Chapter 2
Bayesian Inference

… some rule could be found, according to which we ought to estimate the chance that the probability for the happening of an event perfectly unknown, should lie between any two named degrees of probability, antecedently to any experiments made about it; …
An Essay towards solving a Problem in the Doctrine of Chances
By the late Rev. Mr. Bayes...

The goal of *statistical inference* is to get information from experimental observations about quantities (parameters, models,…) on which we want to learn something, be them directly observable or not. Bayesian inference[1] is based on the *Bayes rule* and considers probability as a measure of the degree of knowledge we have on the quantities of interest. Bayesian methods provide a framework with enough freedom to analyze different models, as complex as needed, using in a natural and conceptually simple way all the information available from the experimental data within a scheme that allows to understand the different steps of the learning process:

(1) state the knowledge we have before we do the experiment;
(2) how the knowledge is modified after the data is taken;
(3) how to incorporate new experimental results.
(4) predict what shall we expect in a future experiment from the knowledge acquired.

It was Sir R.A Fisher, one of the greatest statisticians ever, who said that "The Theory of Inverse Probability (that is how Bayesianism was called at the beginning of the XX century) is founded upon an error and must be wholly rejected" although, as time went by, he became a little more acquiescent with Bayesianism. You will see that Bayesianism is great, rational, coherent, conceptually simple,… "even useful",… and worth to, at least, take a look at it and at the more detailed references on the subject

[1] For a gentle reading on the subject see [1].

C. Maña, *Probability and Statistics for Particle Physics*,
UNITEXT for Physics, DOI 10.1007/978-3-319-55738-0_2

given along the section. At the end, to quote Lindley, "Inside every non-Bayesian there is a Bayesian struggling to get out". For a more classical approach to Statistical Inference see [2] where most of what you will need in Experimental Physics is covered in detail.

2.1 Elements of Parametric Inference

Consider an experiment designed to provide information about the set of parameters $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_k\} \in \Theta \subseteq R^k$ and whose realization results in the random sample $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$. The inferential process entails:

(1) Specification of the probabilistic model for the random quantities of interest; that is, state the joint density:

$$p(\boldsymbol{\theta}, \boldsymbol{x}) = p(\theta_1, \theta_2, \ldots, \theta_k, x_1, x_2, \ldots, x_n); \qquad \boldsymbol{\theta} = \in \Theta \subseteq R^k; \qquad \boldsymbol{x} \in X$$

(2) Conditioning the observed data ($\boldsymbol{x}$) to the parameters ($\boldsymbol{\theta}$) of the model:

$$p(\boldsymbol{\theta}, \boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})$$

(3) Last, since $p(\boldsymbol{\theta}, \boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{x})\, p(\boldsymbol{x})$ and

$$p(\boldsymbol{x}) = \int_{\Theta} p(\boldsymbol{x}, \boldsymbol{\theta})\, d\boldsymbol{\theta} = \int_{\Theta} p(\boldsymbol{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

we have (*Bayes Rule*) that:

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\displaystyle\int_{\Theta} p(\boldsymbol{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}) d\boldsymbol{\theta}}$$

This is the basic equation for parametric inference. The integral of the denominator does not depend on the parameters ($\boldsymbol{\theta}$) of interest; is just a normalization factor so we can write in a general way;

$$p(\boldsymbol{\theta}|\boldsymbol{x}) \propto p(\boldsymbol{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})$$

Let's see these elements in detail:

$p(\boldsymbol{\theta}|\boldsymbol{x})$: This is the *Posterior Distribution* that quantifies the knowledge we have on the parameters of interest $\boldsymbol{\theta}$ conditioned to the observed data $\boldsymbol{x}$ (that is, after the experiment has been done) and will allow to perform inferences about the parameters;

$p(\boldsymbol{x}|\boldsymbol{\theta})$: The *Likelihood*; the sampling distribution considered as a function of the parameters $\boldsymbol{\theta}$ for the *fixed* values (already observed) $\boldsymbol{x}$. Usually, it is written as $\ell(\boldsymbol{\theta}; \boldsymbol{x})$ to stress the fact that it is a function of the parameters. The experimental results modify the prior knowledge we have on the parameters $\boldsymbol{\theta}$ only through the likelihood so, for the inferential process, we can consider the likelihood function defined up to multiplicative factors provided they do not depend on the parameters.

$p(\boldsymbol{\theta})$: This is *a reference function*, independent of the results of the experiment, that quantifies or expresses, in a sense to be discussed later, the knowledge we have on the parameters $\boldsymbol{\theta}$ *before* the experiment is done. It is termed *Prior Density* although, in many cases, it is an improper function and therefore not a probability density.

2.2 Exchangeable Sequences

The inferential process to obtain information about a set of parameters $\boldsymbol{\theta} \in \Theta$ of a model $X \sim p(x|\boldsymbol{\theta})$ with $X \in \Omega_X$ is based on the realization of an experiment $e(1)$ that provides an observation $\{x_1\}$. The $n-$fold repetition of the experiment under the same conditions, $e(n)$, will provide the random sample $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$ and this can be considered as a draw of the n-dimensional random quantity $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ where each $X_i \sim p(x|\boldsymbol{\theta})$.

In Classical Statistics, the inferential process makes extensive use of the idea that the observed sample is originated from a sequence of *independent and identically distributed* (iid) random quantities while Bayesian Inference rests on the less restrictive idea of *exchangeability* [3]. An infinite sequence of random quantities $\{X_i\}_{i=1}^{\infty}$ is said to be *exchangeable* if *any* finite sub-sequence $\{X_1, X_2, \ldots, X_n\}$ is *exchangeable*; that is, if the joint density $p(x_1, x_2, \ldots, x_n)$ is invariant under *any* permutation of the indices.

The hypothesis of *exchangeability* assumes a symmetry of the experimental observations $\{x_1, x_2, \ldots, x_n\}$ such that the subscripts which identify a particular observation (for instance the order in which they appear) are irrelevant for the inferences. Clearly, if $\{X_1, X_2, \ldots X_n\}$ are iid then the conditional joint density can be expressed as:

$$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i)$$

and therefore, since the product is invariant to reordering, is an *exchangeable* sequence. The converse is not necessarily true[2] so the hypothesis of exchangeability is weaker than the hypothesis of independence. Now, if $\{X_i\}_{i=1}^{\infty}$ is an exchangeable

[2]It is easy to check for instance that if X_0 is a non-trivial random quantity independent of the X_i, the sequence $\{X_0 + X_1, X_0 + X_2, \ldots X_0 + X_n\}$ is exchangeable but not iid.

sequence of real-valued random quantities it can be shown that, for *any* finite subset, there exists a parameter $\boldsymbol{\theta} \in \Theta$, a parametric model $p(x|\boldsymbol{\theta})$ and measure $d\mu(\boldsymbol{\theta})$ such that[3]:

$$p(x_1, x_2, \ldots, x_n) = \int_\Theta \prod_{i=1}^{n} p(x_i|\boldsymbol{\theta})\, d\mu(\boldsymbol{\theta})$$

Thus, any finite sequence of exchangeable observations is described by a model $p(x|\boldsymbol{\theta})$ and, if $d\mu(\boldsymbol{\theta}) = p(\boldsymbol{\theta})d\boldsymbol{\theta}$, there is a prior density $p(\boldsymbol{\theta})$ that we may consider as describing the available information on the parameter $\boldsymbol{\theta}$ before the experiment is done. This justifies and, in fact, leads to the Bayesian approach in which, by formally applying Bayes Theorem

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = p(\boldsymbol{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{x})\, p(\boldsymbol{x})$$

we obtain the *posterior density* $p(\boldsymbol{\theta}|\boldsymbol{x})$ that accounts for the degree of knowledge we have on the parameter after the experiment has been performed. Note that the random quantities of the exchangeable sequence $\{X_1, X_2, \ldots, X_n\}$ are *conditionally independent given* $\boldsymbol{\theta}$ *but not iid* because

$$p(x_j) = \int_\Theta p(x_j|\boldsymbol{\theta})\, d\mu(\boldsymbol{\theta}) \left(\prod_{i(\neq j)=1}^{n} \int_{\Omega_X} p(x_i|\boldsymbol{\theta})\, dx_i \right)$$

and

$$p(x_1, x_2, \ldots, x_n) \neq \prod_{i=1}^{n} p(x_i)$$

There are situations for which the hypothesis of exchangeability can not be assumed to hold. That is the case, for instance, when the data collected by an experiment depends on the running conditions that may be different for different periods of time, for data provided by two different experiments with different acceptances, selection criteria, efficiencies,… or the same medical treatment when applied to individuals from different environments, sex, ethnic groups,… In these cases, we shall have different *units of observation* and it may be more sound to assume *partial exchangeability* within each unit (data taking periods, detectors, hospitals,…) and design a *hierarchical structure* with parameters that account for the relevant information from each unit analyzing all the data in a more global framework.

Note 4: Suppose that we have a parametric model $p_1(x|\boldsymbol{\theta})$ and the exchangeable sample $\boldsymbol{x}_1 = \{x_1, x_2, \ldots, x_n\}$ provided by the experiment $e_1(n)$. The inferences on

[3]This is referred as *De Finetti's Theorem* after B. de Finetti (1930s) and was generalized by E. Hewitt and L.J. Savage in the 1950s. See [4].

the parameters $\boldsymbol{\theta}$ will be drawn from the posterior density $p(\boldsymbol{\theta}|\boldsymbol{x}_1) \propto p_1(\boldsymbol{x}_1|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Now, we do a second experiment $e_2(m)$, statistically independent of the first, that provides the exchangeable sample $\boldsymbol{x}_2 = \{x_{n+1}, x_{n+2}, \ldots, x_{n+m}\}$ from the model $p_2(x|\boldsymbol{\theta})$. It is sound to take as prior density for this second experiment the posterior of the first including therefore the information that we already have about $\boldsymbol{\theta}$ so

$$p(\boldsymbol{\theta}|\boldsymbol{x}_2) \propto p_2(\boldsymbol{x}_2|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{x}_1) \propto p_2(\boldsymbol{x}_2|\boldsymbol{\theta})p_1(\boldsymbol{x}_1|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

Being the two experiments statistically independent and their sequences exchangeable, if they have the same sampling distribution $p(x|\boldsymbol{\theta})$ we have that $p_1(\boldsymbol{x}_1|\boldsymbol{\theta})p_2(\boldsymbol{x}_2|\boldsymbol{\theta}) = p(\boldsymbol{x}|\boldsymbol{\theta})$ where $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2\} = \{x_1, \ldots, x_n, x_{n+1}, \ldots, x_{n+m}\}$ and therefore $p(\boldsymbol{\theta}|\boldsymbol{x}_2) \propto p(\boldsymbol{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})$. Thus, the knowledge we have on $\boldsymbol{\theta}$ including the information provided by the experiments $e_1(n)$ and $e_2(m)$ is determined by the likelihood function $p(\boldsymbol{x}|\boldsymbol{\theta})$ and, in consequence, under the aforementioned conditions the realization of $e_1(n)$ first and $e_2(m)$ after is equivalent, from the inferential point of view, to the realization of the experiment $e(n+m)$.

2.3 Predictive Inference

Consider the realization of the experiment $e_1(n)$ that provides the sample $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$ drawn from the model $p(x|\boldsymbol{\theta})$. Inferences about $\boldsymbol{\theta} \in \Theta$ are determined by the posterior density

$$p(\boldsymbol{\theta}|\boldsymbol{x}) \;\propto\; p(\boldsymbol{x}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})$$

Now suppose that, under the same model and the same experimental conditions, we think about doing a new independent experiment $e_2(m)$. What will be the distribution of the random sample $\boldsymbol{y} = \{y_1, y_2, \ldots, y_m\}$ not yet observed? Consider the experiment $e(n+m)$ and the sampling density

$$p(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})$$

Since both experiments are independent and iid, we have the joint density

$$p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}) = p(\boldsymbol{x}|\boldsymbol{\theta})\; p(\boldsymbol{y}|\boldsymbol{\theta}) \qquad \longrightarrow \qquad p(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x}|\boldsymbol{\theta})\; p(\boldsymbol{y}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})$$

and integrating the parameter $\boldsymbol{\theta} \in \Theta$:

$$p(\boldsymbol{y}, \boldsymbol{x}) = p(\boldsymbol{y}|\boldsymbol{x})\; p(\boldsymbol{x}) = \int_{\Theta} p(\boldsymbol{y}|\boldsymbol{\theta})\; p(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} = p(\boldsymbol{x}) \int_{\Theta} p(\boldsymbol{y}|\boldsymbol{\theta})\; p(\boldsymbol{\theta}|\boldsymbol{x})\, d\boldsymbol{\theta}$$

Thus, we have that

$$p(\boldsymbol{y}|\boldsymbol{x}) = \int_{\Theta} p(\boldsymbol{y}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\boldsymbol{x})\, d\boldsymbol{\theta}$$

This is the basic expression for the *predictive inference* and allows us to predict the results $\boldsymbol{y}$ of a future experiment from the results $\boldsymbol{x}$ observed in a previous experiment within the same parametric model. Note that $p(\boldsymbol{y}|\boldsymbol{x})$ is the density of the quantities not yet observed conditioned to the observed sample. Thus, even though the experiments $e(\boldsymbol{y})$ and $e(\boldsymbol{x})$ are statistically independent, the realization of the first one ($e(\boldsymbol{x})$) modifies the knowledge we have on the parameters $\boldsymbol{\theta}$ of the model and therefore affect the prediction on future experiments for, if we do not consider the results of the first experiment or just don't do it, the predictive distribution for $e(\boldsymbol{y})$ would be

$$p(\boldsymbol{y}) = \int_{\Theta} p(\boldsymbol{y}|\boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

It is then clear from the expression of *predictive inference* that in practice it is equivalent to consider as prior density for the second experiment the proper density $\pi(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{x})$. If the first experiment provides very little information on the parameters, then $p(\boldsymbol{\theta}|\boldsymbol{x}) \simeq \pi(\boldsymbol{\theta})$ and

$$p(\boldsymbol{y}|\boldsymbol{x}) \simeq \int_{\Theta} p(\boldsymbol{y}|\boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} \simeq p(\boldsymbol{y})$$

On the other hand, if after the first experiment we know the parameters with high accuracy then, in distributional sense, $\langle p(\boldsymbol{\theta}|\boldsymbol{x}), \cdot\rangle \simeq \langle \delta(\boldsymbol{\theta}_0), \cdot\rangle$ and

$$p(\boldsymbol{y}|\boldsymbol{x}) \simeq \langle \delta(\boldsymbol{\theta}_0),\, p(\boldsymbol{y}|\boldsymbol{\theta})\rangle = p(\boldsymbol{y}|\theta_0).$$

2.4 Sufficient Statistics

Consider m random quantities $\{X_1, X_2, \ldots, X_m\}$ that take values in $\Omega_1 \times \cdots \times \Omega_m$ and a random vector

$$\boldsymbol{T} : \Omega_1 \times \cdots \times \Omega_m \longrightarrow \mathcal{R}^{k(m)}$$

whose $k(m) \leq m$ components are functions of the random quantities $\{X_i\}_{i=1}^m$. Given the sample $\{x_1, x_2, \ldots, x_m\}$, the vector $\boldsymbol{t} = \boldsymbol{t}(x_1, \ldots, x_m)$ is a *$k(m)$-dimensional statistic*. The practical interest lies in the existence of statistics that contain all the relevant information about the parameters so we don't have to work with the whole sample and simplify considerably the expressions. Thus, of special relevance are the *sufficient statistics*. Given the model $p(x_1, x_2, \ldots, x_n|\boldsymbol{\theta})$, the set of statistics

$\boldsymbol{t} = \boldsymbol{t}(x_1, \ldots, x_m)$ is *sufficient* for $\boldsymbol{\theta}$ if, and only if, $\forall m \geq 1$ and any prior distribution $\pi(\boldsymbol{\theta})$ it holds that

$$p(\boldsymbol{\theta}|x_1, x_2, \ldots, x_m) = p(\boldsymbol{\theta}|\boldsymbol{t})$$

Since the data act in the Bayes formula only through the likelihood, it is clear that to specify the posterior density of $\boldsymbol{\theta}$ we can consider

$$p(\boldsymbol{\theta}|x_1, x_2, \ldots, x_m) = p(\boldsymbol{\theta}|\boldsymbol{t}) \propto p(\boldsymbol{t}|\boldsymbol{\theta})\, \pi(\boldsymbol{\theta})$$

and all other aspects of the data but $\boldsymbol{t}$ are irrelevant. It is obvious however that $\boldsymbol{t} = \{x_1, \ldots, x_m\}$ is sufficient and, in principle, gives no simplification in the modeling. For this we should have $k(m) = \dim(\boldsymbol{t}) < m$ (*minimal sufficient statistics*) and, in the ideal case, we would like that $k(m) = k$ does not depend on m. Except some irregular cases, the only distributions that admit a fixed number of sufficient statistics independently of the sample size (that is, $k(m) = k < m\ \forall m$) are those that belong to the exponential family.

Example 2.1 (1) Consider the exponential model $X \sim Ex(x|\theta)$: and the iid experiment $e(m)$ that provides the sample $\boldsymbol{x} = \{x_1, \ldots, x_m\}$. The likelihood function is:

$$p(\boldsymbol{x}|\theta) = \theta^m\, e^{-\theta\,(x_1 + \cdots + x_m)} = \theta^{t_1}\, e^{-\theta\, t_2}$$

and therefore we have the sufficient statistic $\boldsymbol{t} = (m,\ \sum_{i=1}^m x_i) : \Omega_1 \times \cdots \times \Omega_m \longrightarrow \mathcal{R}^{k(m)=2}$

(2) Consider the Normal model $X \sim N(x|\mu, \sigma)$ and the iid experiment $e(m)$ again with $\boldsymbol{x} = \{x_1, \ldots, x_m\}$. The likelihood function is:

$$p(\boldsymbol{x}|\mu, \sigma)\ \propto\ \sigma^{-m} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^m (x_i - \mu)^2\right\} = \sigma^{-t_1} \exp\left\{-\frac{1}{2\sigma^2}(t_3 - 2\,\mu\, t_2 + \mu^2\, t_1)\right\}$$

and $\boldsymbol{t} = (m,\ \sum_{i=1}^m x_i,\ \sum_{i=1}^m x_i^2) : \Omega_1 \times \cdots \times \Omega_m \longrightarrow \mathcal{R}^{k(m)=3}$ a sufficient statistic. Usually we shall consider $\boldsymbol{t} = \{m,\ \overline{x}, s^2\}$ with

$$\overline{x} = \frac{1}{m}\sum_{i=1}^m x_i \qquad \text{and} \qquad s^2 = \frac{1}{m}\sum_{i=1}^m (x_i - \overline{x})^2$$

the sample mean and the sample variance. Inferences on the parameters μ and σ will depend on $\boldsymbol{t}$ and all other aspects of the data are irrelevant.

(3) Consider the Uniform model $X \sim Un(x|0, \theta)$ and the iid sampling $\{x_1, x_2, \ldots, x_m\}$. Then $\boldsymbol{t} = (m,\ \max\{x_i,\ i = 1, \ldots m\}) : \Omega_1 \times \cdots \times \Omega_m \longrightarrow \mathcal{R}^{k(m)=2}$ is a sufficient statistic for θ.

2.5 Exponential Family

A probability density $p(x|\boldsymbol{\theta})$, with $x \in \Omega_X$ and $\boldsymbol{\theta} \in \Theta \subseteq \mathcal{R}^k$ belongs to the *k-parameter exponential family* if it has the form:

$$p(x|\boldsymbol{\theta}) = f(x)\, g(\boldsymbol{\theta}) \exp\left\{\sum_{i=1}^{k} c_i\, \phi_i(\boldsymbol{\theta})\, h_i(x)\right\}$$

with

$$g(\boldsymbol{\theta})^{-1} = \int_{\Omega_x} f(x) \prod_{i=1}^{k} \exp\{c_i\, \phi_i(\boldsymbol{\theta})\, h_i(x)\}\, dx \;\leq\; \infty$$

The family is called *regular* if supp$\{X\}$ is independent of $\boldsymbol{\theta}$; *irregular* otherwise.

If $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$ is an exchangeable random sampling from the k-parameter regular exponential family, then

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \left[\prod\nolimits_{i=1}^{n} f(x_i)\right] [g(\boldsymbol{\theta})]^n \exp\left\{\sum_{i=1}^{k} c_i\, \phi_i(\boldsymbol{\theta}) \left(\sum_{j=1}^{n} h_i(x_j)\right)\right\}$$

and therefore $\boldsymbol{t}(\boldsymbol{x}) = \left\{n, \sum_{i=1}^{n} h_1(x_i), \ldots \sum_{i=1}^{n} h_k(x_i)\right\}$ will be a set of *sufficient statistics*.

Example 2.2 Several distributions of interest, like Poisson and Binomial, belong to the exponential family:

(1) Poisson $Po(n|\mu)$: $P(n|\mu) = \dfrac{e^{-\mu}\,\mu^n}{\Gamma(n+1)} = \dfrac{e^{-(\mu - n\ln\mu)}}{\Gamma(n+1)}$

(2) Binomial $Bi(n|N,\theta)$: $P(n|N,\theta) = \binom{N}{n} \theta^n\, (1-\theta)^{N-n} = \binom{N}{n} e^{n\ln\theta + (N-n)\ln(1-\theta)}$

However, the Cauchy $Ca(x|\alpha,\beta)$ distribution, for instance, does not because

$$p(x_1, \ldots, x_m|\alpha,\beta) \;\propto\; \prod_{i=1}^{n} \left(1 + \beta(x_i - \alpha)^2\right)^{-1} = \exp\left\{\sum_{i=1}^{m} \log(1 + \beta(x_i - \alpha)^2)\right\}$$

can not be expressed as the exponential family form. In consequence, there are no sufficient *minimal* statistics (in other words $\boldsymbol{t} = \{n, x_1, \ldots, x_n\}$ is the sufficient statistic) and we will have to work with the whole sample.

2.6 Prior Functions

In the *Bayes rule*, $p(\boldsymbol{\theta}|\boldsymbol{x}) \propto p(\boldsymbol{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})$, the *prior function* $p(\boldsymbol{\theta})$ represents the knowledge (*degree of credibility*) that we have about the parameters before the experiment is done and it is a necessary element to obtain the *posterior density* $p(\boldsymbol{\theta}|\boldsymbol{x})$ from which we shall make inferences. If we have faithful information on them before we do the experiment, it is reasonable to incorporate that in the specification of the prior density (*informative prior*) so the new data will provide additional information that will update and improve our knowledge. The specific form of the prior can be motivated, for instance, by the results obtained in previous experiments. However, it is usual that before we do the experiment, either we have a vague knowledge of the parameters compared to what we expect to get from the experiment or simply we do not want to include previous results to perform an independent analysis. In this case, all the new information will be contained in the likelihood function $p(\boldsymbol{x}|\boldsymbol{\theta})$ of the experiment and the prior density (*non-informative prior*) will be merely a mathematical element needed for the inferential process. Being this the case, we expect that the whole weight of the inferences rests on the likelihood and the prior function has the smallest possible influence on them. To learn something from the experiment it is then desirable to have a situation like the one shown in Fig. 2.1 where the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x})$ is dominated by the likelihood function. Otherwise, the experiment will provide little information compared to the one we had before and, unless our previous knowledge is based on suspicious observations, it will be wise to design a better experiment.

A considerable amount of effort has been put to obtain reasonable *non-informative priors* that can be used as a standard reference function for the Bayes rule. Clearly, *non-informative* is somewhat misleading because we are never in a state of absolute ignorance about the parameters and the specification of a mathematical model for the process assumes some knowledge about them (masses and life-times take non-negative real values, probabilities have support on [0, 1],…). On the other hand, it doesn't make sense to think about a function that represents ignorance in a formal and objective way so *knowing little a priory* is relative to what we may expect to learn from the experiment. Whatever prior we use will certainly have some effect on the posterior inferences and, in some cases, it would be wise to consider a reasonable set of them to see what is the effect.

The ultimate task of this section is to present the most usual approaches to derive a non-informative prior function to be used as a standard reference that contains little information about the parameters compared to what we expect to get from the experiment.[4] In many cases, these priors will not be Lebesgue integrable (*improper functions*) and, obviously, can not be considered as probability density functions that quantify any knowledge on the parameters (although, with little rigor, sometimes we still talk about prior *densities*). If one is reluctant to use them right the way one can, for instance, define them on a sufficiently large compact support that contains the region where the likelihood is dominant. However, since

[4] For a comprehensive discussion see [5].

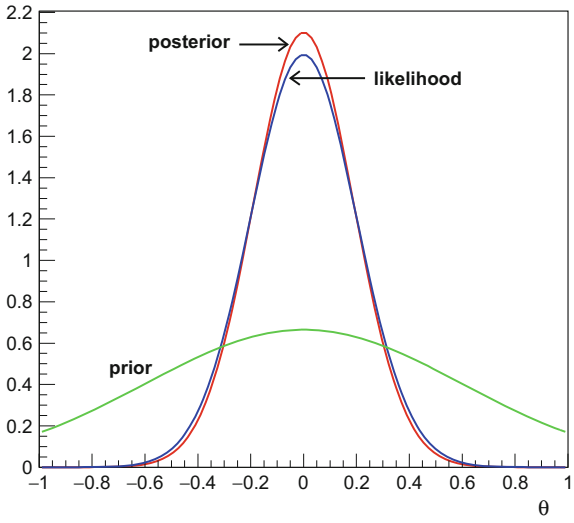

Fig. 2.1 Prior, likelihood and posterior as function of the parameter θ. In this case, the prior is a smooth function and the posterior is dominated by the likelihood

$$p(\boldsymbol{\theta}|\boldsymbol{x})\, d\boldsymbol{\theta} \;\propto\; p(\boldsymbol{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta} = p(\boldsymbol{x}|\boldsymbol{\theta})\, d\mu(\boldsymbol{\theta})$$

in most cases it will be sufficient to consider them simply as what they really are: a measure. In any case, what is mandatory is that the posterior is a well defined proper density.

2.6.1 *Principle of Insufficient Reason*

The *Principle of Insufficient Reason*[5] dates back to J. Bernoulli and P.S. Laplace and, originally, it states that if we have n exclusive and exhaustive hypothesis and there is no special reason to prefer one over the other, it is reasonable to consider them equally likely and assign a prior probability $1/n$ to each of them. This certainly sounds reasonable and the idea was right the way extended to parameters taking countable possible values and to those with continuous support that, in case of compact sets, becomes a uniform density. It was extensively used by P.S. Laplace and T. Bayes, being he the first to use a uniform prior density for making inferences on the parameter of a Binomial distribution, and is usually referred to as the *"Bayes-Laplace Postulate"*. However, a uniform prior density is obviously not invariant under reparameterizations. If prior to the experiment we have a very vague knowledge about the parameter $\theta \in [a, b]$, we certainly have a vague knowledge about $\phi = 1/\theta$ or $\zeta = \log\theta$ and a uniform distribution for θ:

[5] Apparently, *"Insufficient Reason"* was coined by Laplace in reference to the Leibniz's *Principle of Sufficient Reason* stating essentially that every fact has a sufficient reason for why it is the way it is and not other way.

$$\pi(\theta)\, d\theta = \frac{1}{b-a}\, d\theta$$

implies that:

$$\pi(\phi)\, d\phi = \frac{1}{\phi^2}\, d\phi \qquad \text{and} \qquad \pi(\zeta)\, d\zeta = e^{\zeta}\, d\zeta$$

Shouldn't we take as well a uniform density for ϕ or ζ?

Nevertheless, we shall see that a uniform density, that is far from representing ignorance on a parameter, may be a reasonable choice in many cases even though, if the support of the parameter is infinite, it is an improper function.

2.6.2 Parameters of Position and Scale

An important class of parameters we are interested in are those of position and scale. Let's treat them separately and leave for a forthcoming section the argument behind that. Start with a random quantity $X \sim p(x|\mu)$ with μ a *location parameter*. The density has the form $p(x|\mu) = f(x-\mu)$ so, taking a prior function $\pi(\mu)$ we can write

$$p(x, \mu)\, dx\, d\mu = [p(x|\mu)\, dx]\, [\pi(\mu)\, d\mu] = [f(x-\mu)\, dx]\, [\pi(\mu)\, d\mu]$$

Now, consider random quantity $X' = X + a$ with $a \in \mathcal{R}$ a known value. Defining the new parameter $\mu' = \mu + a$ we have

$$p(x', \mu')\, dx'\, d\mu' = \left[p(x'|\mu')\, dx'\right] \left[\pi'(\mu')\, d\mu'\right] = \left[f(x'-\mu')\, dx'\right] \left[\pi(\mu'-a)\, d\mu'\right]$$

In both cases the models have the same structure so making inferences on μ from the sample $\{x_1, x_1, \ldots, x_n\}$ is formally equivalent to making inferences on μ' from the shifted sample $\{x'_1, x'_2, \ldots, x'_n\}$. Since we have the same prior degree of knowledge on μ and μ', it is reasonable to take the same functional form for $\pi(\cdot)$ and $\pi'(\cdot)$ so:

$$\pi(\mu'-a)\, d\mu' == \pi(\mu')\, d\mu' \qquad \forall a \in \mathcal{R}$$

and, in consequence:

$$\pi(\mu) = \text{constant}$$

If θ is a *scale parameter*, the model has the form $p(x|\theta) = \theta f(x\theta)$ so taking a prior function $\pi(\theta)$ we have that

$$p(x, \theta)\, dx\, d\theta = [p(x|\theta)\, dx]\, [\pi(\theta)\, d\theta] = [\theta f(x\theta)\, dx]\, [\pi(\theta)\, d\theta]$$

For the scaled random quantity $X' = a\,X$ with $a \in \mathcal{R}^+$ known, we have that:

$$p(x', \theta')\,dx'\,d\theta' = \left[p(x'|\theta')\,dx'\right]\left[\pi'(\theta')\,d\theta'\right] = \left[\theta' f(x'\theta')\,dx'\right]\left[\pi(a\theta')\,a d\theta\right]$$

where we have defined the new parameter $\theta' = \theta/a$. Following the same argument as before, it is sound to assume the same functional form for $\pi(\cdot)$ and $\pi'(\cdot)$ so:

$$\pi(a\theta')\,a\,d\theta' = \pi(\theta')\,d\theta' \qquad \forall a \in \mathcal{R}$$

and, in consequence:

$$\pi(\theta) = \frac{1}{\theta}$$

Both prior functions are improper so they may be explicited as

$$\pi(\mu, \theta) \;\propto\; \frac{1}{\theta}\,\mathbf{1}_{\Theta}(\theta)\,\mathbf{1}_{M}(\mu)$$

with Θ, M an appropriate sequence of compact sets or considered as prior measures provided that the posterior densities are well defined. Let's see some examples.

Example 2.3 (*The Exponential Distribution*) Consider the sequence of independent observations $\{x_1, x_2, \ldots, x_n\}$ of the random quantity $X \sim Ex(x|\theta)$ drawn under the same conditions. The joint density is

$$p(x_1, x_2, \ldots, x_n|\theta) = \theta^n\, e^{-\theta(x_1 + x_2 + \cdots x_n)}$$

The statistic $t = n^{-1}\sum_{i=1}^{n} x_i$ is sufficient for θ and is distributed as

$$p(t|\theta) = \frac{(n\theta)^n}{\Gamma(n)}\,t^{n-1}\exp\{-n\theta t\}$$

It is clear that θ is a scale parameter so we shall take the prior function $\pi(\theta) = 1/\theta$. Note that if we make the change $z = \log t$ and $\phi = \log\theta$ we have that

$$p(z|\phi) = \frac{n^n}{\Gamma(n)}\exp\{n\left((\phi + z) - e^{\phi+z}\right)\}$$

In this parameterization, ϕ is a position parameter and therefore $\pi(\phi) = \text{const}$ in consistency with $\pi(\theta)$. Then, we have the proper posterior for inferences:

$$p(\theta|t, n) = \frac{(nt)^n}{\Gamma(n)}\exp\{-nt\theta\}\,\theta^{n-1}; \qquad \theta > 0$$

Consider now the sequence of compact sets $C_k = [1/k, k]$ covering R^+ as $k \to \infty$. Then, with support on C_k we have the proper prior density

$$\pi_k(\theta) = \frac{1}{2 \log k} \frac{1}{\theta} \mathbf{1}_{C_k}(\theta)$$

and the sequence of posteriors:

$$p_k(\theta|t, n) = \frac{(nt)^n}{\gamma(n, ntk) - \gamma(n, nt/k)} \exp\{-nt\theta\} \theta^{n-1} \mathbf{1}_{C_k}(\theta)$$

with $\gamma(a, x)$ the Incomplete Gamma Function. It is clear that

$$\lim_{k \to \infty} p_k(\theta|t, n) = p(\theta|t, n)$$

Example 2.4 (*The Uniform Distribution*) Consider the random quantity $X \sim Un(x|0, \theta)$ and the independent sampling $\{x_1, x_2, \ldots, x_n\}$. To draw inferences on θ, the statistics $x_M = \max\{x_1, x_2, \ldots, x_n\}$ is sufficient and is distributed as (show that):

$$p(x_M|\theta) = n \frac{x_M^{n-1}}{\theta^n} \mathbf{1}_{[0,\theta]}(x_M)$$

As in the previous case, θ is a scale parameter and with the change $t_M = \log x_M$, $\phi = \log \theta$ is a position parameter. Then, we shall take $\pi(\theta) \propto \theta^{-1}$ and get the posterior density (Pareto):

$$p(\theta|x_M, n) = n \frac{x_M^n}{\theta^{n+1}} \mathbf{1}_{[x_M, \infty)}(\theta)$$

Example 2.5 (*The one-dimensional Normal Distribution*) Consider the random quantity $X \sim N(x|\mu, \sigma)$ and the experiment $e(n)$ that provides the independent and exchangeable sequence $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$ of observations. The likelihood function will then be:

$$p(\boldsymbol{x}|\mu, \sigma) = \prod_{i=1}^{n} p(x_i|\mu, \sigma) \propto \frac{1}{\sigma^n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\}$$

There is a three-dimensional sufficient statistic $\boldsymbol{t} = \{n, \overline{x}, s^2\}$ where

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad \text{and} \qquad s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

so we can write

$$p(\boldsymbol{x}|\mu,\sigma) \propto \frac{1}{\sigma^n} \exp\left\{-\frac{n}{2\sigma^2}\left(s^2 + (\overline{x}-\mu)^2\right)\right\}$$

In this case we have both position and scale parameters so we take $\pi(\mu,\sigma) = \pi(\mu)\pi(\sigma) = \sigma^{-1}$ and get the proper posterior

$$p(\mu,\sigma|\boldsymbol{x}) \propto p(\boldsymbol{x}|\mu,\sigma)\,\pi(\mu,\sigma) \propto \frac{1}{\sigma^{n+1}} \exp\left\{-\frac{n}{2\sigma^2}\left[s^2 + (\overline{x}-\mu)^2\right]\right\}$$

• **Marginal posterior density of** σ: Integrating the parameter $\mu \in \mathcal{R}$ we have that:

$$p(\sigma|\boldsymbol{x}) = \int_{-\infty}^{+\infty} p(\mu,\sigma|\boldsymbol{x})\,d\mu \propto \sigma^{-n} \exp\left\{-\frac{n\,s^2}{2\sigma^2}\right\} \mathbf{1}_{(0,\infty)}(\sigma)$$

and therefore, the random quantity

$$Z = \frac{n\,s^2}{\sigma^2} \sim \chi^2(z|n-1)$$

• **Marginal posterior density of** μ: Integrating the parameter $\sigma \in [0,\infty)$ we have that:

$$p(\mu|\boldsymbol{x}) = \int_0^{+\infty} p(\mu,\sigma|\boldsymbol{x})\,d\sigma \propto \left(1 + \frac{(\mu-\overline{x})^2}{s^2}\right)^{-n/2} \mathbf{1}_{(-\infty,\infty)}(\mu)$$

so the random quantity

$$T = \frac{\sqrt{n-1}(\mu-\overline{x})}{s} \sim St(t|n-1)$$

It is clear that $p(\mu,\sigma|\boldsymbol{x}) \neq p(\mu|\boldsymbol{x})\,p(\sigma|\boldsymbol{x})$ and, in consequence, are not independent.

• **Distribution of** μ **conditioned to** σ: Since $p(\mu,\sigma|\boldsymbol{x}) = p(\mu|\sigma,\boldsymbol{x})\,p(\sigma|\boldsymbol{x})$ we have that

$$p(\mu|\sigma,\boldsymbol{x}) \propto \frac{1}{\sigma} \exp\{-\frac{n}{2\sigma^2}(\mu-\overline{x})^2\}$$

so $\mu|\sigma \sim N(\mu|\overline{x},\sigma/\sqrt{n})$.

Example 2.6 (*Contrast of parameters of Normal Densities*) Consider two independent random quantities $X_1 \sim N(x_1,|\mu_1,\sigma_1)$ and $X_2 \sim N(x_2,|\mu_2,\sigma_2)$ and the ran-

dom samplings $\boldsymbol{x}_1 = \{x_{11}, x_{12}, \ldots, x_{1n_1}\}$ and $\boldsymbol{x}_2 = \{x_{21}, x_{22}, \ldots, x_{2n_2}\}$ of sizes n_1 and n_2 under the usual conditions. From the considerations of the previous example, we can write

$$p(\boldsymbol{x}_i|\mu_i, \sigma_i) \propto \frac{1}{\sigma_i^{n_i}} \exp\left\{-\frac{n_i}{2\sigma_i^2}\ \left(s_i^2 + (\overline{x_i} - \mu_i)^2\right)\right\}; \qquad i = 1, 2$$

Clearly, (μ_1, μ_2) are position parameters and (σ_1, σ_2) scale parameters so, in principle, we shall take the improper prior function

$$\pi(\mu_1, \sigma_1, \mu_2, \sigma_2) = \pi(\mu_1)\pi(\mu_2)\pi(\sigma_1)\pi(\sigma_2) \propto \frac{1}{\sigma_1\, \sigma_2}$$

However, if we have know that both distributions have the same variance, then we may set $\sigma = \sigma_1 = \sigma_2$ and, in this case, the prior function will be

$$\pi(\mu_1, \mu_2, \sigma) = \pi(\mu_1)\pi(\mu_2)\pi(\sigma) \propto \frac{1}{\sigma}$$

Let's analyze both cases.

• **Marginal Distribution of σ_1 and σ_2**: In this case we assume that $\sigma_1 \neq \sigma_2$ and we shall take the prior $\pi(\mu_1, \sigma_1, \mu_2, \sigma_2) \propto (\sigma_1\sigma_2)^{-1}$. Integrating μ_1 and μ_2 we get:

$$p(\sigma_1, \sigma_2|\boldsymbol{x}_1, \boldsymbol{x}_2) = p(\sigma_1, |\boldsymbol{x}_1)p(\sigma_2, |\boldsymbol{x}_2) \propto \sigma_1^{-n_1}\, \sigma_2^{-n_2} \exp\left\{-\frac{1}{2}\left(\frac{n_1 s_1^2}{\sigma_1^2} + \frac{n_2 s_2^2}{\sigma_2^2}\right)\right\}$$

Now, if we define the new random quantities

$$Z = \frac{s_2^2}{w^2\, s_1^2} = \frac{(\sigma_1/s_1)^2}{(\sigma_2/s_2)^2} \qquad \text{and} \qquad W = \frac{n_1 s_1^2}{{\sigma_1}^2}$$

both with support in $(0, +\infty)$, and integrate the last we get we get that Z follows a Snedecor Distribution $Sn(z|n_2 - 1, n_1 - 1)$ whose density is

$$p(z|\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{(\nu_1/\nu_2)^{\nu_1/2}}{\text{Be}\,(\nu_1/2, \nu_2/2)}\, z^{(\nu_1/2)-1} \left(1 + \frac{\nu_1}{\nu_2}\, z\right)^{-(\nu_1+\nu_2)/2} \mathbf{1}_{(0,\infty)}(z).$$

• **Marginal Distribution of μ_1 and μ_2**: In this case, it is different whether we assume that, although unknown, the variances are the same or not. In the first case, we set $\sigma_1 = \sigma_2 = \sigma$ and take the reference prior $\pi(\mu_1, \mu_2, \sigma) = \sigma^{-1}$. Defining

$$A = n_1\left[s_1^2 + (\overline{x}_1 - \mu_1)^2\right] + n_2\left[s_2^2 + (\overline{x}_2 - \mu_2)^2\right]$$

we can write

$$p(\mu_1, \mu_2, \sigma|\boldsymbol{x}, \boldsymbol{y}) \propto \frac{1}{\sigma^{n_1+n_2+1}} \exp\left\{-\frac{1}{2} A/\sigma^2\right\}$$

It is left as an exercise to show that if we make the transformation

$$w = \mu_1 - \mu_2 \in (-\infty, +\infty); \quad u = \mu_2 \in (-\infty, +\infty) \quad \text{and} \quad z = \sigma^{-2} \in (0, +\infty)$$

and integrate the last two, we get

$$p(w|\boldsymbol{x}_1, \boldsymbol{x}_2) \propto \left(1 + \frac{n_1 n_2}{n_1 + n_2} \frac{[(\overline{x}_1 - \overline{x}_2) - w]^2}{n_1 s_1^2 + n_2 s_2^2}\right)^{-(n_1+n_2-1)/2}$$

Introducing the more usual terminology

$$s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

we have that

$$p(w|\boldsymbol{x}_1, \boldsymbol{x}_2) \propto \left(1 + \frac{n_1 n_2}{n_1 + n_2} \frac{[w - (\overline{x}_1 - \overline{x}_2)]^2}{s^2 (n_1 + n_2 - 2)}\right)^{-[(n_1+n_2-2)+1]/2}$$

and therefore the random quantity

$$T = \frac{(\mu_1 - \mu_2) - (\overline{x}_1 - \overline{x}_2)}{s (1/n_1 + 1/n_2)^{1/2}}$$

follows a Student's Distribution $St(t|\nu)$ with $\nu = n_1 + n_2 - 2$ degrees of freedom.

Let's see now the case where we can not assume that the variances are equal. Taking the prior reference function $\pi(\mu_1, \mu_2, \sigma_1, \sigma_2) = (\sigma_1 \sigma_2)^{-1}$ we get

$$p(\mu_1, \mu_2, \sigma_1, \sigma_2|\boldsymbol{x}_1, \boldsymbol{x}_2) \propto \sigma_1^{-(n_1+1)} \sigma_2^{-(n_2+1)} \exp\left\{-\frac{1}{2} \sum_{i=1}^{2} \frac{s_i^2 + (\overline{x}_i - \mu_i)^2}{\sigma_i^2/n_i}\right\}$$

After the appropriate integrations (left as exercise), defining $w = \mu_1 - \mu_2$ and $u = \mu_2$ we end up with the density

$$p(w, u|\boldsymbol{x}_1, \boldsymbol{x}_2) \propto \left(1 + \frac{(\overline{x}_1 - w - u)^2}{s_1^2}\right)^{-n_1/2} \left(1 + \frac{(\overline{x}_2 - u)^2}{s_2^2}\right)^{-n_2/2}$$

where integral over $u \in \mathcal{R}$ can not be expressed in a simple way. The density

$$p(w|\boldsymbol{x}_1, \boldsymbol{x}_2) \propto \int_{-\infty}^{+\infty} p(w, u|\boldsymbol{x}_1, \boldsymbol{x}_2)\, du$$

is called the *Behrens-Fisher Distribution*. Thus, to make statements on the difference of Normal means, we should analyze first the sample variances and decide how shall we treat them.

2.6.3 Covariance Under Reparameterizations

The question of how to establish a reasonable criteria to obtain a prior for a given model $p(\boldsymbol{x}|\boldsymbol{\theta})$ that can be used as a standard reference function was studied by Harold Jeffreys [6] in the mid XX century. The rationale behind the argument is that if we have the model $p(\boldsymbol{x}|\boldsymbol{\theta})$ with $\boldsymbol{\theta} \in \Omega_\theta \subseteq R^n$ and make a reparameterizations $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta})$ with $\boldsymbol{\phi}(\cdot)$ a one-to-one differentiable function, the statements we make about $\boldsymbol{\theta}$ should be consistent with those we make about $\boldsymbol{\phi}$ and, in consequence, priors should be related by

$$\pi_\theta(\boldsymbol{\theta}) d\boldsymbol{\theta} = \pi_\phi(\boldsymbol{\phi}(\boldsymbol{\theta})) \left| \det \left[\frac{\partial \phi_i(\theta)}{\partial \theta_j} \right] \right| d\boldsymbol{\theta}$$

Now, assume that the Fisher's matrix (see Sect. 4.5)

$$\boldsymbol{I}_{ij}(\boldsymbol{\theta}) = E_{\boldsymbol{X}} \left[\frac{\partial \log\, p(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \log\, p(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \theta_j} \right]$$

exists for this model. Under a differentiable one-to-one transformation $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta})$ we have that

$$\boldsymbol{I}_{ij}(\boldsymbol{\phi}) = \frac{\partial \theta_k}{\partial \phi_i} \frac{\partial \theta_l}{\partial \phi_j} \boldsymbol{I}_{kl}(\boldsymbol{\theta})$$

so it behaves as a covariant symmetric tensor of second order (left as exercise). Then, since

$$\det [\boldsymbol{I}(\boldsymbol{\phi})] = \left| \det \left[\frac{\partial \theta_i}{\partial \phi_j} \right] \right|^2 \det [\boldsymbol{I}(\boldsymbol{\theta})]$$

Jeffreys proposed to consider the prior

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

In fact, if we consider the parameter space as a Riemannian manifold (see Sect. 4.7) the Fisher's matrix is the metric tensor (Fisher-Rao metric) and this is just the invariant

volume element. Intuitively, if we make a transformation such that at a particular value $\boldsymbol{\phi}_0 = \boldsymbol{\phi}(\boldsymbol{\theta}_0)$ the Fisher's tensor is constant and diagonal, the metric in a neighborhood of $\boldsymbol{\phi}_0$ is Euclidean and we have location parameters for which a constant prior is appropriate and therefore

$$\pi(\boldsymbol{\phi}) d\boldsymbol{\phi} \propto d\boldsymbol{\phi} = \left[\det\left[\boldsymbol{I}(\boldsymbol{\theta})\right]^{1/2}\right] d\boldsymbol{\theta} = \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

It should be pointed out that there may be other priors that are also invariant under reparameterizations and that, as usual, we talk loosely about *prior densities* although they usually are improper functions.

For one-dimensional parameter, the density function expressed in terms of

$$\phi \sim \int \left[\boldsymbol{I}(\theta)\right]^{1/2} d\theta$$

may be reasonably well approximated by a Normal density (at least in the parametric region where the likelihood is dominant) because $\boldsymbol{I}(\phi)$ is constant (see Sect. 4.5) and then, due to translation invariance, a constant prior for ϕ is justified. Let' see some examples.

Example 2.7 (*The Binomial Distribution*) Consider the random quantity $X \sim Bi(x|\theta, n)$:

$$p(x|n, \theta) = \binom{n}{x} \theta^k (1-\theta)^{n-k}; \qquad n, k \in N_0;\ k \leq n$$

with $0 < \theta < 1$. Since $E[X] = n\theta$ we have that:

$$\boldsymbol{I}(\theta) = E_X\left[\left(-\frac{\partial^2 \log p(x|n, \theta)}{\partial \theta^2}\right)\right] = \frac{n}{\theta\,(1-\theta)}$$

so the Jeffreys prior (proper in this case) for the parameter θ is

$$\pi(\theta) \propto \left[\theta\,(1-\theta)\right]^{-1/2}$$

and the posterior density will therefore be

$$p(\theta|k, n) \propto \theta^{k-1/2}\,(1-\theta)^{n-k-1/2}$$

that is; a $Be(x|k+1/2, n-k+1/2)$ distribution. Since

$$\phi = \int \frac{d\theta}{\sqrt{\theta\,(1-\theta)}} = 2\,\mathrm{asin}\,(\theta^{1/2})$$

we have that $\theta = \sin^2\phi/2$ and, parameterized in terms of ϕ, $\boldsymbol{I}(\phi)$ is constant so the distribution *"looks"* more Normal (see Fig. 2.2).

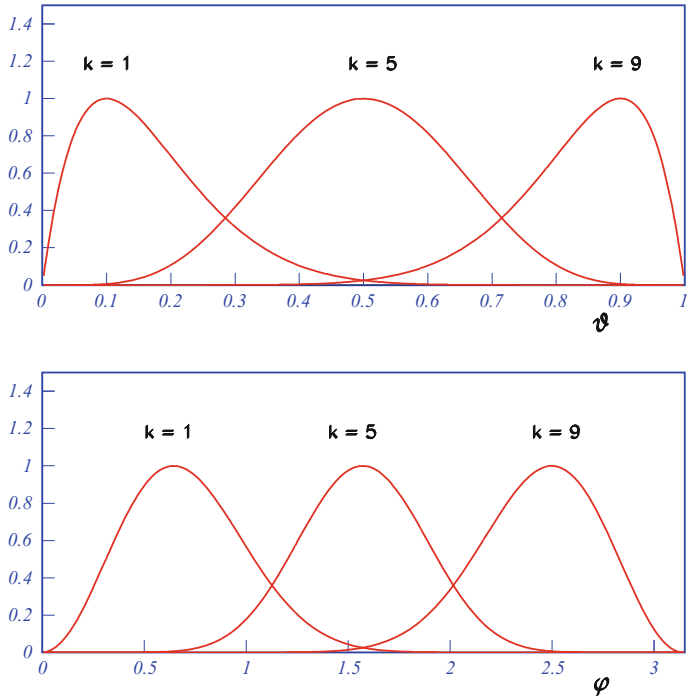

Fig. 2.2 Dependence of the likelihood function with the parameter θ (*upper*) and with $\phi = 2\,\text{asin}(\theta^{1/2})$ (*lower*) for a Binomial process with $n = 10$ and $k = 1, 5$ and 9

Example 2.8 (*The Poisson Distribution*) Consider the random quantity $X \sim Po(x|\mu)$:

$$p(x|\mu) = e^{-\mu}\,\frac{\mu^x}{\Gamma(x+1)}; \qquad x \in N;\ \mu \in R^+$$

Then, since $E[X] = \mu$ we have

$$\boldsymbol{I}(\mu) = E_X\left[\left(-\frac{\partial^2 \log p(x|\mu)}{\partial \mu^2}\right)\right] = \frac{1}{\mu}$$

so we shall take as *prior* (improper):

$$\pi(\mu) = [\boldsymbol{I}(\mu)]^{1/2} = \mu^{-1/2}$$

and make inferences on μ from the proper posterior density

$$p(\mu|x) \propto e^{-\mu}\,\mu^{x-1/2}$$

that is, a $Ga(x|1, x + 1/2)$ distribution.

Example 2.9 (*The Pareto Distribution*) Consider the random quantity $X \sim Pa(x|\theta, x_0)$ with $x_0 \in R^+$ known and density

$$p(x|\theta, x_0) = \frac{\theta}{x_0} \left(\frac{x_0}{x}\right)^{\theta+1} \mathbf{1}_{(x_0,\infty)}(x); \qquad \theta \in R^+$$

Then,

$$\boldsymbol{I}(\theta) = E_X \left[\left(- \frac{\partial^2 \log p(x|\theta, x_0)}{\partial \theta^2} \right) \right] = \frac{1}{\theta^2}$$

so we shall take as *prior* (improper):

$$\pi(\theta) \propto [\boldsymbol{I}(\mu)]^{1/2} = \theta^{-1}$$

and make inferences from the posterior density (proper)

$$p(\theta|x, x_0) = x^{-\theta} \log x$$

Note that if we make the transformation $t = \log x$, the density becomes

$$p(t|\theta, x_0) = \theta\, x_0^{\theta}\, e^{-\theta t} \mathbf{1}_{(\log x_0,\infty)}(t)$$

for which θ is a scale parameter and, from previous considerations, we should take $\pi(\theta) \propto \theta^{-1}$ in consistency with Jeffreys's prior.

Example 2.10 (*The Gamma Distribution*) Consider the random quantity $X \sim Ga(x|\alpha, \beta)$ with $\alpha, \beta \in R^+$ and density

$$p(x|\alpha, \beta) = \frac{\alpha^{\beta}}{\Gamma(\beta)}\, e^{-\alpha x}\, x^{\beta-1}\, \mathbf{1}_{(0,\infty)}(x)$$

Show that the Fisher's matrix is

$$\boldsymbol{I}(\alpha, \beta) = \begin{pmatrix} \beta\alpha^{-2} & -\alpha^{-1} \\ -\alpha^{-1} & \Psi'(\beta) \end{pmatrix}$$

with $\Psi'(x)$ the first derivative of the Digamma Function and, following Jeffreys' rule, we should take the prior

$$\pi(\alpha, \beta) \propto \alpha^{-1} \left[\beta \Psi'(\beta) - 1\right]^{1/2}$$

Note that α is a scale parameter so, from previous considerations, we should take $\pi(\alpha) \propto \alpha^{-1}$. Furthermore, if we consider α and β independently, we shall get

$$\pi(\alpha, \beta) = \pi(\alpha)\pi(\beta) \propto \alpha^{-1} \left[\Psi'(\beta)\right]^{1/2}$$

Example 2.11 (*The Beta Distribution*)
Show that for the $Be(x|\alpha, \beta)$ distribution with density

$$p(x|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \left(x^{\alpha-1} (1 - x)^{\beta-1} \mathbf{1}_{[0,1]}(x); \qquad \alpha, \beta \in R^+\right.$$

the Fisher's matrix is given by

$$\boldsymbol{I}(\alpha, \beta) = \begin{pmatrix} \Psi'(\alpha) - \Psi'(\alpha + \beta) & -\Psi'(\alpha + \beta) \\ -\Psi'(\alpha + \beta) & \Psi'(\beta) - \Psi'(\alpha + \beta) \end{pmatrix}$$

with $\Psi'(x)$ the first derivative of the Digamma Function.

Example 2.12 (*The Normal Distribution*)

Univariate: The Fisher's matrix is given by

$$\boldsymbol{I}(\mu, \sigma) = \begin{pmatrix} \sigma^{-2} & 0 \\ 0 & 2\sigma^{-2} \end{pmatrix}$$

so

$$\pi_1(\mu, \sigma) \propto [\det[\boldsymbol{I}(\mu, \sigma]]^{1/2} \propto \frac{1}{\sigma^2}$$

However, had we treated the two parameters independently, we should have obtained

$$\pi_2(\mu, \sigma) = \pi(\mu)\, \pi(\sigma) \propto \frac{1}{\sigma}$$

The prior $\pi_2 \propto \sigma^{-1}$ is the one we had used in Example 2.5 where the problem was treated as two one-dimensional independent problems and, as we saw:

$$T = \frac{\sqrt{n-1}(\mu - \overline{x})}{s} \sim St(t|n-1) \quad \text{and} \quad Z = \frac{n\, s^2}{\sigma^2} \sim \chi^2(z|n-1)$$

with $E[Z] = n - 1$. Had we used prior $\pi_1 \propto \sigma^{-2}$, we would have obtained that $Z \sim \chi^2(z|n)$ and therefore $E[Z] = n$. This is not reasonable. On the one hand, we know from the sampling distribution $N(x|\mu, \sigma)$ that $E[ns^2\sigma^{-2}] = n - 1$. On the other hand, we have two parameters (μ, σ) and integrate on one (σ) so the number of degrees of freedom should be $n - 1$.

Bivariate: The Fisher's matrix is given by

$$\boldsymbol{I}(\mu_1, \mu_2) = (1-\rho^2)^{-1} \begin{pmatrix} \sigma_1^{-2} & -\rho(\sigma_1\sigma_2)^{-1} \\ -\rho(\sigma_1\sigma_2)^{-1} & \sigma_2^{-2} \end{pmatrix}$$

$$\boldsymbol{I}(\sigma_1, \sigma_2, \rho) = (1-\rho^2)^{-1} \begin{pmatrix} (2-\rho^2)\sigma_1^{-2} & -\rho^2(\sigma_1\sigma_2)^{-1} & -\rho\sigma_1^{-1} \\ -\rho^2(\sigma_1\sigma_2)^{-1} & (2-\rho^2)\sigma_2^{-2} & -\rho\sigma_2^{-1} \\ -\rho\sigma_1^{-1} & -\rho\sigma_2^{-1} & (1+\rho^2)(1-\rho^2)^{-1} \end{pmatrix}$$

$$\boldsymbol{I}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \begin{pmatrix} \boldsymbol{I}(\mu_1, \mu_2) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}(\sigma_1, \sigma_2, \rho) \end{pmatrix}$$

From this,

$$\pi(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) \propto |\det \boldsymbol{I}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)|^{1/2} = \frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)^2}$$

while if we consider $\pi(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \pi(\mu_1, \mu_2)\pi(\sigma_1, \sigma_2, \rho)$ we get

$$\pi(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) \propto \frac{1}{\sigma_1\sigma_2(1-\rho^2)^{3/2}}$$

Problem 2.1 Show that for the density $p(\boldsymbol{x}|\boldsymbol{\theta})$; $\boldsymbol{x} \in \Omega \subseteq R^n$, the Fisher's matrix (if exists)

$$\boldsymbol{I}_{ij}(\boldsymbol{\theta}) = E_X \left[\frac{\partial \log\, p(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \theta_i} \, \frac{\partial \log\, p(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \theta_j} \right]$$

transforms under a differentiable one-to-one transformation $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta})$ as a covariant symmetric tensor of second order; that is

$$\boldsymbol{I}_{ij}(\boldsymbol{\phi}) = \frac{\partial \theta_k}{\partial \phi_i} \, \frac{\partial \theta_l}{\partial \phi_j} \, \boldsymbol{I}_{kl}(\boldsymbol{\theta})$$

Problem 2.2 Show that for $X \sim Po(x|\mu + b)$ with $b \in R^+$ known (Poisson model with known background), we have that $\boldsymbol{I}(\mu) = (\mu + b)^{-1}$ and therefore the posterior (proper) is given by:

$$p(\mu|x, b) \propto e^{-(\mu+b)} \, (\mu + b)^{x-1/2}$$

Problem 2.3 Show that for the one parameter mixture model $p(x|\lambda) = \lambda p_1(x) + (1-\lambda)p_2(x)$ with $p_1(x) \neq p_2(x)$ properly normalized and $\lambda \in (0, 1)$,

$$\boldsymbol{I}(\lambda) = \frac{1}{\lambda(1-\lambda)} \left\{ 1 - \int_{-\infty}^{\infty} \frac{p_1(x)p_2(x)}{p(x|\lambda)} dx \right\}$$

When $p_1(x)$ and $p_2(x)$ are *"well separated"*, the integral is $<< 1$ and therefore $\boldsymbol{I}(\lambda) \sim [\lambda(1-\lambda)]^{-1}$. On the other hand, when they *"get closer"* we can write $p_2(x) = p_1(x) + \eta(x)$ with $\int_{-\infty}^{\infty} \eta(x)dx = 0$ and, after a Taylor expansion for $|\eta(x)| << 1$ get to first order that

$$\boldsymbol{I}(\lambda) \simeq \int_{-\infty}^{\infty} \frac{(p_1(x) - p_2(x))^2}{p_1(x)} dx + \cdots$$

independent of λ. Thus, for this problem it will be sound to consider the prior $\pi(\lambda|a, b) = Be(\lambda|a, b)$ with parameters between $(1/2, 1/2)$ and $(1, 1)$.

2.6.4 Invariance Under a Group of Transformations

Some times, we may be interested to provide the prior with invariance under some transformations of the parameters (or a subset of them) considered of interest for the problem at hand. As we have stated, from a formal point of view the prior can be treated as an absolute continuous measure with respect to Lebesgue so $p(\boldsymbol{\theta}|\boldsymbol{x})\, d\boldsymbol{\theta} \propto p(\boldsymbol{x}|\boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} = p(\boldsymbol{x}|\boldsymbol{\theta})\, d\mu(\boldsymbol{\theta})$. Now, consider the probability space (Ω, B, μ) and a measurable homeomorphism $T : \Omega \to \Omega$. A measure μ on the Borel algebra B would be invariant by the mapping T if for any $A \subset B$, we have that $\mu(T^{-1}(A)) = \mu(A)$. We know, for instance, that there is a unique measure λ on R^n that is invariant under translations and such that for the unit cube $\lambda([0, 1]^n) = 1$: the Lebesgue measure (in fact, it could have been defined that way). This is consistent with the constant prior specified already for position parameters. The Lebesgue measure is also the unique measure in R^n that is invariant under the rotation group $SO(n)$ (see Problem 2.5). Thus, when expressed in spherical polar coordinates, it would be reasonable for the spherical surface S^{n-1} the rotation invariant prior

$$d\mu(\boldsymbol{\phi}) = \prod_{k=1}^{n-1} (\sin \phi_k)^{(n-1)-k}\, d\phi_k$$

with $\phi_{n-1} \in [0, 2\pi)$ and $\phi_j \in [0, \pi]$ for the rest. We shall use this prior function in a later problem.

In other cases, the group of invariance is suggested by the model

$$M : \{p(\boldsymbol{x}|\boldsymbol{\theta}),\ \boldsymbol{x} \in \Omega_X,\ \boldsymbol{\theta} \in \Omega_\Theta\}$$

in the sense that we can make a transformation of the random quantity $\boldsymbol{X} \rightarrow \boldsymbol{X}'$ and absorb the change in a redefinition of the parameters $\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}'$ such that the expression of the probability density remains unchanged. Consider a group of transformations[6] G that acts

$$\text{on the Sample Space:} \quad x \rightarrow x' = g \circ x; \qquad g \in G; x, x' \in \Omega_X$$

$$\text{on the Parametric Space:} \quad \theta \rightarrow \theta' = g \circ \theta; \qquad g \in G; \theta, \theta' \in \Omega_\Theta$$

The model M is said to be invariant under G if $\forall g \in G$ and $\forall \theta \in \Omega_\Theta$ the random quantity $X' = g \circ X$ is distributed as $p(x'|\theta') \equiv p(g \circ x | g \circ \theta)$. Therefore, transformations of data under G will make no difference on the inferences if we assign consistent "prior beliefs" to the original and transformed parameters. Note that the action of the group on the sample and parameter spaces will, in general, be different. The essential point is that, as Alfred Haar showed in 1933, for the action of the group G of transformations there is an invariant measure μ (*Haar measure*; [8]) such that

$$\int_{\Omega_X} f(g \circ x) d\mu(x) = \int_{\Omega_X} f(x') d\mu(x')$$

for any Lebesgue integrable function $f(x)$ on Ω_X. Shortly after, it was shown (Von Neumann (1934); Weil and Cartan (1940)) that this measure is unique up to a multiplicative constant. In our case, the function will be $p(\cdot|\theta)\mathbf{1}_\Theta(\theta)$ and the invariant measure we are looking for is $d\mu(\theta) \propto \pi(\theta) d\theta$. Furthermore, since the group may be non-abelian, we shall consider the action on the right and on the left of the parameter space. Thus, we shall have:

$$\int_\Theta p(\cdot|g \circ \theta)\, \pi_L(\theta)\, d\theta = \int_\Theta p(\cdot|\theta')\, \pi_L(\theta')\, d\theta'$$

if the group acts on the left and

$$\int_\Theta p(\cdot|\theta \circ g)\, \pi_R(\theta)\, d\theta = \int_\Theta p(\cdot|\theta')\, \pi_R(\theta')\, d\theta'$$

if the action is on the right. Then, we should start by identifying the group of transformations under which the model is invariant (if any; in many cases, either there is no invariance or at least not obvious) work in the parameter space. The most interesting cases for us are:

[6] In this context, the use of Transformation Groups arguments was pioneered by E.T. Jaynes [7].

$$\text{Affine Transformations: } x \to x' = g \circ x = a + b\,x$$

$$\text{Matrix Transformations: } x \to x' = g \circ x = R\,x$$

Translations and scale transformations are a particular case of the first and rotations of the second. Let's start with the location and scale parameters; that is, a density

$$p(x|\mu, \sigma)\,dx = \frac{1}{\sigma}\,f\left(\frac{x-\mu}{\sigma}\right)dx$$

the Affine group $G = \{g \equiv (a, b); a \in R; b \in R^+\}$ so $x' = g \circ x = a + bx$ and the model will be invariant if

$$(\mu', \sigma') = g \circ (\mu, \sigma) = (a, b) \circ (\mu, \sigma) = (a + b\mu, b\sigma)$$

Now,

$$\int p(\cdot|\mu', \sigma')\,\pi_L(\mu', \sigma')\,d\mu'\,d\sigma' = \int p(\cdot|g \circ (\mu, \sigma)\,\pi_L(\mu, \sigma)\,d\mu\,d\sigma =$$

$$= \int p(\cdot|\mu', \sigma')\,\left\{\pi_L[g^{-1}(\mu', \sigma')]\,J(\mu', \sigma'; \mu, \sigma)\right\}\,d\mu'\,d\sigma' =$$

$$= \int p(\cdot|\mu', \sigma')\,\left\{\pi_L\left(\frac{\mu'-a}{b}, \frac{\sigma'}{b}\right)\frac{1}{b^2}\right\}\,d\mu'\,d\sigma'$$

and this should hold for all $(a, b) \in R \times R^+$ so, in consequence:

$$d\mu_L(\mu, \sigma) = \pi_L(\mu, \sigma)\,d\mu\,d\sigma \propto \frac{1}{\sigma^2}\,d\mu\,d\sigma$$

However, the group of Affine Transformations is non-abelian so if we study the action on the left, there is no reason why we should not consider also the action on the right. Since

$$(\mu', \sigma') = (\mu, \sigma) \circ g = (\mu, \sigma) \circ (a, b) = (\mu + a\sigma, b\sigma)$$

the same reasoning leads to (left as exercise):

$$d\mu_R(\mu, \sigma) = \pi_R(\mu, \sigma)\,d\mu\,d\sigma \propto \frac{1}{\sigma}\,d\mu\,d\sigma$$

The first one (π_L) is the one we obtain using Jeffrey's rule in two dimensions while π_R is the one we get for position and scale parameters or Jeffrey's rule treating both parameters independently; that is, as two one-dimensional problems instead a one two-dimensional problem. Thus, although from the invariance point of view there is no reason why one should prefer one over the other, the right invariant Haar prior

gives more consistent results. In fact ([9, 10]), a necessary and sufficient condition for a sequence of posteriors based on proper priors to converge in probability to an invariant posterior is that the prior is the right Haar measure.

Problem 2.4 As a remainder, given a measure space $(\Omega, \mathcal{B}, \mu)$ a mapping $T : \Omega \longrightarrow \Omega$ is measurable if $T^{-1}(A) \in \mathcal{B}$ for all $A \in \mathcal{B}$ and the measure μ is invariant under T if $\mu(T^{-1}(A)) = \mu(A)$ for all $A \in \mathcal{B}$. Show that the measure $d\mu(\theta) = [\theta(1-\theta)]^{-1/2} d\theta$ is invariant under the mapping $T : [0, 1] \rightarrow [0, 1]$ such that $T : \theta \rightarrow \theta' = T(\theta) = 4\theta(1-\theta)$. This is the Jeffrey's prior for the Binomial model $Bi(x|N, \theta)$.

Problem 2.5 Consider the n-dimensional spherical surface S_n of unit radius, $\boldsymbol{x} \in S_n$ and the transformation $\boldsymbol{x}' = \boldsymbol{R}\boldsymbol{x} \in S_n$ where $\boldsymbol{R} \in SO(n)$. Show that the Haar invariant measure is the Lebesgue measure on the sphere.

Hint: Recall that $\boldsymbol{R}$ is an orthogonal matrix so $\boldsymbol{R}^t = \boldsymbol{R}^{-1}$; that $|\det \boldsymbol{R}| = 1$ so $J(\boldsymbol{x}'; \boldsymbol{x}) = |\partial \boldsymbol{x}/\partial \boldsymbol{x}')| = |\partial \boldsymbol{R}^{-1}\boldsymbol{x}'/\partial \boldsymbol{x}')| = |\det \boldsymbol{R}| = 1$ and that $\boldsymbol{x}'^t \boldsymbol{x}' = \boldsymbol{x}^t \boldsymbol{x} = 1$.

Example 2.12 (*Bivariate Normal Distribution*) Let $\boldsymbol{X} = (X_1, X_2) \sim N(\boldsymbol{x}|\boldsymbol{0}, \phi)$ with $\phi = \{\sigma_1, \sigma_2, \rho\}$; that is:

$$p(\boldsymbol{x}|\phi) = (2\pi)^{-1} |\det[\boldsymbol{\Sigma}]|^{-1/2} \exp\left\{-\frac{1}{2}\left(\boldsymbol{x}^t\, \boldsymbol{\Sigma}^{-1}\, \boldsymbol{x}\right)\right\}$$

with the covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \quad \text{and} \quad \det[\boldsymbol{\Sigma}] = \sigma_1^2\sigma_2^2(1-\rho)^2$$

Using the Cholesky decomposition we can express $\boldsymbol{\Sigma}^{-1}$ as the product of two lower (or upper) triangular matrices:

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\det[\boldsymbol{\Sigma}]} \begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix} = \boldsymbol{A}^t \boldsymbol{A} \quad \text{with} \quad \boldsymbol{A} = \begin{pmatrix} \frac{1}{\sigma_1} & 0 \\ \frac{-\rho}{\sigma_1\sqrt{1-\rho^2}} & \frac{1}{\sigma_2\sqrt{1-\rho^2}} \end{pmatrix}$$

For the action on the left:

$$\boldsymbol{M} = \boldsymbol{T} = \begin{pmatrix} a & 0 \\ b & c \end{pmatrix}; a, b > 0 \quad \longrightarrow \quad J(\boldsymbol{A}'; \boldsymbol{A}) = a^2 c$$

and, in consequence

$$\pi(aa'_{11}, aa'_{21} + ba'_{22}, ca'_{22})\, ac^2 = \pi(a'_{11}, a'_{21}, a'_{22}) \longrightarrow \pi(a'_{11}, a'_{21}, a'_{22}) \propto \frac{1}{{a'_{11}}^2 a'_{22}}$$

and $\det[\boldsymbol{\Sigma}] = (\det[\boldsymbol{\Sigma}^{-1}])^{-1} = (\det[\boldsymbol{A}])^{-2}$. Thus, in the new parameterization $\boldsymbol{\theta} = \{a_{11}, a_{21}, a_{22}\}$

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = (2\pi)^{-1}\,|\det[\boldsymbol{A}]|\exp\left\{-\frac{1}{2}\left(\boldsymbol{x}^t\,\boldsymbol{A}^t\boldsymbol{A}\,\boldsymbol{x}\right)\right\}$$

Consider now the group of lower triangular $2x2$ matrices

$$G_l = \{\boldsymbol{T}\ \in\ LT_{2x2};\quad T_{ii} > 0\}$$

Since $\boldsymbol{T}^{-1} \in G_l$, inserting the identity matrix $\boldsymbol{I} = \boldsymbol{T}\boldsymbol{T}^{-1} = \boldsymbol{T}^{-1}\boldsymbol{T}$ we have: action

On the Left	On the Right
$\boldsymbol{T}\circ\boldsymbol{x}\rightarrow\boldsymbol{T}\boldsymbol{x} = \boldsymbol{x}'$	$\boldsymbol{x}\circ\boldsymbol{T}\rightarrow\boldsymbol{T}^{-1}\boldsymbol{x} = \boldsymbol{x}'$
$\left[\boldsymbol{x}^t\,(\boldsymbol{T}^t(\boldsymbol{T}^t)^{-1})\,\boldsymbol{A}^t\boldsymbol{A}\,(\boldsymbol{T}^{-1}\boldsymbol{T})\,\boldsymbol{x}\right]$	$\left[\boldsymbol{x}^t\,((\boldsymbol{T}^t)^{-1}\boldsymbol{T}^t)\,\boldsymbol{A}^t\boldsymbol{A}\,(\boldsymbol{T}\boldsymbol{T}^{-1})\,\boldsymbol{x}\right]$
$\boldsymbol{M} = \boldsymbol{T}$	$\boldsymbol{M} = \boldsymbol{T}^{-1}$

Then

$$\boldsymbol{M}\boldsymbol{x} = \boldsymbol{x}';\quad \boldsymbol{x} = \boldsymbol{M}^{-1}\boldsymbol{x}';\quad \boldsymbol{x}'^t = \boldsymbol{x}^t\boldsymbol{M}^t\quad \text{and}\quad d\boldsymbol{x} = \frac{1}{|\det[\boldsymbol{M}]|}d\boldsymbol{x}'$$

so

$$p(\boldsymbol{x}'|\boldsymbol{\theta}) = (2\pi)^{-1}\,\frac{|\det[\boldsymbol{A}]|}{|\det[\boldsymbol{M}]|}\exp\left\{-\frac{1}{2}\left(\boldsymbol{x}'^t\,(\boldsymbol{A}\boldsymbol{M}^{-1})^t(\boldsymbol{A}\boldsymbol{M}^{-1})\,\boldsymbol{x}'\right)\right\}$$

and the model is invariant under G_l if the action on the parameter space is

$$G_l : \boldsymbol{A}\longrightarrow\boldsymbol{A}' = \boldsymbol{A}\boldsymbol{M}^{-1};\quad \boldsymbol{A} = \boldsymbol{A}'\boldsymbol{M};\quad \det[\boldsymbol{A}] = \det[\boldsymbol{A}']\det[\boldsymbol{M}]$$

so

$$p(\boldsymbol{x}'|\boldsymbol{\theta}') = (2\pi)^{-1}\,|\det[\boldsymbol{A}']|\exp\left\{-\frac{1}{2}\left(\boldsymbol{x}'^t\,\boldsymbol{A}'^t\boldsymbol{A}'\,\boldsymbol{x}'\right)\right\}$$

Then, the Haar equation reads

$$\int_\Theta p(\bullet|\boldsymbol{A}')\,\pi(\boldsymbol{A}')\,d\boldsymbol{A}' = \int_\Theta p(\bullet|g\circ\boldsymbol{A})\,\pi(\boldsymbol{A})\,d\boldsymbol{A} = \int_\Theta p(\bullet|\boldsymbol{A}')\,\pi(\boldsymbol{A}'\boldsymbol{M})\,J(\boldsymbol{A}';\,\boldsymbol{A})\,d\boldsymbol{A}$$

and, in consequence, $\forall\boldsymbol{M}\in G$

$$\pi(\boldsymbol{A}'\boldsymbol{M})\,J(\boldsymbol{A}';\,\boldsymbol{A})\,da'_{11}da'_{21}da'_{22} = \pi(\boldsymbol{A}')\,da'_{11}da'_{21}da'_{22}$$

For the action on the left:

$$\boldsymbol{M} = \boldsymbol{T} = \begin{pmatrix} a & 0 \\ b & c \end{pmatrix}; a, b > 0 \quad \longrightarrow \quad J(\boldsymbol{A}'; \boldsymbol{A}) = a^2 c$$

and, in consequence

$$\pi(aa'_{11}, aa'_{21} + ba'_{22}, ca'_{22})\, a^2 c = \pi(a'_{11}, a'_{21}, a'_{22}) \;\longrightarrow\; \pi(a'_{11}, a'_{21}, a'_{22}) \propto \frac{1}{{a'_{11}}^2 a'_{22}}$$

For the action on the right:

$$\boldsymbol{M} = \boldsymbol{T}^{-1} = \begin{pmatrix} a^{-1} & 0 \\ -b(ac)^{-1} & c^{-1} \end{pmatrix} \quad \longrightarrow \quad J(\boldsymbol{A}'; \boldsymbol{A}) = (ac^2)^{-1}$$

and, in consequence

$$\pi\left(\frac{a'_{11}}{a}, \frac{ca'_{21} - ba'_{22}}{ac}, \frac{a'_{22}}{c}\right) \frac{1}{ac^2} = \pi(a'_{11}, a'_{21}, a'_{22}) \;\longrightarrow\; \pi(a'_{11}, a'_{21}, a'_{22}) \propto \frac{1}{a'_{11}{a'_{22}}^2}$$

In terms of the parameters of interest $\{\sigma_1, \sigma_2, \rho\}$, since

$$da_{11} da_{21} da_{22} = \frac{1}{\sigma_1^2 \sigma_2^2 (1 - \rho^2)^2}\, d\sigma_1 d\sigma d\rho$$

we have finally that for invariance under G_l:

$$\pi_L^l(\sigma_1, \sigma_2, \rho) = \frac{1}{\sigma_1 \sigma_2 (1 - \rho^2)^{3/2}} \qquad \text{and} \qquad \pi_R^l(\sigma_1, \sigma_2, \rho) = \frac{1}{\sigma_2^2 (1 - \rho^2)}$$

The same analysis with decomposition in upper triangular matrices leads to

$$\pi_L^u(\sigma_1, \sigma_2, \rho) = \frac{1}{\sigma_1 \sigma_2 (1 - \rho^2)^{3/2}} \qquad \text{and} \qquad \pi_R^u(\sigma_1, \sigma_2, \rho) = \frac{1}{\sigma_1^2 (1 - \rho^2)}$$

As we see, in both cases the left Haar invariant prior coincides with Jeffrey's prior when $\{\mu_1, \mu_2\}$ and $\{\sigma_1, \sigma_2, \rho\}$ are decoupled.

At this point, one may be tempted to use a right Haar invariant prior where the two parameters σ_1 and σ_2 are treated on equal footing

$$\pi(\sigma_1, \sigma_2, \rho) = \frac{1}{\sigma_1 \sigma_2 (1 - \rho^2)}$$

Under this prior, since the sample correlation

$$r = \frac{\sum_i (x_{1i} - \overline{x}_1)(x_{2i} - \overline{x}_2)}{\left(\sum_i (x_{1i} - \overline{x}_1)^2 \sum_i (x_{2i} - \overline{x}_2)^2\right)^{1/2}}$$

is a sufficient statistics for ρ, we have that the posterior for inferences on the correlation coefficient will be

$$p(\rho|\boldsymbol{x}) \propto (1-\rho^2)^{(n-3)/2} F(n-1, n-1, n-1/2; (1+r\rho)/2)$$

with $F(a, b, c; z)$ the Hypergeometric Function.

Example 2.13 If $\theta \in \Theta \longrightarrow g \circ \theta = \phi(\theta) = \theta' \in \Theta$ with $\phi(\theta)$ is a one-to-one differentiable mapping, then

$$\int_\Theta p(\bullet|\theta') d\mu(\theta) = \int_\Theta p(\bullet|\theta')\pi(\theta) d\theta = \int_\Theta p(\bullet|\theta')\pi(\phi^{-1}(\theta')) \left|\frac{\partial \phi^{-1}(\theta')}{\partial \theta'}\right| d\theta' =$$
$$= \int_\Theta p(\bullet|\theta')\pi(\theta') d\theta' = \int_\Theta p(\bullet|\theta') d\mu(\theta')$$

and therefore, Jeffreys' prior defines a Haar invariant measure.

2.6.5 Conjugated Distributions

In as much as possible, we would like to consider reference priors $\pi(\theta|a, b, \ldots)$ versatile enough such that by varying some of the parameters $a, b, \ldots$ we get diverse forms to analyze the effect on the final results and, on the other hand, to simplify the evaluation of integrals like:

$$p(x) = \int p(x|\theta) \cdot p(\theta)\, d\theta \qquad \text{and} \qquad p(y|x) = \int p(y|\theta) \cdot p(\theta|x)\, d\theta$$

This leads us to consider as reference priors the *Conjugated Distributions* [11].

Let $\mathcal{S}$ be a class of sampling distributions $p(x|\theta)$ and $\mathcal{P}$ the class of prior densities for the parameter θ. If

$$p(\theta|x) \in \mathcal{P} \qquad \text{for all} \qquad p(x|\theta) \in \mathcal{S} \quad \text{and} \quad p(\theta) \in \mathcal{P}$$

we say that the class $\mathcal{P}$ is conjugated to $\mathcal{S}$. We are mainly interested in the class of priors $\mathcal{P}$ that have the same functional form as the likelihood. In this case, since both the prior density and the posterior belong to the same family of distributions, we say that they are *closed under sampling*. It should be stressed that the criteria for

taking conjugated reference priors is eminently practical and, in many cases, they do not exist. In fact, only the *exponential family* of distributions has conjugated prior densities. Thus, if $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$ is an exchangeable random sampling from the k-parameter regular exponential family, then

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = f(\boldsymbol{x})\, g(\boldsymbol{\theta}) \exp\left\{\sum_{j=1}^{k} c_j\, \phi_j(\boldsymbol{\theta}) \left(\sum_{i=1}^{n} h_j(x_i)\right)\right\}$$

and the *conjugated prior* will have the form:

$$\pi(\boldsymbol{\theta}|\boldsymbol{\tau}) = \frac{1}{K(\boldsymbol{\tau})}\, [g(\boldsymbol{\theta})]^{\tau_0} \exp\left\{\sum_{j=1}^{k} c_j\, \phi_j(\boldsymbol{\theta})\, \tau_j\right\}$$

where $\boldsymbol{\theta} \in \Theta$, $\boldsymbol{\tau} = \{\tau_0, \tau_1, \ldots, \tau_k\}$ the *hyperparameters* and $K(\boldsymbol{\tau}) < \infty$ the normalization factor so $\int_\Theta \pi(\boldsymbol{\theta}|\boldsymbol{\tau}) d\boldsymbol{\theta} = 1$. Then, the general scheme will be[7]:

(1) Choose the class of priors $\pi(\boldsymbol{\theta}|\boldsymbol{\tau})$ that reflect the structure of the model;
(2) Choose a prior function $\pi(\boldsymbol{\tau})$ for the *hyperparameters*;
(3) Express the posterior density as $p(\boldsymbol{\theta}, \boldsymbol{\tau}|\boldsymbol{x}) \propto p(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\tau})\pi(\boldsymbol{\tau})$;
(4) Marginalize for the parameters of interest:

$$p(\boldsymbol{\theta}|\boldsymbol{x}) \propto \int_\Phi p(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\tau})\pi(\boldsymbol{\tau}) d\boldsymbol{\tau}$$

or, if desired, get the conditional density

$$p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\tau}) = \frac{p(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\tau})}{p(\boldsymbol{x}, \boldsymbol{\tau})} = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\tau})}{p(\boldsymbol{x}|\boldsymbol{\tau})}$$

The obvious question that arises is how do we choose the prior $\pi(\phi)$ for the hyperparameters. Besides *reasonableness*, we may consider two approaches. Integrating the parameters $\boldsymbol{\theta}$ of interest, we get

$$p(\boldsymbol{\tau}, \boldsymbol{x}) = \pi(\boldsymbol{\tau}) \int_\Theta p(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\tau})\, d\boldsymbol{\theta} = \pi(\boldsymbol{\tau})\, p(\boldsymbol{x}|\boldsymbol{\tau})$$

so we may use any of the procedures under discussion to take $\pi(\boldsymbol{\tau})$ as the prior for the model $p(\boldsymbol{x}|\boldsymbol{\tau})$ and then obtain

$$\pi(\boldsymbol{\theta}) = \int_{\Omega_\tau} \pi(\boldsymbol{\theta}|\boldsymbol{\tau})\, \pi(\boldsymbol{\tau})\, d\boldsymbol{\tau}$$

[7]We can go an step upwards and assign a prior to the hyperparameters with hyper-hyperparameters,…

The beauty of Bayes rule but not very practical in complicated situations. A second approach, more ugly and practical, is the so called *Empirical Method* where we assign numeric values to the hyperparameters suggested by $p(\boldsymbol{x}|\boldsymbol{\tau})$ (for instance, moments, maximum-likelihood estimation,…); that is, setting, in a distributional sense, $\pi(\boldsymbol{\tau}) = \delta_{\boldsymbol{\tau}^\star}$ so $\langle \pi(\boldsymbol{\tau}), p(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{\tau})\rangle = p(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{\tau}^\star)$. Thus,

$$p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\tau}^\star) \propto\ p(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\tau}^\star)$$

Obviously, fixing the hyperparameters assumes a perfect knowledge of them and does not allow for variations but the procedure may be useful to guess at least were to go.

Last, it may happen that a single conjugated prior does not represent sufficiently well our beliefs. In this case, we may consider a k-mixture of conjugated priors

$$\pi(\boldsymbol{\theta}|\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_k) = \sum_{i=1}^{k} w_i\, \pi(\boldsymbol{\theta}|\boldsymbol{\tau}_i)$$

In fact [12], any prior density for a model that belongs to the exponential family can be approximated arbitrarily close by a mixture of conjugated priors.

Example 2.14 Let's see the conjugated prior distributions for some models:
• **Poisson model** $Po(n|\mu)$: Writing

$$p(n|\mu) = \frac{e^{-\mu}\,\mu^n}{\Gamma(n+1)} = \frac{e^{-(\mu - n\log\mu)}}{\Gamma(n+1)}$$

it is clear that the Poisson distribution belongs to the exponential family and the conjugated prior density for the parameter μ is

$$\pi(\mu|\tau_1, \tau_2) \propto e^{-\tau_1\mu + \tau_2 \log\mu} \propto Ga(\mu|\tau_1, \tau_2)$$

If we set a prior $\pi(\tau_1, \tau_2)$ for the hyperparameters we can write

$$p(n, \mu, \tau_1, \tau_2)\; p(n|\mu)\, \pi(\mu|\tau_1, \tau_2) = \pi(\tau_1, \tau_2)$$

and integrating μ:

$$p(n, \tau_1, \tau_2) = \left[\frac{\Gamma(n+\tau_2)}{\Gamma(\tau_1)}\,\frac{\tau_1^{\tau_2}}{(1+\tau_1)^{n+\tau_2}}\right]\pi(\tau_1, \tau_2) = p(n|\tau_1, \tau_2)\,\pi(\tau_1, \tau_2)$$

• **Binomial model** $Bi(n|N, \theta)$: Writing

$$P(n|N, \theta) = \binom{N}{n}\,\theta^n\,(1-\theta)^{N-n} = \binom{N}{n}\,e^{n\log\theta + (N-n)\log(1-\theta)}$$

it is clear that it belong to the exponential family and the conjugated prior density for the parameter θ will be:

$$\pi(\theta|\tau_1, \tau_2) = Be(\tau|\tau_1, \tau_2)$$

• **Multinomial model** Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_k) \sim Mn(\boldsymbol{x}|\boldsymbol{\theta})$; that is:

$$\boldsymbol{X} \sim p(\boldsymbol{x}|\boldsymbol{\theta}) = \Gamma(n+1) \prod_{i=1}^{k} \frac{\theta_i^{x_i}}{\Gamma(x_i+1)} \qquad \begin{cases} X_i \in N, & \sum_{i=1}^{k} X_i = n \\ \theta_i \in [0,1], & \sum_{i=1}^{k} \theta_i = 1 \end{cases}$$

The Dirichlet distribution $Di(\boldsymbol{\theta}|\boldsymbol{\alpha})$:

$$\pi(\boldsymbol{\theta}|\boldsymbol{\alpha}) = D(\boldsymbol{\alpha}) \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \qquad \begin{cases} \boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_k), \quad \alpha_i > 0, \quad \sum_{i=1}^{k} \alpha_i = \alpha_0 \\ D(\boldsymbol{\alpha}) = \Gamma(\alpha_0) \left[\prod_{i=1}^{k} \Gamma(\alpha_i) \right]^{-1} \end{cases}$$

is the natural conjugated prior for this model. It is a degenerated distribution in the sense that

$$\pi(\boldsymbol{\theta}|\boldsymbol{\alpha}) = D(\boldsymbol{\alpha}) \left[\prod_{i=1}^{k-1} \theta_i^{\alpha_i - 1} \right] \left[1 - \sum_{i=1}^{k-1} \theta_i \right]^{\alpha_k - 1}$$

The posterior density will then be $\boldsymbol{\theta} \sim Di(\boldsymbol{\theta}|\boldsymbol{x} + \boldsymbol{\alpha})$ with

$$E[\theta_i] = \frac{x_i + \alpha_i}{n + \alpha_0} \qquad \text{and} \qquad V[\theta_i, \theta_j] = \frac{E[\theta_i](\delta_{ij} - E[\theta_j])}{n + \alpha_0 + 1}$$

The parameters $\boldsymbol{\alpha}$ of the Dirichlet distribution $Di(\boldsymbol{\theta}|\boldsymbol{\alpha})$ determine the expected values $E[\theta_i] = \alpha_i/\alpha_0$. In practice, it is more convenient to control also the variances and use the *Generalized Dirichlet Distribution* $GDi(\boldsymbol{\theta}|\boldsymbol{\alpha}, \boldsymbol{\beta})$:

$$\pi(\boldsymbol{\theta}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^{k-1} \frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)} \theta_i^{\alpha_i - 1} \left[1 - \sum_{j=1}^{i} \theta_j \right]^{\gamma_i}$$

where:

$$0 < \theta_i < 1, \quad \sum_{i=1}^{k-1} \theta_i < 1, \quad \theta_n = 1 - \sum_{i=1}^{k-1} \theta_i$$

$$\alpha_i > 0, \quad \beta_i > 0, \quad \text{and} \quad \gamma_i \begin{cases} \beta_i - \alpha_{i+1} - \beta_{i+1}; & i = 1, 2, \ldots, k-2 \\ \beta_{k-1} - 1; & i = k-1 \end{cases}$$

When $\beta_i = \alpha_{i+1} + \beta_{i+1}$ it becomes the Dirichlet distribution. For this prior we have that

$$E[\theta_i] = \frac{\alpha_i}{\alpha_i + \beta_i} S_i \qquad \text{and} \qquad V[\theta_i, \theta_j] = E[\theta_j] \left(\frac{\alpha_i + \delta_{ij}}{\alpha_i + \beta_i + 1} T_i - E[\theta_i] \right)$$

where

$$S_i = \prod_{j=1}^{i-1} \frac{\beta_j}{\alpha_j + \beta_j} \qquad \text{and} \qquad T_i = \prod_{j=1}^{i-1} \frac{\beta_j + 1}{\alpha_j + \beta_j + 1}$$

with $S_1 = T_1 = 1$ and we can have control over the prior means and variances.

2.6.6 Probability Matching Priors

A pragmatic criteria is that of *probability matching priors* for which the one sided credible intervals derived from the posterior distribution coincide, to a certain level of accuracy, with those derived by the classical approach. This condition leads to a differential equation for the prior distribution [13, 14]. We shall illustrate in the following lines the rationale behind for the simple one parameter case assuming that the needed regularity conditions are satisfied.

Consider then a random quantity $X \sim p(x|\theta)$ and an iid sampling $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$ with θ the parameter of interest. The classical approach for inferences is based on the likelihood

$$p(\boldsymbol{x}|\theta) = p(x_1, x_2, \ldots x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$$

and goes through the following reasoning:

(1) Assumes that the parameter θ has the *true* but unknown value θ_0 so the sample is actually drawn from $p(x|\theta_0)$;
(2) Find the estimator $\theta_m(\boldsymbol{x})$ of θ_0 as the value of θ that maximizes the likelihood; that is:

$$\theta_m = \max_{\theta} \{p(\boldsymbol{x}|\theta)\} \quad \longrightarrow \quad \left(\frac{\partial \ln p(\boldsymbol{x}|\theta)}{\partial \theta} \right)_{\theta_m} = 0$$

(3) Given the model $X \sim p(x|\theta_0)$, after the appropriate change of variables get the distribution

$$p(\theta_m|\theta_0)$$

of the random quantity $\theta_m(X_1, X_2, \ldots X_n)$ and draw inferences from it.

The Bayesian inferential process considers a prior distribution $\pi(\theta)$ and draws inferences on θ from the posterior distribution of the quantity of interest

$$p(\theta|\boldsymbol{x}) \;\propto\; p(\boldsymbol{x}|\theta)\,\pi(\theta)$$

Let's start with the Bayesian and expand the term on the right around θ_m. On the one hand:

$$\ln \frac{p(\boldsymbol{x}|\theta)}{p(\boldsymbol{x}|\theta_m)} = \frac{1}{2!}\left(\frac{\partial^2 \ln p(\boldsymbol{x}|\theta)}{\partial\theta^2}\right)_{\theta_m}(\theta-\theta_m)^2 + \frac{1}{3!}\left(\frac{\partial^3 \ln p(\boldsymbol{x}|\theta)}{\partial\theta^3}\right)_{\theta_m}(\theta-\theta_m)^3 + \cdots$$

Now,

$$-\frac{1}{n}\frac{\partial^2 \ln p(\boldsymbol{x}|\theta)}{\partial\theta^2} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2(-\ln p(x_i|\theta))}{\partial\theta^2} \;\overset{n\to\infty}{\longrightarrow}\; \mathrm{E}_X\left[\frac{\partial^2(-\ln p(x|\theta))}{\partial\theta^2}\right] = I(\theta)$$

so we can substitute:

$$\left(\frac{\partial^2 \ln p(\boldsymbol{x}|\theta)}{\partial\theta^2}\right)_{\theta_m} = -n\,I(\theta_m) \qquad \text{and} \qquad \left(\frac{\partial^3 \ln p(\boldsymbol{x}|\theta)}{\partial\theta^3}\right)_{\theta_m} = -n\left(\frac{\partial I(\theta)}{\partial\theta}\right)_{\theta_m}$$

to get

$$p(\boldsymbol{x}|\theta) = e^{\ln p(\boldsymbol{x}|\theta)} \;\propto\; e^{-\frac{nI(\theta_m)}{2}(\theta-\theta_m)^2}\left(1 - \frac{n}{3!}\left(\frac{\partial I(\theta)}{\partial\theta}\right)_{\theta_m}(\theta-\theta_m)^3 + \cdots\right)$$

On the other hand:

$$\pi(\theta) = \pi(\theta_m)\left(1 + \left(\frac{1}{\pi(\theta)}\frac{\partial\pi(\theta)}{\partial\theta}\right)_{\theta_m}(\theta-\theta_m) + \cdots\right)$$

so If we define the random quantity $T = \sqrt{nI(\theta_m)}(\theta-\theta_m)$ and consider that

$$I^{-3/2}(\theta)\,\frac{\partial I(\theta)}{\partial\theta} = -2\,\frac{\partial I^{-1/2}}{\partial\theta}$$

we get finally:

$$p(t|\boldsymbol{x}) = \frac{\exp(-t^2/2)}{\sqrt{2\pi}}\left(1 + \frac{1}{\sqrt{n}}\left[\left(\frac{I^{-1/2}(\theta)}{\pi(\theta)}\frac{\partial\pi(\theta)}{\partial\theta}\right)_{\theta_m} t + \frac{1}{3}\left(\frac{\partial I^{-1/2}}{\partial\theta}\right)_{\theta_m} t^3\right] + O\left(\frac{1}{n}\right)\right)$$

Let's now find

$$P(T \leq z|\boldsymbol{x}) = \int_{-\infty}^{z} p(t|\boldsymbol{x})dt$$

Defining

$$Z(x) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2} \quad \text{and} \quad P(x) = \int_{-\infty}^{x} Z(t)dt$$

and considering that

$$\int_{-\infty}^{z} Z(t)\, t\, dt = -Z(z) \quad \text{and} \quad \int_{-\infty}^{z} Z(t)\, t^3\, dt = -Z(z)\,(z^2+2)$$

it is straight forward to get:

$$P(T \leq z|\boldsymbol{x}) = P(z) - \frac{Z(z)}{\sqrt{n}}\left[\left(\frac{I^{-1/2}(\theta)}{\pi(\theta)}\frac{\partial\pi(\theta)}{\partial\theta}\right)_{\theta_m} + \frac{z^2+2}{3}\left(\frac{\partial I^{-1/2}}{\partial\theta}\right)_{\theta_m}\right] O\left(\frac{1}{n}\right)$$

From this probability distribution, we can infer what the classical approach will get. Since he will draw inferences from $p(\boldsymbol{x}|\theta_0)$, we can take a sequence of proper priors $\pi_k(\theta|\theta_0)$ for $k = 1, 2, \ldots$ that induce a sequence of distributions such that

$$\lim_{k\to\infty}\langle \pi_k(\theta|\theta_0),\, p(\boldsymbol{x}|\theta)\rangle = p(\boldsymbol{x}|\theta_0)$$

In Distributional sense, the sequence of distributions generated by

$$\pi_k(\theta|\theta_0) = \frac{k}{2}\,\mathbf{1}_{[\theta_0-1/k,\theta_0+1/k]}; \qquad k = 1, 2, \ldots$$

converge to the Delta distribution δ_{θ_0} and, from distributional derivatives, as $k\to\infty$,

$$\langle \frac{d}{d\theta}\pi_k(\theta|\theta_0),\, I^{-1/2}(\theta)\rangle = -\langle \pi_k(\theta|\theta_0),\, \frac{d}{d\theta}I^{-1/2}(\theta)\rangle \simeq -\left(\frac{\partial I^{-1/2}(\theta))}{\partial\theta}\right)_{\theta_0}$$

But $\theta_0 = \theta_m + O(1/\sqrt{n})$ so, for a sequence of priors that shrink to $\theta_0 \simeq \theta_m$,

$$P(T \leq z|\boldsymbol{x}) = P(z) - \frac{Z(z)}{\sqrt{n}}\left[\frac{z^2+1}{3}\left(\frac{\partial I^{-1/2}}{\partial\theta}\right)_{\theta_m}\right] + O\left(\frac{1}{n}\right)$$

For terms of order $O(1/\sqrt{n})$ in both expressions of $P(T \leq z|\boldsymbol{x})$ to be the same, we need that:

$$\left(\frac{1}{\sqrt{I(\theta)}}\frac{1}{\pi(\theta)}\frac{\partial\pi(\theta)}{\partial\theta}\right)_{\theta_m} = -\left(\frac{\partial I^{-1/2}}{\partial\theta}\right)_{\theta_m}$$

and therefore

$$\pi(\theta) = I^{1/2}(\theta)$$

that is, Jeffrey's prior. In the case of n-dimensional parameters, the reasoning goes along the same lines but the expressions and the development become much more lengthy and messy and we refer to the literature.

The procedure for a first order *probability matching prior* [15, 16] starts from the likelihood

$$p(x_1, x_2, \ldots x_n|\theta_1, \theta_2, \ldots \theta_p)$$

and then:

(1) Get the Fisher's matrix $\boldsymbol{I}(\theta_1, \theta_2, \ldots \theta_p)$ and the inverse $\boldsymbol{I}^{-1}(\theta_1, \theta_2, \ldots \theta_p)$;
(2) Suppose we are interested in the parameter $t = t(\theta_1, \theta_2, \ldots \theta_p)$ a twice continuous and differentiable function of the parameters. Define the column vector

$$\nabla_t = \left(\frac{\partial t}{\partial\theta_1}, \frac{\partial t}{\partial\theta_2}, \ldots, \frac{\partial t}{\partial\theta_p}\right)^T$$

(3) Define the column vector

$$\eta = \frac{\boldsymbol{I}^{-1}\nabla_t}{(\nabla_t^T\boldsymbol{I}^{-1}\nabla_t)^{1/2}} \qquad \text{so that} \qquad \eta^T\boldsymbol{I}\eta = 1$$

(4) The probability matching prior for the parameter $t = t(\boldsymbol{\theta})$ in terms of $\theta_1, \theta_2, \ldots \theta_p$ is given by the equation:

$$\sum_{k=1}^{p}\frac{\partial}{\partial\theta_k}\left[\eta_k(\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\right] = 0$$

Any solution $\pi(\theta_1, \theta_2, \ldots \theta_p)$ will do the job.
(5) Introduce $t = t(\boldsymbol{\theta})$ in this expression, say, for instance $\theta_1 = \theta_1(t, \theta_2, \ldots \theta_p)$, and the corresponding Jacobian $J(t, \theta_2, \ldots \theta_p)$. Then we get the prior for the parameter t of interest and the nuisance parameters $\theta_2, \ldots \theta_p$ that, eventually, will be integrated out.

Example 2.15 Consider two independent random quantities X_1 and X_2 such that

$$P(X_i = n_k) = Po(n_k|\mu_i).$$

We are interested in the parameter $t = \mu_1/\mu_2$ so setting $\mu = \mu_2$ we have the ordered parameterization $\{t, \mu\}$. The joint probability is

$$P(n_1, n_2|\mu_1, \mu_2) = P(n_1|\mu_1)\, P(n_2|\mu_2) = e^{-(\mu_1 + \mu_2)} \frac{\mu_1^{n_1} \mu_2^{n_2}}{\Gamma(n_1 + 1)\Gamma(n_2 + 1)}$$

from which we get the Fisher's matrix

$$\mathbf{I}(\mu_1, \mu_2) = \begin{pmatrix} 1/\mu_1 & 0 \\ 0 & 1/\mu_2 \end{pmatrix} \qquad \text{and} \qquad \mathbf{I}^{-1}(\mu_1, \mu_2) = \begin{pmatrix} \mu_1 & 0 \\ 0 & \mu_2 \end{pmatrix}$$

We are interested in the parameter $t = \mu_1/\mu_2$, a twice continuous and differentiable function of the parameters, so

$$\nabla_t(\mu_1, \mu_2) = \left(\frac{\partial t}{\partial \mu_1}, \frac{\partial t}{\partial \mu_2} \right)^T = \left(\mu_2^{-1}, -\mu_1 \mu_2^{-2} \right)^T = \begin{pmatrix} \mu_2^{-1} \\ -\mu_1 \mu_2^{-2} \end{pmatrix}$$

Therefore:

$$\mathbf{I}^{-1}\, \nabla_t = \begin{pmatrix} \mu_1 \mu_2^{-1} \\ -\mu_1 \mu_2^{-1} \end{pmatrix} \qquad S = \nabla_t^T\, \mathbf{I}^{-1}\, \nabla_t = \frac{\mu_1(\mu_1 + \mu_2)}{\mu_2^3}$$

$$\eta = \frac{\mathbf{I}^{-1}\, \nabla_t}{(\nabla_t^T \mathbf{I}^{-1}\, \nabla_t)^{1/2}} = \begin{pmatrix} (\mu_1\mu_2)^{1/2}(\mu_1 + \mu_2)^{-1/2} \\ -(\mu_1\mu_2)^{1/2}(\mu_1 + \mu_2)^{-1/2} \end{pmatrix}$$

so that $\eta^T \mathbf{I} \eta = 1$. The probability matching prior for the parameter $t = \mu_1/\mu_2$ in terms of μ_1 and μ_2 is given by the equation:

$$\sum_{k=1}^{2} \frac{\partial}{\partial \mu_k} [\eta_k(\mu)\, \pi(\mu)] = 0$$

so, if $f(\mu_1, \mu_2) = (\mu_1\mu_2)^{1/2}(\mu_1 + \mu_2)^{-1/2}$, we have to solve

$$\frac{\partial}{\partial \mu_1} f(\mu_1, \mu_2)\, \pi(\mu_1, \mu_2) = \frac{\partial}{\partial \mu_2} f(\mu_1, \mu_2)\, \pi(\mu_1, \mu_2)$$

Any solution will do so:

$$\pi(\mu_1, \mu_2) \propto f^{-1}(\mu_1, \mu_2) = \frac{\sqrt{\mu_1 + \mu_2}}{\sqrt{\mu_1 \mu_2}}$$

Substituting $\mu_1 = t\mu_2$ and including the Jacobian $J = \mu_2$ we have finally:

$$\pi(t, \mu_2) \propto \sqrt{\mu_2}\sqrt{\frac{1+t}{t}}$$

The posterior density will be:

$$p(t, \mu_2|n_1, n_2) \propto p(n_1, n_2|t, \mu_2)\pi(t, \mu_2) \propto e^{-\mu_2(1+t)} t^{n_1-1/2} (1+t)^{1/2} \mu_2^{n+3/2-1}$$

and, integrating the nuisance parameter $\mu_2 \in [0, \infty)$, we get the posterior density:

$$p(t|n_1, n_2) = N \frac{t^{n_1-1/2}}{(1+t)^{n+1}}$$

with $N^{-1} = B(n_1 + 1/2, n_2 + 1/2)$.

Example 2.16 (*Gamma distribution*) Show that for $Ga(x|\alpha, \beta)$:

$$p(x|\alpha, \beta) = \frac{\alpha^\beta}{\Gamma(\beta)} e^{-\alpha x} x^{\beta-1} \mathbf{1}_{(0,\infty)}(x)$$

the probability matching prior for the ordering

- $\{\beta, \alpha\}$ is $\pi(\alpha, \beta) = \beta^{-1/2}\left[\alpha^{-1}\sqrt{\beta\Psi'(\beta) - 1}\right]$
- $\{\alpha, \beta\}$ is $\pi(\alpha, \beta) = \left[\alpha^{-1}\sqrt{\Psi'(\beta)}\right]\sqrt{\beta\Psi'(\beta) - 1}$

to be compared with Jeffrey's prior $\pi_2^J(\alpha, \beta) = \alpha^{-1}\sqrt{\beta\Psi'(\beta) - 1}$ and Jeffrey's prior when both parameters are treated individually $\pi_{1+1}^J(\alpha, \beta) = \alpha^{-1}\sqrt{\Psi'(\beta)}$

Example 2.17 (*Bivariate Normal Distribution*)

For the ordered parameterization ρ, σ_1, σ_2: the Fisher's matrix (see Example 2.12) is:

$$\boldsymbol{I}(\rho, \sigma_1, \sigma_2) = (1-\rho^2)^{-1} \begin{pmatrix} (1+\rho^2)(1-\rho^2)^{-1} & -\rho\sigma_1^{-1} & -\rho\sigma_2^{-1} \\ -\rho\sigma_1^{-1} & (2-\rho^2)\sigma_1^{-2} & -\rho^2(\sigma_1\sigma_2)^{-1} \\ -\rho\sigma_2^{-1} & -\rho^2(\sigma_1\sigma_2)^{-1} & (2-\rho^2)\sigma_2^{-2} \end{pmatrix}$$

and the inverse:

$$\boldsymbol{I}^{-1}(\rho, \sigma_1, \sigma_2) = \frac{1}{2} \begin{pmatrix} 2(1-\rho^2)^2 & \sigma_1\rho(1-\rho^2) & \sigma_2\rho(1-\rho^2) \\ \sigma_1\rho(1-\rho^2) & \sigma_1^2 & \rho^2\sigma_1\sigma_2 \\ \sigma_2\rho(1-\rho^2) & \rho^2\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

Then

$$\frac{2}{\rho}\frac{\partial}{\partial\rho}[\pi(1-\rho^2)] + \frac{\partial}{\partial\sigma_1}[\pi\sigma_1] + \frac{\partial}{\partial\sigma_2}[\pi\sigma_2] = 0$$

for which

$$\pi(\sigma_1, \sigma_2, \rho) = \frac{1}{\sigma_1 \sigma_2 (1 - \rho^2)}$$

is a solution.

Problem 2.6 Consider

$$X \sim p(x|a, b, \sigma) = \frac{\sinh[\sigma(b-a)]}{2(b-a)} \frac{1}{\cosh[\sigma(x-a)]\cosh[\sigma(b-x)]} \mathbf{1}_{(-\infty,\infty)}(x)$$

where $a < b \in \mathcal{R}$ and $\sigma \in (0, \infty)$. Show that

$$E[X] = \frac{b+a}{2} \qquad \text{and} \qquad V[x] = \frac{(b-a)^2}{12} + \frac{\pi^2}{12\sigma^2}$$

and that, for known $\sigma >>$, the probability matching prior for a and b tends to $\pi_{pm}(a, b) \sim (b-a)^{-1/2}$. Show also that, under the same limit, $\pi_{pm}(\theta) \sim \theta^{-1/2}$ for $(a, b) = (-\theta, \theta)$ and $(a, b) = (0, \theta)$. Since $p(x|a, b, \sigma) \overset{\sigma >>}{\to} Un(x|a, b)$ discuss in this last case what is the difference with the Example 2.4.

2.6.7 *Reference Analysis*

The expected amount of information (*Expected Mutual Information*) on the parameter θ provided by k independent observations of the model $p(\boldsymbol{x}|\theta)$ relative to the prior knowledge on θ described by $\pi(\theta)$ is

$$I\,[e(k), \pi(\theta)] = \int_\Theta \pi(\theta)\, d\theta \int_{\Omega_X} p(\boldsymbol{z}_k|\theta)\, \log \frac{p(\theta|\boldsymbol{z}_k)}{\pi(\theta)} d\boldsymbol{z}_k$$

where $\boldsymbol{z}_k = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$. If $\lim_{k\to\infty} I\,[e(k), \pi(\theta)]$ exists, it will quantify the maximum amount of information that we could obtain on θ from experiments described by this model relative to the prior knowledge $\pi(\theta)$. The central idea of the *reference analysis* [4, 17] is to take as *reference prior for the model* $p(\boldsymbol{x}|\theta)$ that which maximizes the maximum amount of information we may get so it will be the *less informative* for this model. From Calculus of Variations, if we introduce the prior $\pi^\star(\theta) = \pi(\theta) + \epsilon\eta(\theta)$ with $\pi(\theta)$ an extremal of the expected information $I\,[e(k), \pi(\theta)]$ and $\eta(\theta)$ such that

$$\int_\Theta \pi(\theta) d\theta = \int_\Theta \pi^\star(\theta) d\theta = 1 \quad \longrightarrow \quad \int_\Theta \eta(\theta) d\theta = 0$$

it is easy to see (left as exercise) that

$$\pi(\theta) \propto \exp\left\{\int_{\Omega_X} p(z_k|\theta)\ \log\ p(\theta|z_k)\, dz_k\right\} = f_k(\theta)$$

This is a nice but complicated implicit equation because, on the one hand, $f_k(\theta)$ depends on $\pi(\theta)$ through the posterior $p(\theta|z_k)$ and, on the other hand, the limit $k\to\infty$ is usually divergent (intuitively, the more precision we want for θ, the more information is needed and to know the actual value from the experiment requires an infinite amount of information). This can be circumvented regularizing the expression as

$$\pi(\theta) \propto \pi(\theta_0) \lim_{k\to\infty} \frac{f_k(\theta)}{f_k(\theta_0)}$$

with θ_0 any interior point of Θ (we are used to that in particle physics!). Let's see some examples.

Example 2.18 Consider again the exponential model for which $t = n^{-1}\sum_{i=1}^n x_i$ is sufficient for θ and distributed as

$$p(t|\theta) = \frac{(n\theta)^n}{\Gamma(n)}\, t^{n-1} \exp\{-n\theta t\}$$

Taking $\pi(\theta) = \mathbf{1}_{(0,\infty)}(\theta)$ we have the proper posterior

$$\pi^\star(\theta|t) = \frac{(nt)^{n+1}}{\Gamma(n+1)}\ \exp\{-n\theta t\}\ \theta^n$$

Then $\log \pi^\star(\theta|t) = -(n\theta)t + n\ \log\theta + (n+1)\log t + g_1(n)$ and

$$f_n(\theta) = \exp\left\{\int_{\Omega_X} p(t|\theta)\ \log\ \pi^\star(\theta|t)\, dt\right\} = \frac{g_2(n)}{\theta} \quad\longrightarrow\quad \pi(\theta) \propto \pi(\theta_0) \lim_{n\to\infty} \frac{f_n(\theta)}{f_n(\theta_0)} \propto \frac{1}{\theta}$$

Example 2.19 Prior functions depend on the particular model we are treating. To learn about a parameter, we can do different experimental designs that respond to different models and, even though the parameter is the same, they may have different priors. For instance, we may be interested in the *acceptance*; the probability to accept an event under some conditions. For this, we can generate for instance a sample of N observed events and see how many (x) pass the conditions. This experimental design corresponds to a Binomial distribution

$$p(x|N,\theta) = \binom{N}{x}\ \theta^x\ (1-\theta)^{N-x}$$

with $x = \{0, 1, \ldots, N\}$. For this model, the reference prior (also Jeffrey's and PM) is $\pi(\theta) = \theta^{-1/2}(1-\theta)^{-1/2}$ and the posterior $\theta \sim Be(\theta|x + 1/2, N - x + 1/2)$. Conversely, we can generate events until r are accepted and see how many (x) have we generated. This experimental design corresponds to a Negative Binomial distribution

$$p(x|r,\theta) = \binom{x-1}{r-1} \theta^r \, (1-\theta)^{x-r}$$

where $x = r, r+1, \ldots$ and $r \geq 1$. For this model, the reference prior (Jeffrey's and PM too) is $\pi(\theta) = \theta^{-1}(1-\theta)^{-1/2}$ and the posterior $\theta \sim Be(\theta|r, x - r + 1/2)$.

Problem 2.7 Consider

(1) $X \sim Po(x|\theta) = \exp\{-\theta\}\frac{\theta^x}{\Gamma(x+1)}$ and the experiment $e(k) \stackrel{iid}{\to} \{x_1, x_2, \ldots x_k\}$. Take $\pi^\star(\theta) = \mathbf{1}_{(0,\infty)}(\theta)$, and show that

$$\pi(\theta) \propto \pi(\theta_0) \lim_{k\to\infty} \frac{f_k(\theta)}{f_k(\theta_0)} \propto \theta^{-1/2}$$

(2) $X \sim Bi(x|N,\theta) = \binom{N}{x} \theta^x \, (1-\theta)^{N-x}$ and the experiment $e(k) \stackrel{iid}{\to} \{x_1, x_2, \ldots x_k\}$. Take $\pi^\star(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}\mathbf{1}_{(0,1)}(\theta)$ with $a, b > 0$ and show that

$$\pi(\theta) \propto \pi(\theta_0) \lim_{k\to\infty} \frac{f_k(\theta)}{f_k(\theta_0)} \propto \theta^{-1/2}(1-\theta)^{-1/2}$$

(Hint: For (1) and (2) consider the Taylor expansion of $\log\Gamma(z,\cdot)$ around $E[z]$ and the asymptotic behavior of the Polygamma Function $\Psi^{n)}(z) = a_n z^{-n} + a_{n+1} z^{-(n+1)} + \cdots$).

(3) $X \sim Un(x|0,\theta)$ and the iid sample $\{x_1, x_2, \ldots x_k\}$. For inferences on θ, show that $f_k = \theta^{-1} g(k)$ and in consequence the posterior is Pareto $Pa(\theta|x_M, n)$ with $x_M = \max\{x_1, x_2, \ldots x_k\}$ the sufficient statistic.

A very useful constructive theorem to obtain the *reference prior* is given in [18]. First, a *permissible prior* for the model $p(\boldsymbol{x}|\theta)$ is defined as a strictly positive function $\pi(\theta)$ such that it renders a proper posterior; that is,

$$\forall x \in \Omega_X \qquad \int_\Theta p(\boldsymbol{x}|\theta)\, \pi(\theta)\, d\theta \; < \; \infty$$

and that for some approximating sequence $\Theta_k \subset \Theta$; $\lim_{k\to\infty} \Theta_k = \Theta$, the sequence of posteriors $p_k(\theta|\boldsymbol{x}) \propto p(\boldsymbol{x}|\theta)\pi_k(\theta)$ converges logarithmically to $p(\theta|\boldsymbol{x}) \propto p(\boldsymbol{x}|\theta)\pi(\theta)$. Then, the *reference prior* is just a *permissible prior* that maximizes

the maximum amount of information the experiment can provide for the parameter. The constructive procedure for a one-dimensional parameter consists on:

(1) Take $\pi^\star(\theta)$ as a continuous strictly positive function such that the corresponding posterior

$$\pi^\star(\theta|\boldsymbol{z}_k) = \frac{p(\boldsymbol{z}_k|\theta)\,\pi^\star(\theta)}{\int_\Theta p(\boldsymbol{z}_k|\theta)\,\pi(\theta)\,d\theta}$$

is proper and asymptotically consistent. $\pi^\star(\theta)$ is arbitrary so it can be taken for convenience to simplify the integrals.

(2) Obtain

$$f_k^\star(\theta) = \exp\left\{\int_{\Omega_X} p(\boldsymbol{z}_k|\theta)\,\log\,\pi^\star(\theta|\boldsymbol{z}_k)\,d\boldsymbol{z}_k\right\} \qquad \text{and} \qquad h_k(\theta;\theta_0) = \frac{f_k^\star(\theta)}{f_k^\star(\theta_0)}$$

for any interior point $\theta_0 \in \Theta$;

(3) If

(3.1) each $f_k^\star(\theta)$ is continuous;

(3.2) for any fixed θ and large k, is $h_k(\theta;\theta_0)$ is either monotonic in k or bounded from above by $h(\theta)$ that is integrable on any compact set;

(3.3) $\pi(\theta) = \lim_{k\to\infty} h_k(\theta;\theta_0)$ is a *permissible prior function*

then $\pi(\theta)$ is a reference prior for the model $p(\boldsymbol{x}|\theta)$. It is important to note that there is no requirement on the existence of the Fisher's information $\boldsymbol{I}(\theta)$. If it exists, a simple Taylor expansion of the densities shows that for a one-dimensional parameter $\pi(\theta) = [\boldsymbol{I}(\theta)]^{1/2}$ in consistency with Jeffrey's proposal. Usually, the last is easier to evaluate but not always as we shall see.

In many cases supp(θ) is unbounded and the prior $\pi(\theta)$ is not a propper density. As we have seen this is not a problem as long as the posterior $p(\theta|\boldsymbol{z}_k) \propto p(\boldsymbol{z}_k|\theta)\pi(\theta)$ is propper although, in any case, one can proceed *"more formally"* considering a sequence of proper priors $\pi_m(\theta)$ defined on a sequence of compact sets $\Theta_m \subset \Theta$ such that $\lim_{m\to\infty}\Theta_m = \Theta$ and taking the limit of the corresponding sequence of posteriors $p_m(\theta|\boldsymbol{z}_k) \propto p(\boldsymbol{z}_k|\theta)\pi_m(\theta)$. Usually simple sequences as for example $\Theta_m = [1/m, m]$; $\lim_{m\to\infty}\Theta_m = (0,\infty)$, or $\Theta_m = [-m, m]$; $\lim_{m\to\infty}\Theta_m = (-\infty,\infty)$ will suffice.

When the parameter $\boldsymbol{\theta}$ is n-dimensional, the procedure is more laborious. First, one starts [4] arranging the parameters in decreasing order of importance $\{\theta_1, \theta_2, \ldots, \theta_n\}$ (as we did for the Probability Matching Priors) and then follow the previous scheme to obtain the conditional prior functions

$$\pi(\theta_n|\theta_1,\theta_2,\ldots,\theta_{n-1})\,\pi(\theta_{n-1}|\theta_1,\theta_2,\ldots,\theta_{n-2})\,\ldots\,\pi(\theta_2|\theta_1)\,\pi(\theta_1)$$

For instance in the case of two parameters and the ordered parameterization $\{\theta, \lambda\}$:

(1) Get the conditional $\pi(\lambda|\theta)$ as the reference prior for λ keeping θ fixed;

(2) Find the marginal model

$$p(\boldsymbol{x}|\theta) = \int_{\Phi} p(\boldsymbol{x}|\theta, \lambda)\, \pi(\lambda|\theta)\, d\lambda$$

(3) Get the reference prior $\pi(\theta)$ from the marginal model $p(\boldsymbol{x}|\theta)$

Then $\pi(\theta, \lambda) \propto \pi(\lambda|\theta)\pi(\theta)$. This is fine if $\pi(\lambda|\theta)$ and $\pi(\theta)$ are propper functions; seldom the case. Otherwise one has to define the appropriate sequence of compact sets observing, among other things, that this has to be done for the full parameter space and usually the limits depend on the parameters. Suppose that we have the sequence $\Theta_i \times \Lambda_i \xrightarrow{i\to\infty} \Theta \times \Lambda$. Then:

(1) Obtain $\pi_i(\lambda|\theta)$:

$$\pi_i^\star(\lambda|\theta)\mathbf{1}_{\Lambda_i}(\lambda) \longrightarrow \pi_i^\star(\lambda|\theta, z_k) = \frac{p(z_k|\theta, \lambda)\pi_i^\star(\lambda|\theta)}{\int_{\Lambda_i} p(z_k|\theta, \lambda)\pi_i^\star(\lambda|\theta)\, d\lambda} \longrightarrow \pi_i(\lambda|\theta) = \lim_{k\to\infty} \frac{f_k^\star(\lambda|\Lambda_i, \theta, \ldots)}{f_k^\star(\lambda_0|\Lambda_i, \theta, \ldots)}$$

(2) Get the marginal density $p_i(\boldsymbol{x}|\theta)$:

$$p_i(\boldsymbol{x}|\theta) = \int_{\Lambda_i} p(\boldsymbol{x}|\theta, \lambda)\, \pi_i(\lambda|\theta)\, d\lambda$$

(3) Determine $\pi_i(\theta)$:

$$\pi_i^\star(\theta)\mathbf{1}_{\Theta_i}(\theta) \longrightarrow \pi_i^\star(\theta|z_k) = \frac{p_i(z_k|\theta)\pi_i^\star(\theta)}{\int_{\Theta_i} p_i(z_k|\theta)\pi_i^\star(\theta)\, d\theta} \longrightarrow \pi_i(\theta) = \lim_{k\to\infty} \frac{f_k^\star(\theta|\Theta_i, \Lambda_i, \ldots)}{f_k^\star(\theta_0|\Theta_i, \Lambda_i, \ldots)}$$

(4) The reference prior for the ordered parameterization $\{\theta, \lambda\}$ will be:

$$\pi(\theta, \lambda) = \lim_{i\to\infty} \frac{\pi_i(\lambda|\theta)\, \pi_i(\theta)}{\pi_i(\lambda_0|\theta_0)\, \pi_i(\theta_0)}$$

In the case of two parameters, if Λ is independent of θ the Fisher's matrix usually exists and, if $\boldsymbol{I}(\theta, \lambda)$ and $\boldsymbol{S}(\theta, \lambda) = \boldsymbol{I}^{-1}(\theta, \lambda)$ are such that:

$$\boldsymbol{I}_{22}(\theta, \lambda) = a_1^2(\theta)\, b_1^2(\lambda) \qquad \text{and} \qquad \boldsymbol{S}_{11}(\theta, \lambda) = a_0^{-2}(\theta)\, b_0^{-2}(\lambda)$$

then [19] $\pi(\theta, \lambda) = \pi(\lambda|\theta)\pi(\theta) = a_0(\theta)\, b_1(\lambda)$ is a permissible prior even if the conditional reference priors are not proper. The reference priors are usually *probability matching priors.*

Example 2.20 A simple example is the Multinomial distribution $\boldsymbol{X} \sim Mn(\boldsymbol{x}|\boldsymbol{\theta})$ with $\dim \boldsymbol{X} = k + 1$ and probability

$$p(\boldsymbol{x}|\theta) \propto \theta_1^{x_1}\,\theta_2^{x_2} \ldots \theta_k^{x_k}\,(1-\delta_k)^{x_{k+1}}; \qquad \delta_k = \sum_{j=1}^{k} \theta_j$$

Consider the ordered parameterization $\{\theta_1, \theta_2, \ldots, \theta_k\}$. Then

$$\pi(\theta_1, \theta_2, \ldots, \theta_k) = \pi(\theta_k|\theta_{k-1}, \theta_{k-2} \ldots \theta_2, \theta_1)\,\pi(\theta_{k-1}|\theta_{k-2} \ldots \theta_2, \theta_1) \ldots \pi(\theta_2|\theta_1)\,\pi(\theta_1)$$

In this case, all the conditional densities are proper

$$\pi(\theta_m|\theta_{m-1}, \ldots\, \theta_1) \propto \theta_m^{-1/2}\,(1-\delta_m)^{-1/2}$$

and therefore

$$\pi(\theta_1, \theta_2, \ldots\, \theta_k) \propto \prod_{i=1}^{k} \theta_i^{-1/2}\,(1-\delta_i)^{-1/2}$$

The posterior density will be then

$$p(\boldsymbol{\theta}|\boldsymbol{x}) \propto \left[\prod_{i=1}^{k} \theta_i^{x_i - 1/2}\,(1-\delta_i)^{-1/2}\right] (1-\delta_k)^{x_{k+1}}$$

Example 2.21 Consider again the case of two independent Poisson distributed random quantities X_1 and X_2 with joint density

$$P(n_1, n_2|\mu_1, \mu_2) = P(n_1|\mu_1)\,P(n_2|\mu_2) = e^{-(\mu_1+\mu_2)}\,\frac{\mu_1^{n_1}\,\mu_2^{n_2}}{\Gamma(n_1+1)\Gamma(n_2+1)}$$

We are interested in the parameter $\theta = \mu_1/\mu_2$ so setting $\mu = \mu_2$ we have the ordered parameterization $\{\theta, \mu\}$ and:

$$P(n_1, n_2|\theta, \mu) = e^{-\mu(1+\theta)}\,\frac{\theta^{n_1}\mu^{n}}{\Gamma(n_1+1)\Gamma(n_2+1)}$$

where $n = n_1 + n_2$. Since $E[X_1] = \mu_1 = \theta\mu$ and $E[X_2] = \mu_2 = \mu$ the Fisher's matrix and its inverse will be

$$\mathbf{I} = \begin{pmatrix} \mu/\theta & 1 \\ 1 & (1+\theta)/\mu \end{pmatrix}; \qquad \det(\mathbf{I}) = \theta^{-1} \qquad \text{and} \qquad \mathbf{S} = \mathbf{I}^{-1} = \begin{pmatrix} \theta(1+\theta)/\mu & -\theta \\ -\theta & \mu \end{pmatrix}$$

Therefore

$$S_{11} = \theta(1+\theta)/\mu \qquad \text{and} \qquad F_{22} = (1+\theta)/\mu$$

and, in consequence:

$$\pi(\theta)\, f_1(\mu) \propto S_{11}^{-1/2} = \frac{\sqrt{\mu}}{\sqrt{\theta(1+\theta)}} \qquad \pi(\mu|\theta)\, f_2(\theta) \propto F_{22}^{1/2} = \frac{\sqrt{1+\theta}}{\sqrt{\mu}}$$

Thus, we have for the ordered parameterization $\{\theta, \mu\}$ the reference prior:

$$\pi(\theta, \mu) = \pi(\mu|\theta)\, \pi(\theta) \propto \frac{1}{\sqrt{\mu\theta(1+\theta)}}$$

and the posterior density will be:

$$p(\theta, \mu|n_1, n_2) \;\propto\; \exp\{-\mu(1+\theta)\}\; \theta^{n_1-1/2}\, (1+\theta)^{-1/2}\, \mu^{n-1/2}$$

and, integrating the nuisance parameter $\mu \in [0, \infty)$ we get finally

$$p(\theta|n_1, n_2) = N\, \frac{\theta^{n_1-1/2}}{(1+\theta)^{n+1}}$$

with $\theta = \mu_1/\mu_2$, $n = n_1 + n_2$ and $N^{-1} = B(n_1 + 1/2, n_2 + 1/2)$. The distribution function will be:

$$P(\theta|n_1, n_2) = \int_0^\theta p(\theta'|n_1, n_2)\, d\theta' = \; I(\theta/(1+\theta); n_1 + 1/2, n_2 + 1/2)$$

with $I(x; a, b)$ the Incomplete Beta Function and the moments, when they exist;

$$E[\theta^m] = \frac{\Gamma(n_1 + 1/2 + m)\, \Gamma(n_2 + 1/2 - m)}{\Gamma(n_1 + 1/2)\, \Gamma(n_2 + 1/2)}$$

It is interesting to look at the problem from a different point of view. Consider again the ordered parameterization $\{\theta, \lambda\}$ with $\theta = \mu_1/\mu_2$ but now, the nuisance parameter is $\lambda = \mu_1 + \mu_2$. The likelihood will be:

$$P(n_1, n_2|\theta, \lambda) = \frac{1}{\Gamma(n_1+1)\Gamma(n_2+1)} e^{-\lambda}\, \lambda^n\, \frac{\theta^{n_1}}{(1+\theta)^n}$$

The domains are $\Theta = (0, \infty)$ and $\Lambda = (0, \infty)$, independent. Thus, no need to specify the prior for λ since

$$p(\theta|n_1, n_2) \propto \pi(\theta)\, \frac{\theta^{n_1}}{(1+\theta)^n} \int_\Lambda e^{-\lambda}\, \lambda^n \pi(\lambda) d\lambda \propto \frac{\theta^{n_1}}{(1+\theta)^n}\, \pi(\theta)$$

In this case we have that

$$I(\theta) \propto \frac{1}{\theta\,(1+\theta)^2} \quad \longrightarrow \quad \pi(\theta) = \frac{1}{\theta^{1/2}\,(1+\theta)}$$

and, in consequence,

$$p(\theta|n_1, n_2) = N\,\frac{\theta^{n_1-1/2}}{(1+\theta)^{n+1}}$$

Problem 2.8 Show that the reference prior for the Pareto distribution $Pa(x|\theta, x_0)$ (see Example 2.9) is $\pi(\theta, x_0) \propto (\theta x_0)^{-1}$ and that for an iid sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, if $x_m = \min\{x_i\}_{i=1}^n$ and $a = \sum_{i=1}^n \ln(x_i/x_m)$ the posterior

$$p(\theta, x_0|\boldsymbol{x}) = \frac{n a^{n-1}}{x_m \Gamma(n-1)}\, e^{-a\theta}\theta^{n-1}\left(\frac{x_0}{x_m}\right)^{n\theta-1} \mathbf{1}_{(0,\infty)}(\theta)\mathbf{1}_{(0,x_m)}(x_0)$$

is proper for a sample size $n > 1$. Obtain the marginal densities

$$p(\theta|\boldsymbol{x}) = \frac{a^{n-1}}{\Gamma(n-1)}\, e^{-a\theta}\theta^{n-2}\mathbf{1}_{(0,\infty)}(\theta) \qquad \text{and}$$
$$p(x_0|\boldsymbol{x}) = \frac{n(n-1)}{a}\, x_0^{-1}\left[1 + \frac{n}{a}\ln\left(\frac{x_m}{x_0}\right)\right]^{-n} \mathbf{1}_{(0,x_m)}(x_0)$$

and show that for large n (see Sect. 2.10.2) $E[\theta] \simeq n a^{-1}$ and $E[x_0] \simeq x_m$.

Problem 2.9 Show that for the shifted Pareto distribution (Lomax distribution):

$$p(x|\theta, x_0) = \frac{\theta}{x_0}\left(\frac{x_0}{x + x_0}\right)^{\theta+1} \mathbf{1}_{(0,\infty)}(x); \qquad \theta, x_0 \in R^+$$

the reference prior for the ordered parameterization $\{\theta, x_0\}$ is $\pi_r(\theta, x_0) \propto (x_0\theta(\theta+1))^{-1}$ and for $\{x_0, \theta\}$ is $\pi_r(x_0, \theta) \propto (x_0\theta)^{-1}$. Show that the first one is a first order probability matching prior while the second is not. In fact, show that for $\{x_0, \theta\}$, $\pi_{pm}(x_0, \theta) \propto (x_0\theta^{3/2}\sqrt{\theta+2})^{-1}$ is a matching prior and that for both orderings the Jeffrey's prior is $\pi_J(\theta, x_0) \propto (x_0(\theta+1)\sqrt{\theta(\theta+2)})^{-1}$.

Problem 2.10 Show that for the Weibull distribution

$$p(x|\alpha, \beta) = \alpha\beta\, x^{\beta-1} \exp\left\{-\alpha x^{\beta}\right\} \mathbf{1}_{(0,\infty)}(x)$$

with $\alpha, \beta > 0$, the reference prior functions are

$$\pi_r(\beta, \alpha) = (\alpha\beta)^{-1} \qquad \text{and} \qquad \pi_r(\alpha, \beta) = \left(\alpha\beta\sqrt{\zeta(2) + (\psi(2) - \ln\alpha)^2}\right)^{-1}$$

for the ordered parameterizations $\{\beta, \alpha\}$ and $\{\alpha, \beta\}$ respectively being $\zeta(2) = \pi^2/6$ the Riemann Zeta Function and $\psi(2) = 1 - \gamma$ the Digamma Function.

2.7 Hierarchical Structures

In many circumstances, even though the experimental observations respond to the same phenomena it is not always possible to consider the full set of observations as an exchangeable sequence but rather exchangeability within subgroups of observations. As stated earlier, this may be the case when the results come from different experiments or when, within the same experiment, data taking conditions (acceptances, efficiencies,…) change from run to run. A similar situation holds, for instance, for the results of responses under a drug performed at different hospitals when the underlying conditions of the population vary between zones, countries,… In general, we shall have different groups of observations

$$\begin{aligned}
\boldsymbol{x}_1 &= \{x_{11}, x_{21}, \ldots, x_{n_1 1}\} \\
&\vdots \\
\boldsymbol{x}_j &= \{x_{1j}, x_{2j}, \ldots, x_{n_j j}\} \\
&\vdots \\
\boldsymbol{x}_J &= \{x_{1J}, x_{2J}, \ldots, x_{n_J J}\}
\end{aligned}$$

from J experiments $e_1(n_1), e_2(n_2), \ldots, e_J(n_J)$. Within each sample $\boldsymbol{x}_j$, we can consider that exchangeability holds and also for the sets of observations $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_J\}$ In this case, it is appropriate to consider *hierarchical structures.*

Let's suppose that for each experiment $e(j)$ the observations are drawn from the model

$$p(\boldsymbol{x}_j|\boldsymbol{\theta}_j); \qquad j = 1,\ 2,\ \ldots,\ J$$

Since the experiments are independent we assume that the parameters of the sequence $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_J\}$ are exchangeable and that, although different, they can be assumed to have a common origin since they respond to the same phenomena. Thus,we can set

$$p(\boldsymbol{\theta}_1,\ \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_J|\phi) = \prod_{i=1}^{J} p(\boldsymbol{\theta}_i|\phi)$$

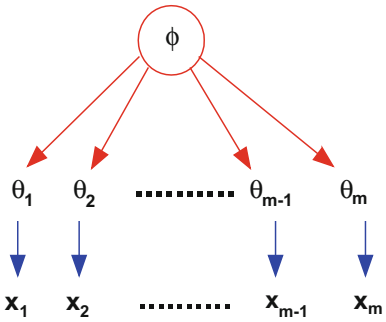

Fig. 2.3 Structure of the hierarchical model

with ϕ the *hyperparameters* for which we take a prior $\pi(\phi)$. Then we have the structure (Fig. 2.3.)

$$p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_J, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J, \phi) = \pi(\phi) \prod_{i=1}^{J} p(\boldsymbol{x}_i|\boldsymbol{\theta}_i)\, \pi(\boldsymbol{\theta}_i|\phi)$$

This structure can be repeated sequentially if we consider appropriate to assign a prior $\pi(\phi|\boldsymbol{\tau})$ to the hyperparameters ϕ so that

$$p(\boldsymbol{x}, \boldsymbol{\theta}, \phi, \boldsymbol{\tau}) = p(\boldsymbol{x}|\boldsymbol{\theta})\, \pi(\boldsymbol{\theta}|\phi)\, \pi(\phi|\boldsymbol{\tau})\, \pi(\boldsymbol{\tau})$$

Now, consider the model $p(\boldsymbol{x}, \boldsymbol{\theta}, \phi)$. We may be interested in $\boldsymbol{\theta}$, in the hyperparameters ϕ or in both. In general we shall need the conditional densities:

- $p(\phi|\boldsymbol{x}) \propto p(\phi) \int p(\boldsymbol{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\phi)\, d\boldsymbol{\theta}$
- $p(\boldsymbol{\theta}|\boldsymbol{x}, \phi) = \dfrac{p(\boldsymbol{\theta}, \boldsymbol{x}, \phi)}{p(\boldsymbol{x}, \phi)}$ and
- $p(\boldsymbol{\theta}|\boldsymbol{x}) = \dfrac{p(\boldsymbol{x}|\boldsymbol{\theta})}{p(\boldsymbol{x})}\, p(\boldsymbol{\theta}) = \dfrac{p(\boldsymbol{x}|\boldsymbol{\theta})}{p(\boldsymbol{x})} \int p(\boldsymbol{\theta}|\phi)\, p(\phi)\, d\phi$

that can be expressed as

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \int \frac{p(\boldsymbol{x}, \boldsymbol{\theta}, \phi)}{p(\boldsymbol{x})}\, d\phi = \int p(\boldsymbol{\theta}|\boldsymbol{x}, \phi)\, p(\phi|\boldsymbol{x})\, d\phi$$

and, since

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{\theta}) \int p(\boldsymbol{\theta}|\phi)\, \frac{p(\phi|\boldsymbol{x})}{p(\boldsymbol{x}|\phi)}\, d\phi$$

we can finally write

$$\frac{p(\boldsymbol{\theta}, \phi)}{p(\boldsymbol{x})} = p(\boldsymbol{\theta}|\phi)\,\frac{p(\phi)}{p(\boldsymbol{x})} = p(\boldsymbol{\theta}|\phi)\,\frac{p(\phi|\boldsymbol{x})}{p(\boldsymbol{x}|\phi)}$$

In general, this conditional densities have complicated expressions and we shall use Monte Carlo methods to proceed (see Gibbs Sampling, Example 3.15, in Chap. 3).

It is important to note that if the prior distributions are not proper we can have improper marginal and posterior densities that obviously have no meaning in the inferential process. Usually, conditional densities are better behaved but, in any case, we have to check that this is so. In general, the better behaved is the likelihood the wildest behavior we can accept for the prior functions. We can also used prior distributions that are a mixture of proper distributions:

$$p(\boldsymbol{\theta}|\phi) = \sum_i w_i\; p_i(\boldsymbol{\theta}|\phi)$$

with $w_i \geq 0$ and $\sum w_i = 1$ so that the combination is convex and we assure that it is proper density or, extending this to a continuous mixture:

$$p(\boldsymbol{\theta}|\phi) = \int w(\boldsymbol{\sigma})\; p(\boldsymbol{\theta}|\phi, \boldsymbol{\sigma})\, d\boldsymbol{\sigma}.$$

2.8 Priors for Discrete Parameters

So far we have discussed parameters with continuous support but in some cases it is either finite or countable. If the parameter of interest can take only a finite set of n possible values, the reasonable option for an *uninformative prior* is a Discrete Uniform Probability $P(X = x_i) = 1/n$. In fact, it is shown in Sect. 4.2 that maximizing the expected information provided by the experiment with the normalization constraint (i.e. the probability distribution for which the *prior* knowledge is minimal) drives to $P(X = x_i) = 1/n$ in accordance with the *Principle of Insufficient Reason*.

Even though finite discrete parameter spaces are either the most usual case we shall have to deal with or, at least, a sufficiently good approximation for the real situation, it may happen that a non-informative prior is not the most appropriate (see Example 2.22). On the other hand, if the parameter takes values on a countable set the problem is more involved. A possible way out is to devise a hierarchical structure in which we assign the discrete parameter θ a prior $\pi(\theta|\boldsymbol{\lambda})$ with $\boldsymbol{\lambda}$ a set of continuous hyperparameters. Then, since

$$p(\boldsymbol{x}, \boldsymbol{\lambda}) = \sum_{\theta \in \Theta} p(\boldsymbol{x}|\theta)\pi(\theta|\boldsymbol{\lambda})\,\pi(\boldsymbol{\lambda}) = p(\boldsymbol{x}|\boldsymbol{\lambda})\,\pi(\boldsymbol{\lambda})$$

we get the prior $\pi(\boldsymbol{\lambda})$ by any of the previous procedures for continuous parameters with the model $p(\boldsymbol{x}|\boldsymbol{\lambda})$ and obtain

$$\pi(\boldsymbol{\theta}) \propto \int_{\Lambda} \pi(\theta|\boldsymbol{\lambda})\, \pi(\boldsymbol{\lambda})\, d\boldsymbol{\lambda}$$

Different procedures are presented and discussed in [20].

Example 2.22 The absolute value of the electric charge (Z) of a particle is to be determined from the number of photons observed by a Cherenkov Counter. We know from test beam studies and Monte Carlo simulations that the number of observed photons n_γ produced by a particle of charge Z is well described by a Poisson distribution with parameter $\mu = n_0 Z^2$; that is

$$P(n_\gamma|n_0, Z) = e^{-n_0 Z^2}\, \frac{(n_0 Z^2)^{n_\gamma}}{\Gamma(n_\gamma + 1)}$$

so $E[n_\gamma|Z = 1] = n_0$. First, by physics considerations Z has a finite support $\Omega_Z = \{1, 2, \ldots, n\}$. Second, we know *a priory* that not all incoming nuclei are equally likely so a *non-informative* prior may not be the best choice. In any case, a discrete uniform prior will give the posterior:

$$P(Z = k|n_\gamma, n_0, n) = \frac{e^{-n_0 k^2}\, k^{2n_\gamma}}{\sum_{k=1}^{n} e^{-n_0 k^2}\, k^{2n_\gamma}}.$$

2.9 Constrains on Parameters and Priors

Consider a parametric model $p(\boldsymbol{x}|\boldsymbol{\theta})$ and the prior $\pi_0(\boldsymbol{\theta})$. Now we have some information on the parameters that we want to include in the prior. Typically we shall have say k constraints of the form

$$\int_{\Theta} g_i(\boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} = a_i; \qquad i = 1, \ldots, k$$

Then, we have to find the prior $\pi(\boldsymbol{\theta})$ for which $\pi_0(\boldsymbol{\theta})$ is the best approximation, in the Kullback-Leibler sense, including the constraints with the corresponding Lagrange multipliers λ_i; that is, the extremal of

$$\mathcal{F} = \int_{\Theta} \pi(\boldsymbol{\theta})\, \log \frac{\pi(\boldsymbol{\theta})}{\pi_0(\boldsymbol{\theta})}\, d\boldsymbol{\theta} + \sum_{i=1}^{k} \lambda_i \left(\int_{\Theta} g_i(\boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} - a_i \right)$$

Again, it is left as an exercise to show that from Calculus of Variations we have the well known solution

$$\pi(\boldsymbol{\theta}) \propto \pi_0(\boldsymbol{\theta})\, \exp\left\{ \sum_{i=1}^{k} \lambda_i\, g_i(\boldsymbol{\theta}) \right\} \qquad \text{where} \qquad \lambda_i \mid \int_{\Theta} g_i(\boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} = a_i$$

Quite frequently we are forced to include constraints on the support of the parameters: some are non-negative (masses, energies, momentum, life-times,…), some are bounded in (0, 1) ($\beta = v/c$, efficiencies, acceptances,…),... At least from a formal point of view, to account for constraints on the support is a trivial problem. Consider the model $p(\boldsymbol{x}|\boldsymbol{\theta})$ with $\boldsymbol{\theta} \in \Theta_0$ and a reference prior $\pi_0(\boldsymbol{\theta})$. Then, our inferences on $\boldsymbol{\theta}$ shall be based on the posterior

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})\,\pi_0(\boldsymbol{\theta})}{\int_{\Theta_0} p(\boldsymbol{x}|\boldsymbol{\theta})\,\pi_0(\boldsymbol{\theta})\,d\boldsymbol{\theta}}$$

Now, if we require that $\boldsymbol{\theta} \in \Theta \subset \Theta_0$ we define

$$g_1(\boldsymbol{\theta}) = \mathbf{1}_{\Theta}(\boldsymbol{\theta}) \longrightarrow \int_{\Theta_0} g_1(\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta} = \int_{\Theta} \pi(\boldsymbol{\theta})\,d\boldsymbol{\theta} = 1 - \epsilon$$
$$g_2(\boldsymbol{\theta}) = \mathbf{1}_{\Theta^c}(\boldsymbol{\theta}) \longrightarrow \int_{\Theta_0} g_2(\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta} = \int_{\Theta^c} \pi(\boldsymbol{\theta})\,d\boldsymbol{\theta} = \epsilon$$

and in the limit $\epsilon \to 0$ we have the *restricted reference prior*

$$\pi(\boldsymbol{\theta}) = \frac{\pi_0(\boldsymbol{\theta})}{\int_{\Theta} \pi_0(\boldsymbol{\theta})\,d\boldsymbol{\theta}}\,\mathbf{1}_{\Theta}(\boldsymbol{\theta})$$

as we have obviously expected. Therefore

$$p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\theta} \in \Theta) = \frac{p(x|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})}{\int_{\Theta} p(x|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}} = \frac{p(x|\theta)\,\pi_0(\boldsymbol{\theta})}{\int_{\Theta} p(x|\boldsymbol{\theta})\,\pi_0(\boldsymbol{\theta})\,d\boldsymbol{\theta}}\,\mathbf{1}_{\Theta}(\boldsymbol{\theta})$$

that is, the same initial expression but normalized in the domain of interest Θ.

2.10 Decision Problems

Even though all the information we have on the parameters of relevance is contained in the posterior density it is interesting, as we saw in Chap. 1, to explicit some particular values that characterize the probability distribution. This certainly entails a considerable and unnecessary reduction of the available information but in the end, quoting Lord Kelvin, *"… when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind"*. In statistics, to specify a particular value of the parameter is termed *Point Estimation* and can be formulated in the framework of *Decision Theory*.

In general, *Decision Theory* studies how to choose the *optimal action* among several possible alternatives based on what has been experimentally observed. Given a particular problem, we have to explicit the set Ω_θ of the possible *"states of nature"*,

the set Ω_X of the possible experimental outcomes and the set Ω_A of the possible actions we can take. Imagine, for instance, that we do a test on an individual suspected to have some disease for which the medical treatment has some potentially dangerous collateral effects. Then, we have:

$$\Omega_\theta = \{\text{healthy, sic}\}$$

$$\Omega_X = \{\text{test positive, test negative}\}$$

$$\Omega_A = \{\text{apply treatment, do not apply treatment}\}$$

Or, for instance, a detector that provides within some accuracy the momentum (p) and the velocity (β) of charged particles. If we want to assign an hypothesis for the mass of the particle we have that $\Omega_\theta = \mathcal{R}^+$ is the set of all possible states of nature (all possible values of the mass), Ω_X the set of experimental observations (the momentum and the velocity) and Ω_A the set of all possible actions that we can take (assign one or other value for the mass). In this case, we shall take a decision based on the probability density $p(m|p, \beta)$.

Obviously, unless we are in a state of absolute certainty we can not take an action without potential losses. Based on the observed experimental outcomes, we can for instance assign the particle a mass m_1 when the *true state of nature* is $m_2 \neq m_1$ or consider that the individual is healthy when is actually sic. Thus, the first element of Decision Theory is the *Loss Function*:

$$l(a, \theta) : (\theta, a) \in \Omega_\theta \times \Omega_A \longrightarrow \mathcal{R}^+ + \{0\}$$

This is a non-negative function, defined for all $\boldsymbol{\theta} \in \Omega_\theta$ and the set of possible actions $\boldsymbol{a} \in \Omega_A$, that quantifies the *loss* associated to take the action $\boldsymbol{a}$ (decide for $\boldsymbol{a}$) when the state of nature is $\boldsymbol{\theta}$.

Obviously, we do not have a perfect knowledge of the *state of nature*; what we know comes from the observed data $\boldsymbol{x}$ and is contained in the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x})$. Therefore, we define the *Risk Function* (*risk* associated to take the action $\boldsymbol{a}$, or decide for $\boldsymbol{a}$ when we have observed the data $\boldsymbol{x}$) as the expected value of the Loss Function:

$$R(\boldsymbol{a}|\boldsymbol{x}) = E_{\boldsymbol{\theta}}[l(\boldsymbol{a}, \boldsymbol{\theta})] = \int_{\Omega_\theta} l(\boldsymbol{a}, \boldsymbol{\theta})\, p(\boldsymbol{\theta}|\boldsymbol{x})\, d\boldsymbol{\theta}$$

Sound enough, the Bayesian decision criteria consists on taking the action $\boldsymbol{a}(\boldsymbol{x})$ (*Bayesian action*) that minimizes the risk $R(\boldsymbol{a}|\boldsymbol{x})$ (*minimum risk*); that is, that minimizes the expected loss under the posterior density function.[8] Then, we shall encounter to kinds of problems:

[8]The problems studied by *Decision Theory* can be addressed from the point of view of *Game Theory*. In this case, instead of *Loss Functions* one works with *Utility Functions* $u(\boldsymbol{\theta}, \boldsymbol{a})$ that, in essence, are nothing else but $u(\boldsymbol{\theta}, \boldsymbol{a}) = K - l(\boldsymbol{\theta}, \boldsymbol{a}) \geq 0$; it is just matter of personal optimism to

- *inferential problems*, where $\Omega_A = \mathcal{R}$ y $\boldsymbol{a}(\boldsymbol{x})$ is a statistic that we shall take as *estimator* of the parameter $\boldsymbol{\theta}$;
- *decision problems* (or *hypothesis testing*) where $\Omega_A = \{\text{accept, reject}\}$ or choose one among a set of hypothesis.

Obviously, the actions depend on the loss function (that we have to specify) and on the posterior density and, therefore, on the data through the model $p(\boldsymbol{x}|\boldsymbol{\theta})$ and the prior function $\pi(\boldsymbol{\theta})$. It is then possible that, for a particular model, two different loss functions drive to the same decision or that the same loss function, depending on the prior, take to different actions.

2.10.1 Hypothesis Testing

Consider the case where we have to choose between two exclusive and exhaustive hypothesis H_1 and $H_2(=H_1{}^c)$. From the data sample and our prior beliefs we have the posterior probabilities

$$P(H_i|\text{data}) = \frac{P(\text{data}|H_i)\, P(H_i)}{P(\text{data})}; \quad i = 1, 2$$

and the actions to be taken are then:

$$\begin{aligned} a_1 &: \text{action to take if we decide upon } H_1 \\ a_2 &: \text{action to take if we decide upon } H_2 \end{aligned}$$

Then, we define the loss function $l(a_i, H_j)$; $i, j = 1, 2$ as:

$$l(a_i|H_j) = \begin{cases} l_{11} = l_{22} = 0 & \text{if we make the correct choice; that is, if we take action } a_1 \text{ when the state of nature is } H_1 \text{ or } a_2 \text{ when it is } H_2; \\ l_{12} > 0 & \text{if we take action } a_1 \text{(decide upon } H_1\text{) when the state of nature is } H_2 \\ l_{21} > 0 & \text{if we take action } a_2 \text{ (decide upon } H_2\text{) when the state of nature is } H_1 \end{cases}$$

(Footnote 8 continued)
work with *"utilities"* or *"losses"*. J. Von Neumann and O. Morgenstern introduced in 1944 the idea of *expected utility* and the criteria to take as *optimal action* hat which maximizes the expected utility.

so the risk function will be:

$$R(a_i|\text{data}) = \sum_{j=1}^{2} l(a_i|H_j)\, P(H_j|\text{data})$$

that is:

$$\begin{aligned} R(a_1|\text{data}) &= l_{11}\, P(H_1|\text{data}) + l_{12}\, P(H_2|\text{data}) \\ R(a_2|\text{data}) &= l_{21}\, P(H_1|\text{data}) + l_{22}\, P(H_2|\text{data}) \end{aligned}$$

and, according to the minimum Bayesian risk, we shall choose the hypothesis H_1 (action a_1) if

$$R(a_1|\text{data}) < R(a_2|\text{data}) \longrightarrow P(H_1|\text{data})\,(l_{11} - l_{21}) < P(H_2|\text{data})\,(l_{22} - l_{12})$$

Since we have chosen $l_{11} = l_{22} = 0$ in this particular case, we shall take action a_1 (decide for hypothesis H_1) if:

$$\frac{P(H_1|\text{data})}{P(H_2|\text{data})} > \frac{l_{12}}{l_{21}}$$

or action a_2 (decide in favor of hypothesis H_2) if:

$$R(a_2, \text{data}) < R(a_1, \text{data}) \longrightarrow \frac{P(H_2|\text{data})}{P(H_1|\text{data})} > \frac{l_{21}}{l_{12}}$$

that is, we take action a_i $(i = 1, 2)$ if:

$$\frac{P(H_i|\text{data})}{P(H_j|\text{data})} = \left[\frac{P(\text{data}|H_i)}{P(\text{data}|H_j)}\right]\left[\frac{P(H_i)}{P(H_j)}\right] > \frac{l_{ij}}{l_{ji}}$$

The ratio of likelihoods

$$B_{ij} = \frac{P(\text{data}|H_i)}{P(\text{data}|H_j)}$$

is called **Bayes Factor** B_{ij} and changes our prior beliefs on the two alternative hypothesis based on the evidence we have from the data; that is, quantifies how strongly data favors one model over the other. Thus, we shall decide in favor of hypothesis H_i against H_j $(i, j = 1, 2)$ if

$$\frac{P(H_i|\text{data})}{P(H_j|\text{data})} > \frac{l_{ij}}{l_{ji}} \qquad \longrightarrow \qquad B_{ij} > \frac{P(H_j)}{P(H_i)}\,\frac{l_{ij}}{l_{ji}}$$

If we consider the same loss if we decide upon the wrong hypothesis whatever it be, we have $l_{12} = l_{21}$ (Zero-One Loss Function). In general, we shall be interested in testing:

(1) **Two simple hypothesis**, H_1 versus H_2, for which the models $M_i = \{X \sim p_i(x|\theta_i)\}$; $i = 1, 2$ are fully specified including the values of the parameters (that is, $\Theta_i = \{\theta_i\}$). In this case, the Bayes Factor will be given by the ratio of likelihoods

$$B_{12} = \frac{p_1(\boldsymbol{x}|\theta_1)}{p_2(\boldsymbol{x}|\theta_2)} \qquad \left(\text{usually } \frac{p(\boldsymbol{x}|\theta_1)}{p(\boldsymbol{x}|\theta_2)}\right)$$

The classical Bayes Factor is the ratio of the likelihoods for the two competing models evaluated at their respective maximums.

(2) **A simple** (H_1) **versus a composite hypothesis** H_2 for which the parameters of the model $M_2 = \{X \sim p_2(x|\theta_2)\}$ have support on Θ_2. Then we have to average the likelihood under H_2 and

$$B_{12} = \frac{p_1(\boldsymbol{x}|\theta_1)}{\int_{\Theta_2} p_2(\boldsymbol{x}|\theta)\pi_2(\theta)d\theta}$$

(3) **Two composite hypothesis**: in which the models M_1 and M_2 have parameters that are not specified by the hypothesis so

$$B_{12} = \frac{\int_{\Theta_1} p_1(\boldsymbol{x}|\theta_1)\pi_1(\theta_1)d\theta_1}{\int_{\Theta_2} p_2(\boldsymbol{x}|\theta_2)\pi_2(\theta_2)d\theta_2}$$

and, since $P(H_1|\text{data}) + P(H_2|\text{data}) = 1$, we can express the posterior probability $P(H_1|\text{data})$ as

$$P(H_1|\text{data}) = \frac{B_{12}\, P(H_1)}{P(H_2) + B_{12}\, P(H_1)}$$

Usually, we consider equal prior probabilities for the two hypothesis ($P(H_1) = P(H_2) = 1/2$) but be aware that in some cases this may not be a realistic assumption.

Bayes Factors are independent of the prior beliefs on the hypothesis ($P(H_i)$) but, when we have composite hypothesis, we average the likelihood with a prior and if it is an improper function they are not well defined. If we have prior knowledge about the parameters, we may take informative priors that are proper but this is not always the case. One possible way out is to consider sufficiently general proper priors (conjugated priors for instance) so the Bayes factors are well defined and then study what is the sensitivity for different reasonable values of the hyperparameters. A more practical and interesting approach to avoid the indeterminacy due to improper priors [21, 22] is to take a subset of the observed sample to render a proper posterior (with, for instance, reference priors) and use that as proper prior density to compute

the Bayes Factor with the remaining sample. Thus, if the sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ consists on iid observations, we may consider $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ and, with the reference prior $\pi(\boldsymbol{\theta})$, obtain the proper posterior

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}_1) = \frac{p(\boldsymbol{x}_1|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})}{\int_\Theta p(\boldsymbol{x}_1|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}}$$

The remaining subsample $(\boldsymbol{x}_2)$ is then used to compute the partial Bayes Factor[9]:

$$B_{12}(x_2|x_1) = \frac{\int_{\Theta_1} p_1(\boldsymbol{x}_2|\boldsymbol{\theta}_1)\,\pi_1(\boldsymbol{\theta}_1|\boldsymbol{x}_1)\,d\boldsymbol{\theta}_1}{\int_{\Theta_2} p_2(\boldsymbol{x}_2|\boldsymbol{\theta}_2)\,\pi_2(\boldsymbol{\theta}_2|\boldsymbol{x}_1)\,d\boldsymbol{\theta}_2} \qquad \left(= \frac{BF(\boldsymbol{x}_1, \boldsymbol{x}_2)}{BF(\boldsymbol{x}_1)}\right)$$

for the hypothesis testing. Berger and Pericchi propose to use the minimal amount of data needed to specify a proper prior (usually $\max\{\dim(\boldsymbol{\theta}_i)\}$) so as to leave most of the sample for the model testing and dilute the dependence on a particular election of the training sample evaluating the Bayes Factors with all possible minimal samples and choosing the truncated mean, the geometric mean or the median, less sensitive to outliers, as a characteristic value (see Example 2.24). A thorough analysis of Bayes Factors, with its caveats and advantages, is given in [23].

A different alternative to quantify the evidence in favour of a particular model that avoids the need of the prior specification and is easy to evaluate is the Schwarz criteria [24] (or *"Bayes Information Criterion (BIC)"*). The rationale is the following. Consider a sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ and two alternative hypothesis for the models $M_i = \{p_i(x|\boldsymbol{\theta}_i); \dim(\boldsymbol{\theta}_i) = d_i\}$; $i = 1, 2$. As we can see in Sect. 4.5, under the appropriate conditions we can approximate the likelihood as

$$l(\boldsymbol{\theta}|\boldsymbol{x}) \simeq l(\widehat{\boldsymbol{\theta}}|\boldsymbol{x}) \exp\left\{-\frac{1}{2}\sum_{k=1}^{d}\sum_{m=1}^{d}(\theta_k - \widehat{\theta}_k)\left[n\boldsymbol{I}_{km}(\widehat{\boldsymbol{\theta}})\right](\theta_m - \widehat{\theta}_m)\right\}$$

so taking a uniform prior for the parameters $\boldsymbol{\theta}$, reasonable in the region where the likelihood is dominant, we can approximate

$$J(\boldsymbol{x}) = \int_{\boldsymbol{\Theta}} p(\boldsymbol{x}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta} \simeq p(\boldsymbol{x}|\widehat{\boldsymbol{\theta}})\,(2\pi/n)^{d/2}\,|\det[\boldsymbol{I}(\widehat{\boldsymbol{\theta}})]|^{-1/2}$$

and, ignoring terms that are bounded as $n \to \infty$, define the $BIC(M_i)$ for the model M_i as

$$2\ln\, J_i(\boldsymbol{x}) \simeq BIC(M_i) \;\equiv\; 2\,\ln\, p_i(\boldsymbol{x}|\widehat{\boldsymbol{\theta}}_i) - d_i\,\ln\, n$$

so:

[9] Essentially, the ratio of the predictive inferences for $\boldsymbol{x}_2$ after $\boldsymbol{x}_1$ has been observed.

$$B_{12} \simeq \frac{p_1(\boldsymbol{x}|\widehat{\boldsymbol{\theta}}_1)}{p_2(\boldsymbol{x}|\widehat{\boldsymbol{\theta}}_2)}\, n^{(d_2-d_1)/2} \quad \longrightarrow \quad \Delta_{12} = 2\ln B_{12} \simeq 2\ln\left(\frac{p_1(\boldsymbol{x}|\widehat{\boldsymbol{\theta}}_1)}{p_2(\boldsymbol{x}|\widehat{\boldsymbol{\theta}}_2)}\right) - (d_1 - d_2)\ln n$$

and therefore, larger values of $\Delta_{12} = BIC(M_1) - BIC(M_2)$ indicate a preference for the hypothesis $H_1(M_1)$ against $H_2(M_2)$ being commonly accepted that for values grater than 6 the evidence is *"strong"*[10] although, in some cases, it is worth to study the behaviour with a Monte Carlo sampling. Note that the last term penalises models with larger number of parameters and that this quantification is sound when the sample size n is much larger than the dimensions d_i of the parameters.

Example 2.23 Suppose that from the information provided by a detector we estimate the mass of an incoming particle and we want to decide upon the two exclusive and alternative hypothesis H_1 (particle of type 1) and $H_2(=H_1{}^c)$ (particle of type 2). We know from calibration data and Monte Carlo simulations that the mass distributions for both hypothesis are, to a very good approximation, Normal with means m_1 and m_2 variances σ_1^2 and σ_2^2 respectively. Then for an observed value of the mass m_0 we have:

$$B_{12} = \frac{p(m_0|H_1)}{p(m_0|H_2)} = \frac{N(m_0|m_1,\sigma_1)}{N(m_0|m_2,\sigma_2)} = \frac{\sigma_2}{\sigma_1}\exp\left\{\frac{(m_0-m_2)^2}{2\,\sigma_2^2} - \frac{(m_0-m_1)^2}{2\,\sigma_1^2}\right\}$$

Taking $(l_{12} = l_{21}; l_{11} = l_{22} = 0)$, the Bayesian decision criteria in favor of the hypothesis H_1 is:

$$B_{12} > \frac{P(H_2)}{P(H_1)} \quad \longrightarrow \quad \ln B_{12} > \ln\frac{P(H_2)}{P(H_1)}$$

Thus, we have a critical value m_c of the mass:

$$\sigma_1{}^2\,(m_c - m_2)^2 - \sigma_2{}^2\,(m_c - m_1)^2 = 2\,\sigma_1{}^2\,\sigma_2{}^2 \ln\left(\frac{P(H_2)\,\sigma_1}{P(H_1)\,\sigma_2}\right)$$

such that, if $m_0 < m_c$ we decide in favor of H_1 and for H_2 otherwise. In the case that $\sigma_1 = \sigma_2$ and $P(H_1) = P(H_2)$, then $m_c = (m_1 + m_2)/2$. This, however, may be a quite unrealistic assumption for if $P(H_1) > P(H_2)$, it may be more likely that the event is of type 1 being $B_{12} < 1$.

Example 2.24 Suppose we have an iid sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ of size n with $X \sim = N(x|\mu, 1)$ and the two hypothesis $H_1 = \{N(x|0, 1)\}$ and $H_2 = \{N(x|\mu, 1);\, \mu \neq 0\}$. Let us take $\{x_i\}$ as the minimum sample and, with the usual constant prior, consider the proper posterior

$$\pi(\mu|x_i) = \frac{1}{\sqrt{2\pi}}\,\exp\{-(\mu - x_i)^2/2\}$$

[10] If $P(H_1) = P(H_2) = 1/2$, then $P(H_1|\text{data}) = 0.95 \longrightarrow B_{12} = 19 \longrightarrow \Delta_{12} \simeq 6$.

that we use as a prior for the rest of the sample $\boldsymbol{x}' = \{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\}$. Then

$$\frac{P(H_1|\boldsymbol{x}', x_i)}{P(H_2|\boldsymbol{x}', x_i)} = B_{12}(i)\,\frac{P(H_1)}{P(H_2)}$$
$$\text{where}\quad B_{12}(i) = \frac{p(\boldsymbol{x}'|0)}{\int_{-\infty}^{\infty} p(\boldsymbol{x}'|\mu)\pi(\mu|x_i)d\mu} = n^{1/2}\,\exp\{-(n\,\overline{x}^2 - x_i^2)/2\}$$

and $\overline{x} = n^{-1}\sum_{k=1}^{n} x_k$. To avoid the effect that a particular choice of the minimal sample ($\{x_i\}$) may have, this is evaluated for all possible minimal samples and the median (or the geometric mean) of all the $B_{12}(i)$ is taken. Since $P(H_1|\boldsymbol{x}) + P(H_2|\boldsymbol{x}) = 1$, if we assign equal prior probabilities to the two hypothesis ($P(H_1) = P(H_2) = 1/2$) we have that

$$P(H_1|\boldsymbol{x}) = \frac{B_{12}}{1 + B_{12}} = \left(1 + n^{-1/2}\,\exp\{(n\,\overline{x}^2 - \text{med}\{x_i^2\})/2\}\right)^{-1}$$

is the posterior probability that quantifies the evidence in favor of the hypothesis H_1. It is left as an exercise to compare the Bayes Factor obtained from the geometric mean with what you would get if you were to take a proper prior $\pi(\mu|\sigma) = N(\mu|0, \sigma)$.

Problem 2.11 Suppose we have n observations (independent, under the same experimental conditions,…) of energies or decay time of particles above a certain known threshold and we want to test the evidence of an exponential fall against a power law. Consider then a sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ of observations with $\text{supp}(X) = (1, \infty)$ and the two models

$$M_1:\ p_1(x|\theta) = \theta\exp\{-\theta(x-1)\}\mathbf{1}_{(1,\infty)}(x) \quad \text{and} \quad M_2:\ p_2(x|\alpha) = \alpha x^{-(\alpha+1)}\mathbf{1}_{(1,\infty)}(x)$$

that is, Exponential and Pareto with unknown parameters θ and α. Show that for the minimal sample $\{x_i\}$ and reference priors, the Bayes Factor $B_{12}(i)$ is given by

$$B_{12}(i) = \left(\frac{x_g\,\ln x_g}{\overline{x} - 1}\right)^n \left(\frac{x_i - 1}{x_i\,\ln x_i}\right) = \frac{p_1(\boldsymbol{x}|\widehat{\boldsymbol{\theta}})}{p_2(\boldsymbol{x}|\widehat{\boldsymbol{\alpha}})}\,\left(\frac{x_i - 1}{x_i\,\ln x_i}\right)$$

where $(\overline{x}, x_g)$ are the arithmetic and geometric sample means and $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\alpha}})$ the values that maximize the likelihoods and therefore

$$\text{med}\{B_{12}(i)\}_{i=1}^n = \left(\frac{x_g\,\ln x_g}{\overline{x} - 1}\right)^n \text{med}\left\{\frac{x_i - 1}{x_i\,\ln x_i}\right\}_{i=1}^n$$

Problem 2.12 Suppose we have two experiments $e_i(n_i)$; $i = 1, 2$ in which, out of n_i trials, x_i successes have been observed and we are interested in testing whether

both treatments are different or not (*contingency tables*). If we assume Binomial models $Bi(x_i|n_i, \theta_i)$ for both experiments and the two hypothesis $H_1 : \{\theta_1 = \theta_2\}$ and $H_2 : \{\theta_1 \neq \theta_2\}$, the Bayes Factor will be

$$B_{12} = \frac{\int_\Theta Bi(x_1|n_1, \theta) Bi(x_2|n_2, \theta)\pi(\theta)d\theta}{\int_{\Theta_1} Bi(x_1|n_1, \theta_1)\pi(\theta_1)d\theta_1 \int_{\Theta_2} Bi(x_2|n_2, \theta)\pi(\theta_2)d\theta_2}$$

We may consider proper Beta prior densities $Be(\theta|a, b)$. In a specific pharmacological analysis, a sample of $n_1 = 52$ individuals were administered a placebo and $n_2 = 61$ were treated with an a priori beneficial drug. After the essay, positive effects were observed in $x_1 = 22$ out of the 52 and $x_2 = 41$ out of the 61 individuals. It is left as an exercise to obtain the posterior probability $P(H_2|\text{data})$ with Jeffreys' ($a = b = 1/2$) and Uniform ($a = b = 1$) priors and to determine the BIC difference Δ_{12}.

2.10.2 Point Estimation

When we have to face the problem to characterize the posterior density by a single number, the most usual *Loss Functions* are:

• **Quadratic Loss**: In the simple one-dimensional case, the Loss Function is

$$l(\theta, a) = (\theta - a)^2$$

so, minimizing the *Risk*:

$$\min \int_{\Omega_\theta} (\theta - a)^2 \, p(\theta|\boldsymbol{x}) \, d\theta \quad \longrightarrow \quad \int_{\Omega_\theta} (\theta - a) \, p(\theta|\boldsymbol{x}) \, d\theta = 0$$

and therefore $a = E[\theta]$; that is, the posterior mean.

In the $k-$dimensional case, if $\mathcal{A} = \Omega_\theta = \mathcal{R}^k$ we shall take as Loss Function

$$l(\boldsymbol{\theta}, \boldsymbol{a}) = (\boldsymbol{a} - \boldsymbol{\theta})^T \, \mathbf{H} \, (\boldsymbol{a} - \boldsymbol{\theta})$$

where $\mathbf{H}$ is a positive defined symmetric matrix. It is clear that:

$$\min \int_{\mathcal{R}^k} (\boldsymbol{a} - \boldsymbol{\theta})^T \, \mathbf{H} \, (\boldsymbol{a} - \boldsymbol{\theta}) \, p(\boldsymbol{\theta}|\boldsymbol{x}) \, d\boldsymbol{\theta} \quad \longrightarrow \quad \mathbf{H}\,\boldsymbol{a} = \mathbf{H}\, E[\boldsymbol{\theta}]$$

so, if $\mathbf{H}^{-1}$ exists, then $\boldsymbol{a} = E[\boldsymbol{\theta}]$. Thus, we have that the Bayesian estimate under a quadratic loss function is the mean of $p(\boldsymbol{\theta}|\boldsymbol{x})$ (… if exists!).

• **Linear Loss**: If $\mathcal{A} = \Omega_\theta = \mathcal{R}$, we shall take the loss function:

$$l(\theta, a) = c_1\,(a-\theta)\,\mathbf{1}_{\theta \le a} + c_2\,(\theta - a)\,\mathbf{1}_{\theta > a}$$

Then, the estimator will be such that

$$\min \int_{\Omega_\theta} l(a,\theta)\,p(\theta|\boldsymbol{x})d\theta = \min\left(c_1 \int_{-\infty}^{a} (a-\theta)\,p(\theta|\boldsymbol{x})d\theta + c_2 \int_{a}^{\infty} (\theta - a)\,p(\theta|\boldsymbol{x})d\theta\right)$$

After derivative with respect to a we have $(c_1 + c_2)\,P(\theta \le a) - c_2 = 0$ and therefore the estimator will be the value of a such that

$$P(\theta \le a) = \frac{c_2}{c_1 + c_2}$$

In particular, if $c_1 = c_2$ then $P(\theta \le a) = 1/2$ and we shall have the median of the distribution $p(\boldsymbol{\theta}|\boldsymbol{x})$. In this case, the Loss Function can be expressed more simply as $l(\theta, a) = |\theta - a|$.

• **Zero-One Loss**: Si $\mathcal{A} = \Omega_\theta = \mathcal{R}^k$, we shall take the Loss Function

$$l(\boldsymbol{\theta}, \boldsymbol{a}) = 1 - \mathbf{1}_{\mathcal{B}_\epsilon(\boldsymbol{a})}$$

where $\mathcal{B}_\epsilon(\boldsymbol{a}) \in \Omega_\theta$ is an open ball of radius ϵ centered at $\boldsymbol{a}$. The corresponding point estimator will be:

$$\min \int_{\Omega_\theta} (1 - \mathbf{1}_{\mathcal{B}_\epsilon(\boldsymbol{a})})\,p(\boldsymbol{\theta}|\boldsymbol{x})\,d\boldsymbol{\theta} = \max \int_{\mathcal{B}_\epsilon(\boldsymbol{a})} p(\boldsymbol{\theta}|\boldsymbol{x})\,d\boldsymbol{\theta}$$

It is clear that, in the limit $\epsilon \to 0$, the Bayesian estimator for the Zero-One Loss Function will be the mode of $p(\boldsymbol{\theta}|\boldsymbol{x})$ if exists.

As explained in Chap. 1, the mode, the median and the mean can be very different if the distribution is not symmetric. Which one should we take then? Quadratic losses, for which large deviations from the *true* value are penalized quadratically, are the most common option but, even if for unimodal symmetric the three statistics coincide, it may be misleading to take this value as a characteristic number for the information we got about the parameters or even be nonsense. In the hypothetical case that the posterior is essentially the same as the likelihood (that is the case for a sufficiently smooth prior), the Zero-One Loss points to the classical estimate of the *Maximum Likelihood Method*. Other considerations of interest in Classical Statistics (like bias, consistency, minimum variance,…) have no special relevance in Bayesian inference.

Problem 2.13 (*The Uniform Distribution*) Show that for the posterior density (see Example 2.4)

$$p(\theta|x_M, n) = n\,\frac{x_M^n}{\theta^{n+1}}\,\mathbf{1}_{[x_M,\infty)}(\theta)$$

the point estimates under quadratic, linear and 0–1 loss functions are

$$\theta_{QL} = x_M\,\frac{n}{n-1}; \qquad \theta_{LL} = x_M\,2^{1/n} \quad \text{and} \qquad \theta_{01L} = x_M$$

and discuss which one you consider more reasonable.

2.11 Credible Regions

Let $p(\boldsymbol{\theta}|\boldsymbol{x})$, with $\boldsymbol{\theta} \in \Omega \subseteq \mathcal{R}^n$ be a posterior density function. A credible region with probability content $1-\alpha$ is a region of $V_\alpha \subseteq \Theta$ of the parametric space such that

$$P(\boldsymbol{\theta} \in V_\alpha) = \int_{V_\alpha} p(\boldsymbol{\theta}|\boldsymbol{x})\,d\boldsymbol{\theta} = 1-\alpha$$

Obviously, for a given probability content credible regions are not unique and a sound criteria is to specify the one that the smallest possible volume. A region C of the parametric space Ω is called *Highest Probability Region* (HPD) with probability content $1-\alpha$ if:

(1) $P(\boldsymbol{\theta} \in C) = 1-\alpha; \quad C \subseteq \Omega;$
(2) $p(\boldsymbol{\theta}_1|\cdot) \geq p(\boldsymbol{\theta}_2|\cdot)$ for all $\boldsymbol{\theta}_1 \in C$ and $\boldsymbol{\theta}_2 \notin C$ except, at most, for a subset of Ω with zero probability measure.

It is left as an exercise to show that condition (2) implies that the HPD region so defined is of minimum volume so both definitions are equivalent. Further properties that are easy to demonstrate are:

(1) If $p(\theta|\cdot)$ is *not uniform*, the HPD region with probability content $1-\alpha$ *is unique*;
(2) If $p(\boldsymbol{\theta}_1|\cdot) = p(\boldsymbol{\theta}_2|\cdot)$, then $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are both either included or excluded of the HPD region;
(3) If $p(\boldsymbol{\theta}_1|\cdot) \neq p(\boldsymbol{\theta}_2\cdot)$, there is an HPD region for some value of $1-\alpha$ that contains one value of $\boldsymbol{\theta}$ and not the other;
(4) $C = \{\boldsymbol{\theta} \in \Theta | p(\boldsymbol{\theta}|\boldsymbol{x}) \geq k_\alpha\}$ where k_α is the largest constant for which $P(\boldsymbol{\theta} \in C) \geq \alpha$;
(5) If $\boldsymbol{\phi} = f(\boldsymbol{\theta})$ is a one-to-one transformation, then
    (a) any region with probability content $1-\alpha$ for $\boldsymbol{\theta}$ will have probability content $1-\alpha$ for $\boldsymbol{\phi}$ but…
    (b) an HPD region for $\boldsymbol{\theta}$ will not, in general, be an HPD region for $\boldsymbol{\phi}$ unless the transformation is linear.

In general, evaluation of credible regions is a bit messy task. A simple way through is to do a Monte Carlo sampling of the posterior density and use the 4th property.

For a one-dimensional parameter, the condition that the HPD region with probability content $1-\alpha$ has the minimum length allows to write a relation that may be useful to obtain those regions in an easier manner. Let $[\theta_1, \theta_2]$ be an interval such that

$$\int_{\theta_1}^{\theta_2} p(\theta|\cdot)\, d\theta = 1-\alpha$$

For this to be an HPD region we have to find the extremal of the function

$$\phi(\theta_1, \theta_2, \lambda) = (\theta_2 - \theta_1) + \lambda \left(\int_{\theta_1}^{\theta_2} p(\theta|\cdot)\, d\theta - (1-\alpha) \right)$$

Taking derivatives we get:

$$\begin{aligned} \left(\frac{\partial \phi(\theta_1, \theta_2, \lambda)}{\partial \theta_i} \right)_{i=1,2} = 0 \qquad &\longrightarrow \qquad p(\theta_1|\cdot) = p(\theta_2 \cdot) \\ \frac{\partial \phi(\theta_1, \theta_2, \lambda)}{\partial \lambda} = 0 \qquad &\longrightarrow \qquad \int_{\theta_1}^{\theta_2} p(\theta)\, d\theta = 1-\alpha \end{aligned}$$

Thus, from the first two conditions we have that $p(\theta_1|\cdot) = p(\theta_2|\cdot)$ and, from the third, we know that $\theta_1 \neq \theta_2$. In the special case that the distribution is unimodal and symmetric the only possible solution is $\theta_2 = 2E[\theta] - \theta_1$.

The HPD regions are useful to summarize the information on the parameters contained in the posterior density $p(\boldsymbol{\theta}|\boldsymbol{x})$ but it should be clear that there is no justification to reject a particular value $\boldsymbol{\theta}_0$ just because is not included in the HPD region (or, in fact, in whatever confidence region) and that in some circumstances (distributions with more than one mode for instance) it may be the union of disconnected regions.

2.12 Bayesian ($\mathcal{B}$) Versus Classical ($\mathcal{F}$) Philosophy

The Bayesian philosophy aims at the right questions in a very intuitive and, at least conceptually, simple manner. However the *"classical"* (frequentist) approach to statistics, that has been very useful in scientific reasoning over the last century, is at present more widespread in the Particle Physics community and most of the stirred up controversies are originated by misinterpretations. It is worth to take a look for instance at [2]. Let's see how a simple problem is attacked by the two schools. "We" are $\mathcal{B}$, "they" are $\mathcal{F}$.

Suppose we want to estimate the life-time of a particle. We both *"assume"* an exponential model $X \sim Ex(x|1/\tau)$ and do an experiment $e(n)$ that provides an iid sample $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$. In this case there is a sufficient statistic $\boldsymbol{t} = (n, \overline{x})$ with $\overline{x}$ the sample mean so let's define the random quantity

$$X = \frac{1}{n}\sum_{i=1}^{n} X_i \sim p(x|n,\tau) = \left(\frac{n}{\tau}\right)^n \frac{1}{\Gamma(n)} \exp\left\{-nx\tau^{-1}\right\} x^{n-1}\, \mathbf{1}_{(0,\infty)}(x)$$

What can we say about the parameter of interest τ?

$\mathcal{F}$ will start by finding the *estimator* (statistic) $\widehat{\tau}$ that maximizes the likelihood (MLE). In this case it is clear that $\widehat{\tau} = \overline{x}$, the sample mean. We may ask about the rationale behind because, apparently, there is no serious mathematical reasoning that justifies this procedure. $\mathcal{F}$ will respond that, in a certain sense, even for us this should be a reasonable way because if we have a smooth prior function, the posterior is dominated by the likelihood and one possible point estimator is the mode of the posterior. Beside that, he will argue that maximizing the likelihood renders an estimator that often has "good" properties like unbiasedness, invariance under monotonous one-to-one transformations, consistency (convergence in probability), smallest variance within the class of unbiased estimators (Cramèr-Rao bound), approximately well known distribution,… We may question some of them (unbiased estimators are not always the best option and invariance… well, if the transformation is not linear usually the MLE is biased), argue that the others hold in the asymptotic limit,… Anyway; for this particular case one has that:

$$E[\widehat{\tau}] = \tau \qquad \text{and} \qquad V[\widehat{\tau}] = \frac{\tau^2}{n}$$

and $\mathcal{F}$ will claim that *"if you repeat the experiment"* many times under the same conditions, you will get a sequence of estimators $\{\widehat{\tau}_1, \widehat{\tau}_2, \ldots\}$ that eventually will cluster around the life-time τ. Fine but we shall point out that, first, although desirable we usually do not repeat the experiments (and under the same conditions is even more rare) so we have just one observed sample ($\boldsymbol{x} \to \overline{x} = \widehat{\tau}$) from $e(n)$. Second, *"if you repeat the experiment you will get"* is a free and unnecessary hypothesis. You do not know what you will get, among other things, because the model we are considering may not be the way nature behaves. Besides that, it is quite unpleasant that inferences on the life-time depend upon what you think you will get if you do what you know you are not going to do. And third, that this is in any case a nice sampling property of the estimator $\hat{\tau}$ but eventually we are interested in τ so, What can we say about it?

For us, the answer is clear. Being τ a scale parameter we write the posterior density function

$$p(\tau|n,\overline{x}) = \frac{(n\,\overline{x})^n}{\Gamma(n)} \exp\left\{-n\,\overline{x}\tau^{-1}\right\} \tau^{-(n+1)}\, \mathbf{1}_{(0,\infty)}(\tau)$$

for the *degree of belief* we have on the parameter and easily get for instance:

$$E[\tau^k] = (n\overline{x})^k \frac{\Gamma(n-k)}{\Gamma(n)} \quad \longrightarrow \quad E[\tau] = \overline{x}\frac{n}{n-1}; \quad V[\tau] = \overline{x}^2 \frac{n^2}{(n-1)^2(n-2)}; \ldots$$

Cleaner and simpler impossible.

To bound the life-time, $\mathcal{F}$ proceeds with the determination of the *Confidence Intervals*. The classical procedure was introduced by J. Neyman in 1933 and rests on establishing, for an specified probability content, the domain of the random quantity (usually a statistic) as function of the possible values the parameters may take. Consider a one dimensional parameter θ and the model $X \sim p(x|\theta)$. Given a desired probability content $\beta \in [0, 1]$, he determines the interval $[x_1, x_2] \subset \Omega_X$ such that

$$P(X \in [x_1, x_2]) = \int_{x_1}^{x_2} p(x|\theta)\, dx = \beta$$

for a particular fixed value of θ. Thus, for each possible value of θ he has one interval $[x_1 = f_1(\theta; \beta), x_1 = f_2(\theta; \beta)] \subset \Omega_X$ and the sequence of those intervals gives a band in the $\Omega_\theta \times \Omega_X$ region of the real plane. As for the *Credible Regions*, these intervals are not uniquely determined so one usually adds the condition:

$$(1) \qquad \int_{-\infty}^{x_1} p(x|\theta)\, dx = \int_{x_2}^{\infty} p(x|\theta)\, dx = \frac{1-\beta}{2} \qquad \text{or}$$

$$(2) \qquad \int_{x_1}^{\theta} p(x|\theta)\, dx = \int_{\theta}^{x_2} p(x|\theta)\, dx = \frac{\beta}{2}$$

or, less often, (3) chooses the interval with smallest size. Now, for an invertible mapping $x_i \longrightarrow f_i(\theta)$ one can write

$$\beta = P(f_1(\theta) \leq X \leq f_2(\theta)) = P(f_2^{-1}(X) \leq \theta \leq f_1^{-1}(X))$$

and get the random interval $[f_2^{-1}(X),\ f_1^{-1}(X)]$ that contains the given value of θ with probability β. Thus, for each possible value that X may take he will get an interval $[f_2^{-1}(X),\ f_1^{-1}(X)]$ on the θ axis and a particular experimental observation $\{x\}$ will single out one of them. This is the *Confidence Interval* that the frequentist analyst will quote. Let's continue with the life-time example and take, for illustration, $n = 50$ and $\beta = 0.68$. The bands $[x_1 = f_1(\tau), x_2 = f_2(\tau)]$ in the (τ, X) plane, in this case obtained with the third prescription, are shown in Fig. 2.4(1). They are essentially straight lines so $P[X \in (0.847\tau, 1.126\tau)] = 0.68$. This is a correct statement, but doesn't say anything about τ so he inverts that and gets $0.89\, X < \tau < 1.18\, X$ in such a way that an observed value $\{x\}$ singles out an interval in the vertical τ axis. We, Bayesians, will argue this does not mean that τ has a 0.68 chance to lie in this interval and the frequentist will certainly agree on that. In fact, this is not an admissible question for him because in the classical philosophy τ is a number, unknown but a *fixed* number. If he repeats the experiment τ will not change; it is the interval that will be different because x will change. They are *random intervals* and what the 68% means is just that if he repeats the experiment a large number N of times, he will end up with N intervals of which ∼68% will contain the true value τ whatever it is. But the experiment is done only once so: Does the interval derived from this observation contain τ or not? We don't know, we have no idea if it does contain τ,

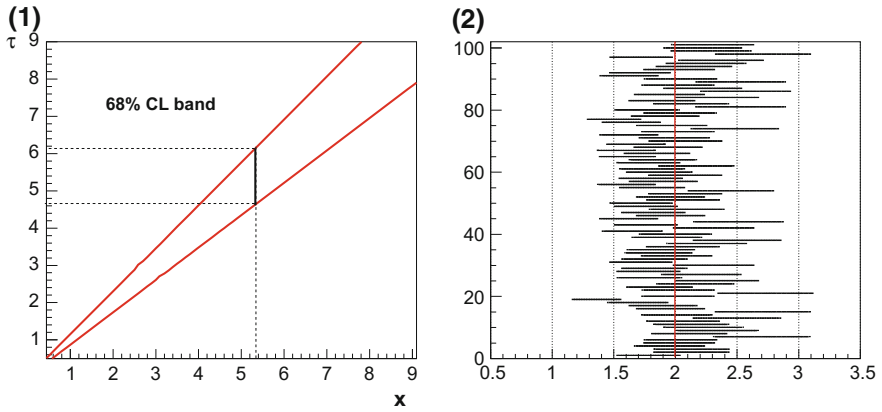

Fig. 2.4 (**1**) 68% confidence level bands in the (τ, X) plane. (**2**) 68% confidence intervals obtained for 100 repetitions of the experiment

if it does not and how far is the unknown true value. Figure 2.4(2) shows the 68% confidence intervals obtained after 100 repetitions of the experiment for $\tau = 2$ and 67 of them did contain the true value. But when the experiment is done once, he picks up one of those intervals and has a 68% chance that the one chosen contains the true value. We $\mathcal{B}$ shall proceed in a different manner. After integration of the posterior density we get the HPD interval $P\,[\tau \in (0.85\,x,\, 1.13\,x)] = 0.68$; almost the same but with a direct interpretation in terms of what we are interested in. Thus, both have an absolutely different philosophy:

$\mathcal{F}$: "Given a particular value of the parameters of interest, How likely is the observed data?"

$\mathcal{B}$: "Having observed this data, What can we say about the parameters of interest?"

… and the probability if the causes, as Poincare said, is the most important from the point of view of scientific applications.

In many circumstances we are also interested in one-sided intervals. That is for instance the case when the data is consistent with the hypothesis $H : \{\theta = \theta_0\}$ and we want to give an upper bound on θ so that $P(\theta \in (-\infty, \theta_\beta] = \beta$. The frequentist rationale is the same: obtain the interval $[-\infty, x_2] \subset \Omega_X$ such that

$$P(X \leq x_2) = \int_{-\infty}^{x_2} p(x|\theta)\, dx = \beta$$

where $x_2 = f_2(\theta)$; in this case without ambiguity. For the random interval $(-\infty, f_2^{-1}(X))$ $\mathcal{F}$ has that

$$P\left(\theta < f_2^{-1}(X)\right) = 1 - P\left(\theta \geq f_2^{-1}(X)\right) = 1 - \beta$$

so, for a probability content α (say 0.95), one should set $\beta = 1 - \alpha$ (=0.05). Now, consider for instance the example of the anisotropy is cosmic rays discussed in the last

Sect. 2.13.3. For a dipole moment (details are unimportant now) we have a statistic

$$X \sim p(x|\theta, 1/2) = \frac{\exp\{-\theta^2/2\}}{\sqrt{2\pi}\theta} \exp\{-x/2\}\sinh(\theta\sqrt{x})\, \mathbf{1}_{(0,\infty)}(x)$$

where the parameter θ is the dipole coefficient multiplied by a factor that is irrelevant for the example. It is argued in Sect. 2.13.3 that the reasonable prior for this model is $\pi(\theta) = \text{constant}$ so we have the posterior

$$p(\theta|x, 1/2) = \frac{\sqrt{2}}{\sqrt{\pi x}\, M(1/2, 3/2, x/2)} \exp\{-\theta^2/2\}\, \theta^{-1} \sinh(\theta\sqrt{x})\, \mathbf{1}_{(0,\infty)}(\theta)$$

with $M(a, b, z)$ the Kummer's Confluent Hypergeometric Function. In fact, θ has a compact support but since the observed values of X are consistent with $H_0 : \{\theta = 0\}$ and the sample size is very large [AMS13],[11] $p(\theta|x, 1/2)$ is concentrated in a small interval $(0, \epsilon)$ and it is easier for the evaluations to extend the domain to $\mathcal{R}^+$ without any effect on the results. Then we, Bayesians, shall derive the one-sided upper credible region $[0, \theta_{0.95}(x)]$ with $\alpha = 95\%$ probability content as simply as:

$$\int_0^{\theta_{0.95}} p(\theta|x, 1/2)\, d\theta = \alpha = 0.95$$

This upper bound shown as function of x in Fig. 2.5 under "*Bayes*" (red line). Neyman's construction is also straight forward. From

$$\int_0^{x_2} p(x|\theta, 1/2)\, dx = 1 - \alpha = 0.05$$

(essentially a χ^2 probability for $\nu = 3$), $\mathcal{F}$ will get the upper bound shown in the same figure under "*Neyman*" (blue broken line). As you can see, they get closer as x grows but, first, there is no solution for $x \leq x_c = 0.352$. In fact, $E[X] = \theta^2 + 3$ so if the dipole moment is $\delta = 0\,(\theta = 0)$, $E[X] = 3$ and observed values below x_c will be an unlikely fluctuation downwards (assuming of course that the model is correct) but certainly a possible experimental outcome. In fact, you can see that for values of x less than 2, even though there is a solution Neyman's upper bound is underestimated. To avoid this "little" problem, a different prescription has to be taken.

The most interesting solution is the one proposed by Feldman and Cousins [25] in which the region $\Delta_X \subset \Omega_X$ that is considered for the specified probability content is determined by the ratio of probability densities. Thus, for a given value θ_0, the interval Δ_X is such that

[11] [AMS13]: Aguilar M. et al. (2013); Phys. Rev. Lett. 110, 141102.

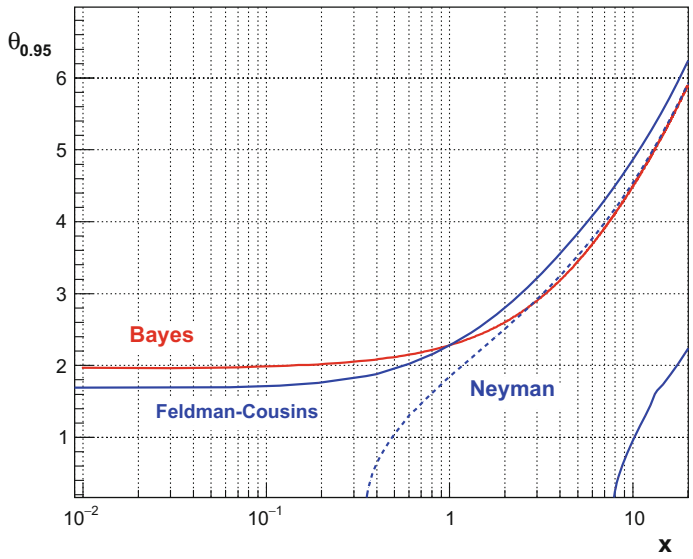

Fig. 2.5 95% upper bounds on the parameter θ following the Bayesian approach (*red*), the Neyman approach (*broken blue*) and Feldman and Cousins (*solid blue line*)

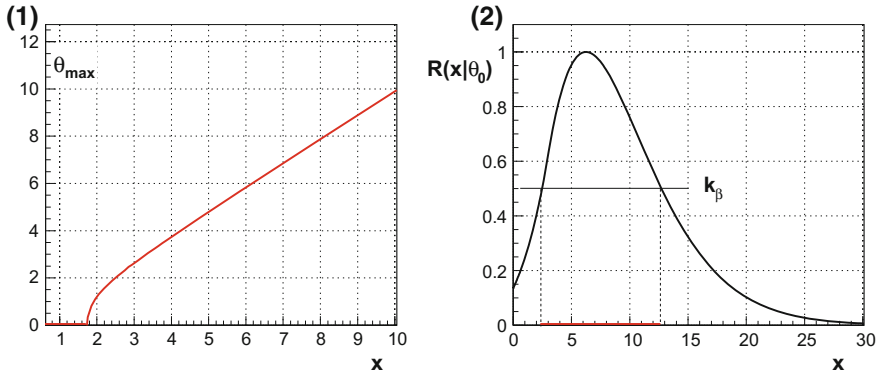

Fig. 2.6 (**1**) Dependence of θ_m with x. (**2**) Probability density ratio $R(x|\theta)$ for $\theta = 2$

$$\int_{\Delta_X} p(x|\theta_0)\,dx = \beta \qquad \text{with} \qquad R(x|\theta_0) = \frac{p(x|\theta_0)}{p(x|\theta_b)} > k_\beta;\ \forall x \in \Delta_X$$

and where θ_b is the *best* estimation of θ for a given $\{x\}$; usually the one that maximizes the likelihood (θ_m). In our case, it is given by:

$$\theta_m = \begin{cases} 0 & \text{if } x \leq \sqrt{3} \\ \theta_m + \theta_m^{-1} - \sqrt{x}\coth(\theta_m\sqrt{x}) = 0 & \text{if } x > \sqrt{3} \end{cases}$$

and the dependence with x is shown in Fig. 2.6(1) ($\theta_m \simeq x$ for $x >>$). As illustration, function $R(x|\theta)$ is shown in Fig. 2.6(2) for the particular value $\theta_0 = 2$. Following this procedure,[12] the 0.95 probability content band is shown in Fig. 2.5 under *"Feldman-Cousins"* (blue line). Note that for large values of x, the confidence region becomes an interval. It is true that if we observe a large value of X, the hypothesis $H_0 : \{\delta = 0\}$ will not be favoured by the data and a different analysis will be more relevant although, by a simple modification of the ordering rule, we still can get an upper bound if desired or use the standard Neyman's procedure.

The Feldman and Cousins prescription allows to consider constrains on the parameters in a simpler way than Neyman's procedure and, as opposed to it, will always provide a region with the specified probability content. However, on the one hand, they are frequentist intervals and as such have to be interpreted. On the other hand, for discrete random quantities with image in $\{x_1, \ldots, x_k, \ldots\}$ it may not be possible to satisfy exactly the probability content equation since for the Distribution Function one has that $F(x_{k+1}) = F(x_k) + P(X = x_{k+1})$. And last, it is not straight forward to deal with nuisance parameters. Therefore, the best advice: "Be Bayesian!".

2.13 Some Worked Examples

2.13.1 Regression

Consider the exchangeable sequence $\boldsymbol{z} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ of n samplings from the two-dimensional model $N(x_i, y_i|\cdot) = N(x_i|\mu_{x_i}, \sigma^2_{x_i})N(y_i|\mu_{y_i}, \sigma^2_{y_i})$. Then

$$p(\boldsymbol{z}|\cdot) \propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left[\frac{(y_i - \mu_{yi})^2}{\sigma^2_{yi}} + \frac{(x_i - \mu_{xi})^2}{\sigma^2_{xi}}\right]\right\}$$

We shall assume that the precisions σ_{xi} and σ_{xi} are known and that there is a functional relation $\mu_y = f(\mu_x; \boldsymbol{\theta})$ with unknown parameters $\boldsymbol{\theta}$. Then, in terms of the new parameters of interest:

$$p(\boldsymbol{y}|\cdot) \propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left[\frac{(y_i - f(\mu_{x_i}; \boldsymbol{\theta}))^2}{\sigma^2_{y_i}} + \frac{(x_i - \mu_{x_i})^2}{\sigma^2_{x_i}}\right]\right\}$$

Consider a linear relation $f(\mu_x; a, b) = a + b\mu_x$ with a, b the unknown parameters so:

[12] In most cases,a Monte Carlo simulation will simplify life.

$$p(\boldsymbol{z}|\cdot) \propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left[\frac{(y_i - a - b\mu_{x_i})^2}{\sigma_{y_i}^2} + \frac{(x_i - \mu_{x_i})^2}{\sigma_{x_i}^2}\right]\right\}$$

and assume, in first place, that $\mu_{x_i} = x_i$ without uncertainty. Then,

$$p(\boldsymbol{y}|a, b) \propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left[\frac{(y_i - a - bx_i)^2}{\sigma_{y_i}^2}\right]\right\}$$

There is a set of sufficient statistics for (a, b):

$$\boldsymbol{t} = \{t_1,\, t_2,\, t_3,\, t_4,\, t_5\} = \left\{\sum_{i=1}^{n}\frac{1}{\sigma_i^2},\, \sum_{i=1}^{n}\frac{x_i^2}{\sigma_i^2},\, \sum_{i=1}^{n}\frac{x_i}{\sigma_i^2},\, \sum_{i=1}^{n}\frac{y_i}{\sigma_i^2},\, \sum_{i=1}^{n}\frac{y_i x_i}{\sigma_i^2}\right\}$$

and, after a simple algebra, it is easy to write

$$p(\boldsymbol{y}|a, b) \propto \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(a-a_0)^2}{\sigma_a^2} + \frac{(b-b_0)^2}{\sigma_b^2} - 2\,\rho\frac{(a-a_0)}{\sigma_a}\frac{(b-b_0)}{\sigma_b}\right]\right\}$$

where the new statistics $\{a_0, b_0, \sigma_a, \sigma_b, \rho\}$ are defined as:

$$\begin{aligned} a_0 &= \frac{t_2 t_4 - t_3 t_5}{t_1 t_2 - t_3^2}, \qquad b_0 = \frac{t_1 t_5 - t_3 t_4}{t_1 t_2 - t_3^2} \\ \sigma_a^2 &= \frac{t_2}{t_1 t_2 - t_3^2}, \qquad \sigma_b^2 = \frac{t_1}{t_1 t_2 - t_3^2}, \qquad \rho = -\frac{t_3}{\sqrt{t_1 t_2}} \end{aligned}$$

Both (a, b) are position parameters so we shall take a uniform prior and in consequence

$$p(a, b|\cdot) = \frac{1}{2\pi\sigma_a\sigma_b\sqrt{1-\rho^2}}\, e^{\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(a-a_0)^2}{\sigma_a^2} + \frac{(b-b_0)^2}{\sigma_b^2} - 2\,\rho\frac{(a-a_0)}{\sigma_a}\frac{(b-b_0)}{\sigma_b}\right]\right\}}$$

This was obviously expected.

When μ_{x_i} are n unknown parameters, if we take $\pi(\mu_{xi}) = \mathbf{1}_{(0,\infty)}(\mu_{xi})$ and marginalize for (a, b) we have

$$p(a, b|\cdot) \propto \pi(a, b)\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - a - b\,x_i)^2}{\sigma_{yi}^2 + b^2\,\sigma_{xi}^2}\right\}\left\{\prod_{i=1}^{n}(\sigma_{yi}^2 + b^2\,\sigma_{xi}^2)\right\}^{-1/2}$$

In general, the expressions one gets for non-linear regression problems are complicated and setting up priors is a non-trivial task but fairly vague priors easy to deal with are usually a reasonable choice. In this case, for instance, one may consider uniform

priors or normal densities $N(\cdot|0, \sigma >>)$ for both parameters (a, b) and sample the proper posterior with a Monte Carlo algorithm (Gibbs sampling will be appropriate).

The same reasoning applies if we want to consider other models or more involved relations with several explanatory variables like $\theta_i = \sum_{j=1}^{k} \alpha_j x_{ij}^{b_j}$. In counting experiments, for example, $y_i \in \mathcal{N}$ so we may be interested in a Poisson model $Po(y_i|\mu_i)$ where μ_i is parameterized as a simple log-linear form $\ln(\mu_i) = \alpha_0 + \alpha_1 x_i$ (so $\mu_i > 0$ for whatever $\alpha_0, \alpha_1 \in \mathcal{R}$). Suppose for instance that we have the sample $\{(y_i, x_i)\}_{i=1}^{n}$. Then:

$$p(\boldsymbol{y}|\alpha_1, \alpha_2, \boldsymbol{x}) \propto \prod_{i=1}^{n} \exp\{-\mu_i\} {\mu_i}^{y_i} = \exp\left\{\alpha_1 s_1 + \alpha_2 s_2 - e^{\alpha_1} \sum_{i=1}^{n} e^{\alpha_2 x_i}\right\}$$

where $s_1 = \sum_{i=1}^{n} y_i$ and $s_2 = \sum_{i=1}^{n} y_i x_i$. In this case, the Normal distribution $N(\alpha_i|a_i, \sigma_i)$ with $\sigma_i >>$ is a reasonable smooth and easy to handle proper prior density for both parameters. Thus, we get the posterior conditional densities

$$p(\alpha_i|\alpha_j, \boldsymbol{y}, \boldsymbol{x}) \propto \exp\left\{-\frac{\alpha_i^2}{2\sigma_i^2} + \alpha_i \left(\frac{a_i}{\sigma_i^2} + s_i\right) - e^{\alpha_1} \sum_{i=1}^{n} e^{\alpha_2 x_i}\right\}; \; i = 1, 2$$

that are perfectly suited for the Gibbs sampling to be discussed in Sect. 4.1 of Chap. 3.

Example 2.25 (*Proton Flux in Primary Cosmic Rays*) For energies between ∼20 and ∼200 GeV, the flux of protons of the primary cosmic radiation is reasonably well described by a power law $\phi(r) = c\, r^{\gamma}$ where r is the *rigidity*[13] and $\gamma = d\ln\phi/ds$, with $s = \ln r$, is the *spectral index*. At lower energies, this dependence is significantly modified by the geomagnetic cut-off and the solar wind but at higher energies, where these effects are negligible, the observations are not consistent with a single power law (Fig. 2.7(1)). One may characterize this behaviour with a simple phenomenological model where the spectral index is no longer constant but has a dependence $\gamma(s) = \alpha + \beta \tanh[a(s - s_0)]$ such that $\lim_{s\to-\infty} \gamma(s) = \gamma_1$ $(r\to 0)$ and $\lim_{s\to\infty} \gamma(s) = \gamma_2$ $(r\to +\infty)$. After integration, the flux can be expressed in terms of 5 parameters $\boldsymbol{\theta} = \{\phi_0, \gamma_1, \delta = \gamma_2 - \gamma_1, r_0, \sigma\}$ as:

$$\phi(r; \boldsymbol{\theta}) = \phi_0\, r^{\gamma_1} \left[1 + \left(\frac{r}{r_0}\right)^{\sigma}\right]^{\delta/\sigma}$$

For this example, I have used the data above 45 GeV published by the AMS experiment[14] and considered only the quoted *statistical errors* (see Fig. 2.7(1)). Last, for a better description of the flux the previous expression has been modified to account for the effect of the solar wind with the force-field approximation in consistency with

[13] The *rigidity* (r) is defined as the momentum (p) divided by the electric charge (Z) so $r = p$ for protons.

[14] [AMS15]: Aguilar M. et al. (2015); PRL 114, 171103 and references therein.

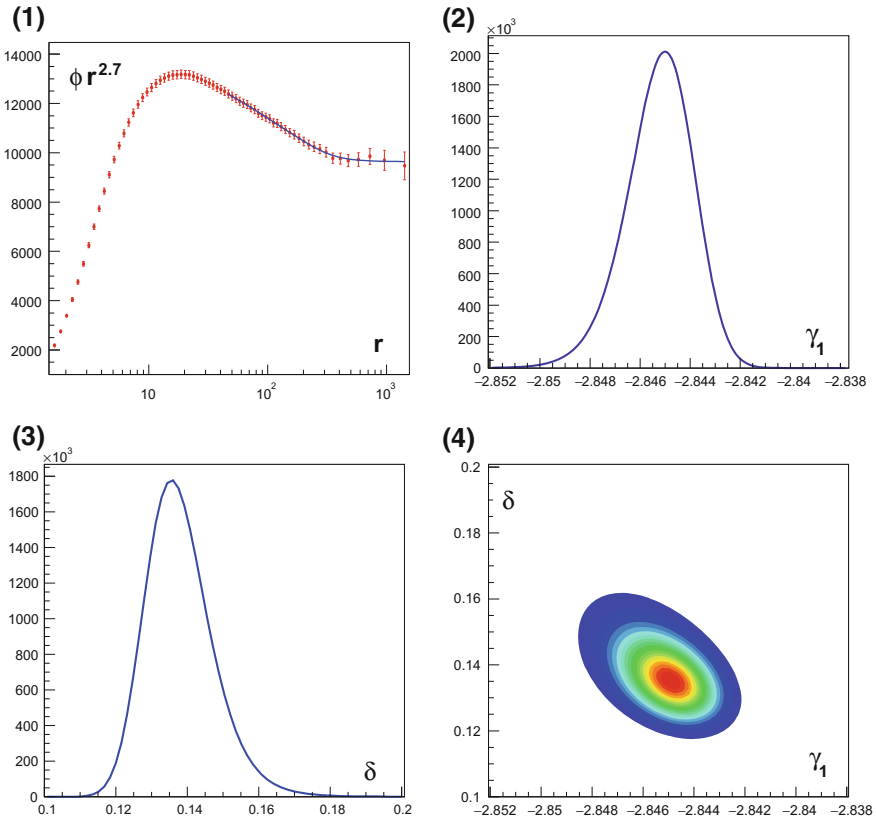

Fig. 2.7 (**1**) Observed flux multiplied by $r^{2.7}$ in $\mathrm{m}^{-2}\mathrm{sr}^{-1}\mathrm{sec}^{-1}\mathrm{GV}^{1.7}$ as given in [AMS15]; (**2**) Posterior density of the parameter γ_1 (arbitrary vertical scale); (**3**) Posterior density of the parameter $\delta = \gamma_2 - \gamma_1$ (arbitrary vertical scale); (4): Projection of the posterior density $p(\gamma_1, \delta)$

[AMS15]. This is just a technical detail, irrelevant for the purpose of the example. Then, assuming a Normal model for the observations we can write the posterior density

$$p(\boldsymbol{\theta}|\text{data}) = \pi(\boldsymbol{\theta}) \prod_{i=1}^{n} \exp\left\{-\frac{1}{2{\sigma_i}^2}\left(\phi_i - \phi(r_i;\boldsymbol{\theta})\right)^2\right\}$$

I have taken Normal priors with large variances ($\sigma_i >>$) for the parameters γ_1 and δ and restricted the support to $\mathcal{R}^+$ for $\{\phi_0, r_0, \sigma\}$. The posterior densities for the parameters γ_1 and δ are shown in Fig. 2.7(2, 3) together with the projection (Fig. 2.7(4)) that gives an idea of correlation between them. For a visual inspection, the phenomenological form of the flux is shown in Fig. 2.7(1) (blue line) overimposed to the data when the parameters are set to their expected posterior values.

2.13.2 Characterization of a Possible Source of Events

Suppose that we observe a particular region Ω of the sky during a time t and denote by λ the rate at which events from this region are produced. We take a Poisson model to describe the number of produced events: $k \sim Po(k|\lambda t)$. Now, denote by ϵ the probability to detect one event (detection area, efficiency of the detector,…). The number of observed events n from the region Ω after an exposure time t and detection probability ϵ will follow:

$$n \sim \sum_{k=n}^{\infty} Bi(k|n,\epsilon)\, Po(k|\lambda t) = Po(n|\lambda t \epsilon)$$

The approach to the problem will be the same for other counting process like, for instance, events collected from a detector for a given integrated luminosity. We suspect that the events observed in a particular region Ω_o of the sky are background events together with those from an emitting source. To determine the significance of the potential source we analyze a nearby region, Ω_b, to infer about the expected background. If after a time t_b we observe n_b events from this region with detection probability e_b then, defining $\beta = \epsilon_b t_b$ we have that

$$n_b \sim Po(n_b|\lambda_b\,\beta) = \exp\{-\beta\lambda_b\}\,\frac{(\beta\lambda_b)^{n_b}}{\Gamma(n_b+1)}$$

At Ω_o we observe n_o events during a time t_o with a detection probability ϵ_o. Since $n_o = n_1 + n_2$ with $n_1 \sim Po(n_1|\lambda_s\alpha)$ signal events ($\alpha = \epsilon_o t_o$) and $n_2 \sim Po(n_2|\lambda_b\alpha)$ background events (assume reasonably that $e_s = e_b = e_o$ in the same region), we have that

$$n_o \sim \sum_{n_1=0}^{n_o} Po(n_1|\lambda_s\alpha)\,Po(n_o - n_1|\lambda_b\alpha) = Po(n_o|(\lambda_s+\lambda_b)\alpha)$$

Now, we can do several things. We can assume for instance that the overall rate form the region Ω_o is λ, write $n_o \sim Po(n_o|\alpha\lambda)$ and study the fraction λ/λ_b of the rates from the information provided by the observations in the two different regions. Then, reparameterizing the model in terms of $\theta = \lambda/\lambda_b$ and $\phi = \lambda_b$ we have

$$p(n_o, n_b|\cdot) = Po(n_o|\alpha\lambda)\,Po(n_b|\beta\lambda_b) \sim e^{-\beta\phi(1+\gamma\theta)}\,\theta^{n_o}\,\phi^{n_o+n_b}$$

where $\gamma = \alpha/\beta = (\epsilon_s t_s)/(\epsilon_b t_b)$. For the ordering $\{\theta, \phi\}$ we have that the Fisher's matrix and its inverse are

$$\mathbf{I}(\theta,\phi) = \begin{pmatrix} \frac{\gamma\beta\phi}{\theta} & \gamma\beta \\ \gamma\beta & \frac{\beta(1+\gamma\theta)}{\phi} \end{pmatrix} \quad \text{and} \quad \mathbf{I}^{-1}(\mu_1,\mu_2) = \begin{pmatrix} \frac{\theta(1+\gamma\theta)}{\phi\gamma\beta} & -\frac{\theta}{\beta} \\ -\frac{\theta}{\beta} & \frac{\phi}{\beta} \end{pmatrix}$$

Then

$$\pi(\theta, \phi) = \pi(\phi|\theta)\, \pi(\theta) \;\propto\; \frac{\phi^{-1/2}}{\sqrt{\theta(1+\gamma\theta)}}$$

and integrating the nuisance parameter ϕ we get finally:

$$p(\theta|n_o, n_b, \gamma) = \frac{\gamma^{n_o+1/2}}{B(n_o+1/2, n_b+1/2)}\, \frac{\theta^{n_o-1/2}}{(1+\gamma\theta)^{n_o+n_b+1}}$$

From this:

$$E[\theta^m] = \frac{1}{\gamma^m}\, \frac{\Gamma(n_o+1/2+m)\,\Gamma(n_b+1/2-m)}{\Gamma(n_o+1/2)\,\Gamma(n_b+1/2)} \;\longrightarrow\; E[\theta] = \frac{1}{\gamma}\, \frac{n_o+1/2}{n_b-1/2}$$

and

$$P(\theta \le \theta_0) = \int_0^{\theta_0} p(\theta|\cdot)\, d\theta = 1 - IB(n_b+1/2, n_o+1/2; (1+\gamma\theta_0)^{-1})$$

with $IB(x, y; z)$ the Incomplete Beta Function. Had we interest in $\theta = \lambda_s/\lambda_b$, the corresponding reference prior will be

$$\pi(\theta, \phi) \;\propto\; \frac{\phi^{-1/2}}{\sqrt{(1+\theta)(\delta+\theta)}} \qquad \text{with} \qquad \delta = \frac{1+\gamma}{\gamma}$$

A different analysis can be performed to make inferences on λ_s. In this case, we may consider as an informative prior for the nuisance parameter the posterior what we had from the study of the background in the region Ω_b; that is:

$$p(\lambda_b|n_b,\ \beta) \;\propto\; \exp\{-\beta\lambda_b\}\ {\lambda_b}^{n_b-1/2}$$

and therefore:

$$p(\lambda_s|\cdot) \;\propto\; \pi(\lambda_s) \int_0^{\infty} p(n_o|\alpha(\lambda_s+\lambda_b))\, p(\lambda_b|n_b,\ \beta)\, d\lambda_b \propto \pi(\lambda_s)\, e^{-\alpha\lambda_s}\, \lambda_s^{n_o} \sum_{k=0}^{n_o} a_k\, {\lambda_s}^{-k}$$

where

$$a_k = \binom{n_o}{k}\, \frac{\Gamma(k+n_b+1/2)}{[(\alpha+\beta)]^k}$$

A reasonable choice for the prior will be a conjugated prior $\pi(\lambda_s) = Ga(\lambda_s|a, b)$ that simplifies the calculations and provides enough freedom analyze the effect of different shapes on the inferences. The same reasoning is valid if the knowledge on λ_b is represented by a different $p(\lambda_b|\cdot)$ from, say, a Monte Carlo simulation. Usual

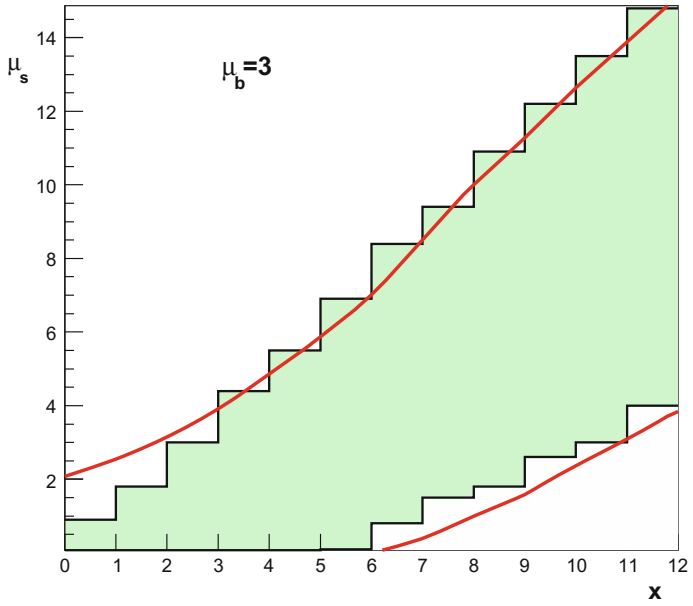

Fig. 2.8 90% Confidence Belt derived with Feldman and Cousins (*filled band*) and the Bayesian HPD region (*red lines*) for a background parameter $\mu_b = 3$

distributions in this case are the Gamma and the Normal with non-negative support. Last, it is clear that if the rate of background events is known with high accuracy then, with $\mu_i = \alpha\lambda_i$ and $\pi(\mu_s) \propto (\mu_s + \mu_b)^{-1/2}$ we have

$$p(\mu_s|\cdot) = \frac{1}{\Gamma(x + 1/2, \mu_b)} \exp\{-(\mu_s + \mu_b)\} (\mu_s + \mu_b)^{x-1/2} \mathbf{1}_{(0,\infty)}(\mu_s)$$

As an example, we show in Fig. 2.8 the 90% HPD region obtained from the previous expression (red lines) as function of x for $\mu_b = 3$ (conditions as given in the example of [25]) and the Confidence Belt derived with the Feldman and Cousins approach (filled band). In this case, $\mu_{s,m} = \max\{0, x - \mu_b\}$ and therefore, for a given μ_s:

$$\sum_{x_1}^{x_2} Po(x|\mu_s + \mu_b) = \beta \quad \text{with} \quad R(x|\mu_s) = e^{(\mu_{s,m} - \mu_s)} \left(\frac{\mu_s + \mu_b}{\mu_{s,m} + \mu_b} \right)^x > k_\beta$$

for all $x \in [x_1, x_2]$.

Problem 2.14 In the search for a new particle, assume that the number of observed events follows a Poisson distribution with $\mu_b = 0.7$ known with enough precision from extensive Monte Carlo simulations. Consider the hypothesis $H_0 : \{\mu_s = 0\}$ and $H_1 : \{\mu_s \neq 0\}$. It is left as an exercise to obtain the Bayes Factor BF_{01} with the proper

prior $\pi(\mu_s|\mu_b) = \mu_b(\mu_s + \mu_b)^{-2}$ proposed in [26], $P(H_1|n)$ and the BIC difference Δ_{01} as function of $n = 1, \ldots 7$ and decide when, based on this results, will you consider that there is evidence for a signal.

2.13.3 Anisotropies of Cosmic Rays

The angular distribution of cosmic rays in galactic coordinates is analyzed searching for possible anisotropies. A well-behaved real function $f(\theta, \phi) \in L_2(\Omega)$, with $(\theta, \phi) \in \Omega = [0, \pi] \times [0, 2\pi]$, can be expressed in the real harmonics basis as:

$$f(\theta, \phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} a_{lm} Y_{lm}(\theta, \phi) \qquad \text{where} \qquad a_{lm} = \int_{\Omega} f(\theta, \phi) Y_{lm}(\theta, \phi)\, d\mu;$$

$a_{lm} \in R$ and $d\mu = \sin\theta d\theta d\phi$. The convention adopted for the spherical harmonic functions is such that (*orthonormal basis*):

$$\int_{\Omega} Y_{lm}(\theta, \phi)\, Y_{l'm'}(\theta, \phi)\, d\mu = \delta_{ll'}\, \delta_{mm'} \qquad \text{and} \qquad \int_{\Omega} Y_{lm}(\theta, \phi)\, d\mu = \sqrt{4\pi}\, \delta_{l0}$$

In consequence, a probability density function $p(\theta, \phi)$ with support in Ω can be expanded as

$$p(\theta, \phi) = c_{00}\, Y_{00}(\theta, \phi) + \sum_{l=1}^{\infty} \sum_{m=-l}^{l} c_{lm}\, Y_{lm}(\theta, \phi)$$

The normalization imposes that $c_{00} = 1/\sqrt{4\pi}$ so we can write

$$p(\theta, \phi|\boldsymbol{a}) = \frac{1}{4\pi}\, (1 + a_{lm}\, Y_{lm}(\theta, \phi))$$

where $l \geq 1$,

$$a_{lm} = 4\pi c_{lm} = 4\pi \int_{\Omega} p(\theta, \phi)\, Y_{lm}(\theta, \phi)\, d\mu = 4\pi\, E_{p;\mu}[Y_{lm}(\theta, \phi)]$$

and summation over repeated indices understood. Obviously, for any $(\theta, \phi) \in \Omega$ we have that $p(\theta, \phi|\boldsymbol{a}) \geq 0$ so the set of parameters $\boldsymbol{a}$ are constrained on a compact support.

Even though we shall study the general case, we are particularly interested in the expansion up to $l = 1$ (dipole terms) so, to simplify the notation, we redefine the indices $(l, m) = \{(1, -1), (1, 0), (1, 1)\}$ as $i = \{1, 2, 3\}$ and, accordingly, the coefficients $\boldsymbol{a} = (a_{1-1}, a_{10}, a_{11})$ as $\boldsymbol{a} = (a_1, a_2, a_3)$. Thus:

$$p(\theta, \phi|\boldsymbol{a}) = \frac{1}{4\pi}\ (1 + a_1 Y_1 + a_2 Y_2 + a_3 Y_3)$$

In this case, the condition $p(\theta, \phi|\boldsymbol{a}) \geq 0$ implies that the coefficients are bounded by the sphere $a_1^2 + a_2^2 + a_3^2 \leq 4\pi/3$ and therefore, the coefficient of anisotropy

$$\delta \overset{def.}{=} \sqrt{\frac{3}{4\pi}}\ \left(a_1^2 + a_2^2 + a_3^2\right)^{1/2} \leq 1$$

There are no sufficient statistics for this model but the Central Limit Theorem applies and, given the large amount of data, the experimental observations can be cast in the statistic $\boldsymbol{a} = (a_1, a_2, a_3)$ such that[15]

$$p(\boldsymbol{a}|\boldsymbol{\mu}) = \prod_{i=1}^{3} N(a_i|\mu_i, \sigma_i^2)$$

with $V(a_i) = 4\pi/n$ known and with negligible correlations ($\rho_{ij} \simeq 0$).

Consider then a k-dimensional random quantity $\boldsymbol{Z} = \{Z_1, \ldots, Z_k\}$ and the distribution

$$p(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\sigma}) = \prod_{j=1}^{k} N(z_j|\mu_j, \sigma_j^2)$$

The interest is centered on the euclidean norm $||\boldsymbol{\mu}||$, with $\dim\{\boldsymbol{\mu}\} = k$, and its square; in particular, in

$$\delta = \sqrt{\frac{3}{4\pi}}||\boldsymbol{\mu}|| \quad \text{for } k = 3 \qquad \text{and} \qquad C_k = \frac{||\boldsymbol{\mu}||^2}{k}$$

First, let us define $X_j = Z_j/\sigma_j$ and $\rho_j = \mu_j/\sigma_j$ so $X_j \sim N(x_j|\rho_j, 1)$ and make a transformation of the parameters ρ_j to spherical coordinates:

$$\begin{aligned}
\rho_1 &= \rho\ \cos\phi_1 \\
\rho_2 &= \rho\ \sin\phi_1 \cos\phi_2 \\
\rho_3 &= \rho\ \sin\phi_1 \sin\phi_2 \cos\phi_3 \\
&\vdots \\
\rho_{k-1} &= \rho\ \sin\phi_1 \sin\phi_2 \ldots \sin\phi_{k-2} \cos\phi_{k-1} \\
\rho_k &= \rho\ \sin\phi_1 \sin\phi_2 \ldots \sin\phi_{k-2} \sin\phi_{k-1}
\end{aligned}$$

The Fisher's matrix is the Riemann metric tensor so the square root of the determinant is the k-dimensional volume element:

[15] Essentially, $a_{lm} = \frac{4\pi}{n} \sum_{i=1}^{n} Y_{lm}(\theta_i, \phi_i)$ for a sample of size n.

$$dV^k = \rho^{k-1} d\rho\, dS^{k-1}$$

with

$$dS^{k-1} = \sin^{k-2}\phi_1\ \sin^{k-3}\phi_2\ \ldots \sin\phi_{k-2}\, d\phi_1\, d\phi_2\ \ldots d\phi_{k-1} = \prod_{j=1}^{k-1} \sin^{(k-1)-j}\phi_j d\phi_j$$

the $k-1$ dimensional spherical surface element, $\phi_{k-1} \in [0, 2\pi)$ and $\phi_{1,\ldots,k-2} \in [0, \pi]$. The interest we have is on the parameter ρ so we should consider the ordered parameterization $\{\rho;\ \boldsymbol{\phi}\}$ with $\boldsymbol{\phi} = \{\phi_1, \phi_2, \ldots, \phi_{k-1}\}$ nuisance parameters. Being ρ and ϕ_i independent for all i, we shall consider the surface element (that is, the determinant of the submatrix obtained for the angular part) as prior density (proper) for the nuisance parameters. As we have commented in Chap. 1, this is just the Lebesgue measure on the $k-1$ dimensional sphere (the Haar invariant measure under rotations) and therefore the natural choice for the prior; in other words, a uniform distribution on the $k-1$ dimensional sphere. Thus, we start integrating the angular parameters. Under the assumption that the variances σ_i^2 are all the same and considering that

$$\int_0^\pi e^{\pm\beta\cos\theta} \sin^{2\nu}\theta d\theta = \sqrt{\pi}\left(\frac{2}{\beta}\right)^\nu \Gamma(\nu+\frac{1}{2})\, I_\nu(\beta) \qquad \text{for} \qquad Re(\nu) > -\frac{1}{2}$$

one gets $p(\phi|\text{data}) \propto p(\phi_m|\phi)\pi(\phi)$ where

$$p(\phi_m|\phi, \nu) = b\, e^{-b(\phi+\phi_m)} \left(\frac{\phi_m}{\phi}\right)^{\nu/2} I_\nu(2b\sqrt{\phi_m}\sqrt{\phi})$$

is properly normalized,

$$\nu = k/2 - 1\ ; \qquad \phi = ||\boldsymbol{\mu}||^2; \qquad \phi_m = ||\boldsymbol{a}||^2; \qquad b = \frac{1}{2\sigma^2} = \frac{n}{8\pi}$$

and $\dim\{\boldsymbol{\mu}\} = \dim\{\boldsymbol{a}\} = k$. This is nothing else but a non-central χ^2 distribution.

From the series expansion of the Bessel functions it is easy to prove that this process is just a compound Poisson-Gamma process

$$p(\phi_m|\phi, \nu) = \sum_{k=0}^{\infty} Po(k|b\phi)\, Ga(\phi_m|b, \nu+k+1)$$

and therefore the sampling distribution is a Gamma-weighted Poisson distribution with the parameter of interest that of the Poisson. From the Mellin Transform:

$$\mathcal{M}(s)_{\langle -\nu,\infty\rangle} = \frac{b\, e^{-b\phi}}{\Gamma(\nu+1)} \frac{\Gamma(s+\nu)}{b^s} M(s+\nu, \nu+1, b\phi)$$

with $M(a, b, z)$ the Kummer's function one can easily get the moments ($E[\phi_m^n] = M(n+1)$); in particular

$$E[\phi_m] = \phi + b^{-1}(\nu+1) \qquad \text{and} \qquad V[\phi_m] = 2\phi b^{-1} + b^{-2}(\nu+1)$$

Now that we have the model $p(\phi_m|\phi)$, let's go for the prior function $\pi(\phi)$ or $\pi(\delta)$. One may guess already what shall we get. The first element of the Fisher's matrix (diagonal) corresponds to the norm and is constant so it would not be surprising to get the Lebesgue measure for the norm $d\lambda(\delta) = \pi(\delta)d\delta = c\, d\delta$. As a second argument, for large sample sizes ($n >>$) we have $b >>$ so $\phi_m \sim N(\phi_m|\phi, \sigma^2 = 2\phi/b)$ and, to first order, Jeffreys' prior is $\pi(\phi) \sim \phi^{-1/2}$. From the reference analysis, if we take for instance

$$\pi^\star(\phi) = \phi^{(\nu-1)/2}$$

we end up, after some algebra, with

$$\pi(\phi) \propto \pi(\phi_0) \lim_{k\to\infty} \frac{f_k(\phi)}{f_k(\phi_0)} \propto \left(\frac{\phi_0}{\phi}\right)^{1/2} \lim_{b\longrightarrow\infty} e^{-3b(\phi-\phi_0)/2 + [I(\phi,b) - I(\phi_0,b)]}$$

where

$$I(\phi, b) = \int_0^\infty p(\phi_m|\phi)\, \log \frac{I_\nu(2b\sqrt{\phi\phi_m})}{I_{\nu/2}(b\phi_m/2)} d\phi_m$$

and ϕ_0 any interior point of $\Lambda(\phi) = [0, \infty)$. From the asymptotic behavior of the Bessel functions one gets

$$\pi(\phi) \propto \phi^{-1/2}$$

and therefore, $\pi(\delta) = c$. It is left as an exercise to get the same result with other priors like $\pi^\star(\phi) = c$ or $\pi^\star(\phi) = \phi^{-1/2}$.

For this problem, it is easier to derive the prior from the reference analysis. Nevertheless, the Fisher's information that can be expressed as:

$$F(\phi;\nu) = b^2 \left\{ -1 + b\frac{e^{-b\phi}}{\phi^{\nu/2+1}} \int_0^\infty e^{-bz}\, z^{\nu/2+1} \frac{I_{\nu+1}^2(2b\sqrt{z\phi})}{I_\nu(2b\sqrt{z\phi})} dz \right\}$$

and, for large b (large sample size), $F(\lambda;\nu) \to \phi^{-1}$ *regardless the number of degrees of freedom* ν. Thus, Jeffrey's prior is consistent with the result from reference analysis. In fact, from the asymptotic behavior of the Bessel Function in the corresponding expressions of the pdf, one can already see that $F(\phi;\nu) \sim \phi^{-1}$. A cross check from a numeric integration is shown in Fig. 2.9 where, for $k = 3, 5, 7$ ($\nu = 1/2, 3/2, 5/2$), $F(\phi;\nu)$ is depicted as function of ϕ compared to $1/\phi$ in black for a sufficiently large

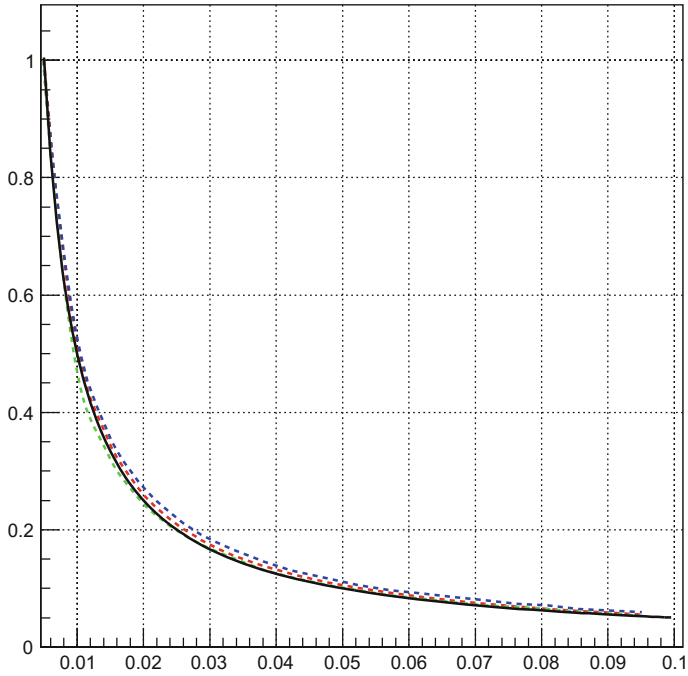

Fig. 2.9 Fisher's information (numeric integration) as function of ϕ for $k = 3, 5, 7$ (*discontinuous lines*) and $f(\phi) = \phi^{-1}$ (*continuous line*). All are scaled so that $F(\phi = 0.005, \nu) = 1$

value of b. Therefore we shall use $\pi(\phi) = \phi^{-1/2}$ for the cases of interest (dipole, quadrupole, … any-pole).

The posterior densities are

• For $\phi = ||\boldsymbol{\mu}||^2$: $\quad p(\phi|\phi_m, \nu) = N\, e^{-b\phi}\, \phi^{-(\nu+1)/2}\, I_\nu(2b\sqrt{\phi_m}\sqrt{\phi})$ with

$$N = \frac{\Gamma(\nu+1)\, b^{1/2-\nu}\, \phi_m^{-\nu/2}}{\sqrt{\pi}\, M(1/2, \nu+1, b\phi_m)}$$

The Mellin Transform is

$$\mathcal{M}_\phi(s)_{\langle 1/2,\infty\rangle} = \frac{\Gamma(s-1/2)}{b^{s-1}\sqrt{\pi}}\, \frac{M(s-1/2, \nu+1, b\phi_m)}{M(1/2, \nu+1, b\phi_m)}$$

and therefore the moments

$$E[\phi^n] = M(n+1) = \frac{\Gamma(n+1/2)\, M(n+1/2, \nu+1, b\phi_m)}{\sqrt{\pi}\, b^n\, M(1/2, \nu+1, b\phi_m)}$$

In the limit $|b\phi_m| \to \infty$, $E[\phi^n] = \phi_m^n$.

• For $\rho = ||\boldsymbol{\mu}|| :\quad p(\rho|\phi_m, \nu) = 2\, N\, e^{-b\rho^2}\, \rho^{-\nu}\, I_\nu(2b\sqrt{\phi_m}\rho)$ and

$$\mathcal{M}_\rho(s) = \mathcal{M}_\phi(s/2 + 1/2) \;\longrightarrow\; E[\rho^n] = \frac{\Gamma(n/2 + 1/2)\, M(n/2 + 1/2, \nu + 1, b\phi_m)}{\sqrt{\pi}\, b^{n/2}\, M(1/2, \nu + 1, b\phi_m)}$$

In the particular case that $k = 3$ (dipole; $\nu = 1/2$), we have for $\delta = \sqrt{3/4\pi}\rho$ that the first two moments are:

$$E[\delta] = \frac{\text{erf}\,(z)}{a\delta_m M(1, 3/2, -z^2)} \qquad E[\delta^2] = \frac{1}{a\, M(1, 3/2, -z^2)}$$

with $z = 2\delta_m\sqrt{b\pi/3}$ and, when $\delta_m \to 0$ we get

$$E[\delta] = \sqrt{\frac{2}{\pi a}} \simeq \frac{1.38}{\sqrt{n}} \qquad E[\delta^2] = \frac{1}{a} \qquad \sigma_\delta \simeq \frac{1.04}{\sqrt{n}}$$

and a one sided 95% upper credible region (see Sect. 2.11 for more details) of $\delta_{0.95} = \dfrac{3.38}{\sqrt{n}}$.

So far, the analysis has been done assuming that the variances σ_j^2 are of the same size (equal in fact) and the correlations are small. This is a very reasonable assumption but may not always be the case. The easiest way to proceed then is to perform a transformation of the parameters of interest ($\boldsymbol{\mu}$) to polar coordinates $\boldsymbol{\mu}(\rho, \Omega)$ and do a Monte Carlo sampling from the posterior:

$$p(\rho, \Omega|z, \boldsymbol{\Sigma}^{-1}) \;\propto\; \left[\prod_{j=1}^{n} N(z_j|\mu_j(\rho, \Omega), \boldsymbol{\Sigma}^{-1})\right] \pi(\rho) d\rho\, dS^{n-1}$$

with a constant prior for δ or $\pi(\phi) \propto \phi^{-1/2}$ for ϕ.

References

1. G. D'Agostini, *Bayesian Reasoning in Data Analysis* (World Scientific Publishing Co, Singapore, 2003)
2. F. James, *Statistical Methods in Experimental Physics* (World Scientific Publishing Co, Singapore, 2006)
3. J.M. Bernardo, The concept of exchangeability and its applications. Far East J. Math. Sci. **4**, 111–121 (1996). www.uv.es/~bernardo/Exchangeability.pdf
4. J.M. Bernardo, A.F.M. Smith, *Bayesian Theory* (Wiley, New York, 1994)
5. R.E. Kass, L. Wasserman, The selection of prior distributions by formal Rules. J. Am. Stat. Assoc. V **91**(453), 1343–1370 (1996)

6. H. Jeffreys, *Theory of Probability* (Oxford University Press, Oxford, 1939)
7. E.T. Jaynes, *Prior Probabilities and Transformation Groups*, NSF G23778 (1964)
8. V.I. Bogachev, *Measure Theory* (Springer, Berlin, 2006)
9. M. Stone, Right haar measures for convergence in probability to invariant posterior distributions. Ann. Math. Stat. **36**, 440–453 (1965)
10. M. Stone, Necessary and sufficient conditions for convergence in probability to invariant posterior distributions. Ann. Math. Stat. **41**, 1349–1353 (1970)
11. H. Raiffa, R. Schlaifer, *Applied Statistical Decision Theory* (Harvard University Press, Cambridge, 1961)
12. S.R. Dalal, W.J. Hall, J.R. Stat, Soc. Ser. B **45**, 278–286 (1983)
13. B. Welch, H. Pears, J.R. Stat, Soc. B **25**, 318–329 (1963)
14. M. Gosh, R. Mukerjee, Biometrika **84**, 970–975 (1984)
15. G.S. Datta, M. Ghosh, Ann. Stat. **24**(1), 141–159 (1996)
16. G.S. Datta, R. Mukerjee, *Probability Matching Priors and Higher Order Asymptotics* (Springer, New York, 2004)
17. J.M. Bernardo, J.R. Stat, Soc. Ser. B **41**, 113–147 (1979)
18. J.O. Berger, J.M. Bernardo, D. Sun, Ann. Stat **37**(2), 905–938 (2009)
19. J.M. Bernardo, J.M. Ramón, The Statistician **47**, 1–35 (1998)
20. J.O. Berger, J.M. Bernardo, D. Sun, Objective priors for discrete parameter spaces. J. Am. Stat. Assoc. **107**(498), 636–648 (2012)
21. A. O'Hagan, J.R. Stat. Soc. **B57**, 99–138 (1995)
22. J.O. Berger, L.R. Pericchi, J. Am. Stat. Assoc. V **91**(433), 109–122 (1996)
23. R.E. Kass, A.E. Raftery, J. Am. Stat. Assoc. V **90**(430), 773–795 (1995)
24. G. Schwarz, Ann. Stat. **6**, 461–464 (1978)
25. Feldman G.J. and Cousins R.D. (1997); arXiV:physics/9711021v2
26. J.O. Berger, L.R. Pericchi, Ann. Stat. V **32**(3), 841–869 (2004)

http://www.springer.com/978-3-319-55737-3

Probability and Statistics for Particle Physics

Maña, C.

2017, X, 244 p. 27 illus., 11 illus. in color., Hardcover
ISBN: 978-3-319-55737-3