3 Prior Distributions

There are several different approaches to specification of prior distributions. We can identify at least three different types of specification:

1. Subjective specification: the prior is chosen by making a serious attempt to quantify a state of knowledge prior to performing the experiment. This is perhaps the "purest" form of Bayesian analysis. This approach is most often used when the prior actually contains non-trivial information that helps inform the analysis.

2. Semi-subjective specification (this is my own term; I think these kinds of priors are widely used, but it isn't a concept that is talked about much) : here the prior isn't actually claimed to reflect the full state of knowledge prior to the experiment, but is chosen to have the property that the resulting *posterior* is nonetheless expected to be a very good approximation to what would have been obtained if a full subjective analysis had been undertaken. We saw an example of this in Savage's potato: using a very flat prior on the potato weight is not a sensible reflection of the "true" prior state of knowledge, but – as Savage argues – it will result in a similar posterior, provided that the "true" prior is approximately flat in the area with appreciable likelihood (and nowhere else much larger). Data-dependent priors (a prior distribution that has been chosen after "peeking" at the data) *can* fall into this category, although not always.

3. Objective specification: the prior is chosen to satisfy some objective criterion. For example, the prior is chosen so that the conclusions of the analysis (i.e. posterior distributions of quantities of interest) will be invariant to choice of parameterization, or to the scale of measurement of the data. Or the prior could be chosen so that (perhaps asymptotically) the resulting posterior has good frequentist properties, such as the correct frequentist coverage for 95% CIs.

In all three cases some consideration may be given to computational convenience. For example, one might take a subjective approach, but restricting oneself to conjugate priors.

Note that these approaches are not mutually exclusive. For example, I might use an Objective specification for parameters where the conclusions are expected to be relatively robust to prior specification (or where I think that they also serve as a semi-subjective prior), but a subjective prior for priors that are going to substantively affect conclusions.

What about performing analyses using more than one prior? Strict subjective Bayesians could argue that one should simply do the analysis with the prior distribution that reflects your prior knowledge/uncertainty. However, it seems to me perfectly reasonable to ask whether individuals with different prior beliefs will come to different conclusions. Thus, in situations where it is plausible that different individuals might have quite different prior beliefs it seems reasonable to report results of multiple analyses using different priors.

A question for thought/discussion: Suppose that an Objective prior specification procedure resulted in a prior that was not either subjective or semi-subjective. Would you be happy to use it anyway? What would the resulting posterior distributions mean? What does a 95% credible interval computed from an "objective prior" mean? (For a subjective prior, it has a relatively straightforward interpretation as representing a certain state of uncertainty in the parameter after viewing data.)

3.1 Subjective Determination of the Prior

In the subjective Bayesian world, the prior distribution is intended to capture the information available about the parameter before the data is observed. We have discussed how use of probability distributions to represent uncertainty, and updating beliefs via Bayes Theorem, leads to coherent, quantitative, belief systems. However, there remains the fact that specifying a prior distribution that encapsulates (even approximately) the appropriate beliefs can be challenging.

When the parameter space is discrete, the specification of a prior distribution boils down to simply comparing a finite number of options. However, if the number of options is sufficiently large it will be impractical to actually compare them all, in which case it is usually helpful to make use of parametric models or make simplifying assumptions (see for example the variable selection in regression example below). Similarly, when the parameter space is continuous, it is usually helpful to restrict oneself to a particular parametric form, and then choose the parameters so that the resulting prior distribution approximates prior beliefs, perhaps by matching moments or (for more robustness) percentiles to prior assessments (see for example the "relative risk" example below).

Examples:

1. Law enforcement agencies are interested in classifying African Elephant ivory as being from Forest or Savanna elephants (two subspecies of African elephant). What is your

prior that a tusk is from Forest? Note that, of course, there is no "right" answer to this question. Let's consider a series of possible situations.

- If we knew nothing more about the problem we might say 0.5 for each. (One possible argument for this is "symmetry", which could be seen as an "Objective" argument. Alternatively we could see it as a directly representation of our subjective knowledge.)
- If the police told us that they had reason to believe that the tusk was from a Forest elephant, because most poaching comes from forest elephants, we might decide to select a prior that favored forest elephants slightly, eg 2/3 to 1/3. On the other hand, we might prefer to stick with the 50-50 prior, to see what the DNA data say in the absence of the police's prior beliefs. If our analysis with a 50-50 prior concluded that the tusk was, with probability 0.99, from savannah, and the police (or someone else) were skeptical of this conclusion, then we might ask the question, "how strong would the prior belief in the forest have to be to outweigh the data?" [*Exercise!*] Note that this might be a general reason to consider priors that do not necessarily represent our actual prior beliefs: to convince a skeptic! In this example, Bayes Factors come in handy for seeing how the prior and data combine.
- Suppose we were considering a whole series of tusks. Let's say there were 100 tusks. Further suppose, for the sake of argument, that the first 99 all turned out to be very clearly from forest elephants on the basis of their DNA. For the last tusk, maybe we would be inclined to use a prior that strongly favored forest: perhaps even something as strong as 0.99 to 0.01? Note: despite the apparent over-simplicity of this example, this kind of "sharing of information" or "borrowing information" across units is a very powerful and widely-used idea. In practice it is usually done more formally, by use of hierarchical models, as we shall see later.

2. At a C/G single nucleotide polymorphism (SNP), there are three possible genotypes ("classes") for each individual: CC,CG and GG. Consider specifying a prior distribution for the frequencies $f = (f_{CC}, f_{CG}, f_{GG})$ for a sample of individuals from China.

   - Someone who knew nothing more about this might start with a uniform prior on f (subject to the constraint that $f_{CC} + f_{CG} + f_{GG} = 1$). Indeed, even someone who knows quite a bit more might use this prior if they had lots of data from which they were about to compute a posterior, and figured that it was not worth worrying too much about more careful prior specification. (In the first case one might view this as a subjective prior; in the second a "semi-subjective" prior).

- Someone who knows a bit about population genetics might decide to use the assumption of "Hardy–Weinberg Equilibrium", which is effectively the assumption that the two alleles carried by each individual are independent draws from some distribution. Thus, if f_C denotes the overall frequency of C in the population, then $f = (f_C^2, 2f_C(1-f_C), (1-f_C)^2)$. This assumption effectively reduces the vector parameter f to a single parameter f_C, so putting a prior on f can now be achieved by putting a prior on f_C. Perhaps we might choose a uniform prior on $[0, 1)$ for f_C.

- Someone who knows a bit more about population genetics might know that, both from mathematical models and in actual data, allele frequencies like f_C tend not to be uniform on $[0, 1)$, but more skewed towards the ends 0 and 1. Perhaps they would use a Beta(0.5,0.5) or Beta(0.1,0.1) distribution for f_C. Better still, perhaps they would go out and gather frequency data on, say, all known SNPs in Chinese, and fit a Beta(α, β) distribution to those data, and use that as a prior for this new SNP. (Note that this strategy is "borrowing information" across SNPs.)

- Finally, suppose that from a large data set on Japanese individuals we know that the frequencies of the three types in Japan are $g = (g_{CC}, g_{CG}, g_{GG})$. Perhaps we could assume a Dirichlet prior for f (this is the conjugate prior for multinomial sampling) in which $E(f) = g$. Thus $f \sim Dir(\alpha g_{CC}, \alpha g_{CG}, \alpha g_{GG})$, where α controls how similar we expect f to be to g, which perhaps again we could get some idea from using data on other SNPs. (Note that this strategy is "borrowing information" across both SNPs and populations).

3. Consider specifying a prior for the "relative risk" (RR) of having a disease given that you carry one of two possible genetic types A or a. Note $RR(A \text{ vs } a) := p(\text{disease}|A)/p(\text{disease}|a)$.

   For most common diseases, such as heart disease, where the factors affecting risk are many and complex, values of RR for a genetic factor are typically small (close to 1): for example, values > 1.2 are generally considered uncommon. Also, the relative risk of A vs a, is the inverse of the relative risk of a vs A. So, unless we have particular reason to distinguish between A and a, our prior on 1/RR should be the same as our prior on RR. One possible choice would be to put a Normal prior on $\log(\text{RR}) \sim N(0, \sigma^2)$, with σ^2 chosen so that $\Pr(\text{RR} > 1.2) \approx 0.01$ (or 0.05? or 0.001?). Note: the normal distribution is a little inflexible, and in particular has tails that decay very quickly. This limits its ability to capture particular prior beliefs. One can get more flexibility by using a t distribution, or (sometimes more conveniently, and almost equivalently) a finite mixture of normals with different variances. See Stephens and Balding (Nature Reviews Genetics, 2009) for further discussion.

Two Rules of Thumb

1. If putting a prior on a parameter that is really a ratio of two symmetric quantities (e.g. the RR above, or the ratio of the expression level of gene A to gene B, where no information is available a priori to distinguish A and B) then take the log of the parameter, and put a symmetric (perhaps normal, with some variance?) prior centered on 0.

2. If putting a prior on a small positive quantity, about which uncertainty spans orders of magnitude, put a prior (perhaps uniform, on some range, say?) on the log of the quantity, and not on the quantity itself. Example: imagine placing a prior on the probability p that an aircraft you are about to board will crash. You decide that, for sure $p < 0.001$, but much more you are not prepared to say. This leads you to consider putting a $U[0, 0.001)$ prior on p. But I would argue that this is a mistake - why?

Prior Elicitation

In many cases a statistician may be in the position of attempting to obtain a prior that captures *someone else's* subjective beliefs. Usually this someone else would be a subject matter expert, who is not a statistician. The process of obtaining priors from subject matter experts is known as *Prior Elicitation*, and is a research topic of its own. See for example the book "Uncertain Judgements: Eliciting Experts' Probabilities", published by Wiley.

3.2 Conjugate Priors

A family $\mathcal{F}$ of prior distributions for θ is said to be **closed under sampling** from a model $f(x \mid \theta)$ if for every prior distribution $f(\theta) \in \mathcal{F}$, the posterior distribution

$$f(\theta \mid x) \propto f(\theta) f(x \mid \theta)$$

is also in $\mathcal{F}$. (It follows, e.g. by iteration, that the posterior from a random sample $x_1, \cdots, x_n$, all with density $f(x \mid \theta)$ will be in $\mathcal{F}$.)

Recall for instance, if X is $N(\theta, \phi)$ given the values of the parameters θ and ϕ. Suppose the variance is known and take a prior for θ which is $N(\theta_0, \phi_0)$. We saw that the posterior for

θ was also Normal. That is, the normal distribution is the conjugate prior for the mean of a normal distribution (with known variance).

Definition: Suppose $x_1, \cdots, x_n$ is a random sample from a regular k-parameter exponential family,

$$f(x \mid \theta) = f(x)g(\theta) \exp\left(\sum_{i=1}^{k} c_i \phi_i(\theta) h_i(x)\right).$$

Then the prior distribution for θ of the form

$$p(\theta \mid \tau) = [K(\tau)]^{-1}[g(\theta)]^{\tau_0} \exp\left(\sum_{i=1}^{k} c_i \tau_i \phi_i(\theta)\right),$$

where τ is such that

$$K(\tau) = \int_{\Theta} g(\theta)^{\tau_0} \exp\left(\sum_{i=1}^{k} c_i \tau_i \phi_i(\theta)\right) d\theta < \infty,$$

is said to be a *conjugate prior.*

Note that the conjugate prior is a distribution for θ. The parameters of the prior, τ, are often referred to as *hyperparameters*. We will prove below that these conjugate priors are closed under sampling.

Examples:

1. Bernoulli Likelihood. Recall that $f(x) = 1$, $g(\theta) = 1 - \theta$, $h(x) = x$, $\phi(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$, $c = 1$. Thus,

$$p(\theta \mid \tau_0, \tau_1) \propto (1-\theta)^{\tau_0} \exp\left(\log(\frac{\theta}{1-\theta})\tau_1\right)$$

$$= \frac{1}{K(\tau_0, \tau_1)} \theta^{\tau_1}(1-\theta)^{\tau_0 - \tau_1}$$

The density will have finite integral iff $\tau_1 > 1$ and $\tau_0 - \tau_1 > 1$, in which case this is the Beta$(\tau_1 + 1, \tau_0 - \tau_1 + 1)$ distribution.

2. Poisson Likelihood. Recall $f(x) = (x!)^{-1}$, $g(\theta) = \exp -\theta$, $h(x) = x$, $\phi(\theta) = \log\theta$, $c = 1$. The conjugate prior is

$$p(\theta \mid \tau_0, \tau_1) \propto \exp(-\tau_0\theta)\exp(\tau_1 \log(\theta))$$
$$= \frac{1}{K(\tau_0, \tau_1)} \theta^{\tau_1} e^{-\tau_0\theta}$$

This is the Gamma$(\tau_1 + 1, \tau_0)$ distribution, assuming that $\tau_1 + 1 > 0$ and $\tau_0 > 0$.

3. Normal Likelihood. For convenience, we write $\lambda = 1/\sigma^2$ for the *precision* of the normal model, and μ for the mean:

$$p(x \mid \mu, \lambda) = (\frac{\lambda}{2\pi})^{1/2} \exp\left(-\frac{\lambda}{2}(x-\mu)^2\right)$$

Note here that $f(x) = 1/\sqrt{2\pi}$, $g(\mu,\lambda) = \sqrt{\lambda}\exp(-\lambda\mu^2/2)$, $h(x) = (x, x^2)$, $\phi(\mu,\lambda) = (\mu\lambda, \lambda)$, $c_1 = 1$, $c_2 = -1/2$. The conjugate prior is then

$$p(\mu, \lambda \mid \tau_0, \tau_1, \tau_2) \propto \left(\sqrt{\lambda}\exp(-\lambda\mu^2/2)\right)^{\tau_0} \exp\left(\mu\lambda\tau_1 - \frac{1}{2}\lambda\tau_2\right)$$

provided τ_0, τ_1, τ_2 are chosen so that this has finite integral.

Thus

$$p(\mu, \lambda \mid \tau_0, \tau_1, \tau_2) \propto \lambda^{\frac{\tau_0+1}{2}-1} \exp\left(-\frac{1}{2}(\tau_2 - \frac{\tau_1^2}{\tau_0})\lambda\right) \lambda^{1/2} \exp\left(-\frac{\lambda\tau_0}{2}(\mu - \frac{\tau_1}{\tau_0})^2\right)$$

On inspection, this has the form of a Gamma$(\frac{1}{2}(\tau_0+1), \frac{1}{2}(\tau_2 - \frac{\tau_1^2}{\tau_0}))$ density for λ and conditional on this a $N(\frac{\tau_1}{\tau_0}, \frac{1}{\lambda\tau_0})$ density for μ. This is sometimes called a *normal-gamma* distribution for (μ, λ). The integral will be finite and hence the prior sensible, provided $\tau_2 > \frac{\tau_1^2}{\tau_0}$, $\tau_0 > 0$.

Proposition: Consider a regular k-parameter exponential family model. The family of conjugate priors defined above for the model is closed under sampling. In fact, writing $p(\theta \mid \tau_0, \tau_1, \ldots, \tau_k)$ for the parameterized conjugate prior:

1) The posterior density for θ is

$$p(\theta \mid x_1, \ldots, x_n, \tau_0, \ldots, \tau_k) = p(\theta \mid \tau_0 + n, \tau_1 + \sum_j h_1(x_j), \ldots, \tau_k + \sum_j h_k(x_j))$$

so we get the same parametric form as the prior, but with parameters $\tau_0 + n$, $\tau_1 + \sum_j h_1(x_j)$, $\ldots$, $\tau_k + \sum_j h_k(x_j)$.

2) The predictive density for future observables $y = (y_1, \cdots, y_m)$ is

$$p(y \mid x_1, \cdots, x_n, \tau_0, \cdots, \tau_k) = p(y \mid \tau_0 + n, \tau_1 + \sum_j h_1(x_j), \ldots, \tau_k + \sum_j h_k(x_j))$$

$$= \prod_{l=1}^{m} f(y_l) \frac{K(\tau_0 + n + m, \tau_1 + \sum h_1(x_j) + \sum h_1(y_l), \ldots, \tau_k + \sum h_k(x_j) + \sum h_k(y_l))}{K(\tau_0 + n, \tau_1 + \sum h_1(x_j), \ldots, \tau_k + \sum h_k(x_j))}.$$

Remarks:

1. Recall that the sufficient statistics for the model are

$$\Big[n, \sum h_1(x_j), \ldots, \sum h_k(x_j)\Big] = [t_0, \ldots, t_k]$$

2. The inference process takes a very simple form. The effect of data $x_1, \cdots, x_n$ is that the labelling parameters of the posterior are changed from those of the prior, $(\tau_0, \ldots, \tau_k)$, simply by the the addition of the sufficient statistics to give parameters $t_0 + \tau_0, t_1 + \tau_1, \ldots, t_k + \tau_k$, for the posterior. The effect of $x_1, \ldots, x_n$ on the predictive distribution is similar.

Proof of the Proposition:

By Bayes Theorem,

$$\begin{aligned} & p(\theta \mid x_1, \ldots, x_n, \tau_0, \ldots, \tau_k) \propto p(x_1, \ldots, x_n \mid \theta) p(\theta \mid \tau_0, \ldots, \tau_k) \\ = \; & \prod_{i=1}^{n} f(x_i)[g(\theta)]^n \exp\left(\sum_{i=1}^{k} c_i \phi_i(\theta) (\sum_{j=1}^{n} h_i(x_j))\right) [K(\tau)]^{-1} [g(\theta)]^{\tau_0} \exp\left(\sum_{i=1}^{k} c_i \phi_i(\theta) \tau_i\right) \\ \propto \; & [g(\theta)]^{n+\tau_0} \exp\left(\sum_{i=1}^{k} c_i \phi_i(\theta) (\tau_i + \sum_{j=1}^{k} h_i(x_j))\right). \end{aligned}$$

This is of the form $p(\theta \mid \tau_0 + n, \tau_1 + \sum h_1(x_j), \ldots, \tau_k + \sum h_k(x_j))$.

For the second part,

$$\begin{aligned} & p(y \mid x_1, \ldots, x_n, \tau_0, \ldots, \tau_k) = \int_\Theta p(y \mid \theta) p(\theta \mid x) d\theta \\ = \; & \prod_{l=1}^{m} f(y_l) \Big(K(\tau_0 + n, \tau_1 + \sum h_1(x_j), \ldots, \tau_k + \sum h_k(x_j))\Big)^{-1} \\ & \times \int_\Theta [g(\theta)]^{\tau_0+m+n} \exp\left(\sum_{i=1}^{k} c_i \phi_i(\theta) \big(\tau_i + \sum_{j=1}^{n} h_i(x_j) + \sum_{l=1}^{m} h_i(y_l)\big)\right) d\theta, \end{aligned}$$

as required.

The use of conjugate priors simplifies Bayesian calculations. They should be thought of as convenient tools. It may be that in a particular problem, one's prior beliefs will be well-approximated by a certain member of the appropriate family of conjugate priors. In fact, mixtures of conjugate priors also lead to a simple analysis (*Exercise*).

3.3 Objective Priors

What's in a name?

Many different terms are used to refer to priors that are chosen for reasons other than strictly representing your (or someone else's) prior uncertainty. Objective is one name, chosen no-doubt because it sounds, well, objective. Other related terms used include "non-informative", and "default" priors. The term "non-informative prior" has been used for a while, but I am not keen on it because in general I think there is no such thing as a "non-informative" prior - the result can always change with a different prior, so in that sense the prior is always "informative". Objective is arguably a better term, but it could be accused of hiding the fact that whether or not a particular objective prior is appropriate for a given application may be subjective! For example, suppose you derive a prior on the basis of its being invariant to certain transformations of the parameter space and/or data. It may be a subjective (or context-dependent) question whether this invariance is desirable in a given context! I like "default" best, as I think it captures what we are trying to achieve. (Imagine distributing a software package for Bayesian analysis of linear regression; what defaults would you have for the prior on the regression parameters? You could give no defaults, but then no-one would use your package....!) But Objective is perhaps the one most modern researchers working in this area use (there are conferences now on "Objective Bayes").

Some other terms that you may come across are "Jeffrey's Priors" and "Reference Priors". These are *particular* approaches to obtaining Objective priors. Jeffrey's priors are due to Harold Jeffreys, and we will look at them in more detail. The term Reference Priors is due to José Bernardo, and have also been worked on by others, including Jim Berger.

People also use terms like "vague prior", "diffuse prior", or "flat prior". I tend to think of these as descriptive terms rather than technical terms. Usually it suggests to me that the authors think that that results will be robust to this choice (e.g. because the data will be highly informative about this parameter), and so did not bother to think too hard about exactly what prior should be chosen. (So if I see those terms, and think that the results will not be robust, then I worry.)

Finally, another related term you need to be familiar with is "improper prior". This is used to refer to a function $p(\theta)$ that does not integrate to a finite quantity (so it is not a distribution, never mind a prior distribution), but which is used in place of the prior distribution in Bayes Theorem, to compute the posterior by $p(\theta|D) \propto p(D|\theta)p(\theta)$. The idea is that even if $p(\theta)$ does not integrate to a finite quantity, if $\int p(D|\theta)p(\theta)d\theta$ is finite then the above produces a "proper" posterior.

1. Improper priors often result from attempts to define Objective priors. See, for example, below.

2. Often, Bayesian analysis with an improper prior can lead to procedures with attractive frequentist properties. In fact, many classical procedures and estimators correspond to Bayesian analysis with particular improper priors.

3. Improper priors can never be "subjective" priors, but they can be "semi-subjective" (e.g. Savage's potato). That is, the posterior that arises from an improper prior may be viewed as a good approximation of a "proper" analyses.

So is it OK to use an improper prior? A strict subjectivist might argue that it is never OK, because it cannot represent your prior uncertainty. Certainly improper priors *can* lead to paradoxes and inconsistencies, which proper priors never do (because they obey the axioms of probability). However, in many cases an improper prior produces a posterior that is a good approximation to the posterior that one would have got with a carefully-constructed subjective proper prior. In these cases (as in the semi-subjective prior above) I would argue it is OK. However, it is wise to be careful before using an improper prior. *In particular you should check that your posterior distribution is proper, as this is not always the case!*

Example: Improper uniform prior for a normal mean

If $x \sim N(\mu, 1)$, and we use an improper prior, $p(\mu) \propto 1$, then show the posterior $p(\mu|x)$ is proper (for any x). Furthermore, this posterior is a reasonable approximation to the posterior you would have gotten with any prior that is "reasonably flat near x". (cf Savage's potato).

Example: The problem with improper priors for mixtures

Consider an observation x from an equal mixture of two normal distributions, one with unknown mean: $x \sim 0.9N(\mu, 1) + 0.1N(0, \sigma^2 = 100)$. (Although this is a toy example, a possible motivation for something along these lines would be to do "robust" estimation of μ, where the second component is to allow for 10% of the observations to be "outliers".) Show that attempting to use an improper prior for μ does not work (i.e. it leads to an improper posterior).

There is no such thing as an non-informative prior!

Before we look at some ways to define Objective/non-informative priors, we should make it clear that in realistic settings there are almost always subjective issues that need addressing.

Consider, for example, the simplest case of a finite parameter space Θ, with $|\Theta| = n$. The "obvious" Objective prior may seem to be to place mass n^{-1} at each possible parameter value. However, when the parameter space has "structure" (e.g. some values of the parameter are more similar to one another than others) then this can be a lot less objective, and a lot less sensible, than it might seem. For example, consider an undirected graph with V vertices. Then the number of possible edges is $E = \binom{V}{2}$, and the number of possible graphs is $G = 2^E$. Now consider putting a prior on graphs. The uniform prior on graphs would put mass $1/2^E$ on each possible graph. This is the same as each edge being present, independently, with probability 0.5. So the number of actual edges present is, a priori, Binomial$(E, 0.5)$. If E is large, as it usually is in this setting, then this prior is very concentrated near $E/2$. So this prior says that, with high probability, close to a half of the possible edges will be present - not very Objective at all (and in many settings not very sensible either - for example, in many applications one expects to see "sparse" graphs).

An alternative, and (in my subjective opinion, generally more sensible) prior here is to assume that each possible edge is present, independently, with probability p, and then to place a Be$(1, 1)$ or Be$(0.5, 0.5)$ prior on p. Alternatively, and equivalently, assume that the number of edges is uniform on 0 to E, and that, conditional on the number of edges, all graphs are equally likely. Note that in some contexts this prior will also not be appropriate. For example, in practice, one often expects some nodes of the graph to be connected to a lot of neighbors, while others will be connected to none. So maybe edge i, j should be present with probability p_{ij} given by $\log[p_{ij}/(1 - p_{ij})] = \mu + \alpha_i + \alpha_j$ where α_i and μ now need some prior specifying.....

This example shows how when attempting to be "uninformative" for one feature (the actual graph) one can easily, unintentionally, be extremely informative about another feature (the number of edges in the graph). Indeed, in general, in complex problems, the first aim of prior specification could be viewed as trying to find a prior that does not have unintended, undesirable, implications.

Bartlett's paradox: problems with non-informative priors and Model choice

When you want to compare models, it can be crucial to give serious consideration to suitable prior specification. In even simple cases there is a subtle trap ready for the unwary who are ready to simply use a "convenient" prior without thinking about it too much, sometimes known as "Bartlett's paradox". (Although it is not a true paradox, it is an unexpected result for many.)

To illustrate the idea, considering $x_1, \dots, x_n \sim N(\mu, 1)$ and consider comparing the hypothesis $H_0 : \mu = 0$ with the alternative $H_1 : \mu \neq 0$. To do this is a Bayesian framework we would place priors on H_0 and H_1 (π_0 and π_1 say), and compute the Bayes Factor for H_0 vs H_1, BF= $p(x|H_1)/p(x|H_0)$, from which we could then compute the posterior probabilities $p(H_0|x$ and $p(H_1|x)$.

However, as we state it above, the alternative H_1 is not sufficiently detailed to compute $p(x|H_1)$. Thus, to compare these hypotheses in a Bayesian framework, one must be more specific about H_1. In particular we need to specify a prior distribution $p_1(\mu)$ for μ under H_1.

Putting this another way, distinguishing between $\mu = 0$ and $\mu \neq 0$ is a question about μ, and so to address it from a Bayesian perspective we have to specify a prior distribution for μ. However, simply saying that with probability π_0 μ is equal to 0, otherwise it is not equal to 0, is incomplete as a specification of a prior distribution! To complete it we have to say how it μ is distributed when it is not equal to 0. That is, we have to specify a prior distribution $p_1(\mu)$ for μ under H_1.

Having specified $p_1(\mu)$, we have the BF

$$\text{BF} = \frac{p(x|H_1)}{p(x|H_0)} = \frac{\int p(x|\mu) p_1(\mu)\, d\mu}{p(x|\mu = 0)}.$$

Suppose now we choose the prior $p_1(\mu)$ to be a normal distribution with mean 0, and variance σ_μ^2. It might be tempting to choose σ_μ^2 to be "big", thinking of this as "non-informative". However, this would be a big mistake. Specifically, under this prior, as $\sigma_\mu^2 \to \infty$, the Bayes Factor tends to 0 (i.e. provides infinite support for H_0, vs H_1).

To prove this, we can use the following trick that can be useful for computing integrals of this sort. Note that, rearranging Bayes Theorem, we get

$$\int p(x|\theta) p(\theta)\, d\theta = p(x|\theta) p(\theta) / p(\theta|x).$$

Note that, despite appearances, the right hand side does not actually depend on θ (since the LHS does not!) So if we know the posterior distribution, the prior distribution, and the likelihood, we also implicitly know the integral.

In the example above, θ is μ. Under H_1, in the limit $\sigma_\mu^2 \to \infty$, the terms $p(x|\mu)$ and $p(\mu|x)$ tend to finite limits, and $p_1(\mu|\sigma_\mu^2) \to 0$, so the integral tends to 0.

Limiting forms of conjugate priors as "non-informative" priors

It often turns out that the hyper-parameters of a conjugate prior have an effect on the posterior that is reasonable straightforward to interpret. In these cases it is often tempting (and sometimes reasonably sensible) to think of "limiting" forms for these priors as being "minimally informative".

Example: Limiting Beta prior for Binomial proportion

Consider observing the number of successes (n_s) and failures (n_f) in a series of n Bernoulli(p) trials. The conjugate prior is $p \sim Beta(\alpha, \beta)$ and corresponding posterior is $p|n_s, n_f \sim Beta(n_s + \alpha, n_f + \beta)$, with posterior mean $(n_s + \alpha)/(n + \alpha + \beta)$. Note that, informally, α and β can be interpreted as an "effective number of successes and failures" that the prior encapsulates, because α is added to n_s and β is added to n_f. Certainly, if α and β are big, then the prior will outweigh the data, and the larger α and β are the more "informative" the prior is. Based on these kinds of argument one might argue that taking α and β to be as small as possible will make the prior "minimally" informative. This could lead you to take the limit $\alpha, \beta \to 0$. The resulting posterior is the same as you would have gotten if you used the improper prior

$$\pi(p) \propto p^{-1}(1-p)^{-1}$$

which is sometimes referred to as Haldane's prior. Note: in this case the posterior is proper provided both n_s and n_f are > 0; otherwise it is improper!

Example: Limiting conjugate prior for Normal mean and Variance

Suppose we observe data $x = x_1, \ldots, x_n$ are i.i.d. $\sim N(\mu, 1/\tau)$ where τ here indicates the inverse of the variance, also known as the "precision". [It turns out that algebra tends to be easier if we work with $\tau = 1/\sigma^2$, so this is common in this context.]

The joint conjugate prior for (μ, τ) is the "Normal-gamma" with hyperparameters (μ_0, n_0, m_0, l_0):

$$\tau \sim \Gamma(m_0/2, l_0/2) \tag{1}$$

$$\mu|\tau \sim N\left(\mu_0, (1/n_0)(1/\tau)\right). \tag{2}$$

With this prior, the posterior can be derived as also being Normal-gamma:

$$\tau|x \sim \Gamma(m_1/2, l_1/2) \tag{3}$$

$$\mu|\tau, x \sim N\left(\mu_1, (1/n_1)(1/\tau)\right). \tag{4}$$

where

$$m_1 = m_0 + n \tag{5}$$

$$l_1 = l_0 + \sum_{i=1}^{n}(x_i - \mu_1)^2 \tag{6}$$

$$\mu_1 = \lambda\bar{x} + (1-\lambda)\mu_0 \tag{7}$$

$$n_1 = n_0 + n \tag{8}$$

where $\lambda = n/(n + n_0)$.

Note that, intuitively, n_0, m_0 can be thought of as representing an "effective number of observations" informing the prior.

Question: what happens to the posterior in the limit $n_0, l_0, m_0 \to 0$? Is there an "improper prior" that gives the same posterior as this limiting posterior? What is it? (Ans: $p(\mu, \tau) \propto \tau^{-1/2}$, or equivalently, $p(\mu, \sigma) \propto 1/\sigma^2$. This is also the multivariate Jeffreys prior, which Jeffrey's ultimately rejected (see later).

A paradox: Limiting conjugate prior for Normal mean and Variance

The above example leads to an interesting paradox. Consider the case $n = 1$ and $n_0 \to \infty$. Then the limiting posterior distribution for $\tau|x$ is $\Gamma((m_0+1)/2, l_0/2)$. Note that this is different from the prior, but does not depend on x_1! So we can simply "imagine" collecting x_1, update our prior, and get information on τ. Note that this paradox occurs only in the limit (or equivalently, only with the use of an improper prior). Paradoxes like this never occur with proper priors. For this reason some people argue that you should never use improper priors.

Note, however, that *in practice*, for realistic $n >> 1$ the limiting posterior will be a very good approximation to the posterior you would get with a proper prior. So I don't think this example completely rules out the use of such a prior in practice, as a way to get a reasonable approximation to your posterior. Rather, it illustrates the *conceptual* or *foundational* difficulty of defining suitable universal "non-informative" priors.

Jeffreys Priors

The above arguments are rather *ad hoc* (if, possibly, intuitively appealing) ways to come up with non-informative priors. In contrast Jeffrey's came up with a much more formal way to derive potentially "non-informative" priors.

Jeffreys wanted to come up with a "rule" for obtaining an objective prior that had the following property: if you applied the rule to θ you would get the same prior for any (monotone differentiable) 1-1 transformation of θ, $h(\theta)$ say, as if you applied the rule to $h(\theta)$. In this sense the prior (or the rule) is "transformation invariant".

Note that the rule "use a uniform prior" doesn't work in general. For example, consider $x \sim$ Binomial(n, p). Now suppose we apply this rule to p, so $p \sim$ U[0,1]. And if we apply the rule to $\sqrt{p}$, we get $\sqrt{(}p) \sim$U[0,1]. But these are different prior assumptions.

Definition: Recall that for a model with parameter space $\Theta \subseteq \mathbf{R}$, the Fisher information is

$$I(\theta) = E_\theta \left(\frac{d \log(f(x \mid \theta))}{d\theta} \right)^2$$

where $f(x \mid \theta)$ is the sampling distribution and the expectation is taken over $f(x \mid \theta)$. Under regularity conditions,

$$I(\theta) = -E_\theta \left(\frac{d^2 \log(f(x \mid \theta))}{d\theta^2} \right).$$

In such a setting, the *Jeffreys Prior* for θ is defined by $\pi(\theta) \propto I(\theta)^{1/2}$, to be proportional to the square root of the Fisher Information at θ. Note that in general the Jeffreys prior may be improper.

Note that by the chain rule,

$$I(\theta) = I(h(\theta)) \Big(\frac{dh}{d\theta} \Big)^2.$$

If θ has the Jeffreys prior and h is a monotone differentiable function of θ, the prior induced on $h(\theta)$ by the Jeffreys prior on θ is

$$\pi(h(\theta)) = \pi(\theta) \big| \frac{dh}{d\theta} \big|^{-1} \propto I(\theta)^{1/2} \big| \frac{dh}{d\theta} \big|^{-1} = I(h(\theta))^{1/2}.$$

Thus the Jeffreys priors are invariant under reparameterization.

Recall the interpretation of $I(\theta)$ as the ability of the data to distinguish between θ and $\theta + d\theta$. If the prior favors values of θ for which $I(\theta)$ is large, the effect is to minimize the effect of the prior distribution relative to the information in the data. In this sense you can think of the prior as attempting to be "uninformative" about θ.

Example:

Suppose $x \sim \text{Binomial}(n, p)$. Then:

$$f(x \mid \theta) = \binom{n}{x} p^x (1-p)^{n-x}$$

so that

$$\frac{d^2 \log(f(x \mid p))}{dp^2} = -\frac{x}{p^2} - \frac{(n-x)}{(1-p)^2}.$$

Thus

$$\begin{aligned} I(p) &= E\left(\frac{x}{p^2} + \frac{n-x}{(1-p)^2}\right) \\ &= \left(\frac{n}{p} + \frac{n}{1-p}\right) = \frac{n}{p(1-p)}. \end{aligned}$$

Thus the Jeffreys prior for p is

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

which is the Beta$(\frac{1}{2}, \frac{1}{2})$ density (and hence proper).

Multivariate Jeffrey's Priors

For $\Theta \subseteq \mathbf{R}^k$, under suitable regularity conditions the Fisher Information matrix has (i, j)th element

$$I_{ij}(\theta) = -E_\theta\left(\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f(x \mid \theta)\right).$$

In this multidimensional case, the Jeffreys prior is defined by

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

It is still invariant under reparameterization. However, it leads to priors that are considered undesirable by some (including Jeffreys himself).

Example (Bernardo and Smith, p361)

Consider $x \sim N(\mu, \sigma^2)$ with $\theta = (\mu, \sigma)$.

Then,

$$\begin{aligned}
I(\theta) &= -E_\theta \begin{pmatrix} \frac{\partial^2}{\partial\mu^2}(-\log\sigma - \frac{(x-\mu)^2}{2\sigma^2}) & \frac{\partial^2}{\partial\mu\partial\sigma}(-\log\sigma - \frac{(x-\mu)^2}{2\sigma^2}) \\ \frac{\partial^2}{\partial\mu\partial\sigma}(-\log\sigma - \frac{(x-\mu)^2}{2\sigma^2}) & \frac{\partial^2}{\partial\sigma^2}(-\log\sigma - \frac{(x-\mu)^2}{2\sigma^2}) \end{pmatrix} \\
&= -E_\theta \begin{pmatrix} -1/\sigma^2 & 2(x-\mu)/\sigma^3 \\ 2(x-\mu)/\sigma^3 & 1/\sigma^2 - 3(x-\mu)^2/\sigma^4) \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}.
\end{aligned}$$

The Jeffreys prior is thus

$$\pi(\mu,\sigma) \propto \left(\frac{1}{\sigma^2}\frac{2}{\sigma^2}\right)^{1/2} \propto \frac{1}{\sigma^2}.$$

Using this prior to analyse data $x_1, \ldots, x_n$ leads to a posterior in which $\sum_{i=1}^n (x_i - \bar{x})^2/\sigma^2$ is χ^2_n. Compare this to an analysis of data with *known* mean 0, and unknown variance; in this case the posterior is such that $\sum_{i=1}^n x_i^2/\sigma^2$ is χ^2_n. Since the degrees of freedom of these χ^2 variables are the same, it seems that not knowing the mean does not lead to any loss of information in the posterior, which is widely recognized as unacceptable.

This kind of problem led Jeffreys to the ad hoc recommendation of treating location and scale parameters independently, applying his rule to each group separately, and multiplying the results together. For the $N(\mu,\sigma^2)$ density this leads to

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

instead of σ^{-2}. This leads to a posterior in which $\sum_{i=1}^n (x_i - \bar{x})^2/\sigma^2$ is χ^2_{n-1}, and we have "lost a degree of freedom" by not knowing the mean, which accords better with intuition.

Marginalisation Paradoxes

Consider example 5.26 in Bernardo and Smith, p362. This example assumes $x_1, \ldots, x_n$ to be a random sample from $N(\mu,\sigma^2)$. The standard noninformative prior for this problem is $\pi(\mu,\sigma) = 1/\sigma$. Although this gives adequate results if one wants to make inferences about either μ or σ, it is quite unsatisfactory if one wants to make inferences about $\psi = \mu/\sigma$. Specifically, it turns out that:

- the posterior for ψ depends on the data x only through the statistics $t = \sum_i x_i / \sum_i x_i^2$.

- the sampling distribution of t depends on μ, σ only through ψ.
- the posterior for ψ is *not* proportional to any prior on ψ times this sampling distribution.

That is, in some sense the posterior that one obtains for ψ using this prior does not correspond to what one would get with any given prior on ψ (Stone and Dawid, 1972)!

Bernardo and Smith note that this type of *marginalisation paradox* appears in many multivariate problems, and makes it "difficult to believe that, for any given model, a *single* prior may be usefully regarded as "universally" non-informative".

Maximum Entropy Priors

The approach of maximum entropy priors was pioneered by ET Jaynes. He assumed that there existed a certain limited amount of information (e.g. the prior mean and variance), and then tried to determine a prior that reflected this initial information, and nothing else (or with minimal additional information). The basic idea is to maximise the uncertainty of the prior (as measured by Entropy, defined below), subject to constraints determined by the initial information. It is a nice idea that works well for discrete parameters, but not continuous ones.

Suppose that Θ is discrete. If π is a probability mass function on Θ then the *entropy* of π, denoted by $\mathcal{E}(\pi)$, is defined by

$$\mathcal{E}(\pi) = -\sum_{\Theta} \pi(\theta_i) \log(\pi(\theta_i))$$

(with x log x $\equiv 0$ for x $= 0$).

Entropy is a natural measure of the amount of uncertainty in an observation from the distribution. For example, if Θ is finite with $|\Theta| = n$, the entropy is largest for the uniform distribution on Θ and smallest whenever $\pi(\theta_i) = 1$ for some θ_i.

Now suppose partial prior information is available so that for m functions $g_1, \ldots, g_m$, we must have

$$E_\pi(g_k(\theta)) \equiv \sum_i \pi(\theta_i) g_k(\theta_i) = \mu_k \tag{9}$$

under the prior π. In choosing amongst the priors which satisfy these constraints, it may be natural (in terms of representing ignorance) to choose the distribution with the maximal entropy. It can be shown that amongst all probability distributions which satisfy the constraints (1), the distribution with maximal entropy is

$$\tilde{\pi}(\theta_i) = \frac{\exp\left(\sum_{k=1}^{m} \lambda_k g_k(\theta_i)\right)}{\sum_i \exp\left(\sum_{k=1}^{m} \lambda_k g_k(\theta_i)\right)}$$

where the λ_i are constants whose values will be determined by the constraints.

Example: Suppose $\Theta = \{0, 1, 2, \ldots\}$ and the prior mean of θ is thought to be 5. This is a constraint with m=1, $g_1(\theta) = \theta, \mu_1 = 5$. Then the maximum entropy prior is

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

a geometric distribution, and we need $e^{\lambda_1} = \frac{1}{6}$ to ensure it has mean 5.

Now consider the case in which Θ is continuous. The use of maximum entropy is more complicated, mostly because there is no longer a completely natural definition of entropy.

Suppose $\pi(\theta)$ is a density. The entropy of π relative to a particular "reference" distribution with density π_0 is

$$\mathcal{E}(\pi) = -E_{\pi}\left(\log \frac{\pi(\theta)}{\pi_0(\theta)}\right) = -\int_{\Theta} \pi(\theta)\left(\log \frac{\pi(\theta)}{\pi_0(\theta)}\right) d\theta.$$

(The definition of entropy in the discrete setting above coincides with this definition if $\pi_0(\theta)$ is taken to be uniform.)

In the discrete setting the natural reference measure π_0 is uniform. For continuous Θ, the choice of reference measure is less clear. Note that in general, a prior which is chosen to have maximum entropy with respect to π_0 subject to certain constraints will depend on the choice of reference measure π_0.

One approach is to use as π_0 the natural "invariant" noninformative prior for the problem. In the presence of partial information of the form

$$\int_{\Theta} g_k(\theta)\pi(\theta)d\theta = \mu_k, \quad k = 1, \ldots, m,$$

the (proper) prior density (satisfying these restrictions) which maximizes entropy relative to π_0 is given (when it exists) by

$$\tilde{\pi}(\theta) = \frac{\pi_0(\theta)\exp\left(\sum_{k=1}^{m} \lambda_k g_k(\theta)\right)}{\int_{\Theta} \pi_0(\theta)\exp\left(\sum_{k=1}^{m} \lambda_k g_k(\theta)\right) d\theta}$$

where the λ_k are constants to be determined by the constraints.

Example 1: Suppose $\Theta = \mathbf{R}$ and that θ is a location parameter. The natural noninformative prior is then $\pi_0(\theta) = 1$. If the prior mean and variance of θ are fixed to be μ and σ^2 respectively, we have $g_1(\theta) = \theta, \mu_1 = \mu$; $g_2(\theta) = (\theta - \mu)^2, \mu_2 = \sigma^2$. The corresponding maximal entropy prior is

$$\begin{aligned}\tilde{\pi}(\theta) &= \frac{\exp(\lambda_1\theta + \lambda_2(\theta-\mu)^2)}{\int_\Theta \exp(\lambda_1\theta + \lambda_2(\theta-\mu)^2)d\theta} \\ &\propto \mathit{exp}(\lambda_1\theta + \lambda_2\theta^2) \\ &\propto \exp(\lambda_2(\theta-\alpha)^2)\end{aligned}$$

for suitable α, from which it follows that $\tilde{\pi}$ is a Normal density. In view of the constraints, it must be the density for $N(\mu, \sigma^2)$.

Example 2: Suppose Θ and θ are as in the previous example, but now only the prior mean of θ is specified. Then $\tilde{\pi}$ must be of the form

$$\tilde{\pi}(\theta) = \frac{\exp(\lambda_1\theta)}{\int_{-\infty}^{\infty} \exp(\lambda_1\theta)d\theta}.$$

No such distribution exists (since the integral is infinite).

Quotes (on the use of subjective priors):

George Box: "In the past, the need for probabilities expressing prior belief has often been thought of, not as a necessity for all scientific inference, but rather as a feature peculiar to Bayesian inference. This seems to come from the curious idea that an outright assumption does not count as a prior belief. ... I believe that it is impossible logically to distinguish between model assumptions and the prior distribution of the parameters."

I.J.Good: "The subjectivist states his judgements, whereas the objectivist sweeps them under the carpet by calling assumptions knowledge, and he basks in the glorious objectivity of science."

3.4 Bayesian Robustness

In some cases, the form chosen for the prior can be important, in the sense that different priors can lead to quite different posteriors (even if, for example, the priors match in certain ways, such as the mean and certain percentiles). In many applications of Bayesian techniques, it is thus important to assess the sensitivity of the conclusions to aspects of the prior. In such analyses it makes sense to identify a range of "reasonable" priors, being priors that you think might reflect opinions of other reasonable individuals (e.g. readers of an article), even if these do not actually reflect your own opinions. In particular, if trying to make a particular case, it can be helpful to consider priors that are skeptical regarding that case, and investigate whether the evidence in the data is enough to overwhelm such skepticism.

Note: if different reasonable priors lead to substantively different conclusions, then you have learned something very important: that what one should believe after viewing the data depends on ones beliefs before viewing the data, or in other words that the data are not sufficiently informative to draw strong conclusions. It is important to note that *this indicates a fundamental limitation of the data*, and *not* a limitation of the Bayesian approach. In particular, it is not a sign that one should discard the Bayesian analysis and look for a different analysis that leads to a more concrete or confident conclusion! It is often the case that a careful Bayesian analysis will result in more nuanced or more uncertain conclusions. This can be a hard sell to collaborators who want to publish strong conclusions- but it is an element that should be embraced, as reducing the risk of publishing *incorrect* strong conclusions!

Toy Example of prior affecting posterior: Suppose that $x \sim N(\theta, 1)$ and that the prior median of θ is 0, the first quartile is -1 and the third quartile is +1. If the prior is taken to be normal, we must have $\theta \sim N(0, 2.14)$. If the prior is Cauchy, it must be C(0,1) (with pdf $1/(\pi(1+\theta^2)))$.

Suppose $x = 4$, an observation quite compatible with the prior information in both cases. The, posterior mean for θ takes the value 2.75 in the Normal analysis and 3.76 in the Cauchy prior analysis.

Note that, in general, sensitivity on the prior can depend on the data. Here are the posterior means for different values of x.

Observed x	0	1	2	4.5	10
Posterior mean for Normal Prior	0	0.69	1.37	3.09	6.87
Posterior mean for Cauchy Prior	0	0.55	1.28	4.01	9.80

Contamination priors

For a more formal robustness analysis, one could specify a class of priors and calculate the extent to which aspects of the posterior change for priors from the class. For details of natural candidate classes, see Robert §3.5, or for more details, Berger §4.7.

Example: (Berger, example 26, page 212). Suppose $x \sim N(\theta, \sigma^2)$, with σ^2 known. Consider initially π_0, a $N(\mu, \tau^2)$ prior for θ. Define an alternative collection of priors

$$Q = \{q_k : q_k \text{ is a uniform } (\mu - k, \mu + k) \text{ density}\}.$$

Now consider the class

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

called an ε-contamination class of priors. (Γ consists of prior beliefs that with probability $1 - \varepsilon$, the prior is π_0, but otherwise it is $q \in Q$.)

Now consider a fixed interval $C = (c_1, c_2)$ and ask about the range of values of

$$P_{\pi(\theta|x)}(\theta \in C), \quad \pi \in \Gamma,$$

the posterior probability that $\theta \in C$ as the prior, π, ranges over Γ.

It follows from Bayes Theorem, that for $\pi = (1 - \varepsilon)\pi_0 + \varepsilon q_k$, where $q_k \in Q$,

$$P_{\pi(\theta|x)}(\theta \in C) = \lambda_k(x) P_0 + (1 - \lambda_k(x)) Q_k$$

where

$$\lambda_k(x) = \left(1 + \frac{\varepsilon}{1 - \varepsilon} \frac{p(x \mid q_k)}{p(x \mid \pi_0)}\right)^{-1},$$

$p(x \mid \cdot)$ is the predictive density for x when the prior is $\cdot$, and P_0 and Q_k are $P_{\pi_0(\theta|x)}(\theta \in C)$ and $P_{q_k(\theta|x)}(\theta \in C)$, the posteriors corresponding to priors π_0 and q_k respectively. (Thus, note that when the prior is a mixture, the posterior is a mixture of the appropriate posteriors with the mixing weights depending on the data.)

Here the predictive density $p(x \mid \pi_0)$ is $N(\mu, \sigma^2 + \tau^2)$,

$$p(x \mid q_k) = \int_{\mu+k}^{\mu+k} \psi\left(\frac{x - \theta}{\sigma}\right) \frac{1}{2k} d\theta,$$

where ψ is standard Normal pdf, and

$$\begin{aligned} Q_k &= P_{q_k(\theta|x)}(\theta \in C) \\ &= \frac{1}{p(x \mid q_k)} \int_{c^\star}^{c^{\star\star}} \psi\left(\frac{x - \mu}{\sigma}\right) \frac{1}{2k} d\theta, \end{aligned}$$

where $c^\star = \max\{c_1, \mu - k\}, c^{\star\star} = \min\{c_2, \mu + k\}$.

As a concrete example, consider $\varepsilon = 0.1, \sigma^2 = 1, \tau^2 = 2$,$\mu = 0, x = 1$, and $c = \{-0.93, 2.27\}$. For these values, C is the 95% credible region for the prior π_0. Then

$$\inf_{\pi \in \Gamma} P_{\pi(\theta|x)}(\theta \in C) = 0.945,$$

acheived at $k = 3.4$, and

$$\sup \pi \in \Gamma P_{\pi(\theta|x)}(\theta \in C) = 0.956,$$

acheived at $k = 0.93$.

Thus, at least for the class Γ, the statement that the posterior probability that $\theta \in C$ is 0.95 is very robust.