LECTURE 3

SOME CONCEPTS OF DECISION THEORY

The Bayesian paradigm tell us that all unknown aspects regarding our phenomena can be resolved throughout the posterior distribution. In the particular case of a given parametric family, $\mathcal{P}_{\mathbb{O}}$, this entitles to resolve the epistemic uncertainty regarding the phenomena at issue. Concretely, given a parametric family with density $f_{X|\Theta}(x \mid \theta)$ (assumed to be absolutely continuous with respect to a σ-finite measure, ν, on $(\mathbb{X}, \mathcal{X})$) and a prior distribution $q_{\Theta}(\theta)$ on $(\mathbb{O}, \mathcal{B}_{\mathbb{O}})$ the Bayesian machinery reduces to compute

$$f_{\Theta|X}(\theta \mid x) \propto f_{X|\Theta}(x \mid \theta)\, q_{\Theta}(\theta) \tag{3.1}$$

From a merely probabilistic viewpoint the job ends here, namely with the knowledge of the distribution of the uncertain given our information. However, once the posterior distribution is at hand one typically faces decisions to take regarding certain features of interest. Also when $\mathbb{O}$ is of high dimension, extracting useful information from the posterior might be cumbersome. Let us use the example displayed in Figure 3.1, where form the posterior distribution a statistician uses the posterior mode as a point estimation of θ, namely she takes the decision $\delta_1(x) = \mathsf{argmax} f_{\Theta|X}(\theta \mid x)$.

But there is some other statistician who decides to use $\delta_2(x) = \bar{x}$. Hence, the natural question is which one is the best decision. Clearly, to reply to such enquiry one must establish a set of preference criteria among decisions that, in particular, depend on θ. This is precisely the objective behind ***decision theory***.

Indeed, decision theory has had important role in statistical inference, to the extend, of being one of the preferred routes to justify Bayesian statistics, at least under the assumption of parametric models.

Within the Bayesian framework the idea is that, upon observing $X = x$, we want to take some ***action*** $a \in \mathbb{A}$, where $\mathbb{A}$ denotes the space of ***possible actions***. This clearly should depend on the parameter $\theta \in \Theta$. So if we work under the thought that every action induces a loss[1] we can motivate the following definition

Definition 11. A ***randomized decision rule***, δ is a mapping from $\mathbb{X}$ to a probability measure on $(\mathbb{A}, \mathcal{A})$ such that for every $A \in \mathcal{A}$, $\delta(A, x)$ is $\mathcal{A}$-measurable. If for every $x \in \mathbb{X}$ there exist $a_x \in \mathbb{A}$ such that $\delta(A, x) = \mathbb{I}_A(a_x)$ then we speak of a ***deterministic decision rule***. In such a case $a_x := \delta(x)$.

The is an axiomatic foundation justifying the necessity of a loss function that respects the notion of rationality behind decision makers (cf. DeGroot, 1970)

Definition 12. A ***loss function*** is a function $\mathsf{L} : \mathbb{O} \times \mathbb{A} \mapsto \mathbb{R}$.

If δ is a randomised decision rule hence we define

$$L(\theta, \delta(x)) = \int_{\mathbb{A}} L(\theta, a)\delta(\mathrm{d}a, x)$$

In what follows we work with nonrandomised decision rules unless otherwise stated.

[1] Equivalently, one can think instead of utility function.

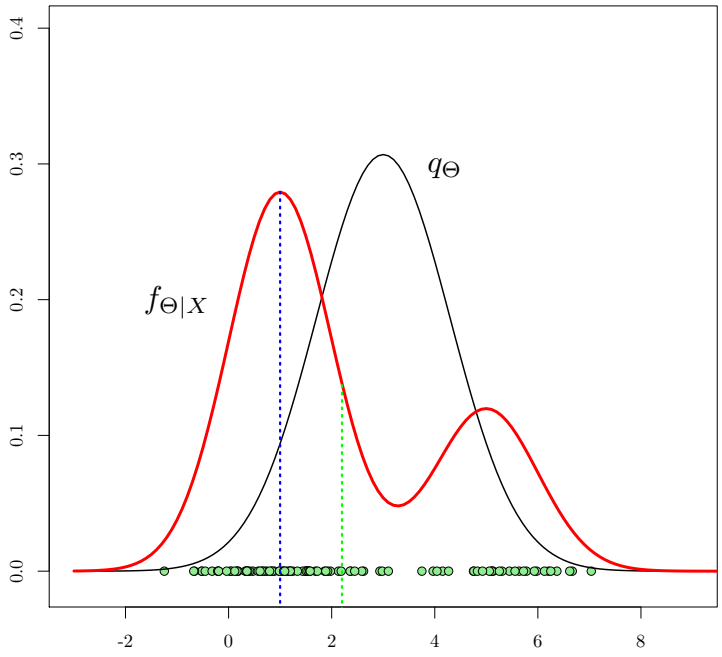

Figure 3.1: Decisions under the Bayesian paradigm. The blue dashed line indicates a possible decision for the value of θ, say the location parameter.

$\mathsf{L}(\theta, a)$ is interpreted as the loss incurred by action a when $\Theta = \theta$. Within the statistical literature, the decisions are typically based on the data $x \in \mathbb{X}$, e.g. $\delta(x) = \hat{\theta}(x) = \bar{x}$, namely the loss of using $\delta(x)$ if θ is the true parameter. However, in general $\delta(x)$ does not need to be an estimate of θ.

Thus, the idea is to disentangle which decision conveys to the smallest loss, even though we do not know the true value of θ. Due to the randomness inherent to Θ is it impossible to uniformly minimize $\delta \mapsto \mathsf{L}(\theta, \delta)$. Hence, here is where division between the frequentist and the Bayesian approaches starts. The frequentist approach seeks to minimize the average loss formalized through the following concept

Definition 13. Given a true, but unknown, value $\theta \in \mathbb{O}$ the ***risk function*** is given by

$$\mathsf{R}(\theta, \delta) = \mathbb{E}_{\mathsf{P}_\theta}[\mathsf{L}(\theta, \delta(X))] = \int_{\mathbb{X}} \mathsf{L}(\theta, \delta(x))\, \mathsf{f}_{X|\Theta}(x \mid \theta)\, \nu(\mathrm{d}x)$$

where as before ν is a reference measure on $(\mathbb{X}, \mathcal{X})$.

Example 8. *Assume $X_1, \ldots, X_n$ ($n > 2$) are iid from $\mathsf{Exp}(\theta)$, we could use $\delta_1(x) = \bar{x}^{-1}$ which corresponds to the maximum likelihood estimation. Hence, if we work under a quadratic loss function, $\mathsf{L}(\theta, \delta) = (\theta - \delta)^2$, we have*

$$\begin{aligned} \mathsf{R}(\theta, \delta_1) &= \mathbb{E}_{\mathsf{P}_\theta}\left[\mathsf{L}(\theta, \delta_1(X))\right] = (\theta - \mathbb{E}_{\mathsf{P}_\theta}[\delta_1(X)])^2 + \mathbb{E}_{\mathsf{P}_\theta}\left[(\delta_1(X) - \mathbb{E}_{\mathsf{P}_\theta}[\delta_1(X)])^2\right] \\ &= \left(\theta - \frac{n\theta}{(n-1)}\right)^2 + \frac{n^2\theta^2}{(n-1)^2(n-2)} = \frac{(n+2)\theta^2}{(n-1)(n-2)} \end{aligned} \tag{3.2}$$

Notice that $\sum_{i=1}^n X_i \sim \mathsf{Ga}(n, \theta)$ (also known as Erlang distribution), thus $\bar{X} \sim \mathsf{Ga}(n, n\theta)$ and $\bar{X}^{-1} \sim \mathsf{IGa}(n, n\theta)$ thus leading to the above result.

Let us consider the case $n = 8$ in the above example, instead of $\delta_1(x)$ one could use different decision rules, namely $\delta_2(x) = \sum_{i=1}^n x_i/(n-4)$ or a sloppy decision such as $\delta_3(x) = 4$. So the question is which of these decisions is better, so far in terms of the risk function. By looking at Figure 3.2, we see that the risk function associated to each of these decisions does not necessarily preserve uniform dominance among them, e.g. we cannot say that $\mathsf{R}(\theta, \delta_1) \leq \mathsf{R}(\theta, \delta_3)$ for all $\theta > 0$.

Since the risk function does not induce a total ordering in the set of all decision rules, thus precluding a direct comparison among rules, the frequentist literature has some criteria to aggregate over the parameter space $\mathbb{O}$ as stated in Definition 15 below.

Definition 14. Let δ be a decision rule. If there exists a decision rule δ_1 such that $\mathsf{R}(\theta, \delta_1) \leq \mathsf{R}(\theta, \delta)$ for all $\theta \in \mathbb{O}$ with strict equality for some θ then we say that δ is ***inadmisible*** and it is ***dominated*** by δ_1. Otherwise δ is ***admisible***.

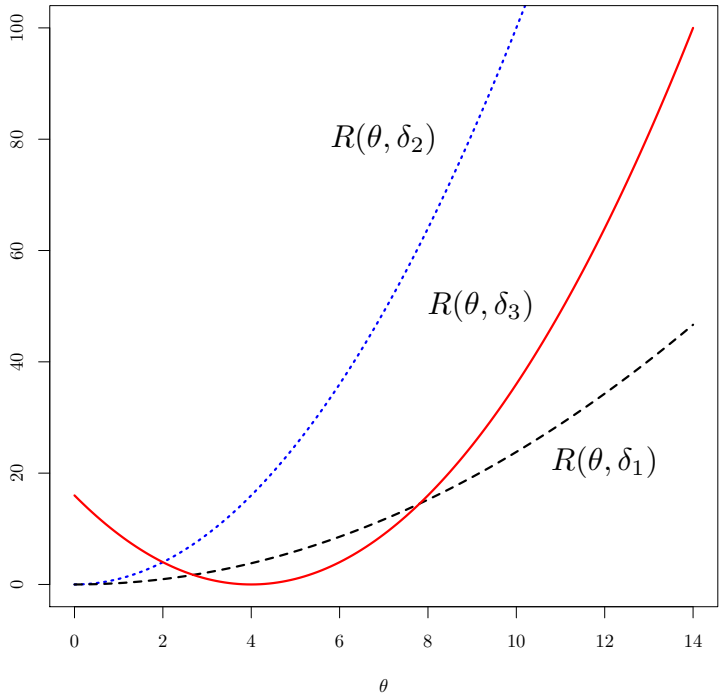

Figure 3.2: Risk function under three different decision rules and quadratic loss function.

In principle, taking admisible decision rules is a good idea to select the best one, however, it can be proved that, in general, it is impossible to minimise $\delta \mapsto \mathsf{R}(\theta, \delta)$ for θ unknown. Under this scenario, the frequentist approach restricts to unbiased decisions, i.e. where $\mathbb{E}[\delta(X)] = \theta$, which avoids having overlapping decisions such as δ_3 in Figure 3.2. Another strategy is to minimise $\sup_{\theta \in \mathbb{O}} \mathsf{R}(\theta, \delta)$, namely to look for the smallest upper bound of the risk function.

Definition 15. A decision rule δ_0 is called ***minimax*** if

$$\sup_{\theta \in \mathbb{O}} \mathsf{R}(\theta, \delta_0) = \inf_{\delta \in \mathbb{D}} \sup_{\theta \in \mathbb{O}} \mathsf{R}(\theta, \delta) \tag{3.3}$$

In particular, it can be seen that if there exist a minimax decision/estimation hence it is admisible. Also if δ_0 is admisible with constant risk then is the unique minimax decision.

1. Bayesian decision theory

Within the Bayesian paradigm it is natural to compute the ***posterior risk*** or ***posterior expected loss***

$$\rho(\delta \mid x) = \mathbb{E}_{\Theta|X}\left[L(\theta, \delta)\right] = \int_{\mathbb{O}} L(\theta, \delta) f_{\Theta|X}(\theta \mid x) \eta(\mathrm{d}\theta) \tag{3.4}$$

Notice that unlike the (frequentist) risk function, that averages over all possible values of $x \in \mathbb{X}$, the posterior risk is a function of the observation at issue. Hence the idea is to choose the decision rule that minimises the posterior risk, i.e.

$$\delta_0(x) = \arg\min_{\delta} \rho(\delta \mid x)$$

for each $x \in \mathbb{X}$

Definition 16. If δ_0 is such that $\rho(\delta_0 \mid x) < \infty$ for all x and $\rho(\delta_0 \mid x) \leq \rho(\delta \mid x)$ for all x and all decision rules δ, then δ_0 is called a (formal) ***Bayes rule*** or ***Bayes action*** .

As an alternative, in a similar spirit of the frequentist approach to aggregate θ, but in this case weighted by the prior q_Θ one can compute the ***integrated risk*** given by

$$r_{q_\Theta}(\delta) \quad = \quad \mathbb{E}_{q_\Theta}\left[\mathsf{R}(\Theta, \delta)\right] \tag{3.5}$$

The integrated risk associates a number with every δ, namely it is not a function of θ. This implies that there is total ordering on the set of decisions/estimators and therefore the existence of a total ordering.

Theorem 5. *An estimator minimising the integrated risk $r_{q_\Theta}(\delta)$ can be obtained by selecting, for every $x \in \mathbb{X}$, the value of $\delta(x)$ which minimises the posterior risk, $\rho(\delta \mid x)$, since*

$$r_{q_\Theta}(\delta) = \int_{\mathbb{X}} \rho(\delta \mid x) f_X(\mathrm{d}x)\nu(\mathrm{d}x)$$

Proof.

$$\begin{aligned} r_{q_\Theta}(\delta) &= \int_{\mathbb{O}} \int_{\mathbb{X}} \mathsf{L}(\theta, \delta(x))\, \mathsf{f}_{X|\Theta}(x \mid \theta)\, \nu(\mathrm{d}x)\, q_\Theta(\theta)\eta(\mathrm{d}\theta) \\ &= \int_{\mathbb{X}} \int_{\mathbb{O}} \mathsf{L}(\theta, \delta(x)) \mathsf{f}_{X|\Theta}(x \mid \theta) q_\Theta(\theta)\eta(\mathrm{d}\theta)\nu(\mathrm{d}x) \\ &= \int_{\mathbb{X}} \int_{\mathbb{O}} \mathsf{L}(\theta, \delta(x)) \mathsf{f}_{\Theta|X}(\theta \mid x)\eta(\mathrm{d}\theta) f_X(\mathrm{d}x)\nu(\mathrm{d}x) \end{aligned} \tag{3.6}$$

Namely

$$\inf_\delta r_{q_\Theta}(\delta) = \int_{\mathbb{X}} \left\{ \inf_\delta \rho(\delta \mid x) \right\} f_X(\mathrm{d}x)\nu(\mathrm{d}x)$$

since $L(\theta, \delta) \geq 0$ implies that $\rho(\delta \mid x) \geq 0$. □

Definition 17. A ***Bayes estimator*** associated with a prior distribution q_Θ and a loss function L, is an estimator δ^q which minimises $r_{q_\Theta}(\delta)$,

$$\arg\min_\delta r_{q_\Theta}(\delta)$$

The value $r(q) := r_{q_\Theta}(\delta^q)$ is called the ***Bayes risk***.

Example 9. *Let $\mathbb{O} \subseteq \mathbb{R}$, $L(\theta, d) = (\theta - d)^2$, hence*

$$\rho(\delta \mid x) = \int_{\mathbb{O}} (\theta - \delta)^2 f_{\Theta|X}(\theta \mid x)\eta(\mathrm{d}\theta) = \mathbb{E}_{\Theta|X}(\Theta^2 \mid X = x) - 2\delta\mathbb{E}_{\Theta|X}(\Theta \mid X = x) + \delta^2,$$

which, when $\mathbb{E}_{\Theta|X}(\Theta^2 \mid X = x) < \infty$, is minimised for $\delta = \mathbb{E}_{\Theta|X}(\Theta \mid X = x)$, namely the Bayes estimator under a quadratic loss is the posterior mean. Notice that such estimator is given by

$$\mathbb{E}_{\Theta|X}(\Theta \mid X = x) = \frac{\int_{\mathbb{O}} \theta\, f_{X|\Theta}(x \mid \theta) q_\Theta(\theta)\eta(\mathrm{d}\theta)}{\int_{\mathbb{O}} f_{X|\Theta}(x \mid \theta) q_\Theta(\theta)\eta(\mathrm{d}\theta)}$$

Proposition 2. *Assume $\mathbb{E}_{\Theta|X}[\Theta] < \infty$. If we further assume the loss function*

$$\mathsf{L}(\theta, a) = c(a - \theta)\mathbb{I}\{a \geq \theta\} + (1 - c)(\theta - a)\mathbb{I}\{a < \theta\}$$

a formal Bayes rule is the $(1 - c)$ quantile of the posterior distribution of Θ. When $c = 0.5$ it reduces to the posterior median.

Proposition 3. *If a Bayes rule δ is unique then is admisible.*

Proof. Assume that δ is a unique Bayes rule and assume that δ^* strictly dominates it, thus

$$r_{q_\theta}(\delta^*) < r_{q_\theta}(\delta)$$

then δ^* improves upon δ or is the Bayes rule, which contradicts the hypothesis. □

Theorem 6. *Let $\mathbb{O} \subset \mathbb{R}^d$. Assume that the risk functions $\mathsf{R}(\theta, \delta)$ are continuous in θ for all decision rules $\delta \in \mathbb{D}$. Moreover, assume that q_Θ places positive mass on any open subset of $\mathbb{O}$. Then a Bayes rule with respect to q_Θ is admissible.*

Proof. Let δ^* be a decision rule that strictly dominates δ. Let $\mathbb{O}_0$ be the set on which $\mathsf{R}(\theta, \delta^*) < \mathsf{R}(\theta, \delta)$. Given a $\theta_0 \in \mathbb{O}_0$, we have $\mathsf{R}(\theta_0, \delta^*) < \mathsf{R}(\theta_0, \delta)$. By continuity, there must exist an $\epsilon > 0$ such that $\mathsf{R}(\theta, \delta^*) < \mathsf{R}(\theta, \delta)$ for all θ satisfying $||\theta - \theta_0|| < \epsilon$. It follows that $\mathbb{O}_0$ is open and hence, by our assumption, $q_\Theta(\mathbb{O}_0) > 0$. Therefore, it must be that

$$\int_{\mathbb{O}_0} \mathsf{R}(\theta, \delta^*) q_\Theta(\theta) \eta(\mathrm{d}\theta) < \int_{\mathbb{O}_0} \mathsf{R}(\theta, \delta) q_\Theta(\theta) \eta(\mathrm{d}\theta)$$

Now observe that

$$\begin{aligned} r_{q_\Theta}(\delta^*) &= \int_{\mathbb{O}} \mathsf{R}(\theta, \delta^*) q_\Theta(\theta) \eta(\mathrm{d}\theta) \\ &= \int_{\mathbb{O}_0} \mathsf{R}(\theta, \delta^*) q_\Theta(\theta) \eta(\mathrm{d}\theta) + \int_{\mathbb{O}_0^c} \mathsf{R}(\theta, \delta^*) q_\Theta(\theta) \eta(\mathrm{d}\theta) \\ &< \int_{\mathbb{O}_0} \mathsf{R}(\theta, \delta) q_\Theta(\theta) \eta(\mathrm{d}\theta) + \int_{\mathbb{O}_0^c} \mathsf{R}(\theta, \delta) q_\Theta(\theta) \eta(\mathrm{d}\theta) \\ &= r_{q_\Theta}(\delta) \end{aligned} \tag{3.7}$$

since

$$\int_{\mathbb{O}_0^c} \mathsf{R}(\theta, \delta^*) q_\Theta(\theta) \eta(\mathrm{d}\theta) < \int_{\mathbb{O}_0^c} \mathsf{R}(\theta, \delta) q_\Theta(\theta) \eta(\mathrm{d}\theta)$$

with strict equality on $\mathbb{O}_0$, which contradicts the assumption that δ is the Bayes rule. □

Definition 18. A prior distribution q_0 on $\mathbb{O}$ is called ***less favorable*** if

$$\inf_\delta r_{q_0}(\delta) = \sup_q \inf_\delta r_q(\delta)$$

The strategy q_0 is sometimes referred to as the ***maximin strategy***.

That is, q_0 is a prior such that the corresponding Bayes rule has the highest possible risk.
Let q_0 and δ_0 a fixed probability and decision rule respectively. It can be seen that

$$\inf_\delta r_{q_0}(\delta) \leq r_{q_0}(\delta_0) \leq \sup_q r_q(\delta_0)$$

so we can introduce the following concepts

Definition 19. Let

$$\underline{V} := \sup_q \inf_\delta r_q(\delta) \leq \inf_\delta r_q(\delta) = \inf_\delta \sup_\theta \mathsf{R}(\theta, \delta) =: \overline{V}$$

The numbers $\underline{V}$ and $\overline{V}$ are respectively called the ***maximin*** and ***minimax*** values of a decision problem.

A way to check that a prior is less favourable and a decision is minimax can be done throughout the following result

Theorem 7. *If δ_0 is a Bayes rule with respect to q_0 and*

$$\mathsf{R}(\theta, \delta_0) \leq r_{q_0}(\delta_0)$$

for all θ, then δ_0 is minimax and q_0 is least favorable.

Proof. Since

$$\overline{V} \leq \sup_{\theta} \mathsf{R}(\theta, \delta_0) \leq r_{q_0}(\delta_0) = \inf_{\delta} r_{q_0}(\delta) \leq \underline{V}$$

and $\underline{V} \leq \overline{V}$ it must be that $\underline{V} = \overline{V}$ so the stated result follows. □

The conundrum here is that is that one of the most preferred estimators, namely the posterior mode, cannot be deduced from a decision theoretical approach, at least not in the general continuous parameter case.