Prior distributions

Bologna, marzo 2010

Brunero Liseo
Dipartimento MEMOTEF
Sapienza Università di Roma
brunero.liseo@uniroma1.it

Outline

- "Truly subjective" priors
- Conjugate priors
- Noninformative priors
  - Introduction
  - Most popular methods (Jeffreys' and Reference priors)

Introduction

- The prior distribution introduces the extra-experimental information in the process.
- It should be entirely subjective (to express personal opinion and knowledge about the problem)
- This is not always easy to do, sometimes impossible!

There is an ongoing and vivid debate among Subjective and Objective Bayesians.

Subjective priors

Some methods to help you in the elicitation process

- Histogram (or Quantile) method
- Relative comparisons method
- Specific functional form

Histogram (or Quantile)

- Partition the parameter space Θ into subsets and evaluate your personal probability of each subset. chiediamo quale probabilità assegnare ai singoli intervalli.
- This is equivalent to choose some quantiles of the prior $\pi(\theta)$.
- Choose a functional form which is compatible with your elicitation and match the values of the (hyper)-parameters.

Relative comparisons

Illustrate the method via a asimple example:

- Suppose $\Theta = [0, 1]$. Try to figure out what is the most "probable" value , say $\bar{\theta}$(and the least probable, say $\underline{\theta}$) of the parameter θ.
- Suppose $\bar{\theta} = 5/6$ and $\underline{\theta}\_0$.
- Suppose also that $\bar{\theta}$ is threee times more probable than $\underline{\theta}$.
- continue to compare, until your observe some familiar functional form for $\pi(\theta)$.

Functional form

Choose a particular functional form you like and fix the values of the k parameters of the distributions by

- k quantile matching
- k moment matching

Gaussian Example

We observe n replications of a measurement. The model is

$$X_i = \theta + \varepsilon_i, \qquad i = 1, \cdots n;$$

$\varepsilon_i \sim N(0, \sigma^2)$ and $\varepsilon_i \perp\!\!\!\perp \varepsilon_j$, $i, j = 1, \ldots, n$. Likelihood is

$$L(\theta; \mathbf{x}) \propto \exp\left\{-\frac{n}{\sigma^2}(\bar{x} - \theta)^2\right\}$$

How to determine $\pi(\theta)$?

- We are practically sure that θ lies between l_1 and l_2, that is

$$\mu_0 \pm 3\sigma_0$$

- we consider more likely the values close to μ_0 than values far from μ_0
- uncertainty around μ_0 is symmetric

A possible (not the only one!) probability law which satisfies the above requirements is a $N(\mu_0, \sigma_0^2)$, with

$$\mu_0 - 3\sigma_0 = l_1, \qquad \mu_0 + 3\sigma_0 = l_2$$

that is

$$\mu_0 = \frac{l_1 + l_2}{2}, \qquad \sigma_0 = \frac{l_1 - l_2}{6}$$

Then, it is enough to elicitate l_1 and l_2 and assume symmetry to obtain a prior for θ.

σ_0 represents a measure of our uncertainty

μ_0 is our prior guess.

Conjugate priors

Let $(X_1, X_2, \cdots, X_n)$ be n r.v. i.i.d. conditionally on $\theta \in \Theta$.

Assumeamo that the r.v. have pdf or pmf denoted by $p(x \mid \theta)$.

The likelihood function for θ is then

$$L(\theta) \propto \prod_{j=1}^{n} p(x_j \mid \theta).$$

A probability distribution $\pi(\theta)$ is conjugate to a given statistical model (or to the corresponding likelihood function) $L(\theta)$, if the functional form of the posterior distribution is the same as the one of the prior distribution, no matter what is the observed sample nor the sample size.

Example [Beta-binomial]**:** We already know that, for a Bernoulli (or Binomial) model,

- a Beta(α, β) prior produces a Beta($\alpha + k$, $\beta + n - k$) posterior, k being the number of successes in n trials.

Example [Exponential-Gamma]:

Let $(X_1, \cdots, X_n) \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$, i.e., $j = 1, \cdots, n$

$$f(x_j \mid \lambda) = \lambda \exp\{-\lambda x_j\} I_{[0,\infty]}(x_j).$$

The likelihood is

$$L(\lambda) \propto \lambda^n \exp\left\{-\lambda \sum_{j=1}^{n} x_j\right\},$$

A prior distribution which is conjugate to $L(\lambda)$ is fiven by the Gamma(ν, α) density

$$\pi(\lambda) = \frac{\alpha^\nu}{\Gamma(\nu)} \exp\{-\alpha\lambda\} \lambda^{\nu-1}.$$

In this parameterization

$$\mathbf{E}(\lambda) = \frac{\nu}{\alpha}; \qquad \text{Var}(\lambda) = \frac{\nu}{\alpha^2}$$

It can be easy checked that

$$\pi(\lambda \mid \mathbf{x}) \propto \lambda^{n+\nu-1} \exp\{-\lambda(\alpha + n\bar{x})\}.$$

Then $\pi(\lambda \mid \mathbf{x})$ is still a Gamma density, that is

$$\lambda \mid \mathbf{x} \sim \text{Gamma}(\nu^*, \alpha^*),$$

with parameters updated by the sample

$$\alpha^* = \alpha + n\bar{x} \quad \text{e} \quad \nu^* = \nu + n.$$

Model	**Prior**	**Posterior**	Notation
$Be(\theta)$	Beta(α, β)	Beta$(\alpha + k, \beta + n - k)$	$k =$ number of successes
$N(\mu, \sigma_0^2)$	$N(\mu_0, , \tau^2)$	$N(\frac{\mu_0\sigma^2 + \bar{x}n\tau^2}{\sigma^2 + n\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2})$	σ_0^2 noto
Poi(θ)	Gamma(λ, δ)	Gamma$(\lambda + n\bar{x}, \delta + n)$	
Esp(θ)	Gamma(λ, δ)	Gamma$(\lambda + n, \delta + n\bar{x})$	
U$(0, \theta)$	Pa(α, ξ)	Pa$(\alpha + n, w)$	$w = \max(x_{(n)}, \xi)$

Tabella 1: Main conjugate distributions

Some comments

- Need elicitations only for some hyper-parameters
- Link with noninformative priors for specific values
- Links with Empirical Bayes

Jeffreys' method

- The Jeffreys general rule prior say that the objective prior must be proportional to the positive square root of the determinant of the Fisher information matrix.

$$\pi^J(\theta) \propto |\det(I(\boldsymbol{\theta}))|^{1/2}$$

- It remains invariant under one-to-one reparameterization.

Proof

$I(\theta) = \mathbf{E}\left(\frac{\partial \ell(\theta)}{\partial \theta}\right)^2$

Let $\phi = g(\theta)$ be a one-to-one transformation (reparametrization).

Let $f^*(\mathbf{X}|\phi) = f(\mathbf{X}|g^{-1}(\phi))$ be the pdf/pmf of $\mathbf{X}$ in the ϕ-parameterization. Then

$$
\begin{aligned}
I_\theta(\theta) &= \mathsf{Var}_\theta\left(\frac{\partial \log f(\mathbf{X}|\theta)}{\partial \theta}\right) \\
&= \mathsf{Var}_\theta\left(\frac{\partial \phi}{\partial \theta}\frac{\partial \log f^*(\mathbf{X}|\phi)}{\partial \phi}\right) \\
&= \left(\frac{\partial g(\theta)}{\partial \phi}\right)\mathsf{Var}_{\theta=\theta(\phi)}\left(\frac{\partial \log f^*(\mathbf{X}|\phi)}{\partial \phi}\right)\left(\frac{\partial g(\theta)}{\partial \phi}\right)^T \\
&= \left(\frac{\partial g(\theta)}{\partial \phi}\right)\mathsf{Var}_\phi\left(\frac{\partial \log f^*(\mathbf{X}|\phi)}{\partial \phi}\right)\left(\frac{\partial g(\theta)}{\partial \phi}\right)^T \\
&= \left(\frac{\partial g(\theta)}{\partial \phi}\right) I_\phi(\phi)\left(\frac{\partial g(\theta)}{\partial \phi}\right)^T. \qquad (1)
\end{aligned}
$$

Let $\pi_J^{\phi}(\phi) = \{\mathsf{det}(I_{\phi}(\phi))\}^{\frac{1}{2}}$ be the Jeffreys' prior in the ϕ-parameterization. Then by (1) we get that

$$\begin{aligned} \pi_J^{\theta}(\theta) &= \{\mathsf{det}(I_{\theta}(\theta))\}^{\frac{1}{2}} = \{\mathsf{det}(I_{\phi}(\phi))\}^{\frac{1}{2}} \mathsf{abs}\Big\{\mathsf{det}\Big(\frac{\partial g(\theta)}{\partial \phi}\Big)\Big\} \\ &= \pi_J^{\phi}(\phi) \mathsf{abs}\Big\{\mathsf{det}\Big(\frac{\partial g(\theta)}{\partial \phi}\Big)\Big\}, \end{aligned}$$

which establishes the invariance of Jeffreys' prior under reparameterization.

Example (Scalar Poisson)

Let $X_1, \cdots, X_n$ be i.i.d. Po(θ).
Likelihood is è

$$L(\theta; \mathbf{x}) \propto \exp\{-n\theta\}\, \theta^{\sum x_i}$$

We need likelihood for $n = 1$, (iid case)

$$L(\theta; x_1) \propto \exp\{-\theta\}\theta^{x_i}$$

Fisher information:

$$\ell(\theta) \propto -\theta + x_i \log\theta$$

$$\frac{\partial \ell(\theta)}{\partial \theta} = -1 + \frac{x_i}{\theta}$$

$$-\frac{\partial^2 \ell(\theta)}{\partial \theta^2} = \frac{x_i}{\theta^2}$$

Then

$$I(\theta) = \mathbf{E}_\theta \left(\frac{X_i}{\theta^2} \right) = \frac{1}{\theta^2} \mathbf{E}_\theta \left(X_i \right) = \frac{1}{\theta}$$

and

$$\pi^J(\theta) \propto \frac{1}{\sqrt{\theta}}$$

The posterior distribution is

$$\pi(\theta \mid \mathbf{x}) \propto \frac{1}{\sqrt{\theta}} \theta^t \exp\{-n\theta\} = \theta^{t-1/2} \exp\{-n\theta\},$$

that is

$$\theta \mid \mathbf{x} \sim Gamma(n, t + \frac{1}{2})$$

- Jeffreys general rule prior enjoys many optimality properties in the absence of nuisance parameters.
- Maximizes the distance between the prior and the posterior in a certain sense.
- Enjoys probability matching property, i.e. the coverage probability of the resulting Bayesian one-sided credible interval matches asymptotically the coverage probability of the corresponding frequentist confidence interval.
- Under a suitable topology, it is the unique invariant uniform prior (J.K. Ghosh et al., 2006).

Jeffreys prior is still the most popular method at least when the dimension of Θ is 1.

Jeffreys itself suggested some modifications to the rule

- in the presence of location and/or scale parameters
- in the multiparameter case

Example 2

Let $X_i \sim N(\mu_i, 1),\ \ i = 1, \cdots, p$ independent of each other.
Parameter of interest is

$$\theta = \frac{1}{p}\sum \mu_i^2 = \frac{1}{p}\|\mu\|^2$$

The Jeffreys' prior here is

$$\pi^J(\mu_1, \cdots, \mu_p) \propto 1$$

since the Fisher matrix is diagonal. Then, a posteriori,

$$(\mu_1, \cdots, \mu_p) = \boldsymbol{\mu} \sim N_p(\mathbf{x}, \mathbf{1})$$

and

$$p\theta \sim \chi^2_p(\sum x_i^2)$$

This is NOT a "good" posterior distribution for θ. (Stein's Paradox (1961), inadmissibility of the sample mean in more than 2 dimensions).

For exemple,

$$\mathbf{E}^{\pi}\left(\theta|\mathbf{x}\right) = 1 + \frac{\sum x_i^2}{p}.$$

and it is easy to prove that this quantity, taken as an estimator, is NOT consistent

$$\lim_{p\to\infty} \mathbf{E}^{\pi}\left(\theta|\mathbf{x}\right) - \theta = 2$$

The same happens for the posterior mode or median.

The previous example highlights one the most important issue in the selection of good noninformative priors.

- The Jeffreys' method seek for **the noninformative prior for the entire vector** of the parameters (in the previous example, $(\mu_1, \ldots, \mu_p)$
- If the parameter of interest is just a function of it, say θ, this introduces a "bias" into the procedure.

This phenomenon is one of motivations of the development of the reference prior method.

Information contained in a probability distribution

The Entropy $\mathcal{E}$ of a probability measure π is given by

$$\mathcal{E} = -\int_{\mathbf{\Omega}} \pi(\omega) \log \pi(\omega) d\omega$$

A measure of the "distance" between two probability measure is the *Kullback-Leibler number*,

$$K(p;q) = \int_{\mathbf{\Omega}} q \log \frac{q}{p}$$

which is zero iff $q = p$ (a.s. q).

The two quantities are important for the definition of information contained in a given experiment, (Lindley, 1956) based on Shannon (1948).

Shannon-Lindley Information

Given the experiment

$$E_k = (\mathcal{X}_k, \mathbf{\Omega}, \mathcal{P})$$

one defines *"Information contained in E_k"*,
wrt a prior distribution π the quantity

$$I_{E_k}(\pi) = \int_{\mathcal{X}_k} \int_{\mathbf{\Omega}} m(\mathbf{x}_k)\pi(\omega|\mathbf{x}_k) \log \frac{\pi(\omega|\mathbf{x}_k)}{\pi(\omega)} d\omega d\mathbf{x}_k \qquad (2)$$

$I_{E_k}(\pi)$ is the expected value (wrt the marginal distribution $m(x)$ of the K-L number of the "prior" to the "posterior".

It is then reasonable to measure the contribution of the prior π in terms of (2).

It is reasonable but not necessary. In the expressions $I_{E_k}(\pi)$ and $\mathcal{E}$ we do not integrate over the ω values. Rather, we integrate over the $\pi(\omega)$ values.

Reference priors

Bernardo (1979) has proposed the reference priors method.

Two main novelties in the search of π^r

- π^r as the prior which maximises $I_{E_k}(\pi)$
- in the multiparameter case, explicit distinction between the parameter of interest and nuisance parameters.

The method has been refined in the 90's

References

Bernardo, J.M. Reference posterior distributions for Bayesian inference. With discussion. *JRSS* B 41 (1979), no. 2, 113-147

Berger, J.O.; Bernardo, J.M. On the development of reference priors. *Bayesian statistics, 4* 35-60, Oxford Univ. Press, New York, 1992.

Berger, J.O.; Bernardo, J.M. Ordered group reference priors with application to the multinomial problem. *Biometrika* 79 (1992), no. 1, 25-37.

Bernardo, J.M. Reference Analysis . In *Handbook of Statistics*, (C.R.Rao and D. Dey eds.) 2005

Berger, J.O., Bernardo, J.M. and Sun, D. The formal definition of reference priors, Ann. Statist. Volume 37, Number 2 (2009), 905-938.

Construction of reference priors

The exact derivation of π^r causes a series of technical problems not always easy to solve.
We will consider only a "regular" model, where

- there exists a sufficient statistic with the same dimension of the parameter
- the MLE has an asymptotic normal distribution.
- the posterior distribution has an asymptotic normal distribution.

The case of a single parameter

Suppose $\omega \in \mathbb{R}$. The quantity $I_{E_k}(\pi)$ represents the average gain of information which the experiment provides when the prior is $\pi(\omega)$.
Per $k \to \infty$, $I_{E_k}(\pi)$ assumes the meaning of *missing information about the parameter* ω.

Than the prior which maximises $I_{E_\infty}(\pi)$ csn be defined the "least informative".

There is a problem: $I_{E_k}(\pi)$ is (in general) not bounded as a function of k.

Bernardo suggestion.

- Maximise $I_{E_k}(\pi)$ for fixed k
  $\Longrightarrow$ k-reference prior $\pi_k^r(\omega)$

- Define the reference prior as the pointwise limit

$$\pi^r(\omega) = \lim_{k\to\infty} \frac{\pi_k^r(\omega)}{\pi_k^r(\Omega_0)}$$

  and Ω_0 is a compact set.

Comments

- The limit may not exist
- The limit is pointwise and it does not guarantee the convergence in the metrics induced by the K-L
- The limit will be often an improper prior

Heuristic computation of π^r

$$I_{E_k}(\pi) =$$
$$\int_{\mathbf{\Omega}} \pi(\omega) \left[\int_{\mathcal{X}_k} p(\mathbf{x}_k|\omega) \left(\log \pi(\omega|\mathbf{x}_k) - \log \pi(\omega) \right) \right] d\omega d\mathbf{x}_k$$

$$= \int_{\mathbf{\Omega}} \pi(\omega) \log \frac{\exp\{\int_{\mathcal{X}_k} p(\mathbf{x}_k|\omega) \log \pi(\omega|\mathbf{x}_k) d\mathbf{x}_k\}}{\pi(\omega)} d\omega$$

This is a problem of the class

$$\sup_f \int f \log g/f$$

which is maximised when $f \propto g$. (Calculus of variations result) It follows that

$$\pi^r(\omega) \propto \exp\{\int_{\mathcal{X}_k} p(\mathbf{x}_k|\omega) \log \pi(\omega|\mathbf{x}_k) d\mathbf{x}_k\}$$

(implicit solution).

In the regular case, there exists an MLE $\hat{\omega}_k$, such that

$$\exp\{\int_{\mathcal{X}_k} p(\mathbf{x}_k|\omega)\log\pi(\omega|\mathbf{x}_k)d\mathbf{x}_k\} =$$

$$\exp\{\int_{\mathbb{R}} p(\hat{\omega}_k|\omega)\log\pi(\omega|\hat{\omega}_k)d\hat{\omega}_k\} \qquad (3)$$

Inoltre

$$\pi(\omega|\hat{\omega}_k) \sim N(\omega; \hat{\omega}_k, [k\pi(\hat{\omega}_k)]^{-1})$$

and the (3) is approximately equal to a

$$\exp\{\int_{\mathbb{R}} p(\hat{\omega}_k|\omega)[\log \pi(\hat{\omega}_k)^{1/2} - \frac{k}{2}\pi(\hat{\omega}_k)(\omega - \hat{\omega}_k)^2]d\hat{\omega}_k\}$$

Assuming that $\hat{\omega}_k$ tends to concentrate around ω, one gets that

$$\pi^r(\omega) \propto \pi(\omega)^{1/2}$$

Commenti

- In the univariate case, under regularity conditions, the reference prior coincides with the Jeffreys' prior.
- A more rigorous derivation of the reference prior is difficult to obtain. One can show that, in some cases, the prior which maximises I_{E_k} is concentrated in a finite number of points.

Un solo parametro di disturbo

Sia ora $\omega = (\theta, \lambda)$ e sia θ il solo parametro di interesse.
In questo caso la Jeffreys' prior è

$$\pi^J(\theta, \lambda) \propto \det(\pi(\theta, \lambda))^{1/2}$$

Nel caso delle reference prior, si cerca quella $\pi(\theta, \lambda)$ che massimizza la distanza d'informazione tra $\pi(\theta|\mathbf{x}_k)$ e $\pi(\theta)$, ovvero

$$\begin{aligned} I_{E_k}(\pi(\theta, \lambda)) &= \int_{\mathbf{\Theta}} \pi(\theta) \left[\int_{\mathcal{X}_k} p(\mathbf{x}_k|\theta) \left(\log \pi(\theta|\mathbf{x}_k)\right.\right. \\ &- \left.\left. \log \pi(\theta)\right)\right] d\omega d\mathbf{x}_k \quad (4) \end{aligned}$$

L'equazione non dipende direttamente da λ (è stato già integrato!) e, analogamente a quanto vista prima,

$$\pi_k^r(\theta) \propto \exp\{ \int_{\mathcal{X}_k} p(\mathbf{x}_k|\theta) \log \pi(\theta|\mathbf{x}_k) d\mathbf{x}_k \}$$

Tale risultato vale qualunque sia la scelta per $\pi(\lambda|\theta)$.
L'algoritmo delle reference priors suggerisce di

- scegliere $\pi_k^r(\lambda|\theta) \propto H_{22}(\theta,\lambda)^{1/2}$
  (la Jeffreys prior per θ noto)
- massimizzare la (4) con $\pi_k^r(\lambda|\theta)$.

Dunque,

$$\begin{aligned} \pi_k^r(\theta) &\propto \exp\{\int_{\hat\theta}\int_{\hat\lambda} p(\hat\theta,\hat\lambda|\theta) \\ &\times \log N(\theta;\hat\theta, S_{11}(\hat\theta,\hat\lambda)d\hat\theta d\hat\lambda \end{aligned}$$

dove $S = H^{-1}$.

$$\begin{aligned} \pi_k^r(\theta) &\propto \exp\{\int_{\hat\theta}\int_{\hat\lambda}\int_{\Lambda} p(\hat\theta,\hat\lambda|\theta,\lambda)\pi^r(\lambda|\theta) \\ &\times \log N(\theta;\hat\theta, S_{11}(\hat\theta,\hat\lambda)d\lambda d\hat\theta d\hat\lambda \end{aligned}$$

$$
\begin{aligned}
&= \exp\{\int_{\hat{\theta}}\int_{\hat{\lambda}} p(\hat{\theta},\hat{\lambda}|\theta,\lambda)\int_{\Lambda}\pi^r(\lambda|\theta) \\
&\times \log N(\theta;\hat{\theta},S_{11}(\hat{\theta},\hat{\lambda})d\lambda d\hat{\theta}d\hat{\lambda}
\end{aligned}
$$

$$
\cong \exp\{\frac{1}{2}\int_{\Lambda}\pi^r(\lambda|\theta)\log S_{11}^{-1}(\theta,\lambda)d\lambda\}
$$

Poiché $S_{11} = \frac{H_{22}}{\det(H)}$,

$$
\pi^r(\theta,\lambda) = \pi(\lambda|\theta)\exp\{\frac{1}{2}\int_{\Lambda}\pi(\lambda|\theta)\log\frac{\det(H)}{H_{22}}d\lambda\}.
$$

Fin qui abbiamo trascurato il problema della non integrabilità delle leggi a priori. All'interno dell'algoritmo tale problema si aggira considerando una successione di compatti che "invadono" Θ e sui quali definiamo una successione di reference priors.

Algoritmo delle reference prior per il caso bidimensionale

1 Si parte dal nucleo della distribuzione non informativa per $\lambda|\theta$,

$$\pi^*(\lambda|\theta) \propto \sqrt{H_{22}(\theta,\lambda)}$$

2 Normalizzazione di $\pi^*(\lambda|\theta)$:

- se $\pi^*(\lambda|\theta)$ è integrabile (propria), poni $\pi(\lambda|\theta) = \pi^*(\lambda|\theta)k(\theta)$ con $k(\theta)^{-1} = \int_\Lambda \pi^*(\lambda|\theta)d\lambda$;
- se $\pi^*(\lambda|\theta)$ non è integrabile (impropria) costruisci una successione di sezioni di s.i. $\Lambda_1(\theta), \Lambda_2(\theta), \cdots, \Lambda_m(\theta), \cdots \to \Lambda$, definiti per ogni θ, tali che, per ogni $m \in \mathbb{N}$, $\pi_m(\lambda|\theta) = \pi^*(\lambda|\theta)k_m(\theta)$ con $k_m(\theta)^{-1} = \int_{\Lambda_m(\theta)} \pi^*(\lambda|\theta)d\lambda$.

3 Determina la marginale di θ, su Λ_m,

$$\pi_m(\theta) \propto \exp\left\{\frac{1}{2}\pi_m(\lambda|\theta)\log\frac{\det H(\theta,\lambda)}{H_{22}(\theta,\lambda)}\right\}d\lambda.$$

4 Poni

$$\pi^r(\theta,\lambda) = \lim_{m\to\infty}\frac{k_m(\theta)\pi_m(\theta)}{k_m(\theta_0)\pi_m(\theta_0)}\frac{\pi_m(\lambda|\theta)}{\pi(\lambda|\theta_0)}$$

Esempio 1 (continua)

La funzione di verosimiglianza è

$$L(\omega) = \omega^t (1-\omega)^{k-t}, \qquad t = \sum x_i,$$

L'informazione di Fisher vale $\pi(\omega) = \frac{1}{\omega(1-\omega)}$ e, di conseguenza

$$\pi^J(\omega) = \pi^r(\omega) = \frac{1}{\pi} \omega^{-1/2} (1-\omega)^{-1/2}$$

Nota: π^J e π^r sono distribuzioni proprie ma non uniformi...

Esempio 3: Modello Trinomiale

Riconsideriamo ora l'esempio precedente ma suddividiamo i risultati possibili non più in due categorie bensì in tre, ovvero

$$X_i = \begin{cases} S & N & F \\ \omega_1 & \omega - 2 & 1 - \omega_1 - \omega_2 \end{cases} \tag{4}$$

Il parametro d'interesse (lo stesso di prima) è ora $\theta = \omega_1$ ma nel modello è presente anche $\lambda = \omega_2$.

$$H(\theta, \lambda) = \frac{1}{1 - \theta - \lambda} \times \begin{pmatrix} \frac{1-\lambda}{\theta} & 1 \\ 1 & \frac{1-\theta}{\lambda} \end{pmatrix}$$

La Jeffreys prior è dunque

$$\pi^J(\theta, \lambda) \propto \frac{1}{\sqrt{\theta\lambda(1 - \lambda - \theta)}}$$

Calcolo della reference prior

- $\pi^*(\lambda|\theta) \propto \sqrt{H_{22}(\theta,\lambda)} = \frac{1}{\sqrt{\lambda(1-\lambda-\theta)}}$
- $\pi^(\lambda|\theta) = k(\theta)\frac{1}{\sqrt{\lambda(1-\lambda-\theta)}}\ I_{[0,1-\theta]}(\lambda)$
- $\pi(\theta) = \exp\left\{\frac{1}{2}\int_{\Lambda(\theta)} k(\theta)\frac{1}{\sqrt{\lambda(1-\lambda-\theta)}}\log\frac{1}{\theta(1-\theta)}d\lambda\right\}$
  $= \frac{1}{\sqrt{\theta(1-\theta)}}$
- $\pi^r(\theta,\lambda) \propto \frac{1}{\sqrt{\theta\lambda(1-\theta)(1-\theta-\lambda)}}$

Confronto tra $\pi^r(\theta, \lambda)$ e $\pi^J(\theta, \lambda)$

La natura differente delle due distribuzioni si può notare considerando le corrispondenti marginali per θ.

- $\pi^r(\theta) = \frac{1}{\pi} \frac{1}{\sqrt{\theta(1-\theta)}}$
  $\Rightarrow E(\theta|\pi^r) = 1/2$

- $\pi^J(\theta) = \frac{1}{2}\theta^{-1/2}$
  $\Rightarrow E(\theta|\pi^J) = 1/3$

Questa differenza è ancor più accentuata nel caso generale con h possibili risultati.

Si vede facilmente che, in questo caso

- $E(\theta_i|\pi^J) = \frac{1}{h}\ i = 1, \cdots, h.$
- $E(\theta_i|\pi^r) = \frac{1}{2^i}\ i = 1, \cdots, h.$

Il problema di Fieller

Siano

$$X \sim N(\omega_1, 1) \quad Y \sim N(\omega_2, 1)$$

- Parametro di interesse $\theta = \frac{\omega_1}{\omega_2}$
- Parametro di disturbo $\lambda = \text{sgn}(\omega_2)\sqrt{\omega_1 + \omega_2}$ (ortogonale)

Matrice d'informazione

$$\pi(\theta, \lambda) = \begin{pmatrix} \frac{\lambda^2}{(1+\theta^2)^2} & 0 \\ 0 & 1 \end{pmatrix}$$

Jeffreys prior

$$\pi^J(\theta, \lambda) \propto \frac{|\lambda|}{1+\theta^2}$$

Reference Prior

$A_m = (-a_m < \omega_1 < a_l) \times (-b_m < \omega_2 < b_l)$
che diventa

- $\frac{a_m}{\theta}\sqrt{(1+\theta^2)} < \lambda < -\frac{a_m}{\theta}\sqrt{(1+\theta^2)}$ per $\theta < -\frac{a_m}{b_m}$
- $-b_m\sqrt{1+\theta^2} < \lambda < b_m\sqrt{1+\theta^2}$ per $|\theta| < \frac{a_m}{b_m}$
- $-\frac{a_m}{\theta}\sqrt{(1+\theta^2)} < \lambda < \frac{a_m}{\theta}\sqrt{(1+\theta^2)}$ per $\theta > \frac{a_m}{b_m}$

Ne segue che

- $\pi^*_m(\lambda|\theta) \propto 1$
- $\pi_m(\lambda|\theta) = k_m(\theta)$

La reference prior è dunque

$$\pi_m(\theta,\lambda) \propto k_m(\theta) \times \exp\left\{\tfrac{1}{2} k_m(\theta) \int_{A_m} \left(\log \lambda^2 - \log(1+\theta^2)^2\right) d\lambda\right\}$$

$$= k_m(\theta) \exp\left\{k_m(\theta) \int_{A_m} \log|\lambda| d\lambda\right\} \tfrac{1}{1+\theta^2} \rightarrow \pi^r(\theta,\lambda) \propto \tfrac{1}{1+\theta^2}$$

Va notato che questa è la reference prior quando il parametro d'interesse è $\theta = \omega_1/\omega_2$. Se ad esempio fossimo interessati a $\xi = \omega_1\omega_2$ il risultato sarebbe differente mentre la Jeffreys prior per ξ si otterrebbe attraverso un cambio di variabile da $\pi^J(\omega_1,\omega_2) \propto 1$.

Esempio 2 (continua)

Sia $X_i \sim N(\mu_i, 1),\ \ i = 1, \cdots, p$. Si vuole stimare

$$\theta = \frac{1}{p}\sum \mu_i^2 = \frac{1}{p}||\boldsymbol{\mu}||^2$$

Una scelta opportuna per il parametro di disturbo è $\lambda = \boldsymbol{\mu}/||\boldsymbol{\mu}||$, la "direzione" di $\boldsymbol{\mu}$ sulla superficie dell'ipersfera di raggio unitario. E' naturale, e l'algoritmo delle reference priors lo conferma, assegnare una legge a priori uniforme per $\lambda|\theta$, e poi determinare la marginale per θ.
Ragionare in termini di (θ, λ) permette di "risolvere" il problema emerso con la Jeffreys prior. L'uso di $\pi(\boldsymbol{\mu}) = 1$ implicava, infatti,

$$\pi(\theta, \lambda) = \left(\frac{1}{\theta}\right)^{\frac{p-2}{2}} \pi(\lambda|\theta)$$

Il caso multiparametrico

Consideriamo il caso $\omega = (\omega_1, \omega_2, \cdots, \omega_k)$. Il metodo si generalizza in modo ovvio.

- [Passo 1] Dividi i k parametri in p gruppi $(\theta_{(1)}, \cdots, \theta_{(p)})$
- [Passo 2] Determina una successione di compatti

$$\mathbf{\Omega}_1 \subseteq \mathbf{\Omega}_2 \subseteq \cdots \rightarrow \mathbf{\Omega}$$

- [Passo 3] Calcola sul generico compatto $\mathbf{\Omega}_m$, la reference prior per $\theta_{(p)}$ dati gli altri, ovvero

$$\pi_m(\theta_{(p)}|\theta(1), \cdots, \theta_{(p-1)})$$

- [Passo 4] Elimina il parametro $\theta_{(p)}$ per integrazione e considera il modello marginale con $p-1$ gruppi di parametri

Algoritmo 2

- [Passo 5] Ripeti i passi 3 e 4 per $\theta_{(j)}$, per $j = p-1, \cdots, 2$.
- [Passo 6] Definisci

$$\begin{aligned} & \pi_m(\boldsymbol{\theta}) = \pi_m(\theta_{(p)}|\theta(1), \cdots, \theta_{(p-1)})) \\ \times \quad & \pi_m(\theta_{(p-1)}|\theta(1), \cdots, \theta_{(p-2)})) \\ \times \quad & \cdots \times \pi_m(\theta_1) \end{aligned}$$

- [Passo 7] Normalizzazione di π_m

$$\pi^r(\boldsymbol{\theta}) = \lim_{m\to\infty} \frac{\pi_m(\boldsymbol{\theta})}{\pi_m(\boldsymbol{\theta}_0)},$$

  dove $\boldsymbol{\theta}_0$ è un opportuno punto interno di $\boldsymbol{\Omega}$.

- [Passo 8] Verifica che

$$\mathbf{E}_\theta \left(KL(\pi_m(\boldsymbol{\theta}|\mathbf{x}), \pi(\boldsymbol{\theta}|\mathbf{x})) \right) \to 0$$

Note tecniche

Nel calcolo effettivo di π^r gli aspetti più complessi riguardano

- [A] Calcolo di $\exp\{\frac{1}{2}\int_{\Lambda_m(\theta)} \pi_m(\lambda|\theta) \log \frac{\det(H)}{H_{22}} d\lambda\}$
- [B]

$$\pi^r(\theta,\lambda) = \lim_{m\to\infty} \frac{k_m(\theta)\pi_m(\theta)}{k_m(\theta_0)\pi_m(\theta_0)} \frac{\pi_m(\lambda|\theta)}{\pi(\lambda|\theta_0)}$$

Il più delle volte il calcolo di [B] semplifica il passo [A]. Infatti

$$[A] \approx K_m + C_m\Psi(\theta) + D_m(\theta)$$

dove $K_m \to \infty$, $C_m \to C$, $D_m \to 0$
Ne segue che la parte relativa ad [A] del limite [B] vale

$$\exp\{\frac{1}{2}C\Psi(\theta)\}$$

Matching Priors

Una interpretazione del termine "non informativa" per una data π è che le inferenze conseguenti l'uso di π abbiano un buon comportamento frequentista. In particolare si guarda alla **Probabilità Frequentista di Ricoprimento** (PFR).
Data una legge a priori π, si considera

$$\pi(\cdot) \longrightarrow \pi(\cdot|X) \longrightarrow C_\pi(X, 1-\alpha)$$

dove C è l'insieme di credibilità ad una coda.
Se la PFR è tale che

$$P(\theta \in C_\pi(X, 1-\alpha)|\theta) = 1 - \alpha + O(n^{-\frac{\gamma}{2}}),$$

allora π è una matching priors di ordine γ.

Tibshirani (1989) ha dimostrato che nel caso di parametro reale d'interesse, ortogonale a tutti i parametri di disturbo,

$$\pi^T(\theta, \lambda) \propto g(\lambda)\sqrt{H_{11}(\theta, \lambda)}$$

è una matching prior $(\forall g)$ di ordine 1.

- Legami con π^J e con π^r