Bayesian Statistics III/IV (MATH3341/4031) Michaelmas term, 2021

Handout 10: Elements of decision theory

Lecturer & author: Georgios P. Karagiannis georgios.karagiannis@durham.ac.uk

Aim: To explain basic concepts and elements of decision theory required for Bayesian inference.

References:

- Berger, J. O. (2013; Sections 1.3, 1.5, 1.6, 2, 4.4, and 4.8). Statistical decision theory and Bayesian analysis. Springer Science & Business Media.
- Robert, C. (2007; Chapter 2). The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer Science & Business Media.
- Raiffa, H., & Schlaifer, R. (1961; Chapter 1). Applied statistical decision theory.
- DeGroot, M. H. (2005; Chapter 8). Optimal statistical decisions (Vol. 82). John Wiley & Sons.
- Ferguson, T. S. (1967; Chapter 1 & 2). Mathematical statistics: A decision theoretic approach. Academic press.

1 Why do we need decision theory in Bayesian Stats?

Another important purpose of statistical analysis, apart from qualifying uncertainty via posterior probability distributions, is to make inference about unknown quantities (e.g., parameters, future outcomes, etc...). In the Bayesian framework, the derivation of inferential tools (such as, estimators, hypothesis tests, etc...) is addressed as a decision theory problem, aka we get decisions about the unknown quantities.

A decision is certainly a function of our knowledge quantified by probabilities. For example, a doctor deciding on the best treatment for a patient will certainly use their knowledge of what illness the patient has. Different illnesses may have different probabilities based on the symptoms, test results, and so on, and it would be foolish to concentrate on the most improbable illness and treat for that. Such questions constitute a decision theory problems.

In statistics, the overall purpose of most inferential studies is to provide the statistician (or a client) with a decision (E.g, what estimator to use? Shall I reject a hypothesis?), it seems reasonable to ask for an evaluation criterion of decision procedures that assesses the consequences of each decision and depends on the parameters of the model, i.e., the true state of the world (or of Nature). As we will see such a criterion is the so called loss function (or more rigorously the utility function) introduced in what follows.

2 Set-up of the decision problem (intro)

Definition 1. The decision problem $(\Theta, \mathcal{D}, \ell)$ involves the following basic elements

- Decision space: $\mathcal{D} = \{d\}$

  The decision maker wishes to select a single decision $d \in \mathcal{D}$ from a space of all possible decisions $\mathcal{D}$.

- State space (Space of the word or Parameter space): $\Theta = \{\theta\}$

The decision process is assumed to be affected by the unknown (uncertain) quantity $\theta \in \Theta$ which signifies the state of the world. The set of all possible states of the world is denoted by Θ. The decision maker perceives that a particular decision $d \in \mathcal{D}$ results in a corresponding state $\theta \in \Theta$.

- Loss function $\ell : \Theta \times \mathcal{D} \to \mathbb{R}_+$

  The decision maker assigns a loss function (or error) $\ell(d, \theta)$ that evaluates the penalty (or error, or suffer, or regret, or etc...) associated with decision d when the parameter takes the value θ.

Definition 2. Statistical decision problem is a decision problem $(\Theta, \mathcal{D}, \ell)$ coupled with an experiment $e \in \mathcal{E}$ ($\mathcal{E}$ denotes the family of experiments) that involves an observable $y \in \mathcal{Y}$ ($\mathcal{Y}$ denotes the sample space) whose sampling distribution $F(y|\theta)$ depends on the state $\theta \in \Theta$ chosen by nature.

Note 3. The main aim in a statistical decision problem $(\Theta, \mathcal{D}, \ell)$ is for the decision maker to choose the/an optimal decision $d^* := d^*(y) \in \mathcal{D}$.

Example 4. Statistical inference consists of taking a decision $d = d(y) \in \mathcal{D}$ elated to the parameter $\theta \in \Theta$ based on the observation $y \in \mathcal{Y}$, where y and θ are related by the sampling distribution $F(y|\theta)$. In many cases, the decision $d \in \mathcal{D}$ will be a procedure performing: Point estimation (what type of estimator to use?), Credible regions (how to find the bounds?), Hypothesis tests (how to reject the hypothesis?), model selection (which model is the 'best'?).

Definition 5. Decision rule is a function which maps the observables $y \in \mathcal{Y}$ to an appropriate decision $d \in \mathcal{D}$. It is defined as $\delta : \mathcal{Y} \to \mathcal{D}$, where $\delta(y) \in \mathcal{D}$, and it is such that

$$\int_{\mathcal{Y}} \ell(\theta, \delta(y)) \mathrm{d}F(y|\theta) < \infty$$

Note 6. Briefly and more rigorously speaking about statistical decision problems,

- The decision maker assigns a utility function $u : \mathcal{E} \times \mathcal{Y} \times \mathcal{D} \times \Theta \to \mathbb{R}$ with $u(e, y, d, \theta)$ which describes the value to perform a particular experiment $e \in \mathcal{E}$, observing a particular observation $y \in \mathcal{Y}$, taking a particular decision $d \in \mathcal{D}$, and then finding that a particular $\theta \in \times$ obtains.
- The existence of the utility function u is introduced in Decision Theory 3, and ignored in this handout.
- Once the utility function u is defined, the loss function ℓ can be specified as

  $$\ell(e, y, d, \theta) = u(e^*, y, d^*, \theta) - u(e, y, d, \theta), \quad \text{or} \quad \ell(e, y, d, \theta) = -u(e, y, d, \theta)$$

  where e^* and d^* are optimal choices of the experiment and decision.
- In this handout, we assume the simplified case where the experiment e is fixed and hence we drop e, y from arguments in u and ℓ; hence

  $$\ell(d, \theta) = u(d^*, \theta) - u(d, \theta), \quad \text{or} \quad \ell(d, \theta) = -u(d, \theta).$$

3 Finding Bayesian optimal decision rules (intro)

Question 7. *Given a statistical decision problem $(\Theta, \mathcal{D}, \ell)$, what is the optimal decision rule $\delta(y) \in \mathcal{D}$?*

Note 8. To get an optimal decision there is a need to derive an effective comparison criterion that orders the possible decisions $d \in \mathcal{D}$ based on the loss function $\ell(\cdot, \cdot)$, (equiv. the utility function $u(\cdot, \cdot)$).

Definition 9. (Frequentist) Risk function $R(\theta, \delta) := R(\theta, \delta(y))$ is defined as

$$R(\theta, \delta) = \mathrm{E}_F(\ell(\theta, \delta(y))|\theta) \quad = \int_{\mathcal{Y}} \ell(\theta, \delta(y)) \mathrm{d}F(y|\theta), \tag{1}$$

where $\delta(\cdot)$ is the decision rule, i.e, the allocation of a decision to each outcome $y \sim F(y|\theta)$ from the random experiment e.

Note 10. For the derivation of inferential tools for θ, such as estimators $\hat{\theta}$, the frequentist approach relies on comparisons based on the risk function (1) in order to choose the 'optimal' decision $\hat{\theta} = \delta^{(\text{FR})}(y)$. The objections are the following:

- Risk function $R(\theta, \delta)$ averages loss function $\ell(\theta, \delta(y))$ over the different values of y proportionally to the density $f(y|\theta)$. Therefore, it seems that the observation y is not taken into account any further. It evaluates procedures on their long-run performance and not directly for the given observation y. To average over all possible values of y, when we know the observed value of y, on one hand may lead to interesting mathematical properties, but on the other hand it is rather a waste of information.
- The frequentist approach implicitly assumes that this problem will be met again and again for the frequency evaluation to make sense. However, there are objections against the this sense of repeatability of experiments. Eg, if new observations become available and one wants to make use of them, this could possibly modify the way the experiment is conducted.
- Risk function $R(\theta, \delta)$ is a function of the parameter θ. There may not exist, in general, an optimal procedure $\delta^{(\text{F})}(\cdot)$ that uniformly minimises $R(\theta, \delta)$ for all $\theta \in \Theta$. The frequentist approach does not induce a total ordering on the set of procedures. For this reason, frequentist approach, requires additional restrictions (in a rather artificial manner), eg, admissibility.

Note 11. To address the decision problem $(\Theta, \mathcal{D}, \ell)$ in the Bayesian paradigm the decision maker assigns a probability distribution $P(\theta, y)$ on the possibility space $\Theta \times \mathcal{Y}$. Recall (Handout 3; Remark 26) that the joint distribution $P(\theta, y)$ determines probability distributions such as

$$\mathrm{d}P(y, \theta) = \mathrm{d}F(y|\theta)\mathrm{d}\Pi(\theta) = \mathrm{d}\Pi(\theta|y)\mathrm{d}F(y)$$

where $\Pi(\theta)$ is the prior distri., $\Pi(\theta|y)$ is the posterior distr., and $F(y)$ is the prior predictive distr.

Note 12. To derive a quantity able to order decisions in the Bayesian paradigm, it is reasonable to integrate out the loss $\ell(\theta, d)$ on the space Θ, (instead of integrating on the space $\mathcal{Y}$) and condition on the observables y. This is because θ is unknown and y is known/observed.

Definition 13. The posterior expected loss of a decision $d \in \mathcal{D}$ when the posterior distribution is $\Pi(\theta|y)$ is defined as

$$\varrho(\pi, d|y) = \mathrm{E}_{\Pi}(\ell(\theta, d)|y) \quad = \int_{\Theta} \ell(\theta, d)\mathrm{d}\Pi(\theta|y) \tag{2}$$

Note 14. Posterior expected loss $\varrho(\pi, d|y)$ integrates the loss $\ell(\theta, d)$ with respect to the posterior distribution $\Pi(\theta|y)$ of the parameter θ, conditionally on the observed value y. It is a function of y but this dependence is not troublesome, because y is known (as it is observed).

Note 15. It is reasonable for a Bayesian Statistician to consider as optimal decision the one that minimizes (2) for a given observable y.

Definition 16. The Bayes risk of a decision $d \in \mathcal{D}$, with respect to a prior $\Pi(\theta)$ on Θ is defined as

$$r(\pi, \delta) = \mathrm{E}_{\Pi}(R(\theta, \delta)) \quad = \int_{\Theta} \int_{\mathcal{Y}} \ell(\theta, \delta(y))\mathrm{d}F(y|\theta)\mathrm{d}\Pi(\theta); \tag{3}$$

It is the risk function $R(\theta, \delta)$ integrated over θ with respect to the prior distribution $\mathrm{d}\Pi(\theta)$

Remark 17. Bayes risk $r(\pi, \delta)$ induces a total ordering on the set of potential decisions, and hence it allows for the direct comparison of decisions unlike the risk function $R(\theta, \delta)$. This is because $r(\pi, \delta)$ associates a real number with every decision d, it is not a function of θ.

Note 18. In the Bayesian framework an optimal decision for the decision problem $(\Theta, \mathcal{D}, \ell)$ can be found in two ways:

1. the extensive form of analysis: it minimizes the posterior expected loss $\varrho(\pi, d|y)$ (2) w.r.t. d

$$\min_d \int_\Theta \ell(\theta, d) \mathrm{d}\Pi(\theta|y)$$

2. the Normal form of analysis: it minimizes the Bayesian risk $r(\pi, \delta)$ (3) w.r.t. δ

$$\min_\delta \int_\Theta \int_\mathcal{Y} \ell(\theta, \delta(y)) \mathrm{d}F(y|\theta) \mathrm{d}\Pi(\theta)$$

Note 19. Theorem 20 provides a constructive tool and a reasonable justification that the two ways are equivalent.

Theorem 20. *A decision minimizing the integrated risk $r(\pi, \delta)$ can be obtained by selecting, for every $y \in \mathcal{Y}$, the value $\delta(y)$ which minimizes the posterior expected loss $\varrho(\pi, d|y)$, since*

$$\min_{\forall \delta \in \mathcal{D}} r(\pi, \delta) = \int_\mathcal{Y} \min_{\forall d \in \mathcal{D}} \varrho(\pi, d|y) dF(y)$$

and

$$r(\pi, \delta) = \int_\mathcal{Y} \varrho(\pi, d|y) dF(y) = \begin{cases} \int_\mathcal{Y} \varrho(\pi, d|y) f(y) dy & , y \text{ cont.} \\ \sum_\mathcal{Y} \varrho(\pi, d|y) f(y) & , y \text{ is disc.} \end{cases}$$

Proof. For this proof assume that the loss function is bounded ($\ell(\theta, d) \geq 0$). Assume that y and θ are continuous for simplicity, e.g.

$$p(\theta, y) = f(y|\theta)\pi(\theta) = \pi(\theta|y) f(y)$$

It is

$$\begin{aligned} r(\pi, \delta) &= \int_\Theta \int_\mathcal{Y} \ell(\theta, \delta(y)) \mathrm{d}F(y|\theta) \mathrm{d}\Pi(\theta) \\ &= \int_\Theta \int_\mathcal{Y} \ell(\theta, d) f(y|\theta) \mathrm{d}y \pi(\theta) \mathrm{d}\theta && (4) \\ &= \int_\mathcal{Y} \int_\Theta \ell(\theta, d) f(y|\theta) \pi(\theta) \mathrm{d}\theta \mathrm{d}y && (5) \\ &= \int_\mathcal{Y} \int_\Theta \ell(\theta, d) \pi(\theta|y) f(y) \mathrm{d}\theta \mathrm{d}y && (6) \\ &= \int_\mathcal{Y} \left[\int_\Theta \ell(\theta, d) \mathrm{d}\Pi(\theta|y) \right] \mathrm{d}F(y) && (7) \\ &= \int_\mathcal{Y} \varrho(\pi, d|y) \mathrm{d}F(y) && (8) \end{aligned}$$

Hence, it is implied that

$$\min_{\forall \delta \in \mathcal{D}} r(\pi, \delta) = \int_\mathcal{X} \min_{\forall d \in \mathcal{D}} \varrho(\pi, d|y) f(y) \mathrm{d}y.$$

Here, we used Fubini's theorem for (4) $\Rightarrow$ (5), and Bayes theorem for (5) $\Rightarrow$ (6). □

Definition 21. A Bayes rule associated with a prior distribution $\Pi(\theta)$ and a loss function $\ell(\cdot, \cdot)$ is any δ^π which minimizes $r(\pi, \delta)$. For every $y \in \mathcal{Y}$, it is given by $\delta^\pi(y)$ such that

$$\delta^\pi(y) = \arg\min_d \varrho(\pi, d|y)$$

Definition 22. The Minimum Bayes risk is defined as the value $r(\pi) = r(\pi, \delta^\pi)$, where δ^π is a Bayes rule.

Remark 23. Operationally, to find the Bayes rule $\delta^\pi(y)$, we will mainly minimize the posterior expected loss $\varrho(\pi, d|y)$.

Remark 24. From a strictly Bayesian point of view, only the posterior expected loss $\varrho(\pi, d|y)$ is important, because the Bayesian paradigm is based on the conditional approach (conditional on the observations).

Remark 25. The 'reasonable' way to view the situation is that minimizing the posterior expected loss $\varrho(\pi, d|y)$. One should condition on what is known (aka the observables y) and integrate/average out on what is unknown (aka θ). Minimizing $r(\pi, \delta^\pi)$ that integrates out y seems bizarre from this perspective.

Remark 26. Unlike the Frequentist one, the Bayesian approach is sufficiently reductive to reach an effective decision δ^π because, by minimizing the posterior expected risk $\varrho(\pi, d|y)$ (w.r.t. δ^π), it minimizes the integrated risk $r(\pi, \delta^\pi)$ (w.r.t. δ^π) which in fact allows a direct comparison of estimators.

Remark 27. The Bayesian approach works conditional upon the actual observation y, as well as it incorporates the probabilistic properties of the sampling distribution $F(y|\theta)$. This is in contrast to the frequentist approach (Note 10) where the observed information y seems to be wasted because it averages over all possible values of y instead of conditioning on the given observed y.

Definition 28. Generalized Bayes rule is called any $\delta^\pi(y)$ that minimizes $r(\pi, \delta)$ or $\varrho(\pi, d|y)$ for each $y \in \mathcal{Y}$ where $f(y) > 0$ but the prior distribution $\Pi(\theta)$ is improper.

Remark 29. Definition 21, and Theorem 20 are valid for proper and improper priors, provided that $r(\pi) < \infty$, otherwise other treatments exist.

4 Admissibility

Contrary to the Bayesian approach, the frequentist approach is not reductive enough to lead to a single optimal rule/estimator, and for this reason frequentist use additional optimality concepts such as admissibility. Here, we study Bayes rules/estimators with respect to the frequentist optimality criterion of admissibility.

Definition 30. A decision rule δ_0 is inadmissible if there exists a decision rule δ_1 which dominates δ_0, namely:

$$R(\theta, \delta_0) \geq R(\theta, \delta_1), \quad \forall \theta \in \Theta$$
$$R(\theta_0, \delta_0) > R(\theta_0, \delta_1), \quad \exists \theta_0 \in \Theta$$

Otherwise, δ_0 is said to be admissible.

Remark 31. Well, I wouldn't use an inadmissible estimator δ because a decision rule with a lower risk could be found...

Admissibility in Bayesian rules

Note 32. One may expect Bayes rule to be admissible because if a rule with better risk $R(\theta, \delta)$ existed, that rule would also have better Bayes risk $r(\pi, \delta) = \mathrm{E}_\Pi(R(\theta, \delta))$. This is always true when the prior π is proper, and we study this case in what follows. Surprisingly, however, this is not always true when π is improper (non-informative).

Theorem 33. *If the Bayes rule associated with a prior π is unique, it is admissible.*

Proof. Let δ^π be a Bayes rule. Assume that δ^π is inadmissible. So let δ^* be any decision rule with $R(\theta, \delta^*) \leq R(\theta, \delta^\pi)$, for all $\theta \in \Theta$. Then

$$r(\pi, \delta^\pi) - r(\pi, \delta^*) = \int_\Theta R(\theta, \delta^*) \mathrm{d}\Pi(\theta) - \int_\Theta R(\theta, \delta^\pi) \mathrm{d}\Pi(\theta) = \int_\Theta (R(\theta, \delta^*) - R(\theta, \delta^\pi)) \mathrm{d}\Pi(\theta) \leq 0$$

and so δ^* is also Bayes. Because δ^π is unique Bayes by assumption, we must have $\delta^*(y) = \delta^\pi(y)$ for all $y \in \mathcal{Y}$. Therefore, δ^π must be admissible. □

Theorem 34. *If a prior distribution π is strictly positive on Θ, with finite Bayes risk and the risk function, $R(\theta, \delta)$, is a continuous function of θ for every δ, the Bayes estimator δ^π is admissible.*

Proof. Let δ^π be a Bayes rule. Assume that δ^π is inadmissible. So let δ^* be any decision rule with $R(\theta,\delta^*) \leq R(\theta,\delta^\pi)$, for all $\theta \in \Theta$, and $R(\theta_0,\delta^*) < R(\theta_0,\delta^\pi)$, for some $\theta_0 \in \Theta$. Let $R(\theta_0,\delta^\pi) - R(\theta_0,\delta^*) = \eta > 0$ for some $\theta_0 \in \Theta$. By continuity of $R(\theta,\delta)$ in θ, $R(\theta,\delta^\pi) - R(\theta,\delta^*)$ is continues as well, and hence

$$(\forall \epsilon > 0)(\exists \zeta > 0)(\forall \theta \in \Theta)(|\theta - \theta_0| < \zeta \Longrightarrow |R(\theta,\delta^\pi) - R(\theta,\delta^*)| < \epsilon)$$

which implies $R(\theta,\delta^\pi) - R(\theta,\delta^*) > \eta - \epsilon = \tilde{\eta} > 0$ Let $A = \{\theta \in \Theta \text{ st } |\theta - \theta_0| < \zeta\}$. Then,

$$\begin{aligned} r(\pi,\delta^\pi) - r(\pi,\delta^*) &= \int_\Theta (R(\theta,\delta^\pi) - R(\theta,\delta^*))\mathrm{d}\Pi(\theta) \\ &= \int_A (R(\theta,\delta^\pi) - R(\theta,\delta^*))\mathrm{d}\Pi(\theta) + \int_{A^\complement} (R(\theta,\delta^\pi) - R(\theta,\delta^*))\mathrm{d}\Pi(\theta) \\ &\geq \int_A (R(\theta,\delta^\pi) - R(\theta,\delta^*))\mathrm{d}\Pi(\theta) > \int_A \tilde{\eta}\mathrm{d}\Pi(\theta) = \tilde{\eta}\mathsf{P}_\Pi(\theta \in A) > 0 \end{aligned}$$

which contradicts that δ^π is Bayes rule. Therefore, δ^π must be admissible. □

Theorem 35. *Assume that Θ is discrete (say $\Theta = \{\theta_1, ...\}$) and that the prior π gives positive probability to each $\theta_i \in \Theta$, then Bayes rule δ^π with respect to π is admissible.*

Proof. Let δ^π be a Bayes rule, and δ^* be any decision rule with $R(\theta,\delta^*) \leq R(\theta,\delta^\pi)$, for all $\theta \in \Theta$, and $R(\theta_k,\delta^*) < R(\theta_k,\delta^\pi)$, for some $\theta_k \in \Theta$. Let $R(\theta_k,\delta^\pi) - R(\theta_k,\delta^*) = \eta > 0$ for some $\theta_0 \in \Theta$. It is

$$\begin{aligned} r(\pi,\delta^\pi) - r(\pi,\delta^*) &= \sum_{i=1}^{\infty} \left(R(\theta_i,\delta^\pi) - R(\theta_i,\delta^*)\right)\pi(\theta_i) \\ &= \sum_{i\neq k} \left(R(\theta_i,\delta^\pi) - R(\theta_i,\delta^*)\right)\pi(\theta_i) + \left(R(\theta_k,\delta^\pi) - R(\theta_k,\delta^*)\right)\pi(\theta_k) \\ &\geq \left(R(\theta_k,\delta^\pi) - R(\theta_k,\delta^*)\right)\pi(\theta_k) \ > \eta\pi(\theta_k) > 0 \end{aligned}$$

which contradicts that δ^π is Bayes estimator. Therefore, δ^π must be admissible. □

Admissibility in Generalised Bayesian rules

Generalized Bayes rules (Definition 28) may be or may not be admissible.

Remark 36. When the loss is positive and

$$r(\pi,\delta^\pi) = \int_\Theta R(\theta,\delta^\pi)\mathrm{d}\Pi(\theta) < \infty,$$

the generalized Bayes rule δ^π (improper prior) can be easily shown to be admissible, similar to the Bayes rule case (proper prior; Theorems 34 & 35). In fact in this case, as mentioned in Remark 29, δ^π minimizes $r(\pi,\delta)$ or $\varrho(\pi,d|x)$ and it can be shown (similar to the Bayes rule) that generalized Bayes rule must be admissible under suitable conditions.

Remark 37. When

$$r(\pi,\delta^\pi) = \int_\Theta R(\theta,\delta^\pi)\mathrm{d}\Pi(\theta) = \infty$$

even super reasonable generalized Bayes rules can be inadmissible; (See Example 38). Unfortunately, the use of improper priors quite often lead to $r(\pi,\delta^\pi) = \infty$, hence they are evil....

Example 38. Consider the Bayesian model

$$\begin{cases} y|b & \sim \text{Ga}(a, 1/b) \\ b & \sim \Pi(b) \end{cases}$$

where the non informative prior such that $\pi(b) \propto \frac{1}{b}$ is used, and $a > 0$ is known.

1. Check if prior $\pi(b)$ is proper.
2. Given the loss function $\ell(b, \delta) = (b - \delta)^2$, find the Bayes rule (estimate).
3. Is that Bayes rule admissible? Check it with the decision rule $\delta_c(y) = cy$ for $c \in \mathbb{R}$.

Hint: Gamma distr.: $x \sim \text{Ga}(a, b)$ has pdf $f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx) 1_{(0,+\infty)}(x)$, $\text{E}(x) = \frac{a}{b}$, and $\text{Var}(x) = \frac{a}{b^2}$.

Hint: Inverse Gamma distr.: $x \sim \text{IG}(a, b)$ has pdf $f(x) = \frac{b^a}{\Gamma(a)} x^{-a-1} \exp(-\frac{b}{x}) 1_{(0,+\infty)}(x)$, and $\text{E}(x) = \frac{b}{a-1}$.

Solution.

1. It is $\int_0^\infty \text{d}\Pi(b) = \int_0^\infty \pi(b)\text{d}b = \int_0^\infty \frac{1}{b}\text{d}b = \infty$ so it is an improper prior.
2. By using Bayes theorem, I can compute the posterior distribution and recognize it as $b|y \sim \text{IG}(a, y)$. So the Bayes rule is

$$\begin{aligned} 0 = \frac{\text{d}}{\text{d}\delta} \varrho(\pi, \delta|y)\Big|_{\delta=\delta^\pi} &= \frac{\text{d}}{\text{d}\delta} \text{E}_{\text{IG}(b|a,y)}(\ell(b, \delta))\Big|_{\delta=\delta^\pi} \\ &= \frac{\text{d}}{\text{d}\delta} \int_\Theta (b - \delta)^2 \text{IG}(b|a, y)\text{d}b\Big|_{\delta=\delta^\pi} = \int_\Theta \frac{\text{d}}{\text{d}\delta}(b - \delta)^2 \text{IG}(b|a, y)\text{d}b\Big|_{\delta=\delta^\pi} \\ &= -2 \int_\Theta b\text{IG}(b|a, y)\text{d}b + 2\delta^\pi = -2\text{E}_{\text{IG}(a,y)}(b) + 2\delta^\pi = -2\frac{y}{a-1} + 2\delta^\pi \implies \\ \delta^\pi(x) &= \frac{y}{a-1} \end{aligned}$$

3. First, I will try to check first if $\delta^\pi(y)$ is inadmissible. In order to show that $\delta^\pi(y)$ is inadmissible, I need to find a decision rule that dominates $\delta^\pi(y)$.

   Consider the decision rule $\delta_c(y) = cy$, where c is some constant. I will try to find a value for c, let's say c_* such that $\delta_{c_*}(y)$ can dominate δ^π. Decision rule $\delta_c(y)$ has risk

$$\begin{aligned} R(b, \delta_c) = \text{E}_{\text{Ga}(a,1/b)}(cy - b)^2 &= \text{E}_{\text{Ga}(a,1/b)}(cy - c\text{E}_{\text{Ga}(a,1/b)}(y) + c\text{E}_{\text{Ga}(a,1/b)}(y) - b)^2 \\ &= c^2 \text{E}_{\text{Ga}(a,1/b)}(y - \text{E}_{\text{Ga}(a,1/b)}(y))^2 + (c\text{E}_{\text{Ga}(a,1/b)}(y) - b)^2 \\ &= c^2 \text{Var}_{\text{Ga}(a,1/b)}(x) + (cab - b)^2 = b^2(c^2 a + (ca - 1)^2) \end{aligned}$$

   To specify δ_{c_*}, a nice value c_* that I can use is the one that minimizes $R(b, \delta_c)$. It can be shown that $R(b, \delta_c)$ has minimum at $c_* = 1/(a + 1)$. In fact, it is

$$\frac{\text{d}}{\text{d}c} R(b, \delta_c)\Big|_{c=c_*} = 0 \implies 2b^2 a(c + (ca - 1))\big|_{c=c_*} = 0 \implies c_* = \frac{1}{a+1}$$

   and

$$\frac{\text{d}^2}{\text{d}c^2} R(b, \delta_c)\Big|_{c=c_*} = 2b^2 a(a + 1) > 0$$

   Now let's check if δ_{c_*} dominates δ^π (and hence δ^π is inadmissible)

$$\frac{R(b, \delta^\pi)}{R(b, \delta_{c_*})} = \frac{R(b, \delta_{\frac{1}{a-1}})}{R(b, \delta_{\frac{1}{a+1}})} \overset{\text{calc.}}{=} \cdots = \frac{a(a-1)^{-2} + (a/(a-1) - 1)^2}{a(a+1)^{-2} + (a/(a+1) - 1)^2} = \frac{(a+1)^2}{(a-1)^2} > 1 \tag{9}$$

   Hence, δ_{c_*} dominates δ^π, for all b. In fact, here, $R(b, \delta_{c_*}) < R(b, \delta^\pi)$ for all b. This shows that the generalised Bayesian rule $\delta^\pi(y) = \frac{y}{a-1}$ is inadmissible ...

From (9), we see that the risk of $\delta^\pi(y)$ significantly worsens compared to that of $\delta_{c_*}(y)$ when a decreases.

Long run analysis of the behavior of Generalised Bayesian estimator δ^π. Consider the case that an objective Bayesian uses non-informative priors (improper in this case) automatically on a routine bases because he/she does not want to contaminate his/her statistical analysis with subjective elements. Due to this repeated use, the Bayesian enters into the frequentist domain; hence it is reasonable to investigate how repeated use of this prior actually performs.

Consider

- a sequence of independent problems $\left((\theta^{(1)}, y^{(1)}), (\theta^{(2)}, y^{(2)}), ...\right)$
- a loss function $\ell(\theta, \delta)$ that measures the performance of the procedure in each problem
- a quantity to compare δ_1 and δ_2, in the limit $N \to \infty$ is

$$S_N = \sum_{i=1}^{N} \left(\ell\left(\theta^{(i)}, \delta_1(y^{(i)})\right) - \ell\left(\theta^{(i)}, \delta_2(y^{(i)})\right)\right)^2$$

the limiting behavior of S_N is related to the risk functions $R(\theta, \delta_1)$, $R(\theta, \delta_2)$ as we will see.

Theorem 39. *Consider $\theta = (\theta^{(1)}, \theta^{(2)}, ...)$ to be any fixed sequence of parameters $\theta^{(i)} \in \Theta$, and suppose random variables $y^{(i)} \in \mathcal{Y}$ are independently generated from the parametric models $f(y^{(i)}|\theta^{(i)})$ (here f is the same for the entire sequence). Define the random variables*

$$Z_i = \ell\left(\theta^{(i)}, \delta_1(x^{(i)})\right) - \ell\left(\theta^{(i)}, \delta_2(x^{(i)})\right)$$

and assume that $Var_F(Z_i|\theta^{(i)}) < \infty$ for all i. If $R(\theta, \delta_1) - R(\theta, \delta_2) > \epsilon > 0$ for all $\theta \in \Theta$, then

$$\lim_{N\to\infty} \inf \frac{1}{N} S_N > \epsilon, \qquad \textit{w.p. } 1 \tag{10}$$

for any sequence of θ.

Proof. It is

$$\mathrm{E}_F\left(Z_i|\theta^{(i)}\right) = R(\theta, \delta_1) - R(\theta, \delta_2)$$

Since $\mathrm{Var}_F(Z_i|\theta^{(i)}) < \infty$, by the SLLN, as $N \to \infty$, it is

$$\frac{1}{N}\sum_{i=1}^{N}\left(Z_i - \mathrm{E}_F(Z_i|\theta^{(i)})\right) \to 0, \qquad \text{w.p. } 1$$

because $\mathrm{E}_F(Z_i|\theta^{(i)}) > \epsilon$, (10) is implied. □

Remark 40. Theorem 39 implies that if our generalized Bayes rule (let's say δ_1) is inadmissible, δ_1 will be always inferior to δ_2 in actual practical use.

Remark 41. If ϵ is too small in (10), we may tolerate to use inadmissible δ_1 in routine use, however, we should proceed consciously. Once again, in Theorem 39, inadmissibility applies only to automated use of δ_1; eg, a computer package.

Question 42. *Practice with Exercise 62 from the Exercise sheet, and the proof of Theorem 34.*