Errata and clarifications by Chapter:

Notation

Throughout the book, we have used the handy notation "prob(x)" to denote either a probability or a probability density. If we draw a datum X from some distribution with a density $g(x)$, what we conventionally mean is that the answer lies in some small range X to $X + dX$, and the probability of this event is $g(X)dX$.
As an example, suppose our density depended on some continuous parameter α, with prior $p(\alpha)$ (so the probability of lying in α to $\alpha + d\alpha$ is $p(\alpha)$). In full, Bayes' theorem would read

$$\begin{aligned} \text{prob}(\alpha|X) &= \frac{\text{prob}(X|\alpha)\text{prob}(\alpha)}{\text{prob}(X)} \\ &= \frac{g(X|\alpha)dX\, p(\alpha)d\alpha}{\int g(X|\alpha)dX\, p(\alpha)\, d\alpha} \end{aligned}$$

and the differentials in X cancel out. If we write the posterior density as $g'(\alpha|X)$ then $\text{prob}(\alpha|X) = g'(\alpha|X)\, d\alpha$ and we just have

$$g'(\alpha|X) \propto g(X|\alpha)p(\alpha)$$

which is how we often express Bayes' Theorem explicitly in terms of densities.

Chapter 5

In ¶5.2.3 (Example) an important point has been glossed over in the interests of space - the normalization of the priors. Essentially this Bayesian model comparison technique uses a ratio which involves likelihoods (where the data enter) and priors on the parameters. For instance: if we use a prior on the mean μ which is uniform from $-\mu_0$ to μ_1, then the prior probability distribution in full is

$$\text{prob}(\mu) = \frac{1}{\mu_1 + \mu_0}.$$

The likelihood associated with this model is $\mathcal{L}(\text{data}|\mu)$ and so the numerator in the Bayes factor is

$$\int_{-\mu_0}^{\mu_1} d\mu\, \frac{1}{\mu_1 + \mu_0} \mathcal{L}(\text{data}|\mu).$$

If we were making a model comparison with some other sort of distribution, characterized by a location parameter λ, then the denominator in the Bayes factor would look like

$$\int_{-\lambda_0}^{\lambda_1} d\lambda \, \frac{1}{\lambda_1 + \lambda_0} \mathcal{L}'(\text{data}|\lambda).$$

Strictly the posterior odds will depend on what we have chosen for μ_0, μ_1, λ_0 and λ_1. In many useful cases, however, the likelihood terms are so peaked that we can extend the integrals from $-\infty$ to ∞, and further our prior information may be so weak that we are happy to accept

$$\frac{\lambda_1 + \lambda_0}{\mu_1 + \mu_0} \to 1$$

which gets back to the treatment given in the book.

There are cases where using improper priors can cause trouble and some famous controversies have revolved around this point (Jaynes 2003, Chapter 15). It is as well to be explicit about what we are doing with priors. Jaynes says "...For many years, the present writer was caught out in this error just as badly as anyone else, because Bayesian calculations with improper priors continued to give just the reasonable and clearly correct results that common sense demanded."

To some extent this is an academic debate which arises because we are dealing in rather abstract examples. In real-world problems we are rarely in the state of virginal ignorance that would encourage us to use improper priors – we always know more than that!

An interesting slant on these kind of problems arises when we treat the number of parameters as itself something we want to estimate – this method applies where we want to decompose some data into a mixture of underlying distributions. See Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with unknown number of components (with discussion). Journal of the Royal Statistical Society, Series B, **59**, 731–92

Chapter 6

¶6.3 and ¶6.7 - same comments apply as for ¶5.2.3

Equation 6.18 – should have $(b - a)$ in the numerator of the r.h.s.

¶6.5.1, first example - exponent on last term should be negative, i.e. $-1/1.5$

¶6.5.2 There are some very useful notes on MCMC methods at
`http://www.maths.soton.ac.uk/staff/sahu/utrecht` and
`http://www.stats.bris.ac.uk/ peter/papers/semstat2.ps`.

¶6.6 The discussion of the bootstrap should make one point clearer – the data have to be *random*. For instance, if we have a set of (x, y) pairs where we have fixed each X by design, and measured a Y with some error, then the data are not random in this sense. To evaluate the errors in a model in this case, use Monte Carlo and your knowledge of the error distribution. The bootstrap applies in the case where you have a randomly-selected set of cases and you measure the (X, Y) with small errors. So the basic assumption is that

the data are drawn from some bivariate distribution of x and y. This is the case in the example in the nice Diaconis & Efron paper, where they are interested in the correlation of two types of scores that were obtained by a sample of students. It wouldn't *necessarily* be the case if you choose X_i and measure Y_i, unless you know that you have chosen representative X_i.

Chapter 8

¶8.2 We did not mention anything about tapering data; this is not really a statistical issue but it is very important. A time series or spectrum should not start or end abruptly at some finite value, or the implied sharp edges will appear as spurious sidelobes on any features in the spectrum. A fundamental paper on optimum tapering is D.J. Thomson, 1982. Proc IEEE, **70**, 1055 "Spectrum estimation and harmonic analysis". A more recent paper on this topic is Reidel & Sidorenko, 1995, IEEE Trans Sig Proc, **43**, 188 "Minimum bias multiple taper spectral estimation".

Chapter 9

Exercise 9.4 - should refer to "w_3" not "w_4".