Principle of Statistics

Michael Li

March 28, 2017

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

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

Decision Theory & Multivariate Analysis
Basic elements of a decision problem, including loss and risk functions. Decision rules, admissibility, minimax and Bayes rules. Finite decision problems, risk set. Stein estimator.Correlation coefficient and distribution of its sample version in a bivariate normal population. Partial correlation coefficients. Classification problems, linear discriminant analysis. Principal component analysis. [9]

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

Contents

1 Maximum Likelihood Principle

Definition. Let $\{f(\cdot,\theta) : \theta \in \Theta\}$ be a statistical model of pdf for the law P of X, and consider observing $X_1, \cdots, X_n$ independent observations of X. The *likelihood function* of the model is:

$$l_n(\theta) = \prod_{i=1}^{n} f(X_i, \theta)$$

The *log-likelihood function* is:

$$l_n(\theta) = \log L_n(\theta) = \sum \log L(x_i, \theta)$$

The *normalized log-likelihood function* is $\bar{l_n}(\theta) = \frac{1}{n} l_n(\theta)$.

Definition. A *maximum likelihood estimator* (MLE) is any value $\hat{\theta} \in \Theta$ for which $L_n(\hat{\theta}) = \max_{\theta\in\Theta} L_n(\theta)$.

To solve for MLE,we take the zero of the *score function*, which is defined as:

$$S_n(\theta) = \nabla_\theta \ln(\theta)$$

1.1 Information Geometry and The Likelihood Function

For X a random variable of law P_θ on $X \subseteq \mathbb{R}^d$ and $g : X \to \mathbb{R}$. We write:

$$\begin{aligned}\mathbb{E}_\theta g(X) = \mathbb{E}_{P_\theta} g(x) &= \int g(x) \mathrm{d}P_\theta(x) \\ &= \int g(x) f(x,\theta) \mathrm{d}x \\ &= \sum g(x) f(x,\theta)\end{aligned}$$

where the last equality only holds if X is discrete.

Maximizing $L_n(\theta)$ is equivalent to maximizing $\bar{l_n}(\theta)$, which is an approximation of

$$l(\theta) = \mathbb{E}_{\theta_0}[\log(f(X,\theta))] = \int \log(f(X,\theta)) f(X,\theta_0) \mathrm{d}x$$

We have

$$l(\theta) - l(\theta_0) = \mathbb{E}_{\theta_0}\left[\log(\frac{f(X,\theta)}{f(X,\theta_0)})\right]$$

Recall Jensen's inequality. Since $\log$ is concave,

$$l(\theta) - l(\theta_0) \leq \log\left[\mathbb{E}_{\theta_0} \frac{f(X,\theta)}{f(X,\theta_0)}\right] = \log \int \frac{f(X,\theta)}{f(X,\theta_0)} f(X,\theta_0) \mathrm{d}X = \log(1) = 0$$

If we make the assumption of *strict identifiability*, as in $l(\theta) = l(\theta_0) \Rightarrow \theta = \theta_0$, then by the strict version of Jenson's inequality, we have $l(\theta) < l(\theta_0)$, so θ_0 is the unique maximizer.

Remark. $l(\theta_0) - l(\theta)$ can be interpreted as a distance between θ and θ_0. This is called the *Kulbach-Leibler distance* or divergence, or the entropy distance between $f(X,\theta)$ and $f(X,\theta_0)$.

1.2 Definitions and Elementary Theorems

Definition. In a parametric model, if $\frac{\partial}{\partial\theta}$ and integration $\int \cdot \mathrm{d}x$ can be interchanged, we say that the model is regular. In a regular model,we have:

$$E_\theta[\frac{\partial}{\partial\theta}\log f(X,\theta)] = \frac{\partial}{\partial\theta}\int f(X,\theta)\mathrm{d}x = 0$$

Now we define an important concept in this course:

Definition. For $\theta \in \int(\theta)$, we set for $\theta \subset \mathbb{R}^p$

$$I(\theta) = E_\theta[[\frac{\partial}{\partial\theta}\log f(x,\theta)][\frac{\partial}{\partial\theta}\log f(x,\theta)]^T]$$

We call $I(\theta)$, the $p \times p$ matrix, the *Fisher Information Matrix*.

Proposition. For all $\theta \in \int(\theta)$, in a regular model, we have:

$$I(\theta) = -E_\theta[\frac{\partial^2}{\partial\theta\partial\theta^T}\log f(x,\theta)]$$

Proof.

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

For the last step, the first integral is just 0 after taking about the derivative, and the second integral is just $I(\theta)$. □

1.3 Cramer-Rao Lower Bound

Theorem. Let $\{f(x,\theta), \theta \in \Theta\}$ be a regular statistical model, and $\hat{\theta}$ an unbiased estimator $\in \mathbb{R}$. Then:

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{nI(\theta)} \forall\theta \in \Theta$$

Proof. Remember the Cauchy-Schwarz inequality:

$$Cov^2(Y,Z) \leq \mathrm{Var}(Y)\,\mathrm{Var}(Z)$$

Let $Y = \hat{\theta}$ and $Z = \frac{\partial}{\partial\theta}\log f(X,\theta)$. Then $Cov(Y,Z) = E[YZ]$ ($E[Z] = 0$) and $\mathrm{Var}(Z) = I(\theta)$. Then:

$$E[YZ] = \int\hat{\theta}(x)\frac{\partial}{\partial\theta}f(x,\theta)\mathrm{d}x = \frac{\partial}{\partial\theta}\int\hat{\theta}(x)f(x,\theta)\mathrm{d}x = 1$$

As the integral is just the expectation of $\hat{\theta}$. Then the CS inequality rearranges to the required form. □

Now you ask, where is the n? This comes from the n samples, and we have to use $Z = \frac{\partial}{\partial\theta}\log\prod_i f(X_i,\theta)$, and $\mathrm{Var}(Z) = nI(\theta)$ here. Of course we also have a easy corollary:

Corollary. If $\hat{\theta}$ is not unbiased, we have $\mathrm{Var}(\hat{\theta}) \geq \frac{(\frac{\partial}{\partial\theta}E_\theta\hat{\theta})^2}{nI(\theta)}$.

1.3.1 Multivariate Cramer-Rao lower bound

Theorem. For $\theta \in \Theta \subseteq \mathbb{R}^p$, $p \geq 1$, consider functionals of the parameter $\Phi : \theta \to \mathbb{R}$. One shows in a similar manner that for any unbiased estimator $\tilde{\Phi}$ based on n iid observations $X_1, \cdots, X_n$ has a lower bound:

$$\mathrm{Var}_\theta(\tilde{\Phi}) \geq \frac{1}{n} \frac{\partial \Phi}{\partial \theta}(\theta)^T I^{-1}(\theta) \frac{\partial \Phi}{\partial \theta}(\theta)$$

For example consider:

$$\Phi(\theta) = \alpha^T \theta = \sum \alpha_i \theta_i \qquad \frac{\partial}{\partial \theta} \Phi(\theta) = \alpha$$

Then $\mathrm{Var}_\theta(\tilde{\Phi}) \geq \frac{1}{n} \alpha^T I^{-1}(\theta) \alpha$.

Example. Let $\binom{x_1}{x_2} = X \sim N(\theta, \Sigma)$ where $\theta = \binom{\theta_1}{\theta_2}$ and where θ is known. For a sample size fo 1:

Case 1 Consider estimation of θ_1, when θ_2 is known. Then, the model is one-dimensional with parameter θ_1 and the Fisher information is $I_1(\theta_1)$.

Case 2 When θ_2 is unknown, we have a two dimensional model and we care about $\theta_1 = \Phi(\theta)$. Applyingt he Cramer-Rao lower bound:

$$\mathrm{Var}_\theta(\tilde{\theta}_1) = \frac{\partial \Phi^T}{\partial \theta}(\theta) I^{-1}(\theta) \frac{\partial \Phi}{\partial \theta}(\theta) = I_\Phi(\theta_1)$$

2 Asymptotic Theory (for MLE)

At the very least, we need $\mathbb{E}(\tilde{\theta}_n) \to \theta$ as $n \to \infty$.

We would like the unbiased estimator $\tilde{\theta}$ to converge to θ when $n \to \infty$ as we take $X_1, \cdots X_n$ samples from P_θ. This is called *consistency*.

The best we can hope for is that $n \,\mathrm{Var}_\theta(\tilde{\theta}_n) \to I^{-1}(\theta)$ as $n \to \infty$. This is called *asymptotic efficiency*.

Definition. Let $(X_n, n \geq 0)$, X be random vectors in a probability space $(\Omega, A, \mathbb{P})$.

- X_n converges to X *almost surely*, denoted $X_n \xrightarrow{a.s.} X$ if:

$$\mathbb{P}(w \in \Omega : \|X_n(w) - X(w)\| \to 0 as n \to \infty) = 1$$

- X_n converges to X in *probability*, denoted $X_n \xrightarrow{P} X$ if:

$$\mathbb{P}(\|X_n - X\| > \epsilon) \to 0$$

as $n \to \infty$ for all $\epsilon > 0$.

Remark. For a vector $X_n \in \mathbb{R}^p$. It is equivalent to have $X_n(j) \to X(j)$ for all $1 \leq j \leq p$ and to have $X_n \to X$.

Definition. Let X_n, X be random vectors in $\mathbb{R}^k$. $X_n \xrightarrow{d} X$ as $n \to \infty$ or X_n converges to Xin *distribution* if $\mathbb{P}(X_n \leq t) \to \mathbb{P}(X \leq t)$ for all t where the map $t \to \mathbb{P}(X \leq t)$ is continuous.

Proposition.

- $X_n \xrightarrow{a.s.} X$ implies $X_n \xrightarrow{P} X$ implies $X_n \xrightarrow{d} X$ as $n \to \infty$.
- X_n, X with values in $\mathcal{X} \subseteq \mathbb{R}^d$ and $g : \mathcal{X} \to \mathbb{R}$ a continuous function. Then $X_n \to X$ a.s./in P/in d implies $g(X_n) \to g(X)$ a.s./in P/in d. This is called the *continuous mapping theorem.*

Slutsky's Lemma Suppose $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$, where c is a deterministic random variable that takes the same value with probability 1. Then as $N \to \infty$:

(i) $Y_n \xrightarrow{P} c$

(ii) $X_n + Y_n \xrightarrow{d} X + c$

(iii) $X_n Y_n \xrightarrow{d} cX$ and if $c \neq 0$, $\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}$.

(iv) If A_n are random matrices such that $(A_n)_{ij} \xrightarrow{P} A_{ij}$ where A_{ij} are deterministic, then
$$A_n X_n \xrightarrow{d} AX$$

(v) If $X_n \xrightarrow{d} X$ as $n \to \infty$, then $(X_n)_{n\in\mathbb{N}}$ is bounded in probability, or $X_n = O_p(1)$. This means that for all $\epsilon > 0$, there exists $M(\epsilon) < \infty$ such that $\mathbb{P}(\|X_n\| > M(\epsilon)) < \epsilon$.

2.1 Law of Large Numbers and Central Limit Theorem

Proposition (Weak law of Large Numbers (WLLN)). For $X_a, \cdots X_n$ iid copies from $X \sim P$ with $\mathrm{Var}(X) < \infty$, we have:

$$\bar{X}_n - \frac{1}{n}\sum X_i \xrightarrow{P} E(X)$$

Proof. $\mathrm{Var}(\frac{1}{n}\sum_i X_i - E(X)) = \frac{\mathrm{Var}(x)}{n}$. Then by Chebyshev's inequality:

$$P(|\frac{1}{n}\sum(X_i - E(X))| > \epsilon) \leq \frac{\mathrm{Var}(x)}{n\epsilon^2}$$

For fixed ϵ, this goes to 0 as $n \to \infty$, so we have convergence in probability. □

Theorem (Strong Law of Large Numbers (SLLN)). $X_1, \cdots X_n$ iid copies of $X \sim P$ in $\mathbb{R}^k$. Assume $E[\|X\|] < \infty$, then:

$$\bar{X}_n = \frac{1}{n}\sum_i X_i \xrightarrow{a.s.} E[X]$$

Nope, no proof. Nada. Can't do.

Theorem (Multivariate Central Limit Theorem (CLT)). Let $X_i, \cdots, X_n$ be iid copies of $X \sim P$ with $Cov(X) = \Sigma$ positive definite. Then we have:

$$\sqrt{n}(\frac{1}{n}\sum X_i - E(X)) \xrightarrow{d} n(0, \Sigma) \quad n \to \infty$$

Definition. In a parametric model, a consistent estimator $\tilde{\theta}_n$ is called asymptotically efficient if:

$$\lim_{n\to\infty} n \operatorname{Var}_{\theta_0}(\tilde{\theta}_n) \to I(\theta_0)^{-1}$$

Theorem. Let $f(\cdot,\theta)$ be a statistical model. Let $X_1, \cdots, X_n \sim P_{\theta_0}$ iid and $\hat{\theta}_n$ is the MLE. then:

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

No proof either.

2.2 Consistency of MLE

Definition. Consider $X_1, \cdots X_n$ iid observations. An estimator $\tilde{\theta}$ is consistent if $\tilde{\theta} \xrightarrow{P} \theta$ for all $\theta \in \Theta$.

Now we will introduce some assumptions for the model that we will assume for the rest of the notes unless *otherwise stated.*

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

(ii) $\int_X f(x,\theta)\mathrm{d}x = 1 \quad \forall \theta \in \Theta$.

(iii) $f(x,\cdot) : \theta \to f(x,\theta)$ is continuous for all $x \in X$.

(iv) Θ is compact. ***This is the assumption that we would violate the most.***

(v) $\theta = \theta' \Leftrightarrow f(\cdot,\theta) = f(\cdot,\theta')$.

(vi) $E_{\theta_0} \sup_{\theta\in\Theta} |\log f(x,\theta)| < \infty \; \forall \theta_0 \in \Theta$.

Theorem (MLE Consistency)**.** Let $X_1, \cdot X_n$ be iid samples from a model satisfying the assumptions above, then an MLE exist, and any MLE is consistent.

Proof. First remember MLE is a maximum of $\bar{l}_n(\theta) = \frac{1}{n}\sum \log f(X_i,\theta)$. Since each term is continuous in θ with Θ compact, there exists a maximizer. Then one can go back to 1.1 to see that assumptions $1, 2, 5, 6$ guarantee the use of the strict Jensen inequality and this is the unique maximizer. Call it θ_0.

Define $\Theta_\epsilon = \{\theta \in \Theta : \|\theta - \theta_0\| \geq \epsilon\}$. Note that θ_ϵ is compact. Since l is continuous, we have that $l(\epsilon) = \sup_{\theta\in\Theta_\epsilon} l(\theta) < l(\theta_0)$ because θ_0 is a unique maximum. We can choose $0 < \delta(\epsilon) < \frac{l(\theta_0)-l(\epsilon)}{2}$. Apply triangle inequality,

$$\sup_{\theta\in\Theta_\epsilon} \bar{l}_n(\theta) = \sup_{\theta\in\Theta_\epsilon} l(\theta) + \sup_{\theta\in\Theta}[\bar{l}_n(\theta) - l(\theta)]$$

Consider events $A_n(\epsilon) = \{\sup_{\theta\in\Theta} |\bar{l}_n(\theta) - l(\theta)| < \delta(\epsilon)\}$. On these events, $\sup_{\theta\in\Theta_\epsilon} \bar{l}_n(\theta) \leq \bar{l}_n(\theta_0)$, which implies that $\|\hat{\theta}_n - \theta_0\| < \epsilon$. (As $\hat{\theta}_n$ uniquely maximizes $\bar{l}_n(\theta)$. Then we use the *uniform law of large numbers*, which states the $P(A_n(\epsilon)) \to 1$ for all ϵ as $N \to \infty$, givevn that Θ is compact, $f(x,\theta)$ is continuous almost everywhere, and there is a a dominating $d(x)$ function of $f(x,\theta)$ with $E[d(X)] < \infty$. Thus we have:

$$P(\|\hat{\theta}_n - \theta_0\| < \epsilon) \to 1$$

Thus we are done. □

Definition. In a parametric model, a consistent estimator $\tilde{\theta}_n$ is called *asymptotically efficient* if $\lim_{n\to\infty} n \operatorname{Var}_{\theta_0}(\tilde{\theta}_n) \to I(\theta_0)^{-1}$.

Theorem. Let the model $\{f(\cdot,\theta), \theta \in \Theta\}$ satisfy assumptions above. Let $X_1, \cdots X_n \sim P_{\theta_0}$ and $\hat{\theta}_n$ is the MLE. Then we have:

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

Proof. It is sufficient to consider this statement on a sequence of events E_n such that $P(E_n^c) \to 0$ as $n \to \infty$. Take a small area K around θ_0 and let $E_n = \{\hat{\theta}_n \in K\}$. Then we have $\frac{\partial}{\partial\theta}\bar{l}_n(\theta)|_{\theta=\hat{\theta}_n} = 0$. Further, we assume that $\bar{l}_n$ is C^2, so the derivative is continuous along all lines in k. In particular consider the line between $\hat{\theta}_n$ and θ_0. We have:

$$0 = \begin{pmatrix} \frac{\partial}{\partial\theta_1}\bar{l}_n(\hat{\theta}_n) \\ \cdots \\ \frac{\partial}{\partial\theta_p}\bar{l}_n(\hat{\theta}_n) \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial\theta_1}\bar{l}_n(\theta_0) \\ \cdots \\ \frac{\partial}{\partial\theta_p}\bar{l}_n(\theta_0) \end{pmatrix} + \begin{pmatrix} \frac{\partial^2}{\partial\theta_1\partial\theta_1}\bar{l}_n(\theta)|_{\theta=\bar{\theta}_1} & \frac{\partial^2}{\partial\theta_1\partial\theta_2}\bar{l}_n(\theta)|_{\theta=\bar{\theta}_2} & \cdots \\ \vdots & \vdots & \vdots \end{pmatrix} \cdot (\hat{\theta}_n - \theta_0)$$

The matrix is denotes $\bar{A}_n$, and we have:

$$\begin{aligned}(\bar{A}_n)_{kj} = \frac{1}{n}\sum_i \left(\frac{\partial^2}{\partial\theta_k\partial\theta_j}\log f(X_i,\bar{\theta}_j) - E_{\theta_0}\frac{\partial^2}{\partial\theta_k\partial\theta_j}\log f(X_i,\bar{\theta}_j)\right) \\ + E_{\theta_0}\frac{\partial^2}{\partial\theta_k\partial\theta_j}\log f(X_i,\bar{\theta}_j) - E_{\theta_0}\frac{\partial^2}{\partial\theta_k\partial\theta_j}\log f(X_i,\bar{\theta}_0) \\ + (-I(\theta_0))_{kj}\end{aligned}$$

So $\bar{A}_n = I + II - I(\theta_0)$. I goes to 0 from uniform law of large numbers, since both functions are continuous. By continuous mapping theorem and consistency of the estimator, II goes to 0.

Thus $\bar{A}_n \xrightarrow{P} -I(\theta_0)$ as $n \to \infty$. By non-singularness of $I(\theta_0)$, $\bar{A}_n$ is eventually non-singular. Thus we have:

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

Expanding the definition of $\bar{l}_n(\theta_0)$, by CLT, we have:

$$\sqrt{n}\frac{\partial}{\partial\theta}\bar{l}_n(\theta_0) \xrightarrow{d} N(0, I(\theta_0))$$

Then using slutsky's lemma, we have the result as required. □

2.3 Plug-in MLE and Delta Method

Theorem (Plug-in Theorem)**.** Let $\Phi : \theta \to \mathbb{R}$ be differentiable at θ_0 and $\hat{\theta}_n$ be estimators such that $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} Z$ when $n \to \infty$. Then:

$$\sqrt{n}(\Phi(\hat{\theta}_n) - \Phi(\theta_0)) \xrightarrow{d} \frac{\partial\Phi^T}{\partial\theta}Z$$

The proof is quite simple, by noting that:

$$\sqrt{n}(\Phi(\hat{\theta}_n) - \Phi(\theta_0)) = \sqrt{n}\frac{\partial\Phi^T}{\partial\theta}(\hat{\theta}_n - \theta_0) + o(\sqrt{n}\|\hat{\theta}_n - \theta_0\|)$$

And using the continuous mapping theorem.

3 Asymptotic Inference with MLE

For a parametric model $\{Lo(\cdot,\theta),\theta\in\Theta\}$, $\theta\subseteq\mathbb{R}^p$ with the same regularity assumptions, suppose we are interested in a confidence interval around $\theta_j=e_j^T\theta$, then we know:

$$\sqrt{N}(\hat{\theta}_{n.j}-\theta_j)\xrightarrow{d}N(0,e_j^TI(\theta)^{-1}e_j)$$

And then we can construct a confidence interval.

Definition. For all $\theta\in\Theta$, the $p\times p$ matrix $l_n(\theta)=\frac{1}{n}\sum\frac{\partial}{partial\theta}\log f(X_i,\theta)\frac{\partial}{\partial\theta}\log f(X_i,\theta)$ is called the *observed Fisher information.*

Proposition. Under regularity assumptions, we have $\hat{l}_n=l_n(\hat{\theta}_{MLE})\xrightarrow{P}I(\theta_0)$.

The proof is almost exactly the same as the one proving consistency of MLE above.

3.1 Hypothesis Testing

First, a review of a theorem that we did not prove in IB:

Theorem (Wilks Theorem)**.** Suppose $\{f(\cdot,\theta)\}$ satisfy regularity assumptions and $\Theta_0=\{\theta_0\}$, then:

$$\Lambda(\theta,\{\theta_0\})=2\log\left(\frac{\prod f(X_i,\hat{\theta}_{MLE})}{\prod f(X_i,\hat{\theta}_{MLE,0})}\right)\xrightarrow{d}\chi_p^2$$

There is an outline of a proof, but it is not important.From IB, we know we can use the score test statistic to test the simple hypothesis. But we introduce a new one:

4 Bayesian Inference

4.1 Basic Ideas, Prior, and Posterior Distribution

Consider a parametric model $\{f(\cdot,\theta),\theta\in\Theta\}$ where we view θ as random, drawn from a distribution Π, the prior on Θ.

Formally, in a parametric model $\{f(\cdot,\theta),\theta\in\Theta\}$. Given θ, the distribution of X is $f(x,\theta)$ Written $X|\theta\sim f(x,\theta)$.

Definition. Let X be a sample space for X, and the product space $X\times\Theta$ with the probability measure Q with distribution $Q(x,\theta)=f(x,\theta)\pi(\theta)$. By rules of probability, we have:

$$X|\theta\sim\frac{f(x,\theta)\pi(\theta)}{\int f(x',\theta)\pi(\theta)\mathrm{d}x'}$$

The construction gives us what we want for $X|\theta$. The distirbution π over Θ is called the prior. And the *posterior distribution* is defined as the law of $\theta|X$:

$$\theta|X\sim\Pi(\theta|X)$$

A *conjugate prior* is a distribution x such that after sampling, the posterior belongs to the same class of distributions.

Example. - Normal Prior + Normal Sampling $\to$ Normal Posterior

- Beta Prior + Binomial sampling $\to$ Beta posterior

Note. We can also have an *improper prior*, where π is not a probability distribution as long as the integral in the denominator is finite. Note this doesn't require the integral of π to be finite.

What do we do with this?

Theorem. For a regular parametric model with a continuous prior π such that $\pi(\theta_0) > 0$ and a posterior density Π_n, and $\hat{\phi}_n = N(\hat{\theta}_{MLE}, \frac{I(\theta_0)^{-1}}{n})$, we have that as $n \to \infty$:

$$\int_{\mathbb{R}^p} |\Pi_n(\theta) - \hat{\phi}_n(\theta)| \mathrm{d}\theta \xrightarrow{a.s.} 0$$

[Proof is only tested for specific cases] Now using this theorem, consider $C_n = \{\theta : |\theta - \hat{\theta}_{MLE}| \leq \frac{R_n}{\sqrt{n}}\}$ where R_n is chosen so that $\Pi(C_n | X_1, \cdots X_n) = 1 - \alpha$.

Now define:

$$\Phi(t) = \int_{-t}^{t} \mathrm{d}N(0, I^{-1}(\theta_0))(v)\mathrm{d}v$$

This is monotonically increasing on $[0, \infty)$ to $[0, 1]$. Thus there exists a well-defined Φ^{-1} function such that $\Phi^{-1}(\Phi(t)) = t$. Then:

$$\Phi(R_n) = \int_{-R_n}^{R_n} \mathrm{d}N(0, I^{-1}(\theta_0))(v)\mathrm{d}v = \int_{C_n} \mathrm{d}N(\hat{\theta}_{MLE}, \frac{I(\theta_0)^{-1}}{n})(\theta)\mathrm{d}\theta$$

by change of variables $v = \sqrt{n}(\theta - \hat{\theta}_{MLE})$. This equals:

$$\int_{C_n} (\hat{\phi} - \pi_n)(\theta)\mathrm{d}\theta + 1 - \alpha$$

Where the first integral goes to 0 by the theorem above. So $\Phi(R_n) \xrightarrow{a.s.} 1 - \alpha$.

5 Decision Theory

Consider an observation x from a parametric model $\{f(\cdot, \theta), \theta \in \Theta\}$. In a decision problem, we consider an action space $\mathcal{A}$, and a decision rule δ with $\delta : X \to \mathcal{A}$.

Example.
- The action space can be binary $\{0, 1\}$, then $\delta(x)$ is a test 0/1 rule.
- The action space is θ and δ is an estimation algorithm.

The performance of a decision rule is measured by a *loss function*:

$$L = A \times \theta \to [0, \infty)$$

Definition. For a decision rule δ, loss function L, and $X \sim P_\theta$, we have:

- A *risk function* is a function $R(\delta, \theta) = E_\theta[L(\delta(x), \theta)]$
- A minimax rule is a minimizer of $R_m(\delta) = \sup R(\delta, \theta)$.
- For any prior π, the π Bayes risk is $R_\pi(\delta) = \int R(\delta, \theta)\pi(\theta)\mathrm{d}\theta$.

- A $\pi-$ Bayes decision rule is a minimizer of $R_\pi(\delta)$.
- A prior λ is called a *least favorable prior* if $R_\lambda(\delta_\lambda) \geq R_{\lambda'}(\delta_{\lambda'})$ for all λ'.

From the definition, we can see a relatively easy proposition:

Proposition. The minimizer of Π-posterior risk minimizes the Bayes risk.

Example. For quadratic risk, the unique Bayes rule is the posterior mean. If we consider the absolute risk, the decision rule is the posterior median.

Proposition. Let $\delta(x)$ be an unbiased estimator of θ for all $\theta \in \Theta$. If δ is also a Bayes rule, then:

$$E[(\delta(x) - \theta)^2] = 0$$

Proof. Now $E[\theta\delta(x)] = E[E^\Pi[\theta\delta(X)|X]] = E[\delta(X)^2]$ as under the quadratic loss , the posterior mean is the decision rule. But similarly we have this equal yo $E[\theta]$ by conditioning on θ. So then:

$$E[(\delta(x) - \theta)^2] = 0$$

□

Proposition. Let λ be a prior on θ such that $R_\lambda(\delta_\lambda) = R_m(\delta_\lambda) = \sup_{\theta\in\theta} R(\delta_\lambda, \theta)$. Then

- δ_λ is minimax.
- If δ_λ is a unique λ Bayes rule, then it is a unique minimax.
- λ is least favorable, meaning, for any other prior, we have $R\lambda(\delta_\lambda) \geq R_{\lambda'}(\delta_{\lambda'})$.

Proof.
- δ_λ is minimax is equivalent to saying that δ_λ is minimizer of $R_m(\delta)$. Let δ be any other decision rule:

$$R_m(\delta) = \sup_{\theta\in\Theta} R_m(\delta, \theta) = \int_\theta R(\delta, \theta)\lambda(\theta)\mathrm{d}\theta \geq \int_\theta R(\delta_\lambda, \theta)\lambda(\theta)\mathrm{d}\theta = R_m(\delta_\lambda)$$

- If δ_λ is unique Bayes, then the inequality above is strict and δ_λ is unique minimax.
- This follows striaght from expanding the definition.

□

A sometimes useful corollary is thus:

Corollary. If a unique Bayes rule δ_λ is with constant risk, it is unique minimax.

5.1 Admissability

Definition. A decision rule δ is inadmissible if there exists δ' such that $R(\delta', \theta) \leq R(\delta, \theta)$ for all $\theta \in \Theta$ and $R(\delta', \theta) < R(\delta, \theta)$ for some $\theta \in \Theta$.

Lemma. In a $N(\theta, 1)$ model, the MLE $\hat{\theta} = \bar{X}_n$ for $\theta \in \Theta = \mathbb{R}$ is admissible and minimax.

Proof. For the MLE, the quadratic risk is $\frac{1}{n}$, which is constant in θ. Thus, since an admissible and constant risk MLE is automatically minimax, we only show that $\bar{X}_n$ is admissible. Let δ be any decision rule for estimating θ and compute its risk:

$$R(\delta, \theta) = E_\theta[(\delta - \theta)^2] = E_\theta[(\delta - E_\theta(\delta))] + (E_\theta[\delta] - \theta)^2 \geq \mathrm{Var}_\theta(\delta) + B^2(\theta)$$

letting $B(\theta) = E_\theta[\delta] - \theta$. Then the CR inequality, in its general version, gives:

$$\mathrm{Var}_\theta(\delta) \geq \frac{(\frac{\mathrm{d}}{\mathrm{d}\theta} E_\theta(\delta))^2}{n} = \frac{(1 + B'(\theta))^2}{n}$$

So if δ has a lower risk thatn $\bar{X}_n$, then:

$$B^2(\theta) + \frac{(1 + B'(\theta))^2}{n} \leq \frac{1}{n}$$

This implies that:

- $|B(\theta)| \leq \frac{1}{\sqrt{n}}$
- $B'(\theta) \leq 0$

Thus there exists a sequence $\theta_i \to -\infty$ for which $B'(\theta_i) \to 0$. Otherwise we have $B'(\theta) \leq -\epsilon$ for some $\epsilon > 0$ and B is unbounded. Looking at the equation above, we need $B(\theta_i) \to 0$.

Applying the same reasoning to $\theta_i \to \infty$, we obtain that $B(\theta) = 0$ for all $\theta \in \Theta$. So if δ has a lower risk than $\bar{X}_n$, then it is unbiased, so $R(\delta, \theta) \geq \frac{1}{n}$ from the CR inequality, so $R(\delta, \theta) \geq R(\bar{X}_n, \theta)$. □

Remark. Here the MLE $\bar{X}_n$ is not a Bayes estimator, for any prior Π. So it is a decision rule that is admissible but not Bayes. It is however, limiting Bayes, in the sense that it is the limit of $\delta_{\Pi_{v^2}} : v^2 \to \infty$ for a $N(\theta, v^2)$ prior.

In general, any admissible decision rule can be obtained as such a limit.

Now on theorem 2, in the case when $p \geq 3$, we introduce the James-Stein estimator, where $\delta_i^{JS} = \left(1 - \frac{p-2}{\|X\|^2}\right) X_i$, where $X \sim N(\theta, I_p)$, $\theta \in \Theta$, and $\|X\|^2 = \sum X_i^2$.

We are going to show that $R(\delta^{JS}, \theta) \leq R(X, \theta) = p$ and $R(\delta^{JS}, \theta) < p$ for some $\theta \in \mathbb{R}^p$. This makes the usual estimator inadmissible!

Lemma (Stein's Lemma)**.** Let $X \sim N(\theta, 1)$ and $g : \mathbb{R} \to \mathbb{R}$ be differentiable and bounded such that $E|g'(x)| < \infty$. Then $E_\theta[g(X)(X - \theta)] = E_\theta[g'(X)]$.

Proof.

$$E_\theta[g(X)(X - \theta)] = \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} g(X)(X - \theta) e^{-(X-\theta)^2/2} \mathrm{d}X$$

By integration by parts, we have:

$$= \frac{-1}{\sqrt{2\pi}} g(x) e^{-(x-\theta)^2/2} |_{-\infty}^{\infty} + \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} g'(X) e^{-(x-\theta)^2/2} \mathrm{d}x$$

Which gives the result as the first term is 0 since g is bounded. □

Now we compute the risk. With some determination, we have:

$$R(\delta^{JS},\theta) = p + (p-2)^2 E[\frac{1}{\|x\|}^2] - 2(p-2)\sum_j E[E[\frac{(x_j-\theta_j)x_j}{\|x\|^2}|X_{(-j)}]]$$

where $X_{(-j)}$ means condition on all observations except j. Then we apply Stein's lemma on the laste term, where $g(X_j) = \frac{X_j}{X_j^2+\sum_{i\neq j}X_i^2}$. Then we have:

$$\begin{aligned} R(\delta^{JS},\theta) &= p + (p-2)^2 E[\frac{1}{\|x\|}^2] - 2(p-2)\sum_j E[\sum \frac{\|X\|^2 - 2X_j^2}{\|X\|^4}] \\ &= p + (p-2)^2 E[\frac{1}{\|x\|}^2] - 2(p-2)\left[pE[\frac{1}{\|x\|^2}] - 2E[\frac{1}{\|x\|^2}]\right] \\ &= p - (p-2)^2 E[\frac{1}{\|X\|^2}] < p \end{aligned}$$

This is fascinating! Actually, later work has shown that the James Stein estimator is in fact *also* inadmissible!

6 Classification Problems

We have two distributions over X, with $X \sim f_1$ or $X \sim f_2$. our job is to decide on the basis of observations which one is it. A *classification rule* δ is given by a region R of X such that:

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

Lets have a priort hat says the probability that it is f_1 is q_1. Given $X = x$, the bayes probability is:

$$\Pi(1|X=x) = \frac{q_1 f_1(x)}{q_1 f_1(x) + q_2 f_2(x)} \qquad \Pi(2|X=x) = 1 - \Pi(1|X=x)$$

So the posterior decision rule is:

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

Proposition. The rule δ_Π is the minimizer of the Bayes classification risk. If

$$P_{f_1}\left(\frac{f_1(X)}{f_2(X)} = \frac{1-q_1}{q_1}\right) = 0$$

Then δ_π is a unique Bayes rule (and thus admissible)

Proof. let J be any classification region.Let $q_1 = q$. The error is:

$$\begin{aligned} & q\int_{J_c} f_1(x)\mathrm{d}x + (1-q)\int_J f_2(x)\mathrm{d}x \\ &= \int_{j_c} \left[qf_1(x) - (1-q)f_2(x)\right]\mathrm{d}x + \int_X (1-q)f_2(x)\mathrm{d}x \end{aligned}$$

The first part is minimized when J^c is exactly the set of $x \in X$ such that $qf_1(x) - (1-q)f_2(x) \leq 0$. This descibrs a unique one if the boundary has probability 0. $\square$

7 Further Topics

7.1 Multivariate Analysis for Statistics

Definition. The correlation is defined as:

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} \qquad \hat{\rho}_{X,Y} = \frac{\sum(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sqrt{\sum(X_i - \bar{X}_n)^2 \sum(Y_i - \bar{Y}_n)^2}}$$

In the case of a Gaussian model, the distirbution of the sample correlation is:

$$f(x) = \frac{\Gamma(\frac{1}{2}(n-1))}{\Gamma(\frac{1}{2}(n-2))}(1-x^2)^{\frac{1}{2}(n-4)}$$

7.2 Resampling Techniques and the Bootstrap

Let $X_1, \cdots X_n$ be iid and T_n an estimator for θ. Let $B_n(\theta) = E_\theta T_n - \theta$. One way to reduce the bias is to estimate it, and to subtract it from T_n. From $1 \leq i \leq n$, we consider $T_{(-i)} = T(X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n)$, so the estimator without observation i.

Definition. The *jacknife bias estimator* is:

$$\hat{B}_n = (n-1)(\frac{1}{n}\sum_i T_{(-i)} - T_n)$$

And the corrected estimator is $\tilde{T} = T_n - \hat{B}_n$. For regular bias function $B_n(\theta) = O(\frac{1}{n})$, we have that $|E\tilde{T} - \theta| = O(\frac{1}{n^2})$.

7.2.1 Bootstrap

What if generating samples are expensive? This would make our asymptotic results invalid. Can we do anything?

Now for a sample X_1, X_n, define a random variable X_n^b with law $P_n(X_n^b = X_i) = \frac{1}{n}$ for all i. We then resample X_i uniformly, independently, with replacement, n times. So we then have $X_{n.i}^b$ drawn from population with mean $\bar{X}_n$ and we have $\bar{X}_n^b = \frac{1}{n}\sum_i X_{n,i}^b$ estimates $\bar{X}_n$. Why are we doing this? We hope the distribution of $\bar{X}_n^b - \bar{X}_n$ estimates the distribution of $\bar{X}_n - \bar{X}$. We have the following theorem:

Theorem (Consistency of Bootstrap of the Mean)**.** Let $X_1, \cdots, X_n$ iif from P with finite variance. Taking Φ the cdf of $N(0, \sigma^2)$, as $N \to \infty$:

$$\sup_{t\in\mathbb{R}} |P_n(\sqrt{n}(\bar{X}_n^b - \bar{X}_n) \leq t | X_1, \cdots, X_n) - \Phi(t)| \xrightarrow{a.s.} 0$$

Proof. We use a lemma to prove this trivially:

Proposition. Let $(Z_{n,i}, i = 1, \cdots, n)$ iid for every fixed n such that

- $\forall \delta > 0$, $nP(|Z_{n1}| > \sqrt{n}\delta) \to 0$ as $n \to \infty$.
- $\mathrm{Var}(Z_{n1}\mathbf{1}\{|Z_{n1}|\sqrt{n}\}) \to \sigma^2$ as $n \to \infty$
- $\sqrt{n}E[Z_{n1}\mathbf{1}\{|Z_{n1}| > \sqrt{n}\}] \to 0$ as $n \to \infty$.

Then:

$$\sqrt{n}\left(\frac{1}{n}\sum_i (Z_{ni} - E[Z_{ni}])\right) \xrightarrow{d} N(0, \sigma^2)$$

Applying this theorem with $Z_{ni} = X_{ni}^b$ gives the result in distribution. An integration then results in the sup result required.

Checking these conditions are just tedious so it will be omitted. No smart trick needed. □

Then we can use bootstrap to:

- Use $\sqrt{n}(\hat{\theta}_n^b - \hat{\theta}_n)$ as a proxy for the derivatives $\sqrt{n}(\hat{\theta}_n - \theta_0)$..
- Now using the original $\hat{\theta}_{MLE}$ we can resample $\hat{X}_i^b$ and substitute this values and calculate a new $\hat{\theta}_{MLE,b}$. Then $\hat{\theta}_{MLE,b} - \hat{\theta}_{MLE}$ as a proxy for $\hat{\theta}_{MLE} - \theta_0$.

7.3 Monte-Carlo Methods

How can a computer sample general distributions?

Definition. We define the generalized inverse of F:

$$F^{-1}(u) = \int \{x : u \leq F(x)\}$$

Proposition. For $U_1, \cdots, U_N$ iid uniform with cdf F and independent $X_i = F^{-1}(U_i)$. For a given function g, we have:

$$\frac{1}{N}\sum_i g \circ F^{-1}(U_i) \to E_F(g(x))$$

Now suppose we have a distribution f that we want to sample, but sampling f is hard and h is easy, and we have $f \leq Mh$ for some $M > 0$. Then we have:

Proposition (Accept-Reject Algorithm). - Generate $X \sim h, U \sim U(0,1)$

- Accept sample if $U \leq \frac{f(x)}{Mh(x)}$ or start again.

This would generate samples with distribution f.

7.4 Nonparametric Models

Now consider a large sample $X_1, \cdot X_n$ and the empirical cdf from this $F_n(t) = \frac{\#\text{ X is} \leq t}{n}$. We have the following theorem:

Theorem (Gliverto-Cartelli). We have, for $X_1, \cdots, X_n \sim P$, when $n \to \infty$:

$$\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \xrightarrow{a.s.} 0$$

Proof. Expand the definition and you have the uniform law of large numbers. Done. □

Theorem (Kolmogorov-Smirnov)**.**

$$\sqrt{n}\sup_{t\in\mathbb{R}}|F_n(t)-F(t)| \xrightarrow{d} \|B\|_\infty$$

Where B is a standard Brownian bridge, which is a brownian walk (continuous random walk) with $B(0)=B(1)=0$.

Note. The actual distribution really doesn't matter. You just need to know that it converges and we can do all the normal statistics inference tests we have.