CH3 PRIOR INFORMATION AND SUBJECTIVE PROBABILITY

Determing prior : discrete variables $\left\{ \begin{array}{l} \text{Scoring} \\ \text{Betting} \end{array} \right.$

§3.2 Determining Prior Density

- Histogram approach
- The relative likelihood approach
- Matching a given functional form
  - Estimating prior moments μ, σ^2

    ☹ The tail of a density can have a drastic effect on its moments Eg. $\int_b^\infty \theta \cdot (k\theta^{-2})\, d\theta = \infty$
  - Estimating fractiles
  - Equivalent sample size / device of imaginary results

    For normal distribution, the posterior with normal prior : $N(\mu, \sigma^2)$

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

    $\sigma^2 = 1/n^*$ equivalent to have $1/\sigma^2$ samples of mean μ.

    ☹ Useful only when certain specific functional forms

    ☹ Tend to considerably underestimate the amount of information carried by a sample of size n.
- CDF Determination (CDF: cumulative distribution function)

  1° Subjectively determine several α-fractiles, $z(\alpha)$.

  2° Plot the points $(\alpha, z(\alpha))$ and sketch a smooth curve joining them

§3.3 Noninformative Priors

No (or minimal) prior information available + Compelling Bayesian analysis

$$\Downarrow$$

Noninformative Prior

Example : Discrete variables: uniform

Example : $\Theta = (-\infty, \infty)$ uniform $\Rightarrow$ $\pi(\theta) = \overset{\overset{1}{\|}}{c} > 0.$

$\int \pi(\theta)\, d\theta = \infty$ may or may not cause problems

Severe (though unjustified) criticism:

Lack of Invarince under Transformation

Example: Let $\eta = \exp\{\theta\}$.

$\pi^*(\eta) = \eta^{-1} \pi(\log \eta)$

Let $y = g(x)$, $x = h(y)$

$(h = g^{-1})$

$f_Y(y) = |h'(y)|\, f_X(h(y))$

Example (Noninformative priors for location problems)

Suppose $\mathcal{X}$ and Θ are subsets of $\mathbb{R}^p$. and desity of $\mathbf{X}$ if of the form $f(\mathbf{x} - \boldsymbol{\theta})$

$\left(\text{E.g. } \mathbf{x} - \boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \underset{\uparrow}{\Sigma})\right)$

Σ fixed.

Imagine that, instead of $\mathbf{X}$, we observe $\mathbf{Y} = \mathbf{X} + \underset{\overset{\downarrow}{c \in \mathbb{R}^p.\ \text{fixed}}}{}\mathbf{c}$.

Then $\mathbf{Y}$ has density $f(\mathbf{y} - \boldsymbol{\eta})$

$\Rightarrow$ The $(\mathbf{X}, \boldsymbol{\theta})$ and $(\mathbf{Y}, \boldsymbol{\eta})$ problems are identical in structure

Noninformative Priors in general settings, please see textbooks pp. 87-8

Let π_1 and π_2 denote the noninformative priors for θ and η

Invariant noninformative assumption : $P^{\pi_1}(\theta \in A) = P^{\pi_2}(\eta \in A) \quad \forall A \text{ in } \mathbb{R}^1$

Then $P^{\pi_2}(\eta \in A) = P^{\pi_1}(\theta + c \in A) = P^{\pi_1}(\theta \in A - c)$

Combining the above equations:

$$P^{\pi_1}(\theta \in A) = P^{\pi_1}(\theta \in A - c)$$

$$\int_A \pi(\theta)\,d\theta = \int_{A-c} \pi(\theta)\,d\theta = \int_A \pi(\theta - c)\,d\theta$$

Unnecessary for intuitive thinking! Insufficient for mathematical proof

It can be shown that

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

Let $\theta = c$, $\quad \pi(c) = \pi(0)$

<u>Example</u> (Noninformative priors for scale problems)

$$y = \frac{x}{\sigma} \qquad f_Y(y) = \sigma^{-1} f_X\!\left(\frac{x}{\sigma}\right)$$

σ : a scale parameter <u>Eg.</u> $\sigma \sim \mathcal{N}(0, \beta^2)$

Imagine that, instead of observing ~~X~~ X, we observe the random variable $Y = cX$ $(c > 0)$

Note that $X \sim \frac{1}{\sigma^{-1}} f\left(\frac{x}{\sigma}\right)$

Let $\eta = c \cdot \sigma$. $\quad Y \sim \eta^{-1} f\left(\frac{y}{\eta}\right)$

$Y = cX \sim \frac{1}{c\sigma^{-1}} f\left(\frac{y}{c\sigma}\right)$

If $\mathcal{X} = \mathbb{R}^1$ of $\mathcal{X} = (0, \infty)$, then

(X, σ) is equivalent to (Y, η)

Denote π_1 and π_2 the prior of σ and η

$$P^{\pi_1}(\sigma \in A) = P^{\pi_2}(\eta \in A)$$

Since $\eta = c\sigma$

$$P^{\pi_2}(\eta \in A) = P^{\pi_1}(\sigma \in c^{-1}A)$$

$c^{-1}A = \{c^{-1}z : z \in A\}$

Combining the above equations

$$P^{\pi}(\sigma \in A) = P^{\pi}(\sigma \in c^{-1}A)$$

Thus

$$\int_A \pi(\sigma)\,d\sigma = \int_{c^{-1}A} \pi(\sigma)\,d\sigma = \int_A \pi(c^{-1}\sigma)\,c^{-1}\,d\sigma$$

Choosing $\sigma = c$ in $\pi(\sigma) = c^{-1}\pi(c^{-1}\sigma)$

$$\pi(c) = c^{-1}\pi(1).$$

Note that $\int_0^\infty \sigma^{-1}\,d\sigma = \infty$. π is an improper prior.

Example (The "Table Entry" Problem)

Observation: The frequencies of the integer 1..9 being the first significant digit of the table entries are $\frac{\log(1+i^{-1})}{\log 10}$.

Explanation by "noninformative priors"

$$\tilde{\pi}(\sigma) = \sigma^{-1}$$

Normalize on $(1,10)$

$$\pi(\sigma) = \frac{\sigma^{-1}}{\log 10}$$

The probability of i being the first significant digit

$$P_i = \int_i^{i+1} [\sigma \log 10]^{-1}\,d\sigma = \frac{\log(1+i^{-1})}{\log 10}$$

May be coincidence, but intriguing

§ 3.4 Maximum Entropy Priors

See Adam L Beger at al., A Maximum Entropy Approach to Natural Language Processing. 1996.

Note: I remember that one or a few formulas in the above paper are wrong, when solving the Lagrangian.

§ 3.5 Using the Marginal Distribution to Determine the Prior

- Definition: The joint density of X and θ is

$$h(x,\theta) = f(x|\theta)\pi(\theta)$$

The marginal density

$$m(x|\pi) = \int_{\Theta} f(x|\theta)\, dF^{\pi}(\theta) = \begin{cases} \int_{\Theta} f(x|\theta)\pi(\theta)\, d(\theta) & \text{(continuous)} \\ \sum_{\Theta} f(x|\theta)\pi(\theta) \end{cases}$$

- Information about m: $\begin{cases} \text{subjective knowledge} \\ \text{data itself (empirical Bayes)} \end{cases}$

- We consider also restricted classes of priors denoted as Γ

1° Priors of a given functional form

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

2° Priors of a given structural form

E.g. θ_i independent: $\Gamma = \{\pi : \pi(\theta) = \prod_{i=1}^{p} \pi_i(\theta_i)\}$

3° Priors close to an elicited prior

$$\Gamma = \{\pi : \pi(\theta) = (1-\varepsilon)\underset{\nearrow}{\pi_0(\theta)} + \varepsilon q(\theta),\ q \in \underset{\uparrow}{\mathcal{Q}}\}$$

elicited prior   class of contamination

- The ML-II approach to prior selection

<u>Definition</u>: Suppose Γ is a class of priors under consideration, and that $\hat{\pi} \in \Gamma$ satisfies (for the observed data x).

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

Then $\hat{\pi}$ will be called the <u>type II maximum likelihood prior</u> or <u>ML-II prior</u>.

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

then $\sup_{\pi \in \Gamma} m(x|\pi) = \sup_{\pi \in \Lambda} m(x|g(\theta|\lambda))$

<u>Example</u>: Let $X \sim \mathcal{N}(\theta, \sigma_f^2)$

$\theta \sim \mathcal{N}(\mu_\pi, \sigma_\pi^2)$

Then $m_0(x_i) = \mathcal{N}(x_i | \mu_\pi, \sigma_\pi^2 + \sigma_f^2) \quad (\forall i)$

$$m(x|\pi) = \prod_{i=1}^{p} m_0(x_i|\pi_0)$$

$$= \prod_{i=1}^{p} \frac{1}{[2\pi(\sigma_\pi^2 + \sigma_f^2)]^{1/2}} \exp\left\{-\frac{(x_i - \mu_\pi)^2}{2(\sigma_\pi^2 + \sigma_f^2)}\right\}$$

$$= [2\pi(\sigma_\pi^2 + \sigma_f^2)]^{-p/2} \exp\left\{-\frac{\sum_{i=1}^{p}(x_i - \mu_\pi)^2}{2(\sigma_\pi^2 + \sigma_f^2)}\right\}$$

$$= [2\pi(\sigma_\pi^2 + \sigma_f^2)]^{-p/2} \exp\left\{-\frac{ps^2}{2(\sigma_\pi^2 + \sigma_f^2)}\right\} \exp\left\{\frac{-p(\bar{x} - \mu_\pi)^2}{2(\sigma_\pi^2 + \sigma_f^2)}\right.$$

where $\bar{x} = \sum_{i=1}^{p} x_i / p$. $\quad s^2 = \sum_{i=1}^{p} (x_i - \bar{x})^2 / p$

To maximize $m(x|\pi)$ with respect to μ_π and σ_π^2, we first observe that μ_π has to be $\bar{x}$

Then we optimize , with respect to σ_π^2,

$$\psi(\sigma_\pi^2) = \left[2\pi(\sigma_\pi^2+\sigma_f^2)\right]^{-p/2} \exp\left\{\frac{-ps^2}{2(\sigma_\pi^2+\sigma_f^2)}\right\}$$

We instead optimize $\log \psi(\sigma_\pi^2)$. Indeed.

$$\frac{d}{d\sigma_\pi^2} \log \psi(\sigma_\pi^2) = \frac{-p/2}{(\sigma_\pi^2+\sigma_f^2)} + \frac{ps^2}{2(\sigma_\pi^2+\sigma_f^2)^2} \overset{\Delta}{=} 0$$

$\Rightarrow \sigma_\pi^2 = s^2 - \sigma_f^2.$

Also we observe that $\sigma_\pi^2 \geq 0.$

Hence $\sigma_\pi^2 = \max\{0, s^2-\sigma_f^2\}$

In conclusion ML-II prior is

$\hat{\pi}_0 = \mathcal{N}(\hat{\mu}_\pi, \hat{\sigma}_\pi^2)$ where $\hat{\mu}_\pi = \bar{x}$ and $\hat{\sigma}_\pi^2 = \max\{0, s^2-\sigma_f^2\}$

<u>Example</u> For any π in the ε-contamination class

$$\Gamma = \{\pi : \pi(\theta) = (1-\varepsilon)\pi_0(\theta) + \varepsilon q(\theta),\ q \in \mathcal{Q}\}$$

$$m(x|\pi) = \int f(x|\theta) \overbrace{\left[(1-\varepsilon)\pi_0(\theta) + \varepsilon q(\theta)\right]}^{\pi(\theta)} d\theta$$

$$= (1-\varepsilon)\, m(x|\pi_0) + \varepsilon \cdot m(x|q)$$

↝ The ML-II prior can be found by maximizing $m(x|q)$ over $q \in \mathcal{Q}$, and using the maximizing $\hat{q}$ in the expression for π.

If $\mathcal{Q}$ is the class of anything.

$$m(x|q) = \int f(x|\theta)\, q(\theta)\, d\theta$$

To maximize this, we choose ~~θ's prior~~ q to be a one-point distribution centered at $\hat{\theta}$ ($\hat{\theta}$ is the ML of θ given data).

Then $$\hat{\pi} = (1-\varepsilon)\pi_0 + \varepsilon\langle\hat{\theta}\rangle$$

- The Moment Approach $\left\{\begin{array}{l}\text{"given functional form" of } \Gamma \\ \text{relate prior moments to moments of marginals}\end{array}\right.$

<u>Lemma</u> Let $\mu_f(\theta)$, $\sigma_f^2(\theta)$ be conditional mean and variance of X (conditioned on θ, i.e., $f(x|\theta)$). Let μ_m and σ_m^2 denote the marginal mean and variance of X (w.r.t. $m(x)$) Assuming these quantities exist, then

$$\mu_m = \mathbb{E}^\pi[\mu_f(\theta)]$$

$$\sigma_m^2 = \mathbb{E}^\pi[\sigma_f^2(\theta)] + \mathbb{E}^\pi[(\mu_f(\theta)-\mu_m)^2]$$

<u>Proof.</u> (the continuous case)

$$\mu_m = \mathbb{E}^m[X] = \int_{\mathcal{X}} x\cdot m(x)\,dx$$

$$= \int_{\mathcal{X}} x\cdot \int_\Theta f(x|\theta)\pi(\theta)\,d\theta\,dx$$

$$= \int_\Theta \pi(\theta) \underbrace{\int_{\mathcal{X}} x f(x|\theta)\,dx}_{\mu_f(\theta)}\,d\theta$$

$$= \int_\Theta \pi(\theta)\cdot \mu_f(\theta)\,d\theta$$

$$\sigma_m^2 = \mathbb{E}^m[(X-\mu_m)^2] = \int_{\mathcal{X}} (x-\mu_m)^2 \int_\Theta f(x|\theta)\pi(\theta)\,d\theta\,dx$$

$$= \int_\Theta \pi(\theta)\cdot \int_{\mathcal{X}} (x-\mu_m)^2 f(x|\theta)\,dx\,d\theta$$

$$= \mathbb{E}^\pi[\,\mathbb{E}^f_\theta[x-\mu_m]^2]$$

$$= \mathbb{E}^\pi[\,\mathbb{E}^f_\theta[(x-\mu_f(\theta)) + (\mu_f(\theta)-\mu_m)]^2]$$

$$= \mathbb{E}^\pi\Big[\ \mathbb{E}^f_\theta[\overbrace{x-\mu_f(\theta)}^{\sigma_f^2(\theta)}]^2 + \overbrace{2\,\mathbb{E}^f_\theta[(x-\mu_f(\theta)\cdot(\mu_f(\theta)-\mu_m)]}^{0} + \mathbb{E}^f_\theta[\mu_f(\theta)-\mu_m]^2\ \Big]$$

$$= \mathbb{E}^\pi[\sigma_f^2(\theta)] + \mathbb{E}^\pi[\mu_f(\theta)-\mu_m]^2$$

~~Collo~~

Corollary i) If $\mu_f(\theta)=\theta$, then $\mu_m=\mu_\pi$. where $\mu_\pi=\mathbb{E}^\pi[\theta]$, prior mean

ii) If, in addition, $\sigma_f^2(\theta)\,\cancel{\theta}=\sigma_f^2$ is a constant independent of θ, then $\sigma_m^2=\sigma_f^2+\sigma_\pi^2$. where σ_π^2 is the prior variance.

→ μ_m, σ_m^2 can usually be estimated by ML-II or subjective experience, We can then solve the prior.

Example Let $X\sim\mathcal{N}(\theta,1)$. $\Gamma=\{\mathcal{N}(\mu_\pi,\sigma_\pi^2)\}$

If we know, either by subjective experince ~~of~~ or type II ML, that preetive density yield $\mu_m=1$, $\sigma_m^2=3$. of X

Using corollary, we have $\underset{\overset{\shortparallel}{1}}{\mu_m}=\mu_\pi$, $\underset{\overset{\shortparallel}{3}}{\sigma_m^2}=1+\sigma_\pi^2$

Thus, we conclude $\pi=\mathcal{N}(1,2)$

- The Distance Approach to Prior Selection

$\Big\{$ Γ not a "given functional form"
considerable information available about m.

⇒ 1° estimate m.

2° use the integral relationship $m(x)=\int_\Theta f(x|\theta)\,dF^\pi(\theta)$ to estimate π

I.e., seek an estimate of π, say $\hat\pi$,

yielding $m_{\hat\pi}(x)=\int_\Theta f(x|\theta)\,dF^{\hat\pi}(\theta)$

is close to $\hat m(x)$.

By "close," ~~by~~ we minimize $KL(\hat m\,\|\,m_{\hat\pi})$, given by

$$KL(\hat m\,\|\,m_{\hat\pi})=\mathbb{E}^{\hat m}\left[\log\frac{\hat m(X)}{m_{\hat\pi}(X)}\right]=\begin{cases}\int_{\mathcal{X}}\hat m(x)\log\left[\frac{\hat m(x)}{m_{\hat\pi}(x)}\right]dx & \text{(continuous)}\\ \sum_{x}\hat m(x)\log\left[\frac{\hat m(x)}{m_{\hat\pi}(x)}\right] & \text{(discrete)}\end{cases}$$

not related to $\hat{\pi}$

$$KL(\hat{m}, m_{\hat{\pi}}) = E^{m}\left[\log \frac{\hat{m}(X)}{m_{\hat{\pi}}(X)}\right] = E^{\hat{m}}[\log \hat{m}(X)] - E^{\hat{m}}[\log m_{\hat{\pi}}(X)]$$

Minimizing $KL(\hat{m} \| m_{\hat{\pi}})$ $\Leftrightarrow$ maximizing $E^{\hat{m}}[\log m_{\hat{\pi}}(X)]$

CASE: $\Theta = \{\theta_1, \cdots \theta_K\}$ finite.

Let $p_i = \hat{\pi}(\theta_i)$.

Then $m_{\hat{\pi}}(x) = \sum_{i=1}^{K} f(x|\theta_i) \cdot p_i$

The problems becomes to maximize

$$E^{\hat{m}}[\log(\sum_{i=1}^{K} f(X|\theta_i)p_i] = \sum_{j=1}^{n} \frac{1}{n}\log(\sum_{i=1}^{k} f(x_j|\theta_i)p_i)$$

CASE: Θ continuous, the problem becomes very difficult.

☹

- Hierarchical Priors

stage 1: $\Gamma = \{\pi_1(\theta|\lambda) : \pi_1$ is of a given functional form, $\lambda \in \Lambda\}$

stage 2: $\pi_2(\lambda)$ on hyperparameter λ

- more robust
- usually use noninformative prior
- more stages are rarely used
- hierarchical prior is a convenient representation

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

is the standard prior distribution.

for criticism