

# Part II — Principles of Statistics

Based on lectures by R. Nickl

Notes taken by Joseph Tedds using Dexter Chua's header and Gilles Castel's snippets.

Michaelmas 2019

These notes are not endorsed by the lecturers, and I have modified them (often significantly) after lectures. They are nowhere near accurate representations of what was actually lectured, and in particular, all errors are almost surely mine.

## **The Likelihood Principle**

Basic inferential principles. Likelihood and score functions, Fisher information, Cramer-Rao lower bound, review of multivariate normal distribution. Maximum likelihood estimators and their asymptotic properties: stochastic convergence concepts, consistency, efficiency, asymptotic normality. Wald, score and likelihood ratio tests, confidence sets, Wilks theorem, profile likelihood. Examples. [8]

## **Bayesian Inference**

Prior and posterior distributions. Conjugate families, improper priors, predictive distributions. Asymptotic theory for posterior distributions. Point estimation, credible regions, hypothesis testing and Bayes factors [3]

## **Decision Theory**

Basic elements of a decision problem, including loss and risk functions. Decision rules, admissibility, minimax and Bayes rules. Finite decision problems, risk set. Stein estimator. [3]

## **Multivariate Analysis**

Correlation coefficient and distribution of its sample version in a bivariate normal population. Partial correlation coefficients. Classification problems, linear discriminant analysis. Principal component analysis. [5]

## **Nonparametric Inference and Monte Carlo Techniques**

GlivenkoCantelli theorem, KolmogorovSmirnov tests and confidence bands. Bootstrap methods: jackknife, roots (pivots), parametric and nonparametric bootstrap. Monte Carlo simulation and the Gibbs sampler. [4]

## Contents

<b>0</b>	<b>Introduction</b>	<b>3</b>
<b>1</b>	<b>Likelihood Principle</b>	<b>4</b>
<b>2</b>	<b>Information geometry</b>	<b>6</b>
<b>3</b>	<b>Asymptotic theory for MLEs</b>	<b>10</b>
3.1	Stochastic convergence: concepts and facts . . . . .	10
3.2	Law of large numbers and central limit theorem . . . . .	11
<b>4</b>	<b>Consistency of MLEs</b>	<b>13</b>
<b>5</b>	<b>Asymptotic distribution of MLEs</b>	<b>17</b>
<b>6</b>	<b>Plug-in MLEs and the Delta-method</b>	<b>21</b>
<b>7</b>	<b>Asymptotic inference with the MLE</b>	<b>22</b>
<b>8</b>	<b>Bayesian Inference</b>	<b>25</b>
8.1	Statistical inference with posterior distributions . . . . .	26
8.2	Frequentist analysis of Bayes methods . . . . .	27
<b>9</b>	<b>Decision theory</b>	<b>30</b>
9.1	Minimax risk . . . . .	31
9.2	Admissibility . . . . .	33
9.3	Classification problems . . . . .	37

## 0 Introduction

Consider a random variable  $X$  defined on some probability space,

$$X : (\Omega, A, P) \mapsto \mathbb{R}.$$

We call  $\Omega$  the set of outcomes,  $A$  is the set of measurable events in  $\Omega$  and  $P$  is our probability measure on  $A$ . with distribution function

$$F(t) = P(\omega \in \Omega : X(\omega) \leq t), \quad t \in \mathbb{R}.$$

If  $X$  is a discrete random variable, then

$$F(t) = \sum_{x \leq t} f(x).$$

where  $f$  is the probability mass function (pmf) and if  $X$  is a continuous random variable, then

$$F(t) = \int_{-\infty}^t f(x) dx.$$

where  $f$  is the probability density function (pdf).

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

**Definition** (Statistical model). A *statistical model* for the law  $P$  of  $X$  is any collection

$$\{f(\theta) : \theta \in \Theta\}, \text{ or } \{P_\theta : \theta \in \Theta\}.$$

of pdf/pmf's or probability distributions. The index set  $\Theta$  is the parameter space

**Example.** (i)  $N(0, 1), \theta \in \Theta = \mathbb{R}$ , or  $\Theta = [-1, 1]$

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

(iii)  $\text{Exp}(\theta), \dots$

**Definition** (Correctly specified). A statistical model  $\{P_\theta : \theta \in \Theta\}$  is *correctly specified* (for the law  $P$  of  $X$ ) if  $\exists \theta \in \Theta$  such that  $P_\theta = P$ . We often write  $\theta_0$  for this specific 'true' value of  $\theta$ . We say that observations  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} P_\theta$  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:

- (i) 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  $x_1, \dots, x_n \stackrel{\text{iid}}{\sim} P_\theta, \forall \theta \in \Theta$ .
- (ii) Hypothesis testing - For  $H_0 : \theta = \theta_0$  vs  $H_1 : \theta \neq \theta_0$ , we want a test (indicator) function  $\psi_n = \psi(x_1, \dots, x_n)$  such that  $\psi_n = 0$  with high probability when  $H_0$  is true, and  $\psi_n = 1$  otherwise.
- (iii) Confidence regions (inference) - Find regions (intervals)  $C_n = C(x_1, \dots, x_n, \alpha) \subseteq \Theta$  of confidence in that

$$P_\theta(\theta \in C_n) \stackrel{(\geq)}{=} 1 - \alpha, \quad \forall \theta \in \Theta.$$

This quantifies the uncertainty in the inference on  $\theta$  by the size (diameter) of  $C_n$ . Here  $0 < \alpha < 1$  is a pre-scribed significance level.

# 1 Likelihood Principle

**Example.** Consider a sample  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Poisson}(\theta)$  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

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

a random function of  $\theta$ .

**Idea** Maximise  $L_n(\theta)$  over  $\Theta$ , and for continuous variables, replace pmf's by pdf's. In the example above, we can equivalently maximise

$$\ell_n(\theta) = \log L_n(\theta) = -n\theta + \log \theta \sum_{i=1}^n X_i - \sum_{i=1}^n \log(x_i!) \text{ over } (0, \infty).$$

Then

$$\ell'_n(\theta) = -n + \frac{1}{\theta} \sum_{i=1}^n X_i \stackrel{\text{FOC}}{=} 0 \iff \hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

Also,

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

**Definition** (Likelihood function). 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  $X_i \stackrel{\text{iid}}{P}$ , the *likelihood function of the model* is defined on

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

Moreover, the *log-likelihood* function is

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

and the *normalised log-likelihood function*

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

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

**Definition** (Maximum likelihood estimator). A *maximum likelihood estimator* (MLE) is any  $\hat{\theta} = \hat{\theta}_{\text{MLE}}(X_1, \dots, X_n) \in \Theta$  such that

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

Equivalently,  $\hat{\theta}$  maximises  $\ell_n$  or  $\bar{\ell}_n$  over  $\Theta$ .

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

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

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

is obtained from simultaneously solving  $\frac{\partial}{\partial \mu} \ell_n(\theta) = \frac{\partial}{\partial \sigma^2} \ell_n(\theta) = 0$

**Remark.** Calculation of 'marginal' MLE's 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*

$$S_n(\hat{\theta}) = 0, \text{ where } S_n : \Theta \mapsto \mathbb{R}^p$$

is the score function

$$S_n(\theta) = \nabla \ell_n(\theta) = \left( \frac{\partial}{\partial \theta_1} \ell_n(\theta), \dots, \frac{\partial}{\partial \theta_p} \ell_n(\theta) \right).$$

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

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

## 2 Information geometry

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

$$\mathbb{E}_\theta [g(X)] = \mathbb{E}_{P_\theta} [g(X)] = \int_\chi g(x) dP_\theta(x)$$

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

Observation 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  $\ell_n(\theta) = \frac{1}{n} \sum_{i=1}^n \log f(x_i, \theta)$  as a sample approximation of

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

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

$$\ell(\theta) = \mathbb{E}_{P_{\theta_0}} [\log f(X, \theta)] = \int_\chi (\log f(x, \theta) f(x, \theta_0)) dx.$$

Next we write

$$\begin{aligned} \ell(\theta) - \ell(\theta_0) &= \mathbb{E}_{\theta_0} \left[ \log \frac{f(X, \theta)}{f(X, \theta_0)} \right] \\ &\stackrel{(\text{Jensen})}{\leq} \log \mathbb{E}_{\theta_0} \left[ \frac{f(X, \theta)}{f(X, \theta_0)} \right] \\ &= \log \int_\chi \frac{f(X, \theta)}{f(X, \theta_0)} f(X, \theta_0) dx \\ &= \log \int_\chi f(x, \theta) dx = 0 \quad \forall \theta \in \Theta \end{aligned}$$

Thus  $\ell(\theta) \leq \ell(\theta_0) \quad \forall \theta \in \Theta$ , and approximately maximising  $\ell(\theta)$  appears sensible. Note next that by the strict version of Jensen's inequality,  $\ell(\theta) = \ell(\theta_0)$  can only occur when  $\frac{f(X, \theta)}{f(X, \theta_0)} = \text{constant}$  (in  $X$ ), which since  $\int_\chi f(x, \theta) dx = 1$  can only happen when  $f(\cdot, \theta) \stackrel{\text{almost surely}}{=} f(\cdot, \theta_0)$  identically.

**Definition** (Identifiable). Let us thus say that the model is *identifiable* if  $f(\cdot, \theta) = f(\cdot, \theta_0)$  (a.s.)  $\iff \theta = \theta_0$ . In this case, the function  $\ell(\theta)$  has a unique maximiser at the true value  $\theta_0$ .

The quantity

$$0 \leq -(\ell(\theta) - \ell(\theta_0)) = \mathbb{E}_{\theta_0} \left[ \log \frac{f(X, \theta_0)}{f(X, \theta)} \right] \equiv \text{KL}(P_{\theta_0}, P_\theta).$$

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

**Definition** (Regular). Let us say that a statistical model  $\{f(\cdot, \theta) : \theta \in \Theta\}$  is *regular* if

$$\frac{\partial}{\partial \theta}, \frac{\partial^2}{\partial \theta \partial \theta^T} = (\nabla_\theta, \nabla_\theta \nabla_\theta^T$$

of  $f(x, \theta)$  can be interchanged with  $\int(\cdot)dx$  integration.

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

$$0 = \frac{\partial}{\partial \theta} 1 = \frac{\partial}{\partial \theta} \int_{\chi} f(\cdot, \theta) dx = \int_{\chi} \frac{\partial}{\partial \theta} [\log f(x, \theta)] f(x, \theta) = \mathbb{E}_\theta \left[ \frac{\partial}{\partial \theta} \log f(X, \theta) \right].$$

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

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

$$I(\theta) = \mathbb{E}_\theta \left[ \frac{\partial}{\partial \theta} \log f(x, \theta) \frac{\partial}{\partial \theta} \log f(x, \theta)^T \right]$$

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

One shows:

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

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

*Proof.* As earlier we write

$$0 = \frac{\partial^2}{\partial \theta \partial \theta^T} 1 = \frac{\partial^2}{\partial \theta \partial \theta^T} \int_{\chi} f(x, \theta) dx = \int_{\chi} \frac{\partial^2}{\partial \theta \partial \theta^T} f(x, \theta) dx \quad (1)$$

Moreover, using the chain product rules, we have

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

Then taking  $\mathbb{E}_\theta$  - expectations and using (1) we see

$$\mathbb{E}_\Theta \left[ \frac{\partial^2}{\partial \theta \partial \theta^T} \log f(X, \theta) \right] = \int_{\chi} \frac{\partial^2}{\partial \theta \partial \theta^T} f(x, \theta) \frac{f(x, \theta)}{f(x, \theta)} dx - \mathbb{E}_\theta \left[ \frac{\partial}{\partial \theta} \log f(X, \theta) \frac{\partial}{\partial \theta} \log f(X, \theta)^T \right].$$

□

**Remark.** (i) When  $p = 1$  the above expressions simplify and we have

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

(ii) 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

$$\begin{aligned} I_n(\theta) &= \mathbb{E}_\theta \left[ \frac{\partial}{\partial \theta} \log f(x_1, \dots, x_n; \theta) \frac{\partial}{\partial \theta} \log f(x_1, \dots, x_n; \theta)^T \right] \\ &= \sum_{i,h=1}^n \mathbb{E}_\theta \left[ \frac{\partial}{\partial \theta} \log f(x_i, \theta) \frac{\partial}{\partial \theta} \log f(x_h, \theta)^T \right] \\ &= \sum_{i=1}^n \mathbb{E}_\theta \left[ \frac{\partial}{\partial \theta} \log f(X_i, \theta) \frac{\partial}{\partial \theta} \log f(X_i, \theta)^T \right] + \sum_{i \neq j} \mathbb{E}_\theta \left[ \frac{\partial}{\partial \theta} \log f(X_i, \theta) \right] \mathbb{E}_\theta \left[ \frac{\partial}{\partial \theta} \log f(X_j, \theta) \right] \\ &= nI_1(\theta) \end{aligned}$$

$I_1(\theta) = I(\theta)$  is the Fisher information 'per observation' i.e. the Fisher information for  $\{f(\cdot, \theta) : \theta \in \Theta\}, x \in \mathbb{R}$ .

**Proposition.** (Cramer-Rao lower bound). Let  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim}$  form 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 (i.e.  $\mathbb{E}_\theta [\tilde{\theta}] = \theta \forall \theta \in \Theta$ ). Then  $\forall \theta \in \text{int}\Theta$

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

*Proof.* Assume wlog  $\text{var}_\theta \tilde{\theta} < \infty$ , and consider first  $n = 1$ . Recall the Cauchy-Schwarz inequality to the effect 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 our observation above and by the preceding remarks,  $\mathbb{E}_\theta [Z] = \text{var}_\theta Z = I(\theta)$ . Thus by the Cauchy-Schwarz inequality.

$$\text{var}(\tilde{\theta}) \geq \frac{\text{Cov}^2(Y, Z)}{I(\theta)} = \frac{1}{I(\theta)}.$$

Since

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



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

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

and use that the Fisher information tensorises.  $\square$

Let us record also

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

$$\text{var}_\theta(\tilde{\theta}) \geq \frac{\left( \frac{d}{d\theta} \mathbb{E}_\theta [\tilde{\theta}] \right)^2}{nI(\theta)} \quad \forall \theta \in \text{int}\Theta, \Theta \in \mathbb{R}.$$

to be called the Cramer-Rao inequality for biased estimators.

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

$$\text{var}_\theta(\tilde{\Phi}) \geq \frac{1}{n} \frac{\partial \Phi^T}{\partial \theta}(\theta) \Phi(\theta)^{-1} \frac{\partial \Phi}{\partial \theta}(\theta) \quad \forall \theta \in \text{int}\Theta.$$

[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

$$\frac{d}{d\theta} \mathbb{E}_\theta [\tilde{\Phi}(\theta)] = \frac{d}{d\theta} \Phi(\theta)$$

and for  $p \geq 1$  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

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

so that

$$\text{cov}_\theta(\tilde{\theta}) - \frac{1}{n} I(\theta)^{-1}$$

is positive semi-definite, hence using the order structure on symmetric  $p \times p$  matrices

$$\text{cov}_\theta(\tilde{\theta}) \geq \frac{1}{n} I(\theta)^{-1}, \quad \forall \theta \in \text{int}\Theta.$$

**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 I Suppose one wants to estimate  $\theta_1$  and  $\theta_2$  is known. Then (see example sheet) one finds the Fisher information  $I_1(\theta_1)$  of this one-dimensional statistical model  $\{f(\cdot, \theta_1) : \theta_1 \in \mathbb{R}\}$  with CRLB  $I_1(\theta_1)^{-1}$ . Case II Now suppose that  $\theta_2$  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 (I)  $\leq$  CRLB (II) unless  $\Sigma$  is diagonal.

### 3 Asymptotic theory for MLEs

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

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

in a sense to be made precise.

#### 3.1 Stochastic convergence: concepts and facts

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

- (i) We say  $X_n \rightarrow X$  *almost surely*,  $X_n \xrightarrow{\text{a.s.}} X$  as  $n \rightarrow \infty$  if

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

$$(\mathbb{P}(\|X_n - X\| \rightarrow 0 \text{ as } n \rightarrow \infty) = 1).$$

- (ii) We say that  $X_n \rightarrow X$  *in probability*,  $X_n \xrightarrow{P} X$  as  $n \rightarrow \infty$  if  $\forall \epsilon > 0$

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

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

**Definition.** We say  $X_n \rightarrow X$  *in distribution* (in law) writing  $X_n \xrightarrow{d} X$  as  $n \rightarrow \infty$ , if

$$P(X_n \leq t) \rightarrow P(X \leq t) \quad \forall t \in \mathbb{R}^k \text{ for which } t \mapsto P(X \leq t) \text{ is continuous.}$$

Recall  $P(Z \leq z) = P(Z_1 \leq z_1, \dots, Z_k \leq z_k)$ .

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

**Proposition.** (i)  $X_n \xrightarrow[n \rightarrow \infty]{\text{a.s.}} X \implies X_n \xrightarrow[n \rightarrow \infty]{P} X \implies X_n \xrightarrow[n \rightarrow \infty]{d} X$  but any converse is false in general.

- (ii) (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

$$X_n \xrightarrow[n \rightarrow \infty]{} X \text{ a.s. / P / in law} \implies g(X_n) \xrightarrow[n \rightarrow \infty]{} g(X) \text{ a.s. / P / in law}$$

respectively.

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

$$- Y_n \xrightarrow[n \rightarrow \infty]{P} C \text{ as } n \rightarrow \infty$$

- $X_n + Y_n \xrightarrow{d} X + C$  as  $n \rightarrow \infty$
  - $X_n Y_n \xrightarrow{d} CX$  and provided  $C \neq 0$ ,  $X_n/Y_n \xrightarrow{d} X/C$  as  $n \rightarrow \infty$
  - If  $(A_n)_{ij}$  are random matrices such that  $(A_n)_{ij} \xrightarrow{P} A_{ij}$ , then  $A_n X_n \xrightarrow{d} AX$  as  $n \rightarrow \infty$
- (iv) If  $X_n \xrightarrow{d} X$  as  $n \rightarrow \infty$ , then  $X_n$  is stochastically bounded ( $Op(1)$ ), that is
- $$\forall \epsilon > 0 \exists M_\epsilon : \forall n \text{ large enough } \mathbb{P}(\|X_n\| > M_\epsilon) < \epsilon.$$

### 3.2 Law of large numbers and central limit theorem

Consider  $X_1, X_2, \dots$  of iid copies of  $X \sim P$  on  $\mathbb{R}^k$ . This sequence can be realised as the coordinate projection of the infinite product probability space

$$(\Omega, \mathcal{A}, P) = (\mathbb{R}^\mathbb{N}, \mathcal{B}^\mathbb{N}, P^\mathbb{N}), \quad P^\mathbb{N} = \otimes_{i=1}^\infty P,$$

where  $P^\mathbb{N}$  is the infinite product probability measure.  $P_r = P^\mathbb{N}$ , under which we can make simultaneous statements about the stochastic behaviour of  $X_1, X_2, \dots$

**Example.** The weak law of large numbers :  
If  $\text{var}(X) < \infty$  (unnecessary ) by Chebyshev,

$$\text{var}\left(\frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right) = \frac{\text{var } X}{n}.$$

$$P_r \left( \left| \frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \right| > \epsilon \right) \leq \frac{\text{var } X}{n\epsilon^2} \xrightarrow{n \rightarrow \infty} 0.$$

This is true for  $P_r$  a.s. but we will omit the proof.

**Theorem** (Strong law of large numbers). Let  $X_1, \dots, X_n$  be iid copies of the integrable random variable  $X \sim P$  on  $\mathbb{R}^k$ . Then

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow[n \rightarrow \infty]{} \mathbb{E}[X] \quad P_r \text{ a.s. .}$$

More is true, the stochastic fluctuations of  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$  around  $\mathbb{E}[X]$  are of order  $\frac{1}{\sqrt{n}}$  and as long as  $\text{var } X < \infty$ , this always look normally distributed.

**Theorem** (Central limit theorem). Let  $X_1, \dots, X_n$  be iid copies of  $X \sim P$  on  $\mathbb{R}$  and  $\text{var } X = \sigma^2 < \infty$ . Then

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

The multivariate version is also true. Recall that  $X \in \mathbb{R}^k$  is multivariate normal if

$$\forall \mathbf{t} \in \mathbb{R}^k, \mathbf{t}^k X$$

is univariate normal and write  $X \sim N_k(\mu, \Sigma)$  where  $\mu = \mathbb{E}[X]$  and  $\Sigma = \text{var } X$  (the covariance matrix). In fact,  $X$  is uniquely characterised as the random

variable on  $\mathbb{R}^k$  such that  $\mathbf{t}^T X \sim N(\mathbf{t}^T \mu, \mathbf{t}^T \Sigma \mathbf{t})$ . If  $\Sigma$  is invertible, the density of  $X$  is

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

Let  $A \in \mathbb{R}^{d \times l}$  and  $\mathbf{b} \in \mathbb{R}^d$ . Then

$$AX + \mathbf{b} \sim N_d(A\mu + \mathbf{b}, A\Sigma A^T).$$

Furthermore if  $A_n \xrightarrow{P} A$  are random matrices and  $X_n \xrightarrow{d} N_k(\mu, \Sigma)$ , then  $A_n X_n \xrightarrow{d} N_d(A\mu, A\Sigma A^T)$ . Lastly,  $\Sigma$  is diagonal  $\implies$  the components of  $X$  are independent.

**Theorem** (Multivariate central limit theorem). Let  $X_1, \dots, X_n$  be iid copies of  $X \sim P$  on  $\mathbb{R}$  and  $\text{var } X = \Sigma$  positive definite (unnecessary). Then,

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

Define, for a sequence  $Y_1, Y_2, \dots$  and  $c_1, c_2, \dots \in \mathbb{R} \setminus \{0\}$ .

$$Y_n = O_{P_r}(c_n) \text{ if } \forall \epsilon > 0 \exists M, N > 0 : P_r \left( \left| \frac{Y_n}{c_n} \right| > M \right) < \epsilon \forall n > N.$$

By Prokhorov's Theorem,

**Corollary.**

$$\bar{X}_n - \mathbb{E}[X] = O_{P_r} \left( \frac{1}{\sqrt{n}} \right).$$

Let  $k = 1$ ,  $X_1, \dots, X_n$  iid copies of  $X \sim P$ ,  $\mu_0 = \mathbb{E}[X]$ ,  $\sigma^2 = \text{var } X$ . Define

$$C_n = \{\mu \in \mathbb{R} : |\bar{X}_n - \mu| \leq \frac{\sigma Z_\alpha}{\sqrt{n}}\},$$

where  $z_\alpha$  is such that  $P_r(|Z| \leq z_\alpha) = 1 - \alpha$ ,  $Z \sim N(0, 1)$   
 $P_{\mu_0} = P$ ,

$$\begin{aligned} P_{\mu_0}^{\mathbb{N}}(\mu_0 \in C_n) &= P_{\mu_0}^{\mathbb{N}}(|\bar{X}_n - \mu_0| \leq \frac{\sigma Z_\alpha}{\sqrt{n}}) \\ &= P_r(|\bar{X}_n - \mathbb{E}[X]| \leq \frac{\sigma z_\alpha}{\sqrt{n}}) \\ &= P_r(\sqrt{n} \left| \frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}[X] \right| \leq z_\alpha) \end{aligned}$$

$$\xrightarrow[n \rightarrow \infty]{\text{CLT}} P_r(|Z| \leq z_\alpha) = 1 - \alpha.$$

by CLT, the continuous mapping theorem for  $|\cdot|$  and because  $z_\alpha$  is a continuity point of the distribution of  $Z \implies C_n$  is an asymptotic confidence interval with confidence level or coverage  $1 - \alpha$  (or size of significance level  $\alpha$ ). When  $\sigma$  is unknown, we replace it (in the definition of  $C_n$ ) by  $S_n$  where

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

and the same conclusion follows using the asymptotic distribution of the  $t$ -statistic

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

## 4 Consistency of MLEs

**Definition** (Consistent). Let  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim}$  form a statistical model  $\{P_\theta : \theta \in \Theta\}$ ,  $\Theta \subseteq \mathbb{R}^p$  then we say that an estimator  $\tilde{\theta}_n = \tilde{\theta}(X_1, \dots, X_n)$  is *consistent* (for the model) if

$$\tilde{\theta}_n \xrightarrow[n \rightarrow \infty]{} \theta \text{ in } (P_\theta^\mathbb{N})\text{-probability } \forall \theta \in \Theta.$$

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

- (i)  $f(x, \theta) > 0 \forall x \in \chi \forall \theta \in \Theta$ .
- (ii)  $\int_\chi f(x, \theta) dx = 1 \forall \theta \in \Theta$ .
- (iii) The map  $\theta \mapsto f(x, \theta)$  is continuous  $\forall x \in \chi$ .
- (iv)  $\Theta \subseteq \mathbb{R}^p$  is compact.
- (v)  $\theta = \theta' \iff f(\cdot, \theta) = f(\cdot, \theta') \forall \theta, \theta' \in \Theta$ .
- (vi)  $\mathbb{E}_\theta [\sup_{\theta \in \Theta} |\log f(x, \theta)|] < \infty \forall \theta \in \Theta$ .

**Remark.** (i) The above conditions justify the application of Jensen's inequality in our first observation in the information geometry section from earlier, in particular the map

$$\theta \mapsto \ell(\theta) \equiv \mathbb{E}_{\theta_0} [\log f(X, \theta)]$$

is uniquely maximised at  $\theta_0 \in \Theta$ .

- (ii) Using the dominated convergence theorem, (probability and measure) one can integrate the limit

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

with respect to  $\int(\cdot) dP_\theta$  and conclude that the map  $\theta \mapsto \ell(\theta)$  is continuous under our assumption.

**Theorem.** Suppose the statistical model  $\{f(\cdot, \theta) : \theta \in \Theta\}$  satisfies our above assumptions. Then a MLE exists and any MLE is consistent.

*Proof.* The map  $\bar{\ell}_n(\theta) = \frac{1}{n} \sum_{i=1}^n \log f(X_i, \theta)$  is continuous on the compact set  $\Theta \subseteq \mathbb{R}^p$  so by the Heine-Borel theorem,  $\bar{\ell}_n$  obtains a maximum on  $\Theta$ , hence a MLE  $\hat{\theta}_n$  exists. Now, let  $\hat{\theta}_n$  be any maximiser and fix a true (arbitrary) value  $\theta_0 \in \Theta$ . We now prove that  $\hat{\theta}_n \rightarrow \theta_0$  in probability as  $n \rightarrow \infty$  (in  $P = P_{\theta_0}^\mathbb{N}$ -probability). The idea is that maximisers  $\hat{\theta}_n$  of  $\bar{\ell}_n$  over  $\Theta$  should converge to the unique maximiser  $\theta_0$  of  $\ell$  over  $\Theta$ , since  $\bar{\ell}_n(\theta) \xrightarrow[n \rightarrow \infty]{P} \ell(\theta)$  by the law of large numbers for all  $\theta \in \Theta$  pointwise. This is generally false unless one has uniform convergence

$$\sup_{\theta \in \Theta} |\bar{\ell}_n(\theta) - \ell(\theta)| \xrightarrow{P} 0 \text{ as } n \rightarrow \infty,$$

(see example sheet for a counter example). We show in a lemma to follow that the above holds under the maintained hypothesis.

Define, for any  $\varepsilon > 0$

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

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

$$c(\varepsilon) = \sup_{\theta \in \Theta_\varepsilon} \ell(\theta) = \ell(\bar{\theta}_\varepsilon) < \ell(\theta_0),$$

since  $\ell$  is maximised uniquely at  $\theta$ . Then we can choose  $\delta(\varepsilon)$  small enough such that

$$c(\varepsilon) + \delta(\varepsilon) < \ell(\theta_0) - \delta(\varepsilon).$$

Now,

$$\sup_{\theta \in \Theta_\varepsilon} \bar{\ell}_n(\theta) = \sup_{\theta \in \Theta_\varepsilon} [\ell(\theta) + \bar{\ell}_n(\theta) - \ell(\theta)] \leq \sup_{\theta \in \Theta_\varepsilon} \ell(\theta) + \sup_{\theta \in \Theta_\varepsilon} |\bar{\ell}_n(\theta) - \ell(\theta)|.$$

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

$$A_n(\varepsilon) = \{\sup_{\theta \in \Theta} |\bar{\ell}_n(\theta) - \ell(\theta)| \leq \delta(\varepsilon)\}.$$

On these events we have

$$\sup_{\theta \in \Theta_\varepsilon} \bar{\ell}_n(\theta) < c(\varepsilon) + \delta(\varepsilon) \leq \ell(\theta_0) - \delta(\varepsilon) \leq \bar{\ell}_n(\theta_0),$$

since on  $A_n(\varepsilon)$  we also have  $|\ell(\theta_0) - \bar{\ell}_n(\theta_0)| < \delta(\varepsilon)$ . Thus if we assume that  $\hat{\theta}_n \in \Theta_\varepsilon$  then by what precedes

$$\bar{\ell}_n(\theta) \leq \sup_{\theta \in \Theta_\varepsilon} \bar{\ell}_n(\theta) < \ell(\theta_0)$$

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

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

Now we can conclude that  $P(A_n(\varepsilon)) \rightarrow 1$  and that  $P(\|\hat{\theta} - \theta_0\| < \varepsilon) \rightarrow 1$  as  $n \rightarrow \infty$  or  $P(\|\theta_n - \theta_0\| \geq \varepsilon) \rightarrow 0$  as  $n \rightarrow \infty$ . Since  $\varepsilon$  was arbitrary,  $\hat{\theta}_n \xrightarrow[n \rightarrow \infty]{P} \theta_0$  and the proof is complete modulo the verification of the next lemma.  $\square$

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

To verify our claim we now make the following digression. For a (measurable)  $\chi \subseteq \mathbb{R}^d$  and a (measurable)  $h : \chi \rightarrow \mathbb{R}$ , and let  $X_1, \dots, X_n$  be iid random variables in  $\chi$  with law  $P$ . Then the  $h(X_i)$ 's are also iid and if  $\mathbb{E}[|h(X)|] < \infty$  where we are using  $\mathbb{E} = \mathbb{E}_P$ , then by the strong law of large numbers

$$\frac{1}{n} \sum_{i=1}^n h(X_i) \xrightarrow[n \rightarrow \infty]{} 0 \text{ a.s. } (P^N).$$

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

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

Moreover,

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

Since

$$Pr\left(\left(\bigcap_{j=1}^N A_j\right)^c\right) = Pr\left(\bigcup_{j=1}^N A_j^c\right) \stackrel{\text{union bound}}{\leq} \sum_{j=1}^N Pr(A_j^c) = 0.$$

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

$$\forall h \in \mathcal{H} \exists j : \underline{h}_j(x) \leq h(x) \leq \overline{h}_j(x) \forall x \in \chi.$$

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

$$(i) \mathbb{E}[\underline{h}_j(X)] < \infty, \mathbb{E}[\overline{h}_j(X)] < \infty$$

$$(ii) \mathbb{E}[\overline{h}_j(X) - \underline{h}_j(X)] < \epsilon$$

Then

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

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

$$Pr\left(\max_{j=1, \dots, N(\frac{\epsilon}{3})} \left| \frac{1}{n} \sum_{i=1}^n \underline{h}_j(X_i) - \mathbb{E}[\underline{h}_j(X)] \right| \leq \frac{\epsilon}{3} \forall n \geq n_0(\epsilon)\right) = Pr(A_\epsilon) = 1$$

and similar for  $\overline{h}_j$ . Note that  $A_\epsilon = (A_m : m \in \mathbb{N})$  Now pick  $h \in \mathcal{H}$  arbitrary and write for the respective bracket  $[\underline{h}_j, \overline{h}_j] \ni h$

$$\frac{1}{n} \sum_{i=1}^n h(X_i) - \mathbb{E}[h(X)] \leq \frac{1}{n} \sum_{i=1}^n \overline{h}_j(X_i) - \mathbb{E}[\overline{h}_j(X)] + \mathbb{E}[\overline{h}_j(X)] - \mathbb{E}[h(X)] \leq \frac{\epsilon}{3} + \mathbb{E}[\overline{h}_j(X) - \underline{h}_j(X)] \leq \frac{2\epsilon}{3}.$$

Likewise we get that

$$\frac{1}{n} \sum_{i=1}^n h(X_i) - \mathbb{E}[h(X)] \geq \frac{1}{n} \sum_{i=1}^n \underline{h}_j(X_i) - \mathbb{E}[\underline{h}_j(X)] + \mathbb{E}[\underline{h}_j(X)] - \mathbb{E}[h(X)] \geq -\frac{2\epsilon}{3}.$$

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

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

and since

$$Pr(A^c) \subseteq \sum_{m=1}^{\infty} Pr(A_m^c) = 0.$$

□

**Proposition.** Let  $X \subseteq \mathbb{R}^d$ ,  $\Theta \subseteq \mathbb{R}^p$  compact, 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  $X_1, \dots, X_n$  are iid copies of  $X$

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^n q(X_i, \theta) - \mathbb{E}[q(X, \theta)] \right| \xrightarrow{n \rightarrow \infty} 0 \quad \text{a.s.}$$

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

**Remark.** The condition  $\mathbb{E}[\sup_{\theta \in \Theta} |q(X, \theta)|] < \infty$  can be seen to be necessary (as  $\mathbb{E}[\|Z\|] < \infty$ ) in the law of large numbers for  $Z_1, \dots, Z_n$  iid in the space  $C(\Theta)$  of countable functions on the compact set  $\Theta$ .



## 5 Asymptotic distribution of MLEs

**Definition** (Asymptotically efficient). We say that an estimator  $\tilde{\theta}_n$  is *asymptotically efficient* in a regular statistical model  $\{f(\cdot, \theta) : \theta \in \Theta\}$  if

$$\lim_{n \rightarrow \infty} n \operatorname{var}_{\theta}(\tilde{\theta}) = I(\theta)^{-1} \quad \forall \theta \in \operatorname{int} \Theta.$$

**Theorem.** Suppose a statistical model  $\{f(\cdot, \theta) : \theta \in \Theta\}$  is regular in the sense that it satisfies Condition B (on the handout). Then if  $\hat{\theta}_n$  is the MLE based on  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim}$  from the model we have

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow[n \rightarrow \infty]{d} N(0, I(\theta_0)^{-1}).$$

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

$$0 = \ell'_n(\hat{\theta}) = \ell'_n(\theta_0) + \ell''_n(\bar{\theta}_n)(\hat{\theta}_n - \theta_0) \quad (\text{MVT}) .$$

So

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{\frac{1}{\sqrt{n}} \ell'_n(\theta_0)}{-\frac{1}{\sqrt{n}} \ell''_n(\bar{\theta}_n)} = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{d}{d\theta} \log f(X_i, \theta_0)}{-\frac{1}{\sqrt{n} (\frac{d^2}{d\theta^2})^2} \log f(X_i, \bar{\theta}_n)} .$$

With the numerator converging in distribution to  $N(0, I(\theta_0))$  and the denominator converging to  $I(\theta_0)$ .

*Proof.* (Of our theorem above).  $Pr = P_{\theta_0}^N, \mathbb{E} = \mathbb{E}_{\theta_0}$

**Lemma.** Our observations from the information geometry section are valid.

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

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

$$|Pr(Z_n \leq t) - Pr(Z_n \leq t, E_n)| \leq Pr(E_n^c) \xrightarrow[n \rightarrow \infty]{} 0.$$

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

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

where we recall that  $\bar{\ell}_n(\theta) = \frac{1}{n} \sum_{i=1}^n \log f(X_i, \theta)$  and  $\frac{\partial}{\partial \theta} \bar{\ell}_n(\hat{\theta}_n) = \frac{\partial}{\partial \theta} \bar{\ell}_n(\theta) \Big|_{\theta=\hat{\theta}_n}$ .

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

$$h(\hat{\theta}_n) = h(\theta_0) + \frac{\partial h}{\partial \theta} \Big|_{\theta=\bar{\theta}} (\hat{\theta}_n - \theta_0),$$

where  $\bar{\theta} = \bar{\theta}(n)$  is some mean value on that line segment. Second derivatives of  $\bar{\ell}_n(\theta)$  are differentials of the map  $u \mapsto \frac{\partial}{\partial \theta} \ell_n(\theta) \Big|_{\theta=u}$  and hence applying what precedes  $p$  times to the vector entries  $\frac{\partial}{\partial \theta_j} \bar{\ell}_n(\hat{\theta})$  we obtain

$$0 = \begin{pmatrix} \vdots \\ \frac{\partial}{\partial \theta_j} \bar{\ell}_n(\hat{\theta}_n) \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ \frac{\partial}{\partial \theta_j} \bar{\ell}_n(\theta_0) \\ \vdots \end{pmatrix} + \underbrace{\begin{pmatrix} \vdots \\ \cdots \quad \frac{\partial}{\partial \theta_i \partial \theta_j} \bar{\ell}_n(\bar{\theta}_{(j)}) \quad \cdots \\ \vdots \end{pmatrix}}_{\equiv \bar{A}_n} (\hat{\theta}_n - \theta_0),$$

where  $\bar{\theta}_{(j)}$  is the  $p \times 1$  vector arising from the  $j$ th application of the MVT. We will show

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

In particular, this implies convergence of

$$\|\bar{A}_n - (-I(\theta_0))\|_{\text{operator norm}} \xrightarrow{P} 0.$$

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

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

and the result follows from the convergence of  $\bar{A}_n$ , Slutsky's lemma and since

$$\sqrt{n} \frac{\partial}{\partial \theta} \bar{\ell}_n(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left( \frac{\partial}{\partial \theta} \log f(X_i, \theta_0) - \underbrace{\mathbb{E} \left[ \frac{\partial}{\partial \theta} \log f(X, \theta_0) \right]}_{=0} \right) \xrightarrow[\text{CLT}]{d} N(0, I(\theta_0)) \text{ as } n \rightarrow \infty.$$

To verify the convergence of  $\bar{A}_n$ , it suffices (see example sheet) to check convergence in probability of  $\bar{A}_{n,jk} \rightarrow (-I(\theta_0))_{jk}$ . Now we write

$$\bar{A}_{n,jk} = \frac{1}{n} \sum_{i=1}^n \frac{\partial^2}{\partial \theta_j \partial \theta_k} \log f(X_i, \bar{\theta}_{(j)}) - \mathbb{E} \left[ \frac{\partial^2}{\partial \theta_j \partial \theta_k} \log f(X, \bar{\theta}_{(j)}) \right] + \mathbb{E} \left[ \frac{\partial^2}{\partial \theta_j \partial \theta_k} \log f(X, \bar{\theta}_{(j)}) \right] - \mathbb{E} \left[ \frac{\partial^2}{\partial \theta_j \partial \theta_k} \log f(X, \theta_0) \right]$$

Where we note that the expectation acts on  $X$  only and  $\bar{\theta}_{(j)}$  is still random and we write the sum as I + II  $-I(\theta_0)_{jk}$  and let us show that I + II  $\xrightarrow[n \rightarrow \infty]{P} 0$ . For I we note that  $\bar{\theta}_{(j)} \in K$  and hence with  $q(x, \theta) = \frac{\partial^2}{\partial \theta_j \partial \theta_k} \log f(X, \theta)$

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

by the uniform law of large numbers.

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

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

completing the proof.  $\square$

**Remark.** (i) 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 family of Laplace distribution (where one may show  $I_n(\theta) = n$ ). However, this cannot be weakened further, and for non-smooth parametrisation the asymptotic theory for MLEs may be different as the example of  $U(0, \theta), \theta \in [0, \infty]$  shows (example sheet).

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

(iii) 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' (in stead 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.

**Example.** Consider a statistical model,  $\{P_\theta : \theta \in \Theta\}$  where  $\Theta \subseteq \mathbb{R}, 0 \in \Theta$  such that

$$\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta) \xrightarrow[n \rightarrow \infty]{d} N(0, I(\theta)^{-1}) \quad \forall \theta \in \text{int}\Theta.$$

[Recall that this implies that  $\sqrt{n}(\hat{\theta}_n - \theta)$  is stochastically bounded i.e.  $\forall \varepsilon > 0 \exists M_\varepsilon :$

$$Pr\left(|\hat{\theta}_n - \theta| > \frac{M_\varepsilon}{\sqrt{n}}\right) < \varepsilon,$$

in particular,  $\hat{\theta}_n \xrightarrow[n \rightarrow \infty]{P} \theta]$  Define

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

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

$$\begin{aligned} P_\theta(\tilde{\theta} \neq \hat{\theta}) &= P_\theta(|\hat{\theta}| < n^{-\frac{1}{4}}) \\ &= P_\theta(|\hat{\theta} - \theta + \theta| < n^{-\frac{1}{4}}) \\ &\leq P_\theta(|\hat{\theta} - \theta| \geq |\theta| - n^{-\frac{1}{4}}) \\ &\stackrel{n=n_\theta \text{ large enough}}{\leq} P_\theta(|\hat{\theta} - \theta| > \frac{|\theta|}{2}) \xrightarrow[n \rightarrow \infty]{} 0. \end{aligned}$$

Where the limit follows since  $\hat{\theta} \xrightarrow{P} \theta$  and  $|\theta| \neq 0$ . So for such  $\theta$  we thus have

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

$$\begin{aligned} P_0(\tilde{\theta} \neq 0) &= P_0(|\hat{\theta}| \geq n^{-\frac{1}{4}}) \\ &= P_0(|\hat{\theta} - \theta| > n^{-\frac{1}{4}}) \\ &= P_0(\sqrt{n}|\hat{\theta} - \theta| > n^{\frac{1}{4}}) \end{aligned}$$

So for any  $\varepsilon > 0$  and  $n$  such that  $n^{\frac{1}{4}} > M_\varepsilon$ , we have by stochastic boundedness of  $\sqrt{n}(\hat{\theta} - \theta)$  that the last probability  $< \varepsilon$ . Hence we conclude that under  $P_0$ ,

$$\sqrt{n}(\tilde{\theta} - \theta) \xrightarrow{d} N(0, 0),$$

so  $\tilde{\theta}$  'beats' the asymptotic efficiency bound  $I(\theta)^{-1}$  at  $\theta = 0$ .

## 6 Plug-in MLEs and the Delta-method

Consider estimating a functional  $\Phi : \Theta \rightarrow \mathbb{R}^k$ ,  $\Theta \subseteq \mathbb{R}^p$  based on  $X_i \stackrel{\text{iid}}{\sim} \{f(\cdot, \theta) : \theta \in \Theta\}$  where  $\hat{\theta}$  is the MLE for  $\theta$ . One can show that a MLE in the model  $\{f(\cdot, \phi) : \phi = \Phi(\theta) \theta \in \Theta\}$  can be obtained from  $\Phi(\hat{\theta})$ . The asymptotic normality and efficiency of  $\hat{\theta}$  then implies the same for  $\Phi(\hat{\theta})$  as long as  $\Phi$  is differentiable.

**Theorem** (Delta-method). Suppose  $\Phi : \Theta \rightarrow \mathbb{R}$  is a continuously differentiable at  $\theta \in \Theta$  with gradient vector  $\frac{\partial \Phi}{\partial \theta}(\theta)$ . Suppose further  $\hat{\theta}_n$  are random vectors in  $\Theta$  such that  $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n \rightarrow \infty]{d} Z$  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  $\{t\hat{\theta} + (1-t)\theta : 0 < t < 1\}$  we can write for mean values  $\bar{\theta}_n$

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

Since  $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n \rightarrow \infty]{d} Z$  we have in particular  $\hat{\theta}_n \xrightarrow[n \rightarrow \infty]{P} \theta$  (by stochastic boundedness) so also  $\bar{\theta}_n \xrightarrow[n \rightarrow \infty]{P} \theta$  hence by the continuous mapping theorem, we also have

$$\frac{\partial \Phi}{\partial \theta}(\bar{\theta}_n) \xrightarrow{P} \frac{\partial \Phi}{\partial \theta}(\theta),$$

hence by Slutsky's lemma, the last displayed expression  $\xrightarrow[n \rightarrow \infty]{d} \frac{\partial \Phi}{\partial \theta}(\theta)^T Z$   $\square$

**Remark.** If  $\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta) \xrightarrow[n \rightarrow \infty]{d} N(0, I(\theta)^{-1})$  then what precedes implies that the plug-in MLE satisfies

$$\sqrt{n}(\Phi(\hat{\theta}_{\text{MLE}}) - \Phi(\theta)) \xrightarrow{d} N(0, \frac{\partial \Phi}{\partial \theta}(\theta)^T I(\theta)^{-1} \frac{\partial \Phi}{\partial \theta}(\theta))$$

in particular the asymptotic covariance attains the CRLB for estimating  $\Phi(\theta)$ .

## 7 Asymptotic inference with the MLE

**Example.** 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  $\theta_i = e_i^T \theta$ ,  $e_i = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}$

with the 1 in the  $i$ th position, and the last theorem

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

This suggests an asymptotic confidence interval

$$C_n = \{v \in \mathbb{R} : |\hat{\theta}_i - v| \leq \frac{(I(\theta)^{-1})_{ii}^{\frac{1}{2}}}{\sqrt{n}} z_\alpha\},$$

where

$$z_\alpha \text{ is defined by } Pr(|Z| \leq z_\alpha) = 1 - \alpha, \quad Z \sim N(0, 1).$$

Indeed, by the continuous mapping theorem,

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

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** (Observed Fisher information). The  $p \times p$  matrix

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

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

**Proposition.** Suppose the statistical model  $\{f(\cdot, \theta) : \theta \in \Theta\}$  satisfies condition B (our second set of assumptions). Then

$$\hat{i}_n \xrightarrow{P} I(\theta_0) \text{ as } n \rightarrow \infty.$$

*Proof.* Just as when proving  $\bar{A}_n \xrightarrow[n \rightarrow \infty]{P} I(\theta_0)$  in the proof of asymptotic normality of  $\hat{\theta}_{\text{MLE}}$  replacing  $\frac{\partial^2}{\partial \theta \partial \theta^T}, \overline{\theta_{(i)}}$  by  $\frac{\partial}{\partial \theta}$  and  $\hat{\theta}_{\text{MLE}}$  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 estimate  $I_n(\theta_0)$  by  $\hat{j} = j_n(\hat{\theta}_{MLE})$  which as before (and by an observation in information geometry) satisfies  $\hat{j} \xrightarrow[n \rightarrow \infty]{P} I(\theta_0)$  as well.

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

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

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

$$W_n(\theta) \xrightarrow[n \rightarrow \infty]{d} \chi_p^2$$

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  $\chi_p^2$  distribution.

Consider next a hypothesis testing problem

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

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

$$P_\theta(\text{reject } H_0) = \mathbb{E}_\theta[\psi_n],$$

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

$$P_\theta(\text{accept } H_0) = \mathbb{E}_\theta[1 - \psi_n].$$

**Definition** (Likelihood ratio test statistic). A general purpose test can be constructed from the *likelihood ratio test statistic*

$$\Lambda_n(\Theta, \Theta_0) = 2 \log \frac{\prod_{i=1}^n f(X_i, \hat{\theta}_{MLE})}{\prod_{i=1}^n f(X_i, \hat{\theta}_{MLE,0})} = 2 \log \frac{\sup_{\theta \in \Theta} \prod_{i=1}^n f(X_i, \theta)}{\sup_{\theta \in \Theta_0} \prod_{i=1}^n f(X_i, \theta)}.$$

**Theorem** (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  $P_{\theta_0}$ ,

$$\Lambda_n(\Theta, \Theta_0) \xrightarrow{d} \chi_p^2, p = \dim \Theta.$$

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

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

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

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

$$\begin{aligned} \Lambda_n(\Theta, \Theta_0) &= 2\ell_n(\hat{\theta}_n) - 2\ell_n(\theta_0) \\ &= (-2\ell_n(\theta_0)) - (-2\ell_n(\hat{\theta}_n)) \\ &= -2 \frac{\partial}{\partial \theta} \ell_n(\hat{\theta}_n) - (\theta_0 - \hat{\theta}_n)^T \frac{\partial^2}{\partial \theta \partial \theta} \ell_n(\bar{\theta})(\theta_0 - \hat{\theta}_n), \end{aligned}$$

where  $\bar{\theta}_n$  are mean values lying on the line segment connecting  $\hat{\theta}_n, \theta_0$  since  $\hat{\theta}_n \in \text{int}\Theta$  so the gradient of maximum must vanish. The second order term can be written as

$$\sqrt{n}(\hat{\theta}_n - \theta_0)^T (\hat{j}(\bar{\theta}) - I(\theta_0)) \xrightarrow[\substack{P \\ \text{previous prop}}]{d} N_{\sqrt{n}(\hat{\theta}_n - \theta_0)} + \sqrt{n}(\hat{\theta}_n - \theta_0)^T I(\theta_0) \sqrt{n}(\hat{\theta}_n - \theta_0).$$

So the LHS converges to 0 in probability by and the RHS converges in distribution to

the continuous map  $X \mapsto X^T I(\theta_0) X$  from  $\mathbb{R}^p \rightarrow \mathbb{R}$ . Moreover, by standard linear algebra

$$Z^T I(\theta_0) Z \propto \sum_{i=1}^p W_i^2, W_i \stackrel{\text{iid}}{\sim} N.$$

So  $\sim \chi_p^2$

□



## 8 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* probability 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

$$P(X = x | H_i) = f_i(x),$$

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

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

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

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

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  $X | \theta \sim f(\cdot, \theta)$  and  $\theta \sim \pi$  on  $\Theta$ , where  $\pi$  is the prior distribution. The *posterior distribution* is then the conditional distribution of  $\theta | X$ . 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 density mass 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

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

Likewise

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

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

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

**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  $X|\theta$ , we see

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

So  $\pi(\cdot|X_1, \dots, X_n) \sim N(\frac{1}{n+1} \sum_{i=1}^n X_i, \frac{1}{n+1})$

One shows more generally that for normal priors 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. We have some other examples

**Example.**    – Beta prior + Binomial sampling  $\rightarrow$  Beta posterior  
                   – Gamma prior + Poisson sampling  $\rightarrow$  Gamma posterior

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

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

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 ( $\sigma^2$  known) one sees that the Jeffrey's prior is  $\propto$  constant, and one shows that the 'improper' posterior equals a  $N(\bar{X}_n, \frac{\sigma^2}{n}) = N(\hat{\theta}_{\text{MLE}}, \frac{\sigma^2}{n})$ , see the example sheet. Note however (see the example 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.

### 8.1 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 for the posterior distribution, the Bayesian version of a confidence 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

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

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

## 8.2 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  $X_i \stackrel{\text{iid}}{\sim} f(\cdot, \theta_0) \theta_0 \in \Theta$ .

**Example.**  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim}$  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  $\bar{\theta}_n = \frac{1}{n+1} \sum_{i=1}^n X_i \rightarrow \theta_0$   $P_{\theta_0}^{\mathbb{N}}$ -a.s., and also that  $\sqrt{n}(\bar{\theta}_n - \theta) \xrightarrow{d} N(0, I(\theta_0)^{-1})$  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** (Bernstein-von Mises). Suppose a statistical model  $\{f(\cdot, \theta) : \theta \in \Theta\}$  satisfies assumptions 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  $N(\hat{\theta}_{\text{MLE}}, \frac{1}{n} I(\theta_0)^{-1})$ . Then as  $n \rightarrow \infty$

$$\|\pi_n - \hat{\phi}_n\|_{L^1} = \int_{\mathbb{R}} |\pi_n(\theta) - \hat{\phi}_n(\theta)| d\theta \rightarrow 0 \text{ } P_{\theta_0}^{\mathbb{N}}\text{-a.s..}$$

*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  $\hat{\theta}_{\text{MLE}} = \bar{X}_n$ . Recall  $\pi_n$  is the pdf of a  $N(\bar{\theta}, \frac{1}{n+1})$  distribution where  $\bar{\theta} = \frac{n}{n+1}$  and so

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

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

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

Where we substituted  $v = \sqrt{n}(\theta - \hat{\theta})$  with  $\frac{dv}{d\theta} = \sqrt{n}$ . Fix  $w \in \Omega_0 \subseteq \Omega$  such that  $P_{\theta_0}^{\mathbb{N}}(\Omega_0) = 1$ , so that  $\sqrt{n}(\hat{\theta}(\omega) - \bar{\theta}(\omega)) \xrightarrow[n \rightarrow \infty]{} 0$  (as scalars) and 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  $\rightarrow 0 \forall \omega \in \Omega_0$  since  $P_{\theta_0}^{\mathbb{N}}(\Omega_0) = 1$  the limit holds a.s.  $\square$

**Remark.** The last theorem remains true when  $\hat{\theta}_{\text{MLE}}$  is replaced by any estimator  $\bar{\theta}_n$  such that

$$\sqrt{n}(\bar{\theta}_n - \bar{\theta}) \xrightarrow[n \rightarrow \infty]{} 0 \text{ } P_{\theta_0}^{\mathbb{N}} \text{ a.s. ,}$$

typically permitting the alternative centering at  $\bar{\theta} = \mathbb{E}^{\pi}[\theta | X_1, \dots, X_n]$

One important consequence of the BvM - theorem is that certain posterior *credible sets* are in fact proper (asymptotic) frequentist confidence sets.

**Example.** Consider

$$C_n = \{\theta : |\theta - \hat{\theta}_{\text{MLE}}| \leq \frac{R_n}{\sqrt{n}}\},$$

where  $R_n$  are random quantile constants chosen such that  $\pi(C_n | X_1, \dots, X_n) = 1 - \alpha, 0 < \alpha < 1$ .

Recall  $\hat{\phi}_n$  was the pdf of  $N(\hat{\theta}_{\text{MLE}}, \frac{1}{n}I(\theta_0)^{-1})$  distribution, and define further  $\phi_0$  to be the pdf of  $Z \sim N(0, I(\theta)^{-1})$ . We can define  $\Phi(t) = Pr(|Z| \leq t) = \int_{-t}^t \phi_0(v) dv$  which is strictly increasing in  $t$  and also continuously differentiable, hence admits a continuous inverse  $\Phi^{-1} : [0, 1] \rightarrow \mathbb{R}$ . Now we can write

$$\Phi(R_n) = \int_{-R_n}^{R_n} \phi_0(v) dv$$

and substituting  $v = \sqrt{n}(\theta - \hat{\theta})$  so that  $-R_n \leq v \leq R_n$  becomes the set  $C_n$  and since

$$v \sim N(0, I(\theta_0)^{-1}) \iff \sqrt{n}(\theta - \hat{\theta}) \sim N(\hat{\theta}, \frac{1}{n}I(\theta_0)^{-1}),$$

the last integral equals

$$\begin{aligned} \int_{C_n} \hat{\phi}_n(\theta) d\theta &= \int_{C_n} (\hat{\phi}_n(\theta) - \pi_n(\theta)) d\theta + \int_{C_n} \pi_n(\theta) d\theta \\ &\leq \int_{\mathbb{R}} |\hat{\phi}_n(\theta) - \pi_n(\theta)| d\theta + 1 - \alpha. \end{aligned}$$

But we know that the first term  $\xrightarrow{n \rightarrow \infty} 0$   $P_{\theta_0}^{\mathbb{N}}$ -a.s. So by the BvM theorem we know that  $\Phi(R_n) \xrightarrow{n \rightarrow \infty} 1 - \alpha$  a.s. under  $P_{\theta_0}$ , hence applying the continuous mapping theorem to  $\Phi^{-1}$  we deduce

$$R_n = \Phi^{-1}(\Phi(R_n)) \xrightarrow[n \rightarrow \infty]{\text{a.s.}} \Phi^{-1}(1 - \alpha) \text{ under } P_{\theta_0}.$$

In particular, by Slutsky's theorem and asymptotic normality of  $\hat{\theta}_{\text{MLE}}$ , we have

$$\frac{\Phi^{-1}(1 - \alpha)}{R_n} \sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta_0) \xrightarrow{d} N(0, I(\theta_0)^{-1}) \text{ as } n \rightarrow \infty \text{ under } P_{\theta_0}.$$

Now

$$\begin{aligned} P_{\theta_0}^{\mathbb{N}}(\theta_0 \in C_n) &= P_{\theta_0}^{\mathbb{N}}\left(|\theta_0 - \hat{\theta}_{\text{MLE}}| \leq \frac{R_n}{\sqrt{n}}\right) \\ &= P_{\theta_0}^{\mathbb{N}}\left(\frac{\Phi^{-1}(1 - \alpha)}{R_n} \sqrt{n}|\hat{\theta}_{\text{MLE}} - \theta_0| \leq \Phi^{-1}(1 - \alpha)\right) \\ &\xrightarrow{n \rightarrow \infty} Pr(|Z| \leq \Phi^{-1}(1 - \alpha)) \\ &= \Phi(\Phi^{-1}(1 - \alpha)) = 1 - \alpha. \end{aligned}$$

Where the limit follows from the line above combined with continuous mapping theorem.

**Remark.** Replacing  $|\cdot|$  by  $\|\cdot\|$ , the argument extends to the parameter spaces  $\Theta \subseteq \mathbb{R}^p$  for  $p > 1$ . We conclude that credible sets computed for priors  $\pi$  with positive continuous density functions on  $\Theta$  give rise to asymptotically exact level  $1 - \alpha$  (frequentist) confidence regions. A version for discrete prior can be proved as well.

## 9 Decision theory

Consider a (single) observation  $X$  from some statistical model  $\{f(\cdot, \theta) : \theta \in \Theta\}$  of pdf / pmf's on  $\chi \subseteq \mathbb{R}^d$ . In a decision problem we consider decision maps  $\delta : \chi \rightarrow \mathcal{A}$  where  $\mathcal{A}$  is some 'action space'.

**Example.** (i)  $\mathcal{A} = \{0, 1\}$ , a binary decision problem, where  $\delta(X) \in \{0, 1\}$  will be a test function or more generally a finite decision problem  $\{1, \dots, M\}$ .

(ii)  $\mathcal{A} = \Theta$ , and decision rules are estimators  $\hat{\theta}(X) \in \Theta$ .

(iii)  $\mathcal{A} = \{\text{all (measurable) subsets of } \Theta\}$ , set valued estimation, decision rules  $\delta(X) = C(X)$  are confidence regions.

For general decision problems, we consider *loss functions*

$$L : \mathcal{A} \times \Theta \rightarrow [0, \infty),$$

measuring the error  $L(\delta(X), \theta)$  incurred by  $\delta(X)$  for observation  $X$  and parameter value  $\theta$ .

**Example.** – Testing -  $L(a, \theta) = 1\{a \neq a(\theta)\}$  where  $a(\theta)$  is the index in  $\mathcal{A}$  corresponding to  $\theta$ .

– Estimation -  $\Theta \subseteq \mathbb{R}$ , absolute loss  $L(a, \theta) = |a - \theta|$ , and  $L(a, \theta) = (a - \theta)^2$  the squared loss

**Definition.** The *risk* of a decision rule  $\delta(X)$  from  $X \sim P_\theta$ , for loss  $L$ , is defined as

$$R(\delta, \theta) = \mathbb{E}_\theta [L(\delta(X), \theta)] = \int_\chi L(\delta(X), \theta) f(x, \theta) dx.$$

**Example.** For squared loss, in estimation problems with estimator  $\bar{\theta}(X) = \delta(X)$

$$R(\delta, \theta) = \mathbb{E}_\theta [(\bar{\theta} - \theta)^2],$$

equals the mean squared error (MSE).

**Definition.** Given a prior  $\pi$  on  $\Theta$ , the  $\pi$ -Bayes risk of a decision rule  $\delta$  is

$$R_\pi(\delta) = \int_\Theta R(\delta, \theta) \pi(\theta) d\theta = \int_\Theta \int_\chi L(\delta(X), \theta) f(x, \theta) \pi(\theta) dx d\theta.$$

A  $\pi$ -Bayes decision rule  $\delta_\pi$  is any decision rule for which  $R_\pi(\delta)$  is minimised in  $\delta$ .

We can rewrite the risk, interchanging the order of integration,

$$\begin{aligned} R_\pi(\delta) &= \int_\chi \int_\Theta L(\delta(x), \theta) \frac{f(x, \theta) \pi(\theta)}{\int_\Theta f(x, v) \pi(v) dv} \left( \int_\Theta f(x, v) \pi(v) dv \right) d\theta dx \\ &= \int_\chi \int_\Theta L(\delta(x), \theta) \pi(\theta|X) m(x) d\theta dx \\ &= \int_\chi E^\pi [L(\delta(x), \theta)|x] m(x) dx, \end{aligned}$$

where  $E^\pi [L(\delta(X), \theta)|X]$  is the posterior risk. Note that  $m(x) \geq 0$ . We conclude that any decision rule  $\bar{\delta}(X)$  which minimises the posterior risk in the sense that  $\forall X, \delta$

$$E^\pi [L(\bar{\delta}(X), \theta)|X] \leq E^\pi [L(\delta(X), \theta)|X],$$

then this inequality can be  $m(x)dx$ -integrated to deduce that  $\bar{\delta}(X)$  is a Bayes rule  $\delta_\pi(X)$  minimising the  $\pi$  Bayes risk.

**Remark.** One shows (example sheet) that for quadratic risk, the unique  $\pi$ -Bayes rule  $\delta_\pi(X)$  equals the posterior mean  $E^\pi [\theta|X]$ , and this is the unique Bayes rule. For absolute loss (risk), the  $\pi$ -Bayes rule will be the posterior median ( $p = 1$ ).

**Proposition.** In an estimation problem, suppose a decision rule  $\delta(X)$  is unbiased for  $\Theta$ , i.e.  $\mathbb{E}_\theta [\delta(X)] = \theta \forall \theta \in \Theta$ . Assume further that  $\delta$  is a  $\pi$ -Bayes rule for some prior  $\pi$  on  $\Theta$  and squared loss (quadratic risk). Then ( $E = E_Q$  where  $Q$  has density on  $\chi \times x\Theta$  given by  $dQ(x, \theta) = f(x, \theta)\pi(\theta)dxd\theta$ )

$$E [(\delta(X) - \theta)^2] = \int_\chi \int_\Theta (\delta(x) - \theta)^2 f(x, \theta)\pi(\theta)d\theta dx = 0.$$

[one says  $\delta(X) = \theta$  a.s. under  $Q$ ]

*Proof.* Recall the 'tower property' of iterated expectations

$$\mathbb{E} [Z(X, \theta)] = \mathbb{E} [\mathbb{E}^\pi [Z(X, \theta)|X]] = \mathbb{E} [\mathbb{E}_\theta [Z(X, \theta)]] .$$

Moreover, by the previous remark,  $\delta(X) = \delta_\pi(X) = \mathbb{E}^\pi [\theta|X]$  (by uniqueness). Thus

$$\mathbb{E} [\delta(X)\theta] = \mathbb{E} [\mathbb{E}^\pi [\theta|X] \delta(X)] = \mathbb{E} [\delta^2(X)]$$

and likewise

$$\mathbb{E} [\delta(X)\theta] = \mathbb{E} [\mathbb{E}_\theta [\delta(X)] \cdot \theta] = \mathbb{E} [\theta^2] .$$

Now

$$\mathbb{E} [(\delta(X) - \theta)^2] = \mathbb{E} [\delta^2(X)] - 2\mathbb{E} [\theta\delta(X)] + \mathbb{E} [\theta^2] = 0.$$

□

From what precedes, unbiased estimators are typically not  $\pi$ -Bayes rules for any prior  $\pi$ .

**Example.** (i)  $\bar{X}_n = \hat{\theta}_{MLE}$  in  $N(\theta, 1), \theta \in \Theta = \mathbb{R}$  is not a Bayes rule for any prior in quadratic risk.

(ii)  $X/n$  in a  $\text{Bin}(n, \theta), \theta \in \Theta = [0, 1]$  is a  $\pi$ -Bayes rule for quadratic risk only for degenerate priors.

## 9.1 Minimax risk

**Definition** (Minimax). A decision rule  $\delta(X)$  in a decision problem is called *minimax* if it attains the *minimax* risk (for loss  $L$ )

$$\inf_{\delta(X)} \sup_{\theta \in \Theta} R(\delta, \theta)$$

(if attained  $\min_{\delta(X)} \max_{\theta \in \Theta} R(\delta, \theta)$ ) where  $R_m(\delta, \Theta) = \sup_{\theta \in \Theta} R(\delta, \theta)$  is the maximal / worst case risk.

clearly, the Bayes risk for any prior is dominated by the minimax risk, since

$$R_m(\delta) = \int_{\theta} R(\delta, \theta) \pi(\theta) d\theta \leq R_m(\delta, \Theta) \times \underbrace{\int_{\Theta} \pi(\theta) d\theta}_{=1}.$$

**Definition** (Least favourable). A prior  $\lambda$  on  $\Theta$  is called *least favourable* (for a decision problem) if

$$R_{\lambda}(\delta_{\lambda}) \geq R_{\lambda'}(\delta_{\lambda'}) \quad \forall \lambda' \text{ priors on } \Theta.$$

**Proposition.** In a decision problem suppose for some prior  $\pi$  on  $\Theta$ , we have that

$$R_{\pi}(\delta_{\pi}) = \sup_{\theta \in \Theta} R(\delta_{\pi}, \theta)$$

(the  $\pi$ -Bayes risk of  $\delta_{\pi}$  coincides with the worst case risk of  $\delta_{\pi}$ ).

Then,

- (i)  $\delta_{\pi}$  is minimax.
- (ii) If  $\delta_{\pi}$  is the unique Bayes rule, then  $\delta_{\pi}$  is the unique minimax.
- (iii)  $\pi$  is least favourable.

*Proof.* (i) Let  $\delta$  be any decision rule with maximal risk

$$\begin{aligned} \sup_{\theta \in \Theta} R(\delta, \theta) &\geq \int_{\Theta} R(\delta, \theta) \pi(\theta) d\theta \\ &= R_{\pi}(\delta) \geq R_{\pi}(\delta_{\pi}) \\ &= \sup_{\theta \in \Theta} R(\delta_{\pi}, \theta). \end{aligned}$$

So taking inf over all  $\delta$  we see

$$\inf_{\delta} \sup_{\theta \in \Theta} R(\delta, \theta) \geq \sup_{\theta \in \Theta} R(\delta_{\pi}, \theta),$$

hence  $\delta_{\pi}$  attains the minimax risk.

- (ii) Moreover, the second preceding inequality is strict when  $\delta_{\pi}$  is the unique  $\pi$ -Bayes rule and if  $\delta \neq \delta_{\pi}$  so that for such  $\delta$  we have

$$\sup_{\theta \in \Theta} R(\delta, \theta) > \sup_{\theta \in \Theta} R(\delta_{\pi}, \theta),$$

hence  $\delta_{\pi}$  is the unique minimax.

- (iii) Let  $\pi'$  be any prior on  $\Theta$  then

$$R_{\pi'}(\delta_{\pi'}) \leq R_{\pi'}(\delta_{\pi}) = \int_{\Theta} R(\delta_{\pi}, \theta) \pi'(\theta) d\theta \leq \sup_{\theta \in \Theta} R(\delta_{\pi}, \theta) \cdot 1 = R_{\pi}(\delta_{\pi}).$$

Hence  $\pi$  maximises the  $\pi$ -risk of the  $\pi$ -Bayes rule among all  $\pi$ 's, i.e. is least favourable.

The corollary also follows since for  $\delta_{\pi}$  with risk constant in  $\theta$  we must have

$$R_{\pi}(\delta_{\pi}) = \int_{\Theta} R(\delta_{\pi}, \theta) \pi(\theta) d\theta = \sup_{\theta \in \Theta} R(\delta_{\pi}, \theta).$$

□



**Corollary.** If  $\delta_\pi$  has constant risk in  $\theta$ , then it is minimax, and if  $\delta_\pi$  is unique then it is unique minimax.

**Remark.** In such situations, the (unique) minimax decision rule is characterised by a (unique)  $\pi$ -Bayes rule corresponding to a least favourable prior  $\pi$ .

**Example.** In a  $\text{Bin}(n, \theta)$  model with  $\theta \in \Theta = [0, 1]$  and quadratic risk, consider priors  $\pi_{a,b}$  arising from the  $\text{Beta}(a, b)$  distribution. In this case, the unique  $\pi_{a,b}$ -Bayes rule equals the posterior mean  $\delta_{a,b}(X) = \mathbb{E}^{\pi_{a,b}}[\theta|X]$ , available in closed form as the mean of an 'updated' Beta distribution (see example sheet). One then solves in  $a, b$  the equation

$$R_{\pi_{a,b}}(\delta_{\pi_{a,b}}, \theta) = C, C \in \mathbb{R}$$

to obtain a unique Bayes rule  $\delta_{\pi_{\bar{a}, \bar{b}}}$  of constant quadratic risk. By what precedes, this gives the unique minimax rule for a  $\text{Bin}(n, \theta)$  model, which is seen to be distinct from the MLE and moreover, biased. One shows further that as  $n \rightarrow \infty$ , the minimax risk of  $\delta_{\pi_{a,b}}$  aligns with the risk of the MLE.

**Remark.** In a  $N_p(\theta, I)$  model,  $\theta \in \Theta = \mathbb{R}^p$ , we will show later that  $\overline{X}_n = \hat{\theta}_{\text{MLE}}$  is minimax, however.

## 9.2 Admissibility

**Definition** (Admissibility). In a decision problem, a decision rule  $\delta$  is called *inadmissible* if  $\exists \delta'$  such that

$$R(\delta', \theta) \leq R(\delta, \theta) \mid \theta \in \Theta$$

and

$$R(\delta', \theta) \not\leq R(\delta, \theta) \text{ for some } \theta.$$

$\delta$  is called *admissible* if no such  $\delta'$  exists. In the former case, we say that  $\delta'$  dominates  $\delta$ .

**Proposition.** (i) Every unique Bayes rule is admissible

(ii) If  $\delta$  is admissible and has risk constant in  $\theta$ , then it is minimax

*Proof.* See the example sheet. □

**Remark.** The unique minimax rule from the previous  $\text{Bin}(n, \theta)$  model is thus also admissible.

**Theorem** (A). Let  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} N(\theta, \sigma^2)$ ,  $\sigma^2$  is known,  $\theta \in \Theta = \mathbb{R}$ . Then  $\hat{\theta}_{\text{MLE}} = \overline{X}_n$  is admissible and minimax for quadratic risk.

**Remark.** Admissibility extends to  $p = 2$ , and minimaxity to any  $p \in \mathbb{N}$ , but this will not be proved.

However, for admissibility in  $p \geq 3$ .

**Theorem** (B). If  $X \sim N_p(\theta, I)$ ,  $\theta \in \Theta = \mathbb{R}^p$ ,  $p \geq 3$ , then  $\hat{\theta}_{\text{MLE}} = X$  is inadmissible for quadratic risk.

*Proof.* (Of A)

We can set  $\sigma^2 = 1$  as the proof will show. Also, the risk of  $\bar{X}_n$  equals

$$R(\bar{X}_n, \theta) = \mathbb{E}_\theta \left[ (\bar{X}_n - \theta)^2 \right] = \frac{1}{n},$$

which is constant in  $\theta$ , and hence it suffices to prove, by the previous proposition, that  $\bar{X}_n$  is admissible. It remains to prove admissibility. So let  $\delta$  be any other decision rule, then

$$\begin{aligned} R(\delta, \theta) &= \mathbb{E}_\theta [(\delta - \theta)^2] \\ &= \mathbb{E}_\theta [(\delta - \mathbb{E}_\theta [\delta])^2] + (\mathbb{E}_\theta [\delta] - \theta)^2 + 2 \underbrace{\mathbb{E}_\theta [\delta - \mathbb{E}_\theta [\delta]]}_{=0} \cdot (\mathbb{E}_\theta [\delta] - \theta) \\ &= \text{var}_\theta(\delta) + B^2(\theta), \end{aligned}$$

where  $B(\theta) = \mathbb{E}_\theta [\delta] - \theta$ . Recall the Cramer-Rao inequality for biased estimators, from the information geometry section, to the effect that

$$\text{var} \delta(X) \geq \frac{\left( \frac{d}{d\theta} \mathbb{E}_\theta [\delta] \right)^2}{nI(\theta)} = \frac{(1 + B'(\theta))^2}{n},$$

by definition of  $B$  and  $I(\theta)$ . Hence, if  $\delta$  dominates  $\bar{X}_n$ , the necessarily

$$\frac{1}{n} = R(\bar{X}_n, \theta) \geq R(\delta, \theta) \geq B^2(\theta) + \frac{(1 + B'(\theta))^2}{n} \quad \forall \theta \in \mathbb{R}.$$

We deduce that  $|B(\theta)| \leq \frac{1}{\sqrt{n}}$ , in particular  $B$  is bounded on  $\mathbb{R}$ . Moreover, we also have

$$(1 + B'(\theta))^2 = 1 + 2B'(\theta) + (B'(\theta))^2 \leq 1,$$

so  $B'(\theta) \leq 0 \quad \forall \theta \in \mathbb{R}$ .

$\forall \varepsilon > 0, i \in \mathbb{N}$  there must exist  $\theta_i$  large enough such that  $B'(\theta_i) \geq -\varepsilon$  as otherwise  $\forall |\theta| \geq \theta_i$  we would have  $B'(\theta) \leq -\varepsilon$  so that by the MVT  $B$  is unbounded ( $B(\theta) = B(\theta_i) + B'(\tilde{\theta})(\theta - \theta_i)$ , which contradicts the bounding. In other words, for these sequences,  $\pm\theta_i \xrightarrow{i \rightarrow \infty} \pm\infty$ , we have  $B'(\theta_i) \rightarrow 0$ . Now, evaluating these

limits in the inequality above, we see  $\lim_{i \rightarrow \infty} \left[ B^2(\theta_i) + \frac{(1 + B'(\theta_i))^2}{n} \leq \frac{1}{n} \right]$  implies  $\lim_{i \rightarrow \infty} B^2(\theta_i) \leq 0$  therefore by monotonicity of  $B$  we deduce that

$$B(-\infty) = B(\theta) = B(\infty) = 0,$$

and the bias vanishes identically, so that returning to the inequality we have proved

$$R(\delta, \theta) \geq \frac{1}{n} = R(\bar{X}_n, \theta) \quad \forall \theta \in \Theta.$$

□

**Remark.** (i) One shows (example sheet) that  $\bar{X}_n$  is not a  $\pi$ -Bayes rule  $\delta_\pi$  for any prior  $\pi$  on  $\Theta = \mathbb{R}$ , hence there exists an admissible minimax decision rule which is not  $\pi$ -Bayes for any  $\pi$ . One may show that  $\bar{X}_n$  is a 'limiting Bayes' in the sense that it is the limit as  $\nu \rightarrow \infty$  of the Bayes rule for a  $N(0, \nu^2)$ -prior on  $\Theta$ .

- (ii) The unboundedness of  $\Theta$  in the last proof is crucial. When  $\Theta = [0, \infty)$  then  $\bar{X}_n$  is inadmissible (it is still minimax, however), and when  $\Theta = [a, b], a < b$ , then  $\bar{X}_n$  is also not minimax any longer (see example sheet).
- (iii) Minimavity of  $\bar{X}_n$  on  $\Theta = \mathbb{R}$  extends to  $p \in \mathbb{N}$  i.e.  $X_i \stackrel{\text{iid}}{\sim} N_p(\theta, I), \theta \in \Theta = \mathbb{R}^p$ .

To prove theorem B, we first need a new estimator

**Definition** (James-Stein estimator). Define the *James-Stein estimator*, for  $X \sim N_p(\theta, I)$

$$\delta^{\text{JS}} = \begin{pmatrix} \delta_1^{\text{JS}} \\ \vdots \\ \delta_p^{\text{JS}} \end{pmatrix}, \delta_j^{\text{JS}} = \left(1 - \frac{p-2}{\|X\|^2}\right) X_j \quad (p \geq 3),$$

where  $\|v\|^2 = \sum_{j=1}^p v_j^2$ .

We now show that the risk of  $\delta^{\text{JS}}$  dominates the (quadratic) risk

$$R(\hat{\theta}_{\text{MLE}}, \theta) = \mathbb{E}_\theta [\|X - \theta\|^2] = \mathbb{E}_\theta \left[ \sum_{j=1}^p (X_j - \theta_j)^2 \right] = p.$$

**Lemma.** (Stein) Let  $X \sim N(\theta, 1), \theta \in \mathbb{R}$  and  $g : \mathbb{R} \rightarrow \mathbb{R}$  is differentiable and such that  $\mathbb{E}_\theta [|g(X)|] < \infty$ . Then  $\forall \theta$

$$\mathbb{E}_\theta [g(X)(X - \theta)] = \mathbb{E}_\theta [g'(X)].$$

*Proof.*

$$\begin{aligned} \mathbb{E}_\theta [g(X)(X - \theta)] &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} g(x)(x - \theta) e^{-\frac{(x-\theta)^2}{2}} dx \\ &= -\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} g(x) \left[ \frac{d}{dx} e^{-\frac{(x-\theta)^2}{2}} \right] dx \\ &= \left[ -\frac{1}{\sqrt{2\pi}} \right]_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} g'(x) e^{-\frac{(x-\theta)^2}{2}} dx \\ &= \mathbb{E}_\theta [g'(X)]. \end{aligned}$$

□

Now

$$\begin{aligned} R(\delta^{\text{JS}}, \theta) &= \mathbb{E}_\theta [\|\delta^{\text{JS}} - \theta\|^2] \\ &= \mathbb{E}_\theta \left[ \left\| \left(1 - \frac{p-2}{\|X\|^2}\right) X - \theta \right\|^2 \right] \\ &= \mathbb{E}_\theta \left[ \left\| X - \theta - \frac{p-2}{\|X\|^2} X \right\|^2 \right] \\ &= \mathbb{E}_\theta [\|X - \theta\|^2] + (p-2)^2 \mathbb{E}_\theta \left[ \frac{\|X\|^2}{\|X\|^4} \right] - 2(p-2) \mathbb{E}_\theta \left[ (X - \theta)^T \frac{X}{\|X\|^2} \right]. \end{aligned}$$

The last expectation can be written

$$\mathbb{E} \left[ \sum_{j=1}^p \mathbb{E}_j \left[ \frac{(X_j - \theta_j)X_j}{\|X\|^2} \right] \right],$$

where  $\mathbb{E}_j = \mathbb{E}[\cdot | X_{(-j)}]$ , with  $X_{(-j)} = \{X_i : i \neq j\}$ . The  $j$ th expectation can further be written as

$$\mathbb{E}_j [(X_j - \theta_j)g(X_j)], g(y) = \frac{y}{y^2 + a}, a = \sum_{i \neq j} X_i^2,$$

which (since  $\mathbb{P}(a = 0) = 0$ ) is a bounded, differentiable function, with

$$g'(y) = \frac{y^2 + a - 2y^2}{(y^2 + a)^2},$$

which is also bounded on  $\mathbb{R}$  so that Stein's Lemma applies to give that the last expectation is

$$\mathbb{E}_j [g'(X_j)] = \mathbb{E} \left[ \frac{X_j^2 + \sum_{i \neq j} X_i^2 - 2X_j^2}{\|X\|^4} \right].$$

When we insert this back into the risk gives

$$\begin{aligned} R(\delta^{\text{JS}}, \theta) &= \mathbb{E}_\theta [\|X - \theta\|^2] + (p-2)^2 \mathbb{E}_\theta \left[ \frac{1}{\|X\|^2} \right] - 2(p-2) \mathbb{E}_\theta \left[ \sum_{j=1}^p \frac{\|X\|^2 - 2X_j^2}{\|X\|^4} \right] \\ &= \mathbb{E}_\theta [\|X - \theta\|^2] + (p-2)^2 \mathbb{E}_\theta \left[ \frac{1}{\|X\|^2} \right] - 2(p-2) \mathbb{E}_\theta \left[ p \frac{1}{\|X\|^2} - 2 \frac{1}{\|X\|^2} \right] \\ &= \mathbb{E}_\theta [\|X - \theta\|^2] - (p-2)^2 \mathbb{E}_\theta \left[ \frac{1}{\|X\|^2} \right] \\ &< p. \end{aligned}$$

Since

$$\mathbb{E}_\theta \left[ \frac{1}{\|X\|^2} \right] = \int_{\mathbb{R}^p} \frac{1}{\|X\|^2} \phi(x - \theta) dx \geq c \int_{c_0 \leq \|X\| \leq c_1} \phi(x - \theta) dx \geq c P_{r_\theta}(\|X\| \in (c_0, c_1)) > 0,$$

for any  $\theta \in \mathbb{R}^p$  with  $\phi$  the pdf of  $N_p(\theta, I)$ .

**Remark.** (i) While  $\delta^{\text{JS}}$  strictly dominates  $\delta_{\text{MLE}} = X$ , the worst case (minimax) risk

$$\sup_{\theta \in \mathbb{R}^p} R(\delta^{\text{JS}}, \theta) = \sup_{\theta \in \mathbb{R}^p} R(X, \theta),$$

see the example sheet.

(ii) The James-Stein estimator itself is also inadmissible, and for instance, dominated by

$$\delta^{\text{JS}+} = \left(1 - \frac{p-2}{\|X\|^2}\right)^+ X$$

known as the 'positive part JS-estimator'.

- (iii) Other shrinkage factors are also permitted, but among the decision rules

$$\delta^{(c)} = (1 - c \frac{p-2}{\|X\|^2})X, c > 0,$$

the choice  $c = 1$  is optimal.

- (iv) While  $\delta^{\text{JS}}$  is attractive from a decision theoretic perspective, its use for inference (confidence revisions and test) is less clear, as its distributional properties are more involved than the one of  $\delta_{\text{MLE}}(X) = X \sim N_p(\theta, I)$

### 9.3 Classification problems