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

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 $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\theta)$ with (unknown) $\theta > 0$. If the actual observed values are $X_1 = x_1, \ldots, X_n = x_n$, then the probability of this particular occurance of $x_1, \ldots, x_n$ as a function of θ is

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

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 $(x_i, i = 1, \ldots, n)$ 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*

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

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

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

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

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 $X_1, \ldots, X_n$ 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 $\sum_{x in X} g(x) f(x_\theta)$

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 $f(\cdot, \theta) = f(\cdot, \theta)$(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 $\forall\ \theta\in$ intΘ (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 $\forall\ \theta\in$ intΘ.

Definition (Fisher information)**.** Let $\Theta\subseteq\mathbb{R}^p,\theta\in$ intΘ, 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 $\forall\ \theta\in$ int$\Theta,\Theta\subseteq\mathbb{R}^p,p\geq1$,

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

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

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

□

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 $X = (X_1, \dots, X_n)$ 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

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

.

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

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

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

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

Since

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

.

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

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

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 $X_i\overset{\mathrm{iid}}{\sim}\{f(\cdot,\theta):\theta\in\Theta\}$, 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 $\tilde{\theta}(X_1, \ldots, X_n)$ 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 $n \to \infty \ \forall \ j = 1, \ldots, k$.

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

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

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

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

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

(ii) (Continuous mapping theorem). If X_n, X take values in $\chi \subseteq \mathbb{R}^k$ and $g : \chi \to \mathbb{R}^d$ 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 $\operatorname{var}(X) < \infty$ (unnecessary) by Chebyshev,

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

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

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 $\operatorname{var} X < \infty$, 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 $\operatorname{var} X = \sigma^2 < \infty$. 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 $\Sigma = \operatorname{var} X$ (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 $A_n X_n \xrightarrow{d} N_d(A\mu, A\Sigma A^T)$. Lastly, Σ is diagonal $\implies$ the components of X are independent.

Theorem (Multivariate central limit theorem)**.** Let $X_1, \ldots, X_n$ 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 $Y_1, Y_2, \ldots$ and $c_1, c_2, \ldots \in \mathbb{R}\setminus\{0\}$.

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

By Prohkorov's Theorem,

Corollary.

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

Let $k = 1$, $X_1, \ldots X_n$ 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 $P_r(|Z| \le z_\alpha) = 1-\alpha$, $Z \sim N(0,1)$
$P_{\mu_0} = P$,

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

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

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 $X_1, \dots, X_n \overset{\text{iid}}{\sim}$ form a statistical model $\{P_\theta : \theta \in \Theta\}, \Theta \subseteq \mathbb{R}^p$ then we say that an estimator $\tilde{\theta}_n = \tilde{\theta}(X_1, \dots, X_n)$ 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.

To verify our claim we now make the following digression. For a (measurable) $\chi \subseteq \mathbb{R}^d$ and a (measurable) $h : \chi \to \mathbb{R}$, and let $X_1, \ldots, X_n$ be iid random variables in χ 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

$$\frac{1}{n} \sum_{i=1}^{n} h(X_i) \underset{n\to\infty}{\to} 0 \text{ a.s. } (P^{\mathbb{N}}).$$

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

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

Moreover,

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

Since

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

To transfer to an infinite collection of h 's, let us say that a family of brackets

$$[\underline{h_j}, \overline{h_j}], \underline{h_j}, \overline{h_j} : \chi \to \mathbb{R}, j = 1, \dots, N$$

covers a class $\mathcal{H}$ of maps on χ 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) $\mathbb{E}\left[\underline{h_j}(X)\right] < \infty, \mathbb{E}\left[\overline{h_j}(X)\right] < \infty$

(ii) $\mathbb{E}\left[\overline{h_j}(X) - \underline{h_j}(X)\right] < \epsilon$

Then

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

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

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

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 $[\underline{h_j}, \overline{h_j}] \ni h$

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

Likewise we get that

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

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

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

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 $\mathbb{E}\left[\sup_{\theta \in \Theta} |q(X, \theta)|\right] < \infty$. If $X_1, \ldots, X_n$ are iid copies of X

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

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

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

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

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

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 $X_1, \ldots, X_n \overset{\text{iid}}{\sim}$ from the model we have

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \underset{n\to\infty}{\overset{d}{\to}} N(0, I(\theta_0)^{-1}).$$

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

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

So

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

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). $Pr = P_{\theta_0}^{\mathbb{N}}$, $\mathbb{E} = \mathbb{E}_{\theta_0}$

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 $Z_n \overset{d}{\to} Z$) it suffices to restrict to any sequence E_n of events (in $\mathbb{R}^{\mathbb{N}}$) such that $Pr(E_) \to 1$. Indeed,

$$|Pr(Z_n \le t) - Pr(Z_n \le t, E_n)| \le Pr(E_n^c) \underset{n\to\infty}{\to} 0.$$

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

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

where we recall that $\overline{\ell}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \log f(X_i,\theta)$ and $\frac{\partial}{\partial\theta}\overline{\ell}_n(\hat{\theta}_n) = \frac{\partial}{\partial\theta}\overline{\ell}_n(\theta)\Big|_{\theta=\hat{\theta}_n}$. For any map $h : U \to \mathbb{R}$ 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 θ_0 and write

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

where $\overline{\theta} = \overline{\theta}(n)$ is some mean value on that line segment. Second derivatives of $\overline{\ell}_n(\theta)$ 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 $\frac{\partial}{\partial\theta_j}\overline{\ell}_n(\hat{\theta})$ we obtain

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

where $\overline{\theta_{(j)}}$ is the $p \times 1$ vector arising from the jth application of the MVT. We will show

$$\overline{A}_n \underset{n\to\infty}{\overset{P}{\to}} -I(\theta_0).$$

In particular, this implies convergence of

$$\|\overline{A}_n - (-I(\theta_0))\| \underset{\text{operator norm}}{\overset{P}{\to}} 0.$$

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

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

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

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

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

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

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

$$|I| = \left|\frac{1}{n}\sum q(X_i,\overline{\theta}_{(j)}) - \mathbb{E}\left[q(X,\overline{\theta}_{(j)}\right]\right| \le \sup_{\theta\in K}\left|\frac{1}{n}\sum_{i=1}^{n}q(X_i,\theta) - \mathbb{E}\left[q(X,\theta)\right]\right| \underset{n\to\infty}{\overset{P}{\to}} 0$$

by the uniform law of large numbers.
For II we notice that $\hat{\theta}_n \overset{P}{\to} \theta_0 \implies \overline{\theta}_{(j)} \overset{P}{\to} \theta_0$ as $n\to\infty$ $\forall$ j, and since $\theta \mapsto \mathbb{E}\left[q(X,\theta)\right]$ is continuous the continuous mapping theorem implies

$$\text{II} = \mathbb{E}\left[q(X,\overline{\theta}_{(j)}\right] - \mathbb{E}\left[q(X,\theta)\right] \underset{n\to\infty}{\overset{P}{\to}} 0,$$

completing the proof. □

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 θ_0 lies at the boundary of Θ, 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

$$\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta) \underset{n\to\infty}{\overset{d}{\to}} N(0, I(\theta)^{-1}) \ \forall\ \theta \in \ \text{int}\Theta.$$

[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, $\hat{\theta}_n \underset{n\to\infty}{\overset{P}{\to}} \theta$] 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_θ

.

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

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

$$\sqrt{n}(\tilde{\theta} - \theta) \underset{n\to\infty}{\overset{d}{\to}} N(0, I(\theta)^{-1}).$$

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

.

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 $\Phi : \Theta \to \mathbb{R}^k, \Theta \subseteq \mathbb{R}^p$ based on $X_i \overset{\text{iid}}{\sim} \{f(\cdot, \theta) : \theta \in \Theta\}$ where $\hat{\theta}$ is the MLE for θ. 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 Φ is differentiable.

Theorem (Delta-method)**.** Suppose $\Phi : \Theta \to \mathbb{R}$ 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 Θ such that $\sqrt{n}(\hat{\theta}_n - \theta) \underset{n\to\infty}{\overset{d}{\to}} Z$ where Z is some random vector in $\mathbb{R}^p$. Then

$$\sqrt{n}(\Phi(\hat{\theta}_n) - \Phi(\theta)) \underset{n\to\infty}{\overset{d}{\to}} \frac{\partial \Phi}{\partial \theta}(\theta)^T Z.$$

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

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

Since $\sqrt{n}(\hat{\theta}_n - \theta) \underset{n\to\infty}{\overset{d}{\to}} Z$ we have in particular $\hat{\theta}_n \underset{n\to\infty}{\overset{p}{\to}} \theta$ (by stochastic boundedness) so also $\overline{\theta}_n \underset{n\to\infty}{\overset{P}{\to}} \theta$ hence by the continuous mapping theorem, we also have

$$\frac{\partial \Phi}{\partial \theta}(\overline{\theta}_n) \overset{P}{\to} \frac{\partial \Phi}{\partial \theta}(\theta),$$

hence by Slutsky's lemma, the last displayed expression $\underset{n\to\infty}{\overset{d}{\to}} \frac{\partial \Phi}{\partial \theta}(\theta)^T Z$ □

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

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

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 θ_i the ith 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 ith position, and the last theorem

$$\sqrt{n}(\hat{\theta}_i - \theta_i) = \sqrt{n}e_i^T(\hat{\theta} - \theta) \underset{n\to\infty}{\overset{d}{\to}} N(0, e_i^T I(\theta)^{-1} e_i).$$

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

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

Indeed, by the continuous mapping theorem,

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

So C_n is a confidence interval of asymptotic level $1-\alpha$. In practice, $I(\theta)$ may still depend on θ and hence needs to be replaced by a consistent estimate $\hat{\imath}_n \underset{n\to\infty}{\overset{P}{\to}} I(\theta)$ (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 θ. We then define $\hat{\imath} = i_n(\hat{\theta}_{\text{MLE}})$, 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

$$\hat{\imath}_n \overset{P}{\to} I(\theta_0) \text{ as } n \to \infty.$$

Proof. Just as when proving $\overline{A}_n \underset{n\to\infty}{\overset{P}{\to}} I(\theta_0)$ 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. □

Remark. The continuous mapping theorem and invertability of $I(\theta_0)$ then also imply that $\hat{\imath}_n^{-1} \underset{n\to\infty}{\overset{P}{\to}} I(\theta_0)^{-1}$, since $A \mapsto A^{-1}$ is continuous on $\{A : \det A \neq 0\}$. Alternatively, one uses

$$j_n(\theta) = -\frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\theta\partial\theta^T} \log f(X_j,\theta)$$

and estimate $I_n(\theta_0)$ by $\hat{\jmath} = j_n(\hat{\theta}_{\text{MLE}})$ which as before (and by an observation in information geometry) satisfies $\hat{\jmath} \underset{n\to\infty}{\overset{P}{\to}} I(\theta_0)$ as well.

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

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

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

$$W_n(\theta) \underset{n\to\infty}{\overset{d}{\to}} \chi^2_p$$

and this entails that the confidence ellipsoid $C_n = \{\theta \in \mathbb{R}^p : W_n(\theta) \leq \Xi_\alpha\}$ has asymptotic coverage $\lim_{n\to\infty} P_\theta(\theta \in C_n) = 1 - \alpha$ if Ξ_α are the $1-\alpha$ quantiles of χ^2_p 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 $\psi_n = \psi(X_1, \ldots, X_n)$ 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*

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

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_{θ_0},

$$\Lambda_n(\Theta, \Theta_0) \overset{d}{\to} \chi^2_p, p = \dim \Theta.$$

Remark.
- One may show more generally, that for $\dim(\Theta_0) = p_0 > 0$ but $< p$ we have
$$\Lambda_n(\Theta, \Theta_0) \underset{n\to\infty}{\overset{d}{\to}} \chi^2_{p-p_0} \text{ under } P_\theta, \theta \in \Theta.$$
- 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 α if Ξ_α are the α-quantiles of a $\chi^2_{p-p_0}$-distribution.

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

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

where $\overline{\theta}_n$ are mean values lying on the lines 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

$$\sqrt{n}(\hat{\theta}_n - \theta_0)^T \underbrace{(\hat{\jmath}(\overline{\theta}) - I(\theta_0))}_{\xrightarrow{P} \text{ previous prop}} \underbrace{\xrightarrow{d} N}_{\sqrt{n}(\hat{\theta}_n - \theta_0)} + \sqrt{n}(\hat{\theta}_n - \theta_0)^T I(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0).$$

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 $\mathbb{R}^p \to \mathbb{R}$. Moreover, by standard linear algebra

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

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 θ as drawn at random from some *prior* probability distribution π on Θ. This may

(i) Model an intrinsically random state of nature Θ.

(ii) Model subjective beliefs about the state of nature Θ.

(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

$$P\left(X = x|H_i\right) = f_i(x),$$

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

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

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

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

If all π_i agree (Θ 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 Θ, where π 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

$$\mathrm{d}Q(x,\theta) = f(x,\theta)\pi(\theta)\mathrm{d}x\mathrm{d}\theta.$$

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

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

Likewise

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

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

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

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

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

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

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 → Beta posterior
- Gamma prior + Poisson sampling → Gamma posterior

Even when π is not a proper probability distribution the expression

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

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 (σ^2 known) one see that the Jeffrey's prior is $\propto$ constant, and one shows that the 'improper' posterior equals a $N(\overline{X}_n, \frac{\sigma^2}{n} = N(\hat{\theta}_{\text{MLE}}, \frac{\sigma^2}{n})$, see the example sheet. Note however (see the example sheet) that uniform priors to 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 $\pi(\cdot|X_1, \ldots, X_n)$ is a (random) probability distribution on Θ and hence can be used im principle to construct inference procedures for θ.

(i) Estimation - One may use the posterior mean $\mathbb{E}^\pi[\theta|X_1, \ldots, X_n]$ as an estimator $\overline{\theta}_n = \overline{\theta}(X_1, \ldots, X_n$ for θ or alternatively, when appropriately defined, the posterior mode or median.

(ii) Uncertainty quantification - Any subset $C_n \subseteq \Theta$ for which $\pi(C_n|X_1,\ldots,X_n) = 1-\alpha$ 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

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

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

8.2 Frequentist analysis of Bayes methods

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

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

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

One shows easily that $\overline{\theta}_n = \frac{1}{n+1}\sum_{i=1}^n X_i \to \theta_0$ $P_{\theta_0}^{\mathbb{N}}$-a.s., and also that $\sqrt{n}(\overline{\theta}_n - \theta) \overset{d}{\to} N(0, I(\theta_0)^{-1})$ 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 $\pi(\cdot|X_1,\ldots,X_n)$.

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 π near θ_0. Denote by $\pi_n = \pi(\cdot|X_1,\ldots,X_n)$ and let $\hat{\phi}_n$ be the pdf of a $N(\hat{\theta}_{\text{MLE}}, \frac{1}{n}I(\theta_0)^{-1})$. Then as $n \to \infty$

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

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

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

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

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

Where we substituted $v = \sqrt{n}(\theta - \hat{\theta})$ with $\frac{\mathrm{d}v}{\mathrm{d}\theta} = \sqrt{n}$. Fix $w \in \Omega_0 \subseteq \Omega$ such that $P_{\theta_0}^{\mathbb{N}}(\Omega_0) = 1$, so that $\sqrt{n}(\hat{\theta}(\omega) - \overline{\theta}(\omega) \underset{n\to\infty}{\to} 0$ (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 $\to 0\ \forall\ \omega \in \Omega_0$ since $P_{\theta_0}^{\mathbb{N}}(\Omega_0) = 1$ the limit holds a.s. □

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

$$\sqrt{n}(\hat{\theta}_{\mathrm{MLE}} = \overline{\theta}_n) \underset{n\to\infty}{\to} 0\ P_{\theta_0}^{\mathbb{N}}\text{ a.s. },$$

typically permitting the alternative centering at $\overline{\theta} = \mathbb{E}^{\pi}\left[\theta | X_1, \ldots, X_n\right]$

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

Example. Consider

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

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

Recall $\hat{\phi}_n$ was the pdf of $N(\hat{\theta_{\mathrm{MLE}}}, \frac{1}{n}I(\theta_0)^{-1})$ distribution, and define further ϕ_0 to be the pdf of $Z \sim N(0, I(\theta)^{-1})$. We can define $\Phi(t) = Pr\left(|Z| \leq t\right) = \int_{-t}^{t}\phi_0(v)\mathrm{d}v$ which is strictly increasing in t and also continuously differentiable, hence admits a continuous inverse $\Phi^{-1} : [0, 1] \to \mathbb{R}$. Now we can write

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

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

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

But we know that the first term $\underset{n\to\infty}{\to} 0$ $P_{\theta_0}^{\mathbb{N}}$-a.s. So by the BvM theorem we know that $\Phi(R_n) \underset{n\to\infty}{\to} 1-\alpha$ a.s. under P_{θ_0} , hence applying the continuous mapping theorem to Φ^{-1} we deduce

$$R_n = \Phi^{-1}(\Phi(R_n)) \overset{\text{a.s.}}{\underset{n\to\infty}{\to}} \Phi^{-1}(1-\alpha) \text{ under} P_{\theta_0}.$$

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

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

Now

.

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

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 π with positive continuous density functions on Θ 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 $\delta : \chi \to \mathcal{A}$ 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) $\mathcal{A} = \{$ 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*

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

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

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

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

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

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

equals the mean squared error (MSE).

Definition. Given a prior π on Θ, the π-Bayes risk of a decision rule δ is

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

A π-Bayes decision rules δ_π is any decision rule for which $R_\pi(\delta)$ is minimised in δ.

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

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

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

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

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

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

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

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

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

Proof. Recall the 'tower property' of iterated expectations

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

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

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

and likewise

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

Now

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

□

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

Example. (i) $\overline{X}_n = \hat{\theta}_{\mathrm{MLE}}$ 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 $\mathrm{Bin}(n,\theta), \theta \in \Theta = [0,1]$ is a π-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

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

Definition (Least favourable)**.** A prior λ on Θ is called *least favourable* (for a decision problem) if

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

Proposition. In a decision problem suppose for some prior π on Θ, we have that

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

(the π-Bayes risk of δ_π coincides with the worst case risk of δ_π).
Then,

(i) δ_π is minimax.

(ii) If δ_π is the unique Bayes rule, then δ_π is the unique minimax.

(iii) π is least favourable.

Proof. (i) Let δ be any decision rule with maximal risk

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

So taking inf over all δ we see

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

hence δ_π attains the minimax risk.

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

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

hence δ_π is the unique minimax.

(iii) Let π' be any prior on Θ then

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

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

The corollary also follows since for δ_π with risk constant in θ we must have

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

□

Corollary. If δ_π has constant risk in θ, then it is minimax, and if δ_π is unique then it is unique minimax.

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

Example. In a $\mathrm{Bin}(n,\theta)$ model with $\theta \in \Theta = [0,1]$ and quadratic risk, consider priors $\pi_{a,b}$ arising from the $\mathrm{Beta}(a,b)$ 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 $\delta_{\pi_{\overline{a},\overline{b}}}$ of constant quadratic risk. By what precedes, this gives the unique minimax rule for a $\mathrm{Bin}(n,\theta)$ model, which is seen to be distinct from the MLE and moreover, biased. One shows further that as $n \to \infty$, 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 $\overline{X_n} = \hat{\theta}_{\mathrm{MLE}}$ is minimax, however.

9.2 Admissibility

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

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

and

$$R(\delta', \theta) \lneq R(\delta, \theta) \text{ for some } \theta.$$

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

Proposition. (i) Every unique Bayes rule is admissible

(ii) If δ is admissible and has risk constant in θ, then it is minimax

Proof. See the example sheet. □

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

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

Proof. (Of A)
We can set $\sigma^2 = 1$ as the proof will show. Also, the risk of $\overline{X}_n$ equals

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

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

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

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

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

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

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

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 $B'(\theta) \leq 0\ \forall\ \theta \in \mathbb{R}$.
$\forall\ \varepsilon > 0, i \in \mathbb{N}$ there must exist θ_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, $\pm\theta_i \underset{i\to\infty}{\to} \pm\infty$, we have $B'(\theta_i) \to 0$. Now, evaluating these limits in the inequality above, we see $\lim_{i\to\infty} \left[B^2(\theta_i) + \frac{(1+B'(\theta_i))^2}{n} \leq \frac{1}{n} \right]$ implies $\lim_{i\to\infty} B^2(\theta_i) \leq 0$ 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

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

□

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

(ii) The unboundedness of Θ in the last proof is crucial. When $\Theta = [0, \infty)$ then $\overline{X}_n$ is inadmissible (it is still minimax, however), and when $\Theta = [a, b], a < b$, then $\overline{X}_n$ is also not minimax any longer (see example sheet).

(iii) Minimaxity of $\overline{X}_n$ on $\Theta = \mathbb{R}$ extends to $p \in \mathbb{N}$ i.e. $X_i \overset{\text{iid}}{\sim} N_p(\theta, I), \theta \in \Theta = \mathbb{R}^p$.

To prove theorem B, we first need a new estimator

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

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

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

We now show that the risk of δ^{JS} dominates the (quadratic) risk

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

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

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

Proof.

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

□

Now

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

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

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

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

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

When we insert this back into the risk gives

.

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

Since

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

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

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

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

see the example sheet.

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

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

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

9.3 Classification problems

Fore two pdf/pmf $f_i i = 1, 2$ defined on χ, consider observing X drawn from f_1 with probability q_1 and from f_2 with probability $q_2 = 1 - q_1$. Given an outcome $X = x$, we wish to classify it into the correct category i. This can be cast as a binary decision problem with $\Theta = \{1, 2\}$ and prior π on $\{q_1 q_2\}$. A classification rule $\delta = \delta_R$ is given by a region $\mathcal{R} \subseteq \chi \subseteq \mathbb{R}^d$ such that

$$\delta_r(X) = \begin{cases} 1 & \text{if } X \in \mathcal{R} \\ 2 & \text{if } X \in \mathcal{R}^c = \chi \setminus R \end{cases}.$$

The classification errors are given by

$$P(2|1, \mathcal{R}) = P_{f_1}(X \in \mathcal{R}^c) = \int_{\mathcal{R}^c} f_1(x)\mathrm{d}x,$$

and

$$P(1|2, \mathcal{R}) = P_{f_2}(X \in \mathcal{R}) = \int_{\mathcal{R}} f_2(x)\mathrm{d}x.$$

The π-Bayes risk now becomes

$$R_\pi(\delta_R) = P(1|2, \mathcal{R})\,\pi(2) + \mathbb{P}(2|1, \mathcal{R})\,\pi(1),$$

where π is a fixed 'prior' assigning the probabilities q_1, q_2. The Bayes-factors are given by

$$\frac{\pi(1|X = x)}{\pi(2|X = x)} = \frac{f_1(x)q_1/\cdot}{f_2(x)q_2/\cdot} = \frac{f_1(x)q_1}{f_2(x)q_2}$$

and the π-Bayes classifier can be shown to choose $\{1\}$ whenever $\frac{f_1(x)}{f_2(x)} = \frac{q_2}{q_1}$.

Proposition. Suppose for all i the $P_{f_i}\left(\frac{f_1}{f_2}(x) = \frac{1-q}{q}\right) = 0$ and $\delta = \delta_R$ arises from the classification region

$$\mathcal{R} = \left\{x \in \chi : \frac{f_1(x)}{f_2(x)} > \frac{1 - q_1}{q_1}\right\},$$

then δ_R is the unique π-Bayes classification rule for prior $\pi(\{1\}) = q$, in particular, δ_R is admissible.

Proof. Let $S \subseteq \chi$ be any other classification region with classification risk

$$q\int_{S^c} f_1(x)\mathrm{d}x+(1-q)\int_S f_2(x)\mathrm{d}x = \int_{S^c}(qf_1(x)-(1-q)f_2(x))\mathrm{d}x+\int_\chi (1-q)f_2(x)\mathrm{d}x.$$

The first term is minimal when S^c includes precisely all $x \in \chi$ such that that integrand $qf_1(x) - (1-q)f_2(x) < 0 \iff S^c = \{x : \frac{f_1(x)}{f_2(x)} < \frac{1-q}{q}\}$. Uniqueness follows since $P\left(qf_1(x) - (1-q)f_2(x) = 0\right) = 0$. Thus $S = R$ and the first claim follows since unique Bayes rules are admissible, the result is proved. □

Similarly, one can find minimax classifiers δ_R by choosing $\mathcal{R}$ such that

$$qP(2|1,\mathcal{R}) + (1-q)P(1|2,\mathcal{R}) = C,\ q \in [0,1].$$

For the case where $f_i \sim N_p(\mu_i, \Sigma)$, these classifiers can be explicitly computed, in dependence of the discriminant function $D = X^T\Sigma(\mu_1 - \mu_2)$ see the example sheet for details.

10 Further topics

10.1 Basic multivariate analysis

Consider random vectors X, Y drawn i.i.d. from $N(\mu_x, \sigma_x), N(\mu_y, \sigma_y)$, then their correlation is

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{var}\, X}\sqrt{\mathrm{var} Y}},$$

which for data $X_1, \ldots, X_n; Y_1, \ldots, Y_n$ jointly i.i.d. Can be estimated by the empirical correlation

$$\hat{\rho}_{X,Y} = \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_i)(Y_i - \overline{Y}_n)}{\sqrt{S_{n,x}}\sqrt{S_{n,y}}}, S_{n,x} = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2.$$

The finite sample distribution of $\hat{\rho}_{X,Y}$ can in principle be computed analytically; it equals the densities

$$f_{\hat{\rho}_{X,Y}}(r) \propto (1-r^2)^{\frac{1}{2}(n-4)}, \ -1 \leq r \leq 1.$$

Alternatively, we can regard $\rho_{X,Y} = \Phi(\theta)$ with samples from $N(\begin{pmatrix}\mu_x \\ \mu_y\end{pmatrix}, \Sigma)$. One shows that $\hat{\rho} = \Phi(\hat{\theta}_{\mathrm{MLE}})$ and by the general theory developed earlier (plus Delta method for Φ) we deduce that

$$\sqrt{n}(\hat{\rho}_{X,Y} - \rho) \xrightarrow[n\to\infty]{d} N(0, \mathrm{var}(\Phi, \theta))$$

for some asymptotic variance $\mathrm{var}(\Phi, \theta)$. More generally, consider a partitioned random vector $X \in \mathbb{R}^p$

$$X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}, X^{(1)} \in \mathbb{R}^q, X^{(2)} \in \mathbb{R}^{p-q}, \ X \sim N_p(0, \Sigma),$$

where Σ is a $p \times p$ positive definite matrix. The conditional covariance of $X^{(1)}|X^{(2)}$ equals

$$\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12} := \Sigma_{11\cdot 2}, \text{ where } \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

One further defines the partial correlations of $X_i, X_j, i \neq j$, $i, j \leq q$, as

$$\rho_{ij\cdot 2} = \frac{(\Sigma_{11\cdot 2})_{ij}}{\sqrt{(\Sigma_{11\cdot 2})_{ii}\,(\Sigma_{11\cdot 2})_{jj}}}.$$

Here likewise, the plug-in MLE $\hat{\rho}_{ij\cdot 2}$ equals the "empirical partial correlation coefficient" equal to $\Phi(\hat{\Sigma}_{\mathrm{MLE}})$ where

$$\left(\hat{\Sigma}_{\mathrm{MLE}}\right)_{ij} = \frac{1}{n}\sum_{m=1}^n (X_m X_m^T)_{ij},$$

where Φ is again a smooth map on $\Theta = \{\Sigma \text{ p.d.}\}$. The previous theory again implies that

$$\sqrt{n}(\hat{\rho}_{ij\cdot 1} - \rho_{ij\cdot 2}) \xrightarrow[n\to\infty]{d} N(0, ?).$$

Finally, consider $X \sim N(0, \Sigma)$, again with $\Sigma, p \times p$ positive definite, symmetric. Then there exists an orthonormal matrix T (i.e. $T^T = T^{-1}$) such that

$$\Sigma = T^T \Lambda T, \text{ where } \Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix}, \lambda_1 > \lambda 2 > \cdots > \lambda_p > 0.$$

In this case, we can define $U = TX \sim N(0, T\Sigma T^T) = N(0, \Lambda)$ and the entries of the random vector U in $\mathbb{R}^p$ are called the principal components corresponding to principal subspaces spanned by the columns vectors of T, arranged in decreasing order of 'explained variance'. Here again, if

$$\left(\hat{\Sigma}_{\text{MLE}}\right)_{ij} = \frac{1}{n} \sum_{m=1}^{n} (X_m X_m^T)_{ij},$$

then the plug-in MLE will give asymptotically efficient estimators for λ_i, μ_i provided there are no multiplicities of λ_i's.

10.2 Monte Carlo methods

We will discuss algorithms to generate random samples from given probability distributions, useful to construct numerical approximations of inference methods (posterior distributions and other examples). One can generate, on a pseudo-random generator (see CS), random samples $U_1, \ldots, U_n \overset{\text{iid}}{\sim} U[0,1], N \in \mathbb{N}$. If F is a cdf of some random variable X, with quantile transform $F^-(u) = \inf\{x : u \leq F(x)\}$, then one shows that $X_i^* = F^-(U_i)$ are iid random variables with distribution F. For normal random variables, F^- is not available in closed form, but one can still simulate $X \sim N_2(0, I)$ starting from $U_1, U_2 \overset{\text{iid}}{\sim} U(0,1)$ by the Box-Mueller transformation (see example sheet). More elaborate MC-algorithms arise from *importance sampling* e.g. as in the following 'ACCEPT-REJECT' algorithm, where we want to sample from some pdf on X and have another density h that we can sample from, such that $f(x) \leq Mh(x) \ \forall \ x \in \chi$, some $M > 0$.

- Step I: Draw $X \sim h$ and $U \sim U[0,1]$
- Step II: Set $Y = X$ if $U \leq \frac{f(X)}{Mh(X)}$ otherwise return to step I

One shows (example sheet) that $Y \sim f$ on χ. In the preceding settings, if we can generate $(X_i^* \ i = 1, \ldots, N)$ iid from a fixed distribution ρ_R then we can numerically approximate integrals

$$\mathbb{E}_{\rho_F}[g(X)] = \int_\chi g(x) \mathrm{d}\rho_F(x), \ g \text{ given}$$

by the MC-average

$$\frac{1}{N} \sum_{i=1}^{N} g(X_i^*) \underset{\text{SLLN}}{\overset{N \to \infty}{\to}} \mathbb{E}_{\rho_F}[g(X)].$$

10.2.1 Markov chain Monte Carlo (MCMC) algorithms

A discrete time Markov chain $(\vartheta_m : m \in \mathbb{N})$ started at ϑ_0 is a sequence of random variables whose 'transition probabilities' are of the form

$$Pr(\vartheta \in B|\vartheta_{m-1}, \vartheta_{m-2}) = Pr\left(\vartheta \in B|\vartheta_{m-1} = t\right) = Pr\left(\vartheta_1 \in B|\vartheta_0 = t\right) = K(t, B),$$

where $B \subseteq \Theta$ (measurable) and where K is a *Markov kernel* such that $\forall\ t$, $K(t, \cdot)$ is a probability distribution on Θ (= state space of (ϑ_m)). A pdf / pmf on Θ is called invariant for K if

$$\int_\Theta K(t, B)\mu(t)\mathrm{d}t = \int Pr\left(\vartheta_1 \in B|\vartheta_0 = t\right)\mu(t)\mathrm{d}t = \mu(B).$$

Under additional hypotheses (ergodicity of (ϑ_m)) one shows that the distribution of ϑ_m 'mixes towards' (converges to) its invariant measure μ, and we can then use MC-averages $(\vartheta_m : m = 1, \ldots, N)$ to approximate $\mathbb{E}_\mu\left[g(X)\right]$ by $\frac{1}{N}\sum_{m=1}^N g(\vartheta_m)$. Am important MCMC method, known as the *Metropolis-Hastings* algorithm requires an auxiliary (conditional) pdf $q(\cdot, t), t \in \Theta$ we can sample from, and proceeds as follows.

- Step I: Given $m \in \mathbb{N}, \vartheta_m \in \Theta$, generate $S_m \sim q(\cdot|\vartheta_m)$
- Step II: Set

$$\vartheta_{m+1} = \begin{cases} s_m & \text{with probability } \rho(\vartheta_m, s_m) \\ \vartheta_m \text{ with probability } 1 - \rho(\vartheta_m, s_m) & \end{cases},$$

  where

$$\rho(t, s) = \min\left(1, \frac{\mu(s)}{\mu(t)}\frac{q(t|s)}{q(s|t)}\right),$$

  and μ a pdf / pmf.

Proposition. For the above Markov chain, assuming $\mu, q(\cdot|t)$ are strictly positive on Θ, the invariant measure of $(\vartheta_m : m \in \mathbb{N})$ equals μ.

An invariant measure is μ, uniqueness requires Probability and Measure. The transition Markov kernel K has 'density' k

$$k(t, s) = \rho(t, s)q(s|t) + (1 - r(t))\mathrm{d}\delta_t(s), \quad s, t \in \Theta,$$

where $\delta_t(s)$ is ("Dirac") point mass at t, where $r(t) = \int_\Theta \rho(t, \tau)q(\tau|t)\mathrm{d}\tau$. Next we have,

$$\begin{aligned}
\rho(t, s)q(s|t)\mu(t) &= \min\left(q(s|t)\mu(t), \frac{\mu(s)q(t|s)}{\mu(t)q(s|t)}q(s|t)\mu(t)\right) \\
&= \min\left(q(s|t)\mu(t), \mu(s)q(t|s)\right) \\
&= \min\left(\frac{q(s|t)\mu(t)}{q(t|s)\mu(s)}q(t|s)\mu(s), \mu(s)q(t|s)\right) \\
&= q(t|s)\mu(s)\min\left(\frac{q(s|t)\mu(t)}{q(t|s)\mu(s)}, 1\right) \\
&= \rho(s, t)q(t|s)\mu(s)
\end{aligned}$$

(detailed balance conditions hold).

$$\begin{aligned}\int_\Theta Pr(\vartheta_1 \in B|\vartheta_0 = t)\mu(t) &= \int_\Theta \int_B \rho(t,s)q(s|t)\mathrm{d}s\mu(t)\mathrm{d}(t) + \int_\Theta \int_B (1-r(t))\mathrm{d}\delta_t(s)\mu(t)\mathrm{d}t \\ &= \int_B \int_\Theta \rho(s,t)q(t|s)\mu(s)\mathrm{d}t\mathrm{d}s + \int_\Theta (1-r(t))1_{\{t\in B\}}\mu(t)\mathrm{d}t \\ &= \int_B r(s)\mu(s)\mathrm{d}s + \int_B (1-r(t))\mu(t)\mathrm{d}t = \int_B \mu(s)\mathrm{d}s = \pi(B).\end{aligned}$$

□

The Metropolis-Hastings Markov chain can be used to approximately compute general posterior distributions arising from data $X_1, \ldots, X_n$ in a statistical model $\{f(\cdot,\theta) : \theta \in \Theta\}, \Theta = \mathbb{R}^p, p \in \mathbb{N}$, when the prior π is $N_p(0,\Sigma)$, where Σ is a $p \times p$ non-singular covariance matrix. In this case, the 'target' density μ should be

$$\pi(\theta|X_1,\ldots,X_n) \propto e^{\ell_n(\theta) - \frac{1}{2}\theta^T\Sigma^{-1}\theta},\ \theta \in \mathbb{R}^p.$$

If, in the Metropolis-Hastings algorithm we choose auxiliary densities

$$q(\cdot|t) \sim N_p(\sqrt{1-2\delta}t, 2\delta\Sigma),$$

possible by results above, then the corresponding steps are

- Step I: Given $\vartheta_m, m \in \mathbb{N}$, generate $\xi \sim N_p(0,\Sigma)$, propose

$$s_m = \sqrt{1-2\delta}\vartheta_m + \sqrt{2\delta}\xi.$$

- Step II: Set

$$\vartheta_{m+1} = \begin{cases} s_m & \text{with } \rho(\vartheta_m, s_m) \\ \vartheta_m & \text{with } 1-\rho(\vartheta_m, s_m) \end{cases},$$

  where $\rho(\vartheta_m, s_m) = \min(1, e^{\ell(s_m)-\ell(\vartheta_m)}$

(This is sometimes called pCN-algorithms). One shows, (example sheet) that a invariant (it is unique by P&M) measure of $(\vartheta_m : m \in \mathbb{N})$ equals $\pi(\cdot|X_1,\ldots,X_n)$ from the 'target' density. This is valid for any 'step size' $\delta > 0$ and we can think of ϑ_m as a Gaussian random walk, which moves forward calibrated by a sequence of corresponding likelihood ratio tests between ϑ_m, s_m. This way we can use Monte Carlo samples $(\vartheta_m : m = M_0, \ldots, M_0 + M)$ where M_0 is some burn in after initialisation of θ_0, to approximate the posterior mean

$$\mathbb{E}\left[\theta|X_1,\ldots,X_n\right] \text{ by } \frac{1}{M}\sum_{m=M_0+1}^{M_0+M} \vartheta_m$$

and likewise the posterior quantiles R_n by empirical quantiles of the chain. In particular, we can approximately compute credible sets

$$C_k\{\theta : \|\theta - \mathbb{E}\left[\theta|X_1\ldots,X_n\right]\| \leq R_n\}, \pi(C_n|X_1,\ldots,X_n) = 1-\alpha$$

which by the Bernstein-von Mises theorem are approximately $1-\alpha$ confidence sets without requiring estimation of $I(\theta)^{-1}$.

Gibbs sampling
In Bayesian statistics, often hierarchical prior specifications are of interest, e.g. $X|\theta \sim N(\theta,1), \theta \sim N(0,\sigma^2), \sigma^2 \sim \pi_\sigma, \pi_\sigma$ a hyperprior. If X, Y are any random variables with joint pdf / pmf $f_{X,Y}$ such that we can sample from the conditional distribution $f_{X|Y}, f_{Y|X}$, then the following 'Gibbs sampling' scheme can be used. Initialise $x = x_0$, draw $Y_1 \sim f_{Y|X}(\cdot|x_0)$, then $X_1 \sim f_{X|Y}(\cdot|Y_1)$ so

$$Y_m \sim f_{Y|X}(\cdot|X_{m-1}),\ X_m \sim f_{X|Y}(\cdot|Y_m), m \in \mathbb{N}.$$

One shows that $(X_m, Y_m), X_m, Y_m$ form Markov chains with invariant densities equal to $f_{X,Y}, f_{X|Y}, f_{Y|X}$, respectively.

10.3 Bootstrap

There are non-Bayesian ways to bootstrap the 'asymptotic' quantiles of confidence sets, known as 'bootstrap methods', the first of which (die to B.Efron) we will study now: Given $X_1, \ldots, X_n$ data, consider conditional on the X_i's, a new sample space $\chi_n^b = \{X_1, \ldots, X_n\}$ and draw bootstrap random variables $X_{nj}^b, j = 1, \ldots, n$ randomly from χ_n^b with replacement, precisely with law $\mathbb{P}_n = \mathbb{P}_n(\cdot|X_1, \ldots, X_n)$

$$\mathbb{P}_n\left(X_{nj}^b = X_i\right) = \frac{1}{n}\ \forall\ i, j(\text{n fixed}).$$

In practice, if $U_1, \ldots, U_n \overset{\text{iid}}{\sim} U(0,1)$ we can take $X_{nj}^b = \sum_{i=1}^n X_i 1_{\{U_j \in (\frac{i-1}{n}, \frac{i}{n}]\}}$. The bootstrap sample mean is

$$\overline{X}_n^b = \frac{1}{n}\sum_{j=1}^m X_{nj}^b$$

has $\mathbb{E}$-expectation

$$\mathbb{E}\left[X_{nj}^b\right] = \sum_{i=1}^n X_i \mathbb{P}\left((X_{nj}^b = X_i\right) = \overline{X_n}.$$

The idea is to use the 'known' distribution of

$$\overline{X_n}^b - \overline{X}_n$$

as a proxy for the unknown distribution of $\overline{X}_n - \mu, \mu = \mathbb{E}[X_i]$. We calculate roots $R_n^b = R^b(X_1, \ldots, X_n)$ such that

$$\mathbb{P}_n\left(\omega \in \chi_n^b : |\overline{X_n^b}(\omega) - \overline{X}_n| \leq \frac{R_n}{\sqrt{n}}|X_1, \ldots, X_n\right) = 1 - \alpha, 0 < \alpha < 1$$

by evaluating the quantiles of the empirical approximation of $\mathbb{P}_n(\cdot)$ obtained from repeated bootstrap sampling (via Monte Carlo). One then proposes a bootstrap confidence set

$$C_n^b = \{v \in R : |\overline{X_n} - v| \leq \frac{R_n}{\sqrt{n}}\}.$$

This approach extends to general statistical models $\{f(\cdot,\theta) : \theta \in \Theta\}$, by computing 'resampled' estimators $\hat{\theta}_n^b = \hat{\theta}(X_{n1}^b, \ldots, X_{nn}^b)$, where $\hat{\theta}(\cdot)$ is the MLE

based on corresponding data points, and where X_{nk}^b are iid $\mathbb{P}_n$, resampled from $\{X_1, \ldots, X_n\}$, where $X_i \overset{\text{iid}}{\sim} f(\cdot, \theta)$. One then uses the known distribution of $\hat{\theta}_n^b - \hat{\theta}_n$, where $\hat{\theta_n} = \hat{\theta}(X_1, \ldots, X_n)$, as a proxy / pivot distribution for the one of $\hat{\theta}_n - \theta_0$. One then defines

$$C_n^b = \{\theta \in \Theta : \|\hat{\theta}_n^b - \hat{\theta}_n\| \leq \frac{R_n}{\sqrt{n}}\},$$

where R_n is an appropriate root of

$$\mathbb{P}_n \left(\omega \in \chi_n^b : |\hat{\theta}_n^b(\omega) - \hat{\theta}_n| \leq \frac{R_n}{\sqrt{n}} \right) = 1 - \alpha, 0 < \alpha < 1.$$

This is called the nonparametric bootstrap, as it resamples from $\{X_1, \ldots, X_n\}$ without using any distributional properties of the model. Alternatively, on my draw bootstrap samples from the distribution $P_{\hat{\theta}_n}$ corresponding to the MLE, known as the parametric bootstrap.
We now prove 'consistency of the bootstrap of the mean $\overline{X}_n$.

Theorem. Let $X_1, \ldots, X_n$ iid P on $\mathbb{R}$ with mean $\mathbb{E}[X_1] = \mu$ and finite $\text{var}(X_1) = \sigma^2$. Denote by $\Phi = \Phi_\sigma$, the cumulative distribution function (cdf) of $N(0, \sigma^2)$. Then

$$\sup_{t \in \mathbb{R}} |\mathbb{P}_n \left(\sqrt{n}(\overline{X_n^b} - \overline{X}_n | X_1, \ldots, X_n \right) - \Phi(t)| \underset{n \to \infty}{\to} 0 \quad P^{\mathbb{N}}\text{-a.s.} \quad (\dagger).$$

Remark. (i) A similar theorem can be proved for $\sqrt{n}(\hat{\theta}_n^b - \hat{\theta}_n)$, $\Phi_{I(\theta)^{-1}}$ replacing $\sqrt{n}(\overline{X}_n^b - \overline{X}_n), \Phi_\sigma$, by using asymptotic normality of $\hat{\theta}_n$ instead of the CLT (in the proof to follow).

(ii) One shows (example sheet) that the theorem implies validity of the above confidence interval that is, if $P = P_\mu$ has mean μ then

$$\mathbb{P}_\mu \left(\mu \in C_n^b \right) \underset{n \to \infty}{\to} 1 - \alpha.$$

The proof is similar to the asymptotic justification of Bayesian credible sets, replacing the Bernstein-von Mises theorem by (†)

Proof. We start with a fact on convergence in distribution: Recall that $Z_n \overset{d}{\to} X$ if $Pr(Z_n \leq x) \underset{n \to \infty}{\to} Pr(Z \leq x)$ when the limit is continuous at x.

Lemma. Suppose F_n, F are cdf's on $\mathbb{R}$ and F is continuous on $\mathbb{R}$. If further $F_n(x) \to F(x)\ \forall\ x \in \mathbb{R}$, then

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x) \underset{n \to \infty}{\to} 0,$$

the Kolmogorov distance between F_n, F

Proof. Since F is continuous, it takes $\mathbb{R}$ onto $[0, 1]$. Thus $\forall\ k \in \mathbb{N}\ \exists$ points $-\infty = x_0 < x_1 < \ldots, x_k = \infty$ such that $F(X_i) = \frac{i}{k}, i = 0, \ldots, k$. Now let $\varepsilon > 0$ and choose $K > \frac{2}{\varepsilon}$. For $x \in [x_i, x_{i+1}]$ we can write

$$F_n(x) - F(x) \leq F_n(x_{i+1}) - F(x_i) = F_n(x_{i+1}) - F(X_{i+1}) + \frac{1}{k}.$$

Likewise,

$$F_n(x) - F(x) \geq F_n(x_i) - F(x_i) - \frac{1}{k},$$

so that choosing $n \geq n(k) = n(\varepsilon)$ large enough implies that

$$|F_n(x) - F(x)| < \frac{2\varepsilon}{2} = \varepsilon \implies \sup_x |F_n(x) - F(x)| \underset{n\to\infty}{\to} 0.$$

□

By the previous lemma, it will suffice to show convergence in distribution of $\sqrt{n}(\frac{1}{n}\sum_{j=1}^{n} X_{n,j}^b - \mathbb{E}[X_{nj}]) \xrightarrow{d} N(0, \sigma^2)$ on an event of $P^{\mathbb{N}}$-probability one. □