MATHEMATICAL STATISTICS — Math21a, O. Knill

AIM. We illustrate how multivariable calculus appears in estimation theory. Estimation theory is a branch of mathematical statistics, where the aim is to estimate continuous or discrete parameters optimally. This leads to extremization problems which can be understood geometrically. For example, in order to know how good we can estimate things, we geometrically estimate an angle in terms of "lengths": the **Fisher information** of the random variables and the **quadratic estimation error** of the estimator. This is a working document. The aim is to make the material accessible to students passing the probability flavoured multivariable calculus track at Harvard.

STATISTICAL MODEL. A collection $(\Omega, \mathcal{A}, P_\theta)$ of probability spaces is called a **statitical model**. If X is a random variable, its expectation with respect to the measure P_θ is denoted by $\mathrm{E}_\theta[X]$. If X is continuous, then its probability density f_θ depends on θ. The parameters θ are taken from a **parameter space** Θ, which is part of the real line. The theory could be extended to the situation, where Θ is a subset of k dimensional space, but it is more convenient to look in this case at one variable estimators $g(\theta)$. A probability distribution $\mu = p(\theta)d\theta$ on Θ is called an **a priori distribution** on Θ. It allows to define the **global expectation** $\mathrm{E}[X] = \int_\Theta E_\theta[X] d\mu(\theta)$.

ESTIMATOR. Given independent and identically distributed random variables $X_1, ..., X_n$ on a probability space $(\Omega, \mathcal{A}, P_\theta)$, where $\theta \in \Delta$ is a **parameter**, we want to estimate the **quantity** $g(\theta)$ using an **estimator**

$$\boxed{T(\omega) = t(X_1(\omega), ..., X_n(\omega))}$$

Often $\theta \in R$ and $g(\theta) = \theta$. Other examples are when θ is a parameter vector and $g(\theta)$ a real-valued function.

The **expectation** and **variance** of a random variable X on $(\Omega, \mathcal{A}, P_\theta)$ is denoted by $\mathrm{E}_\theta[X]$ and $\mathrm{Var}_\theta[X]$. If X has a density f_θ, one has $E_\theta[X] = \int_\Omega f_\theta(x)\, dx$ and $\mathrm{Var}_\theta[X] = E_\theta[(X - E_\theta[X])^2] = \int_\Omega f_\theta^2 dx - (\int f_\theta\, dx)^2$.

ESTIMATE EXPECTATION. If $\theta = \mathrm{E}[X_i]$ is the expectation of the n random variables:

- $T = \frac{1}{n}\sum_{j=1}^n X_i$, the **arithmetic mean**.
- $T =$ **median**, defined as $X_{(n)}$ $X_{(1)} \leq ... \leq X_{(2m+1)}$ is an ordered list of the random variables if n is odd. $(X_{m+1} + X_m)/2$ if $X_{(1)} \leq ... \leq X_{(2m)}$ is the ordered list of the random variables if n is even.

a) $f(x) = \sum_{i=1}^n (x_i - x)^2$ is minimized by the **arithmetic mean**.
b) $f(x) = \sum_{i=1}^n |x_i - x|$ is minimized by the **median**.

Proof. a) $f'(x) = 0$ at $\sum_{i=1}^n (x_i - x) = 0$. To see b), note that $|a - x| + |b - x| = |b - a| + C(x)$, where $C(x)$ is zero if $a \leq x \leq b$ and $C(x) = x - b$ if $x > b$ and $D(x) = a - x$ if $x < a$. If $n = 2m + 1$ is odd, we have $f(x) = \sum_{j=1}^m |x_i - x_{n+1-i}| + \sum_{x_j > x_m} C(x_j) + \sum_{x_j < x_m} D(x_j)$ which is minimized for $x = x_m$. If $n = 2m$, we have $f(x) = \sum_{j=1}^m |x_i - x_{n+1-i}| + \sum_{x_j > x_{m+1}} C(x_j) + \sum_{x_j < x_{m-1}} D(x_j)$ which is minimized for $x \in [x_m, x_{m+1}]$.

BIAS. Define the **bias** of an estimator as $\boxed{B(\theta) = B_\theta[T] = \mathrm{E}_\theta[T] - g(\theta)}$ The bias is also called the **systematic error**. If the bias is zero, the estimator is called **unbiased**. In the case of an apriori distribution on Θ, one can define the **global error** $B(T) = \int_\Theta B(\theta)\, d\mu(\theta)$.

LINEAR ESTIMATORS FOR EXPECTATION. $g(\theta) = \mathrm{E}_\theta[X_i]$.

A linear estimator $T = \sum_{j=1}^n \alpha_i X_i$ with $\sum_i \alpha_i = 1$ is unbiased for the estimate $g(\theta) = E_\theta[X_i]$.

Proof. $\mathrm{E}_\theta[T] = \sum_{j=1}^n \alpha_i \mathrm{E}_\theta[X_i] = \mathrm{E}_\theta[X_i]$.

ESTIMATOR FOR VARIANCE. $g(\theta) = \mathrm{Var}_\theta[X_i]$.

a) If the mean m is known, then $T = \frac{1}{n}\sum_{j=1}^n (X_i - m)^2$ is unbiased.
b) With unknown mean, $T = \frac{1}{n-1}\sum_i (X_i - \overline{X}_i)^2$ is unbiased, where $\overline{X} = \frac{1}{n} X_i/n$.

Proof. a) $E_\theta[T] = \frac{1}{n}\sum_{j=1}^n (X_i - m)^2 = \mathrm{Var}_\theta[T] = g(\theta)$.
b) If we try with $T = \frac{1}{n}\sum_i (X_i - \overline{X}_i)^2$, we get $\mathrm{E}_\theta[T] = \mathrm{E}_\theta[X_i^2] - \mathrm{E}_\theta[\frac{1}{n^2}\sum_{i,j} X_i X_j] = \mathrm{E}_\theta[X_i^2] - \frac{1}{n}\mathrm{E}_\theta[X_i^2] - \frac{n(n-1)}{n^2}\mathrm{E}_\theta[X_i]^2 = (1 - 1/n)\mathrm{E}_\theta[X_i^2] - (n-1)/n\mathrm{E}_\theta[X_i]^2 = \frac{n-1}{n}\mathrm{Var}_\theta[X_i]$. Therefore $n/(n-1)T$ is the correct unbiased estimate.

RISK FUNCTION. The expectation of the quadratic estimation error

$$\boxed{\mathrm{Err}_\theta[T] = \mathrm{E}_\theta[(T - g(\theta))^2]}$$

is called the **risk function** or the **mean square error** of the estimator T. It measures the estimator performance.

REMARK. $\mathrm{Err}_\theta[T] = \mathrm{Var}_\theta[T] + B_\theta[T]$.
EXAMPLE. If T is unbiased, then $\mathrm{Err}_\theta[T] = \mathrm{Var}_\theta[T]$.

EXAMPLE. The arithmetic mean is the "best linear unbiased estimator".

Proof. With $T = \sum_i \alpha_i X_i$, where $\sum_i \alpha_i = 1$ the risk function is $\mathrm{Err}_\theta[T] = \mathrm{Var}_\theta[T] = \sum_i \alpha_i^2 \mathrm{Var}_\theta[X_i]$ which is by Lagrange minimal for $\alpha_i = 1/n$.

MAXIMUM LIKELIHOOD FUNCTION. The **maximum likelyhood function** $t(x_1, ..., x_n)$ is defined as the maximum of

$$\boxed{\theta \mapsto L_\theta(x_1, ..., x_n) = f_\theta(x_1) \cdot ... \cdot f_\theta(x_n)}$$

The **maximum likelyhood estimator** is the random variable $T(\omega) = t(X_1(\omega), ..., X_n(\omega))$. For discrete random variables, $L_\theta(x_1, ..., x_n)$ would be replaced by $P_\theta[X_1 = x_1, ..., X_n = x_n]$.
One also looks at the **maximum a posteriori** estimator, which is the maximum of

$$\boxed{\theta \mapsto L_\theta(x_1, ..., x_n) = f_\theta(x_1) \cdot ... \cdot f_\theta(x_n) p(\theta)}$$

where $p(\theta)\, d\theta$ was the a priori distribution on Θ.

MINIMAX PRINCIPLE. Find $\min_T \max_\theta R(\theta, T)$ to find the worst case.

BAYES PRINCIPLE. Find $\min_T \int_\Theta (R(\theta, T)\, d\mu(\theta)$.

EXAMPLES.

1) $f_\theta(x) = \frac{1}{2}e^{-|x-\theta|}$. The maximum likelihood function $L_\theta(x_1, ..., x_n) = \frac{1}{2^n} e^{-\sum_j |x_i - \theta|}$ is maximal when $\sum_j |x_i - \theta|$ is minimal which means that $t(x_1, ..., x_n)$ is the median of the data $x_1, ..., x_n$.

2) $f_\theta(x) = \theta^x e^{-\theta}/x!$ Poisson distribution. The maximal likelihood function $l_\theta(x_1, ..., x_n) = e^{\sum_i \log(\theta) x_i - n\theta}/(x_1!...x_n!)$ is maximal for $\theta = \sum_i x_i/n$.

3) The maximum likelihood estimator for $\theta = (m, \sigma^2)$ for Gaussian distributed random variables $f_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-m)^2}{2\sigma^2}}$ has the maximum likelyhood function maximized for $t(x_1, ..., x_n) = (\frac{1}{n}\sum_i x_i, \frac{1}{n}\sum_i (x_i - \overline{x})^2)$.

FISHER INFORMATION. Define the **Fisher information** of a random variable X with density f_θ

$$\boxed{I(\theta) = \int (\frac{f_\theta'(x)}{f_\theta(x)})^2 f_\theta(x)\, dx.}$$

Remark. In the multiparameter case, one defines the **Fisher information matrix** $I_{ij}(\theta) = \int \frac{f_{\theta_i} f_{\theta_j}}{f_\theta^2} f_\theta\, dx$.

LEMMA. $I(\theta) = E_\theta[(\frac{f_\theta'}{f_\theta})^2(X)] = \mathrm{Var}_\theta[\frac{f_\theta'}{f_\theta}]$.

Proof. This follows from $\mathrm{E}[\frac{f_\theta'}{f_\theta}(X_1)] = \int f_\theta' dx = 0$.

LEMMA. $I(\theta) = -\mathrm{E}_\theta[(\log(f_\theta)'']$.

Proof. $\mathrm{E}[\log(f_\theta)''] = \int \log(f_\theta)'' f_\theta\ dx = -\int \log(f_\theta)' f_\theta'\, dx = -\int (f_\theta'/f_\theta)^2 f_\theta\, dx$.

THE SCORE FUNCTION is defined as the logarithmic derivative $\rho_\theta = f_\theta'/f_\theta$. One has $I(\theta) = \mathrm{E}_\theta[\rho_\theta^2] = \mathrm{Var}_\theta[\rho_\theta]$.

EXAMPLE. If X is Gaussian, the score function $\rho_\theta = f'(\theta)/f(\theta) = -(x - m)/(\sigma^2)$ is linear and has variance 1. The Fisher information I is $1/\sigma^2$. We see that $\mathrm{Var}[X] = 1/I$. This is a special case $n = 1, T = X, \theta = m$ is the mean. of the following bound.

RAO-CRAMER BOUND. $\boxed{\mathrm{Var}_\theta[T] \geq \frac{(1+B'(\theta))^2}{nI(\theta)}}$ Unbiased case: $\boxed{\mathrm{Err}_\theta[T] \geq \frac{1}{nI(\theta)}}$

Proof.
1) $\theta + B(\theta) = \mathrm{E}_\theta[T] = \int t(x_1,...,x_n) L_\theta(x_1,...,x_n)\, dx_1...dx_n$.
2) $1 + B'(\theta) = \int t(x_1,...,x_n) L'_\theta(x_1,...,x_n)\, dx_1...dx_n = \int t(x_1,...,x_n) \frac{L'_\theta(x_1,...,x_n)}{L_\theta(x_1,...,x_n)}\, dx_1...dx_n = E_\theta[T\frac{L'_\theta}{L_\theta}]$.
3) $1 = \int L_\theta(x_1,...,x_n)\, dx_1...dx_n$ implies $0 = \int L'_\theta(x_1,...,x_n)/L_\theta(x_1,...,x_n) = \mathrm{E}[L'_\theta/L_\theta]$.
4) Using 3) and 2) $\mathrm{Cov}[T, L'_\theta/L_\theta] = E_\theta[TL'_\theta/L_\theta] - 0 = 1 + B'(\theta)$.
5) $(1 + B'(\theta))^2 = \mathrm{Cov}^2[T, \frac{L'_\theta}{L_\theta}] \leq \mathrm{Var}_\theta[T]\mathrm{Var}_\theta[\frac{L'_\theta}{L_\theta}] = \mathrm{Var}_\theta[T] \sum_{i=1}^n E_\theta[(\frac{f'_\theta(x_i)}{f_\theta(x_i)})^2] = \mathrm{Var}_\theta[T]\, nI(\theta)$
where we used 4), the lemma and $L'_\theta/L_\theta = \sum_{i=1}^n f'_\theta(x_i)/f_\theta(x_i)$.

SHANNON ENTROPY. Closely related to the Fisher information is the **Shannon entropy** of a random variable X: $\boxed{S(\theta) = -\int f_\theta \log(f_\theta)\, dx.}$ and the **power entropy** $\boxed{N(\theta) = \frac{1}{2\pi e} e^{2S(\theta)}.}$

INFORMATION INEQUALITIES. If X, Y are independent random variables

a) **Fisher information inequality**: $\boxed{I_{X+Y}^{-1} \geq I_X^{-1} + I_Y^{-1}}$.
b) **Power entropy inequality**: $\boxed{N_{X+Y} \geq N_X + N_Y}$.
c) **Uncertainty property**: $\boxed{I_X N_X \geq 1}$.

In all cases, equality holds if and only if the random variables are Gaussian.
Proof.
a) $I_{X+Y} \leq c^2 I_X + (1-c)^2 I_Y$ is proven using the **Jensen inequality**. Take then $c = I_Y/(I_X + I_Y)$.

RAO-CRAMER BOUND. A random variable X with mean m and variance σ^2 satisfies: $I_X \geq 1/\sigma^2$. Equality holds if and only if X is the Normal distribution.

Proof. This is a special case of Rao-Cramer inequality, where θ is fixed, $n = 1$. The bias is automatically zero. A direct computation giving also uniqueness: $\mathrm{E}[(aX + b)\rho(X)] = \int (ax + b) f'(x)\, dx = -a \int f(x)\, dx = -a$ implies

$$0 \leq \mathrm{E}[(\rho(X) + (X-m)/\sigma^2)^2] = \mathrm{E}[(\rho(X)^2] + 2\mathrm{E}[(X-m)\rho(X)]/\sigma^2 + \mathrm{E}[(X-m)^2/\sigma^4] \leq I_X - 2/\sigma^2 + 1\sigma_2$$

Equality holds if and only if ρ_X is linear, i.e. when X is normal.

COROLLARY: FISHER INFORMATION EXTREMIZATION. The normal distribution has the smallest Fisher information among all distributions with the same variance σ^2.

ENTROPY CHARACTERIZATIONS OF DISTRIBUTIONS. Maximizers of the Shannon entropy are

1) The uniform distribution on $[a, b]$.
2) The exponential distribution on $[0, \infty)$.
3) The Gaussian distribution on the real line.

These are results from the calculus of variations with constraints. For 1), one has to extremize $F(f) = \int_a^b \log(f)\ f\ dx$ under the constraint $G(f) = \int_a^b f(x)\ dx = 1$. The Lagrange equations are $1 - \log(f) = \lambda$, so that $f = 1/(b-a)$ is constant.

LITERATURE: This document partly based on lecture notes of a solid probability and statistics course at ETH given by H. Föllmer, which all mathematics students had to take.