Teoria Bayesiana

Ralph dos Santos Silva

http://www.ralphsilva.org/teoriabayesiana.html

Instituto de Matemática
Centro de Ciências Matemáticas e da Natureza
Universidade Federal do Rio de Janeiro

Sumário

Introduction

Decision Theoretic Foundations

From Prior Information to Prior Distributions

Bayesian Point Estimation

Tests and Confidence Regions

Bayesian Calculations

Introduction

We consider that this second approach (parametric) is more pragmatic, since it takes into account that a finite number of observations can efficiently estimate only a finite number of parameters.

Observations: $x_1, x_2, \ldots, x_n$,

$$x_i \sim f_i(x_i|\theta_i, x_1, \ldots, x_{i-1}) \quad \text{on} \quad \mathbb{R}^p,$$

with f_i known and θ_i unknown.

Definition

A parametric model: $x \sim f(x|\theta)$ with θ unknown, $\theta \in \Theta$ (finite) dimension.

- Inverse probability is the same as Statistics.
- Likelihood function: $\ell(\theta|x) = f(x|\theta)$.

Bayes's theorem

$$P(A|E) = \frac{P(E|A)P(A)}{P(E|A)P(A) + P(E|A^c)P(A^c)} = \frac{P(E|A)P(A)}{P(E)} \quad \text{with} \quad P(E) \neq 0$$

If $P(B) = P(A)$, then

$$\frac{P(A|E)}{P(B|E)} = \frac{P(E|A)}{P(E|B)}.$$

Bayes's theorem (1764):

$$g(y|x) = \frac{f(x|y)g(y)}{\int f(x|y)g(y)dy}.$$

Posterior distribution

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta}.$$

Definition

A Bayesian statistical model is made of a parametric statistical model, $f(x|\theta)$, and a prior distribution on the parameters, $\pi(\theta)$.

Definition

When $x \sim f(x|\theta)$, a function T of x (also called a statistics) is said to be sufficient if the distribution of x conditional upon $T(x)$ does not depend on θ.

The factorization theorem:

$$f(x|\theta) = g(T(x)|\theta)h(x|T(x))$$

if g is the density of $T(x)$.

Sufficiency principle: Two observations x and y factorizing through the same value of a sufficient T, that is, such that $T(x) = T(y)$, must lead to the same inference on θ.

Likelihood principle: The information brought by an observation x about θ is entirely contained in the likelihood function $\ell(\theta|x)$. Moreover, if x_1 and x_2 are two observations depending on the same parameter θ, such that there exists a constant c satisfying

$$\ell_1(\theta|x_1) = c\ell_2(\theta|x_2),$$

for every θ, they then bring the same information about θ and must lead to identical inferences.

Stopping rule principle: If a sequence of experiments, $\varepsilon_1, \ \varepsilon_2, \ \ldots$ is directed by a stopping rule, τ, which indicates when the experiments should stop, inference about θ must depend on τ only through the resulting sample.

Conditionality principle: If two experiments on the parameter θ, ε_1 and ε_2, are available and if one of these two experiments is selected with probability *p*, the resulting inference on θ should only depend on the selected experiment.

Theorem

The Likelihood Principle is equivalent to the conjunction of the Sufficiency and the Conditionality Principles.

Maximum likelihood

When $x \sim f(x|\theta)$ is observed, the maximum likelihood approach considers the following estimator of θ

$$\widehat{\theta} = \arg\sup_{\theta} \ell(\theta|x),$$

i.e., the value of θ that maximizes the density at x, $f(x|\theta)$.

Prior and posterior distributions

- The joint distribution of (θ, x):

$$\varphi(\theta, x) = f(x|\theta)\pi(\theta).$$

- The marginal distribution of x:

$$m(x) = \int \varphi(\theta, x) d\theta = \int f(x|\theta)\pi(\theta) d\theta.$$

- The posterior distribution of θ:

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\displaystyle\int f(x|\theta)\pi(\theta) d\theta}.$$

- The predictive distribution of y with $y \sim g(y|\theta, x)$:

$$g(y|x) = \int g(y|\theta, x)\pi(\theta|x) d\theta.$$

An important aspect of the Bayesian paradigm in nonidentifiable settings is, however, that the prior distribution can be used as a tool to identify the parts of the parameter that are not covered by the likelihood, even though the choice of prior may have a bearing on the identifiable part.

Evaluating estimators

- $\mathcal{D}$: set of possible decisions;
- Decision space;
- In most examples, $\mathcal{D} = \Theta$.

Definition (2.1.1)

A loss function is any function L from $\Theta \times \mathcal{D}$ in $[0, \infty]$.

Notation: $L(\theta, d)$.

The loss function $L(\theta, \delta)$ measures the error made in evaluating $h(\theta)$ by δ. In this case, $\mathcal{D}$ is Θ or $h(\Theta)$.

Statistical inference should start with:

- The distribution family of the observations;
- The prior distribution for the parameters, $\pi(\theta)$; and
- The loss associated with the decision, $L(\theta, \delta)$.

Existence of a utility function

$\mathcal{R}$: space of rewards, e.g. $\mathcal{R} = \mathbb{R}$.

Hypothesis: it is possible to order the rewards (there exist a total ordering).

1. $r_1 \preccurlyeq r_2$ or $r_2 \preccurlyeq r_1$; and
2. If $r_1 \preccurlyeq r_2$ and $r_2 \preccurlyeq r_3$, then $r_1 \preccurlyeq r_3$.

Notation: $\prec$ for strict order, and $\sim$ for equivalence.

Note: One and only one of the following relations is satisfied by any pair (r_1, r_2) in $\mathcal{R}$:

$$r_1 \prec r_2, \quad r_2 \prec r_1, \quad \text{or} \quad r_1 \sim r_2.$$

It is necessary to extend the reward space $\mathcal{R}$ to $\mathcal{P}$, the space of probabilities distributions on $\mathcal{R}$.

Example

In most real-life settings, the rewards associated with an action are not exactly known when the decision is taken or, equivalently, some decisions involve a gambling step. For instance, in finance, the monetary revenue $r \in \mathcal{R} \equiv \mathbb{R}$ derived from stock market shares is not guaranteed when the shareholder has to decide from which company she should buy shares. In this case, $\mathcal{D} = \{d_1, \ldots, d_n\}$, where d_k represents the action "buy the share from company k." At the time of the decision, the rewards associated with the different shares are random dividends, only known by the end of the year.

The order relation $\preccurlyeq$ is also assumed to be available on $\mathcal{P}$.

It is possible to compare two distributions of probability on $\mathcal{R}$, P_1 and P_2.
$\mathcal{A}_1$: $P_1 \preccurlyeq P_2$ or $P_2 \preccurlyeq P_1$; and

$\mathcal{A}_2$: If $P_1 \preccurlyeq P_2$ and $P_2 \preccurlyeq P_3$, then $P_1 \preccurlyeq P_3$.

The existence of the order $\preccurlyeq$ on $\mathcal{P}$ relies on the assumption that there exist an optimal reward.

There exists a function U on $\mathcal{R}$ associated with $\preccurlyeq$, such that $P_1 \preccurlyeq P_2$ is equivalent to

$$\mathrm{E}^{P_1}(U(r)) \leqslant \mathrm{E}^{P_2}(U(r)).$$

This function U is called the utility function.

We only consider the set of *bounded* distributions on $\mathcal{P}_{\mathcal{B}}$ (bounded support), for which there exist r_1 and r_2 such that

$$[r_1, r_2] = \{r : \ r_1 \preccurlyeq r \preccurlyeq r_2\} \quad \text{and} \quad P([r_1, r_2]) = 1.$$

For P_1, P_2 in $\mathcal{P}_{\mathcal{B}}$, we define the mixture $P = \alpha P_1 + (1-\alpha)P_2$ as the distribution that generates a reward from P_1 with probability α and a reward from P_2 with probability $(1-\alpha)$.

For instance, $\alpha r_1 + (1-\alpha)r_2$ is the distribution that gives the reward r_1 with probability α and the reward r_2 with probability $(1-\alpha)$.

There must be *conservation of the ordering under indifferent alternatives*:
$\mathcal{A}_3$: If $P_1 \preccurlyeq P_2$, then $\alpha P_1 + (1-\alpha)P \preccurlyeq \alpha P_2 + (1-\alpha)P$ for every $P \in \mathcal{P}$.

Example

If the share buyers of the previous example can compare two companies with dividend distributions P_1 and P_2, they should be able to keep a ranking of the two companies if there is a chance $(1-\alpha)$ that both dividends are replaced by state bounds with dividend distribution P.

The order relation must also be *connected* (or *closed*):

$\mathcal{A}_4$: If $P_1 \preccurlyeq P_2 \preccurlyeq P_3$, then there exist $\alpha \in (0,1)$ and $\beta \in (0,1)$ such that $\alpha P_1 + (1-\alpha)P_3 \preccurlyeq P_2 \preccurlyeq \beta P_1 + (1-\beta)P_3$.

Lemma

If r_1, r_2, and r are rewards in $\mathcal{R}$ with $r_1 \prec r_2$ and $r_1 \preccurlyeq r \preccurlyeq r_2$, there exists a unique ν $(0 \leqslant \nu \leqslant 1)$ such that $r \sim \nu r_1 + (1-\nu)r_2$.

This lemma is the key to the derivation of the *utility function U* on $\mathcal{R}$. Proof in DeGroot (1970, p. 105).

Lemma

If r_1, r_2, and r_3 are three rewards in $\mathcal{R}$ such that $r_2 \sim \alpha r_1 + (1-\alpha)r_3$, then

$$U(r_2) = \alpha U(r_1) + (1-\alpha)U(r_3).$$

The extension of the definition of the utility function to $\mathcal{P}_{\mathcal{B}}$ calls for an additional assumption.

Given P such that $P([r_1, r_2]) = 1$, define

$$\alpha(r) = \frac{U(r) - U(r_1)}{U(r_2) - U(r_1)} \quad \text{and} \quad \beta = \int_{[r_1, r_2]} \alpha(r) dP(r).$$

$\mathcal{A}_5$: $P \sim \beta\delta_{r_2} + (1 - \beta)\delta_{r_1}$.

It implies that, if r is equivalent to $\alpha(r)r_1 + (1 - \alpha(r))r_2$ for every $r \in [r_1, r_2]$, this equivalence must hold on average.

Notice that β is derived from the expected utility

$$\beta = \frac{\mathrm{E}^P(U(r)) - U(r_1)}{U(r_2) - U(r_1)},$$

and this assumption provides a definition of U on $\mathcal{P}_\mathcal{B}$.

Theorem

Consider P_1 and P_2 in $\mathcal{P}_{\mathcal{B}}$. Then,

$$P_1 \preccurlyeq P_2 \quad \text{if, and only if,} \quad E^{P_1}(U(r)) \leqslant E^{P_2}(U(r)).$$

Moreover, if $U^\star$ is another utility function satisfying the above equivalence relation, there exist $a > 0$ and b such that $U^\star(r) = aU(r) + b$.

Let $\mathcal{P}_\varepsilon$ be the set of distributions P in $\mathcal{P}$ such that $\mathrm{E}^P(U(r))$ is finite.

Theorem

Consider P and Q, two distributions in $\mathcal{P}_\varepsilon$. Then, $P \preccurlyeq Q$ if, and only if,

$$E^P(U(r)) \leqslant E^Q(U(r)).$$

Example (Saint Petersburg Paradox)

Consider a game where a coin is thrown until a head appears. When this event occurs at the nth throw, the player gain is 3^n, leading to an average gain of

$$L = \sum_{n=1}^{+\infty} 3^n \frac{1}{2^n} = +\infty.$$

Every player should then be ready to pay an arbitrarily high entrance fee to play this game, even though there is less than a 0.05 chance to go beyond the fifth throw! This modeling does not take into account that the fortune of a player is necessarily bounded and that he or she can only play a limited number of games. A solution to this paradox is to substitute for the linear utility function a bounded utility function, such as

$$U(r) = \frac{r}{\delta + r}, \qquad (\delta > 0,\ r > -\delta),$$

and $U(r) = -\infty$ otherwise. This construction is quite similar to Laplace's moral expectation. An acceptable entrance fee e will then be such that the expected utility of the game is larger than the utility of doing nothing, i.e.,

$$\mathrm{E}(U(r - e)) \geqslant U(0) = 0.$$

Utility and loss

Three spaces:

$\mathcal{X}$: observation space;

Θ: parameter space; and

$\mathcal{D}$: decision space (or action space).

Note that

- x and θ are related by $f(x|\theta)$;
- d is to evaluate (or estimate) θ or $h(\theta)$ as accurately as possible;
- d can be evaluated and leads to a reward r, with utility $U(r)$ - which exists under the assumption of rationality of the decision-maker;
- The utility is written as $U(\theta, d)$ to stress the dependence on θ and d only;
- When other random factors r are involved in U, we take $U(\theta, d) = \mathrm{E}_{\theta,d}(U(r))$; and
- $U(\theta, d)$ can be seen as a measure of proximity between the proposed estimate d and the true value $h(\theta)$.

The loss function: $L(\theta, d) = -U(\theta, d)$.

In general, the loss function is supposed to be nonnegative $\Rightarrow U(\theta, d) \leqslant 0$. Therefore, there is no decision with infinite utility.

It is generally impossible to uniformly minimize (in d) the loss function $L(\theta, d)$ when θ is unknown.

The frequentist approach proposes to consider instead the average loss (or frequentist risk)

$$R(\theta, \delta) = \mathrm{E}_\theta(L(\theta, \delta(x))) = \int_{\mathcal{X}} L(\theta, \delta(x)) f(x|\theta) dx,$$

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

The function $\delta(x)$, from $\mathcal{X}$ in $\mathcal{D}$, is usually called estimator while the value $\delta(x)$ is called estimate of θ.

Example

Consider x_1 and x_2, two observations from

$$P_\theta(x = \theta - 1) = P_\theta(x = \theta + 1) = 0.5, \qquad \theta \in \mathbb{R}.$$

The parameter of interest in θ ($\mathcal{D} = \Theta$) and it is estimated by estimators δ under the loss

$$L(\theta, \delta) = 1 - \boldsymbol{I}_\theta(\delta) \qquad \text{(0-1 loss)}.$$

Considering the particular estimator

$$\delta_0(x_1, x_2) = \frac{x_1 + x_2}{2},$$

its risk function is

$$R(\theta, \delta_0) = 1 - P_\theta(\delta_0(x_1, x_2) = \theta) = 1 - P_\theta(x_1 \neq x_2) = 0.5.$$

This computation shows that the estimator δ_0 is correct half of the time. The estimator $\delta_1(x_1, x_2) = x_1 + 1$ and $\delta_2(x_1, x_2) = x_2 - 1$ also have risk functions equal to 0.5. Therefore, δ_0, δ_1 and δ_2 cannot be ranked under the 0-1 loss.

The Bayesian approach to Decision Theory integrates on the space Θ since θ is unknown, instead of integrating on the space $\mathcal{X}$ as x is known.

It relies on the posterior expected loss

$$\varrho(\pi, d|x) = \mathrm{E}^{\pi}(L(\theta, d)|x) = \int_{\Theta} L(\theta, d)\pi(\theta|x)d\theta.$$

Given a prior distribution π, it is also possible to define the integrated risk, which is the frequentist risk averaged over the values of θ according to their prior distribution

$$r(\pi, \delta|x) = \mathrm{E}^{\pi}(R(\theta, \delta)) = \int_{\Theta}\int_{\mathcal{X}} L(\theta, \delta(x))f(x|\theta)dx\pi(\theta)d\theta.$$

One particular interest of this second concept is that it associates a real number with every estimator, not a function of θ. It therefore induces a total ordering on the set of estimators, i.e., allows for the direct comparison of estimators.

Theorem

An estimator minimizing the integrated risk $r(\pi, \delta)$ can be obtained by selecting, for every $x \in \mathcal{X}$, the value $\delta(x)$ which minimizes the posterior expected loss, $\varrho(\pi, \delta|x)$ since

$$r(\pi, \delta) = \int_{\mathcal{X}} \varrho(\pi, \delta(x)|x) m(x) dx.$$

Proof: This equality follows directly from Fubini's Theorem since, as $L(\theta, d) \geqslant 0$,

$$\begin{aligned} r(\pi, \delta) &= \int_{\Theta} \int_{\mathcal{X}} L(\theta, \delta(x)) f(x|\theta) dx \pi(\theta) d\theta \\ &= \int_{\mathcal{X}} \int_{\Theta} L(\theta, \delta(x)) f(x|\theta) \pi(\theta) d\theta dx \\ &= \int_{\mathcal{X}} \int_{\Theta} L(\theta, \delta(x)) \pi(\theta|x) d\theta m(x) dx. \end{aligned}$$

Definition (2.3.3)

A Bayes estimator associated with a prior distribution π and a loss function L is any estimator δ^π which minimizes $r(\pi, \delta)$. For every $x \in \mathcal{X}$, it is given by $\delta^\pi(x)$, argument of $\min_d \varrho(\pi, d|x)$. The value $r(\pi) = r(\pi, \delta^\pi)$ is then called the Bayes risk.

The result above is valid for proper and improper priors, as long as the Bayes risk $r(\pi)$ is finite.

Notice that, for strictly convex losses, the Bayes estimators are unique.

Two optimalities: minimaxity and admissibility

We extended the reward space from $\mathcal{R}$ to $\mathcal{P}$, we need to extend the decision space to the set of randomized estimators, taking values in $\mathcal{D}^\star$, space of the probability distributions on $\mathcal{D}$.

To use a randomized estimator $\delta^\star$ means that the action is generated according to the distribution with probability density $\delta^\star(x, .)$, once the observation x has been collected.

The loss of a randomized estimator $\delta^\star$ is then defined as the average loss

$$L(\theta, \delta^\star) = \int_{\mathcal{D}} L(\theta, a)\delta^\star(x, a)da.$$

Theorem

For every prior distribution π on Θ, the Bayes risk on the set of randomized estimators is the same as the Bayes risk on the set of nonrandomized estimators, i.e.,

$$\inf_{\delta\in\mathcal{D}} r(\pi,\delta) = \inf_{\delta^\star\in\mathcal{D}^\star} r(\pi,\delta^\star) = r(\pi).$$

Proof: For every $x \in \mathcal{X}$ and every $\delta^\star \in \mathcal{D}^\star$, we have

$$\begin{aligned}
&\int_\Theta \int_{\mathcal{D}} L(\theta,a)\delta^\star(x,a)da\,\pi(\theta|x)d\theta \\
&= \int_{\mathcal{D}} \int_\Theta L(\theta,a)\pi(\theta|x)d\theta\delta^\star(x,a)da \\
&\geqslant \int_{\mathcal{D}} \inf_a \left\{ \int_\Theta L(\theta,a)\pi(\theta|x)d\theta \right\} \delta^\star(x,a)da \\
&= \varrho(\pi,\delta^\pi|x).
\end{aligned}$$

Definition (2.4.3)

The minimax risk associated with a loss function L is the value

$$\overline{R} = \inf_{\delta \in \mathcal{D}^\star} \sup_\theta R(\theta, \delta) = \inf_{\delta \in \mathcal{D}^\star} \sup_\theta \mathrm{E}_\theta(L(\theta, \delta(x))),$$

and a minimax estimator is any (possibly randomized) estimator δ_0 such that

$$\sup_\theta R(\theta, \delta_0) = \overline{R}.$$

An important difficulty related with minimaxity is that a minimax estimation does not necessarily exist.

Theorem

If $\mathcal{D} \subset \mathbb{R}^k$ is a convex compacted set and if $L(\theta, d)$ is continuous and convex as a function of d for every $\theta \in \Theta$, there exists a nonrandomized minimax estimator.

Lemma

The Bayes risks are always smaller than the minimax risk, i.e.,

$$\underline{R} = \sup_{\pi} r(\pi) = \sup_{\pi} \inf_{\delta \in \mathcal{D}} r(\pi, \delta) \leqslant \overline{R} = \inf_{\delta \in \mathcal{D}^\star} \sup_{\theta} R(\theta, \delta).$$

The first value is called maximin risk and a distribution $\pi^\star$ such that $r(\pi^\star) = \underline{R}$ is called a least *favorable distribution*, when such distributions exist.

Definition

The estimation problem is said to have a value when $\underline{R} = \overline{R}$, i.e.,

$$\sup_{\pi} \inf_{\delta \in \mathcal{D}} r(\pi, \delta) = \inf_{\delta \in \mathcal{D}^\star} \sup_{\theta} R(\theta, \delta).$$

When the problem has a value, some minimax estimators are the Bayes estimators for the least favorable distributions. However, they may be randomized. Therefore, the minimax principle does not always lead to acceptable estimators.

Lemma

If δ_0 is a Bayes estimator with respect to π_0 and if $R(\theta, \delta_0) \leqslant r(\pi_0)$ for every θ in the support of π_0, then δ_0 is minimax and π_0 is the least favorable distribution.

Example

Consider $x \sim \mathcal{B}(n, \theta)$ when θ is to be estimated under the quadratic loss,

$$L(\theta, \delta) = (\delta - \theta)^2.$$

Bayes estimators are then given by posterior expectations and, when

$$\theta \sim \mathcal{B}\mathrm{e}\left(\frac{\sqrt{n}}{2}, \frac{\sqrt{n}}{2}\right),$$

the posterior mean is

$$\delta^\star(x) = \frac{x + \sqrt{n}/2}{n + \sqrt{n}}.$$

Moreover, this estimator has constant risk, $R(\theta, \delta^\star) = 1/4(1 + \sqrt{n})^2$. Therefore, integrating out θ, $r(\pi) = R(\theta, \delta^\star)$ and $\delta^\star$ is minimax (according to the previous Lemma). Notice the difference with the maximum likelihood estimator, $\delta_0(x) = x/n$, for the small values of n, and the unrealistic concentration of the prior around 0.5 for larger values of n.

Lemma

If there exists a sequence (π_n) of proper prior distributions such that the generalized Bayes estimator δ_0 satisfies

$$R(\theta, \delta_0) \leqslant \lim_{n \to \infty} r(\pi_n) < +\infty$$

for every $\theta \in \Theta$, then δ_0 is minimax.

Example

When $x \sim \mathcal{N}(\theta, 1)$, the maximum likelihood estimator $\delta_0(x) = x$ is a generalized Bayes estimator associated with the Lebesgue measure on $\mathbb{R}$ and the quadratic loss. Since $R(\theta, \delta_0) = \mathrm{E}_\theta(x - \theta)^2 = 1$, this risk is the limit of the Bayes risks $r(\pi_n)$ when π_n is equal to $\mathcal{N}(0, n)$, as $r(\pi_n) = n/(n+1)$. Therefore, the maximum likelihood estimator δ_0 is minimax. Note that this argument can be extended directly to the case $x_{\sim}\mathcal{N}_p(\theta, I_p)$ to establish that δ_0 is minimax for every p.

When the space Θ is compact, minimax Bayes rules (or estimators) can be exactly described, owing to the *separated zeros principle* in complex calculus: if $R(\theta, \delta^\pi)$ is not constant and is analytic, the set of θ's where $R(\theta, \delta^\pi)$ is maximal is separated and, in the case of a compact set Θ, is necessarily finite.

Theorem

Consider a statistical problem that simultaneously has a value, a least favorable distribution π_0, and a minimax estimator δ^{π_0}. Then, if $\Theta \subset \mathbb{R}$ is compact and if $R(\theta, \delta^{\pi_0})$ is an analytic function of θ, then either π_0 has a finite support or $R(\theta, \delta^{\pi_0})$ is constant.

Example

Consider $x \sim \mathcal{N}(\theta, 1)$, with $|\theta| \leqslant m$, namely, $\theta \in [-m, m]$. Then, according to previous theorem, least favorable distributions have necessarily a finite support, $\{\pm\theta_i, i \leqslant i \leqslant \omega\}$, with cardinal 2ω and supporting points θ_i depending on m. In fact, the only estimator with constant risk is $\delta_0(x) = x$, which is not minimax in this case. In general, the exact determination of n and of the points θ_i can only be done numerically. For instance, when $m \leqslant 1.06$, the prior distribution with weights 1/2 at $\pm m$ is the unique least favorable distribution. Then, for $1.06 \leqslant m \leqslant 2$, the support of π contains $-m$, 0, and m.

Definition (2.4.19)

An estimator δ_0 is inadmissible if there exists an estimator δ_1 which dominates δ_0, that is, such that, for every θ,

$$R(\theta, \delta_0) \geqslant R(\theta, \delta_1)$$

and, for at least one value θ_0 of the parameter,

$$R(\theta_0, \delta_0) > R(\theta_0, \delta_1).$$

Otherwise, δ_0 is said to be admissible.

However, admissibility alone is not enough to validate the use of an estimator. For instance, constant estimators $\delta(x) = \theta_0$ are usually admissible because they produce the exact value at $\theta = \theta_0$. From a frequentist point of view, it is then important to look for estimators satisfying both optimalities, that is, minimaxity and admissibility.

Proposition

If there exists a unique minimax estimator, this estimator is admissible.

Proof: If $\delta^\star$ is the only minimax estimator, for any estimator $\widetilde{\delta} \neq \delta^\star$,

$$\sup_\theta R(\theta, \widetilde{\delta}) > \sup_\theta R(\theta, \delta^\star).$$

Therefore, $\widetilde{\delta}$ cannot dominate $\delta^\star$.

Notice that the converse to this result is false, since there can exist several minimax admissible estimators.

Proposition

If δ_0 is admissible with constant risk, δ_0 is the unique minimax estimator.

Proof: For any $\theta_0 \in \Theta$, $\sup_\theta R(\theta, \delta_0) = R(\theta_0, \delta_0)$. Therefore, if there exists δ_1 such that $\overline{R} \leqslant \sup_\theta R(\theta, \delta_1) < R(\theta_0, \delta_0)$, δ_0 cannot be admissible. Similarly, if $\overline{R} = \sup_\theta R(\theta, \delta_1) = R(\theta_0, \delta_0)$ and if θ_1 is such that $R(\theta_1, \delta_1) < \overline{R}$, δ_1 dominates δ_0. Therefore, when δ_0 is admissible, the only possible case is that there exists δ_1 such that $R(\theta, \delta_1) = R(\theta, \delta_0)$ for every $\theta \in \Theta$. And this is also impossible when δ_0 is admissible.

Again, notice that the converse of this result is false. There may be minimax estimators with constant risk that are inadmissible: actually, they are certainly inadmissible if there are other minimax estimators.

Proposition

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: Suppose δ^π is inadmissible and consider δ' which uniformly dominates δ^π. Then, for every θ, $R(\theta, \delta') \leqslant R(\theta, \delta^\pi)$ and, in an *open set C* of Θ, $R(\theta, \delta') < R(\theta, \delta^\pi)$. Integrating out this inequality, we derive that

$$r(\theta, \delta') < r(\theta, \delta) = \int_\Theta R(\theta, \delta^\pi)\pi(\theta) d\theta,$$

which is impossible.

Proposition

If the Bayes estimator associated with a prior π is unique, it is admissible.

Even if the Bayes estimator is not unique, it is still possible to exhibit at least one admissible Bayes estimator.

When the loss function is strictly convex, the Bayes estimator is necessarily unique and thus admissible, according to the above proposition.

Example (Continued)

Consider $x \sim \mathcal{B}(n, \theta)$, $L(\theta, \delta) = (\delta - \theta)^2$, $\theta \sim \mathcal{B}\mathrm{e}\left(\frac{\sqrt{n}}{2}, \frac{\sqrt{n}}{2}\right)$ and $\delta^\star(x) = \dfrac{x + \sqrt{n}/2}{n + \sqrt{n}}$. This estimator has constant risk, $R(\theta, \delta^\star) = 1/4(1 + \sqrt{n})^2$. Integrating out θ, $r(\pi) = R(\theta, \delta^\star)$ and $\delta^\star$ is minimax.

Now, the estimator $\delta^\star$ is a (proper) Bayes estimator, therefore admissible, and it has constant risk. Therefore, it is the unique minimax estimator under squared error loss.

Proposition

If a Bayes estimator, δ^π, associated with a (proper or improper) prior π and a strictly convex loss function, is such that the Bayes risk,

$$r(\pi) = \int_\Theta R(\theta, \delta^\pi)\pi(\theta)d\theta,$$

is finite, then δ^π is admissible.

Usual loss functions

The quadratic loss:

$$L(\theta, d) = (\theta - d)^2.$$

- Criticism: penalizes large deviations too heavily.
- Convex loss functions like the above have the incomparable advantage of avoiding the paradox of *risk lovers* and to exclude randomized estimators.
- The quadratic loss may provide a Taylor expansion approximation to more complex symmetric losses.
- The Bayes estimators associated with the quadratic loss are the posterior means.
- Losses leading to posterior means as the Bayes estimators are called proper losses.

Proposition

The Bayes estimator δ^π associated with the prior distribution π and with the quadratic loss $L(\theta, \delta) = (\theta - \delta)^2$, is the posterior expectation

$$\delta^\pi(x) = E^\pi(\theta|x) = \frac{\displaystyle\int_\Theta \theta f(x|\theta)\pi(\theta)d\theta}{\displaystyle\int_\Theta f(x|\theta)\pi(\theta)d\theta}.$$

Proof: Since

$$\mathrm{E}^\pi((\theta-\delta)^2|x) = \mathrm{E}^\pi(\theta^2|x) - 2\delta\mathrm{E}^\pi(\theta|x) + \delta^2,$$

the posterior loss actually attains its minimum at $\delta^\pi(x) = \mathrm{E}^\pi(\theta|x)$.

Corollary

The Bayes estimator δ^π associated with π and with the weighted quadratic loss $L(\theta, \delta) = \omega(\theta)(\theta - \delta)^2$, where $\omega(\theta)$ is a nonnegative function, is

$$\delta^\pi(x) = \frac{E^\pi(\omega(\theta)\theta|x)}{E^\pi(\omega(\theta)|x)}.$$

This corollary exhibits a (weak) duality between loss and prior distribution, in the sense that it is equivalent to estimate θ under $L(\theta, \delta) = \omega(\theta)(\theta - \delta)^2$ with the prior π, or under $L(\theta, \delta) = (\theta - \delta)^2$ with the prior $\pi_\omega(\theta) \propto \pi(\theta)\omega(\theta)$.

Corollary

When $\Theta \in \mathbb{R}^p$, the Bayes estimator δ^π associated with π and with quadratic loss,

$$\delta^\pi(x) = (\theta - \delta)' Q(\theta - \delta),$$

is the posterior mean, $\delta^\pi(x) = E^\pi(\theta|x)$, for every positive-definite symmetric $p \times p$ matrix Q.

The absolute erro loss:

$$L(\theta, d) = |\theta - d|,$$

is an alternative to the quadratic loss in dimension one.

A mixture loss is given by

$$\widetilde{L}(\theta, d) = \begin{cases} (d-\theta)^2, & \text{if } |d-\theta| < k; \\ 2k|d-\theta| - k^2, & \text{otherwise.} \end{cases}$$

A multilinear function (loss)

$$L_{k_1,k_2}(\theta, d) = \begin{cases} k_2(\theta - d), & \text{if } \theta > d; \\ k_1(d-\theta), & \text{otherwise.} \end{cases}$$

Proposition

A Bayes estimator associated with the prior distribution π and the multilinear loss above is a $k_2/(k_1+k_2)$ fractile of $\pi(\theta|x)$.

Proof: The following classical equality

$$\begin{aligned} \mathrm{E}^{\pi}(L_{k_1,k_2}(\theta,d)|x) &= k_1 \int_{-\infty}^{0} (d-\theta)\pi(\theta|x)d\theta + k_2 \int_{0}^{+\infty} (\theta-d)\pi(\theta|x)d\theta \\ &= k_1 \int_{-\infty}^{d} P^{\pi}(\theta<s|x)ds + k_2 \int_{d}^{+\infty} P^{\pi}(\theta>s|x)ds, \end{aligned}$$

is obtained by an integration by parts. Taking derivatives in d, we get

$$k_1 P^{\pi}(\theta<d|x) + k_2 P^{\pi}(\theta>d|x), \quad \text{i.e.,} \quad P^{\pi}(\theta<d|x) = \frac{k_2}{k_1+k_2}.$$

The 0-1 loss:

- Mainly used in the classical approach to hypothesis testing;
- A typical example of a nonquantitative loss.

Example

Consider the test of $H_0; \ \theta \in \Theta_0$ versus $H_1; \ \theta \notin \Theta_0$. Then $\mathcal{D} = \{0, 1\}$, where 1 stands for acceptance of H_0 and 0 for rejection, $I_{\Theta_0}(\theta)$. For the 0-1 loss, i.e.,

$$L(\theta, d) = \begin{cases} 1 - d, & \text{if } \theta \in \Theta_0; \\ d, & \text{otherwise.} \end{cases}$$

the associate risk is

$$R(\theta, \delta) = \mathrm{E}_\theta L(\theta, \delta(x)) = \begin{cases} P_\theta(\delta(x) = 0), & \text{if } \theta \in \Theta_0; \\ P_\theta(\delta(x) = 1), & \text{otherwise.} \end{cases}$$

which are exactly the type-one and type-two errors underlying the Neyman-Pearson theory.

Proposition

The Bayes estimator associated with π and with

$$L(\theta, d) = \begin{cases} 1 - d, & \textit{if } \theta \in \Theta_0; \\ d, & \textit{otherwise.} \end{cases}$$

is

$$\delta^\pi(x) = \begin{cases} 1, & \textit{if } P^\pi(\theta \in \Theta_0|x) > P^\pi(\theta \notin \Theta_0|x); \\ 0, & \textit{otherwise.} \end{cases}$$

i.e., $\delta^\pi(x)$ is equal to 1 if and only if $P^\pi(\theta \in \Theta_0|x) > 1/2$.

Maximum entropy priors

If some characteristics of the prior (moments, quantiles, etc.) are know,

$$\mathrm{E}(g_k(\theta)) = \omega_k, \quad k = 1, \ldots, K,$$

a way to select a prior π satisfying these constraints is the *maximum entropy* method.

In a finite settings, the entropy is defined as

$$\mathcal{E}(\pi) = -\sum_i \pi(\theta_i) \log(\pi(\theta_i)).$$

The maximum entropy distribution (under the restrictions above):

$$\pi^\star(\theta_i) = \frac{\exp\left(\sum_{k=1}^{K} \lambda_k g_k(\theta_i)\right)}{\sum_j \exp\left(\sum_{k=1}^{K} \lambda_k g_k(\theta_j)\right)},$$

where λ_k's are the Lagrange multipliers.

In the continuous we need a reference measure π_0.

The entropy of π is defined as (Kullback-Leibler)

$$\mathcal{E}(\pi) = \mathrm{E}^{\pi_0}\left(\log\left(\frac{\pi(\theta)}{\pi_0(\theta)}\right)\right) = \int \log\left(\frac{\pi(\theta)}{\pi_0(\theta)}\right)\pi_0(d\theta).$$

with the maximum entropy distribution given by

$$\pi^\star(\theta) = \frac{\exp\left(\sum_{k=1}^{K}\lambda_k g_k(\theta)\right)\pi_0(\theta)}{\int \exp\left(\sum_{k=1}^{K}\lambda_k g_k(\eta)\right)\pi_0(d\eta)}.$$

There are problems with the number of constraints.

Example

Consider θ such that $\mathrm{E}^\pi(\theta) = \mu$ and $\pi_0(\theta) \propto 1$, then $\pi^\star(\theta) \propto \exp(\lambda\theta)$ which cannot be normalized. If $\mathrm{E}^\pi(\theta) = \mu$ and $\mathrm{Var}^\pi(\theta) = \sigma^2$, then $\pi^\star(\theta) \propto \exp(\lambda_1\theta + \lambda_2\theta^2)$ which is $\mathcal{N}(\mu, \sigma^2)$.

Parametric approximations

Examples

Let $X_i \sim \text{Bin}(n_i, p_i)$ be the number of passing students in a freshman calculus course of n_i students. Over the previous years, the average of the p_i is 0.70, with variance 0.1. If we assume that the p_i's are all generated according to the same beta distribution, Beta(α, β), the parameters α and β can be estimated through

$$\frac{\alpha}{\alpha+\beta} = 0.7 \quad \text{and} \quad \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = 0.1 \Rightarrow (\alpha, \beta) = (0.77,\ 0.33).$$

It has the advantage of being a conjugate prior.

Conjugate priors

Definition

A family $\mathcal{F}$ of probability distributions on Θ is said to be conjugate (or closed under sampling) for a likelihood function $f(x|\theta)$ if, for every $\pi \in \mathcal{F}$, the posterior distribution $\pi(\theta|x)$ also belongs to $\mathcal{F}$.

Example

- $x|\theta \sim \text{Bin}(n, \theta)$ and $\theta \sim \text{Beta}(a, b)$;
- $x|\theta \sim \text{Poi}(\theta)$ and $\theta \sim \mathcal{G}(a, b)$;
- $x|\theta \sim \mathcal{E}(\theta)$ and $\theta \sim \mathcal{G}(a, b)$;

Exponential families

Definition

Let μ be a σ-finite measure on $\mathcal{X}$, and let Θ be the parameter space. Let C and h be functions, respectively, from $\mathcal{X}$ and Θ to $\mathbb{R}_+$, and let R and T be functions from Θ and $\mathcal{X}$ to $\mathbb{R}^k$. The family of distributions with densities (w.r.t. μ)

$$f(x|\theta) = C(\theta)h(x)\exp\left(R(\theta).T(x)\right)$$

is called an exponential family of dimension k. In the particular case when $\Theta \subset \mathcal{R}^k$, $\mathcal{X} \subset \mathcal{R}^k$ and

$$f(x|\theta) = C(\theta)h(x)\exp(\theta.x),$$

the family is said to be natural.

Theorem (Pitman-Koopman)

If a family of distributions $f(.|\theta)$ *is such that, for a sample size large enough, there exists a sufficient statistic of constant dimension, the family is exponential if the support of* $f(.|\theta)$ *does not depend on* θ*.*

Example

If S is the simplex of $\mathbb{R}^k$,

$$S = \left\{ \omega = (\omega_1, \ldots, \omega_k);\ \sum_{i=1}^{k} \omega_i = 1;\ \omega_i > 0 \right\},$$

the Dirichlet distribution on S, $D_k(\alpha_1, \ldots, \alpha_k)$, is an extension of the beta distribution, which is defined by

$$f(p|\alpha) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_k)} \prod_{i=1}^{k} p_i^{\alpha_i - 1} \boldsymbol{I}_S(p),$$

where $p = (p_1, \ldots, p_k)$. Since

$$f(p|\alpha) = C(\alpha) h(p) \exp \left\{ \sum_{i=1}^{k} \alpha_i \log(p_i) \right\},$$

the Dirichlet distribtuions constitute a natural exponential family for $T(p) = (\log(p_1), \ldots, \log(p_k))$.

Example

Let $x \sim \mathcal{N}(\theta, \sigma^2 I_p)$. Then

$$\begin{aligned} f(x|\theta,\sigma) &= \frac{1}{\sigma^p}\frac{1}{(2\pi)^{p/2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{p}(x_i-\theta_i)^2\right\} \\ &= C(\theta,\sigma)h(x)\exp\{x.(\theta/\sigma)+||x||^2(-1/2\sigma^2)\} \end{aligned}$$

and the normal distribution belongs to an exponential family with natural parameters θ/σ^2 and $-1/2\sigma^2$. Similarly, if $x_1, \ldots, x_n \sim \mathcal{N}(\theta, \sigma^2 I_p)$, the joint distribution satisfies

$$\begin{aligned} f(x|\theta,\sigma) &= C'(\theta,\sigma)h'(x_1,\ldots,x_n) \\ &\times \exp\left\{n\overline{x}.(\theta/\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}||x_i-\overline{x}||^2\right\} \end{aligned}$$

and the statistics $(\overline{x}, \sum_{i=1}^{n}||x_i-\overline{x}||^2)$ is sufficient for all $n \geqslant 2$.

Definition

Let $f(x|\theta) = C(\theta)h(x)\exp\{\theta.x\}$ be a natural exponential family. The natural parameter space is

$$N = \left\{\theta;\ \int e^{\theta.x} h(x) d\mu(x) < +\infty\right\}.$$

The family is said to be regular if N is an open set and minimal if $\dim(N) = \dim(K) = k$, where K is the closure of the convex envelope of the support of μ.

Natural exponential families can also be rewritten under the form

$$f(x|\theta) = h(x)\exp\{\theta.x - \psi(\theta)\}$$

and $\psi(\theta)$ is called the cumulant generating function.

Lemma

If $\theta \in N^0$, the interior set of N, the cumulant generating function ψ is $\mathcal{C}^\infty$ and

$$E_\theta(x) = \nabla\psi(\theta), \qquad cov(x_i, x_j) = \frac{\partial^2\psi}{\partial\theta_i\partial\theta_j}(\theta),$$

where ∇ denotes the gradient operator.

Example

Let $x \sim \text{Poi}(\lambda)$. Then

$$f(x|\lambda) = \exp(-\lambda)\frac{\lambda^x}{x!} = \frac{1}{x!}\exp(\theta.x - \exp(\theta)),$$

and $\psi(\theta) = \exp(\theta)$ for the natural parameter $\theta = \log(\lambda)$. Therefore, $\mathrm{E}_\lambda(x) = \exp(\theta) = \lambda$ and $\mathrm{Var}_\lambda(x) = \lambda$.

Consider $f(x|\theta) = h(x) \exp(\theta.x - \psi(\theta))$.

Proposition

A conjugate family for $f(x|\theta)$ is given by

$$\pi(\theta|\mu, \lambda) = K(\mu, \lambda) \exp(\theta.\mu - \lambda\psi(\theta)),$$

where $K(\mu, \lambda)$ is the normalizing constant of the density. The corresponding posterior distribution is $\pi(\theta|\mu + x, \lambda + 1)$.

The measure defined above is σ-finite; it induces a probability distribution on Θ if and only if

$$\lambda > 0 \quad \text{and} \quad \frac{\mu}{\lambda} \in N^0.$$

Proposition

If Θ is an open set in $\mathbb{R}^k$ and θ has the prior distribution

$$\pi_{\lambda,x_0}(\theta) \propto \exp(\theta.x_0 - \lambda\psi(\theta))$$

with $x_0 \in \mathcal{X}$, then

$$E^\pi(\xi(\theta)) = E^\pi(\nabla\psi(\theta)) = \frac{x_0}{\lambda}.$$

Therefore, if $x_1, \ldots, x_n$ are i.i.d. $f(x|\theta)$,

$$E^\pi(\xi(\theta)|x_1, \ldots, x_n) = \frac{x_0 + n\bar{x}}{\lambda + n}.$$

This result is well known for the normal distributions and can be generalized for all exponential families.

Lemma

Let $\mathcal{F}$ be the natural conjugate family of an exponential family. Then the set of mixtures of N conjugate distributions,

$$\widetilde{\mathcal{F}}_N = \left\{ \sum_{i=1}^{N} \omega_i \pi(\theta|\lambda_i, \mu_i); \ \sum_{i=1}^{N} \omega_i = 1, \ \omega_i > 0 \right\},$$

is also a conjugate family. Moreover, if

$$\pi(\theta) = \sum_{i=1}^{N} \omega_i \pi(\theta|\lambda_i, \mu_i), \quad \textit{then}$$

$$\pi(\theta|x) = \sum_{i=1}^{N} \omega_i'(x) \pi(\theta|\lambda_i + 1, \mu_i + x), \quad \textit{with}$$

$$\omega_i'(x) = \frac{\omega_i K(\mu_i, \lambda_i)/K(\mu_i + x, \lambda_i + 1)}{\sum_{j=1}^{N} \omega_j K(\mu_j, \lambda_j)/K(\mu_j + x, \lambda_j + 1)}.$$

Noninformative prior distributions

Noninformative priors cannot be expected to represent exactly total ignorance about the problem at hand, but should rather be taken as reference or default priors, upon which everyone could fall back when the prior information is missing. In this light, some noninformative priors may be more useful or more efficient than others, but they cannot be said to be less informative than others.

Invariant priors

Example

The family $f(x-\theta)$ is translation invariant, i.e., $y = x - x_0$ has a distribution in the same family for every x_0, $f(y-(\theta - x_0))$; θ is then said to be a *location parameter* and an invariance requirement is that the prior distribution should be translation invariant, i.e., satisfy

$$\pi(\theta) = \pi(\theta - \theta_0)$$

for every θ_0. The solution is $\pi(\theta) = c$, the uniform distribution on Θ.

Example

If the distribution family is parametrized by a *scale parameter*, i.e., is of the form $(1/\sigma)f(x/\sigma)$, $(\sigma > 0)$, it is *scale-invariant*, i.e., $y = x/\sigma \sim f(y)$. A scale invariant prior π satisfies $\pi(A) = \pi(A/c)$ for every measurable set A in $(0, +\infty)$ and $c > 0$, i.e.

$$\pi(\sigma) = \frac{1}{c}\pi\left(\frac{\sigma}{c}\right).$$

This implies $\pi(\sigma) = \alpha/\sigma$, where α is a constant. Therefore, the invariant measure is not constant anymore.

The Jeffreys prior

The Jeffreys noninformative prior distributions are based on Fisher information, given by

$$I(\theta) = \mathrm{E}_\theta(S(\theta,x)S(\theta,x)') \quad \text{where} \quad S(\theta,x) = \frac{\partial \log(f(x|\theta))}{\partial\theta}.$$

Under some regularity assumptions, this information can also be written as

$$I(\theta) = \mathrm{E}_\theta\left(\frac{\partial^2 \log(f(x|\theta))}{\partial\theta\partial\theta'}\right).$$

The Jeffreys prior distribution is

$$\pi^\star(\theta) \propto [\det(I(\theta))]^{1/2}.$$

If $f(x|\theta) = h(x)\exp(\theta.x - \psi(\theta))$, then $I(\theta) = \nabla\nabla'\psi(\theta)$ and

$$\pi^\star(\theta) \propto \left[\prod_{i=1}^{k}\psi_{ii}''(\theta)\right]^{1/2} \quad \text{where} \quad \psi_{ii}''(\theta) = \frac{\partial^2}{\partial\theta_i^2}\psi(\theta).$$

The reference prior

When $x \sim f(x|\theta)$ and $\theta = (\theta_1, \theta_2)$, where θ_1 is the parameter of interest, the reference prior is obtained by first defining $\pi(\theta_2|\theta_1)$ as the Jeffreys prior associated with $f(x|\theta)$ when θ_1 is fixed, then deriving the marginal distribution

$$\widetilde{f}(x|\theta_1) = \int f(x|\theta_1, \theta_2)\pi(\theta_2|\theta_1) d\theta_2,$$

and computing the Jeffreys prior $\pi(\theta_1)$ associated with $\widetilde{f}(x|\theta_1)$.

The principle behind the reference prior is therefore to eliminate the nuisance parameter by using a Jeffreys prior where the parameter of interest remains fixed.

Example

The Neyman–Scott (1948) problem is related to the observation of x_{ij}'s distributed from $\mathcal{N}(\mu_i, \sigma^2)$, $i = 1, \ldots, n$, $j = 1, 2$. The usual Jeffreys prior for this model is $\pi(\mu_1, \ldots, \mu_n, \sigma) = \sigma^{-(n+1)}$ and an inconsistency arises because $\mathrm{E}(\sigma^2 | x_{11}, \ldots, x_{n2}) = s^2/(2n-2)$, where

$$s^2 = \sum_{i=1}^{n} \frac{(x_{i1} - x_{i2})^2}{2},$$

and this posterior expectation converges to $\sigma^2/2$ with n.

The reference prior associated with $\theta_1 = \sigma$ and $\theta_2 = (\mu_1, \ldots, \mu_n)$ gives a flat prior for $\pi(\theta_2|\theta_1)$, since θ_2 is a location parameter. Then

$$\widetilde{f}(x|\theta_1) = \prod_{i=1}^{n} \exp\left(-\frac{(x_{i1} - x_{i2})^2}{4\sigma^2}\right) \frac{1}{2\sigma\sqrt{2\pi}},$$

is a scale family and $\pi(\sigma) = 1/\sigma$. Therefore, $\mathrm{E}(\sigma^2 | x_{11}, \ldots, x_{n2}) = s^2/(n-2)$, which is consistent.

Maximum a Posteriori Estimator

A possible estimator of θ based on $\pi(\theta|x)$ is the maximum a posteriori (MAP) estimator, defined as the posterior mode.

The MAP estimator maximizes $f(x|\theta)\pi(\theta)$, thus bypassing the computation of the marginal distribution $m(x)$.

It is associated with the 0-1 loss. In continuous settings, since, for every $\delta \in \Theta$,

$$\int_\Theta \mathbb{I}_{\delta \neq \theta}(\theta)\pi(\theta|x)d\theta = 1,$$

the 0-1 loss must be replaced by a sequence of losses, $L_\varepsilon(d,\theta) = \mathbb{I}_{||\theta - d|| > \varepsilon}$, and the MAP estimate is then the limit of the Bayes estimates associated with L_ε, when ε goes to 0.

This natural estimator can be expressed as a penalized maximum likelihood estimator in the classical sense.

Notice that the asymptotic optimality properties of the regular maximum likelihood estimator (consistency, efficiency) are preserved for these Bayesian extensions, under a few regularity conditions on f and π

Example

Consider $x \sim \text{Bin}(n, \theta)$ and $\theta \sim \text{Beta}(1/2, 1/2)$ (Jeffreys's prior $\pi^\star$).

Two other noninformative distributions have been proposed in the literature:

$$\pi_1(\theta) = 1 \quad \text{and} \quad \pi_2(\theta) \propto \theta^{-1}(1-\theta)^{-1}.$$

The corresponding MAP estimators are then, for $n > 2$,

$$\begin{aligned}
\delta^\star(x) &= \max\left(\frac{x - 1/2}{n-1}, 0\right), \\
\delta_1(x) &= \frac{x}{n}, \quad \text{and} \\
\delta_2(x) &= \max\left(\frac{x-1}{n-1}, 0\right).
\end{aligned}$$

When $n = 1$, $\delta^\star$ and δ_2 are equal to δ_1.

For $n = 2$ and $x = 1$, the estimator δ_2 is also equal to δ_1, which is the regular maximum likelihood estimator.

Notice that, when n is large, the three estimators are indeed equivalent.

Precision of the Bayes Estimator

Since the whole posterior distribution $\pi(\theta|x)$ is available, it is possible to associate to an estimator $\delta^\pi(x)$ of $h(\theta)$ an evaluation of the precision of the estimation.

For instance, the posterior squared error,

$$\mathrm{E}^\pi([\delta^\pi(x) - h(\theta)]^2|x),$$

equal to $\mathrm{var}^\pi(h(\theta)|x)$ when $\delta^\pi(x) = \mathrm{E}^\pi(h(\theta)|x)$.

In a multidimensional setting, the covariance matrix characterizes the performances of estimators.

These additional indications provided by the posterior distribution illustrate the operational advantage of the Bayesian approach, since the classical approach usually has difficulties motivating the choice of these evaluations.

Moreover, Bayesian evaluation measures are always conditional, while the frequentist approach usually relies on upper bounds through the minimax principle, since the parameter θ is unknown.

Prediction

If $x \sim f(x|\theta)$ and $z \sim g(z|x, \theta)$, where z does not necessarily depend on x, the predictive distribution of z after the observation of x is given by

$$g^{\pi}(z|x) = \int_{\Theta} g(z|x, \theta)\pi(\theta|x)d\theta.$$

It is possible to use the equation above to derive the predictive mean and variance of the random variable z.

Example

A particular case of AR(1) model defines the distribution of a stochastic process $\{x_t\}_{t=1}^{T}$ by a linear dynamic representation conditional on the previous variable x_{t-1}, as

$$x_t = \phi x_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2) \quad \text{i.i.d.}$$

Given a sequence of observations till time $T-1$, $x_{1:(T-1)} = (x_1, \ldots, x_{T-1})$, the predictive distribution of x_T is then given by

$$x_T|x_{1:(T-1)} \sim \int \frac{1}{\sqrt{2\pi}}\sigma^{-1} \exp\left(-\frac{1}{2\sigma^2}(x_T - \phi x_{T-1})\right) \pi(\phi, \sigma|x_{1:(T-1)})d\phi d\sigma.$$

Bayes estimators

Given a loss function $L(\theta, \delta)$ and a prior distribution (or a measure) π, the Bayes rule $\delta^\pi(x)$ is solution of

$$\min_\delta \mathrm{E}^\pi(L(\theta, \delta)|x).$$

Depending on the complexity of the loss L and the posterior distribution $\pi(\theta|x)$, the estimator δ^π will be determined analytically or numerically, similar to maximum likelihood estimation.

The solutions associated with classical losses are formally known and correspond to the natural indicators associated with a distribution (mean, median, mode, quantiles, etc.).

For instance, the Bayes estimator associated with the quadratic loss is the posterior mean. This formal derivation of classical Bayes estimators does not always avoid a numerical approximation of the estimators.

Lemma

Consider $f(x|\theta) = h(x)\exp(\theta.x - \psi(\theta))$, a distribution from an exponential family. For every prior distribution π, the posterior mean of θ is given by

$$\delta^\pi(x) = \nabla \log(m_\pi(x)) - \nabla \log(h(x)),$$

where ∇ denotes the gradient operator and m_π is the marginal distribution associated with π.

Proof: The posterior expectation is given by

$$\begin{aligned}
\mathrm{E}^\pi(\theta_i|x) &= \frac{\displaystyle\int_\Theta \theta_i h(x)\exp(\theta.x - \psi(\theta))\pi(\theta)d\theta}{m_\pi(x)} \\
&= \left[\frac{\partial}{\partial x_i}\int_\Theta h(x)\exp(\theta.x - \psi(\theta))\pi(\theta)d\theta\right]\frac{1}{m_\pi(x)} - \left[\frac{\partial}{\partial x_i}h(x)\right]\frac{1}{h(x)} \\
&= \frac{\partial}{\partial x_i}[\log(m_\pi(x)) - \log(h(x))].
\end{aligned}$$

It is satisfied for every π; it appears as the dual result of the derivation of the moments of $f(x|\theta)$ from the derivative of ψ in exponential families.

Tabela 1: The Bayes estimators of the parameter θ under quadratic loss for conjugate distributions in the usual exponential families.

Distribution	Conjugate prior	Posterior mean
$\mathcal{N}(\theta, \sigma^2)$	$\mathcal{N}(\mu, \tau^2)$	$\dfrac{\mu\sigma^2 + \tau^2 x}{\sigma^2 + \tau^2}$
$\mathcal{P}(\theta)$	$\mathcal{G}(\alpha, \beta)$	$\dfrac{\alpha + x}{\beta + 1}$
$\mathcal{G}(\nu, \theta)$	$\mathcal{G}(\alpha, \beta)$	$\dfrac{\alpha + \nu}{\beta + x}$
$\mathcal{B}(n, \theta)$	$\mathcal{B}e(\alpha, \beta)$	$\dfrac{\alpha + x}{\alpha + \beta + n}$
$\mathcal{N}\mathcal{B}(n, \theta)$	$\mathcal{B}e(\alpha, \beta)$	$\dfrac{\alpha + n}{\alpha + \beta + x + n}$
$\mathcal{M}_k(n, \theta_1, \ldots, \theta_k)$	$\mathcal{D}ir(\alpha_1, \ldots, \alpha_k)$	$\dfrac{\alpha_i + x_i}{n + \sum_j \alpha_j}$
$\mathcal{N}(\mu, 1/\theta)$	$\mathcal{G}(\alpha/2, \beta/2)$	$\dfrac{\alpha + 1}{\beta + (\mu - x)^2}$

Example

If $x_1, \ldots, x_n$ are independent observations from $\mathcal{NB}(m, \theta)$ and if $\theta \sim \mathcal{B}e(\alpha, \beta)$, the posterior distribution of θ is the beta distribution $\mathcal{B}e(\alpha + nm, \beta + \sum_{i=1}^n x_i)$ and

$$\delta^\pi(x_1, \ldots, x_n) = \frac{\alpha + nm}{\beta + \sum_{i=1}^n x_i}.$$

This result is a straightforward consequence of $\sum_{i=1}^n x_i \sim \mathcal{NB}(mn, \theta)$.

Example

Consider n observations $x_1, \ldots, x_n$ from $\mathcal{U}([0,\theta])$ and $\theta \sim \mathcal{P}a(\beta, \alpha)$. Then

$$\theta | x_1, \ldots, x_n \sim \mathcal{P}a(\max(\beta, x_1, \ldots, x_n), \alpha + n)$$

and

$$\delta^\pi(x_1, \ldots, x_n) = \frac{\alpha + n}{\alpha + n - 1} \max(\beta, x_1, \ldots, x_n).$$

Therefore, compared with the maximum likelihood estimator,

$$\delta_0(x_1, \ldots, x_n) = \max(x_1, \ldots, x_n),$$

the Bayes estimator gives a more optimistic estimation of θ, since

$$\frac{\alpha + n}{\alpha + n - 1} > 1.$$

Example

Consider $x \sim \mathcal{G}(\nu, \theta)$ where the shape parameter ν is known, and $\theta \sim \mathcal{G}(\alpha, \beta)$. The parameter of interest is $1/\theta$. Under the quadratic loss

$$L(\theta, \delta) = \left(\delta - \frac{1}{\theta}\right)^2,$$

the Bayes estimator is then

$$\begin{aligned}\delta_1^\pi(x) &= \frac{(\beta + x)^{\alpha+\nu}}{\Gamma(\alpha+\nu)} \int_0^{+\infty} \frac{1}{\theta} \theta^{\alpha+\nu-1} \exp(-(\beta + x)\theta) d\theta \\ &= \frac{\beta + x}{\alpha + \nu - 1}.\end{aligned}$$

Under a renormalized (or weighted) quadratic loss,

$$L(\theta, \delta) = \omega(\theta) ||\delta - \theta||_Q^2,$$

where Q is a $p \times p$ nonnegative symmetric matrix, the corresponding Bayes estimator is

$$\delta^\pi(x) = \frac{\mathrm{E}(\theta\omega(\theta))}{\mathrm{E}(\omega(\theta))}.$$

Example (continued)

A scale-invariant loss does not depend on the unit of measurement and may be more relevant for the estimation of $1/\theta$. For instance, the loss

$$L(\theta, \delta) = \theta^2 \left(\delta - \frac{1}{\theta}\right)^2,$$

gives the Bayes estimator

$$\begin{aligned}
\delta_2^\pi(x) &= \frac{\mathrm{E}(\theta^2/\theta|x)}{\mathrm{E}(\theta^2|x)} \\
&= \frac{\displaystyle\int_0^{+\infty} \theta^{\alpha+\nu+1-1} \exp(-(\beta+x)\theta) d\theta}{\displaystyle\int_0^{+\infty} \theta^{\alpha+\nu+2-1} \exp(-(\beta+x)\theta) d\theta} \\
&= \frac{\beta+x}{\alpha+\nu-1} = \frac{\alpha+\nu-1}{\alpha+\nu+1}\delta_1^\pi(x).
\end{aligned}$$

For a given loss function, $L(\theta, \delta)$, it may also be of interest to assess the performance of the Bayes estimator $\delta^{\pi}(x)$.

This evaluation can be perceived from a decision-theoretic point of view as the estimation of the loss $L(\theta, \delta^{\pi}(x))$ by $\gamma(x)$ under a loss function, like

$$\widetilde{L}(\theta, \delta^{\pi}, \gamma) = [\gamma(x) - L(\theta, \delta^{\pi}(x))]^2.$$

Proposition

The Bayes estimator of the loss $L(\theta, \delta^{\pi}(x))$ *under* $\widetilde{L}(\theta, \delta^{\pi}, \gamma) = [\gamma(x) - L(\theta, \delta^{\pi}(x))]^2$ *for the prior distribution* π *is*

$$\gamma^{\pi}(x) = E(L(\theta, \delta^{\pi}(x))|x).$$

No model is right, but some models are less wrong than others.

Consider a statistical model $f(x|\theta)$ with $\theta \in \Theta_0$. Given a subset of interest of $\Theta_0 \subset \Theta$, the question to be answered is

$$H_0 : \ \theta \in \Theta_0,$$

usually called the null hypothesis.

If additional information such as $\theta \in \Theta_0 \cup \Theta_1 \neq \Theta$ is available, then we define the alternative hypothesis against as

$$H_1 : \ \theta \in \Theta_1.$$

Under this formalization, every test procedure φ appears as an estimator of $\mathbb{I}_{\Theta_0}(\theta)$ and we only need a loss function $L(\theta, \varphi)$ to derive the Bayes estimators.

The loss function proposed by Neyman and Pearson is the 0-1 loss

$$L(\theta, \varphi) = \begin{cases} 1, & \text{if } \varphi \neq \mathbb{I}_{\Theta_0}(\theta); \\ 0, & \text{otherwise.} \end{cases}$$

For this loss, the Bayesian solution is

$$\varphi^\pi(x) = \begin{cases} 1, & \text{if } P^\pi(\theta \in \Theta_0 | x) > P^\pi(\theta \in \Theta_0^c | x); \\ 0, & \text{otherwise.} \end{cases}$$

This estimator is easily justified on an intuitive basis since it chooses the hypothesis with the largest posterior probability.

A generalization of the above loss is to penalize differently errors when the null hypothesis is true or false. The weighted 0-1 losses

$$L(\theta, \varphi) = \begin{cases} 0, & \text{if } \varphi = \mathbb{I}_{\Theta_0}(\theta); \\ a_0, & \text{if } \theta \in \Theta_0 \text{ and } \varphi = 0, \\ a_1, & \text{if } \theta \notin \Theta_0 \text{ and } \varphi = 1, \end{cases}$$

are called "$a_0 - a_1$" for obvious reason.

Proposition

Under the "$a_0 - a_1$", the Bayes estimator associated with a prior distribution π is

$$\varphi^\pi(x) = \begin{cases} 1, & \textit{if } P^\pi(\theta \in \Theta_0 | x) > \dfrac{a_1}{a_0 + a_1}; \\ 0, & \textit{otherwise.} \end{cases}$$

Proof: Since the posterior loss is

$$\begin{aligned} L(\pi, \varphi | x) &= \int_\Theta L(\theta, \varphi) \pi(\theta | x) d\theta \\ &= a_0 P^\pi(\theta \in \Theta_0 | x) \mathbb{I}_{\{0\}}(\varphi) + a_1 P^\pi(\theta \notin \Theta_0 | x) \mathbb{I}_{\{1\}}(\varphi), \end{aligned}$$

the Bayes estimator can be derived directly.

The Bayes Factor

Definition

The Bayes factor is the ratio of the posterior probabilities of the null and the alternative hypotheses over the ratio of the prior probabilities of the null and the alternative hypotheses, i.e.,

$$B_{01}^{\pi}(x) = \frac{P(\theta \in \Theta_0|x)}{P(\theta \in \Theta_1|x)} \bigg/ \frac{\pi(\theta \in \Theta_0)}{\pi(\theta \in \Theta_1)}.$$

In the particular case where $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$, the Bayes factor simplifies to the usual likelihood ratio

$$B_{01}^{\pi}(x) = \frac{f(x|\theta_0)}{f(x|\theta_1)}.$$

The Bayes factor can be perceived as a Bayesian likelihood ratio since, if π_0 is the prior distribution under H_0 and π_1 the prior distribution under H_1, then

$$B_{01}^{\pi}(x) = \frac{\displaystyle\int_{\Theta_0} f(x|\theta_0)\pi_0(\theta)d\theta}{\displaystyle\int_{\Theta_1} f(x|\theta_1)\pi_1(\theta)d\theta} = \frac{m_0(x)}{m_1(x)}.$$

As indicated above, the Bayes factor is, from a decision-theoretic point of view, completely equivalent to the posterior probability of the null hypothesis as, under "$a_0 - a_1$" loss, H_0 is accepted when

$$B_{01}^{\pi}(x) = \frac{a_1}{a_2}\Big/\frac{\eta_0}{\eta_1} = \frac{a_1\eta_1}{a_0\eta_0},$$

where

$$\eta_0 = \pi(\theta \in \Theta_0) \quad \text{and} \quad \eta_1 = \pi(\theta \in \Theta_1) = 1 - \eta_0.$$

This alternative thus provides an illustration of the duality existing between loss and prior distribution.

It shows that it is equivalent to weight both hypotheses equally, $\eta_0 = \eta_1 = 1/2$, and to modify the error penalties into $a_i' = a_i\eta_i$ $(i = 0, 1)$ or to penalize similarly both types of errors $(a_0 = a_1 = 1)$, when the prior distribution incorporates the actual weights in the weighted prior probabilities,

$$\eta_0' = \frac{a_0\eta_0}{a_0\eta_0 + a_1\eta_1} \quad \text{and} \quad \eta_1 = \frac{a_1\eta_1}{a_0\eta_0 + a_1\eta_1}.$$

A scale to judge the evidence in favor of or against H_0 brought by the data, outside a true decision-theoretic setting:

- if $\log_{10}(B_{10}^{\pi})$ varies between 0 and 0.5, the evidence against H_0 is poor,
- if it is between 0.5 and 1, it is is substantial,
- if it is between 1 and 2, it is strong, and
- if it is above 2 it is decisive.

Point-null hypotheses

Considering the point-null hypothesis $H_0: \ \theta = \theta_0$ and $\pi_0(\theta) = \eta_0 \mathbb{I}_{\Theta_0}(\theta) + (1 - \eta_0) g_1(\theta)$. The posterior probability of H_0 is given by

$$\pi(\Theta_0|x) = \frac{f(x|\theta_0)\eta_0}{\int f(x|\theta)\pi(\theta)d\theta} = \frac{f(x|\theta_0)\eta_0}{f(x|\theta_0)\eta_0 + (1-\eta_0)m_1(x)}$$

where

$$g_1(\theta) \propto \pi(\theta)\mathbb{I}_{\Theta_1}(\theta) \quad \text{and} \quad m_1(x) = \int_{\Theta_1} f(x|\theta)g_1(\theta)d\theta.$$

Pseudo-Bayes factors

Definition

Given an improper prior π, a sample $(x_1, \dots, x_n)$ is a training sample if the corresponding posterior $\pi(|x_1, \dots, x_n)$ is proper and is a minimal training sample if no subsample is a training sample.

When facing a hypothesis H_0 with a prior distribution π_0, with a broader alternative H_1 with a prior distribution π_1, if the minimal training sample under H_1 is such that $\pi_0(x_{(-\ell)})$ is also proper, the pseudo-Bayes factor

$$B_{10}^{(\ell)} = \frac{\displaystyle\int_{\Theta_1} f_1(x_{(-\ell)}|\theta_1)\pi_1(\theta_1|x_{(\ell)})d\theta_1}{\displaystyle\int_{\Theta_0} f_0(x_{(-\ell)}|\theta_0)\pi_0(\theta_0|x_{(\ell)})d\theta_0}$$

is then independent from the normalizing constants used in both π_0 and π_1.

Lemma

In the case of independent distributions, the pseudo-Bayes factor can be written as

$$\begin{aligned}
B_{10}^{(\ell)} &= B_{10}(x) \times B_{01}(x_{(\ell)}), \quad \textit{with} \\
B_{10}(x) &= \frac{\displaystyle\int_{\Theta_1} f_1(x|\theta_1)\pi_1(\theta_1)d\theta_1}{\displaystyle\int_{\Theta_0} f_0(x|\theta_0)\pi_0(\theta_0)d\theta_0} \\
B_{01}(x_{(\ell)}) &= \frac{\displaystyle\int_{\Theta_0} f_0(x_{(\ell)}|\theta_0)\pi_0(\theta_0)d\theta_0}{\displaystyle\int_{\Theta_1} f_1(x_{(\ell)}|\theta_1)\pi_1(\theta_1)d\theta_1}
\end{aligned}$$

A way to remove the dependence on the training sample is to average the different pseudo-Bayes factors over all the possible training samples $x_{(\ell)}$. The next difficulty is to decide what kind of averaging should be used here.

- the arithmetic intrinsic Bayes factor

  $$B_{10}^{A} = \frac{1}{L}\sum_{x_{(\ell)}} B_{10}^{(\ell)} = B_{10}(x)\frac{1}{L}\sum_{x_{(\ell)}} B_{01}(x_{(\ell)}),$$

  where L is the number of different training samples;

- the geometric intrinsic Bayes factor

  $$B_{10}^{G} = \exp\left\{\frac{1}{L}\sum_{x_{(\ell)}} \log\left(B_{10}^{(\ell)}\right)\right\} = B_{10}(x)\exp\left\{\frac{1}{L}\sum_{x_{(\ell)}} \log\left(B_{01}(x_{(\ell)})\right)\right) ;$$

- the median intrinsic Bayes factor

  $$B_{10}^{M} = \text{med}\left(B_{10}^{(\ell)}\right) = B_{10}(x)\text{med}\left(B_{01}(x_{(\ell)})\right).$$

Another idea is to use a fraction $0 < b < 1$ of the likelihood to “properize” the prior:

$$\int_{\Theta_0} [f_0(x|\theta_0)]^b \pi_0(\theta_0) d\theta_0 < \infty \quad \text{and} \quad \int_{\Theta_1} [f_1(x|\theta_1)]^b \pi_1(\theta_1) d\theta_1 < \infty.$$

The remaining fraction $(1 - b)$ of the likelihood is then used to run the test.

$$B_{10}^F = \frac{\int_{\Theta_1} [f_1(x|\theta_1)]^{1-b} \pi_1^b(\theta_1|x) d\theta_1}{\int_{\Theta_0} [f_0(x|\theta_0)]^{1-b} \pi_0^b(\theta_0|x) d\theta_0} = B_{10}(x) \frac{\int_{\Theta_0} [f_0(x|\theta_0)]^b \pi_0(\theta_0) d\theta_0}{\int_{\Theta_1} [f_1(x|\theta_1)]^b \pi_1(\theta_1) d\theta_1}.$$

where $\pi_0^b(\theta_0|x)$ and $\pi_1^b(\theta_1|x)$ denote the pseudo-posteriors associated with $[f_0(x|\theta_0)]^b$ and $[f_1(x|\theta_1)]^b$, respectively.

For exponential families, the fraction b clearly corresponds to a training sample size, since for n observations from an exponential family with sufficient statistic T

$$[\exp\{\theta.nT(x) - n\Psi(\theta)\}]^b = \exp\{\theta.[bn]T(x) - [bn]\Psi(\theta)\}.$$

There are enough difficulties with these pseudo-Bayes factors to make us question their use in testing and model choice problems:

- Bayes factors, when associated with proper priors, do satisfy some *coherence* properties such as

  $$B_{12} = B_{10} B_{02} \quad \text{and} \quad B_{01} = 1/B_{10}.$$

  Most pseudo-Bayes factors do not, even though the fractional Bayes factor satisfies $B_{01}^F = 1/B_{10}^F$.
- When the pseudo-Bayes factors can be expressed as true Bayes factors, the corresponding intrinsic priors are not necessarily appealing, and these priors do depend on the choice of the improper reference priors π_0 and π_1, hence are hardly intrinsic.
- The pseudo-Bayes factors may also exhibit a bias in favor of one of the hypotheses, in the sense that they can be expressed as a true Bayes factor multiplied by a certain factor.
- Most often, the pseudo-Bayes factors do not correspond to any true Bayes factor, and they may give strongly biased solutions. For instance, the arithmetic intrinsic Bayes factors do not have intrinsic priors for most one-sided testing problems.

- Pseudo-Bayes factors may simply not exist for a whole class of models, (e.g. mixture of normals model).
- There are many ways of defining pseudo-Bayes factors and, while most are arguably logical, there is no coherent way of ordering them. Pseudo-Bayes factors, as defined here, do agree with the Likelihood Principle, but the multiplication of possible answers, even if those are close, is not a good signal to users. Similarly, there is no clear-cut procedure for the choice of the fraction b in the fractional Bayes factors, since the minimal training sample size is not always clearly defined.
- The issue of computing pseudo-Bayes factors has not been mentioned so far. But each Bayes factor $B_{10}^{(\ell)}$ may be a complex integral and the derivation of the averaged intrinsic Bayes factor may involve $\binom{n}{m}$ integrals of this kind, if m is the minimal training sample size. Fractional Bayes factors are easier to compute in exponential settings, but other distributions are much more difficult to handle.

A second decision-theoretic approach

The testing problem formalized by Neyman and Pearson can be expressed as estimating the indicator function $\mathbb{I}_{\Theta_0}(\theta)$ under the 0-1 loss or, equivalently, the absolute error loss

$$L_1(\theta, \varphi) = |\varphi - \mathbb{I}_{\Theta_0}(\theta)|.$$

We turn to a less restrictive theory, according to which estimators take values in $\mathcal{D} = [0, 1]$ and can be considered as indicators of the degree of evidence in favor of H_0.

The $L_1(\theta, \varphi)$ loss is too similar to the 0-1 loss function as it provides the same Bayes procedures

$$\varphi^\pi(x) = \begin{cases} 1, & \text{if} \quad P^\pi(\theta \in \Theta_0|x) > P^\pi(\theta \notin \Theta_0|x) \\ 0, & \text{otherwise.} \end{cases}$$

In the opposite, strictly convex losses, such as the quadratic loss

$$L_2(\theta, \varphi) = (\varphi - \mathbb{I}_{\Theta_0}(\theta))^2,$$

lead to more adaptive estimators.

Proposition

Under the loss $L_2(\theta, \varphi)$, the Bayes estimator associated with π is the posterior probability

$$\varphi^\pi(x) = P^\pi(\theta \in \Theta_0 | x).$$

The posterior expectation of $\mathbb{I}_{\Theta_0}(\theta)$ is nothing but the posterior probability of Θ_0.

This (proper) loss provides a decision-theoretic foundation to the use of posterior probabilities as Bayesian answers.

Definition

For a one-sided test, i.e., for hypotheses of the form $H_0: \ \theta \leqslant \theta_0$ versus $H_1: \ \theta > \theta_0$, an interval $[t_1, t_2]$ is said to be a truncation set for the estimator φ if $\varphi(t) = 1$ when $t < t_1$ and $\varphi(t) = 0$ when $t > t_2$. For a two-sided test of $H_0: \ \theta \in [\theta_1, \theta_2]$ versus $H_1: \ \theta \notin [\theta_1, \theta_2]$, the interval $[t_1, t_2]$ is said to be a truncation set for the estimator φ if $\varphi(t) = 0$ when $t \notin [t_1, t_2]$.

Theorem

For the two-sided problem $H_0: \ \theta \in [\theta_1, \theta_2]$ versus $H_1: \ \theta \notin [\theta_1, \theta_2]$, an estimator φ with truncation set $[t_1, t_2]$ is admissible if there exist a probability measure π_0 $[\theta_1, \theta_2]$ and a σ-finite measure π_1 on $[t_2, t_2]^c$ such that

$$\varphi(x) = \frac{\displaystyle\int f(x|\theta)\pi_0(\theta)d\theta}{\displaystyle\int f(x|\theta)\pi_0(\theta)d\theta + \int f(x|\theta)\pi_1(\theta)d\theta},$$

for $x \in [t_1, t_2]$. Conversely, if φ is admissible, there exist $[t_1, t_2]$, π_0, and π_1 such that the equality (on $\varphi(x)$) above holds.

In the one-sided case, we can only propose an admissibility necessary condition, but it implies that the generalized Bayes estimators form a complete class.

Theorem

For the one-sided problem $H_0 : \ \theta \leqslant \theta_0$ *versus* $H_1 : \ \theta > \theta_0$*, if* φ *is admissible, there exists an increasing procedure* φ' *such that* φ' *is (risk) equivalent to* φ*. If* φ *is an increasing admissible procedure and* $[t_1, t_2]$ *is a truncation set such that* $0 < \varphi(x) < 1$ *on* $[t_1, t_2]$*, there exist two* σ*-finite measures on* $(-\infty, \theta_0]$ *and* $[\theta_0, +\infty)$*,* π_0 *and* π_1*, such that*

$$1 = \int \exp\{t_0\theta - \psi(\theta)\}(\pi_0(\theta) + \pi_1(\theta))d\theta$$

for $t_1 < t_0 < t_2$ *and* $\varphi(x)$ *given as in the previous theorem.*

These two complete class theorems show that it is sufficient to consider the generalized Bayes estimators to obtain admissible estimators under quadratic loss.

Theorem

For the test $H_0: \ \theta \in [\theta_1, \theta_2]$ versus $H_1: \ \theta \notin [\theta_1, \theta_2]$, when the sampling distribution is continuous with respect to the Lebesgue measure, the p-value is inadmissible for the loss $L_2(\theta, \varphi) = (\varphi - \mathbb{I}_{\Theta_0}(\theta))^2$.

This result shows that *p*-values do not belong to the range of Bayesian answers. It justifies the rejection of *p*-values for two-sided hypotheses.

The inadmissibility of *p*-values can be extended to most bounded proper losses.

Credible intervals

Definition

For a prior distribution π, a set C_x is said to be an α-credible set if

$$P^\pi(\theta \in C_x | x) \geqslant 1 - \alpha.$$

This region is called an HPD α-credible region (for highest posterior density) if it can be written under the form

$$\{\theta : \ \pi(\theta|x) > k_\alpha\} \subset C_x^\pi \subset \{\theta : \ \pi(\theta|x) \geqslant k_\alpha\}$$

where k_α is the largest bound such that

$$P^\pi(\theta \in C_x^\alpha | x) \geqslant 1 - \alpha.$$

Numerical Integration

Polynomial quadrature is intended to approximate integrals involving distributions close to the normal distribution:

$$\int_{-\infty}^{\infty} e^{-t^2/2} f(t) dt \simeq \sum_{i=1}^{n} \omega_i f(t_i),$$

where

$$\omega_i = \frac{2^{n-1} n! \sqrt{n}}{n^2 [H_{n-1}(t_i)]^2}$$

and t_i is the ith zero of the nth Hermite polynomial, $H_n(t)$.

No matter the numerical integration method used, its accuracy dramatically decreases as the dimension of Θ increases.

An empirical rule of thumb is that most standard methods should not be used for integration in dimensions larger than 4, although they keep improving over the years.

Monte Carlo methods

The approximation of the integral

$$\int_{\Theta} g(\theta) f(\theta|x) \pi(\theta) d\theta$$

should take advantage of the fact that π is a probability density (assuming it is a proper prior) or that $f(x|\theta)\pi(\theta)$ is proportional to a density.

If it is possible to generate random variables $\theta_1, \ldots, \theta_m$ from $\pi(\theta)$, the average

$$\frac{1}{m} \sum_{i=1}^{m} g(\theta_i) f(x|\theta_i)$$

converges almost surely to the integral above (Law of Large Numbers).

If it is possible to generate random variables $\theta_1, \ldots, \theta_m$ from $\pi(\theta|x)$, the average

$$\frac{1}{m} \sum_{i=1}^{m} g(\theta_i) \quad \text{converges a.s. to} \quad \frac{\int_{\Theta} g(\theta) f(x|\theta) \pi(\theta) d\theta}{\int_{\Theta} f(x|\theta) \pi(\theta) d\theta}.$$

If the posterior variance $\mathrm{Var}(g(\theta)|x)$ is finite, the Central Limit Theorem applies to the average

$$\frac{1}{m}\sum_{i=1}^{m} g(\theta_i) f(x|\theta_i),$$

which is then asymptotically normal with variance $\mathrm{Var}(g(\theta)|x)$.

Confidence regions can then be built from this normal approximation.

The magnitude of the error remains of order $1/\sqrt{m}$, whatever the dimension of the problem, in opposition with numerical methods.

The implementation of the method requires the production of the θ_i's by computer, using a deterministic pseudo-random generator to mimic generation from $\pi(\theta)$ or $\pi(\theta|x)$ by first replicating i.i.d. sampling from a uniform $\mathcal{U}([0,1])$ distribution and then transforming the uniforms into the variables of interest.

Importance sampling

If h is a probability density with support which includes the support of $g(\theta)f(x|\theta)\pi(\theta)$, then

$$\int_\Theta g(\theta)f(\theta|x)\pi(\theta)d\theta = \int_\Theta \frac{g(\theta)f(\theta|x)\pi(\theta)}{h(\theta)}h(\theta)d\theta.$$

If it is possible to generate random variables $\theta_1, \ldots, \theta_m$ from $h(\theta)$, the average

$$\frac{1}{m}\sum_{i=1}^{m} g(\theta_i)\omega(\theta_i), \quad \text{with the weights} \quad \frac{f(\theta|x)\pi(\theta)}{h(\theta)},$$

which also converges a.s. to the integral above. We also have that

$$\mathrm{E}^\pi(g(\theta)|x) \simeq \frac{\sum_{i=1}^m g(\theta_i)\omega(\theta_i)}{\sum_{i=1}^m \omega(\theta_i)}.$$

and this ratio does not depend on the normalizing constants in either $h(\theta)$, $f(x|\theta)$ or $\pi(\theta)$.

If $\mathrm{E}^h(g^2(\theta)\omega^2(\theta))$ is not finite, the variance of the estimator above is infinite.

Laplace analytic approximation

Consider the posterior expectation of interest

$$\mathrm{E}^{\pi}(g(\theta)|x) = \frac{\displaystyle\int_{\Theta} g(\theta) f(\theta|x)\pi(\theta) d\theta}{\displaystyle\int_{\Theta} f(\theta|x)\pi(\theta) d\theta}.$$

This ratio of integrals can be written as

$$\mathrm{E}^{\pi}(g(\theta)|x) = \frac{\displaystyle\int_{\Theta} b_N(\theta) \exp\{-nh_N(\theta)\} d\theta}{\displaystyle\int_{\Theta} b_D(\theta) \exp\{-nh_D(\theta)\} d\theta},$$

where the dependence on x is suppressed for simplicity's sake and where n is usually the sample size (although it may sometimes correspond to the inverse prior variance).

Given $\widehat{\theta}$ the Laplace expansion of a general function h with a single minimum θ, integral is given by

$$\begin{aligned}\int b(\theta)\exp\{-nh(\theta)\}d\theta &= \sqrt{2\pi}\sigma\exp\{-n\hat{h}\}\left\{\hat{b}+\frac{1}{2n}\left[\sigma^2\hat{b}''-\sigma^4\hat{b}'\hat{h}'''\right.\right.\\ &+ \left.\left.\frac{5}{12}\hat{b}(\hat{h}''')^2\sigma^2-\frac{1}{4}\hat{b}\hat{h}^{(4)}\sigma^4\right]\right\}+O(n^{-2}),\end{aligned}$$

where $\hat{b}$, $\hat{h}$, etc., denote the values of b, h, and of their derivatives for $\theta=\hat{\theta}$ and $\sigma^2=[h''^{(\hat{\theta})}]^{-1}$.

Criticisms:

- Laplace approximation is only justified asymptotically;
- analytical methods always imply delicate preliminary studies about the regularity of the integrated function that are not necessarily feasible;
- the posterior distribution should be similar enough to the normal distribution (for which Laplace approximation is exact); and
- such methods cannot be used in settings where the computation of the maximum likelihood estimator is quite difficult.

Markov chain Monte Carlo methods

It derives its name from the idea that, to produce acceptable approximations to integrals and to other functionals depending on a distribution of interest, it is enough to generate a Markov chain $(\theta^{(m)})_m$ with limiting distribution the distribution of interest.

If the Markov chains $(\theta^{(m)})_m$ produced by MCMC algorithms are irreducible, that is, if they are guaranteed to visit any set A such that $\pi(A|x) > 0$, then these chains are positive recurrent with stationary distribution $\pi(\theta|x)$.

These Markov chains are also ergodic, which means that the distribution of $\theta^{(m)}$ converges to $\pi(.|x)$ for almost every starting value $\theta^{(0)}$, that is, the influence of the starting value vanishes. Therefore, for k large enough, the resulting $\theta^{(k)}$ is approximately distributed from $\pi(\theta|x)$, no matter what the starting value $\theta^{(0)}$ is.

Independence of the $\theta^{(k)}$ is not crucial if we are mainly interested in functionals of $\pi(\theta|x)$, since the Ergodic Theorem implies that the average

$$\frac{1}{K}\sum_{k=1}^{K} g(\theta^{(k)}) \to \mathrm{E}^{\pi}(g(\theta)|x) < \infty \quad \text{as} \quad K \to \infty.$$

Metropolis-Hastings algorithms

Given a density $\pi(\theta)$, known up to a normalizing factor, and a conditional density $q(\theta'|\theta)$, the algorithm generates the chain $(\theta^{(m)})_m$ by:

1. Start with an arbitrary initial value $\theta^{(0)}$;
2. Update from $\theta^{(m)}$ to $\theta^{(m)}$ $(m = 1, 2, \ldots)$ by
   2.1 Generate $\xi \sim q(\xi|\theta^{(m)})$;
   2.2 Define
   $$\alpha = \min\left\{1, \frac{\pi(\xi)q(\theta^{(m)}|\xi)}{\pi(\theta^{(m)})q(\xi|\theta^{(m)})}\right\}$$
   2.3 Take
   $$\theta^{(m+1)} = \begin{cases} \xi & \text{with probability } \alpha; \\ \theta^{(m)} & \text{otherwise.} \end{cases}$$