is given by

$$r(\tau, \delta^\tau(X)) = \sum_{i=1}^{p} \text{var}(\mu_i|X_i) = \sum_{i=1}^{p} \frac{\tau^2}{\tau^2+1} = \frac{p\tau^2}{\tau^2+1}.$$

Marginally the X_i are independent, identically distributed $N(0, \tau^2+1)$, so that $X_i/\sqrt{\tau^2+1} \sim N(0,1)$ and marginally $\|X\|^2/(\tau^2+1) \sim \chi^2_p$. Since we know that $E(1/Z) = 1/(p-2)$ if $Z \sim \chi^2_p$ and $p \geq 3$, we see that taking the expectation with respect to this marginal distribution of X gives

$$\mathbb{E}\left[1 - \frac{(p-2)}{\|X\|^2}\right] = \frac{\tau^2}{\tau^2+1}, \tag{3.4}$$

if $p \geq 3$.

In the case when τ^2 is unknown, estimating $\tau^2/(\tau^2+1)$ by $1-(p-2)/(\|X\|^2)$ yields the James–Stein estimator $d^{p-2}(X)$.

Under our assumed model, the Bayes risk of the James–Stein estimator $d^{p-2}(X)$ is

$$\begin{aligned} r(\tau, d^{p-2}(X)) &= \int R(\mu, d^{p-2}(X))\pi(\mu)d\mu \\ &= \int_{\mathbb{R}^p}\int_{\mathcal{X}} \left[p - \frac{(p-2)^2}{\|x\|^2}\right] f(x|\mu)\pi(\mu)dxd\mu \\ &= \int_{\mathcal{X}} \left\{\int_{\mathbb{R}^p} \left[p - \frac{(p-2)^2}{\|x\|^2}\right]\pi(\mu|x)d\mu\right\} f(x)dx, \end{aligned}$$

where we have used (3.3) and then changed the order of integration. Now, the integrand in the inner integral is independent of μ, and $\int \pi(\mu|x)d\mu$ is trivially equal to 1, and therefore

$$r(\tau, d^{p-2}(X)) = p - (p-2)^2\mathbb{E}\left(\frac{1}{\|X\|^2}\right).$$

Now the expectation is, as in (3.4), with respect to the marginal distribution of X, so that (3.4) immediately gives

$$r(\tau, d^{p-2}(X)) = p - \frac{p-2}{\tau^2+1} = r(\tau, \delta^\tau(X)) + \frac{2}{\tau^2+1}.$$

The second term represents the increase in Bayes risk associated with the need to estimate τ^2: the increase tends to 0 as $\tau^2 \to \infty$.

3.6 Choice of prior distributions

To understand some of the controversies about Bayesian statistics, including various ways of thinking about the choice of prior distributions, it is helpful to know something more of the history of the subject.

Bayesian statistics takes its name from an eighteenth-century English clergyman, the Reverend Thomas Bayes. Bayes died in 1761 but his most famous work, 'An essay towards solving a problem in the doctrine of chances', was published posthumously in the *Philosophical Transactions of the Royal Society* (Bayes, 1763). The problem considered by Bayes was, in modern terminology, the problem of estimating θ in a binomial (n, θ) distribution

and he worked out what we now call the Bayesian solution, under the assumption that θ has a uniform prior density on (0,1), equivalent to $a = b = 1$ in our Beta prior formulation of Example 3.1. This assumption, sometimes called *Bayes' postulate*, is the controversial assumption in the paper (not Bayes' Theorem itself, which is just an elementary statement about conditional probabilities). Some authors have held, though modern scholars dispute this, that Bayes' dissatisfaction with this assumption is the reason that he did not publish his paper during his lifetime. Whether this is correct or not, it is the case that much of the paper is devoted to justifying this assumption, for which Bayes gave an ingenious physical argument. However, Bayes' argument is difficult to generalise to other situations in which one might want to apply Bayesian statistics.

At the time, Bayes' paper had very little influence and much of what we now call Bayesian statistics was developed, independently of Bayes, by the French mathematician Laplace (resulting in this *Théorie Analytique des Probabilités*, published in 1812, though the bulk of the work was done in the 1770s and 1780s). Laplace widely used the 'principle of insufficient reason' to justify uniform prior densities: we do not have any reason to think that one value of θ is more likely than any other, therefore we should use a uniform prior distribution. One disadvantage of that argument is that if we apply the principle of insufficient reason to θ^2, say, this results in a different prior from the same principle applied to θ. The argument used by Bayes was more subtle than that, and did lead to a uniform prior on θ itself rather than some transformation of θ, but only for a specific physical model.

By the time more-modern theories of statistical inference were being developed, starting with the work of Francis Galton and Karl Pearson in the late nineteenth century, Bayesian ideas were under a cloud, and R.A. Fisher, arguably the greatest contributor of all to modern statistical methods, was vehemently anti-Bayesian throughout his career. (Fisher never held an academic post in statistics or mathematics, but for many years was Professor of Genetics in Cambridge, and a Fellow of Gonville and Caius College.) However, the tide began to swing back towards Bayesian statistics beginning with the publication of Jeffreys' book *Theory of Probability* in 1939. Jeffreys was also a Cambridge professor, most famous for his contributions to applied mathematics, geophysics and astronomy, but he also thought deeply about the foundations of scientific inference, and his book, despite its title, is a treatise on Bayesian methods. Following in the tradition of Laplace, Jeffreys believed that the prior distribution should be as uninformative as possible, and proposed a general formula, now known as the Jeffreys prior, for achieving this. However, his arguments did not convince the sceptics; Fisher, in a review of his book, stated that there was a mistake on page 1 (that is the use of a Bayesian formulation) and this invalidated the whole book!

One feature of the arguments of Laplace and Jeffreys is that they often result in what we have termed improper priors. Suppose we use the principle of insufficient reason to argue in favour of a uniform prior for a parameter θ. When the range of θ is the whole real line (for instance, if θ is the unknown mean of a normal distribution) then this would lead to a prior which cannot be normalised to form a proper density. The limit in Example 3.2 above, where $\sigma_0^2 \to \infty$, is a case in point. However, in many such cases the *posterior* density is still proper, and can be thought of as a limit of posterior densities based on proper priors. Alternatively, a decision rule of this form is extended Bayes. Most modern Bayesians do not have a problem with improper prior distributions, though with very complicated problems there is a danger that an improper prior density will result in an improper posterior density, and this must of course be avoided!

While Jeffreys was developing his theory, Neyman and Egon Pearson (son of Karl) had published their theory of hypothesis testing (Neyman and Pearson, 1933), which also avoided any reference to Bayesian ideas. (Fisher also disagreed with Neyman's approach, but the source of their disagreement is too complicated to summarise in a couple of sentences. The one thing they agreed on was that Bayesian ideas were no good.) The ideas started by Neyman and Pearson were taken up in the United States, in particular by Abraham Wald, whose book *Statistical Decision Functions* (1950) developed much of the abstract theory of statistical decisions which we see in this text.

However at about the same time B. de Finetti (in Italy) and L.J. Savage (in the USA) were developing an alternative approach to Bayesian statistics based on subjective probability. In the UK, the leading exponent of this approach was D.V. Lindley. According to de Finetti, Savage and Lindley, the only logically consistent theory of probability, and therefore of statistics, is one based on personal probability, in which each individual behaves in such a way as to maximise his/her expected utility according to his/her own judgement of the probabilities of various outcomes. Thus they rejected not only the whole of classical (non-Bayesian) statistics, but also the 'uninformative prior' approach of Laplace and Jeffreys. They believed that the only way to choose a prior distribution was subjectively, and they had no problem with the fact that this would mean different statisticians reaching different conclusions from the same set of data.

There are many situations where subjective judgement of probability is essential. The most familiar situation is at a racetrack! When a bookmaker quotes the odds on a horse race, he is using his subjective judgement, but a bookmaker who did not consistently get the odds right (or very nearly right) would soon go out of business. American weather forecasters also make widespread use of subjective probabilities, because their forecasts always include statements like 'the chance of rain is 40%'. Although they have all the modern tools of computer-based weather forecasting to help them, the actual probability quoted is a subjective judgement by the person making the forecast, and much research has been done on assessing and improving the skills of forecasters in making these subjectively based forecasts.

Thus there are many situations where subjective probability methods are highly appropriate; the controversial part about the theories of de Finetti and Savage is the assertion that *all* probabilistic and statistical statements should be based on subjective probability.

From the perspective of present-day statistics, Bayesian and non-Bayesian methods happily co-exist most of the time. Some modern theoreticians have taken a strongly pro-Bayesian approach (see, for example, the introduction to the 1985 second edition of Berger's book) but much of the modern interest in Bayesian methods for applied statistics has resulted from more pragmatic considerations: in the very complicated models analysed in present-day statistics, often involving thousands of observations and hundreds of parameters, Bayesian methods can be implemented computationally using devices such as the Gibbs sampler (see Section 3.7), whereas the calculation of, for instance, a minimax decision rule, is too complicated to apply in practice. Nevertheless, the arguments are very far from being resolved. Consider, for example, the problem of estimating a density $f(x)$, when we have independent, identically distributed observations $X_1, \ldots, X_n$ from that density, but where we do not make any parametric assumption, such as normal, gamma, etc. This kind of problem can be thought of as one with an infinite-dimensional unknown parameter, but in that case it is a hard problem (conceptually, not just practically) to formulate the kind of prior

distribution necessary to apply Bayesian methods. Meanwhile, comparisons of different estimators by means of a criterion such as mean squared error are relatively straightforward, and some modern theoreticians have developed ingenious minimax solutions to problems of this nature, which have no counterpart in the Bayesian literature.

Thus, the main approaches to the selection of prior distributions may be summarised as:

(a) physical reasoning (Bayes) – too restrictive for most practical purposes;
(b) flat or uniform priors, including improper priors (Laplace, Jeffreys) – the most widely used method in practice, but the theoretical justification for this approach is still a source of argument;
(c) subjective priors (de Finetti, Savage) – used in certain specific situations such as weather forecasting (though even there it does not tend to be as part of a formal Bayesian analysis with likelihoods and posterior distributions) and for certain kinds of business applications where prior information is very important and it is worthwhile to go to the trouble of trying to establish ('elicit' is the word most commonly used for this) the client's true subjective opinions, but hardly used at all for routine statistical analysis;
(d) prior distributions for convenience, for example conjugate priors – in practice these are very often used just to simplify the calculations.

3.7 Computational techniques

As mentioned previously, one of the main practical advantages of Bayesian methods is that they may often be applied in very complicated situations where both X and θ are very high dimensional. In such a situation, the main computational problem is to compute numerically the normalising constant that is required to make the posterior density a proper density function.

Direct numerical integration is usually impracticable in more than four or five dimensions. Instead, *Monte Carlo methods* – in which random numbers are drawn to simulate a sample from the posterior distribution – have become very widely used. These methods use computational algorithms known as *pseudorandom number generators* to obtain streams of numbers, which look like independent, identically distributed uniform random numbers over (0,1), and then a variety of transformation techniques to convert these uniform random numbers to any desired distribution.

3.7.1 Gibbs sampler

One computational technique in common use is the *Gibbs sampler*. Suppose θ is d-dimensional: $\theta = (\theta_1, \ldots, \theta_d) \in \Theta \subseteq \mathbb{R}^d$. We know that

$$\pi(\theta|X = x) \propto \pi(\theta) f(x; \theta),$$

but we have no practical method of computing the normalising constant needed to make this into a proper density function. So, instead of doing that, we try to generate a pseudorandom sample of observations from $\pi(\cdot|x)$, sampling from the distribution of θ, holding x fixed. If we can do that, then we can easily approximate probabilities of interest (for example what is $\Pr\{\theta_1 > 27.15|X = x\}$?) from the empirical distribution of the simulated sample.