Principles of Statistics **Part II - Michaelmas 2018**

Lecture 1: Introduction

Lecturer: Quentin Berthet

This course is concerned with presenting some of the mathematical principles of statistical theory. One of the general objectives of statistics is to "reverse-engineer" probability, i.e. to make a statement about an unknown probability distribution, given access to draws from this distribution. To make this statement more precise, we describe the following formalism

Consider a real-valued random variable X, on a probability space Ω, with distribution defined for all $t \in \mathbf{R}$ by

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

When X is discrete it is equal to

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

and f is called the probability mass function of X (p.m.f.). When X is continuous it is equal to

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

and f is called the probability density function of X (p.d.f.). Many problems in statistics are concerned with determining the distribution of a *sample*, n independent copies $X_1, \cdots, X_n$ of X. We refer to n as the *sample size*. Often, the distribution belongs to a certain class that can be parametrized by an unknown θ.

Definition 0.1. A *statistical model* for a sample from X is any family

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

of p.m.f. or p.d.f. $f(\theta, \cdot)$, or of probability distribution P_θ for the law of X. The index set Θ is called the *parameter space*

Example 0.1. Some statistical models and their parameter spaces

i) $\mathcal{N}(\theta, 1)\,;\ \theta \in \Theta = \mathbf{R}$.

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

Informal notes, based on past lecture notes by Richard Nickl. Please let me know of any errors.

iii) $\text{Exp}(\theta)$; $\theta \in \Theta = (0, \infty)$.

iv) $\mathcal{N}(\theta, 1)$; $\theta \in \Theta = [-1, 1]$.

DEFINITION 0.2. For a variable X with distribution P, we say that the model $\{P_\theta : \theta \in \Theta\}$ is *correctly specified* if there exists $\theta \in \Theta$ such that $P_\theta = P$.

We will often write θ_0 for the true value of θ to distinguish it from other elements of the parameter space Θ. We will say that the X_i are i.i.d from the model $\{P_\theta : \theta \in \Theta\}$ in this case. As an example, if $X \sim \mathcal{N}(2, 1)$ the model in *i)* is correctly specified but the model in *iv)* is not.

Somewhat more formally, some of the main goals of statistics are the following problems:

1. ESTIMATION: Constructing $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$, i.e. a function of the observations, such that for all $\theta \in \Theta$, when $X_i \sim P_\theta$, the *estimator* $\hat{\theta}$ is close to θ.
2. TESTING HYPOTHESES: Determining whether we are under the null hypothesis $H_0 : \theta = \theta_0$ or the alternative $H_1 : \theta \neq \theta_0$, by a *test* $\psi_n = \psi(X_1, \dots, X_n)$ such that $\psi_n = 0$ when H_0 is true, and $\psi_n = 1$ when H_1 is true, with high probability.
3. INFERENCE: Find intervals, or sets, of *confidence* $\mathcal{C}_n = \mathcal{C}(X_1, \dots, X_n)$ such that for $0 < \alpha < 1$ we have $P_\theta(\theta \in \mathcal{C}_n) = 1 - \alpha$ (or $\geq 1 - \alpha$), for all $\theta \in \Theta$, where α is the significance level. This is useful for uncertainty quantification.

1. The likelihood principle.

1.1. *Basic ideas and concepts.* We study the following example: let $X_1, \dots, X_n$ be i.i.d. from a Poisson model $\{\text{Poi}(\theta) : \theta \geq 0\}$, with numerical values $X_i = x_i$ for $1 \leq i \leq n$. The joint distribution of the sample is

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

It is the probability of occurence of this particular sample $(X_1 = x_1, \dots, X_n = x_n)$, as a function of the unknown parameter $\theta \geq 0$. One of the first principles of this course is that it is helpful to think of $L_n(\cdot)$ as a random function of Θ to $\mathbf{R}$, the randomness coming from the X_i.

The idea of the likelihood principle is to find θ that maximizes the above probability. If the X_i are continuous, we use the p.d.f for $f(x_i, \theta)$. In the current example, it is equivalent to maximize $\log(L_n(\theta)) = \ell_n(\theta)$ over $(0, \infty)$.

$$\ell_n(\theta) = -n\theta + \log(\theta) \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \log(x_i!)\,.$$

Taking a first order condition *i.e.* setting $\ell_n'(\theta) = 0$ gives the equation

$$-n + \frac{1}{\theta} \sum_{i=1}^{n} x_i = 0\,,$$

which has solution $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} X_i$, the sample mean. This is indeed optimal as one can check that $\ell_n''(\theta) < 0$ for all $\theta > 0$. The case where all the x_i are zero can be checked by hand: in this case, maximizing ℓ_n is directly equivalent to maximizing $-n\theta$ and $\hat{\theta} = 0$.

Principles of Statistics **Part II - Michaelmas 2018**

Lecture 2: Maximum likelihood estimator

Lecturer: Quentin Berthet

We have recalled in the past lecture, through an example, the principle of maximizing the likelihood. Formally, we define the following.

Definition 1.1. Let $\{f(\cdot, \theta) : \theta \in \Theta\}$ be a statistical model of p.d.f./p.m.f. $f(x, \theta)$ for the distribution P of a random variable X and consider observing n realisations x_i, for $1 \leq i \leq n$ of i.i.d. copies $X_1, \dots, X_n$ of X.

The *likelihood function* of the model is defined as

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

The *log-likelihood function* of the model is defined as

$$\ell_n(\theta) = \log(L_n) = \sum_{i=1}^{n} \log(f(x_i, \theta))\,.$$

The *normalized log-likelihood function* of the model is defined as

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

Definition 1.2. We define a *maximum likelihood estimator* as any element $\hat{\theta} = \hat{\theta}_{MLE} = \hat{\theta}_{MLE}(X_1, \dots, X_n) \in \Theta$ for which

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

Remark. By definition of this estimator and of the functions above,

- It is equivalent to maximize L_n, ℓ_n or $\bar{\ell}_n$, so any of these functions can be used in the definition of $\hat{\theta}_{MLE}$.

- The estimator $\hat{\theta}_{MLE}$ is a function of $X_1, \dots, X_n$ only.

- The above definitions can be generalized whenever a joint p.m.f./p.d.f. for $X_1, \dots, X_n$ can be specified, even without the i.i.d. assumption.

Informal notes, based on past lecture notes by Richard Nickl. Please let me know of any errors.

Example 1.1. For $X_i \sim \text{Poi}(\theta),\ \theta \geq 0,\ \hat{\theta}_{MLE} = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}_n$. (see Lecture 1)

Example 1.2. For $X_i \sim \mathcal{N}(\mu, \sigma^2),\ \theta = (\mu, \sigma^2)^\top \in \mathbf{R} \times (0, \infty)$, (see Example sheet)

Example 1.3. In the Gaussian linear model $Y = \mathbb{X}\theta + \varepsilon$, known $\mathbb{X} \in \mathbf{R}^{n \times p}$, unknown $\theta \in \mathbf{R}^p$ and $\varepsilon \sim \mathcal{N}(0, I_n)$, the observations $Y_i = X_i^\top \theta + \varepsilon_i$ are not identically distributed, but a joint distribution $f(Y_1, \ldots, Y_n, \theta)$ can still be specified, and the MLE coincides with the least-squares estimator (see Example sheet).

In these examples and several other cases, the maximum likelihood estimator is found as the unique zero of the gradient of the log-likelihood.

Definition 1.3. For $\Theta \subseteq \mathbf{R}^p$ and ℓ_n differentiable in θ the *score function* S_n is defined as

$$S_n(\theta) = \nabla_\theta\, \ell_n(\theta) = \Big[\frac{\partial}{\partial \theta_1}\ell_n(\theta), \ldots, \frac{\partial}{\partial \theta_p}\ell_n(\theta)\Big]^\top .$$

Remark. As noted above, one of the main uses of this function is to look for the MLE $\hat{\theta}$ as a solution to $S_n(\hat{\theta}) = 0$, which is a common heuristic to maximize ℓ_n. We will often consider situations such that this is a necessary and sufficient condition.

It is very important to remember the point of view that ℓ_n and S_n are functions of the parameter θ, the randomness being in the values of the X_i. Therefore, derivatives (and gradients) are taken with respect to θ, not the x_i. To remember this, it can help to notice that after observation, the values x_i are fixed, but the value of θ is unknown. It makes more sense to let θ vary, which is what derivatives and gradients mean.

1.2. *Information geometry & likelihood function.* Given access to a sample from a distribution, the likelihood principle is to build a random function $\bar{\ell}_n$ out of the values of the sample, and to maximize it. In order to understand the behaviour of the MLE, it is natural to examine what would happen if we were to maximize the expectation of this random function.

Definition 1.4. We recall that for a variable X with distribution P_θ on $\mathcal{X} \subseteq \mathbf{R}^d$, and $g : \mathcal{X} \to \mathbf{R}$, we have

$$\mathbf{E}_\theta[g(X)] = \int_{\mathcal{X}} g(x) dP_\theta(x) = \int_{\mathcal{X}} g(x) f(x, \theta) dx\,,$$

when $\mathcal{X}$ is continuous and P_θ has p.d.f. $f(x, \theta)$ and

$$\mathbf{E}_\theta[g(X)] = \sum_{x \in \mathcal{X}} g(x) f(x, \theta)\,,$$

when $\mathcal{X}$ is discrete and P_θ has p.m.f. $f(x, \theta)$.

Theorem 1.1. *For a model $\{f(\cdot,\theta) : \theta \in \Theta\}$, and a variable $X \sim P$ such that $\mathbf{E}[|\log(f(X,\theta)|] < \infty$, if the model is well specified with $f(x,\theta_0)$ as p.d.f. of P, the function ℓ defined by*

$$\ell(\theta) = \mathbf{E}_{\theta_0}[\log(f(X,\theta)]$$

is maximized at θ_0.

Remark. This theorem suggests that if we had knowledge of the function ℓ, we could recover exactly θ_0, the "true value" of the unknown parameter. Since we do not have access to this function, we maximize instead a *sample approximation* of this function $\bar{\ell}_n$. Indeed, we recall the definition

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

which is the empirical average of i.i.d variables with mean $\ell(\theta)$.

Proof of Theorem 1.1. For all $\theta \in \Theta$, we have

$$\ell(\theta) - \ell(\theta_0) = \mathbf{E}_{\theta_0}[\log(f(X,\theta)] - \mathbf{E}_{\theta_0}[\log(f(X,\theta_0)] = \mathbf{E}_{\theta_0}\Big[\log\frac{f(X,\theta)}{f(X,\theta_0)}\Big]\,.$$

For a concave function φ, Jensen's inequality gives $\mathbf{E}[\varphi(Z)] \le \varphi(\mathbf{E}[Z])$. Therefore, by concavity of the logarithm,

$$\ell(\theta) - \ell(\theta_0) \le \log \mathbf{E}_{\theta_0}\Big[\frac{f(X,\theta)}{f(X,\theta_0)}\Big] = \log\int_{\mathcal{X}} \frac{f(x,\theta)}{f(x,\theta_0)}\, f(x,\theta_0)dx = \log(1) = 0\,.$$

Indeed, by definition of a p.d.f., we have that $\int_{\mathcal{X}} f(x,\theta)dx = 1$ for all $\theta \in \Theta$. □

Remark. If the variable X is on a discrete space, the same proof applies, with p.m.f. and sum rather than with p.d.f. and integral.

If the assumption *strict identifiability of the model parametrization* is satisfied, i.e. $f(\cdot,\theta) = f(\cdot,\theta') \iff \theta = \theta'$, we are not in the equality case of Jensen's, and the inequality is strict. In this case, maximizing $\bar{\ell}_n(\theta)$ will approximately maximize $\ell(\theta)$ and thus approximate the true value θ_0.

Information theory gives another interpretation: the quantity $\ell(\theta_0) - \ell(\theta)$ defined as

$$\mathsf{KL}(P_{\theta_0}, P_\theta) = \int_{\mathcal{X}} f(x,\theta_0)\log\frac{f(x,\theta_0)}{f(x,\theta)}dx\,,$$

is called the *Kullback-Leibler divergence*, or entropy between the distributions P_{θ_0} and P_θ. It can be thought of as a "distance" between distributions, and the reformulation $\ell(\theta) = \ell(\theta_0) - \mathsf{KL}(P_{\theta_0}, P_\theta)$ shows that maximizing the likelihood is akin to minimizing an approximate "distance" to θ_0.

Principles of Statistics **Part II - Michaelmas 2018**

Lecture 3: Fisher information

Lecturer: Quentin Berthet

We have seen in the past lecture that the MLE $\hat{\theta}$ can often be obtained as a solution to the equation $S_n(\hat{\theta}) = \nabla_\theta \bar{\ell}_n(\hat{\theta}) = 0$, i.e. that

$$\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \log f(x_i, \hat{\theta}) = 0\,,$$

by exchanging finite sum and derivatives. Furthermore, we have also seen that the expectation ℓ of the function ℓ_n is maximized at θ_0, which suggests that a similar equation must hold with expectation $\mathbf{E}_{\theta_0}$.

Theorem 1.2. *For a parametric model $\{f(\cdot,\theta) : \theta \in \Theta\}$ regular enough that integration and differentiation can be exchanged, we have for all $\theta \in int(\Theta)$*

$$\mathbf{E}_\theta[\nabla_\theta \log f(X,\theta)] = 0\,.$$

Proof. We compute explicitly the expectation

$$\begin{aligned}\mathbf{E}_\theta[\nabla_\theta \log f(X,\theta)] &= \int_{\mathcal{X}} \big(\nabla_\theta \log f(x,\theta)\big) f(x,\theta)dx \\ &= \int_{\mathcal{X}} \Big(\nabla_\theta f(x,\theta)\frac{1}{f(x,\theta)}\Big) f(x,\theta)dx \\ &= \int_{\mathcal{X}} \nabla_\theta f(x,\theta)dx \\ &= \nabla_\theta \int_{\mathcal{X}} f(x,\theta)dx = 0\,.\end{aligned}$$

The last two equalities are due, successively, to the assumed regularity of the model and to the fact that $\int_{\mathcal{X}} f(x,\theta)dx = 1$ for all $\theta \in \Theta$. □

As a consequence, we have in particular $\mathbf{E}_{\theta_0}[\nabla_\theta \log f(X,\theta_0)] = 0$.

Definition 1.5. For a parameter space $\Theta \subseteq \mathbf{R}^p$, we define for all $\theta \in \mathrm{int}(\Theta)$ the *Fisher information matrix* as

$$I(\theta) = \mathbf{E}_\theta[\nabla_\theta \log f(X,\theta)\, \nabla_\theta \log f(X,\theta)^\top]\,,$$

Informal notes, based on past lecture notes by Richard Nickl. Please let me know of any errors.

i.e. coefficient-wise

$$I_{ij}(\theta) = \mathbf{E}_\theta\Big[\frac{\partial}{\partial\theta_i}\log f(X,\theta)\,\frac{\partial}{\partial\theta_j}\log f(X,\theta)\Big]\,.$$

Remark. In dimension 1, we have

$$I(\theta) = \mathbf{E}_\theta\Big[\Big(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X,\theta)\Big)^2\Big] = \mathrm{Var}_\theta\Big[\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X,\theta)\Big]\,,$$

as it is a centered random variable. In particular, $I(\theta_0)$ describes the random variations of $S_n(\theta_0)$ around zero, its mean. This helps to quantify the precision of $\hat\theta$ around θ_0 as a solution of $S_n(\hat\theta) = 0$.

Theorem 1.3. *With the same regularity assumptions, for all $\theta \in int(\Theta)$, we have*

$$I(\theta) = -\mathbf{E}_\theta[\nabla^2_\theta \log f(X,\theta)]\,.$$

i.e. coefficient-wise

$$I_{ij}(\theta) = -\mathbf{E}_\theta\Big[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(X,\theta)\Big]\,.$$

Proof. We develop the term in the expectation

$$\begin{aligned}\nabla^2_\theta \log f(X,\theta) &= \nabla_\theta\Big(\frac{1}{f(X,\theta)}\nabla_\theta f(X,\theta)\Big)\\ &= \frac{1}{f(X,\theta)}\nabla^2_\theta f(X,\theta) - \frac{1}{f^2(X,\theta)}\nabla_\theta f(X,\theta)\nabla_\theta f(X,\theta)^\top\end{aligned}$$

Taking expectation yields

$$\begin{aligned}-\mathbf{E}_\theta[\nabla^2_\theta \log f(X,\theta)] &= -\int_{\mathcal{X}}\frac{1}{f(x,\theta)}\nabla^2_\theta f(x,\theta)\,f(x,\theta)\mathrm{d}x + \mathbf{E}_\theta\Big[\frac{1}{f^2(X,\theta)}\nabla_\theta f(X,\theta)\nabla_\theta f(X,\theta)^\top\Big]\\ &= 0 + E_\theta\Big[\Big(\frac{1}{f(X,\theta)}\nabla_\theta f(X,\theta)\Big)\Big(\frac{1}{f(X,\theta)}\nabla_\theta f(X,\theta)\Big)^\top\Big]\\ &= \mathbf{E}_\theta[\nabla_\theta \log f(X,\theta)\,\nabla_\theta \log f(X,\theta)^\top]\\ &= I(\theta)\,.\end{aligned}$$

The integral is cancelled by exchanging integral and derivatives on $\int_{\mathcal{X}} f(x,\theta)dx = 1$. □

Remark. In dimension 1, we extend the remark above to

$$I(\theta) = \mathrm{Var}_\theta\Big[\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X,\theta)\Big] = -\mathbf{E}_\theta\Big[\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f(X,\theta)\Big]$$

This shows a relationship between the variance of the score and the curvature of ℓ, clearly two important quantities when describing the quality of $\hat\theta$ (the maximum of ℓ_n) as an approximation of θ_0 (the maximum of ℓ).

Definition 1.6. For a random vector $X = (X_1, \dots, X_n) \in \mathbf{R}^n$, the Fisher information matrix is naturally defined as

$$I_n(\theta) = \mathbf{E}_\theta[\nabla_\theta \log f(X_1, \dots, X_n, \theta)\, \nabla_\theta \log f(X_1, \dots, X_n, \theta)^\top]\,,$$

Proposition 1.1. *When the vector $X = (X_1, \dots, X_n) \in \mathbf{R}^n$ is composed of n i.i.d. copies of a random variable and $\{f(\cdot, \theta)\,;\, \theta \in \Theta\}$ is a model for X, the Fisher information* tensorizes*, i.e.*

$$I_n(\theta) = n\, I(\theta)\,,$$

where $I(\theta)$ is the Fisher information for one copy X_i.

Proof. Since $f(X_1, \dots, X_n, \theta) = \prod_{i=1}^n f(X_i, \theta)$ by independence of the X_i, it holds that $\log f(X_1, \dots, X_n, \theta) = \sum_{i=1}^n \log f(X_i, \theta)$ and

$$I_n(\theta) = \mathbf{E}_\theta\Big[\sum_{i=1}^n \sum_{j=1}^n \nabla_\theta \log f(X_i, \theta)\, \nabla_\theta \log f(X_j, \theta)^\top\Big]$$

For all X_i, we recall that $\mathbf{E}_\theta[\nabla_\theta \log f(X_i, \theta)] = 0$, so by independence, only the n cross terms of this sum remain, giving the desired result. □

Principles of Statistics **Part II - Michaelmas 2018**

Lecture 4: Cramer-Rao bound and Convergence

Lecturer: Quentin Berthet

The following result formalizes the link between Fisher information and precision of estimation.

THEOREM 1.4 (Cramèr–Rao lower bound). *Let $\{f(\cdot,\theta) : \theta \in \Theta\}$ be a "regular" statistical model with $p = 1$ and $\Theta \subseteq \mathbf{R}$. Let $\tilde\theta = \tilde\theta(X_1, \dots, X_n)$ be an unbiased estimator based on the observation of n i.i.d. X_i from this model. We have, for all $\theta \in int(\Theta)$*

$$Var_\theta(\tilde\theta) = \mathbf{E}_\theta[(\tilde\theta - \theta)^2] \geq \frac{1}{nI(\theta)} \,.$$

PROOF. For $\mathrm{Var}_\theta(\tilde\theta) < \infty$, we first treat the case of $n = 1$. We recall that by the Cauchy–Schwarz inequality, it holds for all Y, Z that

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

Taking $Y = \tilde\theta$ and $Z = \frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X, \theta)$, we have

$$\mathrm{Var}_\theta(\tilde\theta) \geq \frac{\mathrm{Cov}_\theta(\tilde\theta, Z)^2}{\mathrm{Var}_\theta(Z)}\,.$$

We recall that $E[Z] = 0$ and that $\mathrm{Var}_\theta(Z) = I(\theta)$ by the results and definitions above. As a consequence, $\mathrm{Cov}_\theta(\tilde\theta, Z) = \mathbf{E}_\theta[\tilde\theta\, Z]$ and we have

$$\begin{aligned} \mathbf{E}_\theta\Big[\tilde\theta \frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X,\theta)\Big] &= \int_{\mathcal{X}} \tilde\theta(x) \frac{\mathrm{d}}{\mathrm{d}\theta} f(x,\theta) \frac{1}{f(x,\theta)} f(x,\theta)\mathrm{d}x \\ &= \frac{\mathrm{d}}{\mathrm{d}\theta} \int_{\mathcal{X}} \tilde\theta(x) f(x,\theta)\mathrm{d}x \\ &= \frac{\mathrm{d}}{\mathrm{d}\theta}\theta = 1\,. \end{aligned}$$

As a consequence, for $n = 1$, we have

$$\mathrm{Var}_\theta(\tilde\theta) \geq \frac{1}{I(\theta)}$$

For general n, we use $Z = \frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X_1, \dots, X_n, \theta)$, note that $\mathrm{Var}_\theta(Z) = I_n(\theta) = nI(\theta)$, and using the same simplification as above to show that $\mathbf{E}_\theta[\tilde\theta\, Z] = 1$ (Example sheet). □

The assumption that the model is *regular enough* is for now equivalent to saying that integrals and derivatives can be exchanged. We will be a bit more precise later, and describe some conditions that guarantee such regularity. These subtleties are studied in *Probability and Measure*, and are not examinable in this course.

The Cramèr–Rao lower bound is a statement about the variance of an estimator, so it is univariate in nature. It is possible to say something in the multivariate parameter case, for $\theta \in \mathbf{R}^p, p \geq 1$. Indeed, for any differentiable functional $\Phi : \Theta \to \mathbf{R}$, we consider $\tilde{\Phi}$ an unbiased estimator of $\Phi(\theta)$ based the on observation of $X_1, \ldots, X_n$, n i.i.d. copies from model $\{f(\cdot, \theta) : \theta \in \Theta\}$.

Proposition 1.2. *For all $\theta \in int(\Theta)$, with the definitions above, we have*

$$Var_\theta(\tilde{\Phi}) \geq \frac{1}{n} \nabla_\theta \Phi(\theta)^\top I^{-1}(\theta) \nabla_\theta \Phi(\theta) .$$

As an example, considering $\Phi(\theta) = \alpha^\top \theta = \sum_{i=1}^p \alpha_i \theta_i$, we have $\nabla_\theta \Phi(\theta) = \alpha$, so the lower bound implies $\mathrm{Var}_\theta(\tilde{\Phi}) \geq \frac{1}{n} \alpha^\top I^{-1}(\theta) \alpha$.

Example 1.4. Let $(X_1, X_2)^\top \sim \mathcal{N}(\theta, \Sigma)$, where $\theta = (\theta_1, \theta_2)^\top$ and where Σ is a known positive definite matrix, the sample size being $n = 1$.

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

- Case 2: Consider the estimation of θ_1 when θ_2 is *unknown*. We can consider the functional $\Phi(\theta) = \theta_1$ and the result above to establish the Fisher information in this new statistical model.

The comparison of these quantities is studied in one of the questions in the example sheet. Of particular interest is the case where Σ is diagonal, corresponding to independent X_1 and X_2.

2. Asymptotic Theory for MLE.

Not all estimators are unbiased for fixed sample size, but a desirable property that we could expect is that all reasonable estimators satisfy

$$\mathbf{E}_\theta[\tilde{\theta}_n] \to \theta \,, \quad \text{when } n \to \infty \text{ and sampling from } P_\theta.$$

A stronger, but closely related concept is that of *consistency* of $\tilde{\theta}$.

$$\tilde{\theta}_n \overset{?}{\to} \theta \,, \quad \text{when } n \to \infty \text{ and sampling from } P_\theta.$$

In a sense that should be specified, as we are evaluating the convergence of a random variable.

Informal notes, based on past lecture notes by Richard Nickl. Please let me know of any errors.

2.1. *Stochastic convergence concepts.*

Definition 2.1 (Convergence almost surely and in probability). Let $(X_n)_{n\geq 0}, X$, be random vectors in $\mathbf{R}^k$, defined on a probability space $(\Omega, \mathcal{A}, \mathbf{P})$.

i) We say that X_n converges to X *almost surely*, or $X_n \xrightarrow{a.s.} X$ as $n \to \infty$, if

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

ii) We say that X_n converges to X *in probability*, or $X_n \xrightarrow{P} X$ as $n \to \infty$, if for all $\varepsilon > 0$

$$\mathbf{P}(\|X_n - X\| > \varepsilon) \to 0\,.$$

Remark. Convergence of vectors is equivalent to convergence of each coefficient, for both these definitions. This follows naturally from the definition for almost sure convergence, and addressed in the Example Sheet for convergence in probability.

Definition 2.2 (Convergence in distribution). Let $(X_n)_{n\geq 0}, X$, be random vectors in $\mathbf{R}^k$, defined on a probability space $(\Omega, \mathcal{A}, \mathbf{P})$. We say that X_n converges to X *in distribution*, or $X_n \xrightarrow{d} X$ as $n \to \infty$ if

$$\mathbf{P}(X_n \preceq t) \to \mathbf{P}(X \preceq t)\,,$$

for all t where the map $t \mapsto \mathbf{P}(X \preceq t)$ is continuous.

Remark. We write $\{X \preceq t\}$ as shorthand for $\{X_{(1)} \leq t_1, \ldots, X_{(k)} \leq t_k\}$. For $k = 1$, the definition becomes $\mathbf{P}(X_n \leq t) \to \mathbf{P}(X \leq t)$.

The following facts about stochastic convergence follow from these definitions and can be proved in measure theory.

Proposition 2.1. $X_n \xrightarrow{a.s.} X \;\Rightarrow\; X_n \xrightarrow{P} X \;\Rightarrow\; X_n \xrightarrow{d} X$ *as* $n \to \infty$.

Proposition 2.2 (Continuous mapping theorem). *If* X_n, X *take values in* $\mathcal{X} \subset \mathbf{R}^d$ *and* $g : \mathcal{X} \to \mathbf{R}$ *is continuous, then* $X_n \xrightarrow{a.s./P/d} X$ *implies* $g(X_n) \xrightarrow{a.s./P/d} g(X)$.

Principles of Statistics **Part II - Michaelmas 2018**

Lecture 5: Central limit theorem and inference

Lecturer: Quentin Berthet

Proposition 2.3 (Slutsky's lemma). *Let $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$ where c is deterministic (or non-stochastic). Then, as $n \to \infty$*

- *a) $Y_n \xrightarrow{P} c$*
- *b) $X_n + Y_n \xrightarrow{d} X + c$*
- *c) $(k = 1)$ $X_n Y_n \xrightarrow{d} cX$ and if $c \neq 0$, $X_n / Y_n \xrightarrow{d} X/c$.*
- *d) If $(A_n)_{n\geq 0}$ are random matrices such that $(A_n)_{ij} \xrightarrow{P} A_{ij}$, where A is deterministic (or non-stochastic), then $A_n X_n \xrightarrow{d} AX$.*

Proposition 2.4. *If $X_n \xrightarrow{d} X$ as $n \to \infty$, then $(X_n)_{n\geq 0}$ is bounded in probability, or $X_n = O_P(1)$, i.e.*

$$\forall \varepsilon > 0 \, \exists M(\varepsilon) < \infty \quad \textit{such that for all } n \geq 0, \quad \mathbf{P}\big(\|X_n\| > M(\varepsilon)\big) < \varepsilon\,.$$

Many estimators in statistics are based on, or related to, the mean of i.i.d. random variables. A very important result regarding their convergence is the law of large numbers. in its most simple form, it can be proved with elementary tools.

Proposition 2.5 (Weak law of large numbers). *Let $X_1, \ldots, X_n$ be i.i.d. copies of X with Var$(X) < \infty$. It holds that*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{P} \mathbf{E}[X]$$

Proof. We apply Chebyshev's inequality to the centered random variable of interest $Z_n = \frac{1}{n}\sum_{i=1}^n (X_i - \mathbf{E}[X])$

$$\mathbf{P}\Big(|\bar{X}_n - \mathbf{E}[X]| > \varepsilon\Big) = \mathbf{P}\Big(|\frac{1}{n}\sum_{i=1}^n (X_i - \mathbf{E}[X])| > \varepsilon\Big) \leq \frac{\mathrm{Var}(Z_n)}{\varepsilon^2}\,.$$

By independence of the X_i, the variance of Z_n satisfies $\mathrm{Var}(Z_n) = \mathrm{Var}(X)/n$. As a consequence

$$\mathbf{P}\Big(|\bar{X}_n - \mathbf{E}[X]| > \varepsilon\Big) \leq \frac{\mathrm{Var}(X)}{\varepsilon^2}\,\frac{1}{n} \to 0\,.$$

□

The strong law of large numbers requires less assumptions on X, and states a stronger convergence (almost sure). It is admitted here.

Theorem 2.1 (Strong law of large numbers). *Let $X_1, \dots, X_n$ be i.i.d. copies of $X \sim P$ on $\mathbf{R}^k$ and assume $\mathbf{E}[\|X\|] < \infty$. We have as $n \to \infty$*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mathbf{E}[X] \,.$$

2.2. *Law of large numbers and Central limit theorem.*
In the previous lecture, we recalled that the average $\bar{X}_n$ of n i.i.d copies of X converges almost surely to the mean $\mathbf{E}[X]$.The stochastic fluctuations of $\bar{X}_n$ around $\mathbf{E}[X]$ are of order $1/\sqrt{n}$ and look normally distributed whenever $\mathrm{Var}(X) = \sigma^2 < \infty$.

Theorem 2.2 (Central limit theorem). *Let $X_1, \dots, X_n$ be i.i.d. copies of $X \sim P$ on $\mathbf{R}$ and assume $Var(X) = \sigma^2 < \infty$. We have as $n \to \infty$*

$$\sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbf{E}[X]\Big) \xrightarrow{d} \mathcal{N}(0, \sigma^2) \,.$$

To give a multidimensional version, we recall the following

Definition 2.3. A random variable $X \in \mathbf{R}^k$ has a normal distribution with mean $\mu \in \mathbf{R}^k$ and $k \times k$ covariance matrix Σ, denoted by $X \sim \mathcal{N}(\mu, \Sigma)$, if

- Its probability density function is

$$f(x) = \frac{1}{(2\pi)^{k/2}} \frac{1}{|\det(\Sigma)|^{1/2}} \exp\Big(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\Big) \,.$$

- It is the unique random variable such that for all linear forms $\alpha^\top X \sim \mathcal{N}(\alpha^\top \mu, \alpha^\top \Sigma \alpha)$, a valid definition for singular Σ.

The following basic facts are recalled

Proposition 2.6. - *For A a $d \times k$ matrix and $b \in \mathbf{R}^d$*

$$AX + b \sim \mathcal{N}(A\mu + b, A\Sigma A^\top) \,.$$

- *If $A_n \xrightarrow{P} A$ are random matrices and $X_n \xrightarrow{d} \mathcal{N}(\mu, \Sigma)$, then $A_n X_n \xrightarrow{d} \mathcal{N}(A\mu, A\Sigma A^\top)$.*

- *If Σ is diagonal, all the coefficients $X_{(j)}$ are independent.*

The multivariate version of the central limit theorem is

Informal notes, based on past lecture notes by Richard Nickl. Please let me know of any errors.

Theorem 2.3. *Let $X_1, \dots, X_n$ be i.i.d. copies from $X \sim P$ on $\mathbf{R}^k$ with $Cov(X) = \Sigma$ positive definite. We have as $n \to \infty$*

$$\sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^n X_i - \mathbf{E}[X]\Big) \xrightarrow{d} \mathcal{N}(0, \Sigma)\,.$$

As a consequence of Proposition 2.1, we can bound in probability the deviations

Corollary 2.1. *Under the conditions of Theorem 2.3, we have*

$$\frac{1}{n}\sum_{i=1}^n X_i - \mathbf{E}[X] = O_P(1/\sqrt{n})\,.$$

In light of these results, a reasonable optimality criterion for estimators is hence *asymptotic efficiency.*

$$n\mathrm{Var}_{\theta_0}(\tilde{\theta}) \to I^{-1}(\theta_0), \quad \text{when } n \to \infty \text{ and sampling from } P_{\theta_0}.$$

Indeed, the Cramèr-Rao lower bound hints at the fact that this is the smallest variance that can be asymptotically achieved. In this section, we show that the maximum likelihood estimator (MLE) is asymptotically efficient. In fact, under suitable assumptions,

$$\hat{\theta}_{MLE} \approx \mathcal{N}\big(\theta, I^{-1}(\theta_0)/n\big)\,,$$

for any $\theta_0 \in \Theta$. This implies efficiency, but is also useful to for inference; the construction of confidence regions.

Example 2.1 (Confidence interval). Let $X_1, \dots$ be a sequence of i.i.d copies of $X \sim P$, real-valued random variable with mean μ_0 and variance σ^2. For $\alpha \in (0, 1)$, we define the confidence region

$$\mathcal{C}_n = \Big\{\mu \in \mathbf{R} : |\mu - \bar{X}| \le \frac{\sigma z_\alpha}{\sqrt{n}}\Big\}$$

where α is taken such that $\mathbf{P}(|Z| \le z_\alpha) = 1 - \alpha$, for $Z \sim \mathcal{N}(0, 1)$. To show that $\mathcal{C}_n$ is a good confidence region, we compute the probability that μ_0 belongs to it

$$\begin{aligned}
\mathbf{P}(\mu_0 \in \mathcal{C}_n) &= \mathbf{P}(|\bar{X}_n - \mu_0| \le \frac{\sigma z_\alpha}{\sqrt{n}})\\
&= \mathbf{P}(|\frac{1}{n}\sum_{i=1}^n \frac{X_i - \mu}{\sigma}| \le \frac{z_\alpha}{\sqrt{n}})\\
&= \mathbf{P}(\sqrt{n}|\frac{1}{n}\sum_{i=1}^n \tilde{X}_i - \mathbf{E}[\tilde{X}]| \le z_\alpha)\\
&\to \mathbf{P}(|Z| \le z_\alpha) = \alpha\,.
\end{aligned}$$

by central limit theorem and continuous mapping theorem. As a consequence, $\mathcal{C}_n$ is an *asymptotic level* $1 - \alpha$ *confidence set.* When σ is unknown, it can be replaced by an estimate of it (see Example sheet).

Remark (Discussion). This example shows that to estimate the mean of a distribution, the empirical average $\bar{X}_n$ of the observations is a suitable estimator. Indeed, the law of large numbers shows that it converges to the true value $\mathbf{E}[X] = \mu_0$. It is also possible to describe the variations of $\bar{X}_n$ around its limit, with the central limit theorem, which is useful for inference. In the following lectures, we will show how the same tools can be used to prove similar properties for the maximum likelihood estimator, under some assumptions on the parametric model.

Principles of Statistics **Part II - Michaelmas 2018**

Lecture 6: Consistency of the MLE

Lecturer: Quentin Berthet

2.3. *Consistency of the MLE.*

DEFINITION 2.4 (Consistency). Consider $X_1, \dots, X_n$ i.i.d. arising from the parametric model $\{P_\theta : \theta \in \Theta\}$. An estimator $\tilde{\theta}_n = \tilde{\theta}_n(X_1, \dots, X_n)$ is called consistent if $\tilde{\theta}_n \to \theta_0$ in probability for $n \to \infty$, whenever the X_i are drawn from P_{θ_0}.

REMARK. We often write simply $\tilde{\theta}_n \xrightarrow{P_\theta} \theta$.

Under some regularity assumptions on the model, we show here that the maximum likelihood estimator $\hat{\theta}_{MLE}$ is unique and consistent. For the version of the theorem shown here, we use the following set of assumptions

ASSUMPTION 2.1 (Regularity for consistency). Let $\{f(\cdot, \theta) : \theta \in \Theta\}$ be a statistical model of p.d.f./p.m.f. on $\mathcal{X} \subseteq \mathbf{R}^d$ such that

1. $f(x, \theta) > 0$ for all $x \in \mathcal{X}, \theta \in \Theta$.
2. $\int_{\mathcal{X}} f(x, \theta) \mathrm{d}x = 1$ for all $\theta \in \Theta$.
3. The function $f(x, \cdot) : \theta \mapsto f(x, \theta)$ is continuous for all $x \in \mathcal{X}$.
4. $\Theta \subseteq \mathbf{R}^p$ is compact.
5. For any $\theta, \theta' \in \Theta$, $f(\cdot, \theta) = f(\cdot, \theta') \Rightarrow \theta = \theta'$.
6. $\mathbf{E}_\theta \sup_{\theta'} |\log f(X, \theta)| < \infty$.

REMARK. - Assumptions 1., 2., 5., and 6. guarantee that we can apply the strict version of Jensen's inequality (see reference), so that θ_0 is the unique maximum of the function ℓ defined by $\ell(\theta) = \mathbf{E}_{\theta_0} \log f(X, \theta)$.

- With these hypotheses (particularly 6.), we have that the continuity of the function $\theta \mapsto \log f(x, \theta)$ carries over to continuity of $\theta \mapsto \mathbf{E}_{\theta_0} \log f(X, \theta)$. This is known as the *dominated convergence theorem*, studied in details in the course *Probability and Measure*.

Informal notes, based on past lecture notes by Richard Nickl. Please let me know of any errors.

- These subtleties are included in the spirit of rigor and exhaustiveness, so that the interested reader can connect this result to other courses. However, they are not the main focus of this course, and these assumptions are not examinable. They are often referred to in the lecture notes, as well as in examination questions as *usual regularity assumptions*.

THEOREM 2.4 (Consistency of the MLE). *Let $X_1, \dots, X_n$ be i.i.d from the model $\{f(\cdot, \theta) : \theta \in \Theta\}$ satisfying Assumption 2.1. Then, an MLE exists and any MLE is consistent.*

PROOF OF EXISTENCE. The mapping $\theta \mapsto \bar{\ell}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \log f(X_i, \theta)$ is continuous on the compact set Θ. As a consequence, a maximizer exists, so an MLE is well-defined □

The idea behind the proof of consistency is that for all θ, $\bar{\ell}_n(\theta)$ converges to $\ell(\theta)$ in probability, by the law of large numbers since

$$\frac{1}{n}\sum_{i=1}^n \log f(X_i, \theta) \to \mathbf{E}_{\theta_0}[\log f(X, \theta)]\,.$$

As a consequence, we expect $\hat{\theta}_n$ to converge to θ_0, the maximizer of $\ell(\theta)$. While this intuition is false in general, it holds under a stronger fact: the uniform convergence of $\bar{\ell}_n$ to ℓ. Under Assumption 2.1, we have as $n \to \infty$

$$\sup_{\theta \in \Theta} |\bar{\ell}_n(\theta) - \ell(\theta)| \xrightarrow{P} 0\,,$$

a *uniform law of large numbers*. We assume for now this result, used in the following proof.

PROOF OF CONSISTENCY. Define $\Theta_\varepsilon = \{\theta \in \Theta : \|\theta - \theta_0\| \geq \varepsilon\}$ for arbitrary $\varepsilon > 0$. Note that Θ_ε is compact as intersection of Θ with a closed set. The function ℓ is continuous on Θ_ε, therefore it attains its maximum on it, so there exists θ_ε such that

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

as θ_0 is the unique maximum of ℓ on Θ. There therefore exists $\delta(\varepsilon) > 0$ such that $c(\varepsilon) + \delta(\varepsilon) < \ell(\theta_0) - \delta(\varepsilon)$. By the triangle inequality, we have

$$\begin{aligned}\sup_{\theta \in \Theta_\varepsilon} \bar{\ell}_n(\theta) &= \sup_{\theta \in \Theta_\varepsilon} \left[\bar{\ell}_n(\theta) - \ell(\theta) + \ell(\theta)\right] \\ &\leq \sup_{\theta \in \Theta_\varepsilon} \ell(\theta) + \sup_{\theta \in \Theta} |\bar{\ell}_n(\theta) - \ell(\theta)|\,.\end{aligned}$$

We consider the sequence of events

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

On these events, the following sequence of inequalities hold

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

the last inequality holding by definiton of $\delta(\varepsilon)$. On $A_n(\varepsilon)$, we also have that $\ell(\theta_0) - \bar{\ell}_n(\theta_0) \leq \delta(\varepsilon)$, leading to

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

As a consequence, on $A_n(\varepsilon)$, $\hat{\theta}_n$ cannot lie in Θ_ε, as this would lead to the contradiction $\bar{\ell}_n(\hat{\theta}_n) < \bar{\ell}_n(\theta_0)$. As a consequence, we have that $A_n(\varepsilon) \subseteq \{\|\hat{\theta}_n - \theta_0\| < \varepsilon\}$. Since $\mathbf{P}(A_n(\varepsilon)) \to 1$ as $n \to \infty$ by the uniform LLN, we have that

$$\mathbf{P}(\|\hat{\theta}_n - \theta_0\| < \varepsilon) \to 1\,.$$

□

Remark. This proof can be simplified under additional properties of the likelihood function, such as differentiability (see Example sheet). This can be useful in situations where Θ is not compact, as this assumption is then no longer required.

Principles of Statistics **Part II - Michaelmas 2018**

Lecture 7: Asymptotic normality of the MLE

Lecturer: Quentin Berthet

Digression: the uniform law of large numbers. As hinted in the previous lecture, one of the important notions in statistics is the uniform convergence of a class of variables to some limits: It is not sufficient that for all $\theta \in \Theta$, the sequence $\bar{\ell}_n(\theta)$ converges to $\ell(\theta)$ almost surely, we use the stronger fact that

$$\sup_{\theta \in \Theta} |\bar{\ell}_n(\theta) - \underbrace{\mathbf{E}_{\theta_0}[\ell_n(\theta)]}_{\ell(\theta)}| \xrightarrow{a.s.} 0 .$$

Written in this form, it can be understood as a uniform law of large numbers.

Observation (Finite case). Let $X_1, \dots, X_n$ be *i.i.d.* in $\mathcal{X} \subseteq \mathbf{R}^d$ and $h : \mathcal{X} \to \mathbf{R}$ a function. The variables $h(X_i)$ are also i.i.d., so if $\mathbf{E}|h(X)| < \infty$ then by the strong law of large numbers

$$\frac{1}{n}\sum_{i=1}^n h(X_i) - \mathbf{E}[h(X)] \xrightarrow{a.s.} 0 .$$

If $h_1, \dots, h_M$ is a finite class of such functions, this applies for each $1 \le j \le M$ and we have on events A_j such that $\mathbf{P}(A_j^c) = 0$,

$$\frac{1}{n}\sum_{i=1}^n h_j(X_i) - \mathbf{E}[h_j(X)] \to 0 .$$

Note that this is a simple convergence in the reals: implicitly it means that for all $\omega \in A_j$, the display above holds for $X_i(\omega)$. Hence, on $A = \cap_{j=1}^M A_j$, we have that

$$(\dagger) \qquad \max_{1 \le j \le M} \Big| \frac{1}{n}\sum_{i=1}^n h_j(X_i) - \mathbf{E}[h_j(X)] \Big| \to 0 .$$

Furthermore, we have that

$$\mathbf{P}(A^c) = \mathbf{P}(\cup_{j=1}^M A_j^c) \le \sum_{j=1}^M \mathbf{P}(A_j) = 0 ,$$

so that (†) holds almost surely. Therefore, the uniform law of large number holds in the finite case. In order to extend this result to an infinite class of functions, as needed here for all the variables $\log f(X_i, \theta)$ indexed over $\theta \in \Theta$ (instead of $1 \le j \le M$), we use the continuous "analogue" to finiteness: compactness.

Informal notes, based on past lecture notes by Richard Nickl. Please let me know of any errors.

Theorem 2.5 (Uniform law of large numbers). *Let Θ be a compact set in $\mathbf{R}^p$ and $q : \mathcal{X} \times \Theta \to \mathbf{R}$ be continuous in θ for all x such that $\mathbf{E}[\sup_{\theta\in\Theta} |q(X,\theta)|] < \infty$. Then, as $n \to \infty$, we have*

$$\sup_{\theta\in\Theta} \Big|\frac{1}{n}\sum_{i=1}^{n} q(X_i,\theta) - \mathbf{E}[q(X,\theta)]\Big| \xrightarrow{a.s} 0\,.$$

The idea of the proof is to transfer the result over a finite set Θ' - as shown in the observation - to an infinite, but compact, set Θ. By compactness, the set Θ can be covered by a finite subset Θ', up to any fixed precision $\delta > 0$. By continuity in θ, the uniform convergence over Θ' implies uniform convergence over Θ: any $\theta \in \Theta$ is δ-close to some $\theta \in \Theta'$, so $q(X,\theta)$ is close to $q(X,\theta')$.

2.4. *Asymptotic normality of the MLE.* The maximum likelihood estimator $\hat\theta_{MLE}$ therefore converges in probability to the true value of the parameter θ_0, i.e. the MLE is consistent. This is based on the property that $\hat\theta_{MLE}$ and θ_0 maximize, respectively, $\bar\ell_n$ and ℓ. It is possible to quantify this maximization, by studying the behavior of the gradient and the Hessian of these functions about their maximizers, in order to obtain a finer understanding of the convergence of $\hat\theta_{MLE} - \theta_0$.

Assumption 2.2. Let $\{f(\cdot,\theta) : \theta \in \Theta\}$ be a statistical model of p.d.f./p.m.f. on $\mathcal{X} \subseteq \mathbf{R}^d$ such that in addition to those of Assumption 2.1 (which ensure consistency) we have

1. The true θ_0 belongs to the interior of Θ.
2. There exists $U \subseteq \Theta$ open containing θ_0 such that the function
$$\theta \mapsto f(x,\theta)$$
is for every $x \in \mathcal{X}$, twice continuously differentiable with respect to $\theta \in U$.
3. The $p \times p$ Fisher information matrix $I(\theta_0)$ is non-singular and we have that $\mathbf{E}_{\theta_0}[\|\nabla_\theta \log f(X,\theta_0)\|] < \infty$.
4. There exists a compact ball K in U of non-empty interior that is centered at θ_0 such that
$$\mathbf{E}_{\theta_0} \sup_{\theta\in K} \|\nabla^2_\theta \log f(X,\theta)\| < \infty$$
$$\int_{\mathcal{X}} \sup_{\theta\in K} \|\nabla_\theta \log f(X,\theta)\| \mathrm{d}x < \infty$$
$$\int_{\mathcal{X}} \sup_{\theta\in K} \|\nabla^2_\theta \log f(X,\theta)\| \mathrm{d}x < \infty\,.$$

Remark. We note once again that these are included to give a rigorous statement, so that the interested reader can connect this result to other courses. However, they are not the main focus of this course, and these assumptions are not examinable. They are often referred to in the lecture notes, as well as in examination questions as *usual regularity assumptions.*

Theorem 2.6. *Let the statistical model $\{f(\cdot,\theta) : \theta \in \Theta\}$ satisfy the properties of Assumption 2.2, and $\hat{\theta}_n$ be the MLE based on n i.i.d observations $X_1,\dots,X_n$ with distribution P_{θ_0}. We have, as $n \to \infty$*

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

Proof. On events of probability going to 1, $\hat{\theta}_n$ belongs to the interior of Θ, since $\hat{\theta}_n$ converges in probability to θ_0, contained in the interior. On these events, we have by the regularity assumptions that $\nabla_\theta \bar{\ell}_n(\hat{\theta}_n) = 0$, by first-order condition on the maximum. By regularity of $\theta \mapsto \nabla \bar{\ell}_n(\theta)$ and applying the mean value theorem on each coordinate between θ_0 and $\hat{\theta}_n$, we have

$$0 = \nabla_\theta \bar{\ell}_n(\hat{\theta}_n) = \nabla_\theta \bar{\ell}_n(\theta_0) + \underbrace{\bar{A}_n}_{\approx \nabla^2_\theta \bar{\ell}_n(\theta_0)} (\hat{\theta}_n - \theta_0)\,,$$

where $\bar{A}_n$ is defined, coefficient-wise by

$$\big(\bar{A}_n\big)_{ij} = \frac{\partial^2}{\partial\theta_i \partial\theta_j} \bar{\ell}_n(\bar{\theta}^{(j)})\,, \qquad \text{for } \theta^{(j)} \in [\theta_0, \hat{\theta}_n].$$

Assuming that $\bar{A}_n$ converges in probability to $\mathbf{E}_{\theta_0}[\nabla^2_\theta \bar{\ell}_n(\theta_0)] = -I(\theta_0)$, this yields that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \underbrace{(-\bar{A}_n^{-1})}_{\xrightarrow{P_{\theta_0}} I(\theta_0)^{-1}} \sqrt{n}\, \nabla_\theta \bar{\ell}_n(\theta_0)\,.$$

Furthermore, by definition of $\bar{\ell}_n$ and X, we have

$$\sqrt{n}\, \nabla_\theta \bar{\ell}_n(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \Big(\nabla_\theta \log f(X_i, \theta) - \underbrace{\mathbf{E}_{\theta_0}[\nabla_\theta \log f(X, \theta)]}_{=0} \Big)$$

As a consequence, we have by the central limit theorem that

$$\sqrt{n}\, \nabla_\theta \bar{\ell}_n(\theta_0) \xrightarrow{d} \mathcal{N}\big(0, \underbrace{\mathrm{Cov}_{\theta_0}(\nabla_\theta \log f(X, \theta))}_{I(\theta_0)}\big)$$

Therefore, by convergence of $\bar{A}_n^{-1}$ to $-I(\theta_0)^{-1}$, this yields by Slutsky's lemma that

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

Proof of convergence of $\bar{A}_n$ (not examinable)

We have, for every entry of $\bar{A}_n$

$$\begin{aligned}(\bar{A}_n)_{kj} = \frac{1}{n}\sum_{i=1}^{n}\Big(&\frac{\partial^2}{\partial\theta_k\partial\theta_j}\log f(X_i,\theta^{(j)}) - \mathbf{E}_{\theta_0}\Big[\frac{\partial^2}{\partial\theta_k\partial\theta_j}\log f(X_i,\theta^{(j)})\Big]\Big)\\ &+ \mathbf{E}_{\theta_0}\Big[\frac{\partial^2}{\partial\theta_k\partial\theta_j}\log f(X_i,\theta^{(j)})\Big] - \mathbf{E}_{\theta_0}\Big[\frac{\partial^2}{\partial\theta_k\partial\theta_j}\log f(X_i,\theta_0)\Big]\\ &+ (-I(\theta_0))_{kj}\,.\end{aligned}$$

writing $q(x,\theta) = \frac{\partial^2}{\partial\theta_k\partial\theta_j}\log f(x,\theta)$, we note that the regularity assumptions imply continuity of $q(x,\theta)$ and $\mathbf{E}_{\theta_0}[q(X,\theta)]$ for all $x \in \mathcal{X}$. We can therefore conclude that the first line of the above display can be upper bounded by

$$\sup_{\theta\in K}\Big|\frac{1}{n}\sum_{i=1}^{n} q(X_i,\theta) - \mathbf{E}_{\theta_0}[q(X,\theta)]\Big| \xrightarrow{P_{\theta_0}} 0\,,$$

by the uniform law of large numbers, whose assumptions are verified. The second line of the display can also be upper bounded by

$$|\mathbf{E}_{\theta_0}[q(X,\theta^{(j)})] - \mathbf{E}_{\theta_0}[q(X,\theta_0)| \xrightarrow{P_{\theta_0}} 0\,,$$

since $\theta^{(j)} \xrightarrow{P_{\theta_0}} \theta_0$ by consistency of $\hat{\theta}_n$ and the continuous mapping theorem. This yields the desired result. □

Principles of Statistics **Part II - Michaelmas 2018**

Lecture 8: Discussion of asymptotic normality

Lecturer: Quentin Berthet

Definition 2.5 (Asymptotic efficency). In a parametric model $\{f(\cdot, \theta) : \theta \in \Theta\}$, a consistent estimator $\tilde{\theta}_n$ is called *asymptotically efficient* if $n \operatorname{Var}_{\theta_0}(\tilde{\theta}_n) \to I(\theta_0)^{-1}$ for all $\theta \in int(\Theta)$ for $p = 1$, or $n \operatorname{Cov}_{\theta_0}(\tilde{\theta}_n) \to I(\theta_0)^{-1}$ in $\mathbf{R}^{p \times p}$.

Remark. In reference to the result and assumptions of Theorem 2.6, we make the following remarks

- At the expense of more sophisticated proofs, one can reduce the regularity assumptions required of the function $\theta \mapsto f(x, \theta)$. This allows in particular to cover the case of Laplace distributions.

- Some notion of regularity is however required, as shown in the example of the uniform distribution over $[0, \theta]$, whose density $f(x, \theta) = \frac{1}{\theta}\mathbf{1}_{[0,\theta]}(x)$ produces a discontinuous likelihood function (see Example sheet).

- For θ_0 at the boundary of the parameter space Θ, the asymptotics need not be normal. For instance, in the model of $\mathcal{N}(\theta, 1)$ for $\theta \in [0, \infty)$, the case $\theta_0 = 0$ is a counter-example (see Example sheet).

- This result confirms the intuition given by the Cramèr-Rao lower bound that the "optimal" variance for an estimator is given by $I^{-1}(\theta_0)$. The lower bound holds only for unbiased estimators, and giving a rigorous theory of asymptotic efficiency requires more precision, as shown by the *Hodge estimator*, obtained from an estimator $\hat{\theta}_n$ over $\mathbf{R}$ with asymptotic normality, and by setting $\tilde{\theta}_n = \hat{\theta}_n$ if $|\hat{\theta}_n| > n^{-1/4}$, and 0 otherwise (see Example sheet).

- Later in the course, we will consider several criteria to compare estimators, including the worst performance for all values of the "true value" θ_0, which will shed some light on these apparent paradoxes.

Informal notes, based on past lecture notes by Richard Nickl. Please let me know of any errors.

2.5. *Plug-in MLE and Delta method.* It is often practical to think of estimation problems in the following way: for a parametric model $\{f(\cdot,\theta) : \theta \in \Theta\}$, we consider the estimation of $\Phi(\theta)$, for a functional $\Phi : \Theta \to \mathbf{R}^k$ and $\Theta \subseteq \mathbf{R}^p$. We first consider a special case, and introduce the following definition

Definition 2.6. For $\Theta = \Theta_1 \times \Theta_2$, and $\theta = (\theta_1, \theta_2)^\top$, we define the *profile likelihood*, for $\phi(\theta) = \theta_1$, by

$$L^{(p)}(\theta_1) = \sup_{\theta_2 \in \Theta_2} L((\theta_1, \theta_2)^\top)\,.$$

Remark. Note that maximizing the profile likelihood in θ_1 is equivalent to maximizing the likelihood in θ, and to take the first argument of the maximizer.

More generally, one shows that a MLE in the new parametrization in ϕ, given by $\{f(\cdot,\phi) : \phi = \Phi(\theta) \text{ for some } \theta \in \Theta\}$ is obtained by taking $\Phi(\hat\theta_{MLE})$. (see Examples sheet)

Definition 2.7. For a statistical model $\{f(\cdot,\theta) : \theta \in \Theta\}$ and $\Phi : \Theta \to \mathbf{R}^k$, the *plug-in MLE* of $\Phi(\theta_0)$ is the estimator $\Phi(\hat\theta_{MLE})$.

Using the known limiting distribution of an estimator to find the limiting distribution of another estimator, function of the initial one, is known as the Delta method.

Theorem 2.7 (Delta method). *Let $\Phi : \Theta \to \mathbf{R}$ be continuously differentiable at θ_0 with gradient satisfying $\nabla_\theta \Phi(\theta_0) \neq 0$. Let $\hat\theta_n$ be a sequence of random variables (estimator) such that $\sqrt{n}(\hat\theta_n - \theta_0) \xrightarrow{d.} Z$, where Z is a random variable in $\mathbf{R}^p$. We have that*

$$\sqrt{n}\big(\Phi(\hat\theta_n) - \Phi(\theta_0)\big) \xrightarrow{d.} \nabla_\theta \Phi(\theta_0)^\top Z\,.$$

Proof. We have, by definition of differentiability, for some $\tilde\theta_n$ in the segment $[\theta_0, \hat\theta_n]$

$$\sqrt{n}\big(\Phi(\hat\theta_n) - \Phi(\theta_0)\big) = \nabla_\theta \Phi(\tilde\theta_n)^\top \sqrt{n}\,(\hat\theta_n - \theta_0)$$

Since $\sqrt{n}(\hat\theta_n - \theta_0) \xrightarrow{d} Z$, it is bounded in probability and $\|\hat\theta_n - \theta_0\| = O_P(1/\sqrt{n})$ and $\theta_n \xrightarrow{P} \theta_0$, so $\tilde\theta_n \xrightarrow{P} \theta_0$ and by the continuous mapping theorem, $\nabla_\theta \Phi(\tilde\theta_n) \xrightarrow{P} \nabla_\theta \Phi(\theta_0)$. The desired result is therefore a direct consequence of Slutsky's lemma. □

Remark. This result is given in this form as it will be usually applied to estimators with deviations of order $1/\sqrt{n}$.

- It can be generalized to other estimators, taking a sequence $r_n \to \infty$ instead of $\sqrt{n}$.

- Note that in the case where $\hat\theta_n$ is a maximum likelihood with asymptotic normality (as in Theorem 2.6), this implies

$$\sqrt{n}\big(\Phi(\hat\theta_n) - \Phi(\theta_0)\big) \xrightarrow{d.} \mathcal{N}(0, \nabla_\theta \Phi(\theta_0)^\top I^{-1}(\theta_0) \nabla_\theta \Phi(\theta_0))\,.$$

In dimension 1, this yields

$$\sqrt{n}\big(\Phi(\hat{\theta}_n) - \Phi(\theta_0)\big) \xrightarrow{d.} \mathcal{N}(0, \Phi'(\theta_0)^2 I^{-1}(\theta_0))\,.$$

- This shows that the plug-in MLE is asymptotically efficient, as its limiting covariance matrix matches that of the Cramèr-Rao lower bound

$$\lim_{n\to\infty} n\mathrm{Var}_{\theta_0}(\Phi(\hat{\theta}_n)) = \nabla_\theta \Phi(\theta_0)^\top I^{-1}(\theta_0)\nabla_\theta\Phi(\theta_0)\,.$$

Principles of Statistics **Part II - Michaelmas 2018**

Lecture 9: Asymptotic inference with the MLE

Lecturer: Quentin Berthet

2.6. *Asymptotic inference with the MLE.* In the discussion on confidence intervals in Example 2.1, we are interested in estimating the mean of a random variable, and to construct confidence intervals, based on the central limit theorem. It is possible to do the same thing in general, for any coefficient θ_j of a parameter θ. Indeed, considering e_j the j-th vector of the canonical basis, we have for the MLE $\hat{\theta}_n$, under regularity assumptions

$$\sqrt{n}(\hat{\theta}_{n,j} - \theta_{0,j}) = e_j^\top \sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d.} \mathcal{N}(0, \underbrace{e_j^\top I(\theta_0)^{-1} e_j}_{(I^{-1}(\theta_0))_{jj}}) .$$

Using the same logic as in Example 2.1, we can therefore consider

$$\mathcal{C}_n = \left\{ \nu \in \mathbf{R} \, : \, |\nu - \hat{\theta}_{nj}| \leq (I^{-1}(\theta_0))_{jj}^{1/2} z_\alpha / \sqrt{n} \right\} ,$$

for z_α such that $P(|Z| \leq z_\alpha) = 1 - \alpha$, and $Z \sim \mathcal{N}(0,1)$. We compute the limit of the probability that $\theta_{0,j}$ is in the confidence interval $\mathcal{C}_n$

$$\mathbf{P}_{\theta_0}(\theta_{0,j} \in \mathcal{C}_n) = \mathbf{P}_{\theta_0}(\sqrt{n}(I^{-1}(\theta_0))_{jj}^{-1/2} |\hat{\theta}_{n,j} - \theta_{0,j}| \leq z_\alpha) \to 1 - \alpha .$$

by limiting distribution of $\sqrt{n}(\hat{\theta}_{n,j} - \theta_{0,j})$ and continuous mapping theorem.

Remark. We note that in order to construct the confidence interval given here, one must know the Fisher information $I(\theta_0)$, or at least its j-th diagonal coefficient. In general, this quantity depends on θ_0 and it is unreasonable to assume that it is known. In the example related to the mean, we saw that it was possible to replace it by an estimate. For the general case, it is therefore useful to introduce the following notion

Definition 2.8. We define the *observed Fisher information* as the $p \times p$ matrix

$$i_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta \log f(X_i, \theta) \, \nabla_\theta \log f(X_i, \theta)^\top .$$

It is common to use $\hat{i}_n = i_n(\hat{\theta}_{MLE})$ as an estimator of $I(\theta_0)$, as in the following proposition

Informal notes, based on past lecture notes by Richard Nickl. Please let me know of any errors.

Proposition 2.7. *Under the assumptions of Theorem 2.6, we have as $n \to \infty$, that $\hat{i}_n \xrightarrow{P_{\theta_0}} I(\theta_0)$*

Proof. We have, for all $\theta \in \Theta$, noting $q(X, \theta) = \nabla_\theta \log f(X, \theta)^\top \nabla_\theta \log f(X, \theta)$, that

$$i_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} q(X_i, \theta) \ , \quad \text{and} \quad I(\theta) = \mathbf{E}_{\theta_0}[q(X, \theta)] \, .$$

We therefore have

$$\hat{i}_n - I(\theta_0) = \left[i_n(\hat{\theta}_{MLE}) - I(\hat{\theta}_{MLE}) \right] + \left[I(\hat{\theta}_{MLE}) - I(\theta_0) \right] .$$

The first term is upper bounded as

$$\left| i_n(\hat{\theta}_{MLE}) - I(\hat{\theta}_{MLE}) \right| \leq \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} q(X_i, \theta) - \mathbf{E}_{\theta_0}[q(X, \theta)] \right| \xrightarrow{P_{\theta_0}} 0 \, ,$$

The second term converges in probability to 0, by consistency of $\hat{\theta}_{MLE}$ and by the continuous mapping theorem. □

Remark. It is also possible to use $\hat{j}_n = j_n(\hat{\theta}_{MLE})$, where

$$j_n(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \nabla_\theta^2 \log f(X_i \theta) \, .$$

This is also a consistant estimator of $I(\theta_0)$, with a similar proof.

Definition 2.9 (Wald statistic). For all $\theta \in \Theta$, we define the Wald statistic as

$$W_n(\theta) = n(\hat{\theta}_{MLE} - \theta)^\top \hat{i}_n (\hat{\theta}_{MLE} - \theta) \, .$$

This is a quadratic form, with a semidefinite positive $\hat{i}_n$, its level sets are ellipsoids that can be used to construct confidence regions.

Proposition 2.8 (Confidence ellipsoids). *Under the same assumptions, the confidence region $\mathcal{C}_n$ defined by*

$$\mathcal{C}_n = \left\{ \theta \, : \, W_n(\theta) \leq \xi_\alpha \right\} ,$$

for ξ_α satisfying $\mathbf{P}(\chi_p^2 \leq \xi_\alpha) = 1 - \alpha$, is an α-level asymptotic confidence region.

Proof. We compute $\mathbf{P}(\theta_0 \in \mathcal{C}_n) = \mathbf{P}(W_n(\theta_0) \leq \xi_\alpha)$, using that under these assumptions, $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d.} \mathcal{N}(0, I(\theta_0)^{-1})$ and $\hat{i}_n \xrightarrow{P_{\theta_0}} I(\theta_0)$. We decompose the statistic $W_n(\theta_0)$ as

$$W_n(\theta) = \sqrt{n}(\hat{\theta}_n - \theta_0)^\top I(\theta_0) \sqrt{n}(\hat{\theta}_n - \theta_0) + \sqrt{n}(\hat{\theta}_n - \theta_0)^\top (\hat{i}_n - I(\theta_0)) \sqrt{n}(\hat{\theta}_n - \theta_0)$$

By continuous mapping theorem, the first term converges in distribution to $U^\top U = U_1^2 + \ldots + U_p^2$ with $U \sim \mathcal{N}(0, I_p)$. As a sum of squares of p standard centered normal variables, it has distribution χ_p^2. The second term is a product of $\sqrt{n}(\hat{\theta}_n - \theta_0)^\top (\hat{\imath}_n - I(\theta_0))$, converging to 0 in distribution (hence also in probability) by Slutsky's lemma, and of $\sqrt{n}(\hat{\theta}_n - \theta_0)$. As such, the second term also converges to 0 in distribution by a second application of Slutsky's lemma, while a third application (on sums) yields that $W_n(\theta_0) \xrightarrow{d.} \chi_p^2$, giving the desired result. □

This statistic can therefore also be used to design a test for the hypothesis testing problem $H_0 : \theta = \theta_0$ against $H_1 : \theta \in \Theta \setminus \{\theta_0\}$, since $\mathbf{P}(W_n(\theta_0) > \xi_\alpha) \to \alpha$.

Principles of Statistics **Part II - Michaelmas 2018**

Lecture 10: Introduction to Bayesian statistics

Lecturer: Quentin Berthet

This method can be generalized to the hypothesis testing problem $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta \setminus \Theta_0$, for $\Theta_0 \subseteq \Theta \subseteq \mathbf{R}^p$. For these types of problem, the objective of a decision rule ψ_n - a function of the sample taking values in $\{0, 1\}$ - is to output 0 with high probability under H_0 and 1 under H_1. In order to measure its performance, we consider the

- type-one error (false positive): $\mathbf{P}_\theta(\underbrace{\psi_n = 1}_{\text{reject } H_0}) = \mathbf{E}_\theta[\psi_n] \,, \theta \in \Theta_0$.
- type-two error (false negative): $\mathbf{P}_\theta(\underbrace{\psi_n = 0}_{\text{accept } H_0}) = \mathbf{E}_\theta[1 - \psi_n] \,, \theta \in \Theta_1$.

Definition 2.10 (Likelihood ratio test). We define the *likelihood-ratio statistic* as

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

where $\hat{\theta}_{MLE,0}$ is the maximum likelihood estimator restricted to the set Θ_0.

Theorem 2.8 (Wilks theorem). *Let $\{f(\cdot, \theta) : \theta \in \Theta\}$ be a statistical model satisfying Assumptions 2.2, and a hypothesis testing problem where $\Theta_0 = \{\theta_0\}$, for some fixed $\theta_0 \in int(\Theta)$. We have, as $n \to \infty$*

$$\Lambda_n(\Theta, \Theta_0) \xrightarrow{d.} \chi^2_p \,.$$

Proof. Considering events where $\hat{\theta}_n \in \text{int}(\Theta)$, we have by definition of the likelihood ratio

$$\begin{aligned} \Lambda_n(\Theta, \Theta_0) &= 2\ell_n(\hat{\theta}_n) - 2\ell_n(\theta_0) \\ &= -(-2\ell_n(\theta_0)) - (-2\ell_n(\hat{\theta}_n)) \\ &= -2 \underbrace{\nabla_\theta \ell_n(\hat{\theta}_n)^\top}_{=0} (\theta_0 - \hat{\theta}_n) + \sqrt{n}(\theta_0 - \hat{\theta}_n)^\top \bar{B}_n \sqrt{n}(\theta_0 - \hat{\theta}_n) \,, \end{aligned}$$

where $\bar{B}_n$ is defined coefficient-wise, by Taylor approximation with remainder as

$$\big(\bar{B}_n\big)_{ij} = \frac{\partial^2}{\partial \theta_i \partial \theta_j} \bar{\ell}_n(\bar{\theta}) \,, \qquad \text{for } \bar{\theta} \in [\theta_0, \hat{\theta}_n].$$

Informal notes, based on past lecture notes by Richard Nickl. Please let me know of any errors.

As in the proof of Theorem 2.6 and similarly to Proposition 2.7, we have that B_n converges to $I(\theta_0)$ so the last term converges in distribution to a χ^2_p. □

Remark. This result gives another statistic with asymptotic distribution χ^2_p.

- The test $\psi_n = \mathbf{1}\{\Lambda_n(\Theta, \Theta_0) \geq \xi_\alpha\}$ controls the type-one error at asymptotic level $1-\alpha$.

- When Θ_0 has dimension $p_0 < p$, we have under the same assumptions that the statistic converges in distribution to a $\chi^2_{p-p_0}$

3. Bayesian inference. For a given parametric model $\{f(\cdot, \theta) : \theta \in \Theta\}$, there are many situations where it is convenient to consider θ as a random variable with distribution π on Θ. This can be motivated by some intrinsic randomness in the data generating process, can represent subjective beliefs or side information about the true value θ_0, or can be a methodological practice to construct statistical decision rules.

Example 3.1. Consider a finite parameter space $\Theta = \{\theta_1, \ldots, \theta_k\}$ and possible hypotheses $H_i : \theta = \theta_i$ for $1 \leq i \leq k$, with prior beliefs $\pi_i = \mathbf{P}(H_i)$. If the true hypothesis is H_i, the distribution of the observation X is $f_i(x)$, i.e.

$$\mathbf{P}(X = x \,|\, H_i) = f_i(x)\,.$$

By Bayes rule, when observing $X = x$, we have

$$\mathbf{P}(H_i \,|\, X = x) = \frac{\mathbf{P}(X = x \text{ and } H_i)}{\mathbf{P}(X = x)} = \frac{\pi_i f_i(x)}{\sum_j \pi_j f_j(x)}\,.$$

We will prefer H_i over H_j given this observation if

$$\frac{\mathbf{P}(H_i \,|\, X = x)}{\mathbf{P}(H_j \,|\, X = x)} = \frac{f_i(x)}{f_j(x)} \frac{\pi_i}{\pi_j} \geq 1\,.$$

If all the π_i are equal, this will be driven only by the likelihood ratio $f_i(x)/f_j(x)$. Otherwise, the priors give a way to update this rule, according to the prior knowledge or information given by π.

3.1. *Basic ideas, prior and posterior.* In a statistical model $\{f(\cdot, \theta) : \theta \in \Theta\}$, we say that the law of X given θ is $X \,|\, \theta \sim f(x, \theta)$. The posterior distribution is defined as the law of $\theta \,|\, X$.

Definition 3.1. For a sample space $\mathcal{X}$ where the observation X takes values, we consider the product space $\mathcal{X} \times \Theta$ and a probability measure Q with p.d.f./p.m.f.

$$Q(x, \theta) = f(x, \theta)\, \pi(\theta)$$

The distribution π is the *prior distribution* of θ. As expected, it has conditional probability

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

with a sum for a p.m.f. The *posterior distribution* is defined as

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

If $X = (X_1,\ldots,X_n)^\top$, all i.i.d. copies of law $f(x,\theta)$, then

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

Principles of Statistics **Part II - Michaelmas 2018**

Lecture 11: Between prior and posterior

Lecturer: Quentin Berthet

As seen in the previous lecture, in a statistical model $\{f(\cdot,\theta) : \theta \in \Theta\}$ with prior distribution $\pi(\theta)$, the *posterior distribution* is defined as

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

In the case where $X = (X_1,\dots,X_n)^\top$, for X_i i.i.d. copies of law $f(x,\theta)$, then

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

Remark. The posterior distribution given n i.i.d. observations is simply a reweighted (and renormalized) version of the likelihood function. Note also that the denominator is a normalizing constant: it does not describe the dependency of θ of this distribution. In practice it can therefore be ignored in computations, as seen in the following example.

Example 3.2. Let $X\,|\,\theta \sim \mathcal{N}(\theta,1)$ with prior $\theta \sim \mathcal{N}(0,1)$. The numerator of the posterior distribution, as a function of θ is therefore proportional to

$$\begin{aligned} e^{-\frac{\theta^2}{2}} \prod_{i=1}^n \exp\big(-\frac{(X_i-\theta)^2}{2}\big) &\propto \exp\big(-n\theta\bar{X} - \frac{n\theta^2}{2} - \frac{\theta^2}{2}\big) \\ &\propto \exp\big(-n\theta\bar{X} - \frac{(n+1)\theta^2}{2}\big) \\ &\propto \exp\big(-\frac{(\theta\sqrt{n+1} - n\bar{X}/\sqrt{n+1})^2}{2}\big) \\ &\propto \exp\big(-\frac{(\theta - n\bar{X}/(n+1))^2}{2/(n+1)}\big) \end{aligned}$$

Using that a distribution has normalization 1, the posterior distribution gives

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

The general case of $\mathcal{N}(\theta,\sigma^2)$ with prior $\mathcal{N}(\mu,\nu^2)$ is given in Examples sheet.

Informal notes, based on past lecture notes by Richard Nickl. Please let me know of any errors.

We remark that in this example, the posterior distribution (our belief about the value of θ after observation of the data) is in the same class of distribution than the prior distribution (our belief before observation). The distribution is still normal, the parameters have been updated based on the X_i.

Definition 3.2. In a statistical model $\{f(\cdot, \theta) : \theta \in \Theta\}$, when the prior $\pi(\theta)$ and $\Pi(\theta|X)$ belong to the same family of distributions, it is called a *conjugate prior*.

Example 3.3. Some of the following examples are covered in examples sheet

- Normal prior and normal sampling for a normal posterior distribution.
- Beta prior and binomial sampling for a Beta posterior distribution.
- Gamma prior and Poisson sampling for Gamma posterior.

Note that the definition of the posterior distribution can be extended to the case when π is not a probability distribution, i.e. does not integrate to 1. This is obvious when the integral of $\pi(\theta)$ over Θ is finite, as a simple renormalization will yield a proper prior without affecting the posterior distribution. More generally, it is only required that $f(X, \theta)\pi(\theta)$ has a finite integral over Θ.

Definition 3.3. A prior nonnegative function with infinite integral over Θ is called an *improper prior*.

This can be very useful in order to build a prior that is as uninformative as possible, by example by taking $\pi(\theta) = 1$, not assigning more weight to some values of θ than others. The following way allows to build a prior that is invariant to reparametrization.

Definition 3.4. The prior $\pi(\theta)$ proportional to $\sqrt{\det(I(\theta))}$ is called the *Jeffreys prior*.

Example 3.4. In a $\mathcal{N}(\mu, \tau)$ model with $\theta = (\mu, \tau)^\top \in \mathbf{R} \times (0, \infty)$, the Fisher information matrix is equal to

$$I(\theta) = I(\mu, \tau) = \begin{pmatrix} \frac{1}{\tau} & 0 \\ 0 & \frac{1}{2\tau^2} \end{pmatrix} .$$

As a consequence, the Jeffreys prior is given by $\pi(\mu, \tau) \propto 1/\tau^{3/2}$, which is constant in μ. In this case, the posterior marginal distribution for μ is $\mathcal{N}(\bar{X}_n, \tau/n)$.

3.2. *Statistical inference with the posterior distribution.* The posterior distribution $\Pi(\cdot \,|\, X_1, \dots, X_n)$ is a random probability measure on the parameter space Θ. It can be used to address several statistical questions about θ.

Definition 3.5. For a Bayesian model with posterior $\Pi(\cdot \,|\, X_1, \dots, X_n)$

- Estimation: We can take, as an estimator of θ, the *posterior mean* $\bar{\theta}$ defined as

$$\bar{\theta}(X_1, \dots, X_n) = \mathbf{E}_\Pi(\theta | X_1, \dots, X_n) .$$

- Uncertainty Quantification: Any subset $\mathcal{C}_n \subseteq \Theta$ such that

  $$\Pi(\mathcal{C}_n \,|\, X_1, \dots, X_n) = 1 - \alpha\,,$$

  is level $1-\alpha$ *credible set* for θ.

- Hypothesis Testing: As in the motivating example from the previous lecture, the *Bayes factor* satisfies

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

Note that for all of these tasks, there is in general no guarantee that the proposed method will have a satisfactory performance. In Example 3.4, the posterior mean of $\theta = (\mu, \tau)$ gives $\bar{\mu}_n = \bar{X}_n$, which is also the MLE, while in the case of Example 3.2, it is equal to $\frac{n}{n+1}\bar{X}_n$. In both cases, we see that the posterior mean will converge to the true value θ (i.e. the realization of the random variable), and that we know the limiting distribution of its deviations around this value.

Principles of Statistics **Part II - Michaelmas 2018**

Lecture 12: Frequentist analysis of Bayesian methods

Lecturer: Quentin Berthet

3.3. *Frequentist behavior of posterior distributions.* The inference procedures based on the posterior distribution, as described in the past lecture, can be analyzed from a frequentist point of view, assuming that $X_i \overset{i.i.d.}{\sim} f(x, \theta_0)$.

EXAMPLE 3.5. Sampling X from $\mathcal{N}(\theta, 1)$ where $\theta \sim \mathcal{N}(0, 1)$ gives, as seen in last lecture

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

The posterior mean is given by $\bar{\theta}_n = \mathbf{E}_\Pi[\theta \,|\, X_1, \dots, X_n] = \frac{1}{n+1}\sum_{i=1}^n X_i = \frac{n}{n+1}\bar{X}_n$. It is not exactly equal to the MLE $\hat{\theta}_n = \bar{X}_n$, but it is very close. In particular we have under the assumption $X_i \overset{i.i.d.}{\sim} \mathcal{N}(\theta_0, 1)$

$$\bar{\theta}_n = \frac{n}{n+1}\hat{\theta}_n \xrightarrow{P_{\theta_0}} \theta_0\,,$$

by Slutsky's lemma. We can also expect the deviations to be of the same order under $X_i \overset{i.i.d.}{\sim} \mathcal{N}(\theta_0, 1)$

$$\sqrt{n}(\bar{\theta}_n - \theta_0) = \sqrt{n}(\bar{\theta}_n - \hat{\theta}_n) + \sqrt{n}(\hat{\theta}_n - \theta_0)\,.$$

The second term satisfies $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d.} \mathcal{N}(0, 1)$. Expanding the first term gives

$$\sqrt{n}(\bar{\theta}_n - \hat{\theta}_n) = \sqrt{n}\Big(\frac{1}{n+1} - \frac{1}{n}\Big)\sum_{i=1}^n X_i = -\frac{\sqrt{n}}{n+1}\big(\bar{X}_n - \theta_0 + \theta_0\big) \xrightarrow{P_{\theta_0}} 0\,,$$

by Slutsky's lemma. Applying it a second time yields that the sum of the two terms converges in distribution to $\mathcal{N}(0, 1)$. One of the consequences is that $\bar{\theta}_n$ can replace $\hat{\theta}_n$ as the "center" of confidence regions, i.e. we can use

$$\mathcal{C}_n = \Big\{\nu : |\nu - \bar{\theta}_n| \leq \frac{I(\theta_0)^{-1/2} z_\alpha}{\sqrt{n}}\Big\}\,,$$

in a one-dimensional model. Here $I(\theta_0) = 1$ does not depend on the true value of θ so there is no issue with estimating it as well. However, in Bayesian analysis, inference is based not on the asymptotic distribution of an estimator, but on the posterior

Informal notes, based on past lecture notes by Richard Nickl. Please let me know of any errors.

distributions, and are considered credible sets of the form

$$\mathcal{C}_n = \big\{\nu : |\nu - \hat{\theta}_n| \leq \frac{R_n}{\sqrt{n}}\big\} \quad \text{or} \quad \big\{\nu : |\nu - \bar{\theta}_n| \leq \frac{R_n}{\sqrt{n}}\big\}\,,$$

taking R_n such that $\Pi(\mathcal{C}_n \,|\, X_1, \ldots, X_n) = 1 - \alpha$. Note that R_n is a random variable depending on the observations X_i. In order to prove that credible sets of these type are frequentists confidence sets, i.e. that $\mathbf{P}_{\theta_0}(\theta_0 \in \mathcal{C}_n) \to 1 - \alpha$, it is therefore important to understand the behavior of $\Pi_n = \Pi(\cdot \,|\, X_1, \ldots, X_n)$ as a random probability distribution.

The Bernstein–von Mises theorem states that the posterior distribution behaves, for large n, like a normal distribution centered at an efficient estimator, such as the MLE $\hat{\theta}_n$.

Theorem 3.1 (Bernstein–von Mises). *For a parametric model $\{f(\cdot, \theta) : \theta \in \Theta\}$ of $\Theta \subseteq \mathbf{R}$ satisfying the regularity assumptions of Theorem 2.6, a prior with continuous density π at θ_0 with $\pi(\theta_0) > 0$, and the associated posterior $\Pi_n = \Pi(\cdot \,|\, X_1, \ldots, X_n)$. Let ϕ_n be the random distribution $\mathcal{N}(\hat{\theta}_n, I(\theta_0)^{-1}/n)$. We have as $n \to \infty$*

$$\|\Pi_n - \phi_n\|_{TV} = \int_\Theta |\Pi_n(\theta) - \phi_n(\theta)| \mathrm{d}\theta \xrightarrow{a.s.} 0\,.$$

Remark. We recall that the setting adopted here is the following. Given observations $x_1, \ldots, x_n$, the Bayesian formalism gives a posterior distribution for the parameter θ. When the observations are $X_i \overset{i.i.d.}{\sim} f(x, \theta)$, the posterior is a random distribution. The distribution ϕ_n is also random, and this result states that these two distributions look increasingly alike when $n \to \infty$.

This implies that for any subset $A \subseteq \Theta$, we have $\Pi_n(A) - \phi_n(A) \to 0$, almost surely. As a consequence, for any credible set $\mathcal{C}_n$, we have that $\phi_n(\mathcal{C}_n) \to 1 - \alpha$, which is helpful in showing that they are frequentist confidence regions of level $1 - \alpha$.

Proof beginning. Since Π_n and ϕ_n are probability distributions, they integrate to 1, and we have

$$\int_\Theta (\Pi_n(\theta) - \phi_n(\theta)) \mathrm{d}\theta = 1 - 1 = 0\,.$$

This means that the integral of the positive and negative parts of $(\Pi_n(\theta) - \phi_n(\theta))$ are equal, so the integral of the absolute value is equal to twice the integral of the positive part, i.e.

$$\begin{aligned}\int_\Theta |\Pi_n(\theta) - \phi_n(\theta)| \mathrm{d}\theta &= 2 \int_\Theta (\Pi_n(\theta) - \phi_n(\theta))^+ \mathrm{d}\theta \\ &= 2 \int_\Theta (\phi_n(\theta) - \Pi_n(\theta))^+ \mathrm{d}\theta \\ &= 2 \int_\Theta \Big(1 - \frac{\Pi_n(\theta)}{\phi_n(\theta)}\Big)^+ \phi_n(\theta) \mathrm{d}\theta\,.\end{aligned}$$

Since the function $x \mapsto (1-x)^+$ is bounded by 1, if it can be shown that $\Pi_n(\theta)/\phi_n(\theta)$ converges almost surely to 1 for all $\theta \in \Theta$, the dominated convergence theorem (from *Probability and Measure*) guarantees that the result holds. At this level, the result is assumed, and special cases will be covered in Examples sheet. □

Remark. While a full proof is beyond the scope of this course, some informal intuition can be given to illustrate this result. Note that it does not in any case constitute a formal proof, and explicit examples can be done by hand. Those in the Examples sheet must be done rigorously.

We have that

$$\Pi_n(\theta) = \frac{\pi(\theta)\prod_{i=1}^n f(X_i,\theta)}{Z_n}\,,$$

where Z_n is a normalization factor independent of θ. The distribution of the variations of $V = \sqrt{n}(\theta - \hat\theta_n)$ have density $\Pi_{n,V}$ that satisfies

$$\Pi_{n,V}(v) = \frac{1}{\sqrt{n}}\Pi_n\big(\hat\theta_n + \frac{v}{\sqrt{n}}\big)$$

As a consequence, taking logarithms yields

$$\begin{aligned}\log\Pi_{n,V}(v) &= \log\Pi_n\big(\hat\theta_n + \frac{v}{\sqrt{n}}\big) + \log\frac{1}{\sqrt{n}} = \log\pi\big(\hat\theta_n + \frac{v}{\sqrt{n}}\big) + \ell_n(\hat\theta_n + \frac{v}{\sqrt{n}}\big) - \log Z_n' \\ &\approx \log\pi(\theta_0) + \ell_n(\hat\theta_n) + \ell_n'(\hat\theta_n)\frac{v}{\sqrt{n}} + \frac{1}{2}\ell_n''(\hat\theta_n)\frac{v^2}{n} - \log Z_n' \approx -\frac{1}{2}I(\theta_0)v^2 - \log\tilde Z_n\,,\end{aligned}$$

with the approximation being valid for large n Note that under $\theta \sim \mathcal{N}(\hat\theta_n, I(\theta_0)^{-1})$, we have $V \sim \mathcal{N}(0, I(\theta_0)^{-1})$ and

$$\log\phi_{n,V} = -\frac{1}{2}I(\theta_0)v^2 - \log C(\theta_0)\,.$$

The main idea is based on the fact that a function of the form $e^{nf(\theta)}$ is exponentially smaller than its maximum for all θ such that $f(\theta) < f(\theta^*)$, where θ^* maximizes f. If the approximation

$$f(\theta^* + h) \approx f(\theta^*) + \nabla_\theta f(\theta^*)\cdot h + \frac{1}{2}h^\top\nabla_\theta^2 f(\theta^*)v\,,$$

holds for small h, the optimality conditions $\nabla_\theta f(\theta^*) = 0$ and $\nabla_\theta^2 f(\theta^*) = -Q \preceq 0$ yield for "constant x"

$$f(\theta^* + \frac{x}{\sqrt{n}}) \approx f(\theta^*) - \frac{1}{2n}x^\top Qx\,,\quad \text{and}\quad e^{nf(\theta^*+\frac{x}{\sqrt{n}})} \approx e^{nf(\theta^*)}e^{-\frac{1}{2}x^\top Qx}\,.$$

This can be useful when computing integrals or approximating distributions, and is known as Laplace's method.

Principles of Statistics **Part II - Michaelmas 2018**

Lecture 13: Credible sets as confidence sets

Lecturer: Quentin Berthet

Our objective is to show that credible sets of the form

$$\mathcal{C}_n = \big\{\nu : |\nu - \hat{\theta}_n| \leq \frac{R_n}{\sqrt{n}}\big\}\,,$$

with R_n chosen such that $\Pi_n(\mathcal{C}_n) = \Pi(\mathcal{C}_n \,|\, X_1, \dots, X_n) = 1-\alpha$ are also frequentist confidence sets, i.e. that for $X_i \sim f(x, \theta_0)$, we have that $\mathbf{P}_{\theta_0}(\theta_0 \in \mathcal{C}_n) \to 1-\alpha$ when $n \to \infty$. The proof can be done in two parts, first by showing that if R_n converges almost surely to its frequentist equivalent, this probability converges to $1-\alpha$, then by showing that R_n converges indeed to this limit.

Definition 3.6. For all $t > 0$, we define the function Φ_0 by

$$\Phi_0(t) = \mathbf{P}(|Z_0| \leq t) = \int_{-t}^{t} \varphi_0(x) \mathrm{d}x\,,$$

for $Z_0 \sim \mathcal{N}(0, I(\theta_0)^{-1})$. It is an increasing, continuous one-to-one mapping from $[0, \infty)$ to $[0, 1)$. Its well-defined functional inverse is also continuous and denoted by Φ_0^{-1}.

Lemma 3.1. *Under the assumptions above, we have that* $R_n \xrightarrow{a.s.} \Phi_0^{-1}(1-\alpha)$, *as* $n \to \infty$.

Proof. We have

$$\begin{aligned}\Phi_0(R_n) &= \int_{-R_n}^{R_n} \varphi_0(v)\mathrm{d}v \\ &= \int_{\hat{\theta}_n - R_n/\sqrt{n}}^{\hat{\theta}_n + R_n/\sqrt{n}} \phi_n(\theta)\mathrm{d}\theta \quad \text{for} \quad v = \sqrt{n}(\theta - \hat{\theta}_n) \\ &= \phi_n(\mathcal{C}_n) - \Pi_n(\mathcal{C}_n) + \Pi_n(\mathcal{C}_n)\end{aligned}$$

By the Bernstein–von Mises theorem, the first difference converges to 0 almost surely, and the second term is equal to $1-\alpha$. As a consequence, when $n \to \infty$, $\Phi_0(R_n) \xrightarrow{a.s.} 1-\alpha$. Applying the continuous mapping theorem with Φ_0^{-1} allows us to conclude. □

Informal notes, based on past lecture notes by Richard Nickl. Please let me know of any errors.

Theorem 3.2. *Under the assumptions above, for $\alpha \in (0,1)$ and $n \to \infty$, we have that $\mathbf{P}_{\theta_0}(\theta_0 \in \mathcal{C}_n) \to 1-\alpha$.*

Proof. By Slutsky's lemma, given that $\Phi_0^{-1}(1-\alpha) > 0$, we have that

$$\frac{\Phi_0^{-1}(1-\alpha)}{R_n}\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d.} \mathcal{N}(0, I(\theta_0)^{-1}).$$

As a consequence, we have that

$$\begin{aligned}\mathbf{P}_{\theta_0}(\theta_0 \in \mathcal{C}_n) &= \mathbf{P}_{\theta_0}(|\hat{\theta}_n - \theta_0| \le R_n/\sqrt{n}) \\ &= \mathbf{P}_{\theta_0}\Big(\frac{\Phi_0^{-1}(1-\alpha)}{R_n}\sqrt{n}|\hat{\theta}_n - \theta_0| \le \Phi_0^{-1}(1-\alpha)\Big) \\ &\to \mathbf{P}(|Z_0| \le \Phi_0^{-1}(1-\alpha)) = \Phi_0(\Phi_0^{-1}(1-\alpha)) = 1-\alpha.\end{aligned}$$

□

Remark. A similar result holds with $\bar{\theta}_n$ (the posterior mean) instead of $\hat{\theta}_n$ (see Examples sheet).

4. Decision theory. Given a statistical model $\{f(\cdot,\theta) : \theta \in \Theta\}$ and an observation sample $X \in \mathcal{X}$, we can phrase many statistical problems as *decision problems*, with an *action space* $\mathcal{A}$ and decision rules

$$\delta : \mathcal{X} \to \mathcal{A}.$$

Example 4.1. Explicitly, for several statistical problems

- In a hypothesis testing problem, $\mathcal{A} = \{0,1\}$ the decision $\delta(X)$ is a test.
- In an estimation problem, $\mathcal{A} = \Theta$ and $\delta(X) = \hat{\theta}(X)$ is an estimation problem.
- In inference problems, $\mathcal{A} =$ "subsets of Θ" and $\delta(X) = \mathcal{C}(X)$ is a confidence set.

The performance of a decision rule is assessed by a *loss function*, that determines the value of a particular action, for a given θ.

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

Example 4.2.

- In an hypothesis testing problem, the answer is right or wrong, and for $\theta \in \{0,1\}$ representing the index of the hypothesis, we can take

$$L(a,\theta) = \mathbf{1}\{a \ne \theta\}.$$

- In an estimation problem, we are concerned with the distance between our estimate and the true value, so we can take the absolute error or the squared error

$$L(a,\theta) = |a-\theta| \quad \text{or} \quad |a-\theta|^2.$$

For any decision rule, we can consider the average loss under the distribution of X

Definition 4.1. For a loss function L, and a decision rule δ we have for $X \sim P_\theta$

$$R(\delta, \theta) = \mathbf{E}_\theta[L(\delta(X), \theta)] = \int_{\mathcal{X}} L(\delta(x), \theta) f(x, \theta) \mathrm{d}x \,.$$

Example 4.3.

- In a hypothesis testing problem, $R(\delta, \theta) = \mathbf{E}_\theta \mathbf{1}\{\delta(X) \neq \theta\} = \mathbf{P}_\theta(\delta(X) \neq \theta)$ describes the probability of error (type I/II).

- In an estimation problem, the *quadratic risk* is the mean squared error (MSE) equal to $\mathbf{E}_\theta[(\delta(X) - \theta)^2] = \mathbf{E}_\theta[(\hat{\theta}(X) - \theta)^2]$

- For $X \sim \mathrm{Bin}(n, \theta)$ and $\theta \in [0, 1]$, we can take $\hat{\theta}(X) = X/n$

  $$R(\hat{\theta}, \theta) = \mathbf{E}_\theta[(\hat{\theta}(X) - \theta)^2] = \frac{\theta(1-\theta)}{n} \,,$$

  If we had taken a naive estimator $\eta(X) = 1/2$, we would have

  $$R(\hat{\eta}, \theta) = \mathbf{E}_\theta[(\hat{\eta}(X) - \theta)^2] = (\theta - 1/2)^2 \,.$$

As shown in the example above, it is not possible in general to compare uniformly the performance of two estimators: $R(\hat{\theta}, \cdot)$ and $R(\hat{\eta}, \cdot)$ cannot be uniformly compared in $\theta \in [0, 1]$.

Principles of Statistics **Part II - Michaelmas 2018**

Lecture 14: Bayesian risk

Lecturer: Quentin Berthet

4.1. *Bayes rule for risk minimization.*

Definition 4.2. Given a prior π on Θ, the π-*Bayes risk* of δ for the loss function L is defined as

$$R_\pi(\delta) = \mathbf{E}_\pi[R(\delta,\theta)] = \int_\Theta R(\delta,\theta)\pi(\theta)\mathrm{d}\theta = \int_\Theta \int_{\mathcal{X}} L(\delta(x),\theta)\pi(\theta)f(x,\theta)\mathrm{d}x\mathrm{d}\theta\,.$$

A π-*Bayes decision rule* δ_π is any decision rule that minimizes $R_\pi(\delta)$.

Example 4.4. In a binomial model $X \sim \mathrm{Bin}(n,\theta)$ with uniform prior on $[0,1]$ for θ,we have for the quadratic loss $R(X/n,\theta) = \theta(1-\theta)/n$

$$R_\pi(X/n) = \mathbf{E}_\pi\Big[\frac{\theta(1-\theta)}{n}\Big] = \frac{1}{n}\int_0^1 \theta(1-\theta)\mathrm{d}\theta = \frac{1}{6n}\,.$$

Definition 4.3. For a Bayesian model, the *posterior risk* R_Π is defined as the average loss under the posterior distribution for some observation $x \in \mathcal{X}$.

$$R_\Pi(\delta) = \mathbf{E}[L(\delta(x),\theta)|x]\,.$$

Remark. We recall that the expectation is taken over θ here, not over the observation X. In the binomial model, with a quadratic loss, we have for any estimator $\delta(X)$

$$\begin{aligned} R_\Pi(\delta) = \mathbf{E}_\Pi[(\delta(x)-\theta)^2|x] &= \mathbf{E}_\Pi[\delta(x)^2 - 2\delta(x)\theta + \theta^2|x] \\ &= \delta(x)^2 - 2\delta(x)\mathbf{E}_\Pi[\theta|x] + \mathbf{E}_\Pi[\theta^2|x]\,. \end{aligned}$$

Proposition 4.1. *An estimator δ that minimizes the Π-posterior risk R_Π also minimizes the π-Bayes risk R_π.*

Informal notes, based on past lecture notes by Richard Nickl. Please let me know of any errors.

Proof. The π-Bayes risk can be rewritten as

$$\begin{aligned}
R_\pi(\delta) &= \int_\Theta \mathbf{E}_\theta[L(\delta(X),\theta)]\pi(\theta)\mathrm{d}\theta \\
&= \int_\Theta \int_{\mathcal{X}} L(\delta(x),\theta) f(x,\theta)\pi(\theta)\mathrm{d}x\,\mathrm{d}\theta \\
&= \int_{\mathcal{X}} \int_\Theta L(\delta(x),\theta) \frac{f(x,\theta)\pi(\theta)}{\int_\Theta f(x,\theta')\pi(\theta')\mathrm{d}\theta'} \times \underbrace{\int_\Theta f(x,\theta')\pi(\theta')\mathrm{d}\theta'}_{:=m(x)\geq 0} \times \mathrm{d}x\,\mathrm{d}\theta \\
&= \int_{\mathcal{X}} \mathbf{E}_\Pi[L(\delta(x),\theta)|x]\, m(x)\mathrm{d}x\,.
\end{aligned}$$

Let δ_Π be a decision rule that minimizes the posterior risk, i.e. such that for all $x \in \mathcal{X}$

$$\mathbf{E}_\Pi[L(\delta_\Pi(x),\theta)|x] \leq \mathbf{E}_\Pi[L(\delta(x),\theta)|x]$$

Multiplying by $m(x) \geq 0$ and integrating on both sides over $\mathcal{X}$ yields the desired result. □

Example 4.5. For the quadratic risk with the squared-loss, the posterior risk is minimized by taking $\delta(X) = \mathbf{E}_\Pi[\theta|X]$, by minimizing the quadratic function in δ. Other losses will give over ways to minimize the posterior risk, and other Bayes decision rules.

Proposition 4.2. *Let δ be an unbiased decision rule for θ, i.e. $\mathbf{E}_\theta[\delta(X)] = \theta$ for all $\theta \in \Theta$. If δ is also a Bayes rule for some prior π in the quadratic risk, then*

$$\mathbf{E}_Q[(\delta(X)-\theta)^2] = \int_\Theta \mathbf{E}_\theta[(\delta(X)-\theta)^2]\pi(\theta)\mathrm{d}\theta = 0\,,$$

where $\mathbf{E}_Q$ is the expectation taken with respect to (X,θ) under the joint distribution $Q(x,\theta) = f(x,\theta)\pi(\theta)$. In particular $\delta(X) = \theta$ with Q-probability 1.

Proof. We recall that for any random variable $Z(x,\theta)$, by applying the "tower rule" in two different manners, we have

$$\begin{aligned}
\mathbf{E}_Q[Z(X,\theta)] &= \mathbf{E}_Q[\mathbf{E}_\Pi[Z(X,\theta)|X]] \\
&= \mathbf{E}_Q[\mathbf{E}_\theta[Z(X,\theta)]]\,.
\end{aligned}$$

For a π-decision rule δ for the quadratic risk, we recall that $\delta(X) = \mathbf{E}_\Pi[\theta|X]$. As a consequence, taking $Z(X,\theta) = \theta\delta(X)$ in the expressions above gives

$$\mathbf{E}_Q[\theta\delta(X)] = \mathbf{E}_Q[\mathbf{E}_\Pi[\theta\delta(X)|X]] = \mathbf{E}_Q[\delta(X)\mathbf{E}_\Pi[\theta|X]] = \mathbf{E}_Q[\delta(X)^2]$$

and

$$\mathbf{E}_Q[\theta\delta(X)] = \mathbf{E}_Q[\mathbf{E}_\theta[\theta\delta(X)]] = \mathbf{E}_Q[\theta\,\mathbf{E}_\theta[\delta(X)]] = \mathbf{E}_Q[\theta^2]$$

by unbiasedness. We therefore have

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

□

Remark. A direct consequence is that unbiased estimators are typically disjoint from Bayes estimators.

- In a $\mathcal{N}(\theta, 1)$ model, the MLE $\bar{X}_n$ is not a Bayes estimator for any prior π.
- In a $\mathrm{Bin}(n, \theta)$ model, the MLE X/n is only Bayes in very degenerate cases (see Examples sheet).

Definition 4.4. A prior λ is called *least favorable* if for every prior λ'

$$R_\lambda(\delta_\lambda) \geq R_{\lambda'}(\delta_{\lambda'}) \,,$$

corresponding to the worst case Bayesian risk.

Principles of Statistics **Part II - Michaelmas 2018**

Lecture 15: Minimax risk and admissibility

Lecturer: Quentin Berthet

4.2. *Minimax risk.* We have seen that the Bayesian risk allows us to consider the average loss of estimators over values of θ, by taking a prior $\pi(\theta)$. Another approach is to consider the *worst case* over all values of θ.

Definition 4.5. The *maximal risk* of the decision rule δ over the parameter space Θ is defined as

$$R_m(\delta, \Theta) = \sup_{\theta \in \Theta} R(\delta, \theta)\,.$$

Definition 4.6 (Minimax risk). The *minimax risk* is defined as the infimum (or "min") of the maximum risk

$$\inf_{\delta} \sup_{\theta \in \Theta} R(\delta, \theta) = \inf_{\delta} R_m(\delta, \Theta)\,.$$

A decision rule that attains this maximum risk is called *minimax.*

Proposition 4.3. *For any prior θ and decision rule δ, we have*

$$R_\pi(\delta) \le R_m(\delta, \Theta)\,.$$

Proof. We have that

$$R_\pi(\delta) = \mathbf{E}_\pi[R(\delta, \theta)] \le \sup_{\theta \in \Theta} R(\delta, \theta) = R_m(\delta, \Theta)\,.$$

□

As an average risk, the Bayes risk is never greater than the worst case maximum risk.

Definition 4.7. A prior λ is called *least favorable* if for every prior λ'

$$R_\lambda(\delta_\lambda) \ge R_{\lambda'}(\delta_{\lambda'})\,,$$

corresponding to the worst case Bayesian risk.

Informal notes, based on past lecture notes by Richard Nickl. Please let me know of any errors.

PROPOSITION 4.4. *Let λ be a prior on Θ such that*

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

where δ_λ is a λ-Bayes rule. Then it holds that

1. *The rule δ_λ is minimax*

2. *If δ_λ is unique Bayes then it is unique minimax.*

3. *The prior λ is least favorable.*

PROOF. 1. Let δ be any decision rule. Then

$$\begin{aligned}
\sup_{\theta\in\Theta} R(\delta,\theta) &\geq \mathbf{E}_\lambda[R(\delta,\theta)] \\
\int_\Theta R(\delta,\theta)\lambda(\theta)\mathrm{d}\theta \\
&\geq \mathbf{E}_\lambda[R(\delta_\lambda,\theta)] \\
\int_\Theta R(\delta_\lambda,\theta)\lambda(\theta)\mathrm{d}\theta \\
&= R_\lambda(\delta_\lambda) \\
&= \sup_{\theta\in\Theta} R(\delta_\lambda,\theta)
\end{aligned}$$

Taking infimum over δ gives the result.

2. If δ_λ is unique Bayes the second inequality is strict for any $\delta' \neq \delta_\lambda$.

3. For any prior λ', we have

$$\begin{aligned}
R_{\lambda'}(\delta_{\lambda'}) &= \mathbf{E}_{\lambda'}[R(\delta_{\lambda'},\theta)] \\
&\leq \mathbf{E}_{\lambda'}[R(\delta_\lambda,\theta)] \quad \text{by definition of } \delta_{\lambda'} \\
&\leq \sup_{\theta\in\Theta} R(\delta_\lambda,\theta) \\
&= \mathbf{E}_\lambda[R(\delta_\lambda,\theta)]\,.
\end{aligned}$$

□

COROLLARY 4.1. *If a (unique) Bayes rule δ_λ has constant risk in θ, then it is (unique) minimax.*

PROOF. If a Bayes rule δ_λ has constant risk, then

$$R_\lambda(\delta_\lambda) = \mathbf{E}_\lambda[\underbrace{R(\delta_\lambda,\theta)}_{\text{const. in }\theta}] = \sup_{\theta\in\Theta} R(\delta_\lambda,\theta)\,,$$

so the hypothesis in Proposition 4.7 is satisfied. Uniqueness of the Bayes rule implies uniqueness of the minimax rule, as in part 2. of the proposition. □

Example 4.6. Hence if the maximal risk of a Bayes rule δ_λ equals the Bayes risk, then λ is least favorable and the corresponding Bayes rule is minimax.

- In a $\text{Bin}(n, \theta)$ model, let $\pi_{a,b}$ be a $\text{Beta}(a, b)$ prior on $\theta \in [0, 1]$. Then the unique Bayes rule for $\pi_{a,b}$ over the quadratic risk is the posterior mean $\delta_{a,b} = \bar{\theta}_{a,b}$. Trying to solve the equation

$$R(\delta_{a,b}, \theta) = \text{const.} \quad \forall \theta \in [0, 1]$$

  we can find a prior π_{a^*,b^*} and corresponding Bayes rule $\delta_{\pi_{a^*,b^*}}$ of constant risk. It is therefore unique minimax, and different from the MLE (see Examples sheet).

- In a $\mathcal{N}(\theta, 1)$ model, $\bar{X}_n$ is minimax, as proved later.

4.3. *Admissibility.*

Definition 4.8. A decision rule δ is *inadmissible* if there exists δ' such that

$$\begin{aligned} R(\delta', \theta) &\leq R(\delta, \theta) \quad \forall \theta \in \Theta \quad \text{and} \\ R(\delta', \theta) &< R(\delta, \theta) \quad \text{for some } \theta \in \Theta. \end{aligned}$$

Remark.

- The intuition is that there is no reason to chose an inadmissible estimator or decision rule: it would always be better to chose another estimator that dominates it.

- Admissibility is not the only criterion to evaluate an estimator: In most cases, a constant estimator will be admissible for the quadratic risk, but it is often not reasonable.

Proposition 4.5.

i) A unique Bayes rule is admissible.

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

Proof is done in the Examples sheet.

Principles of Statistics **Part II - Michaelmas 2018**

Lecture 16: Admissibility in the Gaussian model

Lecturer: Quentin Berthet

As seen in the previous lecture (with proof left to the Examples sheet), an important manner of showing that an estimator - or decision rule in general - is minimax, is to show that it has constant risk and that it is admissible. An example, given for illustration, is that of the binomial model, where a certain Beta prior can be shown to have a corresponding Bayes rule with constant risk. In this lecture, we study a case where an estimator that is *not Bayes* for any prior can be shown to be directly admissible.

Proposition 4.6. *Let $X_1, \dots, X_n$ be i.i.d. from a Gaussian $\mathcal{N}(\theta, \sigma^2)$, with known σ^2 and $\theta \in \Theta = \mathbf{R}$. Then $\hat{\theta}_{MLE} = \bar{X}_n$ is admissible and minimax for estimating θ in quadratic risk.*

Proof. We treat in this proof the case of $\sigma^2 = 1$ to simplify notation. The general case follows exactly the same proof.

We first remark that this estimator has constant risk

$$R(\hat{\theta}_{MLE}, \theta) = \mathbf{E}_\theta[L(\bar{X}_n, \theta)] = \mathbf{E}_\theta[(\bar{X}_n - \theta)^2] = \mathrm{Var}_\theta(\bar{X}_n) = \frac{1}{n}\,.$$

For any decision rule δ, we have that

$$R(\delta, \theta) = \mathbf{E}_\theta[L(\delta(X), \theta)] = \mathbf{E}_\theta[(\delta(X) - \theta)^2] = \big(\underbrace{\mathbf{E}_\theta[\delta(X)] - \theta}_{B(\theta)}\big)^2 + \mathrm{Var}_\theta(\delta(X))\,,$$

where $B(\theta) = \mathbf{E}_\theta[\delta(X)] - \theta$ denotes the bias of the estimator δ. We recall that the Cramèr–Rao lower bound only applies to unbiased estimators, but an examination of the proof gives

$$\mathrm{Var}_\theta(\delta) \geq \frac{\big(\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbf{E}_\theta[\delta]\big)^2}{nI(\theta)} = \frac{\big(1 + B'(\theta)\big)^2}{n}\,,$$

since $I(\theta) = 1$ in this model. As a consequence, if δ dominates $\bar{X}_n$, we have $R(\delta, \theta) \leq 1/n$ for all $\theta \in \mathbf{R}$ and

$$(\dagger) \qquad B(\theta)^2 + \frac{\big(1 + B'(\theta)\big)^2}{n} \leq \frac{1}{n}\,.$$

The differentiability of B is a consequence of the regularity of the Gaussian model. This immediately yields that $B(\theta)$ is bounded above and below, and that $B'(\theta) \leq 0$, so B

Informal notes, based on past lecture notes by Richard Nickl. Please let me know of any errors.

is nonincreasing. There are therefore two sequences $(\theta_n)_{n\geq 1}$ going one to $-\infty$ and the other to $+\infty$ such that $B'(\theta_n) \to 0$. Otherwise, $B'(\theta)$ would be bounded away from 0 for θ small and large enough, and $B(\theta)$ would be unbounded. As a consequence of the inequality (†), for both of those sequences, $B(\theta_n) \to 0$. Since B is nonincreasing, this yields that $B(\theta) = 0$ for all $\theta \in \mathbf{R}$ and the Cramèr–Rao lower bound applies.

$$\mathrm{Var}_\theta(\delta) \geq \frac{1}{n}$$

and $R(\delta, \theta) = 1/n$ for all θ. We have proved that if $R(\delta, \theta)$ is uniformly smaller than $1/n$, it is equal to $1/n$. Therefore δ does not dominate $\hat{\theta}_{MLE}$ and the MLE is admissible. Since it also has constant risk, it is also minimax, by Proposition 4.5. □

Remark. The decision rule studied here is not a Bayes rule for any prior, as seen in Examples sheet. It is however, in some sense, the limit of the Bayes rules δ_{ν^2} for prior $\mathcal{N}(0, \nu^2)$, when $\nu \to \infty$. It can be shown that all minimax rules are limits of Bayes rule. The result shown here can be extended to dimension $p = 2$, with a model $\mathcal{N}(\theta, I_2)$, where $\theta \in \Theta = \mathbf{R}^2$. It is however false for $p \geq 3$, which is the focus of the following.

4.4. *The James–Stein phenomenon.* At first sight, it appears intuitive that the result above, in dimension 1, should extend to any dimension: when estimating several quantities subject to independent Gaussian errors, if the "best" way to estimate each of them (under quadratic risk) is to take the given observation, why should one do any differently when considering them altogether? In this section, we show that in the model $X \sim \mathcal{N}(\theta, I_p)$, for $p \geq 3$, the MLE given by $\hat{\theta} = X$ is not admissible. The idea is that it is possible to construct a new estimator δ^{JS}, the James–Stein estimator which uses the whole vector to estimate any coordinate

Definition 4.9. For a vector $X \in \mathbf{R}^p$,the *James–Stein estimator* is defined as

$$\delta^{JS}(X) = \Big(1 - \frac{p-2}{\|X\|^2}\Big) X\,.$$

In a Gaussian model $X \sim \mathcal{N}(\theta, I_p)$ for $\theta \in \mathbf{R}^p$ (with a single observation, to simplify notation), the risk of the MLE is given by

$$R(\hat{\theta}_{MLE}, \theta) = \mathbf{E}_\theta[\|X - \theta\|^2] = \sum_{j=1}^{p} \mathbf{E}_\theta[(X_i - \theta_i)^2] = p\,.$$

In order to show that δ^{JS} dominates $\hat{\theta}_{MLE}$, we can explicitly compute the risk given by $R(\delta^{JS}, \theta) = \mathbf{E}_\theta[\|\delta^{JS}(X) - \theta\|^2]$. In this computation, we will use the following lemma.

Lemma 4.1 (Stein's lemma). *Let $X \sim \mathcal{N}(\theta, 1)$ and $g : \mathbf{R} \to \mathbf{R}$ be a bounded, differentiable function such that $\mathbf{E}|g'(X)| < \infty$. We have*

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

Proof. We compute explicitly

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

Integration by parts is taken on $\mathbf{R}$, but taking it on $[-N, N]$ and letting N go to ∞ yields the same result. The first term is equal to 0 because g is bounded, and e^{-x^2} goes to 0 at ±∞. □

Remark.

- One can think of this formula as of "Gaussian integration by parts": we have for the standard Gaussian density φ that $\mathcal{A}(\varphi)(x) = \varphi'(x) + x\,\varphi(x) = 0$, and integration by parts guarantees that for such functions g and regular densities p

$$\langle g, \mathcal{A}(p)\rangle = \langle \mathcal{A}^*(g), p\rangle\,,$$

  where $\mathcal{A}^*(g)(x) = -g'(x) + x\,g(x)$. The fact that $\mathcal{A}(\varphi) = 0$ implies $\langle \mathcal{A}^*(g), \varphi\rangle = 0$. This interpretation, from the point of view of analysis of functions can be used to show the converse of this result. Informally, if $\langle \mathcal{A}^*(g), p\rangle = 0$ over a large class of functions, then $\langle g, \mathcal{A}(p)\rangle = 0$. If this class is large enough, this implies that $\mathcal{A}(p)(x) = 0$, and the differential equation implies that $p = \varphi$,i.e. that the density is Gaussian.

- Since the converse is true, this can be used to show that some distributions are "almost Gaussian", if the identity above is "almost satisfied". This can be used to show that a distribution converges to a standard Gaussian, when it is too complicated to do so directly (e.g. sum of dependent random variables).

Principles of Statistics **Part II - Michaelmas 2018**

Lecture 17: Risk of the James–Stein estimator

Lecturer: Quentin Berthet

PROPOSITION 4.7. *Let $X \sim \mathcal{N}(\theta, I_p)$, for $p \geq 3$. The risk of the James–Stein estimator satisfies forall $\theta \in \mathbf{R}^p$*

$$R(\delta^{JS}, \theta) < p\,.$$

PROOF. We compute the risk directly

$$\begin{aligned} R(\delta^{JS}, \theta) &= \mathbf{E}_\theta[\|\delta^{JS}(X) - \theta\|^2] = \mathbf{E}_\theta\big[\big\|X - \theta - \frac{p-2}{\|X\|^2}X\big\|^2\big] \\ &= \mathbf{E}_\theta\big[\big\|X-\theta\big\|^2\big] + (p-2)^2\mathbf{E}_\theta\big[\big\|\frac{X}{\|X\|^2}\big\|^2\big] - 2(p-2)\mathbf{E}_\theta\big[\frac{X^\top(X-\theta)}{\|X\|^2}\big] \\ &= p + (p-2)^2\mathbf{E}_\theta\big[\frac{1}{\|X\|^2}\big] - 2(p-2)\mathbf{E}_\theta\big[\frac{X^\top(X-\theta)}{\|X\|^2}\big] \end{aligned}$$

The last term is the only potentially negative term here, we develop it explicitly

$$\mathbf{E}_\theta\big[\frac{X^\top(X-\theta)}{\|X\|^2}\big] = \sum_{j=1}^p \mathbf{E}_\theta\big[\frac{X_j(X_j-\theta_j)}{\|X\|^2}\big] = \sum_{j=1}^p \mathbf{E}_\theta\big[\mathbf{E}_j\big[(X_j-\theta_j)\frac{X_j}{X_j^2 + \sum_{i\neq j} X_i^2}|X_{(-j)}\big]\big]\,,$$

the last equality being an application of the tower property, conditioning on all the other coefficients $X_{(-j)} = X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_p$. Each of these conditional expectations is computed as

$$\mathbf{E}_j\big[(X_j-\theta_j)\frac{X_j}{X_j^2 + \sum_{i\neq j} X_i^2}|X_{(-j)}\big] = \mathbf{E}[(X_j-\theta_j)g_j(X_j)]$$

where $X_j \sim \mathcal{N}(\theta_j, 1)$ and $g_j(x) = x/(x^2 + \sum_{i \neq j} X_i^2)$ is a bounded function, with probability 1. We have that

$$g_j'(x) = \frac{x^2 + \sum_{i\neq j} X_i^2 - 2x^2}{(x^2 + \sum_{i\neq j} X_i^2)^2}$$

The derivative is also bounded, so $\mathbf{E}|g_j'(X_j)| < \infty$ and Stein's lemma applies, we have

$$\mathbf{E}[(X_j-\theta_j)g_j(X_j)] = \mathbf{E}[g'(X_j)] = \mathbf{E}_j\big[\frac{X_j^2 + \sum_{i\neq j} X_i^2 - 2X_j^2}{(X_j^2 + \sum_{i\neq j} X_i^2)^2}|X_{(-j)}\big]$$

Informal notes, based on past lecture notes by Richard Nickl. Please let me know of any errors.

and

$$\mathbf{E}_\theta\big[\mathbf{E}_j\big[(X_j-\theta_j)\frac{X_j}{X_j^2+\sum_{i\neq j}X_i^2}|X_{(-j)}\big]\big] = \mathbf{E}_\theta[\frac{X_j^2+\sum_{i\neq j}X_i^2-2X_j^2}{(X_j^2+\sum_{i\neq j}X_i^2)^2}] = \mathbf{E}_\theta\big[\frac{1}{\|X\|^2}-\frac{2X_j^2}{\|X\|^4}\big]$$

summing this from 1 to p yields that

$$\mathbf{E}_\theta\big[\frac{X^\top(X-\theta)}{\|X\|^2}\big] = \sum_{j=1}^p \mathbf{E}_\theta\big[\frac{1}{\|X\|^2}-\frac{2X_j^2}{\|X\|^4}\big] = (p-2)\mathbf{E}_\theta\big[\frac{1}{\|X\|^2}\big]\,.$$

Plugging this in the computation of the risk gives

$$R(\delta^{JS},\theta) = p-(p-2)^2\mathbf{E}_\theta[\frac{1}{\|X\|^2}] < p\,.$$

Indeed, the term $\mathbf{E}_\theta[1/\|X\|^2]$ is always positive, it can even be lower bounded in the following way, where ϕ is the density of the p-variate standard normal.

$$\mathbf{E}_\theta[1/\|X\|^2] = \int_{\mathbf{R}^p}\frac{1}{\|x\|^2}\phi(x-\theta)\mathrm{d}x \geq \int_{c1\leq\|x\|\leq c_2}\frac{1}{\|x\|^2}\phi(x-\theta)\mathrm{d}x \geq \frac{1}{c_2^2}\mathbf{P}_\theta(\|x\|\in[c_1,c_2]) > 0\,.$$

This direct computation can be helpful to analyze $R(\delta^{JS},\theta)$ when $\|\theta\|\to\infty$. □

Remark.

- Even though $R(\delta^{JS},\theta) < R(X,\theta)$, meaning that X is not admissible, they have the same maximal risk: indeed, when $\|\theta\|\to\infty$, $R(\delta^{JS},\theta)\to p$.

- While δ^{JS} is an improvement on X because it dominates it, it is itself not admissible, as it is dominated by δ^{JS+} defined by

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

  It can be shown that admissible estimators must be smooth in the observation. As a consequence, this is itself an inadmissible estimator.

- In practice, X can be much easier to work with, in particular to design tests or confidence regions, since its distribution is easily tractable, which is not the case of the James–Stein estimator.

4.5. *Classification problems.* A decision problem of great practical importance is that of classification. It is a type of regression problem, with a variable $X\sim P_X$ on a set $\mathcal{X}$, and a binary response $Y\in\{0,1\}$, with probabilities depending on the value of X. The joint distribution can therefore be written as $Q(x,y)$, with

$$X\sim P_X\,,\quad\text{and}\quad \mathbf{P}(Y=1\,|\,X=x) = \mathbf{E}[Y\,|\,X=x] = \eta(x)\,.$$

The objective in this problem is to predict the value of Y given X. This setting has various applications.

Example 4.7. In a football match between two teams, the information $x \in \mathbf{R}^p$ (number of goals scored by each of the players last year, age of the referee, etc.) can be used to predict the probability that the home team wins. If it is modelled as $\eta(x)$, we are in the model above.

Another equivalent way to formulate this problem is closer to hypothesis testing: A variable Y is drawn from $\{0,1\}$ with probability (π_0, π_1) and

$$X \,|\, Y = 0 \sim f_0(x) \quad \text{or} \quad X \,|\, Y = 1 \sim f_1(x)\,,$$

for two distributions assumed for now to be known. In practice, the distributions f_0 and f_1 are often unknown, and estimated based on samples.

Example 4.8. Based on a vector of medical observation $X = (X_1, \dots, X_p)^\top$ about a patient, we would like to determine if the patient has the flu (in which case $X \sim f_1$) or not (and $X \sim f_0$).

As in hypothesis testing, the decision rule of this problem only has two possible actions, predicting either the value 0 or the value 1.

Definition 4.10. A *classification rule* δ is a function $\delta : \mathcal{X} \to \{0,1\}$. It is equivalently defined by a region $\mathcal{R}$ such that

$$\delta(X) = \delta_{\mathcal{R}}(X) = \begin{cases} 1 & \text{if } x \in \mathcal{R}, \\ 0 & \text{if } x \in \mathcal{R}^c\,. \end{cases}$$

The probability of error of $\delta_{\mathcal{R}}$ is therefore given by two quantities:

- The probability of misclassifying $X \sim f_1$

$$\mathbf{P}(X \in \mathcal{R}^c \,|\, Y = 1) = \int_{\mathcal{R}^c} f_1(x)\mathrm{d}x = \mathbf{P}_1(X \in \mathcal{R}^c)\,.$$

- The probability of misclassifying $X \sim f_0$

$$\mathbf{P}(X \in \mathcal{R} \,|\, Y = 0) = \int_{\mathcal{R}} f_0(x)\mathrm{d}x = \mathbf{P}_0(X \in \mathcal{R})\,.$$

Together, they describe the whole risk function of $\delta_{\mathcal{R}}$. Under a prior $\pi = (\pi_0, \pi_1)$ the Bayes risk is therefore equal to

$$R_\pi(\delta_{\mathcal{R}}) = \pi_0 \mathbf{P}(X \in \mathcal{R} \,|\, Y = 0) + \pi_1 \mathbf{P}(X \in \mathcal{R}^c \,|\, Y = 1)\,.$$

Intuitively, good decision rules will correspond to regions where $X \in \mathcal{R}$ is more likely under f_1 than under f_0, and vice-versa.

Principles of Statistics **Part II - Michaelmas 2018**

Lecture 18: Classification problems

Lecturer: Quentin Berthet

REMARK. Consider the joint distribution Q on $\mathcal{X} \times \{0,1\}$ defined by

$$Q(x,y) = f(x,y)\pi(y)$$

where $\pi(1) = \pi_1$, $\pi(0) = \pi_0$, $f(x,1) = f_1(x)$ and $f(x,0) = f_0(x)$. This is the joint distribution of Y - the index of f_i - and of X, as introduced before. We recall that two interpretations of this problem are equivalent.

- The variable Y is drawn randomly from π, and X with distribution $f(x,y)$, conditionally on the value y of Y. This is the point of view given in the previous lecture.

- The variable X is drawn from its marginal distribution P_X given by

  $$P_X(x) = \sum_y Q(x,y) = \pi_0 f_0(x) + \pi_1 f_1(x)\,.$$

  The variable Y is drawn from a posterior distribution with

  $$\Pi(1|X=x) = \frac{\pi_1 f_1(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)}\,, \quad \text{and} \quad \Pi(0|X=x) = \frac{\pi_0 f_0(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)}\,.$$

  A common notation for $(\Pi(1|X=x), \Pi(0|X=x))$ is $(\eta(x), 1-\eta(x))$. It can be interpreted as a local probability for the value of the label.

Conceptually, the second point of view can give a better understanding in some other models, where the goal is to predict the value of the label, based on side information X.

PROPOSITION 4.8. *The classification error, or Bayes classification risk, satisfies*

$$R_\pi(\delta) = \mathbf{P}_Q(\delta(X) \neq Y)\,.$$

PROOF.

$$\begin{aligned}\mathbf{P}_Q(\delta(X) \neq Y) &= \mathbf{P}_Q(Y=1 \text{ and } \delta(X) \neq 1) + \mathbf{P}_Q(Y=0 \text{ and } \delta(X) \neq 0)\\ &= \pi_1 \mathbf{P}_1(X \notin \mathcal{R}^c) + \pi_0 \mathbf{P}_0(X \notin \mathcal{R}) = R_\pi(\delta)\,.\end{aligned}$$

Informal notes, based on past lecture notes by Richard Nickl. Please let me know of any errors.

This concludes the proof. To provide further explanation, the other formulation gives

$$\mathbf{P}_Q(\delta(X) \neq Y) = \mathbf{E}_Q[\mathbf{1}\{\delta(X) \neq Y\}] = \int_{\mathcal{X}} \Pi(\delta^c(x)|x)\mathrm{d}P_X(x)\,,$$

where $\delta^c = 1 - \delta$, the complement of δ in $\{0, 1\}$. □

These two formulations give the same intuition: in order to minimize the probability of error we should pick δ such that it is equal to 1 when $\eta(x)$ is large, 0 otherwise. This is equivalent to taking $\mathcal{R}$ such that the probabilities of $X \in \mathcal{R}$ under $\mathbf{P}_1$ and of $X \in \mathcal{R}^c$ under $\mathbf{P}_0$ are low.

Definition 4.11. For a prior $\pi = (\pi_0, \pi_1)$, with $\pi_1 \in (0, 1)$, the *Bayes classifier* is given by $\delta_\pi = \delta_{\mathcal{R}}$

$$\delta_{\mathcal{R}}(X) = \begin{cases} 1 & \text{if } x \in \mathcal{R}, \\ 0 & \text{if } x \in \mathcal{R}^c\,, \end{cases}$$

where

$$\mathcal{R} = \Big\{x \in \mathcal{X} \,:\, \frac{\pi_1 f_1(x)}{(1-\pi_1)f_0(x)} \geq 1\Big\}\,.$$

Proposition 4.9. *The Bayes classifier δ_π is the rule that minimizes the Bayes classification risk. If*

$$\mathbf{P}_i\Big(\frac{\pi_1 f_1(x)}{(1-\pi_1)f_0(x)} = 1\Big) = 0\,,$$

then the Bayes rule is unique.

Proof. Let $\mathcal{J} \subseteq \mathcal{X}$ be a classification region. The classification error associated to this region is

$$\begin{aligned} R_\pi(\delta_{\mathcal{J}}) &= \pi_1 \int_{\mathcal{J}^c} f_1(x)\mathrm{d}x + (1-\pi_1)\int_{\mathcal{J}} f_0(x)\mathrm{d}x \\ &= \int_{\mathcal{J}^c} [\pi_1 f_1(x) - (1-\pi_1)f_0(x)]\mathrm{d}x + \underbrace{(1-\pi_1)\int_{\mathcal{X}} f_0(x)\mathrm{d}x}_{\text{ind. of } \mathcal{J}}\,. \end{aligned}$$

The first term is minimized when the integrated term is nonpositive, i.e. when $\delta_{\mathcal{J}}$ is equal to the Bayes classification rule. It is a unique Bayes rule when the boundary has probability 0. □

Remark. Since a unique Bayes rule is admissible, the Bayes classifier is admissible in this case. Using this property can be useful to find a minimax classifier. For any $q \in (0, 1)$, let δ_q be the associated Bayes classifier for the prior $(q, 1-q)$ and $\mathcal{R}_q$ the corresponding classifying region. The risk consists of two values, the probabilities of error $\mathbf{P}(\mathcal{R}_q^c \,|\, 1)$ and $\mathbf{P}(\mathcal{R}_q \,|\, 0)$, so finding q such that

$$\mathbf{P}(\mathcal{R}_q^c \,|\, 1) = \mathbf{P}(\mathcal{R}_q \,|\, 0) = \text{const.}$$

we find a Bayes rule that has constant risk and is minimax.

Example 4.9. Consider the example of two normal distributions

$$X \sim f_0 = \mathcal{N}(\mu_0, \Sigma) \qquad \text{or} \qquad X \sim f_1 = \mathcal{N}(\mu_1, \Sigma)\,,$$

where $\mu_i \in \mathbf{R}^p$ and Σ is a $p \times p$ covariance matrix. One can show that any Bayes rule or minimax classifier depends on the data X only through the discriminant function

$$D(X) = X^\top \Sigma(\mu_1 - \mu_0)\,,$$

which is linear in X. This method is known as *Linear discriminant analysis.*

Remark. In practice, the two distributions and the prior are unknown, and so is the posterior η. In statistical learning theory, the objective is to "estimate" η from past observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, with $Y_i \in \{0, 1\}$. Since $\mathbf{P}_Q(\delta(X) \neq Y)$ cannot be directly minimized (as Q is unknown), we can minimize instead a sample version

$$\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{\delta(X_i) \neq Y_i\}\,.$$

In practice, the decision rules considered are parametrized of the form

$$\delta_\beta(x) = \mathbf{1}\{h_\beta(x) \geq 1/2\}\,,$$

where β is chosen to minimize the observed probability error above. It is equivalent to minimize

$$\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{\delta(X_i) \neq Y_i\} = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{|h_\beta(X_i) - Y_i| \geq 1/2\}$$

In order to simplify this optimization procedure, it is possible to consider a smooth version

$$\frac{1}{n}\sum_{i=1}^n \ell\big(h_\beta(X_i), Y_i\big)\,.$$

examples include $\ell(h, y) = (h - y)^2$ or $\log(1 + e^{hy})$. Recent progress in artificial intelligence and machine learning is sometimes based on these classical ideas.

Principles of Statistics **Part II - Michaelmas 2018**

Lecture 19: Multivariate analysis and PCA

Lecturer: Quentin Berthet

5. Further topics.

5.1. *Multivariate analysis & PCA.*

Definition 5.1. We recall that for two real-valued random variables X, Y, their covariance is defined as

$$\mathrm{Cov}(X,Y) = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])]$$

and their correlation is

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

Given observations $(X_1, Y_1), \dots, (X_n, Y_n)$ the *sample correlation coefficient* is

$$\hat{\rho}_{X,Y} = \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sqrt{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2 \cdot \frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y}_n)}}$$

By standard results of this course, if Var(X),Var(X) are positive and finite, all the sample versions of Var(X), Var(X), and Cov(X,Y) are consistent, and $\hat{\rho}_{X,Y}$ as well (proof is left as exercise).

Remark.

- If the model is $\mathcal{N}(\mu, \Sigma)$, this is equal to the MLE ρ_{MLE}.
- In a $\mathcal{N}(\mu, \Sigma)$ for $X = (X^{(1)}, \dots, X^{(p)})^\top$, $\mathrm{Cov}(X^{(i)}, X^{(j)}) = \Sigma_{ij}$ and if Σ is positive definite
$$\rho_{X^{(i)},X^{(j)}} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii} \cdot \Sigma_{jj}}}\,.$$
- The matrix $[\rho]_{ij} = \rho_{X^{(i)},X^{(j)}}$ has coefficients in $[-1,1]$, with diagonal coefficients equal to 1, and is semidefinite positive.

Informal notes, based on past lecture notes by Richard Nickl. Please let me know of any errors.

Proposition 5.1.

- *Given Σ a $p \times p$ semidefinite positive matrix, there exists $X \in \mathbf{R}^p$ such that $\mathbf{E}[X] = 0$ and $\Sigma = Cov(X)$.*

- *Given ρ a $p \times p$ semidefinite positive matrix with ones on the diagonal, there exists $X \in \mathbf{R}^p$ such that $\mathbf{E}[X] = 0$ and*

$$\rho_{X^{(i)},X^{(j)}} = \rho_{ij}$$

Remark.

- The proof is left to Examples sheet.

- These conditions are necessary and sufficient.

- These characterization are helpful when estimating such matrices: optimization problems on these sets will often be tractable, and it is practical to have an explicit description of "the set of covariance (or of correlation) matrices".

Proposition 5.2. *Under a $\mathcal{N}(0, I_p)$ model the distribution of $\hat{\rho}_{X,Y}$ is given by the following density*

$$f_{\hat{\rho}}(r) = \frac{\Gamma(\frac{1}{2}(n-1))}{\Gamma(\frac{1}{2}(n-2))}(1-r^2)^{\frac{1}{2}(n-4)}, \quad for\ -1 \leq r \leq 1\,.$$

Remark. It can be used to design hypothesis tests or confidence regions for ρ.

Proposition 5.3. *For a given random vector $X \in \mathbf{R}^p$, and $X \sim \mathcal{N}(\mu, \Sigma)$ such that*

$$X^\top = (\underbrace{\cdots\ X^{(1)\top} \cdots}_{\in \mathbf{R}^q}, \underbrace{\cdots\ X^{(2)\top} \cdots}_{\in \mathbf{R}^{p-q}})$$

with covariance matrix given by

$$\Sigma = \left(\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right)$$

the covariance of $X^{(1)}|X^{(2)}$ is given by

$$\Sigma_{11|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\,.$$

The partial correlation of $X^{(1)(i)}$,$X^{(1)(j)} for 1 \leq i, j \leq q$ is given by

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

It can also be estimated by plug-in MLE.

5.1.1. *Principal component analysis (PCA).* Principal component analysis is a common method in data analysis to reduce the dimension of a dataset, while trying to preserve as much information as possible. This is done by projecting the observations on directions with maximum variance.

From a mathematical statistics point of view, we want to recover information about the leading eigenspaces of the covariance matrix Σ.

For a random vector $X \in \mathbf{R}^p$, $\mathbf{E}[X] = 0$, and $\Sigma = \mathbf{E}[XX^\top]$, there is an orthogonal matrix V (i.e. $VV^\top = I_p$) such that

$$\Sigma = V\Lambda V^\top \,, \quad \text{where} \quad \Lambda = \operatorname{diag}\big(\{\lambda_i\}\big) \quad \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \,.$$

This can be rewritten as

$$\Sigma = \sum_{i=1}^{p} \lambda_i v_i v_i^\top \,.$$

If we define $U = V^\top X$, the vector of coefficients of X in the basis of the v_i's, then

$$\mathbf{E}[UU^\top] = \mathbf{E}[V^\top XX^\top V] = V^\top \mathbf{E}[XX^\top]V = V^\top \Sigma V = V^\top V\Lambda V^\top V = \Lambda \,.$$

Writing the random vector X in the basis described by V gives a vector of *uncorrelated* coefficients with variances λ_i. As a direct consequence, the following holds in the Gaussian case.

Proposition 5.4. *If $X \sim \mathcal{N}(0, \Sigma)$, then $U = V^\top X \sim \mathcal{N}(0, \Lambda)$, so the U_i are independent.*

Furthermore, if $Z \sim \mathcal{N}(0, \Lambda)$, $X = VZ$ in distribution

$$X = \sum_{i=1}^{p} \underbrace{v_i}_{\in \mathbf{R}^p,\ \text{determ.}} \underbrace{Z_i}_{\in \mathbf{R},\ \text{r.v.}} = \sum_{i=1}^{p} v_i \sqrt{\lambda_i}\, G_i \,,$$

where the G_i are i.i.d. $\mathcal{N}(0,1)$ (or $G \sim \mathcal{N}(0, I_p)$).

This representation gives an interpretation of principal component analysis. The vector X is the sum of orthogonal effects, along vector v_i's, with different amplitudes λ_i. The main directions, or *principal components* are those with greatest variance.

Principles of Statistics **Part II - Michaelmas 2018**

Lecture 20: Resampling principles & the bootstrap

Lecturer: Quentin Berthet

5.2. *Resampling principles & the bootstrap.* One of the informal intuition given by our analysis of statistics thus far is that the greater the *sample size* n is, the more information about our model we have, and the easier most statistical tasks get. The idea between resampling techniques is that it is sometimes possible to reuse some information from the sample by drawing again partially from the sample. The first case is given as an illustrating example.

Let $T_n = T(X_1, \dots, X_n)$ be an estimator of a parameter θ, with bias $B(\theta) = \mathbf{E}_\theta[T_n] - \theta$. If the estimator is biased, it is possible to achieve a bias reduction as follows

Definition 5.2. Let $T_{(-i)} = T(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$ be the estimator with the i-th observation removed. The *jacknife bias estimate* is defined as

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

and the jacknife bias corrected estimate of θ is

$$\tilde{T} = \tilde{T}_{JACK} = T_n - \hat{B}_n \,.$$

Proposition 5.5. *For regular bias functions $B(\theta)$, it holds that*

$$|\mathbf{E}[\tilde{T}_{JACK}] - \theta| = O\Big(\frac{1}{n^2}\Big)\,,$$

see Examples sheet.

Remark. A bias reduction does not necessarily imply a smaller risk, but the resampling idea can be generalized to statistical inference.

Remark (Motivation). In the statistical problem of estimating the mean of a distribution P based on $X_1, \dots, X_n \sim P$ i.i.d., where $\mathbf{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$, in order to build the asymptotic confidence interval

$$C_n = \{\nu \in \mathbf{R} \,:\, |\nu - \bar{X}_n| \leq \sigma z_\alpha/\sqrt{n}\}$$

Informal notes, based on past lecture notes by Richard Nickl. Please let me know of any errors.

one needs to know the variance σ^2, or to replace it by an estimate $\hat{\sigma}^2$. In order to show that it has the proper asymptotic probability, one also needs to know the limiting distribution $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d.} \mathcal{N}(0,\sigma^2)$. Instead of relying these principles, we can rely on the *bootstrap*, by reusing information from the sample.

Definition 5.3. For fixed observations $X_1,\dots,X_n$, we define a discrete probability distribution $\mathbf{P}_n = \mathbf{P}_n(\cdot \mid X_1,\dots,X_n)$ that generates $(X^b_{n,i} : 1 \le i \le n)$, n indendent copies of X^b_n with law

$$\mathbf{P}_n(X^b_n = X_i) = \frac{1}{n} \quad \text{for} \quad 1 \le i \le n\,.$$

In other words, we sample n values uniformly at random, independently (with replacement) from the existing X_i.

Proposition 5.6.

$$\mathbf{E}[X^b_n] = \sum_{i=1}^n X_i\,\mathbf{P}(X^b_n = X_i) = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}_n\,.$$

The $X^b_{n,i}$ are therefore drawn from a population with mean $\bar{X}_n$ (conditionally on the X_i's) and the *bootstrap sample mean* $\bar{X}^b_n = \frac{1}{n}\sum_{i=1}^n X^b_{n,i}$ "estimates" $\bar{X}_n$.

Remark. In order to build a valid confidence interval, it is important to understand, at least approximately, the distribution of $\bar{X}_n - \mu$. The main idea behind the bootstrap principle is that the distribution of

$$\bar{X}^b_n - \bar{X}_n = \bar{X}^b_n - \mathbf{E}[\bar{X}^b_n]$$

is approximately the same as the one of

$$\bar{X}_n - \mu = \bar{X}_n - \mathbf{E}[\bar{X}_n]\,.$$

This supports the following construction.

Definition 5.4. Let $R^b_n = R^b_n(X_1,\dots,X_n)$ satisfy

$$\mathbf{P}_n\big(|\bar{X}^b_n - \bar{X}_n| \le R^b_n/\sqrt{n} \mid X_1,\dots,X_n\big) = 1-\alpha$$

The *bootstrap confidence set* $\mathcal{C}^b_n$ is defined as

$$\mathcal{C}^b_n = \{\nu \in \mathbf{R} \,:\, |\nu - \bar{X}_n| \le R^b_n/\sqrt{n}\}\,.$$

Remark. This confidence set can be computed without estimating σ or even knowing the asymptotic distribution of $\bar{X}_n$. On the other hand, for fixed $X_1,\dots,X_n$, the distribution $\mathbf{P}_n$ is known, and its quantiles can be determined exactly, or approximated from simulations. Theoretically, it is not technically necessary to actually create the bootstrap sample, but just to use it as a thought experiment to compute the root R^b_n.

Remark. We note that $\mathbf{P}_n(\bar{X}_n^b \in \mathcal{C}_n^b) = 1 - \alpha$ by definition. Our objective is to show that $P(\mu \in \mathcal{C}_n^b) \to 1 - \alpha$ as $n \to \infty$, to show that it is a proper frequentist confidence interval. This follows from the following theorem.

Theorem 5.1. *Let $X_1, \ldots, X_n$ be drawn i.i.d. from P with mean μ and finite variance σ^2. We have, as $n \to \infty$*

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

where Φ is the c.d.f. of a $\mathcal{N}(0, \sigma^2)$ distribution.

The proof of this theorem is the subject of the following lecture.

Principles of Statistics **Part II - Michaelmas 2018**

Lecture 21: Validity of the bootstrap

Lecturer: Quentin Berthet

In the previous lecture, we stated that the c.d.f. of $\sqrt{n}(\bar{X}_n^b - \bar{X}_n)$ converges uniformly to that of a $\mathcal{N}(0, \sigma^2)$. As in the case of the Bernstein–von Mises theorem, this can be used to show that $P(\mu \in \mathcal{C}_n^b) \to 1 - \alpha$. The idea of the proof given in this lecture is as follows.

For a fixed infinite sequence $X_1, X_2, \ldots$ (equivalent to a fixed ω in the probability space of sequences from P), we have a sequence of distributions $\mathbf{P}_n$ conditioned on $X_1, \ldots, X_n$

- We show that for almost all ω, for a sequence of random variables A_n with distributions Q_n, we have a CLT-type result as $n \to \infty$

$$A_n \xrightarrow{d.} \mathcal{N}(0, \sigma^2)\,.$$

- We show that for these ω, this implies the uniform convergence of the c.d.f. to that of $\mathcal{N}(0, \sigma^2)$.

- This implies the desired result. There are "two layers" of randomness here: the infinite sequence of $(X_i)_{i\geq 1}$ and the drawing of the bootstrap sample under $\mathbf{P}_n$, conditionally on the infinite sequence. What we show here is that for some $(X_i)_{i\geq 1}$ satisfying some properties, the resulting sequence of $\mathbf{P}_n$ will satisfy the uniform convergence of the c.d.f. (this is a deterministic result). Furthermore, almost all the sequences satisfy these properties, implying the almost sure convergence.
  In order to show a result on sequence of random $\mathbf{P}_n$, we show that it holds for all ω satisfying some properties, and that almost all

Lemma 5.1. *If $A_n \sim f_n$ with c.d.f. F_n and $A \sim f$ with continuous c.d.f. F. We have, as $n \to \infty$*

$$A_n \sim f_n \xrightarrow{d.} A \sim f \quad \Rightarrow \quad \sup_{t\in\mathbf{R}} |F_n(t) - F(t)| \to 0\,.$$

Proof. Since F is monotonous and continuous, for all integer k and $1 \leq i \leq k-1$, there exists $x_i \in \mathbf{R}$ such that $F(x_i) = i/k$, and $-\infty = x_0 < x_1 < \ldots < x_k = +\infty$.

Informal notes, based on past lecture notes by Richard Nickl. Please let me know of any errors.

As a consequence, for all $x \in [x_{i-1}, x_i)$, we have

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

Taking k large enough such that $1/k < \varepsilon/2$ and $n \geq N_k$, by convergence in distribution $|F_n(x_i) - F(x_i)| < \varepsilon/2$ For all $0 \leq i \leq k$. It therefore holds that

$$\sup_{x \in \mathbf{R}} |F_n(x) - F(x)| \leq \varepsilon/2 + 1/k < \varepsilon\,.$$

□

Definition 5.5 (i.i.d. triangular arrays). The sequence $(Z_{n,i}\,;\, 1 \leq i \leq n)_{n \geq 1}$ is a *triangular array* of i.i.d. random variables if: for all $n \geq 1$, $Z_{n,1}, \ldots, Z_{n,n}$ are n i.i.d. draws from a distribution Q_n.

Remark. The name comes from writing each of the sequences $Z_{n,1}, \ldots, Z_{n,n}$ on a row in order, following matrix notation. On *each* row, the variables are i.i.d. but no assumption is made *across* rows: the distribution can change at each line.

Proposition 5.7 (CLT for triangular arrays). *Let $(Z_{n,i}\,;\, 1 \leq i \leq n)_{n \geq 1}$ be a triangular array of i.i.d. random variables all with finite variance such that Var$(Z_{n,i}) = \sigma_n^2 \to \sigma^2$ as $n \to \infty$. Then, under assumptions (1-3), we have*

$$\sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^{n} Z_{n,i} - \mathbf{E}_n[Z_{n,i}]\Big) \xrightarrow{d.} \mathcal{N}(0, \sigma^2)\,.$$

Assumptions:

1) For all $\delta > 0$, $nP_n(|Z_{n,1} > \delta\sqrt{n}) \to 0$ as $n \to \infty$.

2) Var$(Z_{n,1}\mathbf{1}\{|Z_{n,1} \leq \sqrt{n}\}) \to \sigma^2$ as $n \to \infty$.

3) $\sqrt{n}\,\mathbf{E}[Z_{n,1}\mathbf{1}\{|Z_{n,1} > \sqrt{n}\}] \to 0$.

This result, as well as its assumptions, are not examinable, and the proof is not shown here. It is simply used in the proof of Theorem 5.1.

Proof of Theorem 5.1. Fix $(X_i)_{i \geq 1}$ (equivalent to fixing ω in the original probability space).

$*$ Under the distribution $\mathbf{P}_n(\cdot \mid X_1, \ldots, X_n)$, with $Z_{n,i} = X_i^{b(n)}$, and $\mathbf{E}_n[Z_{n,i}] = \mathbf{E}[X_i^{b(n)}] = \bar{X}_n$. The sequence $(X_i^{b(n)}\,;\, 1 \leq i \leq n)_{n \geq 1}$ is a triangular array of i.i.d. variables. We have that

$$\begin{aligned}\mathrm{Var}(X_i^{b(n)}) &= \mathbf{E}_n[(X_i^{b(n)})^2] - \mathbf{E}_n[X_i^{b(n)}]^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big)^2 \\ &= \sigma_n^2\,.\end{aligned}$$

$*$ For ω such that $\sigma_n^2 \to \sigma^2$ (and for which the assumptions 1-3 are satisfied), we have

$$\sqrt{n}(\bar{X}_n^{b(n)} - \underbrace{\bar{X}_n}_{\mathbf{E}[X_n^{b(n)}]}) \xrightarrow{d.} \mathcal{N}(0, \sigma^2)\,.$$

$*$ By Lemma 5.1, this that implies for such ω

$$\sup_{t \in \mathbf{R}} |\mathbf{P}_n\big(\sqrt{n}(\bar{X}_n^b - \bar{X}_n) \leq t | X_1, \ldots, X_n\big) - \Phi(t)| \to 0\,,$$

$*$ By the law of large numbers, $\sigma_n^2 \to \sigma^2$ for almost all ω, the assumptions 1-3 being also satisfied, we obtain the final result. □

Remark. This shows the validity of the bootstrap method for confidence intervals for the mean. In a general estimation, this method can be extended in the following manner.

In a parametric model $\{P_\theta : \theta \in \Theta\}$, the ideas can be extended in at least two ways.

1) One resamples $(X_{n,i}^b : 1 \leq i \leq n)$from $\mathbf{P}_n$ and computes the MLE $\hat{\theta}_n^b$ based on the bootstrap sample. Akin to using $\bar{X}_n^b - \bar{X}_n$ as a proxy for $\bar{X}_n - \mathbf{E}[X]$, we can use $\sqrt{n}(\hat{\theta}_n^b - \hat{\theta}_n)$ as a pivot for $\sqrt{n}(\hat{\theta}_n - \theta_0)$, around the MLE $\hat{\theta}_n$ and find roots R_n such that

$$\mathbf{P}_n\Big(\|\hat{\theta}_n^b - \hat{\theta}_n\| \leq \frac{R_n}{\sqrt{n}} \,|\, X_1, \ldots, X_n\Big) = 1 - \alpha\,.$$

It is possible to show that the bootstrap confidence set

$$\mathcal{C}_n^b = \Big\{\theta : \|\hat{\theta}_n^b - \hat{\theta}_n\| \leq \frac{R_n}{\sqrt{n}}\Big\}$$

satisfies $P_{\theta_0}(\theta_0 \in \mathcal{C}_n^b)$. It is not necessary to estimate the Fisher information or to know the asymptotic distribution of $\hat{\theta}_n$. This is known as the *nonparametric bootstrap* as we resample directly from the X_i without using the parametric model.

2) In contrast, in the *parametric bootstrap*, one samples directly from $P_{\hat{\theta}_n}$, using the MLE $\hat{\theta}_n$, and uses the same ideas as above.

Principles of Statistics **Part II - Michaelmas 2018**

Lecture 22: Monte-Carlo methods

Lecturer: Quentin Berthet

5.3. *Monte-Carlo methods.* To apply in practice many statistical techniques studied in this course, one often needs access to quantities related to a fixed, known distribution. Examples include the mean of a posterior distribution, level sets of a multivariate Gaussian distribution, quantiles of a bootstrap resampling distribution, etc. Often, there is no explicit, closed-form formula for these quantities, and one must use simulation techniques in order to approximate them.

Definition 5.6. A pseudo-random uniform sample is a a collection of variables $U_1^*, \ldots, U_N^*$ such that for all $u_1, \ldots, u_N \in [0,1]$

$$\mathbf{P}(U_1^* \leq u_1, \ldots, U_N^* \leq u_N) \approx \mathbf{P}(U_1 \leq u_1, \ldots, U_N \leq u_N) = u_1 \ldots u_N \,.$$

Remark. The sign "≈" means that the equality holds up to machine precision. For the purposes of this course, one can consider that this is an "=" sign, and that $U_1^*, \ldots, U_N^* \sim U[0,1]$ i.i.d. This can be used as a starting point to generate samples from other distributions

Proposition 5.8. *Let $U_1, \ldots, U_N \overset{i.i.d.}{\sim} U[0,1]$ and X_i be defined, for $1 \leq i \leq N$ as*

$$X_i = \sum_{j=1}^n x_j \mathbf{1}\Big\{U_i \in \Big(\frac{j-1}{n}, \frac{j}{n}\Big]\Big\}\,.$$

Then, the X_i are i.i.d. uniform over the set $\{x_1, \ldots, x_n\}$.

Proof. The independence of the X_i stems directly from the independence of the U_i. One can check directly that for all i, j

$$\mathbf{P}(X_i = x_j) = \mathbf{P}\Big(U_i \in \Big(\frac{j-1}{n}, \frac{j}{n}\Big]\Big) = \frac{1}{n}\,.$$

□

Remark. This proposition shows how one can generate i.i.d. samples from a uniform distribution over any finite set. In particular this can be used to generate a bootstrap sample, by taking the x_i equal to the initial sample of size n. We note that this

Informal notes, based on past lecture notes by Richard Nickl. Please let me know of any errors.

result can easily be generalized to any discrete distribution, by taking segments of different lengths. It can also be generalized to continuous distributions, as shown in the following.

Definition 5.7. Let F be the c.d.f of a distribution on the reals. The *generalized inverse* F^- of F is defined for $u \in (0,1)$ as

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

Remark. For continuous c.d.fs, this is equivalent to the functional inverse, and can be seen as a "quantile function": $F^-(u)$ gives the u-th quantile of the distribution. It is used to generate variables with any given distribution, based on uniform random variables.

Proposition 5.9. *For any distribution P with c.d.f F and generalized inverse F^-, and U uniform over $[0,1]$, we have that*

$$X = F^-(U) \sim P\,.$$

Proof. By definition of F^-, we have

$$P(X \leq t) = P(F^-(U) \leq t) = P(U \leq F(t)) = F(t)\,.$$

Since F is the c.d.f of P, we have $X \sim P$. □

Note that for a pseudo-random sample $(U_1^*, \ldots, U_N^*)$, this implies that $(X_1^*, \ldots, X_N^*) = (F^-(U_1^*), \ldots, F^-(U_N^*))$ is an i.i.d. sample from P. This can be used to approximate integrals or expectations, as a direct corollary of the law of large numbers, since

$$\frac{1}{N}\sum_{i=1}^N g(X_i^*) \xrightarrow{a.s.} \mathbf{E}_P[g(X)]\,.$$

In certain situations, it is not possible to compute explicitely F^- (in a $\mathcal{N}(\mu,\sigma^2)$ model for instance), and one must resort to other methods.

5.3.1. *Importance sampling.* Let P have density f, from which it is hard to simulate samples, and h be a density whose support includes that of f from which it is easier to simulate samples. It holds that

$$\mathbf{E}_h[\frac{g(X)}{h(X)}f(X)] = \int_{\mathcal{X}} \frac{g(x)}{h(x)} f(x)h(x)\mathrm{d}x = \int_{\mathcal{X}} g(x)f(x)\mathrm{d}x = \mathbf{E}_f[g(X)]\,.$$

As a consequence, we have, for $(X_1^*, \ldots, X_N^*)$ generated from the distribution with density h

$$\frac{1}{N}\sum_{i=1}^N \frac{g(X_i^*)}{h(X_i^*)} f(X_i^*) \xrightarrow{a.s.} \mathbf{E}_f[g(X)]\,.$$

5.3.2. *Accept/reject algorithm.* In a similar setup, with densities f and h, satisfying $f \leq Mh$, for some $M > 0$.

- Step 1 generate $X \sim h$ and $U \sim U(0,1)$.
- Step 2 If $U \leq f(X)/(Mh(X))$, take $Y = X$
  else, return to Step 1.

One show that Y has distribution with density f. The time taken to generate a sample is however random, and depends on M.

5.3.3. *Gibbs sampler.* When dealing with joint distributions, in particular when dealing with posterior distributions of a multivariate parameter θ that is hard to sample from, it is possible to use the *Gibbs sampler* to generate approximate samples. In the bivariate case (X, Y), one starts at some $X_0 = x$ and applies the steps for $t \geq 1$

- $Y_t \sim f_{Y|X}(\cdot|X = X_{t-1})$,
- $X_t \sim f_{X|Y}(\cdot|Y = Y_t)$.

This algorithm generates a sequence $(X_t, Y_t)_{1 \leq t \leq N}$. One shows that (X_t, Y_t), X_t, Y_t, are all Markov processes with respectively invariant distributions $f_{X,Y}, f_X, f_Y$. As a consequence of the ergodic theorem, when $N \to \infty$, we have

$$\frac{1}{N} \sum_{t=1}^{N} g(X_t, Y_t) \xrightarrow{a.s.} \mathbf{E}_{(X,Y)}[g(X,Y)]\,.$$

This can be used to compute posterior means, and can be generalized to larger number of variables, e.g. for a parameter vector $(\theta_1, \ldots, \theta_d)$ by cycling through the coefficients.

Principles of Statistics **Part II - Michaelmas 2018**

Lecture 23: Introduction to nonparametric statistics

Lecturer: Quentin Berthet

5.4. *Nonparametric methods.* Informally, if the goal of statistical inference is to gain information on a distribution based on samples from it, it is done in this course mainly through a *parameter*. For a parameter set Θ, usually a subset of $\mathbf{R}^d$, the distribution is equal to P_θ, for some $\theta \in \Theta$. Finding the distribution is therefore reduced to finding the unknown parameter, or an approximate value of it. However, it is possible to directly estimate the distribution, without a parametric model. Formally, the goal is to estimate the c.d.f. F of a distribution P on the reals, based on i.i.d. samples $X_1, \dots, X_n$ from it. We recall that for all $t \in \mathbf{R}$

$$F(t) = \mathbf{P}(X \leq t) = \mathbf{E}[\mathbf{1}_{(-\infty,t]}(x)]\,.$$

Definition 5.8. The *empirical distribution function* F_n of a sample $X_1, \dots, X_n$ is given, for all $t \in \mathbf{R}$, by

$$F_n(t) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{(-\infty,t]}(X_i)] = \frac{\#\{i\,:X_i \leq t\}}{n}$$

The law of large numbers guarantees, for any given t, that $F_n(t)$ almost surely of to $F(t)$. It is possible to give a stronger result, about *uniform convergence*

Theorem 5.2 (Glivenko-Cantelli). *For any F, we have as $n \to \infty$, almost surely*

$$\sup_{t \in \mathbf{R}} |F_n(t) - F(t)| \to 0\,.$$

It is a type of uniform law of large numbers. It is further possible to improve this result in a uniform version of a central limit theorem, that involves the notion of *Brownian bridge*

Definition 5.9 (Brownian motion). A Brownian motion (or Wiener process) is a continuous process $(W_t)_{t\geq 0}$ satisfying

- $W_0 = 0$

- $W_t - W_s \sim \mathcal{N}(0, t-s)$ for $s < t$, independently of $(W_{s'})_{s \leq s'}$.

Informal notes, based on past lecture notes by Richard Nickl. Please let me know of any errors.

Remark. A more formal description, and proof of existence of this process can be found in measure theory or probability courses. These are the properties needed here. Informally, it can be thought of as the limit of a random walk with independent steps, when the time steps goes to 0.

Definition 5.10 (Brownian bridge). A Brownian bridge is a continuous process $(B_t)_{0\leq t\leq 1}$ equal to a Brownian motion conditioned on $B_1 = 0$. It satisfies

- $B_0 = B_1 = 0$

- $B_t \sim \mathcal{N}(0, t(1-t))$, and $Cov(B_s, B_t) = s(1-t)$ for $s \leq t$.

It can be constructed from a Brownian motion W_t by taking $B_t = W_t - tW_1$.

Theorem 5.3 (Donsker-Kolmogorov-Doob). *As $n \to \infty$, we have that*

$$\sqrt{n}(F_n - F) \xrightarrow{d.} \mathcal{G}_F\,,$$

where $\mathcal{G}_F$ is a Gaussian process such that $\mathcal{G}_F(t) = B_{F(t)}$, that satisfies

$$Cov(\mathcal{G}_F(s), \mathcal{G}_F(t)) = F(s)(1 - F(t))\,.$$

Remark. If $X_1, \ldots, X_n \sim U[0,1]$ and $F(t) = t$, then $\mathcal{G}_F(t) = B_t$ and $\sqrt{n}(F_n - F)$ converges to a standard Brownian bridge: its values are close to 0 for t close to 0 and 1, and the deviations have larger amplitude towards the middle of the segment.

For general F, the supremum of the deviations is independent F, since $\mathcal{G}_F$ is a just a reparametrization of a Brownian bridge, as given in the following

Theorem 5.4 (Kolmogorov-Smirnov). *As $n \to \infty$, we have that*

$$\sqrt{n}\|F_n - F\|_\infty \xrightarrow{d.} \|\mathcal{G}_F\|_\infty = \sup_{t\in[0,1]} |B_t|\,.$$

Remark. We note again that the limit distribution is independent of F, which is of key importance in applications of this result

To test the null $H_0 : F = F_0$ against the alternative $H_1 : F \neq F_0$, the statistic $\sqrt{n}\|F_n - F\|_\infty$ can be compared to the quantiles of $\|B\|_\infty$, for which there are tables.

To construct a uniform confidence band $\mathcal{C}_n$ for F, where for all $x \in \mathbf{R}$, $\mathcal{C}_n(x)$ is centered on $F_n(x)$ with amplitude given by the quantiles of $\|B\|_\infty$ such that

$$\mathbf{P}(F(x) \in \mathcal{C}_n(x) \forall x \in R) \to 1 - \alpha\,.$$

❍