Statistical Decision Theory

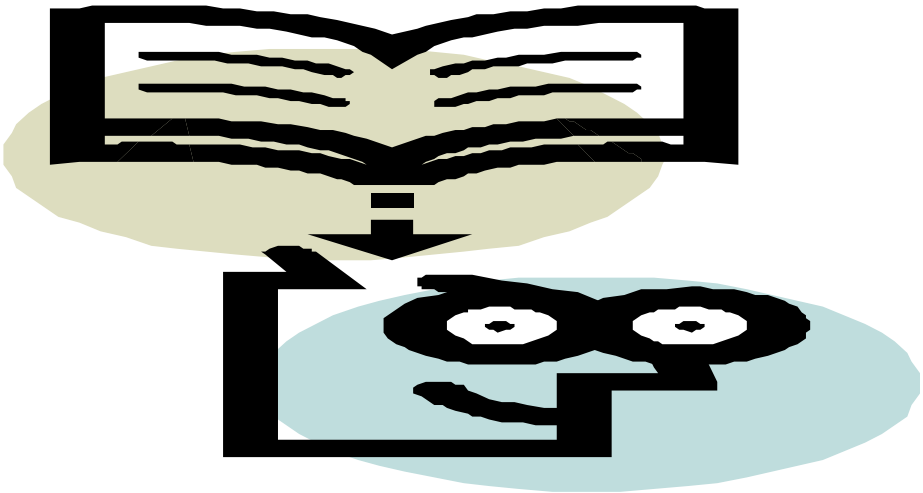
Prior Information and Subjective Probability

Jiangsheng Yu

©School of Electronics Engineering and Computer Science

Peking University, Beijing, 100871

yujs@pku.edu.cn, http://icl.pku.edu.cn/yujs



Topics

1. Subjective probability
2. Subjective determination of prior density
3. Noninformative priors
4. Maximum entropy priors
5. Using marginal distribution to determine the prior
6. Hierarchical priors
7. Criticisms
8. Conclusion
9. References

One-time Event

Problem 1 Frequentist fails at explaining the probability of one time event.

1. What is the probability of John's unborn baby being a girl?
2. How about the probability of raining tomorrow?

Note The assumption of repeatable experiments is not feasible. Bayesian thinks the probability of an event as the belief degree of its occurrence.

Subjective Probability

Gamble way of understanding subjective probability:

- Lose z if E occurs, where $0 \leq z \leq 1$.
- Win $(1 - z)$ if E^c occurs.

Choose z so that the gamble is fair (i.e., the overall utility is zero), resulting in the equation

$$\begin{aligned} 0 &= \text{expected utility of the gamble} \\ &= \mathsf{U}(-z)\mathsf{P}(E) + \mathsf{U}(1-z)(1-\mathsf{P}(E)) \end{aligned} \tag{1}$$

Suppose that z is small, solving for $\mathsf{P}(E)$ yields,

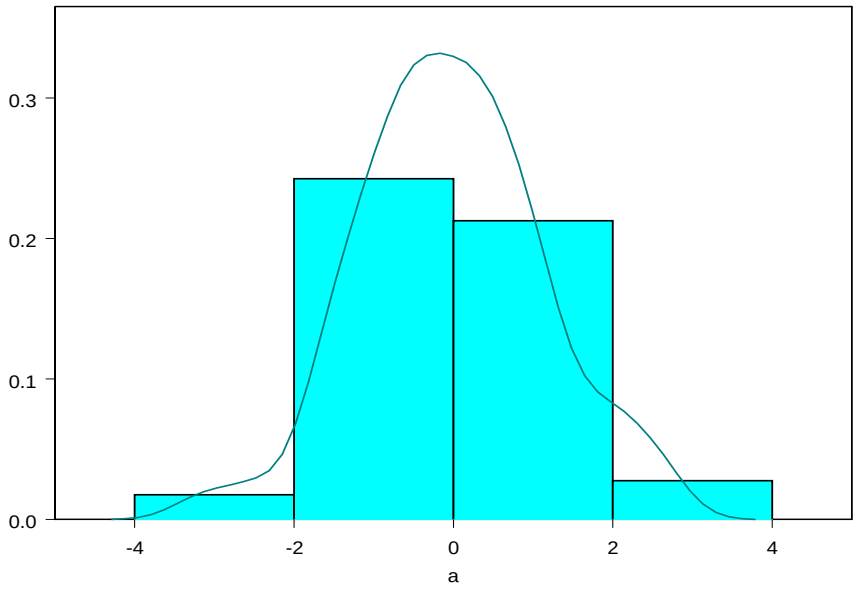
$$\mathsf{P}(E) = \frac{\mathsf{U}(1-z)}{\mathsf{U}(1-z) - \mathsf{U}(-z)} \approx 1 - z \tag{2}$$

Subjective Prior Density

1. Histogram approach
2. Relative likelihood approach
3. Matching a given functional form
4. CDF determination

Shortcomings of histogram approach:

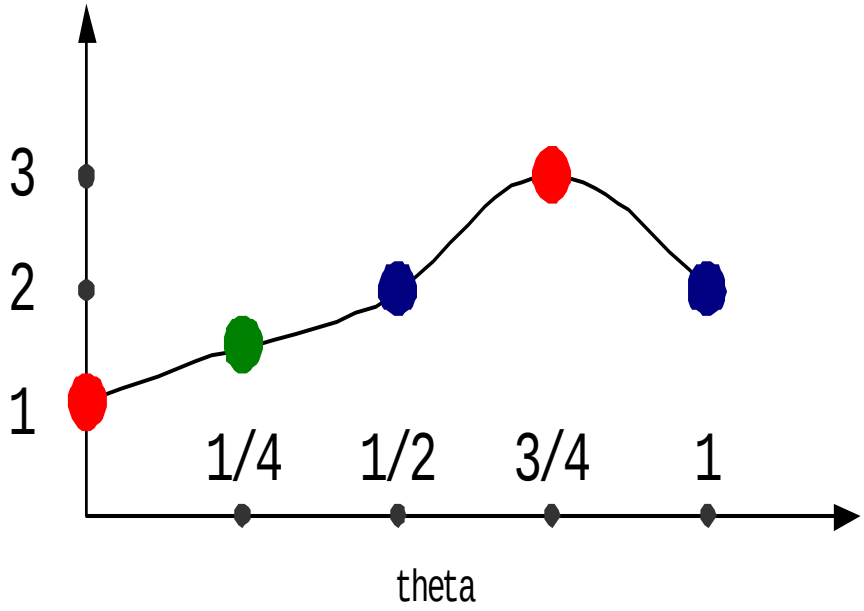
- No clearcut rule that determines how many intervals, what size intervals, etc
- Not practicable
- No tail



Relative Likelihood Approach

Problem 2 Let $\Theta = [0, 1]$ and $\mathsf{P}(\theta = \frac{3}{4}) = 3\mathsf{P}(\theta = 0)$, where $\operatorname{argmax}_\theta \mathsf{P}(\theta) = \frac{3}{4}$, $\operatorname{argmin}_\theta \mathsf{P}(\theta) = 0$.

If we know $\mathsf{P}(\theta = \frac{1}{2}) = \mathsf{P}(\theta = 1) = 2\mathsf{P}(\theta = 0), \mathsf{P}(\theta = \frac{3}{4}) = 2\mathsf{P}(\theta = \frac{1}{4})$, then the density is fitted.



Note It does not matter that $\int_\Theta \pi(\theta) \neq 1$.

Matching a Functional Form

Example 1 Given a functional form[a] of $\pi(\theta)$.

1. Suppose $\theta \sim \mathcal{N}(\mu, \sigma^2)$. One need only decide the prior mean and prior variance to specify $\pi(\theta)$.
2. Suppose $\theta \sim \beta(r, s)$, the prior mean μ and prior variance σ^2. By $\mu = \frac{r}{r+s}, \sigma^2 = \frac{rs}{(r+s)^2(r+s+1)}$, we can specify $\pi(\theta)$.
3. Let $\Theta = (-\infty, +\infty)$ and the prior distribution is $\mathcal{N}(0, \sigma^2)$. If we know the $\frac{1}{4}$-fractile and $\frac{3}{4}$-fractile are -1 and 1, we can get $\sigma^2 = 2.19$.

Example 2 The density with tail of $K\theta^{-2}$ on $(0, +\infty)$ has no moments.

---

[a] Unfortunately, the estimation of prior moments is often an extremely uncertain undertaking. The difficulty is that the tails of density can have a drastic effect on its moments.

Equivalent Sample Size

Example 3 Assume a sample $X_1, X_2, \cdots, X_n$ from a $\mathcal{N}(\theta, 1)$ distribution is observed, then $\bar{X} \sim \mathcal{N}(\theta, \frac{1}{n})$. It is easy to determine the mean μ of $\pi(\theta)$, but difficult to determine the prior variance σ^2.

We will see[a], the mean of posterior distribution is

$$\bar{x}\left(\frac{\sigma^2}{\sigma^2 + 1/n}\right) + \mu\left(\frac{1/n}{\sigma^2 + 1/n}\right)$$

[a] This suggests that the prior variance, σ^2, plays the same role as $1/n$ in the determination of θ. Hence, the idea of equivalent sample size is to determine n^* s.t. a sample of that size would make $\bar{x}$ as convincing an estimate of θ as a subjective guess μ. Then, $\sigma^2 = 1/n^*$ would be an appropriate prior variance.

CDF Determination

- Subjectively determine several α-fractiles, $z(\alpha)$.
- Plot the points $(\alpha, z(\alpha))$ and sketch a smooth curve joining them.

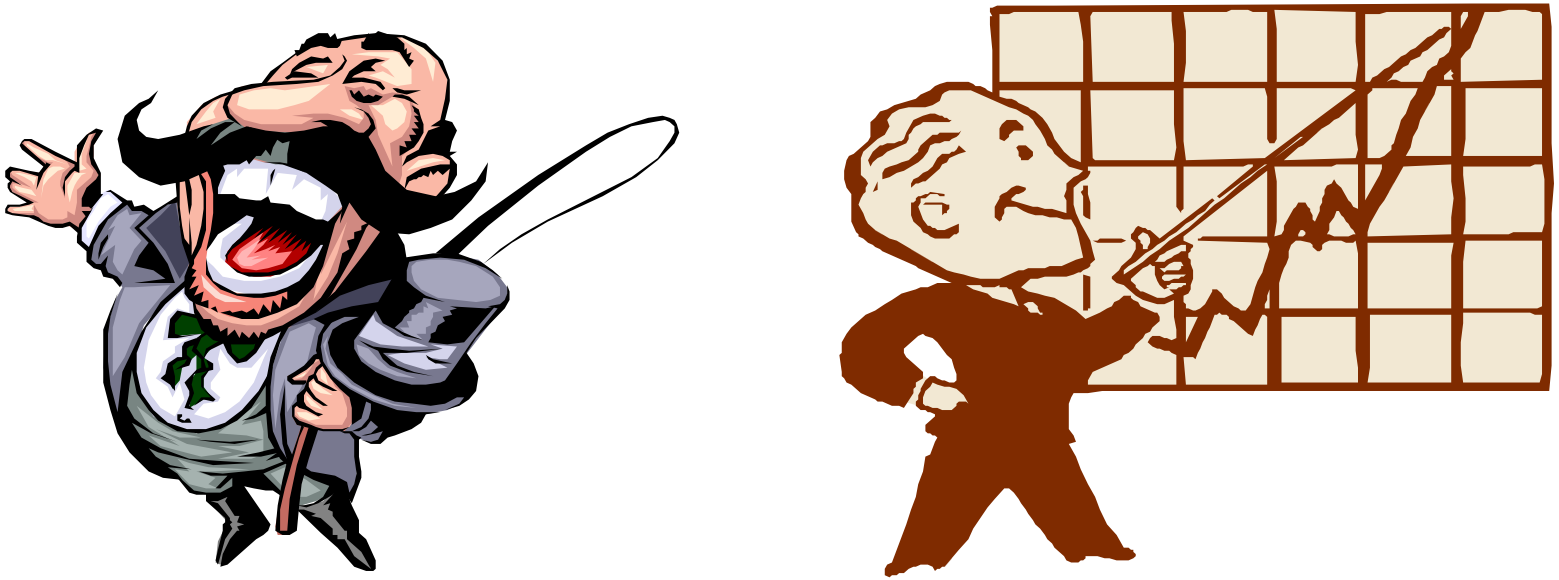


Figure 1: Training is propitious to prior density

Noninformative Prior

Example 4 Consider the mean of normal population $\theta \in \Theta = (-\infty, +\infty)$, whose noninformative prior (NP) is $\pi(\theta) = 1$,[a] that is improper density because

$$\int \pi(\theta) d\theta = \infty \tag{3}$$

Example 5 If Θ is finite, it sounds reasonable that the noninformative prior of θ is the uniform distribution on Θ.

Example 6 Let $\eta = e^{\theta}$ and $\pi(\theta) = 1$ on $\mathbb{R}$. Then $\pi^*(\eta) = \eta^{-1}\pi(\log \eta) = \eta^{-1}$.

---

[a] Called uniform density on $\mathbb{R}$, used by Laplace (1812) first. Any value of θ is not a particular favor.

NP of Parameter— My Opinion

1. The parameter θ is a random variable whose knowledge is its distribution. In the case of knowing nothing about θ, why should we prefer the uniform distribution?
2. My opinion of knowing nothing about θ is that the noninformative prior distribution of θ is uniformly distributed on the set of all possible distributions.

Note Any distribution can be as the prior knowledge of θ if it is completely unknown, which at least manifests the taste of decision maker.

NP for Location Problem

Definition 1 Let $\mathcal{X}, \Theta \subset \mathbb{R}^p$. If the density of $\boldsymbol{X}$ is $f(\boldsymbol{X} - \boldsymbol{\theta})$, then f is called location density and $\boldsymbol{\theta}$ is called the location parameter. For instance, $\mathcal{N}_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ ($\boldsymbol{\Sigma}$ fixed) is a location density.

Note If we observe $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{c}$ instead of $\boldsymbol{X}$ where $\boldsymbol{c} \in \mathbb{R}^p$, and let $\boldsymbol{\eta} = \boldsymbol{\theta} + \boldsymbol{c}$, then $\boldsymbol{Y}$ has density $f(\boldsymbol{Y} - \boldsymbol{\eta})$. Thus, $(\boldsymbol{X}, \boldsymbol{\theta})$ and $(\boldsymbol{Y}, \boldsymbol{\eta})$ problems are identical in structure, and it seems reasonable to insist that they have the same NP. Let π, π^* denote the NPs in the $(\boldsymbol{X}, \boldsymbol{\theta})$ and $(\boldsymbol{Y}, \boldsymbol{\eta})$ problems, then $\forall \boldsymbol{A} \subset \mathbb{R}^p$, we have

$$\begin{aligned} \mathsf{P}^{\pi}(\boldsymbol{\theta} \in \boldsymbol{A}) &= \mathsf{P}^{\pi^*}(\boldsymbol{\eta} \in \boldsymbol{A}) \\ &= \mathsf{P}^{\pi}(\boldsymbol{\theta} \in \boldsymbol{A} - \boldsymbol{c}) \end{aligned} \tag{4}$$

Location Invariant Prior

Definition 2 The density π is called location invariant prior if it satisfies (4) or $\forall \boldsymbol{A} \subset \mathbb{R}^p$,

$$\begin{aligned} \int_{\boldsymbol{A}} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} &= \int_{\boldsymbol{A}-\boldsymbol{c}} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{A}} \pi(\boldsymbol{\theta}-\boldsymbol{c}) d\boldsymbol{\theta} \end{aligned} \tag{5}$$

Thus, $\forall \boldsymbol{\theta} \in \mathbb{R}^p, \pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}-\boldsymbol{c})$. Let $\boldsymbol{\theta} = \boldsymbol{c}$, we have $\pi(\boldsymbol{c}) = \pi(\mathbf{0})$ for all $\boldsymbol{c} \in \mathbb{R}^p$. Consequently, π should be a constant. $\pi(\boldsymbol{\theta}) = 1$ is reasonable.[a]

[a] More general case can be found in pp86-87 in [2].

Scale Parameter

Definition 3 A (one-dimension) scale density is a density of the form $\sigma^{-1} f(x/\sigma)$, where $\sigma > 0$. The parameter σ is called a scale parameter. For example, $\mathcal{N}(0, \sigma^2)$.

Note Observe $Y = cX$ $(c > 0)$ instead of X. Let $\eta = c\sigma$, then the density of Y is $\eta^{-1} f(y/\eta)$. Let π and π^* denote the priors in the (X, σ) and (Y, η) respectively.

$$\begin{aligned} \mathsf{P}^{\pi}(\sigma \in A) &= \mathsf{P}^{\pi^*}(\eta \in A) \\ &= \mathsf{P}^{\pi}(\sigma \in c^{-1}A) \end{aligned} \tag{6}$$

where $\forall A \subset \mathbb{R}$. Any distribution π satisfying (6) is called scale invariant.

NP for Scale Parameter

Suppose π is a scale invariant parameter, then

$$\begin{aligned} \int_A \pi(\sigma)d\sigma &= \int_{c^{-1}A} \pi(\sigma)d\sigma \\ &= \int_A \pi(c^{-1}\sigma)c^{-1}d\sigma \end{aligned} \tag{7}$$

Consequently, $\pi(\sigma) = c^{-1}\pi(c^{-1}\sigma)$. Let $\sigma = c$, we get $\pi(c) = c^{-1}\pi(1)$. Set $\pi(1) = 1$, the noninformative prior of σ is $\pi(\sigma) = \sigma^{-1}$, which is also an improper density since $\int_0^\infty \sigma^{-1}d\sigma = \infty$.

Table Entry Problem

Problem 3 The relative frequencies of the integers 1 through 9 in the first significant digit of the table entries are $\ln(1+i^{-1})/\ln 10$, where $i = 1, 2, \cdots, 9$.

Solution Assume that the distribution of table entries is scale invariant. The normalized prior π on $(1, 10)$ is $\pi(\sigma) = \sigma^{-1}/\ln 10$. So, σ will have first digit i when it lies in the interval $[i, i+1)$, whose probability is

$$p_i = \int_i^{i+1} [\sigma \ln 10]^{-1} d\sigma = \frac{\ln(1+i^{-1})}{\ln 10}$$

It may be coincidence, but intriguing.

Jeffreys' NP (1961)

- When θ is r.v., Jeffreys' NP is

$$\pi(\theta) = \sqrt{\mathsf{I}(\theta)} \tag{8}$$

where $\mathsf{I}(\theta)$ is the expected Fisher's information

$$\mathsf{I}(\theta) = -\mathsf{E}_\theta \left[\frac{\partial^2 \ln f(X|\theta)}{\partial \theta^2} \right] \tag{9}$$

- When $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)^\mathsf{T}$, Jeffreys' NP is

$$\pi(\boldsymbol{\theta}) = \sqrt{\det \mathsf{I}(\boldsymbol{\theta})} \tag{10}$$

where $\mathsf{I}(\boldsymbol{\theta})$ is the expected Fisher's information matrix

$$\mathsf{I}_{ij}(\boldsymbol{\theta}) = -\mathsf{E}_{\boldsymbol{\theta}} \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln f(X|\boldsymbol{\theta}) \right] \tag{11}$$

Location-Scale Parameters

Definition 4 A location-scale density is a density of the form $\sigma^{-1} f((x-\theta)/\sigma)$, for instance, $\mathcal{N}(\theta, \sigma^2)$ where $\boldsymbol{\theta} = (\theta, \sigma)$. Fisher information matrix is

$$\begin{aligned}
\mathsf{I}(\boldsymbol{\theta}) &= -\mathsf{E}_{\boldsymbol{\theta}} \begin{pmatrix} \frac{\partial^2}{\partial\theta^2} & \frac{\partial^2}{\partial\theta\partial\sigma} \\ \frac{\partial^2}{\partial\theta\partial\sigma} & \frac{\partial^2}{\partial\sigma^2} \end{pmatrix} \left[-\tfrac{(X-\theta)^2}{2\sigma^2} \right] \\
&= -\mathsf{E}_{\boldsymbol{\theta}} \begin{pmatrix} -1/\sigma^2 & 2(\theta - X)/\sigma^3 \\ 2(\theta - X)/\sigma^3 & -3(X-\sigma)^2/\sigma^4 \end{pmatrix} \\
&= \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 3/\sigma^2 \end{pmatrix}
\end{aligned}$$

Hence, $\pi(\boldsymbol{\theta}) = \sqrt{\frac{1}{\sigma^2} \cdot \frac{3}{\sigma^2}} \propto \frac{1}{\sigma^2}$, which is also improper.

Discussion on NP

Example 7 Let θ be a binomial parameter, then $\Theta = [0, 1]$. The plausible NPs for θ are

- $\pi_1(\theta) = 1$ (Bayes 1763, Laplace 1812)
- $\pi_2(\theta) = \theta^{-1}(1-\theta)^{-1}$ (Novick 1965)
- $\pi_3(\theta) \propto [\theta(1-\theta)]^{-1/2}$ (Jeffreys 1968)
- $\pi_4(\theta) \propto \theta^{\theta}(1-\theta)^{1-\theta}$ (Zellner 1977)

where π_1, π_3, π_4 are proper densities (π_3, π_4 upon suitable normalization).

Note The Bayesian argued that operationally it is rare for the choice of an NP to markedly affect the answer.

Maximum Entropy Prior (MEP)

Definition 5 Let π be a probability density on discrete Θ.

$$\mathcal{H}(\pi) = -\sum_{\Theta} \pi(\theta_i) \log \pi(\theta_i) \tag{12}$$

Theorem 1 Given the partial prior information about θ in the form of restrictions

$$\mathsf{E}^{\pi}[g_k(\theta)] = \sum_{i} \pi(\theta_i) g_k(\theta_i) = \mu_k \tag{13}$$

where $k = 1, 2, \cdots, m$. Then the MEP is

$$\bar{\pi}(\theta_i) = \frac{\exp\{\sum_{k=1}^{m} \lambda_k g_k(\theta_i)\}}{\sum_i \exp\{\sum_{k=1}^{m} \lambda_k g_k(\theta_i)\}} \tag{14}$$

where λ_k are constants determined by (13).

Example of MEP

Example 8 Assume $\theta = \mathbb{N}$ and given $\mathsf{E}^{\pi}(\theta) = 5$. By (13), $m = 1, g_1(\theta) = \theta, \mu_1 = 5$. The MEP is

$$\bar{\pi}(\theta) = \frac{e^{\lambda_1 \theta}}{\sum_{\theta=0}^{\infty} e^{\lambda_1 \theta}} = (1 - e^{\lambda_1}) e^{\lambda_1 \theta}$$

Thus, $\mathsf{E}^{\bar{\pi}}(\theta) = (1 - e^{\lambda_1})/e^{\lambda_1}$. Setting this equal to $\mu_1 = 5$, and solving, we have $\bar{\pi}(\theta) = 5/6^{\theta+1}$ or $\theta \sim \mathcal{NB}(1, 5/6)$.

Note If Θ is continuous, there is no longer a completely natural definition of entropy.

Jaynes Entropy (1968)

Definition 6 Let $\pi_0(\theta)$ be the natural invariant NP.

$$\begin{aligned} \mathcal{H}(\pi) &= -\mathsf{E}^{\pi}\left[\log \frac{\pi(\theta)}{\pi_0(\theta)}\right] \\ &= -\int \pi(\theta) \log \frac{\pi(\theta)}{\pi_0(\theta)} d\theta \end{aligned} \tag{15}$$

Theorem 2 The MEP restricted by (13) is

$$\bar{\pi}(\theta) = \frac{\pi_0(\theta) \exp\{\sum_{k=1}^{m} \lambda_k g_k(\theta)\}}{\int_{\Theta} \pi_0(\theta) \exp\{\sum_{k=1}^{m} \lambda_k g_k(\theta)\} d\theta} \tag{16}$$

where λ_k are constants determined by (13).

Note When Θ is unbounded and the specified restrictions are specifications of fractiles, (16) often nonexists.

Example of Jaynes' MEP

Example 9 Assume $\Theta = \mathbb{R}$, and θ is a location parameter. The natural NP is then $\pi_0(\theta) = 1$. Let the restrictions be

$$\begin{cases} g_1(\theta) = \theta, & \mu_1 = \mu \text{ (mean)} \\ g_2(\theta) = (\theta - \mu)^2, & \mu_2 = \sigma^2 \text{ (variance)} \end{cases}$$

$$\bar{\pi}(\theta) = \frac{\exp\{\lambda_1\theta + \lambda_2(\theta - \mu)^2\}}{\int_{-\infty}^{\infty} \exp\{\lambda_1\theta + \lambda_2(\theta - \mu)^2\} d\theta}$$

Intuitively, $\bar{\pi}(\theta) = \mathcal{N}(\mu - \frac{\lambda_1}{2\lambda_2}, -\frac{1}{2\lambda_2})$. Hence, $\lambda_1 = 0, \lambda_2 = -\frac{1}{2\sigma^2}$, i.e., $\theta \sim \mathcal{N}(\mu, \sigma^2)$.

Marginal Distribution

Definition 7 If X has density $f(x|\theta)$ and $\theta \sim \pi(\theta)$, then the joint density of X and θ is

$$h(x, \theta) = f(x|\theta)\pi(\theta) \tag{17}$$

Definition 8 The marginal density of X is

$$\begin{aligned} m(x) &= \int_{\Theta} f(x|\theta) dF^{\pi}(\theta) \\ &= \begin{cases} \int_{\Theta} f(x|\theta)\pi(\theta) d\theta \\ \sum_{\Theta} f(x|\theta)\pi(\theta) \end{cases} \end{aligned} \tag{18}$$

Information about m

There are several possible sources of information about marginal density:

- Subjective knowledge (θ is not intuitive)
- Data (Empirical Bayes proposed by Robbins)

Example 10 Assume $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)$ iid from π_0 and $\boldsymbol{X} = (X_1, \cdots, X_p)$, then data $\boldsymbol{x}$ can be used to estimate m_0 (and hence m):

$$m_0(x_i) = \int f(x_i|\theta_i) dF^{\pi_0}(\theta_i)$$

$$m(\boldsymbol{x}) = \prod_{i=1}^{p} m_0(x_i)$$

Restricted Classes of Priors

Priors of a given functional form: Given a prescribed function g, consider the priors

$$\Gamma = \{\pi | \pi(\theta) = g(\theta|\lambda), \lambda \in \Lambda\} \tag{19}$$

where λ is called the hyperparameter of the prior.

Priors of a given structural form: Some relationship among $\boldsymbol{\theta}$ is suspected. For instance, assume $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)$ are iid,

$$\Gamma = \{\pi | \pi(\boldsymbol{\theta}) = \prod_{i=1}^{p} \pi_0(\theta_i)\} \tag{20}$$

where π_0 is an arbitrary density.

Priors close to an elicited prior: ϵ-contamination class

$$\Gamma = \{\pi | \pi(\boldsymbol{\theta}) = (1-\epsilon)\pi_0(\theta_i) + \epsilon q(\theta), q \in \mathcal{Q}\} \tag{21}$$

ML-2 Prior

Definition 9 Suppose Γ is a class of priors under consideration, and the $\hat{\pi} \in \Gamma$ satisfies (for the observed data x)

$$m(x|\hat{\pi}) = \sup_{\pi \in \Gamma} m(x|\pi) \tag{22}$$

then $\hat{\pi}$ will be called the type 2 maximum likelihood prior, or ML-2 prior for short.

Example 11 See pp99-101 in [2].

Further reading pp101-104 in [2].

- The moment approach to prior selection
- The distance approach to prior selection

Marginal Exchangeability

Example 12 (Coin Tossing) It is diffi cult, maybe impossible, to give precise and operationally realizable defi nitions of independence and of θ that are not subjective.

Theorem 3 (deFietti, 1937) The fundamental entity should be a (subjective) probability distribution, m, describing the actual sequence of H and T that would be anticipated.

$$\begin{aligned} m(\boldsymbol{x}) &= m(\boldsymbol{x}') \\ &= \int_0^1 \left[\prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} \right] dF^{\pi}(\theta) \end{aligned} \tag{23}$$

where $\boldsymbol{x}'$ is any permutation of $\boldsymbol{x}$ and $F^{\pi}(\theta) = \lim_{n\to\infty} \mathsf{P}^m \left(\frac{1}{n} \sum_{i=1}^{n} X_i \leq \theta \right)$.

Consequently, $f(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i}$.

Hierarchical Priors

One may have the structural and subjective prior information at the same time, and it is often convenient to model this stages.

1. $\Gamma = \{\pi_1(\boldsymbol{\theta}|\lambda)|\lambda \in \Lambda\}$, where the functional form of π is known.
2. $\pi_2(\lambda)$, the prior of hyperparameters λ, can be chosen according to the subjective belief.
3. Computation of $\pi(\boldsymbol{\theta})$:

$$\pi(\boldsymbol{\theta}) = \int_{\Lambda} \pi_1(\boldsymbol{\theta}|\lambda) dF^{\pi_2}(\lambda) \tag{24}$$

Note Hierarchical is more robust than single.

Example of Hierarchical Priors

1. Functional model

$$\Gamma = \{\pi_1(\boldsymbol{\theta}|\lambda)|\pi_1(\boldsymbol{\theta}|\lambda) = \prod_{i=1}^{p} \pi_0(\theta_i), \pi_0 \text{ being } \mathcal{N}(\mu_\pi, \sigma_\pi^2), \mu_\pi \in \mathbb{R}, \sigma_\pi^2 > 0\}$$

2. Hyperprior: $\pi_2(\lambda) = ?$
3. Computation of $\pi(\boldsymbol{\theta})$.[a]

[a] See pp107-109 in [2].

Bayesian Opinions

- Box (1980) said, "· · · , *I believe that it is impossible logically to distinguish between model assumptions and the prior distribution of the parameter.*"
- More bluntly, Good (1976) said, "*The subjectivist states his judgements, whereas the objectivist sweeps them under the carpet by calling assumptions knowledge, and he basks in the glorious objectivity of science.*"
- Savage (1962) said, "*It takes a lot of self-discipline not to exaggerate the probabilities you would have attached to hypotheses before they were suggested to you.*"

References

1. P. J. Bickel and K. A. Doksum (2001), *Mathematical Statistics — Basic Ideas and Selected Topics* (Second Edition). Prentice-Hall, Inc.
2. J. O. Berger (1985), *Statistical Decision Theory and Bayesian Analysis*. Springer Verlag, New York.

Thank you

for your attention!