Principles of Statistics

1 Introduction

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

If X is a discrete r.v. Then

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

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

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

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

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

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

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

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

Definition 1.2. A statistical model $\{P_\theta : \theta \in \Theta\}$ is correctly specified (for the law P of X) if $\exists \theta \in \Theta$ s.t. $P_\theta = P$. We often write θ_0 for this specific 'true' value of θ. We say that observations $X_1, \ldots, X_n \overset{iid}{\sim}$ arise from the model $\{P_0 : \theta \in \Theta\}$ in this case. We refer to n as the sample size.

The tasks of statistical inference comprise at least:

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

2) Hypothesis Testing: For $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$, we want a test (indicator) function $\psi_n = \psi(X_1, \ldots, X_n)$ s.t $\psi_n = 0$ with high probability when H_0 is true, and $\psi_n = 1$ otherwise.

3) Confidence Regions (Inference): Find regions (intervals) $C_n = C(X_1, \ldots, X_n) \subseteq \Theta$ of confidence in that $P_\theta(\theta \in C_n) \geq 1 - \alpha, \forall \theta \in \Theta$. This quantifies the uncertainty in the inference on Θ by the size/diameter of C_n. Here, $0 < \alpha < 1$ is a pre-described significance level.

2 Likelihood Principle

Example. Consider a sample $X_1, \ldots, X_n \sim^{iid} Poi(\theta)$ 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 occurrence of $x_1, \ldots, x_n$ as a function of θ is

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

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

In the example, we can equivalently maximise

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

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

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

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

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

Moreover, the *log-likelihood* is

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

and the normalised log-likelihood is

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

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

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

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

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

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

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

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

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

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

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

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

3 Information Geometry

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

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

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

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

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

If the model is correctly specified with any true value θ_0 s.t. $P = P_{\theta_0}$, then we can rewrite

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

Next, we write

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

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

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

The quantity

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

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

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

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

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

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

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

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

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

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

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

Proof. As earlier, we write

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

Moreover, using the chain/product rule, we have

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

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

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

Hence by $(*)$, (3.1) holds. □

Remark.

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

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

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

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

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

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

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

$$Var_\theta \tilde{\theta} \geq \frac{1}{nI(\theta)} \; \forall n \in \mathbb{N}$$

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

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

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

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

and use that Fisher information tensorises. □

Let us recall also

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

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

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

A multi-dimensional version of the CRLB can be obtained from considering estimation of general differentiable functions $\Phi : \Theta \to \mathbb{R}, \Theta \in \mathbb{R}^p$. Then one shows that for any unbiased estimator

$\tilde{\Phi} = \tilde{\Phi}(X_1, \dots, X_n)$, where $X_i \overset{iid}{\sim} \{f(\cdot, \theta) : \theta \in \Theta\}$, we have

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

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

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

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

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

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

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

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

Case 1: Suppose one wants to estimate θ_1, and θ_2 is known. Then (see example sheet), one finds the Fisher information $I_n(\theta)$ of this one-dimensional statistical model $\{f(\cdot, \theta_1), \Theta_1 \in \mathbb{R}\}$, with CRLB $I_1(\theta)^{-1}$.

Case 2: Now suppose 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 $CRLB(1) < CRLB(2)$ unless Σ is diagonal

4 Asymptotic Theory for MLEs

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

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

in a sense to be made precise later.

Stochastic Convergence: Concepts and Facts

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

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

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

2) We say $X_n \to X$ in probability, $X_n \to^P X$ as $n \to \infty$, if $\forall \varepsilon > 0$,

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

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

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

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

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

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

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

Proposition 4.3. 1) $X_n \underset{n\to\infty}{\overset{a.s}{\to}} X \Rightarrow X_n \underset{n\to\infty}{\to^P} X \Rightarrow X_n \underset{n\to\infty}{\to^d} X$, but any converse is false in general.

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

3) (Slutsky's Lemma) Suppose $X_n \underset{n\to\infty}{\to^d} X, Y_n \underset{n\to\infty}{\to^d} c$, c constant (non-stochastic), then

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

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

$$A_n X_n \to^d AX \text{ as } n \to \infty$$

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

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

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

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

Theorem 5.1. (Weak Law of Large Numbers)

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

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

Theorem 5.2. (Strong Law of Large Numbers)

Let $X_1, \dots, X_n \overset{iid}{\sim} X \sim P$ on $\mathbb{R}^k$ such that $\mathbb{E}\left(\|X\|\right) < \infty$. Then $\overline{X} \overset{a.s^P}{\to} \mathbb{E}(X)$ as $n \to \infty$

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

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

Theorem 5.3. (Univariate Central Limit Theorem)

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

$$\sqrt{n}(\overline{X} - \mathbb{E}(X)) \underset{n\to\infty}{\to^d} N(0, \sigma^2)$$

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

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

If Σ is invertible, then X has pdf

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

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

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

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

Theorem 5.4. (Multivariate Central Limit Theorem)

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

$$\sqrt{n}\left(\overline{X} - \mathbb{E}(X)\right) \underset{n\to\infty}{\to^d} N_k(0, \Sigma)$$

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

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

$$Y_n = O_P(c_n)$$

if

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

(By Prokhorov's Theorem)

Corollary 5.6.

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

Example. (Confidence Intervals)

Let $X_1, \dots, X_n \overset{iid}{\sim} X \sim P$ on $\mathbb{R}$ with mean μ_0 and variance σ^2.

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

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

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

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

$$t_n := \frac{\sqrt{n}(\overline{X} - \mathbb{E}(X)}{S_n} \underset{n\to\infty}{\to^d} N(0,1)$$

6 Consistency of MLEs

Definition 6.1. Let $X_1, \ldots, X_n \overset{iid}{\sim} X$ form a statistical model $\{P_\theta : \theta \in \Theta\, \Theta \subseteq \mathbb{R}^p\}$. Then we say that an estimator $\tilde{\theta}_n = \tilde{\theta}(x_1, \ldots, x_n)$ is consistent (for that model) if $\tilde{\theta}_n \underset{n\to\infty}{\to} \theta$ in $(P_\theta^{\mathbb{N}})$-probability, $\forall \theta \in \theta$.

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

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

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

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

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

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

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

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

Now, let $\hat{\theta}_n$ be any MLE, and for a true (arbitrary) value $\theta_0 \in \Theta$ we now prove that $\hat{\theta}_n \underset{n\to\infty}{\to^P} \theta_0$ (in $P_{\theta_0}^{\mathbb{N}}$ probability). The idea is that maximisers $\hat{\theta}_n$ of $\overline{l_n}$ over Θ should converge to the unique maximiser θ_0 of l over Θ, since $\overline{l_n}(\theta) \underset{n\to\infty}{\to^p} l(\theta)$ by the LLN, for all $\theta \in \Theta$ pointwise.

This is generally false unless one has uniform convergence.

$$(*) sup_{\theta\in\Theta} |\overline{l_n}(\theta) - l(\theta)| \underset{n\to\infty}{\to^P} 0 \text{ [see Ex sheet for counterexample]}$$

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

Define for any $\varepsilon > 0$

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

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

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

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

$$c(\varepsilon) + \delta(\varepsilon) < l(\theta_0) - \delta(\varepsilon) \tag{†}$$

Now, $sup_{\theta \in \Theta_\varepsilon} \overline{l_n} = sup_{\theta \in \Theta_\varepsilon} l(\theta) + (\overline{l_n}(\theta) - l(\theta) \leq sup_{\theta \in \Theta_\varepsilon} l(\theta) + sup_{\theta \in \Theta_\varepsilon} |\overline{l_n}(\theta) - l(\theta)|$

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

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

On these events, we have

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

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

$$\overline{l_n}(\hat{\theta_n} \leq sup_{\theta \in \Theta_\varepsilon} \overline{l_n}(\theta) < l(\theta_0) \text{ on } A_n(\varepsilon)$$

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

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

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

$$\mathbb{P}\left(\|\hat{\theta}_n - \theta_0\| < \varepsilon\right) \underset{n \to \infty}{\to} 1 \text{ or}$$
$$\mathbb{P}\left(\|\hat{\theta}_n - \theta_0\| < \varepsilon\right) \underset{n \to \infty}{\to} 0$$

Since ε is non arbitrary, $\hat{\theta}_n \underset{n \to \infty}{\to^P} \theta_0$, and the proof is complete modulo the verification of $(*)$. □

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

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

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

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

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

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

Moreover,

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

since

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

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 $\forall h \in \mathcal{H}, \exists j$ s.t $\underline{h_j} \leq h(x) \leq \overline{h^j} \forall x \in \chi$.

Proposition 6.3. Suppose that $\forall \varepsilon > 0$ there exist brackets $[\underline{h_j}, \overline{h_j}], j = 1, \dots, N(\varepsilon)$ covering $\mathcal{H}$ and such that

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

Then

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

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

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

and equivalently for $\overline{h_j}$

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

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

Likewise, we get

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

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

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

and since

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

the proof is complete. □

From this proposition, we deduce

Proposition 6.4. Let $\chi \subseteq \mathbb{R}^d, \Theta \in \mathbb{R}^p$ compact, and suppose $\theta \mapsto q(x, \theta)$ is continuous $\forall x$ (and x measurable $\forall \theta$), and that $\mathbb{E} sup_{\theta \in \Theta} |q(X, \theta)| < \infty$. If $X_1, \ldots, X_n \overset{iid}{\sim} X$, then

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

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

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

7 Asymptotic Distribution of MLEs

Definition 7.1. We say that an estimator $\tilde{\theta}_n$ is asymptotically efficient in a regular statistical model $\{f(\cdot, \theta) : \theta \in \Theta\}$ if $lim_{n \to \infty} n Var_\theta(\tilde{\theta}) = I(\theta)^{-1}, \forall \theta \in \text{int}\, \Theta$.

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

Let θ_0 be a fixed ('true') value, and assume

1. $\theta_0 \in \operatorname{int}\Theta$
2. There exists an open set U satisfying $\theta_0 \in U \subseteq \Theta$ such that $f(x,\theta)$ is, for every $x \in \chi$, twice continuously differentiable wrt θ in U
3. the $p \times p$ matrix $\mathbb{E}_{\theta_0}\frac{\partial^2 \log f(X,\theta_0)}{\partial\theta\partial\theta^T}$ is non singular, and
$$\mathbb{E}_{\theta_0}\left\|\frac{\partial \log f(X,\theta_0)}{\partial\theta}\right\|^2 < \infty$$
4. There exists a compact ball $K \subset U$ (with non-empty interior) centered at θ_0 such that
$$\mathbb{E}_{\theta_0}\left\|\frac{\partial^2 \log f(X,\theta_0)}{\partial\theta\partial\theta^T}\right\| < \infty,$$
$$\int_\chi sup_{\theta\in K}\left\|\frac{\partial f(x,\theta)}{\partial\theta}\right\|dx < \infty \text{ and } \int_\chi sup_{\theta\in K}\left\|\frac{\partial^2 \log f(X,\theta_0)}{\partial\theta\partial\theta^T}\right\|dx < \infty$$
5. Suppose the MLE $\hat\theta_n$ in the model $\{f(\cdot,\theta) : \theta \in \Theta\}$ based on the sample $X_1, \dots, X_n$ exists and is consistent, i.e. $\hat\theta_n \xrightarrow[n\to\infty]{P_{\theta_0}} \theta_0$.

Theorem 7.2. Suppose a statistical model $\{f(\cdot,\theta) : \theta \in \Theta\}$ is regular in the sense that it satisfies the conditions B on the handout. Then, if $\hat\theta_n$ is the MLE, based on $X_1, \dots, X_n \overset{iid}{\sim}$ from the model, we have $\sqrt{n}(\hat\theta_n - \theta_0) \xrightarrow[n\to\infty]{d} N(0, I(\theta)^{-1})$

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

$$0 = l_n'(\hat\theta) = l_n'(\theta_0) + l_n''(\overline{\theta}_n(\hat\theta_n - \theta_0) \text{ (MVT) },$$

so

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

Lemma 7.3. Observations 2 and 3 from the Information Geometry section are valid.

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

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

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

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

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

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

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

For any map $h : U \to \mathbb{R}$, we can apply the mean value theorem along the line segment $\{t\hat{\theta}_n + (1-t)\hat{\theta}_0 : 0 < t < 1\}$ connecting $\hat{\theta}_n$ and θ_0, and write

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

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

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

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

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

We will show

$$\overline{A}_n \underset{n\to\infty}{\to^P} -I(\theta_0) \tag{†}$$

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

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

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

and the theorem follows from (†), Slutsky's Lemma, and since

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

To verify (†), it suffices (see Ex. sheet) to check convergence in probability of $\overline{A}_{n_{jk}} \to (-I_(\theta_0))_{jk}$. Now, we write

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

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

$$|(1)| = |\frac{1}{n}\sum q(X_i,\overline{\theta}_j) - \mathbb{E}q(X,\overline{\theta}_{(j)})| \leq sup_{\theta\in K}|\frac{1}{n}\sum q(X_i,\overline{\theta}_j) - \mathbb{E}q(X,\overline{\theta}_{(j)})| \underset{n\to\infty}{\to^P} 0 \text{ by the uniform LLN}$$

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

$$(2) = \mathbb{E}q(X,\overline{\theta}_j) - \mathbb{E}q(X,\theta_0) \underset{n\to\infty}{\to^P} 0,$$

completing the proof of (†)

□

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

2. If the 'true' value θ_0 lies at the boundary of Θ, then the MLE is also not asymptotically normal (Ex. sheet $N(\theta,1), \theta \in \Theta = [0,\infty)$)

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

   Define

$$\tilde{\theta} = \theta\tilde{}_{Hodges} = |\hat{\theta} \text{ if } |\hat{\theta}| > n^{-\frac{1}{4}} \qquad\qquad 0 \text{ if } |\hat{\theta}| < n^{-\frac{1}{4}}$$

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

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

So for such θ we thus have $\sqrt{n}(\tilde{\theta} - \theta) \underset{n\to\infty}{\to^d} N(0, I(\theta)^{-1})$.

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

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

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

8 Plug-in MLEs and the Delta method

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

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

Theorem 8.1. (Delta-method)

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

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 value $\overline{\theta}$

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

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

□

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

9 Asymptotic inference with the MLE

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

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

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

Indeed by the continuous mapping theorem,

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

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

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

Definition 9.1. The $p \times p$ matrix

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

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

Proposition 9.2. Suppose the statistical model $\{f(\cdot, \theta) : \theta \in \Theta\}$ satisfies Assumption B. Then $\hat{i}_n \underset{n\to\infty}{\to^P} I(\theta_0)$.

Proof. Just as when proving $\overline{A}_n \underset{n\to\infty}{\to^P} I(\theta_0)$ in the proof of asymptotic normality of $\hat{\theta}_{MLE}$, replacing $\frac{\partial^2}{\partial\theta\partial\theta^T}, \overline{\theta}_{ij}$ by $\frac{\partial}{\partial\theta}$ and $\hat{\theta}_{MLE}$ respectively. □

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

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

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

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

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

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

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

Consider next a hypothesis testing problem

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

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

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

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

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

Definition 9.3. A general purpose test can be constructed from the Likelihood ratio test statistic

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

Theorem 9.4 (Wilks'). In a statistical model $\{f(\cdot, \theta) : \theta \in \Theta\}$ satisfying Assumption B, and for $\Theta_0 = \{\theta_0\}$, for $\theta_0 \in \Theta$, we have under P_{θ_0},

$$\Lambda_n(\Theta, \Theta_0) \underset{n\to\infty}{\to^d} \chi_p^2, p = \dim\Theta$$

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

$$\Lambda_n(\Theta, \Theta_0) \underset{n\to\infty}{\to^d} \chi^2_{p-p_0} \text{ under } P_\theta, \theta \in \Theta_0$$

2. 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 $\chi^2_{p-p_0}$-distribution.

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

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

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

$$\sqrt{n}(\hat{\theta}_n - \theta_0)(j_n(\theta) - I(\theta_0))\sqrt{n}(\hat{\theta}_n - \theta_0) + \sqrt{n}(\hat{\theta}_n - \theta_0)(I(\theta_0))\sqrt{n}(\hat{\theta}_n - \theta_0) \underset{n\to\infty}{\to^d} Z^T I(\theta_0)Z$$

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

10 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* 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 $\pi_i, \sum_{i\in\Theta} \pi = 1$, and such that

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

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

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

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

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

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

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

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

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

Likewise,

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

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

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

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 that

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

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

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

Example. We have some other examples of conjugate priors.

i) Beta prior + Binomial sampling → Beta posterior

ii) Gamma prior + Poisson sampling → Gamma posterior

Even when π is not a proper probability distribution, the expression

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

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

10.1 Statistical Inference with Posterior Distributions

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

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

ii) Uncertainty quantification - Any subset $C_n \subseteq \Theta$ for which $\pi(C_n|X_1,\dots,X_n) = 1-\alpha$ is a level $1-\alpha$ *credible* set (but it has, a fortiori, no interpretation in terms of coverage probabilities $P_\theta(\theta \in C_n)$)

iii) Hypothesis testing - Given $\Theta_0, \Theta_1 \subseteq \Theta$, we can compute Bayes-factors

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

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

10.2 Frequentist Analysis of Bayes Methods

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

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

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

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

Theorem 10.1. (Bernstein-von Misses)

Suppose a statistical model $\{f(\cdot,\theta) : \theta \in \Theta\}$ satisfies assumption 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 $N(\hat{\theta}_{MLE}, \frac{1}{n}I(\theta_0)^{-1})$. Then

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

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

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

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

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

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

The last theorem remains true when $\hat{\theta}_{MLE}$ is replaced by any estimator $\overline{\theta}_n$ s,t $\sqrt{n}(\hat{\theta}_{MLE} - \overline{\theta}_n) \underset{n\to\infty}{\to} 0 P_{\theta_0}^{\mathbb{N}}$-a.s, typically permitting the alternative centring at $\overline{\theta} = \mathbb{E}^{\pi}\theta(X_1, \ldots, X_n)$. One important consequence of the BvM-theorem is that certain posterior *credible sets* are in fact proper (asymptotic) frequentist confidence sets.

For instance, consider $C_n = \{\theta : |\theta - \hat{\theta}_{MLE} \leq \frac{R_n}{\sqrt{r}}\}$, where R_n are random quantile constants chosen s.t $\prod(C_n|X_1, \ldots, X_n) = 1 - \alpha, 0 < \alpha < 1$. Recall $\hat{\phi}_n$ was the pdf of $N(\hat{\theta}_{MLE}, \frac{1}{n}I(\theta_0)^{-1})$distribution, and defined further ϕ_0 to be the pdf of $Z \sim N(0, I(\theta)^{-1})$. We can define $\Phi(t) = \mathbb{P}(|Z| \leq t) = \int_{-t}^{t} \phi_0(v)dv$, which is strictly increasing in t, and also continuously differentiable, hence admits a continuous inverse $\Phi^{-1} : [0,1] \to \mathbb{R}$.

Now, we can write $\Phi(R_n) = \int_{-R_n}^{R_n} \phi_0(v)dv$, and substituting $v = \sqrt{n}(\theta - \hat{\theta})$ so that $-R_n \leq v \leq R_n$ becomes the set C_n and since $v\ N(0, I(\theta_0)^{-1}) \Leftrightarrow \sqrt{n}(\theta - \hat{\theta}) \sim N(\hat{\theta}, \frac{1}{n}I(\theta_0)^{-1})$, the last integral gives

$$\begin{aligned}
\Phi(R_n) &= \int_{C_n} \hat{\phi}_n(\theta)d\theta \\
&= \int_{C_n} (\hat{\phi}_n(\theta) - \pi_n(\theta))d\theta + \int_{C_n} \pi_n(\theta) \\
&= \pi(C_n|X_1, \dots, X_n) \text{ by BvM theorem} \\
= 1 - \alpha \forall n &
\end{aligned}$$

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

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

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

$$\frac{\Phi^{-1}(1-\alpha}{R_n}\sqrt{n}(\hat{\theta_{MLE}} - \theta_0) \underset{n\to\infty}{\to}^{d} N(0, I(\theta_0)^{-1}) \text{ under } P_{\theta_0}$$

Now,

$$\begin{aligned}
P_{\theta_0}^{\mathbb{N}}(\theta_0 \in C_n) &= P_{\theta_0}^{\mathbb{N}}\left(|\theta_0 - \hat{\theta}_{MLE}| \leq \frac{R_n}{\sqrt{n}}\right) \\
&= P_{\theta_0}^{\mathbb{N}}\left(\frac{\Phi^{-1}(1-\alpha)}{R_n}\sqrt{n}|\hat{\theta}_{MLE} - \theta_0| \leq \Phi^{-1}(1-\alpha)\right) \\
&\underset{n\to\infty}{\to} \mathbb{P}\left(|Z| \leq \Phi^{-1}(1-\alpha)\right) \text{ by } (\dagger) \text{ and continuous mapping theorem on } |\cdot| \\
&= \Phi(\Phi^{-1}(1-\alpha)) \\
&= 1-\alpha
\end{aligned}$$

Replacing $|\cdot|$ by $\|\cdot\|$, the argument extends to parameter spaces $\Theta \subseteq \mathbb{R}^p$.

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 priors can be proved as well]

11 Decision Theory

Consider a (single) observation X from some statistical model $\{f(\cdot, \theta) : \theta \in \Theta\}$ of pdf/pmfs on $\chi \subseteq \mathbb{R}^d$. In a decision 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 $\mathcal{A} = \{1, \dots, M\}$

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

iii) $\mathcal{A} = \{\text{All (meas) subsets of } \Theta\}$, set-valued estimation, decision rule $\delta(X) = C(X)$ are confidence regions.

For general decision problems, we consider *loss functions*

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

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

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

ii) (Estimation) $\Theta \subset \mathbb{R}$, absolute loss $L(a, \theta) = |a - \theta|$, squared loss $L(a, \theta) = (a - \theta)^2$

Definition 11.1. The risk of a decision rule $\delta(X)$ from $X \sim P_\theta$, for loss L, is defined as

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

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

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

I.e our risk is the mean squared error (MSE) (quadratic risk)

Definition 11.2. Given a prior π on Θ, the π-Bayes risk in a decision rule δ is

$$R_\pi(\delta) = \int_\Theta R(\delta, \theta) \pi(\theta) d\theta$$

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

We can rewrite, interchanging the order of integration,

$$\begin{aligned}
R_\pi(\delta) &= \int_\chi \int_\Theta L(\delta(x),\theta) \frac{f(x,\theta)\pi(\theta)}{\int_\Theta f(x,v)\pi(v)dv} \left(\int_\Theta f(x,v)\pi(v)dv \right) d\theta dx \\
&:= \int_\chi \int_\Theta L(\delta(x),\theta)\pi(\theta|x)m(x)d\theta dx \text{ where } m \geq 0 \\
&= \int_\chi \mathbb{E}^\pi \left[L(\delta(x),\theta)|x \right] m(x)dx
\end{aligned}$$

where $\mathbb{E}^\pi \left[L(\delta(x),\theta)|x \right]$ is the posterior risk.

We conclude that any decision rule $\overline{\delta}(X)$ which minimises the posterior risk in the sense that $\forall X, \delta$,

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

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

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

Proposition 11.3. In an estimation problem, suppose a decision rule $\delta(X)$ is unbiased for Θ, i.e $\mathbb{E}_\theta \delta(X) = \theta \ \forall \theta \in \Theta$. Assume further that δ is a π-Bayes rule for some prior π on Θ with squared loss. Then, where Q has density on $\chi \times \Theta$ given by $dQ(x,\theta) = f(x,\theta)\pi(\theta)dxd\theta$,

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

[It is sometimes said that $\delta(X) = \theta$ almost surely under Q]

Proof. Recall the 'tower property' of iterated expectations

$$\begin{aligned}
\mathbb{E}[Z(X,\theta)] &= \mathbb{E}[\mathbb{E}^\pi[|(X,\theta)|X]] \\
&\overset{\text{or}}{=} \mathbb{E}[\mathbb{E}_\theta[Z(X,\theta)]]
\end{aligned}$$

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

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

and likewise

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

Now,

$$\begin{aligned}
\mathbb{E}[\delta(X) - \theta]^2 &= \mathbb{E}[\delta^2(X)] - 2E[\theta\delta(X)] + \mathbb{E}[\theta^2] \\
&= 0
\end{aligned}$$

□

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

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

Example. $\frac{X}{n}$ in a $Bin(n, \theta), \theta \in \Theta = [0, 1]$ is a π-Bayes rule for quadratic risk only for degenerate pairs.

11.1 Minimax Risk

Definition 11.4. A decision rule $\delta(X)$ in a decision problem is called *minimax* if it attains the *minimax risk* (for loss L)

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

where $R_m(\delta, \Theta) = \sup_{\theta\in\Theta}$ is the maximal/worst case risk.

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

$$R_\pi(\delta) = \int_\Theta R(\delta, \theta)\pi(\theta)d\theta \leq R_m(\delta, \Theta) \int_\Theta \pi(\theta)d\theta = R_m(\delta, \Theta)$$

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

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

(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.

Corollary 11.6. If δ_π has constant risk in θ, then it is minimax, and if δ_π is unique, then it is unique minimax.

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

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

So, taking inf over all δ we see

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

and 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 $\delta \neq \delta_\pi$ so that for such δ we have

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

so δ_π is the unique minimax.

iii) Let π' be any prior on Θ. Then

$$\begin{aligned} R_{\pi'}(\delta_{\pi'}) &\leq R_{\pi'}(\delta_\pi) \\ &= \int_\Theta R(\delta_\pi, \theta)\pi'(\theta) d\theta \\ &\leq sup_{\theta \in \Theta} R(\delta_\pi, \theta) \\ &= R_\pi(\delta_\pi) \text{ by assumption} \end{aligned}$$

]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 equality. □

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 $\text{Bin}(n, \theta)$ model with $\theta \in \Theta = [0,1]$ and quadratic risk, consider priors $\pi_{a,b}$ arising from the $\text{Beta}(a, b)$ distribution. In this case, the unique $\pi_{a,b}$-Bayes rule is given by 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 (ex. sheet).

One then solves in a,b the equation

$$R_{\pi_{a,b}}(\delta_{\pi_{a,b}}, \theta) = \text{ const},$$

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

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

11.2 Admissibility

Definition 11.7. In a decision problem, a decision rule δ is called *inadmissible* if $\exists \delta'$ s.t $R(\delta', \theta) \leq \mathbb{R}(\delta, \theta)$ $\forall \theta \in \Theta$, and $R(\delta', \theta) < R(\delta, \theta)$ for some θ. We say that δ' dominates δ.

δ is called admissible if no such δ' exists.

Proposition 11.8. i) Every unique Bayes rule is admissible

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

Proof. See example sheet. □

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

Theorem 11.9. Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, \sigma^2)$, where σ^2 is known, $\theta \in \Theta = \mathbb{R}$. Then $\hat{\theta}_{MLE} = \overline{X}_n$ 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 $p \geq 3$ we have...

Theorem 11.10. If $X \sim N_p(\theta, I_p), \theta \in \Theta = \mathbb{R}^p, p \geq 3$, then $\hat{\theta}_{MLE} = \overline{X}_n$ is inadmissible for quadratic risk.

Proof. (12.9) We can set $\sigma^2 = 1$ as the proof will show. Also, the risk of $\overline{X}_n$ is given by

$$R(\overline{X}_n, \theta) = \mathbb{E}_\theta(\overline{X}_n - \theta)^2 = \frac{1}{n} \text{ which is constant in } \theta,$$

hence by the previous proposition, it suffices to prove that $\overline{X}_n$ is admissible.

So, let δ be any other decision rule. Then

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

for $B(\theta) = \mathbb{E}_\theta\delta - \theta$

Recalling the Cramer-Rao inequality for biased estimators, we know that

$$var_\theta\delta(X) \geq \frac{(\frac{d}{d\theta}\mathbb{E}_\theta\delta)^2}{nI(\theta)} = \frac{(1 + B'(\theta))^2}{n},$$

by definition of B and $I(\theta)$. Hence, if δ dominates $\overline{X}_n$, then 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. In other words, for these sequences $B'(\theta_i) \to 0$.

Now, evaluating these limits in the above inequality, we see that $\lim_{i\to\infty}\left[B^2(\theta_i)+\frac{(1+B'(\theta_i))^2}{n}\right]\leq\frac{1}{n}$ gives as that $\lim_{i\to\infty} B^2(\theta_i)=0$. Therefore by monotonicity $B(-\infty)=B(\theta)=B(\infty)=0$, and so the bias vanishes identically.

So,

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

□

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 and 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 is 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 $\Theta=[a,b]$, then $\overline{X}_n$ is also no longer minimax (see example sheet).

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

To prove theorem 12.10, we first need a new estimator.

Definition 11.11. (James-Stein Estimator) Define the *James-Stein Estimator*, for $X\sim N_p(\theta,I)$ by

$$\delta^{JS}:=\begin{pmatrix}\delta_1^{JS}\\ \vdots\\ \delta_p^{JS}\end{pmatrix},\delta_j^{JS}=\left(1-\frac{p-2}{\|X\|_2^2}\right)X(p\geq 3)$$

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

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

Lemma 11.12. (Stein)

Let $X \sim N(\theta, 1)$ and let $g : \mathbb{R} \to \mathbb{R}$ be differentiable and such that $\mathbb{E}_\theta g(X) < inf$. Then $\forall \theta \in \mathbb{R}$,

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

Proof.

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

□

Proof. (12.10)

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

Now, the last expectation can be written as

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

where $\mathbb{E}_j = \mathbb{E}[\cdot | X_{(-j)}]$, with $X_{(-j)} = \{X_i : -i \neq j\}$. The j^{th} expectation can be written as $\mathbb{E}_j[(X_j - \theta_j)g(X_j)]$, where $g(y) = \frac{y}{y^2+1}$, $a = \sum_{i \neq j} X_i^2$, 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 $\mathbb{R}$.

Hence we can apply Stein's lemma to this expectation, giving

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

So going back to our original statement,

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

Since

$$\begin{aligned}
\mathbb{E}_\theta\frac{1}{\|X\|^2} &= \int_{\mathbb{R}^p}\frac{1}{\|X\|^2}\phi(x-\theta)dx \\
&\geq c\int_{c_0\leq\|X\|\leq c_1}\phi(x-\theta)dx \\
&\geq c\mathbb{P}_\theta(\|X\|\in(c_0,c_1)) \\
&> 0
\end{aligned}$$

where here ϕ is our $N(\theta, I_p)$ pdf. □

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

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

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

$$\delta^{JS,+} = \left(1-\frac{p-2}{\|X\|^2}\right)^+ X, \text{ where } (\cdot)^+ = \max(\cdot, 0),$$

known as the 'positive part JS-estimator. This is also not admissible.

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

$$\delta^c := \left(1-c\frac{p-2}{\|X\|^2}\right)X, 0<c,$$

the choice $c = 1$ is optimal.

iv) While δ^{JS} is attractive from a decision-theoretic perspective, its use for inference (confidence regions and tests) is less clear, as its distributional properties are more involved than those of $\delta_{MLE}(X) = X \sim N(\theta, I_p)$.

11.3 Classification Problems

For two pdf/pmfs f_1, f_2 defined on χ, consider observing X drawn from f_i 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 $R \subseteq \chi (\subseteq \mathbb{R}^d)$ such that

$$\delta_R(X) = \begin{cases} 1 \text{ if } X \in R \\ 2 \text{ if } X \in R^c \end{cases}$$

The classification errors are given by

$$\mathbb{P}\left(2|1, R)\right) = P_{f_1}(X \in R^c = \int_{R^c} f_1(x)dx,$$

and by

$$\mathbb{P}\left(1|2, R\right) = p_{f_2}(x \in R) = \int_R f_2(x)dx$$

The π-Bayes risk now becomes

$$R_\pi(\delta_R) = \mathbb{P}\left(1|2, R\right)\pi_2) + \mathbb{P}\left(2|1, R\right)\pi(1),$$

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

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

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 11.13. Suppose that for all i, $P_{f_i}\left(\frac{f_1(x)}{f_2(x)} = \frac{1-q}{q}\right) = 0$. Then the unique π-Bayes classification rule for prior $\pi(\{1\}) = q$arises from

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

and in fact δ_R is also admissible.

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

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

The first term is minimal when S^c includes precisely all $x \in \chi$ s.t. the integrand

$$qf_1(x) - (1-q)f_2(x) < 0 \Leftrightarrow \frac{f_1(x)}{f_2(x)} < 1 - \frac{q}{q}$$

(Uniqueness since $\mathbb{P}\left(\frac{f_1(x)}{f_2(x)} = \frac{1-q}{q}\right) = 0$. 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 R s.t.

$$q\mathbb{P}\left(2|1,R\right)+(1-q)\mathbb{P}\left(1|2,R\right)= \text{ const in } q\in[0,1]$$

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

12 Further Topics

12.1 Basic Multivariate Analysis

Consider random vectors X,Y, here assumed to be from $N(\mu_x,\sigma_x),N(\mu_y,\sigma_y)$. [The assumption of normality is not required if we instead assume that we can work to just the order of the first two moments]. Then their correlation is $\rho_{x,y}=\frac{Cov(X,Y)}{\sqrt{VarX}\sqrt{VarY}}$. For $X_1,\ldots,X_n;Y_1,\ldots,Y_n$ jointly iid, this can be estimated by the empirical correlation

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

where $S_{n,x}=\frac{1}{n}\sum_{i=1}^n(X_i-\overline{X}_n)^2$ etc.

The finite sample distribution of $\hat{\rho}_{x,y}$ can in principle be computed analytically; it is given by $f_{\hat{\rho}_{x,y}}(r)\propto(1-r^2)^{\frac{1}{2}(n-4)}, -1\le r\le 1$.

Alternatively, we can regard $\rho_{x,y}=\Phi(\theta)$, with samples from $N\left(\begin{pmatrix}\mu_x\\ \mu_y\end{pmatrix},\Sigma\right)$, where $\theta=\left(\begin{pmatrix}\mu_x\\ \mu_y\end{pmatrix},\Sigma\right)$.

One shows that $\hat{\rho}=\Phi(\hat{\theta}_{MLE})$, and by the general theory developed earlier (plus Delta method for Φ), we deduce $\sqrt{n}(\hat{\rho}_{x,y}-\rho)\underset{n\to\infty}{\to^d}N(0,Var(\Phi,\theta))$ for some asymptotic variance $Var(\Phi,\theta)$

More generally, consider a partitioned random vector $X\in\mathbb{R}^p$

$$X=\begin{pmatrix}X^{(1)}\\ X^{(2)}\end{pmatrix}, X\sim N(0,\Sigma),$$

such that $X\in\mathbb{R}^q\times\mathbb{R}^{p-q}$, with Σ a $p\times p$ positive definite matrix. The conditional variance of $X^{(1)}|X^{(2)}$ is given by $\underbrace{\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}}_{\equiv\Sigma_{11\cdot 2}}$, where $\Sigma=\begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{12} & \Sigma_{22}\end{pmatrix}$.

One further defines the partial correlations of $X_i,X_j, i\ne j, i,j\le q$, 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' $\Phi(\hat{\Sigma}_{MLE})$, where $\left(\hat{\Sigma}\right)_{ij}=\frac{1}{n}\sum_{k=1}^n(X_kX_k^T)_{ij}$, where Φ is again a smooth map on $\Theta=\{\Sigma \text{ positive definite}\}$

The previous theory again implies that $\sqrt{n}(\hat{\rho}_{ij\cdot 2}-\rho_{ij\cdot 2})\underset{n\to\infty}{\to^d}N(0,?)$

Finally, consider $X \sim N(0, \Sigma)$, again with Σ a $p \times p$ positive definite symmetric. Then there exists an orthonormal matrix T s.t.

$$\Sigma = T^T \Lambda T,\ \Lambda = diag(\lambda_1, \dots, \lambda_p), \lambda_1 > \lambda_2 > \dots > \lambda_p > 0.$$

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 the principle subspaces spanned by the column vectors of T (eigenvectors of Σ), arranged in decreasing order of 'explained variance'. Here again, if $\left(\hat{\Sigma}\right)_{ij} = \frac{1}{n} \sum_{k=1}^{n} (X_m, X_m^T)_{ij}$, then the plug-in MLE will give asymptotically efficient estimators λ_i, u_i provided that Σ has no multiplicity in its eigenvalues.

12.2 Monte Carlo Methods

We will discuss algorithms to generate random samples from given probability distributions, which are useful to construct numerical approximations of inference methods (e.g. non-conjugate posterior distributions).

One can generate, with a pseudorandom number generator (see Computer Science), random samples $U_1, \dots, U_N \overset{iid}{\sim} U[0,1], N \in \mathbb{N}$. If F is a cdf of some r.v. X, with quantile transform

$$F^{-}(u) = inf\{x_i : u \leq F(x)\},$$

then one shows that $X_i^* = F^{-}(U_i)$ are iid r.v's with distribution F

For normal r.v's, F^{-} is not available in closed form, but one can still simulate $X \sim N_2(0, I_2)$ starting from $U_1, U_2 \overset{iid}{\sim} U(0,1)$ by the Box-Müller transform (see ex. 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 f on χ, and have another density h that we can sample from, s.t.

$$f(x) \leq Mh(x)\ \forall x \in \chi, \text{ some } M > 0$$

1. Draw $X \sim h$, and independently $U \sim U[0,1]$
2. Set $Y = X$ if $U \leq \frac{f(X)}{Mh(X)}$, otherwise return to step 1.

One shows (ex sheet) that $Y \sim f$ on χ.

In the preceding settings, if we can generate $(X_i^*)_{i=1,\dots,N}$ iid from a fixed distribution P_F, then we can numerically approximate integrals

$$\mathbb{E}_{P_F} g(X) = \int_\chi g(x) dP_F(x),\ g \text{ given}$$

by the MC-average

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

12.3 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

$$\begin{aligned}\mathbb{P}\left(\vartheta_m \in \mathcal{B} | \vartheta_{m-1}, \vartheta_{m-2}, \ldots\right) &= \mathbb{P}\left(\vartheta_m \in \mathcal{B} | \vartheta_{m-1} = t\right) \\ &= \mathbb{P}\left(\vartheta_1 \in \mathcal{B} | \vartheta_0 = t\right) \\ &= K(t, \mathcal{B})\end{aligned}$$

where $\mathcal{B} \subseteq \Theta$ (meas), and where K is a *Markov kernel* s.t. $K(t, \cdot)$ is a pdf on Θ, the state space of (ϑ_m).

A pdf/pmf μ on Θ is called invariant for K if

$$\begin{aligned}\int_\Theta K(t, B)\mu(t)dt &= \int_\Theta \mathbb{P}\left(\vartheta_1 \in \mathcal{B} | \vartheta_0 = t\right)\mu(t)dt \\ &= \mu(t)\end{aligned}$$

Under additional hypotheses (ergodicity of (ϑ_m)), one shows that the distribution of ϑ_m 'mixes towards' (converges to in some sense) its invariant measure μ, and we can then use MC-averages $(\vartheta_m : m = 1, \ldots, N)$ to approximate $\mathbb{E}_\mu g(X)$ by $\frac{1}{N}\sum_{i=1}^N g(\vartheta_m)$

An important MCMC method, known as the *Metropolis-Hastings* algorithm, requires an auxiliary (conditional) pdf $q(\cdot, t), t \in \Theta$ that we can sample from, and proceeds as follows:

1. Given $m \in \mathbb{N}, \vartheta_m \in \Theta$, generate $s_m \sim q(\cdot|\vartheta_m)$.
2. 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

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

and μ is a prescribed pdf/pmf (in fact our invariant measure).

Proposition 12.1. 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.

Proof. (Uniqueness requires Probability and Measure).

The transition Markov kernel K has 'density' k,

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

where δ_τ is the Dirac point mass probability measure at τ, and where $r(t) = \int_\Theta \rho(t, \tau)q(\tau|t)d\tau$.

We have that

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

So our detailed balance conditions hold.

Then

$$\begin{aligned}\int_\Theta \mathbb{P}\left(\vartheta_1 \in \mathcal{B}|\vartheta_0 = t\right)\mu(t)dt &= \int_\Theta\int_\mathcal{B} \rho(t,s)q(s|t)\mu(t)dsdt + \int_\Theta\int_\mathcal{B}(1-r(t))d\delta_t(s)\mu(t)dt\\ &= \int_\mathcal{B}\int_\Theta \rho(s,t)q(t|s)\mu(s)dtds + \int_\Theta 1_{\{t\in\mathcal{B}\}}(1-r(t))\mu(t)dt\\ &= \int_\mathcal{B} r(s)\mu(s)ds + \int_\mathcal{B}(1-r(t))\mu(t)dt\\ &= \int_\mathcal{B}\mu(s)ds\\ &= \mu(\mathcal{B})\end{aligned}$$

□

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

$$\pi(\theta|X_1,\dots,X_n) \propto e^{l_n(\theta)-\frac{1}{2}\theta^T\Sigma_p^{-1}\theta}, \theta\in\mathbb{R}^p$$

If in the Metropolis-Hastings algorithm, we choose auxiliary densities $q(\cdot,t)\sim N(\sqrt{1-2\delta}t, 2\delta\Sigma_p)$, which is possible by results from previous lectures, and then the corresponding steps are

1. Given $\vartheta_m, m\in\mathbb{N}$, generate $\xi\sim N(0,\Sigma_p)$, define

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

2. 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 $\rho(\vartheta_m,S_m) = min(1, e^{l(S_m)-l(\vartheta_m)}$

This is sometimes called pCN (preconditioned Crank-Nicolson) algorithm.

One shows (ex sheet) that an invariant (it is unique by PM) measure of $(\vartheta_m : m\in\mathbb{N})$ is given by $\pi(\cdot|X_1,\dots,X_n)$ from our target density μ. This is valid for any $\delta>0$, 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 and S_m. This way we can use Monte Carlo samples $(\vartheta_m : m = M_0,\dots,M_0+M)$ where M_0 is some turn after initialisation at ϑ_0, to approximate the posterior mean $\mathbb{E}[\theta|X_1,\dots,X_n]$ 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

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

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

12.4 Gibbs Sampling

In Bayesian statistics, often hierarchical prior specifications are of interest, e.g.

$$X|\theta \sim N(\theta, 1), \theta \sim N(0, \sigma^2), \sigma^2 \sim \pi_\sigma, \pi_\sigma \text{ a } \textit{hyperprior}.$$

If X, Y are any r.v's with joint pdf/pmf $f_{X,Y}$ s.t. we can sample from the conditional distributions $f_{X|Y}, F_{Y|X}$, then the following 'Gibbs sampling' scheme can be used.

Initialise $X = x_0$, then draw $Y_1 \sim f_{Y|X}(\cdot|x_0)$, then $X_1 \sim f_{X|Y}(\cdot, y_1)$. Repeating, we have $Y_m \sim f_{Y|X}(\cdot|X_m - 1), X_m \sim f_{X|Y}(\cdot|Y_m), m \in \mathbb{N}$.

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.

13 Bootstrap Methods

There are non-Bayesian ways to bootstrap the 'asymptotic' quantiles of confidence sets, known often as 'bootstrap methods', the first of which (due to B. Efron) we will now study.

Given data $X_1, \dots, X_n$, consider, conditional on the observed values, a new sample space $\chi_n^b = \{X_1, \dots, X_n\}$, and draw bootstrap r.v's $X_{nj}^b, j = 1, \dots, n$ randomly from χ_n^b with replacement, i.e. precisely with law $\mathbb{P}_n = \mathbb{P}_n(\cdot|X_1, \dots, X_n)$, $\mathbb{P}_n(X_{nj}^b = X_i) = \frac{1}{n} \forall i, j$ (n fixed) (This is easily done by sampling uniform variables and partitioning $[0, 1]$ into n equal parts).

The bootstrap sample mean

$$\overline{X_n^b} = \frac{1}{n} \sum_{j=1}^n X_{nj}^b$$

has $\mathbb{E}$-expectation

$$\mathbb{E} X_{nj}^b = \sum_{i=1}^n X_i \mathbb{P}(X_{nj}^b = X_i) = \overline{X}_n$$

The idea is to use the 'known' distribution of $\overline{X_n^b} - \overline{X}_n$ as a proxy for the unknown distribution of $\overline{X}_n - \mu, \mu - \mathbb{E}X_i$.