smfd5.tex
Day 5. 23.5.2012

Then from (*), to get the posterior density $f(\theta|x)$ we have to take the product $f(\theta)f(x|\theta)$ above, and divide by $f(x)$ – a function of x *only* (θ has been integrated out to get it). So: the posterior density $f(\theta|x)$ is itself of the form above, as a function of θ (with a different constant and a different function of x – but these do not matter, as our interest is in θ).

We can now recognise the posterior density $f(\theta|x)$ – it is *normal*. We can read off:
(i) its mean, $\mu + (x - \mu)/(c\sigma^2/n)$,
(ii) its variance, $1/c$. Thus the *posterior precision* is c. But from the definition of c, this is the sum of $1/\tau^2$, the *prior precision*, and $1/(\sigma^2/n)$, the *data precision*. By (i), the mean is

$$\mu[1 - \frac{\text{data precision}}{\text{posterior precision}}] + x.[\frac{\text{data precision}}{\text{posterior precision}}],$$

or

$$\mu[\frac{\text{prior precision}}{\text{posterior precision}}] + x.[\frac{\text{data precision}}{\text{posterior precision}}].$$

This is a *weighted average* of the prior mean μ and the data value x (the sample mean of the n readings), *weighted according to their precisions*. So:
(a) the form, mean and variance (or precision) of the posterior density are intuitive, statistically meaningful and easy to interpret,
(b) the conclusions above show clearly how the Bayesian procedure synthesises prior and data information to give a compromise,
(c) the family of normal distributions is closed in the above example: a normal prior and normal data give a normal posterior. This is an example of *conjugate priors*, to which we return later.
Note. The example above on the normal distribution makes another important point: often θ will be a vector parameter, $\theta = (\theta_1, \cdots, \theta_p)$. For, to specify a normal distribution $N(\mu, \sigma^2)$, we have to specify both the mean $\theta_1 = \mu$ and the variance σ^2 – or its reciprocal, the precision, either of which we can call θ_2. For simplicity, the variance σ^2 in the above was taken known. But in general, we will not know σ^2. Instead, we should include it in the Bayesian analysis, representing our uncertainty about it in the prior density. We then arrive at a posterior density $f(\theta|x)$ for the vector parameter $\theta = (\theta_1, \cdots, \theta_p)$. If our interest is in, say, θ_1, we want the posterior density of θ_1, $f(\theta_1|x)$. We

get this just as in classical statistics we get a marginal density out of a joint density – by *integrating out the unwanted variables*.

In the normal example above, Ex. 2, we could impose a prior density on σ without assuming it known. This can be done ([O'H], Ex. 1.6 p.8, Lee [L], S2.12), but there is no obvious natural choice, so we shall not do so here.

Example 3. The Dirichlet distribution ([O'H], Ex. 1.7 p.10, S10.2-6). Consider the density in $\theta = (\theta_1, \cdots, \theta_k)$ on the region

$$\theta_1, \cdots, \theta_k \geq 0, \qquad \theta_1 + \cdots + \theta_k = 1$$

(a *simplex* in k dimensions), with density

$$f(\theta) \propto \Pi_{i=1}^{k} {\theta_i}^{a_i - 1}$$

for constants a_i. We quote that the constant of proportionality is

$$\Gamma(a_1 + \cdots + a_k)/\Gamma(a_1) \cdots \Gamma(a_k),$$

by *Dirichlet's integral*, an extension of the Eulerian integral for the gamma function (see [O'H] 10.4, or, say, 12.5 of
WHITTAKER, E. T. & WATSON, G. N.: *Modern analysis*, 4th ed., 1927/1963, CUP).
Thus the *Dirichlet density* $D(a_1, \cdots, a_k)$ with *parameters* $\theta_1, \cdots, \theta_k$ is

$$f(\theta) := \frac{\Gamma(a_1 + \cdots + a_k)}{\Gamma(a_1) \cdots \Gamma(a_k)} . {\theta_1}^{a_1 - 1} \cdots {\theta_k}^{a_k - 1}.$$

Now draw a random sample of size n from a population of k distinct types of individuals, with proportions θ_i of type i ($i = 1 \cdots k$). Then the likelihood is

$$f(x|\theta) = \frac{n!}{x_1! \cdots x_k!} . {\theta_1}^{x_1} \cdots {\theta_k}^{x_k},$$

the multinomial distribution. So

$$f(x|\theta) f(\theta) = const. {\theta_1}^{x_1 + a_1 - 1} \cdots {\theta_k}^{x_k + a_k - 1},$$

and the posterior density $f(\theta|x)$ is also of this form, with a different constant (making it a density - i.e., integrating to 1). We recognise the functional form of a Dirichlet density, with a_i replaced by $a_i + x_i$. So

$$f(\theta|x) = \frac{\Gamma(a_1 + \cdots + a_k + n)}{\Gamma(a_1 + x_1) \cdots \Gamma(a_k + x_k)} . {\theta_1}^{a_1 + x_1 - 1} \cdots {\theta_k}^{a_k + x_k - 1}$$

(as $x_1 + \cdots x_k = n$, the sample size).
Mode. To find the mode (the maximum of the density) of the Dirichlet density, we have to maximise $f(\theta) = const.\Pi{\theta_i}^{a_i - 1}$ subject to $\theta_1 + \cdots + \theta_k = 1$, i.e. to maximise $\log f(\theta) = const + \Sigma(a_i - 1)\log\theta_i$ subject to $\theta_1 + \cdots + \theta_k = 1$ (and $\theta_i > 0$). Using a Lagrange multiplier λ, maximise

$$\begin{aligned} g(\theta) &:= \log f(\theta) + \lambda(\theta_1 + \cdots \theta_k - 1) \\ &= const + \Sigma_1^k (a_i - 1)\log\theta_i + \lambda(\theta_1 + \cdots + \theta_k - 1). \end{aligned}$$

$$\partial g/\partial\theta_i = 0 : \qquad (a_i - 1)/\theta_i + \lambda = 0 : \qquad \theta_i = -\frac{1}{\lambda}(a_i - 1) = \frac{1}{\lambda}(1 - a_i) \qquad (i = 1, \cdots k).$$

Sum:

$$\Sigma_1^k \theta_i = 1 = \frac{1}{\lambda}\Sigma_1^k (1 - a_i) = \frac{1}{\lambda}(k - \Sigma a_i) : \qquad \lambda = (k - \Sigma a_i).$$

So the mode is at

$$\theta_i = (1 - a_i)/(k - \Sigma a_i) \qquad (i = 1, \cdots, k).$$

Example 4. Poisson and Gamma distributions ([O'H], Ex. 1.1, 1.2 p.21). Data: $x = (x_1, \cdots, x_n)$, x_i independent, Poisson distributed with parameter θ:

$$f(x|\theta) = \Pi_1^n f(x_i|\theta) = \theta^{x_1 + \cdots + x_n} e^{-n\theta}/x_1! \cdots x_n! = \theta^{n\bar{x}} e^{-n\theta}/\Pi x_i!,$$

where $\bar{x} := \frac{1}{n}\Sigma x_i$ is the sample mean.
Prior: the Gamma density $\Gamma(a, b) \quad (a, b > 0)$:

$$f(\theta) = \frac{a^b \theta^{b-1}}{\Gamma(b)} e^{-a\theta} \qquad (\theta > 0).$$

So

$$f(x|\theta) f(\theta) = \frac{a^b}{\Gamma(b)\Pi x_i!} \theta^{n\bar{x} + b - 1} e^{-(n+a)\theta},$$

$$f(\theta|x) \propto f(x|\theta) f(\theta) = const.\theta^{n\bar{x} + b - 1} e^{-(n+a)\theta}.$$

This has the form of a Gamma density. So, it *is* a Gamma density, $\Gamma(n + a, n\bar{x} + b)$:

$$f(\theta|x) = \frac{(n+a)^{n\bar{x}+b}}{\Gamma(n\bar{x} + b)} .\theta^{n\bar{x} + b - 1} e^{-(n+a)\theta} \qquad (\theta > 0).$$

Means. For $\Gamma(a,b)$, the mean is

$$E\theta = \int_0^\infty \theta f(\theta)d\theta = \frac{a^b}{\Gamma(b)}.\int_0^\infty \theta^b e^{-a\theta}d\theta.$$

Substituting $t := a\theta$, the integral is $\Gamma(b+1)/a^{b+1}$, which is $b\Gamma(b)/a^{b+1}$ using the functional equation for the Gamma function:

$$\Gamma(x+1) = x\Gamma(x) \qquad (x>0).$$

So the mean is $E\theta = b/a$. Similarly,

$$E\theta^2 = \int_0^\infty \theta^2 f(\theta)d\theta = \Gamma(b+2)/a^{b+2},$$

so $var\theta = E(\theta^2) - [E\theta]^2 = b(b+1)/a^2 - (b/a)^2 = b/a^2$.

So by above, the prior mean is b/a; the posterior mean is $(n\bar{x}+b)/(n+a)$; the data mean is $\bar{x}$. Write

$$\lambda := a/(n+a), \qquad \text{so } 1-\lambda = n/(n+a):$$

since

$$\frac{n\bar{x}+b}{n+a} = \frac{a}{n+a}.\frac{b}{a} + \frac{n}{n+a}.\bar{x},$$

posterior mean $(n\bar{x}+b)/(n+a) = \lambda$. prior mean $b/a+(1-\lambda)$. sample mean $\bar{x}$.

Again, this is a weighted average, with weights proportional to n and a. Now n is the sample size, a measure of the precision of the data, and a is the rate of decay of the Gamma density, a measure of the precision of the prior information.

5. PROS AND CONS

Advantages of the Bayesian paradigm

1. *Updating.* Bayesian procedures provide an efficient algorithm for updating prior information as new data information is obtained. This is attractive theoretically: it reflects the way we all constantly update our thinking in the light of new experience, and it works well in a range of examples, as IV.4 shows. It also works well in many practical situations. It is particularly well suited to situations involving *time*, when new information is constantly coming in. Recursive algorithms exist for handling such situations *on-line*, or *in real time*, using computers. Such algorithms are typically Bayesian; an

example is the *Kalman filter*, used for on-line control problems (e.g., adjusting orbits of satellites) from the 1960s on.

2. *Uncertainty.* We have seldom used the words 'probability' or 'random' in the above. Technically, Bayesian statistics differs from classical statistics by treating parameters, not as unknown constants, but – in effect (and explicitly, in [O'H]) – as random variables. This is necessary: only random variables can have distributions, prior and/or posterior.

This change of view – away from thinking of random variables and parameters as separate, towards treating them on the same footing, thinking about uncertainty - is often helpful, *provided* one takes the trouble to get used to it. This chapter is designed to do just that!

Some Bayesians carry this shift away from probability language to surprising extremes. An example is the famous dictum by the father of 20th century Bayesian statistics, Bruno de FINETTI (1906-1985):

PROBABILITY DOES NOT EXIST!

We would not go so far, but do recommend the Bayesian viewpoint as being useful and workable.

3. *Subjectivity.* The information in the data is objective: it is the same to all statisticians following the same procedure and obtaining that data. By contrast, the information used in the choice of prior is subjective: it reflects the experience/knowledge/beliefs of the statistician (or his client). This subjectivity persists into the posterior distribution after we use Bayes' Theorem: the entire analysis has been personalised, to suit the statistician (or his client).

4. *Decision Theory.* The Bayesian formulation (or paradigm) combines well with the ideas of Decision Theory. For this important subject, see e.g.
D. V. LINDLEY, *Making decisions*, 2nd ed., Wiley, 1985.

One context in which the Bayesian/decision-theoretic approach is useful is in business/finance/investment. Suppose one is faced with the need to take major business decisions – e.g., whether/where/when to drill for oil. Drilling is very expensive, and may well produce no return on the large investment of capital in the shape of exploitable oil reserves. But, commercially viable oil reserves can be profitably exploited – and necessarily have to be found by risky exploratory drilling. Nothing venture, nothing win.

In such situations, the Bayesian approach quantifies the statistician's (or client's) uncertainty: decision theory then helps him to act rationally given his beliefs.

5. *Output.* The end-product of a Bayesian analysis is a *posterior distribution.* This is more informative than
(i) a number [point-estimate: e.g., a maximum-likelihood estimate],
(ii) two numbers [interval estimate: e.g., a confidence interval].
It also depends continuously on what it depends on – the prior information and the data information. The discontinuous 'accept or reject' framework of hypothesis testing is avoided.
6. *Nuisance parameters.* A *nuisance parameter* is what its name implies: a parameter which is present in the formulation of the model, but absent from the question of interest. The parameter(s) in which we are interested are called, by contrast, *parameters of interest* or *interest parameters.*
E.g.: *Testing for equality of two normal means.*
The usual classical assumption for testing $H_0 : \mu_1 = \mu_2$ v. $H_1 : \mu_1 \neq \mu_2$, for two normal populations $N(\mu_i, \sigma_i^2)$, is to assume *equality of variances*: $\sigma_1 = \sigma_2$. Testing for equality of means *without* assuming equality of variances is a famous statistical problem, the *Behrens-Fisher problem.* It has a satisfactory solution (Scheffé's solution) when the two sample sizes n_1, n_2 are equal, but not in general.
E. g.: *Testing for normality.* Is this population normal? Here *both* μ *and* σ are nuisance parameters. It is much easier to ask: is this population $N(\mu_0, \sigma_0)$ for *specified* μ_0, σ_0? than to ask: is it $N(\mu, \sigma)$ for *some* μ, σ? One approach would be to estimate the mean and variance from the data, and then 'plug in' these estimates to try to reduce the second question to the first – but this sort of procedure can be hard to justify.

In principle, nuisance parameters are easily handled in Bayesian statistics. If $\theta = (\theta_1, \theta_2)$ with θ_1 the interest parameter and θ_2 the nuisance parameter (either or both of θ_1, θ_2 can be several-dimensional), one finds the posterior density $f(\theta|x)$ as usual. This is the *joint* density of θ_1 and θ_2 (given the data x), so one extracts the *marginal* density of θ_1 (given x) as usual, by integrating out the unwanted variable θ_2:

$$f(\theta_1|x) = \int_{-\infty}^{\infty} f(\theta|x) d\theta_2 = \int_{-\infty}^{\infty} f(\theta_1, \theta_2|x) d\theta_2.$$

Of course, the integration may be difficult to perform – it may, in practice, need to be done numerically. But such problems are quite general, and not the fault of Bayesian statistics!
7. *The Likelihood Principle.* As the fundamental formula of Bayesian statis-

tics,

posterior density is proportional to prior density times likelihood

shows, the data only enters a Bayesian analysis through the likelihood. The *Likelihood Principle* (LP), formulated by G. A. BARNARD (1915-2002) (in a series of papers, 1947-1962) and A. BIRNBAUM (1962) says that the data should only enter any statistical analysis through the likelihood. Thus

Bayesian statistics satisfies the Likelihood Principle.

Classical statistics, however, violates the LP. O'Hagan, for instance, discusses a number of examples.
Ex. ([O'H], 33): Bernoulli trials, success probability θ. Consider two situations:
(a) n trials; you observe r successes;
(b) toss till you observe the rth success: you need n trials.
The two likelihoods are the same [apart from constant factors, arising because in (b), but not in (a), the last toss must be a success]: to a Bayesian statistician, these situations are equivalent. To a classical statistician, however, they are quite different. For instance, the stopping rules are quite different [the area of statistics where one continues sampling until something happens and then stops is called sequential analysis, and has been extensively developed].

[O'H] (5.14-15) points out that
(a) the minimum variance unbiased estimators of θ differ in these two cases,
(b) the very concept of unbiasedness itself violates the LP. For, it involves an expectation over the distribution of x - the bias in a statistic $T(x)$ is

$$b := E[T(x)|\theta] - \theta$$

– and this involves values of x we could have seen but didn't. The LP insists we take account only of the values of x we did see.

A full-length account of the LP, arguing persuasively for Bayesian statistics, is given by
BERGER, J. O. & WOLPERT, R. L. (1988): *The Likelihood Principle* (2nd ed.), Institute of Mathematical Statistics.

Disadvantages of the Bayesian paradigm.
1. *Choice of prior.* A Bayesian analysis cannot even begin without a choice of prior density (or distribution). This may well be problematic:

(a) we may have little prior information,
(b) what prior information we have may not suggest a mathematically convenient, or even tractable, choice of prior,
(c) the choice may be to some extent arbitrary,
(d) different choices of prior may (will) lead to different conclusions,
(e) we may have too sparse a collection of suitable families of priors to hand.
Of course, problems of this sort affect classical parametric statistics too. But classical statistics can fall back in such cases on a non-parametric approach, for which Bayesian treatments are less well developed, and in any case the problem is more acute in Bayesian statistics, as we have to choose suitable forms for both the prior and the likelihood.
Undoubtedly, choice of prior is the hardest thing in many – or even most – Bayesian analyses, and is the feature of Bayesian statistics most objectionable to non-Bayesians.
2. *Prior ignorance.* The less a Bayesian knows, the harder he finds it to choose a prior. The worst-case scenario for a Bayesian is little (or even no) prior knowledge. To a non-Bayesian, this is a non-problem: simply use a classical analysis, relying on the data (which is all we've got).

If θ belongs to a finite interval, $[a, b]$ say, there is a natural choice of prior to represent prior ignorance: the uniform density on $[a, b]$:

$$f(\theta) := 1/(b-a) \qquad \text{if } a \leq \theta \leq b, \text{ 0 else.}$$

But, there is no analogous density in an infinite interval - the real line, say. If $f(\theta) \equiv c > 0$, then *either* $c = 0$, when $\int_{-\infty}^{\infty} f(\theta)d\theta = 0$, *or* $c > 0$, when $\int_{-\infty}^{\infty} f(\theta)d\theta = +\infty$. It is impossible to get $\int_{-\infty}^{\infty} f(\theta)d\theta = 1$, the condition for $f(\theta) \geq 0$ to be a density, without $f(\theta)$ varying with θ. But this treates some θ-values differently from others, which is inconsistent with prior ignorance, when we have no grounds to discriminate between different values of θ.
Note. Some Bayesian statisticians have advocated using *improper priors* (allowing $\int_{-\infty}^{\infty} f(\theta)d\theta = +\infty$) in such cases, for this reason. But this is hard to justify, and is becoming less common nowadays.
3. *Objectivity.* The Bayesian paradigm is well suited to situations where a subjective view is appropriate – particularly where a decision-taker has to act in the face of uncertainty, as in Decision Theory. Typical examples include businessmen facing management decisions about investment (whether/where/when to drill for oil, for instance). The manager's judgement is fed into the choice of prior, and he stands or falls by it. The subjective view is appropriate here.

By contrast, in science, one seeks objectivity. Whether or not Nature works in a certain way depends on Nature (or God), not on our opinions or beliefs [we leave to one side foundational questions about quantum mechanics, and whether or not a quantum formulation necessarily involves the mind of the observer]. Consequently, the Bayesian paradigm has met with more resistance in science than in business, because of the higher value put there on objectivity as against subjectivity.

Note. Lee's book makes telling use of examples about dating rocks in geology. Obviously the age of a rock (some hundreds of millions of years old) is completely objective – it hs nothing to do with us or our opinions. Indeed, it is hard to imagine anything more indifferent to us than a chunk of rock. It has a definite age; God (or Nature) knows this, but won't tell us. We thus have no means, even in principle, of assessing the age of a rock sample (which long predates humanity!) other than our own experimentation, observation and analysis, which will provide partial knowledge with remaining uncertainty. The Bayesian paradigm does provide a sensible way of expressing this. So, despite the obvious objection about subjectivity, a Bayesian approach is quite defensible where, as here, it produces sensible results and there is nothing else to do.

4. *Summary statistics and dimensionality.* For a one-dimensional parameter θ, the output is a posterior density, which we can graph. This is an advantage: 'One picture is worth a thousand words'! The advantage is particularly telling if, as we assume, a computer graphics capability is available.

For a two-dimensional parameter θ, the output is a posterior density in the plane, which we can 'graph' in three dimensions, using a suitable computer graphics package. Again, this is an advantage.

In *three* dimensions, graphics are no longer applicable, because *four* dimensions would be needed.

In higher dimensions, the situation rapidly gets even worse. We cannot *graph* the output; it becomes increasingly difficult even to *visualise* the output. Instead, we seek to *summarise* the output, using suitable summary statistics (e.g., mean/median/mode, covariance matrix, measures of skewness/kurtosis, ...). Thus the extra information in the Bayesian output (posterior density), over and above that from a classical output (summary statistics), is no longer an advantage – because we cannot use it – but actually a drawback – because we have to work to get back to summary statistics, such as a classical treatment provides anyway.

Note. 1. Summarisation methods are discussed in detail in [O'H], 2.1 – 2.24.

2. The dimensional aspects above underscore the *principle of parsimony*: one should seek to work in as low a dimensionality (i.e., with as few parameters) as possible. [It is quite common to find the complexity of a theory growing uncontrollably with increase in dimension. This phenomenon is called the *curse of dimensionality*, a term due to Richard Bellman.]
3. If the right dimensionality is not clear, we may be able to formalise the trade-off between the better fit a higher dimension can provide against the extra complexity by using methods such as Akaike's Information Criterion (AIC): see e.g. [O'H], Ch. 7.
5. *Integration.* Bayesian statistics involves the need for integration in several ways:
(i) to get $f(x) = \int f(x|\theta) f(\theta) d\theta$,
(ii) to get marginal posterior densities from joint posterior densities - e.g., to eliminate nuisance parameters,
(iii) to produce summary statistics as above - e.g., posterior means, etc.
Such integrations may be hard or impossible to do analytically. We may need to integrate numerically. This may be computer-intensive, and involves a good knowledge of, e.g.,
(a) numerical analysis as a branch of mathematics,
(b) computer implementation - e.g., by using the NAG Library [NAG = Numerical Algorithms Group, Oxford University].

Within the last few years, much theoretical and practical progress in such areas has been made, using techniques such as the *Gibbs sampler* and *Markov Chain Monte Carlo* (MCMC) methods. Such methods are still being intensively developed, but are too advanced for courses at this level. However, MCMC often provides the crucial numerical breakthrough needed to make implementation of a complicated Bayesian analysis practically feasible.