th}$  component of  $\theta \in \mathbb{R}^p$ , from a regular statistical model  $\{f(\cdot, \theta) : \theta \in \Theta\}$ . Then  $\theta_i = e_i^T \theta$ , and by the last theorem,

$$\sqrt{n}(\hat{\theta}_i - \theta_i) = \sqrt{n}e_i^T(\hat{\theta} - \theta) \xrightarrow[n \rightarrow \infty]{d} N(0, e_i^T I(\theta)^{-1} e_i) = N(0, I(\theta)_{ii}^{-1})$$

This suggests an asymptotic confidence interval (CI)  $C_n = \{v \in \mathbb{R} : |\hat{\theta}_i - v| \leq \frac{(I(\theta)_{ii}^{-1})^{\frac{1}{2}}}{\sqrt{n}} z_\alpha\}$

### Lecture 10

Indeed by the continuous mapping theorem,

$$\begin{aligned} \mathbb{P}(\theta_j \in C_n) &= \mathbb{P}\left(\sqrt{n}(I(\theta)^{-1})_{jj}^{-\frac{1}{2}} |\hat{\theta}_{n,j} - \theta_j| \leq z_\alpha\right) \\ &\xrightarrow[n \rightarrow \infty]{} \mathbb{P}(|Z| \leq z_\alpha) \\ &= 1 - \alpha \end{aligned}$$

So  $C_n$  is a confidence interval of asymptotic level  $1 - \alpha$ .

In practice,  $I(\theta)$  may still depend on  $\theta$  and hence needs to be replaced by a consistent estimate  $\hat{i}_n \xrightarrow[n \rightarrow \infty]{P} I(\theta)$  (in which case, by Slutsky's lemma, the new CI again has asymptotic coverage level  $1 - \alpha$ ).

**Definition 10.1.** The  $p \times p$  matrix

$$i_n(\theta) = \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta} \log f(X_i, \theta) \frac{\partial}{\partial \theta} \log f(X_i, \theta)^T$$

is called the observed Fisher information (at  $\theta$ ). We then defined  $\hat{i}_n = i_n(\hat{\theta}_{MLE})$ , an estimator of  $I(\theta_0)$ .

**Proposition 10.2.** Suppose the statistical model  $\{f(\cdot, \theta) : \theta \in \Theta\}$  satisfies Assumption B. Then  $\hat{i}_n \xrightarrow[n \rightarrow \infty]{P} I(\theta_0)$ .

*Proof.* Just as when proving  $\bar{A}_n \xrightarrow[n \rightarrow \infty]{P} I(\theta_0)$  in the proof of asymptotic normality of  $\hat{\theta}_{MLE}$ , replacing  $\frac{\partial^2}{\partial \theta \partial \theta^T}, \bar{\theta}_{ij}$  by  $\frac{\partial}{\partial \theta}$  and  $\hat{\theta}_{MLE}$  respectively.  $\square$

**Remark.** The continuous mapping theorem and invertability of  $I(\theta_0)$  then also imply that  $\hat{i}_n^{-1} \xrightarrow[n \rightarrow \infty]{P} I(\theta_0)^{-1}$ , since  $A \mapsto A^{-1}$  is continuous on  $\{A : \det A \neq 0\}$ .

Alternatively, one uses  $j_n(\theta) = -\frac{1}{n} \sum_{i=1}^n \frac{\partial^2}{\partial \theta \partial \theta^T} \log f(X_i, \theta)$ , and estimates  $I_n(\theta_0)$  by  $\hat{j}_n = j_n(\hat{\theta}_{MLE})$ , which as before (and by Observation 3 from information geometry) satisfies again  $\hat{j}_n \xrightarrow[n \rightarrow \infty]{P} I(\theta_0)$ .

To make inference on the entire parameter  $\theta \in \Theta \subseteq \mathbb{R}^p$ , one can use the Wald-Statistic

$$W_n(\theta) = n(\hat{\theta}_n - \theta)^T \hat{i}_n(\hat{\theta}_n - \theta), \theta \in \Theta,$$

with  $\hat{i}_n$  possibly replaced by  $i_n(\theta)$ . One shows (Ex sheet), that under  $P_\theta$  then

$$W_n(\theta) \xrightarrow[n \rightarrow \infty]{d} \chi_p^2 \text{ (Chi-squared with } p \text{ degrees of freedom),}$$

and this entails that the confidence ellipsoid  $C_n = \{\theta \in \mathbb{R}^p : W_n(\theta) \leq \xi_\alpha\}$  has asymptotic coverage  $\lim_{n \rightarrow \infty} P_\theta(\theta \in C_n) = 1 - \alpha$  if  $\xi_\alpha$  are the  $1 - \alpha$  quantiles of the  $\chi_p^2$  distribution.

Consider next a hypothesis testing problem

$$H_0 : \theta \in \Theta_0 \subset \Theta \text{ vs } H_1 : \theta \in \Theta \setminus \Theta_0$$

We wish to construct a test  $\Psi_n = \Psi(X_1, \dots, X_n)$  which takes value 0 to indicate  $H_0$  is true, and takes value 1 otherwise (to indicate  $H_1$  is true). The type-I error of any such test is, for  $\theta \in \Theta_0$ ,

$$P_\theta(\text{reject } H_0) = E_\theta \Psi_n,$$

and the type-II error, for  $\theta \in \Theta \setminus \Theta_0$  is

$$P_\theta(\text{accept } H_0) = E_\theta(1 - \Psi_n)$$

**Definition 10.3.** A general purpose test can be constructed from the Likelihood ratio test statistic

$$\begin{aligned} \Lambda_n(\Theta, \Theta_0) &= 2 \log \frac{\prod_{i=1}^n f(X_i, \hat{\theta}_{MLE})}{\prod_{i=1}^n f(X_i, \hat{\theta})} \\ &= 2 \log \frac{\sup_{\theta \in \Theta} \prod_{i=1}^n f(X_i, \theta)}{\sup_{\theta \in \Theta_0} \prod_{i=1}^n f(X_i, \theta)} \end{aligned}$$

**Theorem 10.4** (Wilks'). In a statistical model  $\{f(\cdot, \theta) : \theta \in \Theta\}$  satisfying Assumption B, and for  $\Theta_0 = \{\theta_0\}$ , for  $\theta_0 \in \Theta$ , we have under  $P_{\theta_0}$ ,

$$\Lambda_n(\Theta, \Theta_0) \xrightarrow[n \rightarrow \infty]{d} \chi_{p, p}^2 = \dim \Theta$$

**Remark.** 1. One can show more generally that for  $\dim(\Theta_0) = p_0 > 0$  but  $p_0 < p$ , we have

$$\Lambda_n(\Theta, \Theta_0) \xrightarrow[n \rightarrow \infty]{d} \chi_{p-p_0}^2 \text{ under } P_{\theta}, \theta \in \Theta_0$$

2. We can construct a test  $\Psi_n = 1_{\{\Lambda_n(\Theta, \Theta_0) > \xi_{\alpha, p-p_0}\}}$  for  $H_0$ , where type-I errors are controlled at asymptotic level  $\alpha$  if  $\xi_{\alpha}$  are the  $\alpha$ -quantiles of  $\chi_{p-p_0}^2$ -distribution.

*Proof.* We restrict to events  $\hat{\theta}_n \in \text{int } \Theta$  (of probability approaching 1). Then since  $\hat{\theta}_{n,0} - \theta_0$ , we can write

$$\begin{aligned} \Lambda_n(\Theta, \Theta_0) &= 2 \log(\hat{\theta}_n) - 2 \log(\theta_0) \\ &= (-2 \log(\theta_0)) - (-2 \log(\hat{\theta}_n)) \\ &= -2 \frac{\partial}{\partial \theta} \log(\hat{\theta}_n) - \frac{2}{2} (\theta_0 - \hat{\theta}_n)^T \frac{\partial^2}{\partial \theta \partial \theta} \log(\bar{\theta}_n) (\theta_0 - \hat{\theta}_n) \text{ (Taylor expansion)} \\ &= 0 - (\theta_0 - \hat{\theta}_n)^T \frac{\partial^2}{\partial \theta \partial \theta} \log(\bar{\theta}_n) (\theta_0 - \hat{\theta}_n) \text{ as } \hat{\theta}_n \in \text{int } \Theta, \text{ as the gradient at maximiser must vanish,} \end{aligned}$$

where  $\bar{\theta}_n$  are mean values lying on the line segment connecting  $\hat{\theta}_n, \theta_0$ . The second order term can be written

$$\sqrt{n}(\hat{\theta}_n - \theta_0)(j_n(\theta) - I(\theta_0))\sqrt{n}(\hat{\theta}_n - \theta_0) + \sqrt{n}(\hat{\theta}_n - \theta_0)(I(\theta_0))\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow[n \rightarrow \infty]{d} Z^T I(\theta_0) Z$$

by liberal usage of Slutsky's lemma, the previous proposition, and the continuous mapping theorem for the map  $x \mapsto x^T I(\theta_0) x$  from  $\mathbb{R}^p \rightarrow \mathbb{R}$ . Moreover, by standard linear algebra,  $Z^T I_{\theta_0} Z = \sum_{i=1}^p W_i^2$ ,  $W_i \stackrel{iid}{\sim} N$ , so  $\sim \chi_p^2$   $\square$

## Lecture 11

### 11 Bayesian Inference

For a given statistical model  $\{f(\cdot, \theta) : \theta \in \Theta\}$ ,  $\Theta \subseteq \mathbb{R}^p$ , we will now regard  $\theta$  as drawn at random from some *prior* distribution  $\pi$  on  $\Theta$ . This may

- i) Model an intrinsically random state of nature  $\Theta$ .
- ii) Model subjective beliefs about the state of nature  $\Theta$ .
- iii) Serve as a way to generate statistical decision rules/estimators in our inference problem.

**Example.** Consider a countable set of (scientific) hypotheses  $H_i, i \in \Theta$ , about the state of nature, each of prior probability  $\pi_i$ ,  $\sum_{i \in \Theta} \pi_i = 1$ , and such that

$$\mathbb{P}(X = x \mid H_i) = f_i(x),$$

where  $x$  is a random outcome that can be measured. Then by the Bayes rule for conditional probabilities,

$$\mathbb{P}(H_i \mid X = x) = \frac{f_i(x)\pi_i}{\sum_{j \in \Theta} f_j(x)\pi_j}$$

To check whether  $H_i$  is more likely than  $H_j$  given  $X = x$ , we compare

$$\frac{\mathbb{P}(H_i \mid X = x)}{\mathbb{P}(H_j \mid X = x)} = \frac{f_i(x)\pi_i}{f_j(x)\pi_j}$$

If all  $\pi_i$  agree ( $\Theta$  is finite), then this just reduces to the likelihood ratio test. In a general setting,  $\{f(\cdot, \theta) : \theta \in \Theta\}$ , we wish to model the observation  $X \mid \theta \sim f(\cdot, \theta)$  and  $\theta \sim \pi$  on  $\Theta$ , where  $\pi$  is the prior distribution. The *posterior distribution* is then the conditional distribution of  $\theta \mid X$ . To make this rigorous, consider a sample space  $\chi \subseteq \mathbb{R}^d$  supporting  $\{f(\cdot, \theta) : \theta \in \Theta \subseteq \mathbb{R}^p\}$ , and on the product space  $\chi \times \Theta (\subseteq \mathbb{R}^d \times \mathbb{R}^p)$  consider a probability distribution  $Q$  with pmf  $f$

$$dQ(x, \theta) = f(x, \theta)\pi(\theta)dx d\theta.$$

By the usual rules for conditional densities, if  $(X, \Theta) \sim Q$ , then

$$X \mid \Theta \sim \frac{f(x, \theta)\pi(\theta)}{\int_{\chi} f(x, \theta)\pi(\theta)dx} = f(x, \theta)$$

Likewise,

$$\theta \mid X \sim \frac{f(x, \theta)\pi(\theta)}{\int_{\Theta} f(x, \theta)\pi(\theta)d\theta} = \pi(\theta \mid X)$$

is the pdf/pmf of the posterior distribution. If  $X_1, \dots, X_n$  are i.i.d copies of  $X \mid \theta$  then the same argument gives that the posterior distribution is given by

$$\theta \mid (X_1, \dots, X_n) \sim \frac{\prod_{i=1}^n f(X_i, \theta)\pi(\theta)}{\int_{\Theta} \prod_{i=1}^n f(X_i, \theta)\pi(\theta)d\theta} = \pi(\theta \mid X_1, \dots, X_n)$$

**Example.** Consider a  $N(\theta, 1)$  model with prior  $\pi \sim N(0, 1)$  on  $\Theta = \mathbb{R}$ . Given  $X_1, \dots, X_n$  i.i.d copies of  $X \mid \theta$ , we see that

$$\begin{aligned} \pi(\theta \mid X_1, \dots, X_n) &\propto e^{\frac{1}{2} \sum_{i=1}^n (X_i - \theta)^2} e^{-\frac{\theta^2}{2}} \\ &= e^{-\frac{1}{2} \sum_{i=1}^n X_i^2 + \sum_{i=1}^n X_i \theta - \frac{n\theta^2}{2} - \frac{\theta^2}{2}} \\ &\propto e^{n\bar{X}\theta - \frac{n+1}{2}\theta^2} \\ &\propto e^{-\frac{1}{2} \left( \frac{n}{\sqrt{n+1}} \bar{X} - \sqrt{n+1}\theta \right)^2} \\ &= e^{-\frac{n+1}{2} \left( \frac{n}{n+1} \bar{X} - \theta \right)^2} \end{aligned}$$

And so  $\pi(\theta \mid X_1, \dots, X_n) \sim N(\frac{1}{n+1} \sum_{i=1}^n X_i, \frac{1}{n+1})$ .

One shows more generally that for normal prior and normal 'sampling' models  $\{f(\cdot, \theta) : \theta \in \Theta\}$ , the posterior distribution is again a normal distribution. This is an example of a *conjugate prior* where the posterior distribution after sampling belongs to the same family of probability distributions.

**Example.** We have some other examples of conjugate priors.

- i) Beta prior + Binomial sampling  $\rightarrow$  Beta posterior
- ii) Gamma prior + Poisson sampling  $\rightarrow$  Gamma posterior

Even when  $\pi$  is not a proper probability distribution, the expression

$$\pi(\theta|X_1, \dots, X_n) = \frac{\prod_{i=1}^n f(X_i, \theta)\pi(\theta)}{\int_{\Theta} \prod_{i=1}^n f(X_i, \theta)\pi(\theta)d\theta}$$

may still be well defined, in which case we speak of a posterior distribution arising from an 'improper' prior. Specifically, the family of 'Jeffrey's' priors which are such that  $\pi(\theta) \propto \sqrt{\det(I(\theta))}$  often fall into this class. For instance, for the  $N(\theta, \sigma^2)$  model (with  $\sigma^2$  known), one sees that the Jeffrey's prior is proportionally constant, and one shows that the 'improper' posterior is equal to a  $N(\bar{X}_n, \frac{\sigma^2}{n}) = N(\hat{\theta}_{MLE}, \frac{\sigma^2}{n})$  distribution (see ex. sheet). Note however (see ex. sheet) that uniform priors do not necessarily return 'Bayes estimators' that coincide with MLEs, as the  $\text{Bin}(n, p)$  model with  $p \sim U(0, 1)$  prior shows.

## 11.1 Statistical Inference with Posterior Distributions

The posterior distribution  $\pi(\cdot, X_1, \dots, X_n)$  is a (random) probability distribution on  $\Theta$ , and hence can be used in principle to construct inference procedures for  $\theta$ .

- i) Estimation - One may use the posterior mean  $\mathbb{E}^\pi[\theta|X_1, \dots, X_n]$  as an estimator  $\bar{\theta}_n = \bar{\theta}(X_1, \dots, X_n)$  for  $\theta$ , or alternatively (when appropriately defined) the posterior mode or median.
- ii) Uncertainty quantification - Any subset  $C_n \subseteq \Theta$  for which  $\pi(C_n|X_1, \dots, X_n) = 1 - \alpha$  is a level  $1 - \alpha$  *credible* set (but it has, a fortiori, no interpretation in terms of coverage probabilities  $P_\theta(\theta \in C_n)$ )
- iii) Hypothesis testing - Given  $\Theta_0, \Theta_1 \subseteq \Theta$ , we can compute Bayes-factors

$$\frac{\pi(\Theta|X_1, \dots, X_n)}{\pi(\Theta_1|X_1, \dots, X_n)} = \frac{\int_{\Theta_0} \prod_{i=1}^n f(X_i, \theta)\pi(\theta)d\theta}{\int_{\Theta_1} \prod_{i=1}^n f(X_i, \theta)\pi(\theta)d\theta} = \frac{\mathbb{P}(X_1, \dots, X_n|\Theta_0)}{\mathbb{P}(X_1, \dots, X_n|\Theta_1)}$$

So we may 'test' for (choose to prefer)  $H_0$  if  $\phi_n 1_{\{\text{Bayes factor} < 1\}}$

## 11.2 Frequentist Analysis of Bayes Methods

Bayesian inference procedures  $\bar{\theta}(X_1, \dots, X_n), C(X_1, \dots, X_n), \Psi(X_1, \dots, X_n)$  can be analysed as statistical algorithms in their own right under the *frequentist* sampling assumption that  $X_i \stackrel{iid}{\sim} f(\cdot, \theta_0), \theta_0 \in \Theta$ .



**Example.**  $X_1, \dots, X_n \stackrel{iid}{\sim}$  copies of  $X|\theta \sim N(\theta, 1)$  with  $\theta \sim N(0, 1)$  prior. Then the posterior is

$$\theta|X_1, \dots, X_n \sim N\left(\frac{1}{n+1} \sum_{i=1}^n X_i, \frac{1}{n+1}\right)$$

One shows easily that  $\bar{\theta}_n = \frac{1}{n+1} \sum_{i=1}^n X_i \xrightarrow{a.s.} \theta_0$  under  $P_{\theta_0}^{\mathbb{N}}$ , and also that  $\sqrt{n}(\bar{\theta}_n - \theta) \rightarrow^d N(0, I(\theta_0)^{-1})$  under  $P_{\theta_0}^{\mathbb{N}}$ . To corroborate Bayesian credible sets, however, more is required, as these are based not on the 'limit distribution'  $N(0, I(\theta_0)^{-1})$ , but on  $\pi(\cdot, X_1, \dots, X_n)$ .

**Theorem 11.1.** (Bernstein-von Mises)

Suppose a statistical model  $\{f(\cdot, \theta) : \theta \in \Theta\}$  satisfies assumption B, and let the prior have a continuous and positive density  $\pi$  near  $\theta_0$ . Denote by  $\pi_n = \pi(\cdot, X_1, \dots, X_n)$ , and let  $\hat{\phi}_n$  be the pdf of a  $N(\hat{\theta}_{MLE}, \frac{1}{n}I(\theta_0)^{-1})$ . Then

$$\|\pi_n - \hat{\phi}_n\|_{L^1} = \int_{\mathbb{R}} |\pi_n - \hat{\phi}_n(\theta)| d\theta \xrightarrow[n \rightarrow \infty]{a.s.} 0$$

*Proof.* The general proof requires LeCam theory, so we only prove  $X|\theta \sim N(\theta, 1)$  with  $\theta \sim N(0, 1)$ , in which case  $I(\theta) = 1$  and  $\hat{\theta}_{MLE} = \bar{X}_n$ . Recall  $\pi_n$  is the pdf of a  $N(\bar{\theta})$  distribution where  $\bar{\theta} = \frac{n}{n+1}$ , and so

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta) = -\sqrt{n} \frac{1}{n+1} (\bar{X}_n - \theta_0 + \theta_0) \xrightarrow[n \rightarrow \infty]{P} 0 \text{ by SLLN.}$$

Since  $\int_{\mathbb{R}} \pi_n - \hat{\phi}_n(\theta) d\theta = 1 - 1 = 0$ , we have

$$\begin{aligned} \int_{\mathbb{R}} |\pi_n(\theta) - \hat{\phi}_n(\theta)| d\theta &= 2 \int_{\mathbb{R}} \left( \pi_n(\theta) - \hat{\phi}_n(\theta) \right)^+ d\theta \\ &= 2 \int_{\mathbb{R}} \left( 1 - \frac{\pi_n(\theta)}{\hat{\phi}_n(\theta)} \right)^+ \hat{\phi}_n(\theta) d\theta \\ &= 2 \int_{\mathbb{R}} \left( 1 - \frac{\sqrt{\frac{n+1}{2\pi}} \exp\left(-\frac{n+1}{2}(\theta - \hat{\theta} + \hat{\theta} - \hat{\theta})^2\right)}{\sqrt{\frac{n}{2\pi}} \exp\left(-\frac{n}{2}(\theta - \bar{\theta})^2\right)} \right)^+ \sqrt{\frac{n}{2\pi}} \exp\left(-\frac{n}{2}(\theta - \hat{\theta})^2\right) d\theta \\ &= 2 \int_{\mathbb{R}} \left( 1 - \sqrt{\frac{n+1}{n}} \frac{\exp\left(-\frac{n+1}{2n}(v + \sqrt{n}(\hat{\theta} - \theta))^2\right)}{\exp\left(-\frac{1}{2}v^2\right)} \right)^+ \frac{1}{\sqrt{2\pi}} e^{-\frac{v^2}{2}} dv \end{aligned}$$

Where we substituted  $v = \sqrt{n}(\theta - \hat{\theta})$ . Fixing  $\omega \in \Omega_0 \subseteq \Omega$  such that  $\mathbb{P}(\Omega_0) = 1$ , so  $\sqrt{n}(\hat{\theta}(\omega) - \bar{\theta}(\omega)) \xrightarrow[n \rightarrow \infty]{} 0$  (as scalars), note that for  $Z \geq 0$ ,  $(1 - Z)^+ \in [0, 1]$ , so that by integrability of  $e^{-\frac{v^2}{2}}$  on  $\mathbb{R}$ , an application of the dominated convergence theorem implies that the whole last integral tends to 0  $\forall \omega \in \Omega_0$ . Since  $\mathbb{P}(\Omega_0) = 1$ , the limit holds almost surely.  $\square$