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

0 Introduction

Consider a random variable X defined on some probability space,

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

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

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

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

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

where f is the probability density function (pdf).
We typically only write $F(t) = P\left(X \leq t\right)$, 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 Θ 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) $\mathrm{Exp}(\theta), \ldots$

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 θ_0 for this specific 'true' value of θ. We say that observations $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P_\theta$ 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 $\hat{\theta}_n = \hat{\theta}(x_1, \ldots, x_n) \in \Theta$ that is close with high probability to θ when $x_1, \ldots, x_n \overset{\text{iid}}{\sim} P_\theta,\ \forall\, \theta \in \Theta$.

(ii) Hypothesis testing - For $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$, we want a test (indicator) function $\psi_n = \psi(x_1, \ldots, x_n)$ 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) $C_n = C(x_1, \ldots, x_n, \alpha) \subseteq \Theta$ of confidence in that

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

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

1 Likelihood Principle

Example. Consider a sample $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\theta)$ 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 occurance of $x_1, \dots, x_n$ as a function of θ is

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

a random function of θ.

Idea Maximise $L_n(\theta)$ over Θ, 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

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

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 $X_i \overset{\text{iid}}{P}$, 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*

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

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

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 ℓ_n or $\overline{\ell}_n$ over Θ.

Example. For 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

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

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

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_θ on $\chi \subseteq \mathbb{R}^d$ and let $g : \chi \to \mathbb{R}$ be given. We will write

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

which in the continuous case equals $\int_\chi g(x) f(x, \theta), \mathrm{d}x$, and in the discrete case is $\sum_{xinX} g(x) f(x_\theta)$

Observation Consider a model $\{f(\cdot, \theta) : \theta \in \Theta\}$ for X of law P on χ, and assume $\mathbb{E}_P [|\log f(x, \theta)|] < \infty$. Then $\overline{\ell}_n(\theta) = \frac{1}{n} \sum_{i=1}^n \log f(x_i, \theta)$ 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 θ_0 such that $P = P_{\theta_0}$, then we can rewrite

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

Next we write

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

.

Thus $\ell(\theta) \leq \ell(\theta_0) \ \forall \ \theta \in \Theta$, 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 $\frac{f(X,\theta)}{f(X,\theta_0)}$ = constant (in X), which since $\int_\chi f(x, \theta) \mathrm{d}x = 1$ can only happen when $f(\cdot, \theta) \overset{\text{almost surely}}{=} f(\cdot, \theta_0)$ identically.

Definition (Identifiable)**.** Let us thus say that the model is *identifiable* if $f(\cdot, \theta) = f(\cdot, \theta)(\text{a.s}) \iff \theta = \theta_0$. In this case, the function $\ell(\theta)$ has a unique maximiser at the true value θ_0.

The quantity

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

is called the Kullback-Leibler divergence (entropy-distance), which builds the basis of statistical information theory. In particular, the differential geometry of the maps $\theta \mapsto \mathrm{KL}(P_{\theta_0}, P_\theta)$ 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 $\int(\cdot)\mathrm{d}x$ integration.

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

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

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

Definition (Fisher information)**.** Let $\Theta \subseteq \mathbb{R}^p, \theta \in \mathrm{int}\Theta$, the 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 θ.

One shows:

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

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

Proof. As earlier we write

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

.

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

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

$\square$

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

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

(ii) If $X = (X_1, \ldots, X_n)$ consists of iid copies of X so that its pdf/pmf equals

$$f(x_1, \ldots, x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta).$$

then the Fisher information tensorises, that is

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

.

$I_1(\theta) = I(\theta)$ 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 $X_1, \ldots, X_n \overset{\text{iid}}{\sim}$ form a regular statistical model $\{f(\cdot, \theta) : \theta \in \Theta\}, \Theta \in \mathbb{R}$ and suppose $\tilde{\theta} = \tilde{\theta}(X_1, \ldots, X_n)$ is any unbiased estimator (i.e. $\mathbb{E}_\theta \left[\tilde{\theta}\right] = \theta \ \forall\ \theta \in \Theta$). Then $\forall\ \theta \in \operatorname{int}\Theta$

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

Proof. Assume wlog $\operatorname{var}_\theta \tilde{\theta} < \infty$, and consider first $n = 1$. Recall the Cauchy-Schwarz inequality to the effect that

$$\operatorname{Cov}^2(Y, Z) \leq \operatorname{var} Y \operatorname{var} Z.$$

For $Y = \tilde{\theta}$ and for $Z = \frac{\mathrm{d}}{\mathrm{d}\theta} \log f(X, \theta)$. Then $\mathbb{E}_\theta [Z] = 0$ by our observation above and by the preceeding remarks, $\mathbb{E}_\theta [Z] = \operatorname{var}_\theta Z = I(\theta)$. Thus by the Cauchy-Schwarz inequality.

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

Since

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

.

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

$$\mathbb{E}_\theta\left[g(X_1,\ldots,X_n)\right] = \int_{\mathcal{X}} g(x_1,\ldots,x_n)\prod_{i=1}^n f(x_i,\theta)\mathrm{d}x_1\cdots\mathrm{d}x_n.$$

and use that the Fisher information tensorises. □

Let us record also

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

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

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 $\Phi:\Theta\to\mathbb{R},\Theta\subseteq\mathbb{R}^p$. Then one shows that for any unbiased estimator $\tilde{\Phi}=\tilde{\Phi}(X_1,\ldots,X_n)$ for $\Phi(\theta)$, where $X_i\overset{\text{iid}}{\sim}\{f(\cdot,\theta):\theta\in\Theta\}$, we have

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

[Indeed, for $p=1$, the proof is the same, but replacing $\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}_\theta\left[\tilde{\theta}\right]=\frac{\mathrm{d}}{\mathrm{d}\theta}\theta=1$ by

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

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

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

so that

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

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

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

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 θ_1 and θ_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 θ_2 is unknown, then one can compute the 2×2 information matrix $I_2(\theta)$, and the CRLB for estimating θ_1 is, with $\Phi(\theta)=\theta_1$

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

One can see CRLB (I) ¡ CRLB (II) umless Σ 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 $n \to \infty$.The main goal will be to prove

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

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 $X_n \to X$ *almost surely*, $X_n \overset{\text{a.s.}}{\to} X$ as $n \to \infty$ if

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

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

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

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

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

Definition. -We say $X_n \to X$ *in distribution* (in law) writing $X_n \overset{\text{d}}{\to} X$ as $n \to \infty$, if

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

Recall $P(Z \le z) = P(Z_1 \le z_1, \dots, Z_k \le z_k$.
The following facts on stochastic convergence will be frequently used, and can be proved with measure theory.

Proposition. (i) $X_n \underset{n\to\infty}{\overset{\text{a.s}}{\to}} X \implies X_n \underset{n\to\infty}{\overset{P}{\to}} \implies X_n \underset{n\to\infty}{\overset{d}{\to}}$ but any converse is false in general.

(ii) (Continuous mapping theorem). If X_n, X take values in $\chi \subseteq \mathbb{R}^k$ and $g : \chi \to \mathbb{R}^d$ is continuous, then

$$X_n \underset{n\to\infty}{\to} X \text{ a.s / P / in law } \implies g(X_n) \underset{n\to\infty}{\to} g(X) \text{ a.s. / P / in law}$$

respectively.

(iii) (Slutsky's Lemma) Suppose $X_n \underset{n\to\infty}{\overset{d}{\to}} X, Y_n \overset{d}{\to} C, C$ is a constant (non-stochastic) then

– $Y_n \overset{P}{\to} C$ as $n \to \infty$

- $X_n + Y_n \xrightarrow{d} X + C$ as $n \to \infty$
- $X_n Y_n \xrightarrow{d} CX$ and provided $C \neq 0$, $X_n / Y_n \xrightarrow{d} X/C$ as $n \to \infty$
- 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 $n \to \infty$

(iv) If $X_n \xrightarrow{d} X$ as $n \to \infty$, then X_n is stochastically bounded ($Op(1)$), that is

$$\forall\ \epsilon > 0\ \exists\ M_\epsilon :\ \forall\ n \text{ large enough } \mathbb{P}\left(\|X_n\| > M_\epsilon\right) < \epsilon.$$

3.2 Law of large numbers and central limit theorem

Consider $X_1, X_2, \ldots$ 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

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

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 $X_1, X_2, \ldots$.

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

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

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

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

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

$$\frac{1}{n}\sum_{i=1}^{n} X_i : \underset{n\to\infty}{\to} \mathbb{E}\left[X\right] \quad P_r \text{ a.s. }.$$

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

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

$$\sqrt{n}(\overline{X})_n - \mathbb{E}\left[X\right]) \underset{n\to\infty}{\xrightarrow{d}} N(0, \sigma^2).$$

The multivariate version is also true. Recall that $X \in R^k$ 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 $\mu = \mathbb{E}\left[X\right]$ and $\Sigma = \mathrm{var}\, X$ (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 Σ 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 $A_n \xrightarrow{P_r} A$ are random matrices and $X_n \xrightarrow{d} N_k(\mu,\Sigma)$, then $A_nX_n \xrightarrow{d} N_d(A\mu, A\Sigma A^T)$. Lastly, Σ 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 $\operatorname{var} X = \Sigma$ positive definite (unnecessary). Then,

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

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 Prohkorov's Theorem,

Corollary.

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

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

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

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

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

$$\underset{n\to\infty}{\xrightarrow{\text{CLT}}} P_r(|Z| \le z_\alpha) = 1-\alpha.$$

by CLT, the continuous mapping theorem for $|\cdot|$ and because z_α 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 α). When σ is unknown, we replace it (in the definition of C_n) by S_n where

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^n \left(X - \overline{X}_i\right)^2$$

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

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

4 Consistency of MLEs

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

$$\tilde{\theta}_n \underset{n\to\infty}{\to} \theta \text{ in } (P_\theta^{\mathbb{N}})\text{-probability } \forall\, \theta \in \Theta.$$

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

(i) $f(x, \theta) > 0 \;\forall\, x \in \chi \;\forall\, \theta \in \Theta$.

(ii) $\int_\chi f(x, \theta)\mathrm{d}x = 1 \;\forall\, \theta \in \Theta$.

(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) $\mathbb{E}_\theta \left[\sup_{\theta\in\Theta} |\log f(x, \theta)|\right] < \infty \;\forall\, \theta \in \Theta$.

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

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

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

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

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

with respect to $\int (\cdot)\, \mathrm{d}P_\theta$ 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. Them a MLE exists and any MLE is consistent.

Proof. The map $\overline{\ell}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \log f(X_i, \theta)$ is continuous on the compact set $\Theta \in \mathbb{R}^p$ so by the Heine-Borel theorem, $\overline{\ell}_n$ obtains a maximum on Θ, 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 $\hat{\theta}_n \to \theta_0$ in probability as $n \to \infty$ (in $P = P_{\theta_0}^{\mathbb{N}}$-probability). The idea is that maximisers $\hat{\theta}_n$ of $\overline{\ell}_n$ over Θ should converge to the unique maximiser θ_0 of ℓ over Θ, since $\overline{\ell}_n(\theta) \underset{n\to\infty}{\overset{P}{\to}} \ell(\theta)$ by the law of large numbers for all $\theta \in \Theta$ pointwise. This is generally false unless one has uniform convergence

$$\sup_{\theta\in\Theta} |\overline{\ell}_n(\theta) - \ell(\theta)| \overset{P}{\to} 0 \text{ as } n \to \infty,$$

(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 Θ_ε, so

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

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

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

Now,

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

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

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

On these events we have

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

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

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

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 $P(A_n(\varepsilon)) \to 1$ and that $P(\|\hat{\theta} - \theta_0 < \varepsilon\| \to 1$ as $n \to \infty$ or $P(\|\theta_n - \theta_0\| \geq \varepsilon) \to 0$ as $n \to \infty$. Since ε was arbitrary, $\hat{\theta}_n \overset{P}{\underset{n\to\infty}{\theta_0}}$ and the proof is complete modulo the verification of the next lemma. □

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