Computationally-Intensive Statistical Methods

1 What is 'Computationally-Intensive Statistics'?

'Computationally-intensive statistics' is statistics that could only be done with *'modern'* computing resources, typically either

- Statistical inference on small problems which needs a lot of computation to do at all, or to do well. Quite small datasets can need complex models to explain, and even simple models can need a lot of computation for a realistic analysis (especially where dependence is involved).
- Statistical inference on 'huge' problems.

All of these terms are relative, and I was reminded of just how relative by Sir David Cox's comment in the verbal discussion of Ripley (2005) that when he was a PhD student inverting a 5×5 matrix was the work of hours.

One very important idea for doing statistical inference 'well' on analytically intractable statistical models (that is, most real-world ones) is to make use of *simulation*. So most of this module could be subtitled *simulation-based inference*, as in Geyer (1999)'s comments about MCMC for spatial point processes:

> *If you can write down a model, I can do likelihood inference for it, not only maximum likelihood estimation, but also likelihood ratio tests, likelihood-based confidence intervals, profile likelihoods, whatever. That includes conditional likelihood inference and inference with missing data.*
>
> *This is overstated, of course. ... But analyses that can be done are far beyond what is generally recognized.*

(even 20 years after my thesis work).

These lecture notes go beyond what will be covered in lectures – they are intended to give pointers to further issues and to the literature.

2 Simulation-based Inference

The basic idea is quite simple – simulate data from one or more plausible models (or for a parametric model, at a range of plausible parameter values), apply the same (or similar) procedure to the simulated datasets as was applied to the original data, and then analyse the results. In this section we consider some of the 'classical' applications, but bootstrapping is another.

The main reference for this section is Ripley (1987, §7.1).

Monte-Carlo tests

Suppose we have a fully-specified null hypothesis, and a test statistic T for which small values indicate departures from the null hypothesis. We can always simulate m samples $t_1, \ldots, t_m$ under the null hypothesis, and use these to obtain an indication of where the observed value T lies on the null distribution. For example, consider a dataset on the amounts of shoe wear in an experiment reported by Box, Hunter and Hunter (1978). There were two materials (A and B) that were randomly assigned to the left and right shoes of 10 boys.

Table 1: Data on shoe wear from Box, Hunter and Hunter (1978).

boy	A	B
1	13.2 (L)	14.0 (R)
2	8.2 (L)	8.8 (R)
3	10.9 (R)	11.2 (L)
4	14.3 (L)	14.2 (R)
5	10.7 (R)	11.8 (L)
6	6.6 (L)	6.4 (R)
7	9.5 (L)	9.8 (R)
8	10.8 (L)	11.3 (R)
9	8.8 (R)	9.3 (L)
10	13.3 (L)	13.6 (R)

A paired t-test gives a t-value of -3.3489 and two-tailed p-value of 0.85% for no difference between the materials. The sample size is rather small, and one might wonder about the validity of the t-distribution. An alternative for a randomized experiment such as this is to base inference on the permutation distribution of `d = B-A`. Figure 1 shows that the agreement is very good.

Monte Carlo tests[1] are a closely related (but not identical) idea. If the null hypothesis is true, we have $m + 1$ samples from the null distribution, one natural and m by simulation. Thus the probability that T is the kth smallest or smaller is *exactly* $k/(m + 1)$ provided we can ignore ties. To do so we now assume that T has a continuous null distribution.

By choosing k and m suitably, say $(1, 19)$, $(5, 99)$, $(50, 999)$ we can derive an exact significance test at any desired level. Note that the experiment can be stopped early if k simulated

[1] Usually attributed to a comment by George Barnard in 1963.

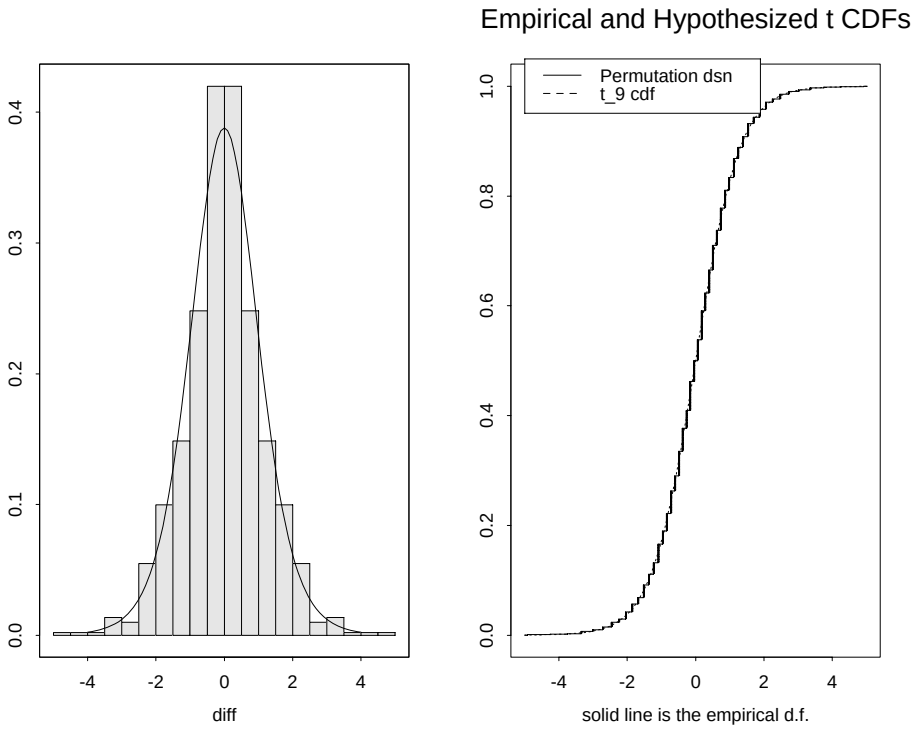

Figure 1: Histogram and empirical CDF of the permutation distribution of the paired t–test in the shoes example. The density and CDF of t_9 are shown overlaid. (Figure 5.5 of Venables and Ripley (2002).)

values less than T have been observed – if the null hypothesis is true this will happen after an average $2k$ trials. However, doing the test indicates some evidence against the null hypothesis, so we should not expect early stopping to be typical.

Power considerations

One common objection to Monte-Carlo tests is that different statisticians will get different results. One answer is

> *So what?q; they would have used different test statistics, or deleted different outliers, or chosen different significance levels or . . .*

Effectively the actual significance level conditional on the simulations varies and only has average α. This will be reflected in a loss in power. Detailed calculations by Jöckel (1986) give

$$\frac{\text{power of MC test}}{\text{power of exact test}} \geq 1 - \frac{E|Z-\alpha|}{2\alpha} \approx 1 - \left[\frac{1-\alpha}{2\pi m\alpha}\right]^{1/2}$$

where $Z \sim \text{beta}(\alpha(m+1), (1-\alpha)(m+1))$. This is a *lower* bound, and ranges, for $\alpha = 5\%$, from 64% for $m = 19$ through 83% for $m = 99$ to 94.5% for $m = 999$. Asymptotic results show better behaviour if the statistic is asymptotically normal, for example.

Note that our discussion has been entirely about simple null hypotheses. There have been some suggestions about how to use Monte Carlo tests for composite null hypotheses, and of course there are many standard arguments to reduce composite null hypotheses to simple ones.

as named by Buckland (1984). These differ in a small but important detail from the bootstrap confidence intervals.

Monte Carlo tests are only defined for a single null hypothesis, so can not easily be inverted to form a confidence interval. Some pivotal quantity is needed. Suppose $\widehat{\theta}$ is a consistent estimator of θ with corresponding CDF F_θ. Let θ^* be a sample from $F_{\widehat{\theta}}$. We want to use the variation of θ^* about $\widehat{\theta}$ to infer the variation of $\widehat{\theta}$ about θ.

Suppose we have a location family. Then

$$\widehat{\theta} - \theta \sim F_0, \qquad \theta^* - \widehat{\theta} \sim F_0$$

so we can obtain upper and lower prediction limits for θ^* by

$$\begin{aligned} L &= F_{\widehat{\theta}}^{-1}(\alpha/2) = \widehat{\theta} + F_0^{-1}(\alpha/2) \\ U &= F_{\widehat{\theta}}^{-1}(1-\alpha/2) = \widehat{\theta} + F_0^{-1}(1-\alpha/2) \end{aligned}$$

either analytically or *via* simulation from the empirical CDF of θ^*. The conventional $(1-\alpha)$ confidence interval for θ is

$$\theta \in \left(\widehat{\theta} - F_0^{-1}(1-\alpha/2), \widehat{\theta} - F_0^{-1}(\alpha/2)\right) = (2\widehat{\theta} - U, 2\widehat{\theta} - L)$$

Thus we get a confidence interval for θ by reflecting the distribution of θ^* about $\widehat{\theta}$. This is the Monte-Carlo confidence interval. If the family is only locally a location family, the confidence interval is only approximately correct. With local scale families, the same arguments apply to $\log\theta$.

Note carefully the difference between this and the bootstrap. The bootstrap resamples (with replacement) from the data. These methods sample from the fitted distribution—sometimes they are called the *parametric bootstrap*, as distinct from the *non-parametric bootstrap*. If we had fitted a completely general class of distributions, the fitted distribution $F_{\widehat{\theta}}$ would be the empirical distribution function F^* which assigns mass $1/n$ to each data point. Then independent sampling from F^* is bootstrap resampling, and the Monte-Carlo confidence intervals are what are known as *basic* bootstrap intervals.

Monte Carlo Likelihood

as termed by Geyer and Thompson (1992). In general the likelihood equations for a canonical exponential family equate the observed value to the expectation of a sufficient statistic, the parameter controlling the expectation. The difficulty can arise in evaluating the expectation.

Consider the so-called Strauss model of spatial inhibition (Ripley, 1988, §4), which has pdf for n points $x_i \in D \subset \mathbb{R}^d$ of

$$f_\theta(x_1, \ldots, x_n) = a(\theta)\theta^{t(x_1,\ldots,x_n)} \tag{1}$$

where $t()$ computes the number of pairs of points closer than distance R, and $0 \le \theta \le 1$. (θ corresponds to c in the preliminary material.) The log-likelihood includes $\log a(\theta)$ and this is unknown.

However, we do not actually need $a(\theta)$, for the MLE of θ satisfies

$$t(x_1, \ldots, x_n) = E_{\widehat{\theta}}[t(X_1, \ldots, X_n)] \tag{2}$$

where the right-hand side can not be expressed in a simple form. We can however estimate the RHS by simulation from the density (1), as in the preliminary material. An example is given in Figure 2, for a data set in which the observed value of $t(x_1, \ldots, x_n)$ was 30. Note that by importance sampling we can estimate the RHS of (2) for a range of θ from simulations at one value, an idea sometimes known as *polysampling*. The idea is that

$$E_\theta[t(X_1, \ldots, X_n)] = E_{\theta_0}\left[t(X_1, \ldots, X_n)\frac{f_\theta(X_1, \ldots, X_n)}{f_{\theta_0}(X_1, \ldots, X_n)}\right]$$

so we can take a series of samples at $\theta = \theta_0$, replace the expectation on the RHS by an average over those samples, and thereby estimate the LHS for any θ. The rub of course is that the estimator is likely to be a good estimator only for θ near θ_0. What does 'near' mean? Well, this is an experiment and standard statistical methods (e.g. *response surface designs*) can be employed to answer such questions. So-called *bridge sampling* (Meng and Wong, 1996) uses this idea for simulations at two values of θ. Note that these ideas *do* need at least ratios of $a(\theta)$. See exercise 3.

Stochastic Approximation

An alternative is to solve equation (2) by iterative methods, usually called *Robbins-Monro* methods or *stochastic approximation*. Suppose we seek to solve

$$\Phi(\theta) = E\phi(\theta, \epsilon) = 0$$

for increasing Φ, and that we can draw independent samples from $\phi(\theta, \epsilon)$. A sequence of estimates is defined recursively by

$$\theta_{n+1} = \theta_n - a_n\phi(\theta_n, \epsilon_n)$$

for $a_n \to 0$, e.g. $a_n \propto n^{-\gamma}$ for $0 < \gamma \leq 1$. Kushner and Lin (2003) and Ripley (1987, p. 185) gives further details and more sophisticated variants, which include averaging over recent values of θ_n.

The `SIENA` program[2] for fitting models of social networks is almost entirely based on these ideas. These are networks with a finite set of nodes (actors) but with links that evolve through time (e.g. who is 'best friends' with whom in a school). Snijders (2006) writes

> *These models can be simulated on computers in rather straightforward ways (cf. Snijders, 2005). Parameter estimation, however, is more complicated, because the likelihood function or explicit probabilities can be computed only for uninteresting models. This section presents the Methods of Moments estimates proposed in Snijders (2001). [...]*

This is just a Big Name for the idea we have illustrated for the Strauss model, equating empirical and simulated moments, mainly by using stochastic approximation (other ideas including using polysampling and MCMC to approximate likelihoods are under investigations).

[2] 'Simulation Investigation for Empirical Network Analysis': `http://stat.gamma.rug.nl/snijders/siena.html`.

Figure 4.5 Plot of $E_c Y(R)$ against c estimated by simulation. The solid line is without edge correction, the dashed line with toroidal edge correction.

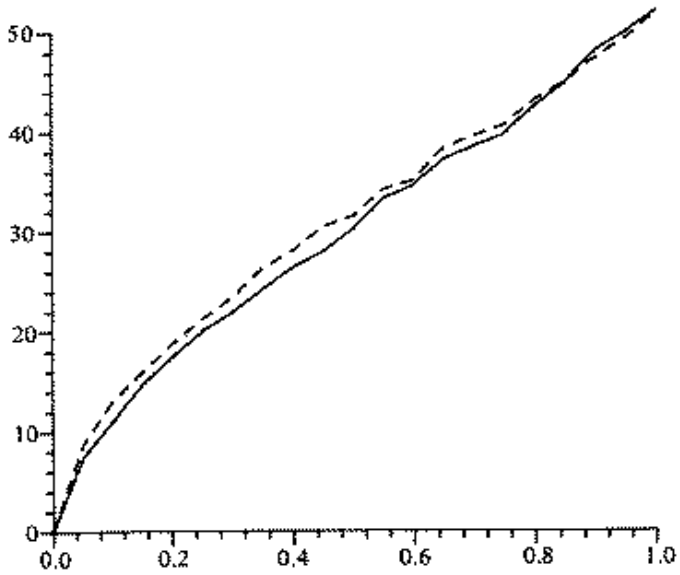

Figure 4.6 An enlarged part of figure 4.5 for toroidal edge correction. The line was fitted by regression.

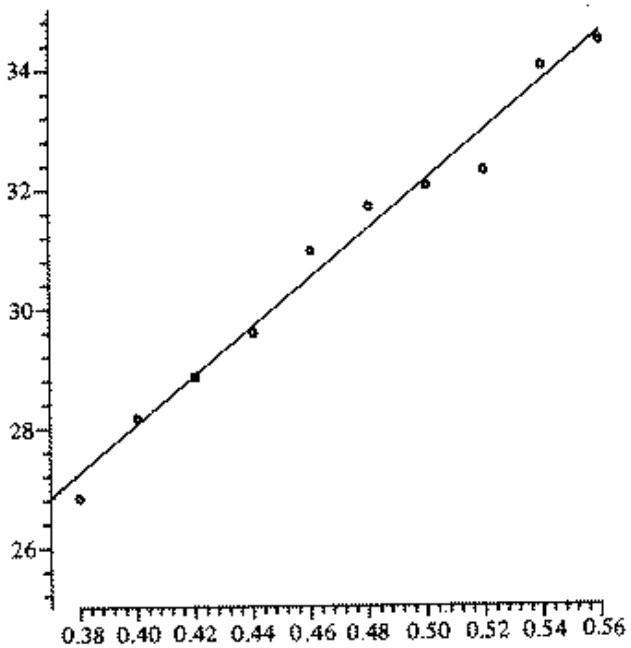

Figure 2: Figures from (Ripley, 1988, p. 72) on fitting the Strauss model (1). These will be explored in the first practical.

Simulated annealing

Simulated annealing is an idea for optimizing functions of many variables, most often discrete variables so a *combinatorial optimization* problem. The name comes from Kirkpatrick *et al.* (1983) and from *annealing*, a process in a metallurgy in which molten metal is cooled extremely slowly to produce a (nearly) stress-free solid. Since annealing is a process to produce a low-energy configuration of the atoms, it is natural[3] to consider its application to optimization of complex problems.

The ground was set by Pincus (1970), based on the idea that if f is continuous over a compact set D and has a unique global maximum at x^* then

$$x^* = \lim_{\lambda \to \infty} \frac{\int_D x \exp\ \lambda f(x)\, dx}{\int_D \exp\ \lambda f(x)\, dx}$$

So if we take a series of samples from density proportional to

$$\exp\ \lambda f(x)$$

for increasing λ, then the distribution of the samples will become increasingly concentrated about x^*. And this procedure is particularly suited to the iterative simulation methods of MCMC since we can use the sample(s) at the previous value of λ to start the iterative process. However, the rate at which λ needs to be increased is very slow, with some studies suggesting that $\lambda \propto \log(1 + t)$ with t the number of iterative steps completed.

Despite the unpromising theoretical behaviour, simulated annealing has proved useful in finding improved solutions to both continuous and combinatorial optimization problems – see e.g. Aarts and Korst (1989)

Finding marginals

A great deal of statistics is about finding marginal distributions of quantities of interest. This occurs in both frequentist and Bayesian settings—especially the latter, where almost all questions boil down to finding a marginal distribution.

Finding those marginals is often difficult, and textbook examples are chosen so that the integrations needed can be done analytically. A great deal of ingenuity has been used in finding systematic ways to compute marginals: examples include the Lauritzen and Spiegelhalter (1988) message-passing algorithm for graphical models.

It is an almost trivial remark that simulation provides a very simple way to compute marginals. Suppose we have a model that provides a joint distribution for a (finite) collection (X_i) of random variables. Then if we have a way to simulate from the joint distribution, taking a subset of the variables provides a painless way to get a marginal distribution of that subset. You should be used to thinking of distributions as represented by samples and so know many ways to make use of that sample as a surrogate for the distribution.

[3] at least to those with some knowledge of statistical physics.

Note that this does not apply directly to marginals in conditional distributions, as we would need to be able to simulate from the conditional distribution. For example, the Lauritzen–Spiegelhalter message-passing algorithm's *raison d'être* is to be able to compute marginals after conditioning on evidence. This is not a problem in the standard Bayesian context where we simulate from the posterior distribution, that is the distribution conditional on the observed data. It is an issue when exploring model fits, where we often want to explore how much one (or more) observation is influencing the results, or even to correct data after discovering large influence.

In the examples we will be using anywhere from a handful to 10,000 samples to represent a marginal distribution. It is important to remember that we only have an approximation to the distribution. A few thousand samples seems like a lot when we are looking at univariate marginals (as people almost invariably do), but we are most often looking at univariate marginals because this is easy to do, not because they are the sole or main interest. In the preliminary exercises you were asked to compare simulations of 71 points in a square with some data – this is a 142-dimensional problem and we have[4] sophisticated multi-dimensional ways to compare such patterns. For another example, ways to look for outliers in multidimensional datasets (Cook and Swayne, 2007) may screen 1,000s or more two-dimensional projections.

SIR

The so-called *sampling-importance resampling* is a technique for improving on an approximate distribution. Suppose we have M samples x_i simulated from an approximation q to a target distribution p. Then *importance sampling* is the idea of estimating

$$E\,h(X) = \int h(x)\,dx = \int h(x)\frac{p(x)}{q(x)}\,dx$$

by the weighted average of $h(x_i)$ with weights $w_i = p(x_i)/q(x_i)$. So we can represent distribution p by a weighted sample from distribution q. For many purposes it is more convenient to have an unweighted sample, and SIR achieves this by taking a subsample of size $m < M$ by weighted sampling without replacement from the current sample. That is we repeat m times

> Select one of the remaining x_i with probability proportional to w_i and remove it from the (x_i).

(See Gelman *et al.*, 2004, pp. 316f, 450.) (Others, including Rubin's original version[5] in the discussion of Tanner and Wong (1987) describe SIR as the version with replacement: the difference will be small if $m \ll M$.) Despite the name, this is a form of rejection sampling.

We have already seen importance sampling used to explore nearby parameter values, and resampling can be used in the same way. Both can be used to perturb Bayesian analyses, e.g. to vary the prior (perhaps away from one chosen for tractability towards something more realistic), as changing the prior just re-weights the posterior samples.

[4] including the human visual system.

[5] by this name: the idea is older.

3 Bootstrapping

Suppose we were interested in inference about the correlation coefficient θ of n IID[6] pairs (x_i, y_i) for moderate n, say 15 (as Efron (1982) apparently was). If we assume that the samples are jointly normally distributed, we might know that there is some approximate distribution theory (using Fisher's inverse tanh transform), but suppose we do not wish to assume normality?

We could do a simulation experiment: repeat R times sampling n pairs and compute their correlation, which gives us a sample of size R from the population of correlation coefficients. **But** to do so, we need to assume both a family of distributions for the pairs **and** a particular parameter value (or a distribution of parameter values).

The *bootstrap* procedure is to take m samples from $\boldsymbol{x}$ *with replacement* and to calculate $\widehat{\theta}^*$ for these samples, where conventionally the asterisk is used to denote a bootstrap resample. Note that the new samples consist of an integer number of copies of each of the original data points, and so will normally have ties. Efron's idea[7] was to assess the variability of $\widehat{\theta}$ about the unknown true θ by the variability of $\widehat{\theta}^*$ about $\widehat{\theta}$. For example, the bias of $\widehat{\theta}$ might be estimated by the mean of $\widehat{\theta}^* - \widehat{\theta}$.

How can we use the bootstrap resamples to do inference on the original problem, and under what circumstances is such inference valid? Note that the bootstrap resample is unlike the original sample in many ways, perhaps most obvious that (with very high probability) it contains ties.

Bootstrapping is most commonly used

- as part of a procedure to produce more accurate confidence intervals for parameters than might be obtained by classical methods (including those based on asymptotic normality). Often this is very similar to using more refined asymptotic statistical theory, and I once heard bootstrapping described as

  'a way to do asymptotics without employing the services of Peter Hall'

- to alleviate biases.

In part because it is so simple to describe and easy to do, bootstrapping has become popular and is often used when it is not valid. For a careful account by two authorities on the subject, see Davison and Hinkley (1997). For another viewpoint by an evangelist of bootstrapping for model validation, see Harrell (2001, Chapter 5). Efron and Tibshirani (1993), Shao and Tu (1995) and Chernick (2008) are complementary material. Hall (1992) covers the (asymptotic) theory. *Statistical Science* **18(2)** (May 2003) has several articles commemorating the *Silver Anniversary of the Bootstrap*.

Efron's original bootstrap is an IID sample of the same size as the original dataset from the *empirical distribution function*. so it is another example of simulation-based inference, using a *non-parametric* rather than *parametric* model of the data. The idea can easily be extended

[6] independent and identically distributed.

[7] Others have claimed priority: for example Simon claims in the preface of Simon (1997) to have discovered it in 1966. See also Hall (2003).

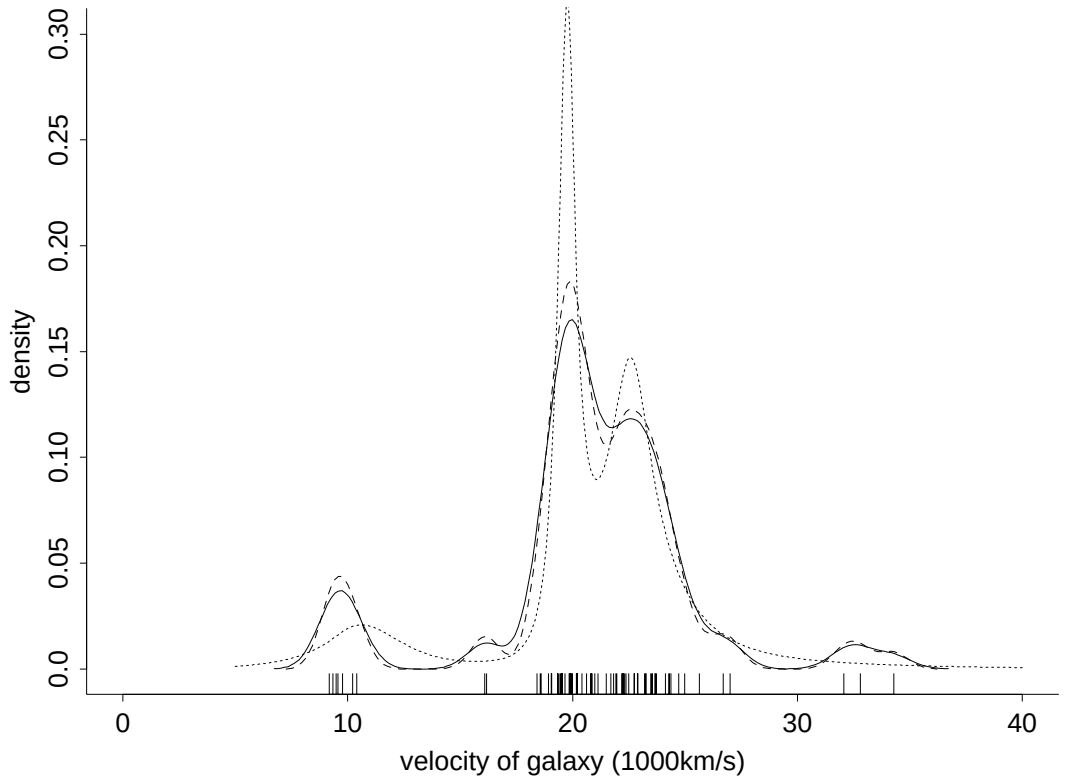

Figure 3: Density estimates for the 82 points of the `galaxies` data. The solid and dashed lines are Gaussian kernel density estimates with bandwidths chosen by two variants of the Sheather–Jones method. The dotted line is a logspline estimate. From Venables and Ripley (2002).

to other non-parametric models, and using a kernel-density estimate[8] of the underlying distribution is called a *smoothed bootstrap*. So that instead of assuming a particular parametric fit, we assume a particular non-parametric fit. This may be more flexible[9], but in both cases we have a *plug-in* estimator—that is we fit a single model from our family and act as if it were the true model.

As a very simple first example, suppose that we needed to know the median m of the `galaxies` data of Roeder (1990), Figure 3. The obvious estimator is the sample median, which is 20 833 km/s. How accurate is this estimator? The large-sample theory says that the median is asymptotically normal with mean m and variance $1/4n\,f(m)^2$. But this depends on the unknown density at the median. We can use our best density estimators to estimate $f(m)$, but we can find considerable bias and variability if we are unlucky enough to encounter a peak (as in a unimodal symmetric distribution). The density estimates give $f(m) \approx 0.13$. Let us try the bootstrap:

```
1/(2*sqrt(length(gal))*0.13)
[1] 0.42474
> library(boot)
> gal.boot <- boot(gal, function(x,i) median(x[i]), R=1000)
```

[8]Note that sampling from a constant-bandwidth kernel-density estimate amounts to resampling from the original data and then adding a random variable drawn from the kernel as a density, so this is the procedure known as *jittering*.

[9]although large parametric families such as spline models for log densities and neural networks can be arbitrarily flexible

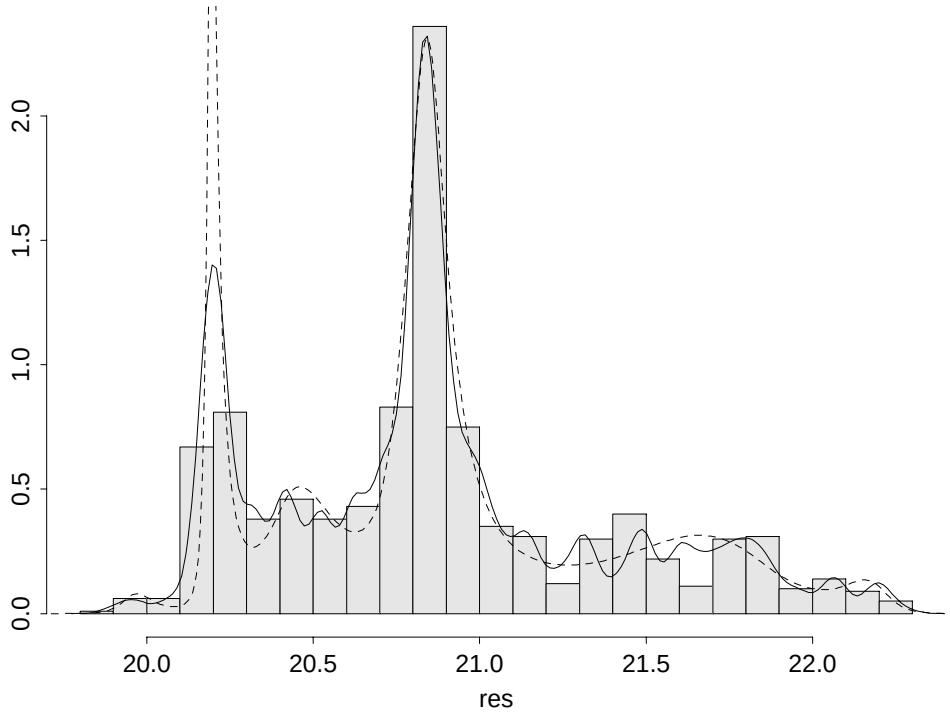

Figure 4: Histogram of the bootstrap distribution for the median of the `galaxies` data, with a kernel density estimate (solid) and a logspline density estimate (dashed). From Venables and Ripley (2002).

```
> gal.boot
Bootstrap Statistics :
    original   bias    std. error
t1*   20.834 0.038747     0.52269
```

which was effectively instant and confirms the adequacy of the large-sample mean and variance for our example (if Efron's idea is correct). In this example the bootstrap resampling can be avoided, for the bootstrap distribution of the median can be found analytically (Efron, 1982, Chapter 10; Staudte and Sheather, 1990, p. 84), at least for odd n. The bootstrap distribution of $\widehat{\theta}_i$ about $\widehat{\theta}$ is far from normal (Figure 4). Choosing the median (a discontinuous function of the data) disadvantaged the simple bootstrap and e.g. a smoothed bootstrap would be preferred in practice.

Note that the bootstrap principle, that the variability of $\widehat{\theta}$ about the unknown true θ can be assessed by the variability of $\widehat{\theta}^*$ about $\widehat{\theta}$, is *not* always valid. For example, the bias of $\widehat{\theta}$ is not mirrored in the bias of $\widehat{\theta}^*$ when bootstrapping density estimation or curve fitting (Bowman and Azzalini, 1997, pp. 44, 82). To quote lecture notes by Peter Hall

> *There is a "meta theorem" which states that the standard bootstrap, which involves constructing a resample that is of (approximately) the same size as the original sample, works (in the sense of consistently estimating the limiting distribution of a statistic) if and only if that statistic's distribution is asymptotically Normal.*
>
> *It does not seem possible to formulate this as a general, rigorously provable result, but it nevertheless appears to be true.*

Although unstated, it seems clear that this is not intended to cover semi-parametric problems

such as density estimation. It hints at other exceptions, for example extreme-value statistics.[10] For more examples, see Davison *et al.* (2003).

The principle is valid often enough that users miss the exceptions and apply it uncritically. To quote the wisdom of Davison and Hinkley (1997, p. 4)

> *Despite its scope and usefulness, resampling must be carefully applied. Unless certain basic ideas are understood, it is all too easy to produce a solution to the wrong problem, or a bad solution to the right one. Bootstrap methods are intended to help avoid tedious calculations based on questionable assumptions, and this they do. But they cannot replace clear critical thought about the problem, appropriate design of the investigation and data analysis, and incisive presentation of conclusions.*

Perhaps the only point here that is particularly apposite to bootstrapping is the "all too easy". It is also easy to apply regression methods to datasets that have other structure (for example, were collected in groups, so mixed-effects models might be more appropriate), and indeed outside simple textbook problems choosing a suitable non-parametric resampling model is no easier than choosing a suitable parametric model. Think about survival problems or time series or spatial patterns or complex surveys for example. Even if missing data are present we have to model the way in which it might occur in other samples.

Contrast this with the embittered comments of Simon (1997)

> *The simple fact is that resampling devalues the knowledge of conventional mathematical statisticians, and especially the less competent ones. By making it possible for each user to develop her/his own method to handle each particular problem, the priesthood with its secret formulaic methods is rendered unnecessary.*

This seems more generally aimed at simulation-based inference than just resampling.

Performance assessment

which is the main part of what Harrell (2001) calls *model validation*. Suppose we have selected a model for, say, regression or classification. Then we expect the model to perform better on our dataset than in future, because both the model (variables used, transformations etc) and the parameter values have been chosen by looking at that dataset. Part of performance assessment is to predict how well the chosen procedure will do in real-world testing—most of the many approaches are discussed in Ripley (1996, §2.7).

One computer-intensive approach is *cross-validation*, to repeatedly keep back a small part of the dataset, do the model selection on the rest and then predict performance on the part held back and then (in some sense) average to estimate the 'out-of-sample' performance. That is a very general method and used sensibly gives very reliable results—but that has not stopped some developers of competitor methods giving it a bad press.

To be concrete, suppose we have a classification procedure and the performance measure is the error rate, the proportion of examples incorrectly classified. Then the 'apparent' misclas-

[10] for which the 'm-out-of-n bootstrap' when $m < n$ and $m/n \to 0$ provides an alternative with valid theory: see Politis *et al.* (1999).

sification rate on the training data will clearly be biased downwards. Estimating how much bias was a simple (and early) application of the bootstrap. Take a series of new training sets by resampling the original, and fit a model to each new training set, and predict at the original training set. The problem here is that the new and original training sets are not distinct, and Efron (1983); Efron and Tibshirani (1997) proposed the '.632' bootstrap. This weights the apparent error rate and the error rate on those original examples which do not appear in the resampled training set as $0.368 : 0.632$. Here 0.632 is shorthand for $(1 - 1/e)$, the large-sample probability that a given example appears in the resampled training set.

Note that here we are trying to estimate what Harrell calls the *optimism* of the whole model fitting procedure. He rightly points out in a quote that

> *In spite of considerable efforts, theoretical statisticians have been unable to analyse the sampling properties of [usual multistep modeling strategies] under realistic conditions*

but then fallaciously goes on to conclude

> *that the modeling strategy must be completely specified and then bootstrapped to get consistent estimates of variances and other sampling properties.*

The fallacy is asserting that bootstrapping is the only game in town, whereas many other simulation-based inference methods could be (and have been) used.

Confidence intervals

One approach to a confidence interval for the parameter θ is to use the quantiles of the bootstrap distributions; this is termed the *percentile confidence interval* and was the original approach suggested by Efron. The bootstrap distribution in our example is quite asymmetric, and the intervals based on normality are not adequate. The 'basic' intervals are based on the idea that the distribution of $\widehat{\theta}^* - \widehat{\theta}$ mimics that of $\widehat{\theta} - \theta$. If this were so, we would get a $1 - \alpha$ confidence interval as

$$1 - \alpha = P(L \le \widehat{\theta} - \theta \le U) \approx P(L \le \widehat{\theta}^* - \widehat{\theta} \le U)$$

so the interval is $(\widehat{\theta} - U, \widehat{\theta} - L)$ where $L + \widehat{\theta}$ and $U + \widehat{\theta}$ are the $\alpha/2$ and $1 - \alpha/2$ points of the bootstrap distribution, say $k_{\alpha/2}$ and $k_{1-\alpha/2}$. (It will be slightly more accurate to estimate these as the $(R+1)\alpha/2$th and $(R+1)(1-\alpha/2)$th ordered values of a sample of size R of $\widehat{\theta}^*$.) Then the basic bootstrap interval

$$(\widehat{\theta} - U, \widehat{\theta} - L) = (\widehat{\theta} - [k_{1-\alpha/2} - \widehat{\theta}], \widehat{\theta} - [k_{\alpha/2} - \widehat{\theta}]) = (2\widehat{\theta} - k_{1-\alpha/2}, 2\widehat{\theta} - k_{\alpha/2})$$

which is the percentile interval reflected about the estimate $\widehat{\theta}$. (This is the same derivation as the Monte-Carlo confidence interval, applied to a non-parametric model.) In asymmetric problems the basic and percentile intervals will differ considerably (as here), and the basic intervals seem more rational. For our example we have

```
boot.ci(gal.boot, conf=c(0.90, 0.95), type=c("norm","basic","perc","bca"))

Level     Normal          Basic         Percentile          BCa
90%   (19.94, 21.65)  (19.78, 21.48)  (20.19, 21.89)  (20.18, 21.87)
95%   (19.77, 21.82)  (19.59, 21.50)  (20.17, 22.07)  (20.14, 21.96)
```

The BC_a intervals are an attempt to shift and scale the percentile intervals to compensate for their biases, apparently unsuccessfully in this example. The idea is that if for some unknown increasing transformation g we had $g(\widehat{\theta}) - g(\theta) \sim F_0$ for a *symmetric* distribution F_0, the percentile intervals would be exact. Suppose more generally that if $\phi = g(\theta)$,

$$g(\widehat{\theta}) - g(\theta) \sim N(-w\,\sigma(\phi), \sigma^2(\phi)) \qquad \text{with } \sigma(\phi) = 1 + a\,\phi$$

Let $U = g(\widehat{\theta})$. Then $U = \phi + (1 + a\,\phi)(Z - w)$ for $Z \sim N(0, 1)$ and hence

$$\log(1 + a\,U) = \log(1 + a\,\phi) + \log(1 + a\,(Z - w))$$

which is a pivotal equation in $\zeta = \log(1 + a\,\phi)$. Thus we have a α confidence limit for ζ, as

$$\zeta_\alpha = \log(1 + a\,u) - \log((1 + a\,(-z_\alpha - w))$$

(using $-z_\alpha = z_{1-\alpha}$) and hence one for ϕ as

$$\phi_\alpha = u + (1 + a\,u)\frac{w + z_\alpha}{1 - a\,(w + z_\alpha)}$$

Then the limit for θ is $\theta_\alpha = g^{-1}(\phi_\alpha)$. We do not know g, but if $u = g(\widehat{\theta})$ and $U^* = g(\widehat{\theta^*})$

$$P^*(\widehat{\theta^*} < \theta_\alpha \,|\, \widehat{\theta}) = P^*(U^* < \phi_\alpha \,|\, u) = \Phi\left(w + \frac{\phi_\alpha - u}{1 + a\,u}\right) = \Phi\left(w + \frac{w + z_\alpha}{1 - a(w + z_\alpha)}\right)$$

Thus the α confidence limit for θ is given by the $\widehat{\alpha}$ percentile of the bootstrap distribution, where

$$\widehat{\alpha} = \Phi\Big(w + \frac{w + z_\alpha}{1 - a(w + z_\alpha)}\Big)$$

Thus if $a = w = 0$ the percentile interval is exact. This is very unlikely, but we can estimate a and w from the bootstrap samples. Now w essentially measures the offset of centre of the distribution, and

$$P^*(\widehat{\theta^*} < \widehat{\theta} \,|\, \widehat{\theta}) = P^*(U^* < u \,|\, u) = \Phi(w)$$

and so w can be estimated by

$$\widehat{w} = \Phi^{-1}\left(\frac{\#\{\widehat{\theta^*} \le \widehat{\theta}\}}{R + 1}\right)$$

Estimating a is a little harder: we make a linear approximation to $\widehat{\theta}$ (as a function of the data points) and take one sixth its skewness (third moment divided by standard deviation cubed) in the bootstrap distribution.

For a smaller, simpler example, consider the data set on page 2.

```
> t.test(B - A)
95 percent confidence interval:
 0.133 0.687
> shoes.boot <- boot(B-A, function(x,i) mean(x[i]), R=1000)
> boot.ci(shoes.boot, type = c("norm", "basic", "perc", "bca"))

Level    Normal          Basic         Percentile         BCa
95%  (0.186, 0.644)  (0.180, 0.650)  (0.170, 0.640)  (0.210, 0.652)
```

There is a fifth type of confidence interval that `boot.ci` can calculate, which needs a variance v^* estimate of the statistic $\widehat{\theta}^*$ from each bootstrap sample. Then the confidence interval can be based on the basic confidence intervals for the *studentized* statistics $(\widehat{\theta}^* - \widehat{\theta})/\sqrt{v^*}$.

```
mean.fun <- function(d, i) {
  n <- length(i)
  c(mean(d[i]), (n-1)*var(d[i])/n^2)
}
> shoes.boot2 <- boot(B - A, mean.fun, R = 1000)
> boot.ci(shoes.boot2, type = "stud")

Level  Studentized
95%   (0.138, 0.718)
```

Some caution is needed here. First, despite three decades of work and lots of theory that suggest bootstrap methods are well-calibrated, we can get as large discrepancies as we have here. Second, this is only univariate statistics, and standard bootstrap resampling is only applicable to IID samples.

Note that the bootstrap distribution provides some diagnostic information on the assumptions being made in the various confidence intervals.

Theory for bootstrap confidence intervals

There are two ways to compare various types of bootstrap confidence intervals. One is empirical comparisons as above. Another is to work out the asymptotic theory to a high enough level of detail to differentiate between the methods.

The principal property of a confidence interval or limit is its coverage properties. We want for an upper confidence limit $\widehat{\theta}_\alpha$ that

$$P_\theta(\theta \le \widehat{\theta}_\alpha) = \alpha + O(n^{-a})$$

for a large a. Obviously we would like exact confidence limits (no remainder term), but they are in general unattainable, and so we aim for the best possible approximation.

Two points to note: a confidence interval or limit can have good coverage but be far from optimal (as measured by length, say), and a confidence interval can have good coverage but be far from equi-tailed.

We can consider a hierarchy of methods of increasing accuracy.

- A normal-based confidence interval, for example with standard deviation based on the bootstrap distribution, and with a bootstrap bias correction. This has $a = 1/2$ (in general, but we won't keep mentioning that).
- Basic bootstrap confidence limits. These have $a = 1/2$, *but* confidence intervals have $a = 1$ and are said to be *second-order accurate*.
- Percentile limits and intervals. The same story as the basic limits and intervals. As we have seen empirically, both of these tend to fail to centre the interval correctly, so achieve second-order accuracy for intervals at the expense of having unequal tails.
- BC_a limits and intervals both have $a = 1$, provided a and w are estimated to $O(n^{-1/2})$.

- The studentized method also has $a = 1$, provided the variance estimate used is accurate to $O(n^{-1/2})$.

To make the limitations of these results clearer, note that they apply equally to any monotone transformation $\phi(\theta)$ (for example a log or arcsin or logistic transformation), but empirical studies (e.g. Davison and Hinkley, 1997, §5.7) show that using the right transformation can be crucial.

Double bootstrapping

Another idea is to re-calibrate a simpler confidence limit or interval, that is to use $\widehat{\theta}_\beta$ for some $\beta \neq \alpha$ to achieve more accurate coverage properties. How do we choose β? The double bootstrap uses a second tier of bootstrapping to estimate the coverage probability of $\widehat{\theta}_\beta$, and then solves for β on setting this estimate equal to α.

It transpires that applying this idea to the normal confidence interval produces the studentized confidence interval, but applied to the basic confidence interval it produces coverage accurate to $a = 2$.

Double bootstrapping appears to require large numbers of replications, say a million samples if we take 1000 in each of the tiers. Fortunately this can be reduced to more manageable numbers by using *polysampling* (page 5, Davison and Hinkley (1997, §9.4.4) and Morgenthaler and Tukey (1991)).

Bootstrapping linear models

In statistical inference we have to consider what might have happened but did not. Linear models can arise exactly or approximately in a number of ways. The most commonly considered form is

$$Y = X\beta + \boldsymbol{\epsilon}$$

in which only $\boldsymbol{\epsilon}$ is considered to be random. This supposes that in all (hypothetical) repetitions the same x points would have been chosen, but the responses would vary. This is a plausible assumption for a designed experiment and for an observational study with pre-specified factors.

Another form of regression is sometimes referred to as the *random regressor* case in which the pairs (x_i, y_i) are thought of as a random sample from a population and we are interested in the regression function $f(x) = E\{Y \mid X = x\}$ which is assumed to be linear. However, it is common to perform conditional inference in this case and condition on the observed xs, converting this to a fixed-design problem. For example, in the Scottish hill races dataset[11] the inferences drawn depend on whether certain races, notably Bens of Jura, are included in the sample. As they were included, conclusions conditional on the set of races seems most pertinent.

[11] Venables and Ripley (2002, p. 8–10, 152–5).

These considerations are particularly relevant when we consider bootstrap resampling. The most obvious form of bootstrapping is to randomly sample pairs (x_i, y_i) with replacement,[12] which corresponds to randomly weighted regressions. However, this may not be appropriate in not mimicking the assumed random variation and in some examples of producing singular fits with high probability. The main alternative, *model-based resampling*, is to resample the residuals. After fitting the linear model we have

$$y_i = x_i\widehat{\beta} + e_i$$

and we create a new dataset by $y_i = x_i\widehat{\beta} + e_i^*$ where the (e_i^*) are resampled with replacement from the residuals (e_i). There are a number of possible objections to this procedure. First, the residuals need not have mean zero if there is no intercept in the model, and it is usual to subtract their mean. Second, they do not have the correct variance or even the same variance. Thus we can adjust their variance by resampling the *modified residuals* $r_i = e_1/\sqrt{1-h_{ii}}$ which have variance σ^2.

We see bootstrapping as having little place in least-squares regression. If the errors are close to normal, the standard theory suffices. If not, there are better methods of fitting than least-squares! One issue that is often brought up is that of heteroscedasticity, which some bootstrap methods accommodate—but then so do Huber–White 'sandwich' estimators.

The distribution theory for the estimated coefficients in robust regression is based on asymptotic theory, so we could use bootstrap estimates of variability as an alternative. Resampling the residuals seems most appropriate for the `phones` data of Venables and Ripley (2002, p. 157)

```
library(MASS); library(boot)
fit <- lm(calls ~ year, data=phones)
summary(fit)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -260.059    102.607  -2.535   0.0189
year           5.041      1.658   3.041   0.0060

ph <- data.frame(phones, res=resid(fit), fitted=fitted(fit))
ph.fun <- function(data, i) {
  d <- data
  d$calls <- d$fitted + d$res[i]
  coef(update(fit, data=d))
}
ph.lm.boot <- boot(ph, ph.fun, R=999)
ph.lm.boot
    ....
     original   bias    std. error
t1* -260.0592  6.32092    100.3970
t2*    5.0415 -0.09289      1.6288

fit <- rlm(calls ~ year, method="MM", data=phones)
summary(fit)
```

[12] Davison and Hinkley (1997) call this *case-based resampling*. Shao and Tu (1995) call it the *paired bootstrap*, in contrast to the *residual bootstrap* we consider next.

```
Coefficients:
            Value    Std. Error t value
(Intercept) -52.4230   2.9159   -17.9783
year          1.1009   0.0471    23.3669

ph <- data.frame(phones, res=resid(fit), fitted=fitted(fit))
ph.rlm.boot <- boot(ph, ph.fun, R=999)
ph.rlm.boot
    ....
    original     bias  std. error
t1* -52.4231  3.232142   30.01894
t2*   1.1009 -0.013896    0.40648
```

(The `rlm` bootstrap is starting to be computer-intensive at 45 secs, but took about 10 mins in S-PLUS for the third edition in 1997.) These results suggest that the asymptotic theory for `rlm` is optimistic for this example, but as the residuals are clearly serially correlated the validity of the bootstrap results is equally in doubt. Statistical inference really does depend on what one considers might have happened but did not.

Shao and Tu (1995, Chapters 7–8) and Davison and Hinkley (1997, Chapters 6–7) consider linear models and extensions such as GLMs and survival models.

How many bootstrap resamples?

Our examples have been based on 1000 resamples θ^*. Largely that figure was plucked from thin air: it was computationally feasible. Is it enough? To answer that question we need to consider the various sources of error. The Monte-Carlo error due to the resampling is one, and since the resampling is independent, the size of the Monte Carlo error will be $O_P(R^{-1/2})$ for R samples. On the other hand, the size of the confidence interval will be $O_P(n^{-1/2})$ and this suggests that to make the Monte Carlo error negligible we should take R to be some multiple of n. Some calculations (Davison and Hinkley, 1997, pp. 35–6) suggest a multiple of 10–40.

However, this is not all there is to it. If we want to compute confidence intervals, we need to be able to estimate fairly extreme quantiles, which suggests we need R around 1000. ($R = 999$ is popular as the $(R+1)p$th quantile is a data point for most popular p.) Moreover, the BC_a method often needs considerably more extreme quantiles than the originals, and so can require very large bootstrap samples.

Further, there are other sources of (systematic) error, not least the extent to which the distribution of $\theta^* - \widehat{\theta}$ mimics that of $\widehat{\theta} - \theta$. It may be more important to choose a (monotone invertible) transformation $\phi(\theta)$ for which that approximation is more accurate, in particular to attempt variance stabilisation. Finally, if the samples were not independent, then simple bootstrapping will be inappropriate.

Diagnostics

It is (I hope) familiar that after fitting a parametric model such as a regression we look at diagnostics to see if the assumptions have been violated.

In one sense the problem is easier with bootstrap models as they are non-parametric and so make fewer assumptions. However *fewer* but often *more* critical ones, e.g. independent and identically distributed. So we would still like to know if there are observations that are particularly influential on our conclusions. That is hard enough for linear regression!

There is one general approach to bootstrap diagnostics, the use of *jackknife-after-bootstrap* (Davison and Hinkley, 1997, §3.10). We could consider dropping each observation j in turn and redoing the analysis (including all the resampling). However, we can avoid this by looking only at bootstrap resamples in which observation j does not occur and applying the jackknife method of the next section.

The Jackknife

The *jackknife*[13] is an older resampling idea, most often used to attempt bias reduction.

Consider an i.i.d. sample $X_1, \ldots, X_n \sim F$, and consider a statistic $\theta_n = \theta(F_n)$ for the empirical CDF F_n.. Then the 'true value' is $\theta(F)$. Any statistic which is independent of the ordering of the sample can be written in this form.

The jackknife estimates the bias from the n sub-samples of size $n-1$. Let $\widehat{\theta}_{(i)}$ denote the estimate from the sample omitting X_i. Then the estimate of bias is

$$\widehat{BIAS} = (n-1)\Big(\frac{1}{n}\sum \widehat{\theta}_{(i)} - \widehat{\theta}\Big)$$

It can be shown that if

$$\text{bias}(\widehat{\theta}) = E(\widehat{\theta}) - \theta(F) = a_1/n + a_2/n^2 + \cdots$$

then $\widehat{\theta} - \widehat{BIAS}$ has a bias of $O(n^{-2})$. For example, applied to the variance functional it replaces the divisor n by $n-1$. The assumption will be true for smooth functionals θ, but not, for example, for the median.

Another way to look at the jackknife is *via* the *pseudo-values* given by

$$\tilde{\theta}_i = \widehat{\theta} + (n-1)(\widehat{\theta} - \widehat{\theta}_{(i)})$$

Then these can be regarded as a new 'sample', and the mean and variance estimated from them. Then $\widehat{BIAS}$ is $\widehat{\theta}$ minus the mean of the pseudo-values.

The jackknife estimate of the variance of a statistic is the variance of the pseudo-values divided by n. The pseudo-values could be used to give a confidence interval, but that has proved to be no better than the normal-based confidence intervals based on the mean and variance of the pseudo-values.

Jackknife ideas have been frutiful for bias and (especially) variance estimation, for example in sample surveys—see Rao and Wu (1988) and Shao and Tu (1995, Chapter 6). (Note that this is another area where naïve applications of the bootstrap may be invalid.)

[13] a Tukey-ism for an idea of Quenouille

Bootstrapping as simulation

Bootstrapping is 'just' simulation-based inference, sampling from a particularly simple model. As such it is subject to all the ways known in the simulation literature[14] to reduce variability. Davison and Hinkley (1997, Chapter 9) discuss many of them, but there are few examples in the many application studies using bootstrapping. If you use bootstrapping in your work, please take heed!

Software

Basic bootstrap resampling is easy, as you saw in the preliminary notes—e.g. just use the R function `sample`.

As for most uses of simulation-based inference the task for software falls into two halves

- Generate the simulations, in this case the resamples.
- Analyse the simulated data.

Once we move away from looking at univariate IID sampling, both become more complicated and less general.

R ships with a package `boot` which is support software for Davison and Hinkley (1997) written[15] largely by Angelo Canty. This has a few functions to do bootstrapping such as `boot`, `censboot` and `tsboot`. The workhorse here is `boot`, which allows several types of resampling/simulation.

`sim="ordinary"` which has `stype` as one of `"i"` (give indices into the data set), `"f"` (give frequencies for each item) and `"w"` (normalized to one)

`sim="parametric"` see below.

`sim="balanced"` stratification.

`sim="permutation"` resampling without replacement.

`sim="antithetic"` induce negative correlations in pairs of bootstrap resamples.

It is also possible to specify strata, and importance sampling weights.

`sim="parametric"` is intended for parametric bootstrapping, but is a completely general mechanism. Here is how it is used for the smoothed bootstrap in the solutions to the preliminary exercises:

```
s <- 0.1 # the standard deviation of a normal kernel
ran.gen <- function(data, mle) {
    n <- length(data)
    rnorm(n, data[sample(n, n, replace=TRUE)], mle)
}
out3 <- boot(nerve, median, R=1000, sim = "parametric",
             ran.gen = ran.gen, mle = s)
```

[14] and sketched in the preliminary material.

[15] for S-PLUS, and ported to R by me.

Here `mle` is an object representing the parameters to be passed to the simulation routine. The main analysis function is `boot.ci`.

4 Markov Chain Monte Carlo

The idea of *Markov Chain Monte Carlo* is to simulate from a probability distribution as the stationary distribution of a Markov process. This is normally[16] employed for quite highly structured problems, typically involving large numbers of dependent random variables. Such problems first arose in statistical physics, and the ideas were re-discovered in spatial statistics in the 1970s and 1980s. Then those wanting to implement Bayesian models jumped on the bandwagon around 1990, rarely giving credit to those whose work in spatial statistics they had taken the ideas from.

The key questions about MCMC from a practical viewpoint are

1. How do we find a suitable Markov process with our target distribution π as its stationary distribution?
2. Assuming we cannot start from the stationary distribution (since if we could we would know another way to simulate from the process), how rapidly does the process reach equilibrium? And how can we know that it is already close to equilibrium?
3. How correlated are successive samples from the process, or (to put it another way), how far apart do we need to take samples for them to contain substantially different information?

These points are all interrelated—a good MCMC sampling scheme will be one for which each step is computationally quick, and which *mixes* well, that is traverses the sample space quickly.

The previous paragraph assumes that these goals are achievable, but people do attempt to use MCMC in problems with millions of random variables. Almost inevitably there are some aspects of the process that mix slowly and some that mix fast, and so the choice of MCMC sampling scheme does often need to be linked to the questions of interest.

MCMC can be approached from several angles. The preliminary material took on one approach based on my personal experience, and for variety these notes take another.

Some of the statements made here about convergence need technical conditions which are omitted. It is generally accepted that the cases that are being excluded are pathological, and since MCMC allows a lot of freedom to design a suitable scheme the conditions are easily satisfied in practice. The clearest and most accessible account of the relevant theory I have seen is Roberts and Rosenthal (1998).

Data augmentation

Suppose we have a parametric model $p(Y \mid \theta)$ for some observable random variables Y. It is rather common for this to be the manifestation of a richer model $p(Y, Z \mid \theta)$ for both the *manifest* variables Y and some *latent* (unobserved, 'missing') variables Z. This can arise in many ways, including

- Missing data, so Z represents e.g. responses from a survey that were unobserved.

[16] apart from in textbook examples and exercises.

- Partial observation, e.g. in social networks we only observe the links at some times: in family studies we have genetic data on only some members.
- Censored data, e.g. lifetimes in which all we know for some subjects is that they were still alive on a particular date. So for each subject we have two pieces of information, whether they were alive at the end of the study, and the actual date of death. For all subjects the first is part of Y whereas for some the second is part of Y and for some part of Z.
- Latent variable/class problems in which Z is some unobserved 'true' characteristic such as intelligence or the component of a mixture distribution. In genetics Y might be the phenotype and Z the genotype.

For simplicity of exposition we will take a Bayesian viewpoint with a prior probability distribution on θ, and the main object of interest is then the posterior distribution $g(\theta) = p(\theta \mid Y)$. Note that

$$g(\theta) = p(\theta \mid Y) = \int p(\theta \mid Y, Z) p(Z \mid Y)\, dZ$$

and

$$P(Z \mid Y) = \int p(Z \mid \theta, Y) p(\theta \mid Y)\, d\theta$$

and hence g satisfies

$$g(\theta) = \int K(\theta, \phi)\, g(\phi)\, d\phi, \quad \text{where} \quad K(\theta, \phi) = \int p(\theta \mid Y, Z)\, p(Z \mid \phi, Y)\, dZ \tag{3}$$

Under mild conditions we can solve (3) by successive substitution,[17] but we do have to integrate out the unobserved variables Z. Tanner and Wong (1987) (see also Tanner, 1996) call a Monte Carlo version *data augmentation*. This alternates the steps

a. Generate a sample (z_i) of size m from the current approximation to $p(Z \mid Y)$. This will probably be done by first sampling θ_i^* from the current approximation $g(\theta)$ and then sampling z_i from $p(z \mid \theta_i^*, Y)$.

b. Use this sample to update the approximation to $g(\theta) = p(\theta \mid Y)$ as the average of $p(\theta \mid z_i, Y)$.

So what this is doing is approximating $p(\theta \mid Y)$ by a finite mixture from $(p(\theta \mid z, Y))$. As iteration progresses we might want to take larger and larger samples to get better approximations.

This is closely related to the notion of *multiple imputation* in the analysis of sample surveys, where missing data are replaced by a sample of their uncertain values. So data augmentation alternates between multiple imputation of the unobserved variables in the model and inference based on the augmented data. From a theoretical viewpoint, the multiple imputations are being used to approximate the integral in the definition of K at (3) by an average over samples.

However, we can take another point of view, as K is the transition kernel of a Markov chain, and successive substitution will converge to the stationary distribution of that Markov chain. Suppose that we just simulate from the Markov chain? This alternates

[17] start with some candidate g for $p(\theta \mid Y)$, and repeatedly use (3) to obtain a new and better candidate. Under mild conditions this does work – there is a unique solution, the new candidate is closer in L_1 norm to that solution and convergence is geometric.

a. Generate a single sample z from $p(Z \mid \theta, Y)$ with the current θ.

b. Use z to sample θ from $p(\theta \mid z, Y)$.

In this version we give up both multiple imputation and any attempt to keep probability distributions in partially analytical form—rather we represent distributions by a single sample, and run the Markov chain as a stochastic process on parameter values θ (rather than iterating an integral operator). This variant is called *chained data augmentation* by Tanner (1996). Clearly we would eventually want more than one sample, but we can get that by simulating the whole Markov chain multiple times, rather than simulating each step multiple times.

In a *particle filter*[18] evolving distributions are represented by a finite set of values, not just one, that is by a finite mixture, usually but not always unweighted.

The observable data Y have played a passive rôle throughout this subsection: what we have been considering is a way to simulate from the joint distribution of (θ, Z) conditional on Y. So we do not need an explicit Y, and 'chained data augmentation' gives us a way to simulate from any joint distribution of two (groups of) random variables by alternately simulating from each of the two conditional distributions of one conditioned on the other.

Detailed balance

Data augmentation and the spatial birth-and-death processes of the preliminary notes provide 'mechanistic' approaches to developing an MCMC algorithm, but in general MCMC algorithms can be unrelated to any hypothesized stochastic generative mechanism. Especially in such cases, we need to be able to show formally[19] that we do indeed have a Markov process with the desired stationary distribution, and that the stationary distribution is the limiting distribution.

A key concept is *detailed balance*, which is connected to *reversibility* of the Markov process. Reversibility just means that the joint distribution of the process at a series of times is unchanged if the direction of time is reversed—clearly this only makes sense for a stationary process as for any other Markov process the convergence towards equilibrium reveals the direction of time.

For a discrete-time discrete-state-space Markov process reversibility entails

$$P(X_t = i, X_{t+1} = j) = P(X_{t+1} = i, X_t = j) = P(X_t = j, X_{t+1} = i)$$

so if (π_i) is the stationary distribution,

$$\pi_i P_{ij} = \pi_j P_{ji} \tag{4}$$

for transition matrix P_{ij}. This equation is known as *detailed balance*.

If we know there is a unique stationary distribution, and we can show detailed balance for our distribution π, we have shown that it is the unique stationary distribution. If we also know[20]

[18] http://en.wikipedia.org/wiki/Particle_filter, Robert and Casella (2004, Chapter 14).

[19] for some value of 'formal'!

[20] e.g. by showing it is aperiodic and irreducible, and for continuous state-spaces Harris recurrent.

that the Markov process converges to its stationary distribution, we have a valid MCMC sampling scheme.

Similar considerations apply to continuous-state-space Markov processes, e.g. detailed balance can apply to the density of the stationary distribution.

Gibbs sampler

so named by Geman and Geman (1984) but published some years earlier by Ripley (1979) and as examples in earlier papers.

It applies to a multivariate distribution, so we can think of Y as m-dimensional. The simplest Gibbs sampler consists of selecting a random component i of Y, and replacing Y_i by a sample from $p(Y_i \mid Y_{-i})$, where Y_{-i} denotes all the variables *except* Y_i.

This can easily be shown to satisfy detailed balance.

Chained data augmentation is a simple example of the Gibbs sampler. It alternately samples from the conditional distributions of Z and θ given the remaining variables.

In practice the Gibbs sampler is often used with a systematic selection of i rather than a random one (as in chained data augmentation). The theory is then not so simple as the process is no longer necessarily reversible—this is discussed in Geman and Geman (1984) and some[21] of the references. One simple modification that makes the process reversible is to use a systematic order of the m components, and then run through them in reverse order (chained DA is an example).

When we have an m-dimensional distribution, it is not necessary to think of each component in the Gibbs sampler as a single random variable. Sometimes the variables naturally form blocks, and it is the blocks to which the Gibbs sampler should be applied. Once again, chained DA provides the simplest example.

Note that the Gibbs sampler does not necessarily converge to the stationary distribution: there are conditions which need to be checked and are related to when a joint distribution is determined by all of its univariate conditionals. Consider the simple example of a two-dimensional joint distribution of (X, Y) in which X has a standard normal distribution and $Y = X$.

Metropolis-Hasting schemes

A general way to construct a Markov chain with a given stationary distribution π was given by Metropolis *et al.* (1953) which was given added flexibility by Hastings (1970).

These MCMC schemes start with a transition kernel $q(x, y)$ of a Markov process on the state space. Given a current state Y_t this is used to generate a candidate next state Y^*. Then *either* the transition is accepted and $Y_{t+1} = Y^*$ *or* it is not when $Y_{t+1} = Y_t$. The probability that the move is accepted is $\alpha(Y_t, Y^*)$ where

$$\alpha(x, y) = \min\left\{1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right\}$$

[21] e.g. Gamerman and Lopes (2006, §5.3.2).

It is a simple exercise to show that this satisfies detailed balance. For the stationary distribution to be also the limiting distribution we need the chain to be aperiodic: note that it *will* be aperiodic if there is a positive probability of rejecting a move.

The original Metropolis *et al.* scheme had a symmetric transition kernel, so the move is accepted with probability $\min\{1, \pi(x)/\pi(y)\}$. That is, all moves to a more or equally plausible state are accepted, but those to a less plausible state are accepted only with a probability less than one, the ratio of the probabilities.

That only the ratio of the probabilities enters is often exploited. If x is a high-dimensional state vector, choosing transitions such that y differs from x only in one or a few components can simplify greatly the computation of $\pi(Y^*)/\pi(Y_t)$, and also avoid rejecting most proposed moves (which will happen if $\pi(Y^*)$ is almost always very much smaller than $\pi(Y_t)$). Indeed, the Gibbs sampler is a special case of the Metropolis-Hastings sampler in which only single-component moves are considered, and $q(x, y) = p(x_i \,|\, x_{-i})$ where i is the chosen component (and hence $\alpha(x, y) \equiv 1$).

A couple of other special cases are worth mentioning. One suggested by Hastings (1970) and others is a *random-walk sampler* in which q specifies a random walk (and so makes most sense when the state space is a vector space, but could apply to a lattice). Another is an *independence sampler* in which $q(x, y) = q(y)$, so the proposed move is independent of the current state.

For a gentle introduction to the many choices in implementing a Metropolis-Hastings MCMC scheme see Chib and Greenberg (1995).

Other schemes

The only limit on the plethora of possible MCMC schemes is the ingenuity of developers. We saw another scheme, spatial birth-and-death processes, in the preliminary notes. A similar idea, the *reversible jump MCMC* of Green (1995), has been applied to model choice in a Bayesian setting.

We do not even need to confine attention to Markov processes which jump: Grenander and Miller (1994) and others have used Langevin methods, that is diffusions. See Robert and Casella (2004, §7.8.5) for a brief account.

Using a MCMC sampler

So far we have described using a Markov chain to obtain a single sample from a stochastic process by running it for an infinite number of steps. In practice we run it for long enough to get close to equilibrium (called a 'burn-in' period) and then start sampling every $m \geq 1$ steps (calling *thinning*). We can estimate any distributional quantity *via* the law of large numbers

$$\frac{1}{N} \sum_{i=1}^{N} h(X_{mi}) \to E\, h(\boldsymbol{X})$$

for any m, so if $h()$ is cheap to compute we may as well average over all steps. In practice we often take m large enough so that samples are fairly dissimilar—thinning is also used to reduce storage requirements.

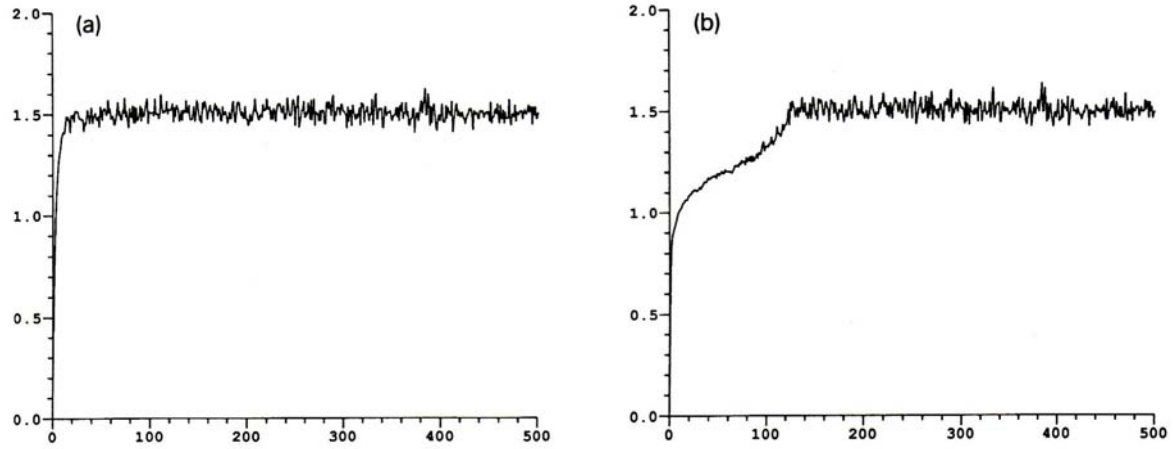

Fig. 2. Pseudolikelihood (a) and (asymptotic) maximum likelihood (b) estimates for 500 sweeps of Metropolis' method.

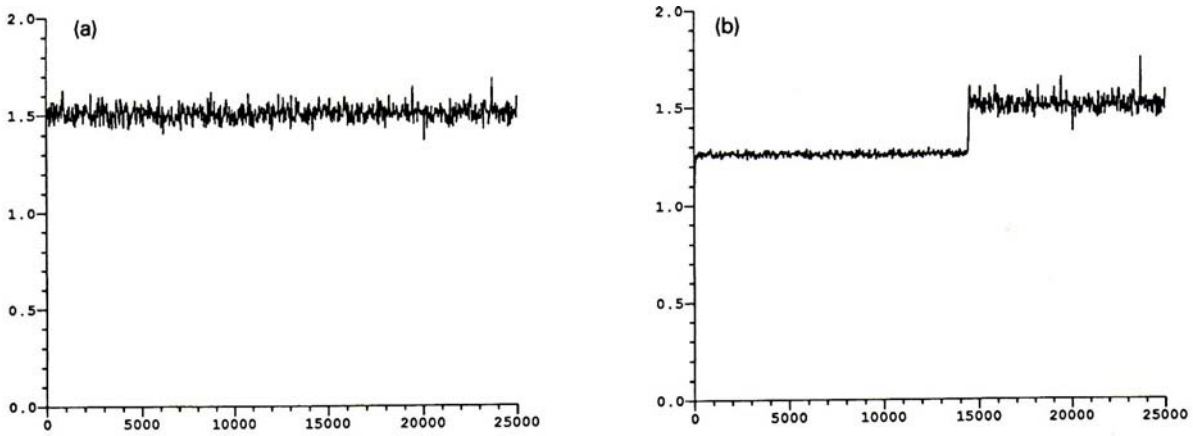

Fig. 3. As Fig. 2, but another sample with 25,000 sweeps!

Figure 5: Diagnostic plots from two realizations of an MCMC simulation. Note the different scales. These are for two estimators of a quantity known to be $\beta = 1.5$. From Ripley and Kirkland (1990).

There are many practical issues – where do we start? How do we know when we are 'close to equilibrium'? And so on. Note that the issue of whether we are yet close to equilibrium is critical if we are simulating to get an idea of how the stochastic process behaves – Geman and Geman (1984) based all their intuition on processes which were far from equilibrium, but incorrect intuition led to interesting statistical methods.

A run of an MCMC algorithm provides a time series of correlated observations. There is a lot of earlier work on analysing such time series from other simulation experiments, for example of queueing problems: see Ripley (1987, Chapter 6). Most of these need a Central Limit Theorem, which hold if the Markov chain is geometric ergodic, for example. (Roberts and Rosenthal (1998, p.10) give an example of an MCMC scheme where the CLT fails to hold.)

Convergence diagnostics

Or 'How do we know when we are close to equilibrium?'

This led to much heated discussion in the early 1990s, and several survey papers. The scale

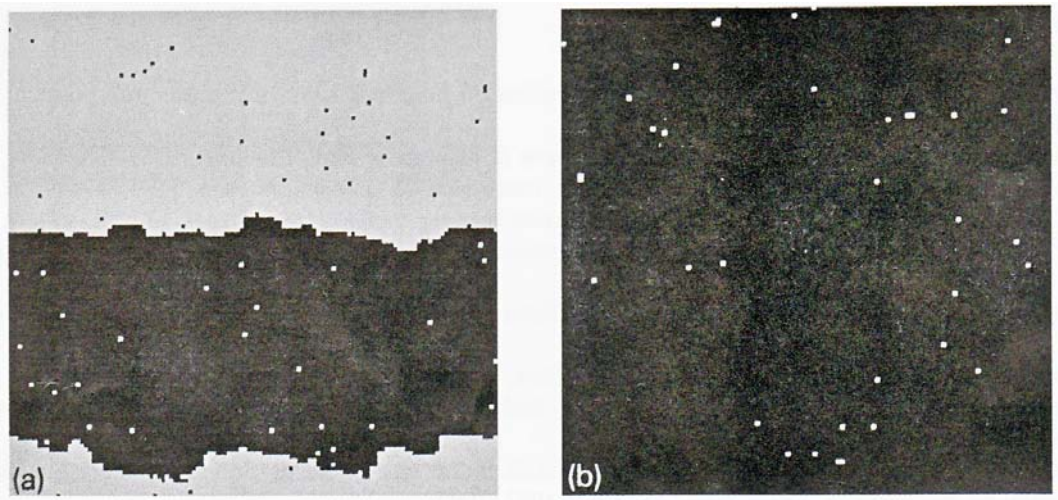

Fig. 4. The realization after 10,000 (a) and 25,000 (b) sweeps of Metropolis' method.

Figure 6: Two snapshots of the second MCMC simulation. From Ripley and Kirkland (1990).

of the problem is often dramatically underestimated – twenty years ago we found an example (Ripley and Kirkland, 1990) in which the Gibbs sampler appeared to have converged after a few minutes, but jumped to a very different state after about a week. In statistical physics such behaviour is sometimes call *metastability*.

The proponents have split into two camps, those advocating running a single realization of the chain, and after a 'burn-in' period sampling it every m steps, and those advocating running several parallel realizations, and taking fewer samples from each. Note that the computing environment can make a difference, as the simplest and computationally most efficient way to make use of multiple CPUs is to use parallel runs.

Writing about this, Robert and Casella (2004) (which is a second edition) say (p. X)

> *We also spend less time on convergence control, because some of the methods presented [in the 1999 first edition] did not stand the test of time. The methods we preserved in Chapter 12 have been sufficiently tested to be considered reliable.*

and (p. 512)

> *Chapter 12 details the difficult task of assessing the convergence of an MCMC sampler, that is, the validity of the approximation $\theta^{(T)} \sim \pi(x)$, and, correspondingly, the determination of a "long enough" simulation time. Nonetheless, the tools presented in that chapter are mostly hints of (or lack of) convergence, and they very rarely give a crystal-clear signal that $\theta^{(T)} \sim \pi(x)$ is valid for all simulation purposes.*

Readers of other accounts (including their first edition) may come away with a very different impression.

If we knew something about the rate of convergence of the Markov chain to equilibrium we could use such knowledge to assess how long the 'burn-in' period needed to be. But this is very rarely helpful, for

(i) we rarely have such knowledge,

(ii) when we do it is in the form of upper bounds on convergence rates and those upper bounds are normally too crude, and

(iii) the theory is about convergence from any initial distribution of all aspects of the distribution. Many of the MCMC schemes converge fast for some aspects of the target distribution and slowly for others—hopefully the scheme was chosen so that the former are the aspects we are interested in.

Nevertheless, that are some exceptions: e.g. simple ones in Roberts and Rosenthal (1998, §5) and an application to randomized graph-colouring algorithms in Jerrum (1995) (see also Asmussen and Glynn, 2007, §XIV.3).

After all those notes of caution, here are some of the main ideas. Let (X_t) be the output from a single MCMC run, possibly sub-sampled every m steps and of one (usually) or more aspects of interest.

- *Tests of stationarity*. If the output is stationary, we can divide into two or more parts which will have the same distribution, and apply a test for equality of distribution such as the Kolomogorov–Smirnov test. Such tests are usually most sensitive to changes in location (which is normally of most interest here), and for IID samples (and so need adjustment). Tests of drift such as CUSUM charts (Yu and Mykland, 1998) come into this category.

- *Regeneration*. Some of the most powerful ideas in the analysis of discrete-event simulations (Ripley, 1987, Chapter 6) are based on the idea that the process will from time-to-time come back to an identifiable state and excursions from that state are independent (by the strong Markov property). (Think for example of a queueing system emptying completely.) Regeneration may be too rare to be useful, but this is one of the few fully satisfactory approaches.

- *Coverage*. The idea is to assess how much of the total mass of π has been explored. For a one-dimensional summary and sorted values $X_{(t)}$ the Riemann sum

$$\sum_{t=1}^{T} |X_{(t+1)} - X_{(t)}| \, \pi(X_{(t)})$$

  provides an approximation to $\int \pi(x)dx = 1$, and so its convergence to one is a measure of coverage of the MCMC to date. This is only applicable if there is a one-dimensional summary of which we know the marginal distribution explicitly (so we can evaluate $\pi(X_{(t)})$), and it only tells us about coverage of that marginal.

- *Multiple chains*. If we have a small number of runs from suitable starting points we can compare the variability within and between runs, and when the between-run variability has reduced to that predicted from the within-run variability all the runs should be close to equilibrium. The series (X_t) is autocorrelated, and we need to take that into account in assessing the within-run variability: but that is a standard problem in the simulation literature. This approach is principally associated with Gelman and Rubin (1992). The problem is to choose suitable starting points so the runs considered do representatively sample π.

- *Discretization*. Some methods look at a discretization of (X_t) to a process with a small number of states. The original proposal by Raftery & Lewis was reduce to a two-state

process. The discretized process will not normally be Markov, but a sub-sampled process (every m steps) might be approximately so and *if so* we know enough about two-state Markov chains to study their convergence, estimating the two parameters of the transition matrix from the observed data. The issues are the Markov approximation and whether convergence of the discretized version tells us enough useful about convergence of the original (although non-convergence definitely does).

Another cautionary note: these diagnostic tests must not be used as stopping rules, as that would introduce bias.

Quite a lot of software has been written for convergence diagnostics. Two of the main suites, `coda` and `boa`, are available as R packages.

There are some methods for using MCMC to produce a sample exactly from π. Propp and Wilson (1996) called these *exact sampling*, but Wilfrid Kendall's term *perfect simulation* has stuck (see, e.g. Kendall, 2005). They cover only a limited set of circumstances and are most definitely computer-intensive.[22] So these are not techniques for mainstream use (and probably never will be), but they could be used for example

- as a reference against which to compare cheaper simulation schemes, and
- to provide a small number (e.g. one) of samples from which to start an MCMC sampler.

See also Asmussen and Glynn (2007), Casella *et al.* (2001) and Robert and Casella (2004, Chapter 13). One possibly more practical idea that arises from Propp & Wilson's work is the idea of *monotonicity* of MCMC samplers. Suppose there are some extreme states for the distribution of interest, e.g. an image coloured entirely white or black. Then if we start an MCMC scheme at those states, and the realizations become 'similar', there is some hope that realizations starting from any initial state would have become similar by that time. 'Monotonicity' provides a theoretical guarantee of this and it (or similar ideas) underlies most perfect sampling schemes.

Further reading

MCMC can be approached from wide range of viewpoints – from theoretical to practical, as a general technique or purely Bayesian, and at different levels (especially in probability background). Texts which have interesting perspectives include Chen *et al.* (2000), Gamerman and Lopes (2006), Gelman *et al.* (2004), Gilks *et al.* (1996), Liu (2001) and Robert and Casella (2004). Roberts and Tweedie (2005) cover the Markov chain theory. As a topic in simulation, it is covered in Ripley (1987) and Dagpunar (2007),[23] and as a method of integration in Evans and Swartz (2000).

For those unfamiliar with applied Bayesian work, Albert (2007) and (especially) Gelman *et al.* (2004) provide accessible introductions to the computational aspects with non-trivial worked examples.

[22] I understood (from Persi Diaconis) that Propp & Wilson ran a simulation for *six weeks* without any knowledge of how long it would actually take to reach an exact sample.

[23] and at a higher mathematical level, Asmussen and Glynn (2007).

Software

Because MCMC is a meta-algorithm, there are very many specific applications and corresponding software. (I counted 29 such R packages on CRAN just by looking at their `DESCRIPTION` files. See also http://cran.r-project.org/web/views/Bayesian.html.)

Creating general software for MCMC is close to impossible, and all attempts known to me restrict themselves in one or both of two ways. Some confine attention to a family of sampling schemes—e.g. R package `mcmc` works with *"normal random-walk" Metropolis* and perhaps the best-known software, BUGS, works with the Gibbs sampler. Others confine attention to a particular class of statistical models and to a particular way to approach inference on those models. One common restriction is to the Bayesian analysis of hierarchical or graphical models.

BUGS was a program written from 1989 at MRC's BSU in Cambridge in an arcane language that has restricted the platforms it could run on. It uses an S-like language to specify graphical models for which it then creates a Gibbs sampler, plus the ability to simulate from the created sampler. It spawned WinBUGS[24] and OpenBUGS[25] with GUI interfaces.

JAGS[26] is an Open Source program written in C that re-implements the BUGS language.

There are R packages `BRugs`, `R2WinBUGS` and `rjags` to interface with OpenBUGS, WinBUGS/OpenBUGS and JAGS respectively. Gelman *et al.* (2004, Appendix C) discusses driving WinBUGS from R *via* `R2WinBUGS`: it is available on-line at http://www.stat.columbia.edu/~gelman/bugsR/software.pdf.

From the BUGS site

> ### Health warning
>
> *'The programs are reasonably easy to use and come with a wide range of examples. There is, however, a need for caution. A knowledge of Bayesian statistics is assumed, including recognition of the potential importance of prior distributions, and MCMC is inherently less robust than analytic statistical methods. There is no in-built protection against misuse.'*
>
> About JAGS:
>
> *'JAGS uses essentially the same model description language but it has been completely re-written. Independent corroboration of MCMC results is always valuable!'*

Note that all the BUGS-like programs require a proper Bayesian model, so exclude improper priors.

R package `LearnBayes` is a companion to Albert (2007) which includes examples of MCMC both *via* `LearnBayes` and *via* `WinBUGS`. However, the code quality seems low.

[24] which as its name suggests is for Windows only, http://www.mrc-bsu.cam.ac.uk/bugs. People have managed to run it on `ix86` Linux *via* WINE.

[25] http://mathstat.helsinki.fi/openbugs/; this is Open Source but is written in a language "Component Pascal" with a proprietary compiler and so is *de facto* also restricted to Windows—there is a `i386` Linux version that few people have managed to get to work.

[26] http://www-ice.iarc.fr/~martyn/software/jags/.

5 MCMC examples

This section sets the background for the examples of MCMC to be used in the practicals. Many of you will have seen an example in the *Statistical Modelling* module practicals, and more can be found in the on-line practicals for Davison (2003) (http://statwww.epfl.ch/davison/SM/SM_Practicals.1.pdf and R package SMPracticals) and in Gelman *et al.* (2004) and Albert (2007).

Binomial logistic regression

Venables and Ripley (2002, §7.2) explore the following example.

> Consider first a small example. Collett (1991, p. 75) reports an experiment on the toxicity to the tobacco budworm *Heliothis virescens* of doses of the pyrethroid *trans*-cypermethrin to which the moths were beginning to show resistance. Batches of 20 moths of each sex were exposed for three days to the pyrethroid and the number in each batch that were dead or knocked down was recorded. The results were

	Dose					
Sex	1	2	4	8	16	32
Male	1	4	9	13	18	20
Female	0	2	6	10	12	16

> The doses were in μg. We fit a logistic regression model using $\log_2$(dose) since the doses are powers of two.

The interest is in estimating the dose required for a particular probability p of death, especially that for $p = 0.5$ called LD50. A frequentist analysis using `glm` is given in Venables and Ripley (2002), but here we consider a Bayesian analysis.

We start by using R package `MCMCpack`: this works with Bernoulli and not binomial data, so first we disaggregate the results. The default prior for β is an improper uniform prior, but others can be supplied – see `?MCMClogit`.

```
ldose <- rep(0:5, 2)
numdead <- c(1, 4, 9, 13, 18, 20, 0, 2, 6, 10, 12, 16)
sex <- factor(rep(c("M", "F"), c(6, 6)))
SF <- cbind(numdead, numalive = 20 - numdead)
resp <- rep(rep(c(1,0), 12), times = t(SF))
budworm <- data.frame(resp, ldose = rep(ldose, each = 20),
                      sex=rep(sex, each = 20))
glm(resp ~ sex*ldose, family = binomial, data = budworm)

library(MCMCpack)  # loads package 'coda'
fit <- MCMClogit(resp ~ sex*ldose, data = budworm)
summary(fit)
plot(fit)
acfplot(fit) # suggests thinning
fit <- MCMClogit(resp ~ sex*ldose, data = budworm, thin = 20)
```

There is an issue with LD50, pointed out by Gelman *et al.* (2004, p. 93): we are really only interested in positive slopes. In this example the chance of a negative fitted slope is negligible, but otherwise LD50 would be a complex non-linear function of the parameters, something simulation-based inference takes in its stride.

Our second approach uses BUGS.[27] We need to specify the BUGS model, which we will do in a file[28] `budworm.bug`

```
model {
   for(i in 1:6) {
      numdead[i] ~ dbin(p[i], 20)
      logit(p[i]) <- alphaM + betaM * ldose[i]
   }
   for(i in 7:12) {
      numdead[i] ~ dbin(p[i], 20)
      logit(p[i]) <- alphaF + betaF * ldose[i]
   }
   betaM ~ dnorm(0.0, 0.001)
   alphaM ~ dnorm(0.0, 0.001)
   betaF ~ dnorm(0.0, 0.001)
   alphaF ~ dnorm(0.0, 0.001)
}
```

This is simple rather than general, and specifies rather vague independent priors for the parameters. The syntax is deceptively similar to S, but note that `dnorm` has arguments mean and precision (reciprocal variance).

To run the MCMC we use

```
library(R2WinBUGS
budworm.sim <- openbugs(list("numdead", "ldose"),
                        list(alphaM = 0, betaM = 0, alphaF = 0, betaF = 0),
                        c("alphaM", "alphaF", "betaM", "betaF"),
                        model.file = "budworm.bug",
                        n.chain = 1, n.iter = 10000, DIC = FALSE)
budworm.sim
plot(budworm.sim)
```

with printout

```
Inference for Bugs model at "budworm.bug", fit using OpenBUGS,
 1 chains, each with 10000 iterations (first 5000 discarded), n.thin = 5
 n.sims = 1000 iterations saved
         mean  sd 2.5%  25%  50%  75% 97.5%
alphaM   -2.9 0.6 -4.2 -3.3 -2.9 -2.5  -1.9
alphaF   -3.1 0.6 -4.2 -3.4 -3.0 -2.7  -2.0
betaM     1.3 0.2  0.9  1.2  1.3  1.4   1.8
betaF     0.9 0.2  0.6  0.8  0.9  1.0   1.3
```

We should explore other starting points, and will do so in the practical.

Looking at the posterior simulations shows a potential problem with naïve use of the Gibbs sampler—the intercept and slope are quite correlated. Only extreme correlations will give

[27] In these examples we call OpenBUGS (*via* package `BRugs`) rather than the default WinBUGS solely to avoid having to install (and license) another program.

[28] model files should have extension `bug` or `txt`.

problems in a classical analysis of a GLM, but here quite modest correlations can slow down convergence of the automatically constructed Gibbs sampler.

Poisson change-point models

Consider the much-used data set of annual counts of British coal mining 'disasters' from 1851 to 1962.[29] Looking at the data suggests that the rate of accidents decreased sometime around 1900, so a suitable model is that the counts are independent Poisson with mean λ_1 before time τ and mean λ_2 from time τ onwards, where we expect $\lambda_2 < \lambda_1$. This is the simplest possible case, and we could consider more than one changepoint.

For a Bayesian analysis we need a prior distribution on the three parameters. If we take them as independent and of conjugate form, the posterior can be found analytically (Gamerman and Lopes, 2006, pp. 143ff), but a realistic prior will have a dependent distribution of (λ_1, λ_2). That is easy to do in the MCMC framework—for a more complex application to radiocarbon dating see Gilks *et al.* (1996, Chapter 25).

We will consider computing posterior distributions *via* MCMC in two ways. R package `MCMCpack` has a function `MCMCpoissonChangepoint` with an MCMC scheme coded in C++, implementing the method of Chib (1998). This has independent gamma priors for the rates and beta priors for the transition point(s). The R code is simple:

```
## D is an integer vector of 113 counts.
library(MCMCpack)
fit <- MCMCpoissonChangepoint(D, m = 1, c0 = 1, d0 = 1,
                              burnin = 10000, mcmc = 10000)
plot(fit)
summary(fit)
plotState(fit)
plotChangepoint(fit)
```

The arguments say that we are looking for `m = 1` changepoints, and specify a gamma$(1, 1)$ prior for the mean counts λ_i. In that approach, someone else has done all the work in designing and coding a suitable MCMC scheme—this is fast but not general. `MCMCpack` produces samples ready for analysis by R package `coda`.

Our second approach follows Albert (2007, §11.4) and uses vague priors. Rather than use specific code, we use BUGS and hence a Gibbs sampler somewhat tailored by the program to the problem. The first step is to tell BUGS what the model is using model file

```
model {
  for(year in 1:N) {
       D[year] ~ dpois(mu[year])
       log(mu[year]) <- b[1] + step(year - changeyear) * b[2]
  }
  for (j in 1:2) {b[j] ~ dnorm(0.0, 1.0E-6)}
  changeyear ~ dunif(1,N)
}
```

This version uses a slightly different parametrization, with the first element of `b` as $\log \lambda_1$ and the second as $\log \lambda_2 - \log \lambda_1$. The data are the counts in `D` and the number of years, `N`.

[29] These were derived from Jarrett (1979) and refer to explosions.

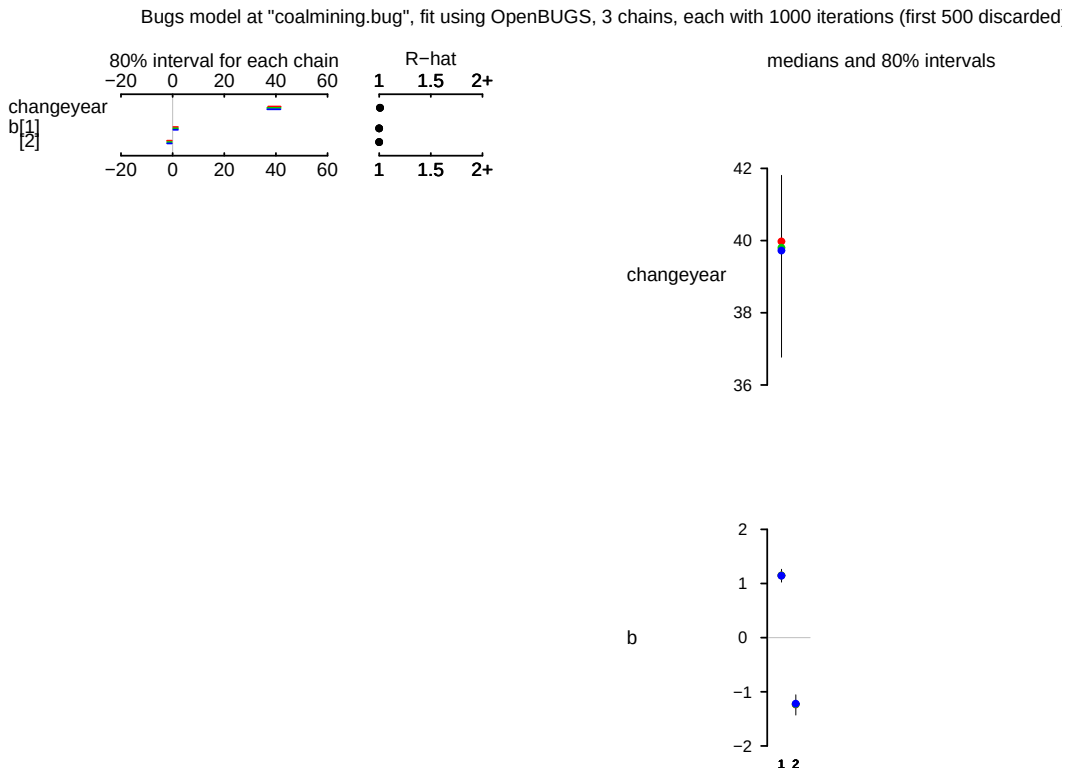

Figure 7: A plot for bugs output for the coal mining disasters problem.

We can then use function openbugs to ask OpenBUGS (*via* package BRugs) to simulate from the posterior distribution of this model, by e.g.

```
library(R2WinBUGS)
inits <- list(list(b=c(0,0), changeyear=50),
              list(b=rnorm(2), changeyear=30),
              list(b=rnorm(2), changeyear=70))
coalmining.sim <-
    openbugs(list("N", "D"), inits, c("changeyear","b"),
             "coalmining.bug",
             n.chains = 3, n.iter = 1000, DIC = FALSE)
```

As MCMC is an iterative scheme we have to supply initial values of the parameters: it is possible to supply (as a list of lists as here) separate starting values for each run, or a function that will give a random list result.

The result object can be printed and plotted. The printout looks like

```
> coalmining.sim
Inference for Bugs model at "coalmining.bug", fit using OpenBUGS,
 3 chains, each with 1000 iterations (first 500 discarded)
 n.sims = 1500 iterations saved
            mean  sd  2.5%   25%   50%   75% 97.5% Rhat n.eff
changeyear  39.7 2.1  36.1  38.0  39.9  40.8  44.8    1   500
b[1]         1.1 0.1   1.0   1.1   1.1   1.2   1.3    1  1500
b[2]        -1.3 0.2  -1.5  -1.4  -1.3  -1.2  -1.0    1  1100

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).
```

Hierarchical models

MCMC is widely used in hierarchical Bayesian models. Here is a very simple example considered by Gelman *et al.* (2004, §§5.6, 6.8, 17.4).

The US Educational Testing Service investigated the effect of coaching for SAT-V tests in 8 schools. The tests are supposed to be insensitive to coaching, but these 8 schools claimed to have an effective coaching program. The SAT-V scores have a range of 200–800. This is a meta-analysis: for each school we have estimates of the mean effect of coaching and of the standard deviation of the effect estimate *via* a within-school analysis of covariance.

Gelman *et al.* (2004, Appendix C) provide R code for several analyses based on Gibbs sampling. The mode can be defined by the BUGS model file `schools.bug`:

```
model {
    for (j in 1:J) {
        y[j] ~ dnorm(theta[j], tau.y[j])
        theta[j] ~ dnorm(mu.theta, tau.theta)
        tau.y[j] <- pow(sigma.y[j], -2)
    }
    mu.theta ~ dnorm(0, 1.0E-6)
    tau.theta <- pow(sigma.theta, -2)
    sigma.theta ~ dunif(0, 1000)
}
```

So there are `J = 8` schools, and each has a mean θ_i and precision τ_i, with the per-schools means being modelled as draws from a normal population. This the parameters are the 8 θ_i and the two hyperparameters for the population distribution of means. An alternative model we can consider is a t_ν distribution for the population of means, with a known or unknown ν.

Simulation-based inference makes it easy to draw inferences about non-linear functions of the parameters, for example of the largest effect $\max_i \theta_i$.

We can fit this model in OpenBUGS using different random starting values for each run by

```
library(R2WinBUGS)
data <- list ("J", "y", "sigma.y")
inits <- function()
    list (theta = rnorm(J,0,100), mu.theta = rnorm(1,0,100),
          sigma.theta = runif(1,0,100))
parameters <- c("theta", "mu.theta", "sigma.theta")
schools.sim <- openbugs(data, inits, parameters, "schools.bug",
                        n.chains = 3, n.iter = 10000, DIC = FALSE)
```

and in JAGS by

```
library(rjags)
foo <- jags.model("schools.bug", inits = inits, nchain = 3)
z <- coda.samples(foo, n.iter = 10000, thin = 10)
```

`rjags` produces samples ready for analysis by `coda`.

Survival

Parametric survival models are not easily fitted in a Bayesian setting, and we consider fitting a Weibull accelerated life model to a subset of the Australian AIDS data of Venables and Ripley (2002, §13.5). To reduce computation time we consider only the 1116 patients from NSW and ACT (an enclave within NSW). To take account of the introduction of Zidovudine (AZT), time was run at half speed from July 1987.

```
library(MASS); library(survival)
Aidsp <- make.aidsp() # MASS ch13 script
fit <- survreg(Surv(survtime + 0.9, status) ~ T.categ + age,
               data = Aidsp, subset = (state=="NSW"))
summary(fit)
```

This model has 9 coefficients in the linear prediction, and one (σ) for the shape of the Weibull, which is very close to exponential.

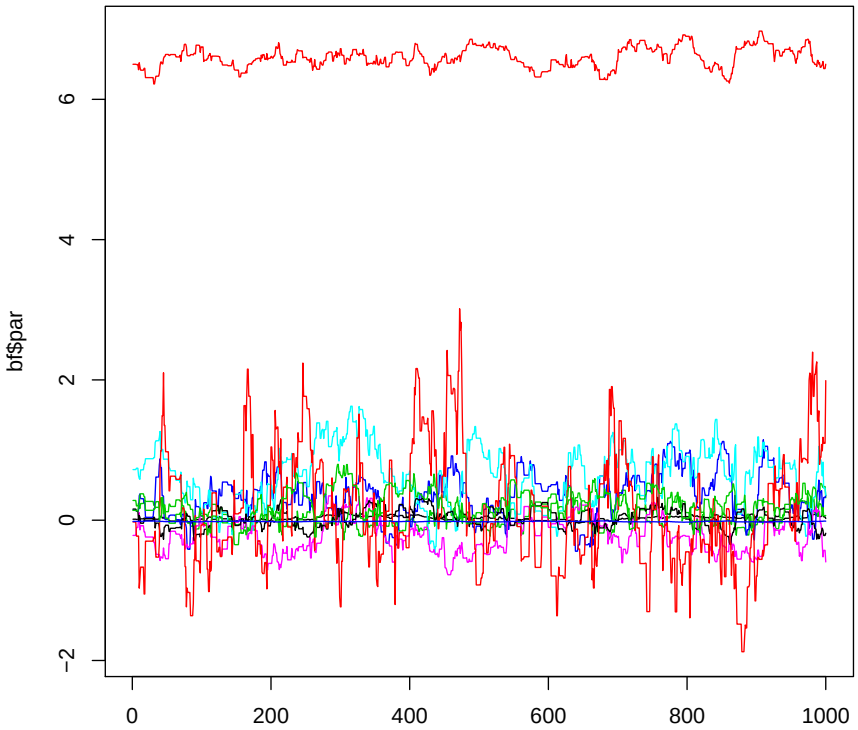

Figure 8: Diagnostic traces from MCMC estimation of a Weibull survival model for the NSW/ACT AIDS data.

We can make use of improved (and corrected) versions of the code in Albert (2007):

```
library(LearnBayes)
weibullregpost <- function(theta, data)
{
    sigma <- exp(theta[1]); beta <- theta[-1]
```

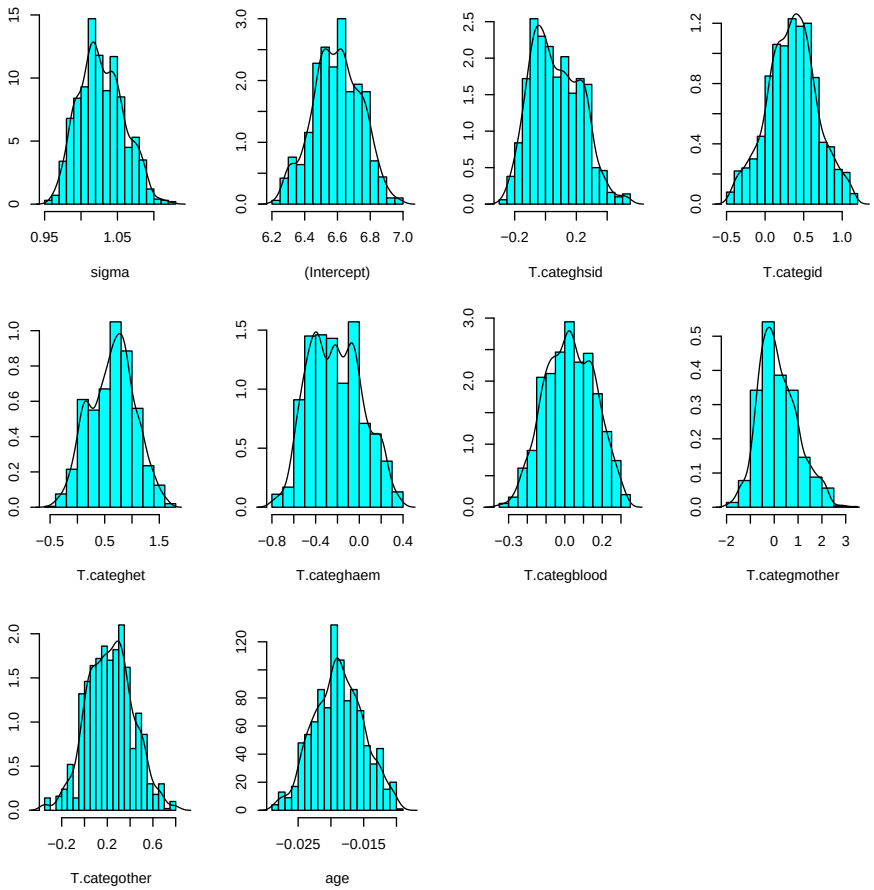

Figure 9: Histograms with overlaid kernel density estimates of the univariate posteriors from MCMC estimation of a Weibull survival model for the NSW/ACT AIDS data.

```
    lp <- drop(data[, -(1:2), drop=FALSE] %*% beta)
    zi <- (log(data[,1]) - lp)/sigma
    fi <- 1/sigma * exp(zi - exp(zi))
    Si <- exp(-exp(zi))
    sum(log(ifelse(data[,2], fi, Si)))
}
start <- t(c(log(fit$scale), coef(fit)))
d <- cbind(model.frame(fit)[[1]], model.matrix(fit))
fit0 <- laplace(weibullregpost, start, d)
```

This computed the posterior density for the parameters $(\log \sigma, \beta)$ for a vague (improper) prior, and then finds the posterior mode (which is essentially the MLE).

Simulation is then done by a random-walk Metropolis algorithm, with a multivariate normal step with variance proportional to the `fit0$var`, that is to the estimated covariance of the MLEs. Tuning constant `scale` is the proportionality factor (on standard-deviation scale), and needs to be chosen by trial-and-error. Starting with a smallish value we have

```
proposal <- list(var=fit0$var, scale=0.5)
bf <- rwmetrop(weibullregpost, proposal, fit0$mode, 1000, d)
bf$accept
matplot(bf$par, type="l", lty=1)
op <- par(mfrow=c(3,4), mar=c(5,4,1,1))
nm <- c("sigma", names(coef(fit)))
```

```
for(i in 1:10) {
    x <- bf$par[, i]
    if(i == 1) x <- exp(x)
    truehist(x, xlab=nm[i], main="")
    lines(density(x), xpd=NA)
}
par(op)
```

This results in about 45% acceptance in the Metropolis step and apparently reasonable convergence well within 1000 steps.

Linear models

Straightforward linear models of the from

$$Y = X\beta + \epsilon, \qquad \epsilon \sim N(0, \sigma^2 I)$$

have $p+1$ parameters $\theta = (\beta, \sigma^2)$. A Bayesian analysis needs a prior for θ, and for convenient priors the posterior can be found explicitly. However, if we allow non-IID errors so $\epsilon \sim N(0, \kappa\Sigma(\psi))$ then simulation-based methods become much more convenient, and perhaps essential.

One common way for such structured variance matrices to arise is what is called in the classical literature *mixed effects* models. Suppose that

$$Y = X\beta + Z\gamma + \epsilon, \qquad \epsilon \sim N(0, \sigma^2 I)$$

where γ is regarded as random vector. In a Bayesian setting β is already regarded as a random vector so this is no real change, but impact comes from thinking of this hierarchically. In the simplest case, suppose we have two levels of units, say observations on classes in schools or repeated measurements on individuals. Rather than the exchangeability that the IID assumption entails, we now have a multi-level invariance amongst groups of observations. With two levels of units, the Bayesian model has three groups of random variables

- data points Y_{ij}, observed at the lowest level,
- random effects η_i on level-one units, unobserved, and
- parameters in the distribution.

Note though that it is just a linear model for the observed data with a parametrized variance matrix of correlated errors.

This is fertile ground for use of a Gibbs sampling scheme to simulate from the posterior distribution given the observed data, so many schemes have been proposed. Here are the basic ideas.

(a) a grouped Gibbs sampler, in which all the variables in one of the three groups are updated at once. Generally the conditional distributions of the groups given the rest are simple to simulate from, although the variance parameters may need a Metropolis step. How well this works depends strongly on how well the hierarchical model mimics the real dependence structure.

(b) an 'all-at-once' Gibbs sampler. This flattens the hierarchical model to two levels, the linear coefficients and the distributional parameters, and alternates between them. Effectively it fits a weighted regression for known variance parameters, then simulates variance parameters conditional on the residuals from the weighted regression. It is in general easy to implement and quick to converge, but the flattened model can be much larger.

(c) a single-variable Gibbs sampler, updating one variable at a time. Again this is usually simple to implement, and simulating the individual regression parameters is fast. The problem is that the latter can be highly correlated and so the Gibbs sampler moves slowly: this can often be overcome by a linear transformation of the regression parameters, one which can be approximated by a pilot run.

(d) parameter expansion. All of these schemes can be slow to converge when estimated hierarchical variance parameters are near zero, since this will ensure that the corresponding random effects are estimated as rather similar and then at the next step the variance parameters is estimated as small. We can circumvent this by adding further parameters, e.g. a multiplicative effect on all the random effects in a group. For the SAT-V data the model becomes

$$\begin{aligned} Y_j &\sim N(\mu + \alpha\gamma_j, \sigma_j^2) \\ \gamma_j &\sim N(0, \sigma_\gamma^2) \\ p(\mu, \alpha, \sigma_\gamma) &\propto 1 \end{aligned}$$

(or some other prior on the parameters). The analysis discussed above was for the model with $\alpha \equiv 1$. Note that σ_γ^2 is no longer the random effects variance, rather $\alpha^2\sigma_\gamma^2$. (Some care is needed as aspects of the posterior here are improper—it would be better to use a proper prior for α.)

For more details on simulation-based Bayesian approaches to regression-like problems see Gelman and Hill (2007) and Gelman *et al.* (2004). Gamerman and Lopes (2006, §6.5) have complementary material.

6 Large datasets

What is 'large' about large datasets?

- Many cases. This is perhaps the most common, for example
  - Oxford's library catalogue has five million items, Harvard's nine million and the British Library's thirteen million.
  - Insurance companies have records for all their customers, and all those they have offered quotes to, and they tend to share information with other companies. So there is a database with about 70 million records on US drivers.
  - Medical registries (in countries which have them) will have pretty much complete coverage of particular diseases (e.g. haemophilia) and typically these are of up to tens of thousands of subjects.
  - Inventories and sales records are often for many thousands of items.
  - Banks have records for every customer (several million in large countries) which they use to target their marketing (and especially their mailshots).
- Many pieces of information per case. This tends to be rarer, but with genome-wide screening it is common to have thousands (and sometimes tens of thousands) of pieces of information for each of a modest number (perhaps 100) subjects.

  We have been working on group studies in fMRI.[30] These have at most tens of subjects, and maybe thousands of brain images each of a hundred thousand voxels for each subject.

  Remote sensing provides another example—complex images on few occasions.
- Many cases and many pieces of information on each. This is currently unusual, and the only one I have encountered is CRM.[31] You probably all have a 'loyalty' card—that is a means to bribe you to collect information about you. So the consortium owning the card know the buying patterns of every customer but also associations between items. You have probably used sites such as Amazon and been offered suggestions, maybe

  > People who purchased this book also purchased . . . '

  Perhaps national censuses come into this category: although full censuses usually have a modest number of questions, it is common to ask more of a, say, 10% sample and perhaps even follow those up yearly.

A typical dataset can be thought of as a rectangular table: the most common case is a 'long' table with the second and third bullet points corresponding to 'wide' and 'both long and wide' tables.

The largest dataset I have been involved with is one being planned by my colleagues in genetics who are envisaging 20TB[32] of raw data.

[30] functional Magnetic Resonance Imaging.

[31] 'Customer Relationship Management'.

[32] 20 terabytes, about 20,000 GB or 2×10^{13} bytes or 80 typical discs.

And what is meant by 'large'? Unwin *et al.* (2006, Chapter 1) traces some history and quotes the following table (from Huber in 1992 and Wegman in 1995)

Size	Description	Bytes
Tiny	Can be written on a blackboard	10^2
Small	Fits on a few pages	10^4
Medium	Fills a floppy disk	10^6
Large	Fills a tape	10^8
Huge	Needs many tapes	10^{10}
Monster		10^{12}

To put this in a bit more recent perspective, a CD-ROM stores about 5×10^8 bytes, a double-layer DVD-ROM about 10^{10} bytes. In 1995 I was offered some remote-sensed data on CD-ROMs—around 2000 of them, and so 'monster'.

Hardware considerations

Currently most computers are '32-bit', but we are in the time of transition to '64-bit'. These transitions happen every few years—Windows went from 16-bit to 32-bit with Windows 95.[33] 64-bit versions of Windows have been available for some time but are rarely seen (and 64-bit software remains rare). Other OSes have been transitioning to 64-bit for a long time: I had a 64-bit Solaris system in 1997, and have used a 64-bit Linux desktop for the last 2.5 years. Mac OS X is still struggling with 64-bit support which in theory was completed in 'Leopard' last autumn.

What is it that is '32-bit'? Several things. Almost all current computers work with *bytes*, units of 8-bits. So '32-bits' refers to the way that bytes are addressed—by using a 32-bit pointer we can address 2^{32} separate bytes, that is about 4 billion. What are these bytes? At least

- Virtual memory for a user process (the address space).
- Virtual memory for the operating system.
- Bytes in a file.

However, because programmers used signed integers which can hold numbers $-2^{31} \ldots 2^{31}-1$, most effective limits for 32-bit systems were 2GB.

2GB seemed large until recently, but with current disc sizes of around 250GB, 2GB files are no longer rare. Most 32-bit OSes have a means to support larger files, and R has made use of those facilities for some years.

These days 2GB of RAM is just above entry-level,[34] so it is reasonable to expect to use 2GB of memory in a single process. In fact 32-bit OSes limit per-process user address spaces to that or a bit more (maybe up to 3.5GB), and it is that limit that is pushing the move to 64-bit OSes. (Most CPUs currently on sale are 32/64-bit, even those in the most modest of laptops.)

[33] in 1995, with some 32-bit support in previous versions.

[34] thanks to the memory usage of Windows Vista.

DBMSs

Large amounts of data are not usually stored in simple files but in *databases*. Generally a *database* is thought of as the actual numeric or character data plus metadata.

Although some people use Excel spreadsheets as databases, professional-level uses of databases are *via* DataBase Management Systems (DBMS), which are designed to efficiently retrieve (and in some cases update) parts of the data. DBMSs lie behind a great deal of what we do: when a call centre says 'I will just bring up your record',[35] what they are actually doing is using a DBMS to extract records from several tables in one or more databases. Also, when you ask for a page on many websites, it is retrieved from various tables in a DBMS.

DBMSs vary greatly in scale, from personal (such as Microsoft Access, MySQL) through department-level servers (Microsoft SQL Server, MySQL, PostgreSQL) to enterprise-wide (Oracle, DB2 and upscaled department-level systems).

Most of these systems work with SQL (Structured Query Language)[36] to access parts of the data. There is a series of ISO standards for SQL, but unfortunately most of the DBMS vendors have their own dialects.

It has been a long time coming, but it will become increasingly common for statisticians to be working with data stored in a DBMS. So some knowledge of SQL will become increasingly valuable.

Strategies for handling large datasets

The increase in volume of data available has been driven by automated collection, but computer power is growing faster than human activity. So in many fields we have already reached or are close to having all the data that will be relevant. For example, the motor insurance databases are as large as they are ever going to be, the medical registries have all current cases of rare diseases (and new cases are by definition rare), and so on.

So already some of the strategies needed in the past are no longer required.

A decade ago, Bill Venables and I heard a talk at a conference about some new capabilities in a software package for 'large data' regressions, illustrated by a simulated regression problem with 10,000 cases. Over dinner we discussed if we had even seen such a problem in real life (no) and if we ever would (we thought not). The issue is that large regression problems are almost never homogeneous—they lump together data from different groups (e.g. different centres in a clinical trial). So the first strategy is

> *Divide the dataset into naturally occurring groups, analyse each group separately and do a meta-analysis.*

In a random regression problem each case adds the same amount of (Fisher) information, so collecting more cases reduces the variance of the estimator of the parameters at a known rate. As the regression model is false,[37] there will be a fixed amount of bias in its predictions

[35] usually followed by 'the system is rather slow today'.

[36] http://en.wikipedia.org/wiki/SQL.

[37] apart from in a perfectly constructed simulation

irrespective of the sample size and for large enough sample sizes the bias will dominate the variance. So

> *With homogeneous datasets we can often achieve close to maximally accurate*[38] *results using a small sample of the dataset.*

It is hard to give detailed guidance as large homogeneous datasets are so rare, but it seems exceptional to need to do a regression on more than 1000 cases. Such problems may exist, but all those we have been offered as counter-examples have crumbled on close examination.

Do be careful in sampling heterogeneous datasets. We once had a motor insurance database of 700,000 cases to which we were fitting binomial and gamma generalized linear models. Because the Fisher information per case is not uniform in such models, some observations are much more important than others, and we found that using a 10% random sample was giving much less good predictions than the whole dataset, and we needed to use a stratified sample. But that was in 2001 and the computers used had about 256MB of RAM, so sampling would no longer be needed. In fact a very important strategy is

> *Get a bigger computer, or even several of them.*

As a student it may be hard to appreciate that the time of analysts (and particularly trained statisticians) is very valuable, and computers are cheap. If someone with a large dataset or a computer-intensive method does not have use of a computer with several GB of RAM 24 hours/day, then their time is not being valued correctly.

If these strategies are not sufficient, we need to consider how to do the actual computations. In some statistical problems, only some summaries of the data are required to do the computation. It is tempting to think that this is the case in statistical models with low-dimensional sufficient statistics, but that is not in general the case as 'data' is not necessarily regarded as random in the model. Consider first a regression problem with response vector Y and an $n \times p$ data matrix X. The sufficient statistics are X^TY, but that is not enough information to find the parameter estimates. We can however find the parameter estimates from X^TX and X^TY, involving dimensions of size p but not n. However, to compute residuals, we need to go back to the data matrix X.

Now consider fitting a GLM. As you know, the commonest method is Iteratively (Re)Weighted Least Squares, which involves solving problems equivalent to

$$(X^TWX)b = X^TWY$$

where W is a diagonal matrix which varies for each iteration. We can use the summarization approach here, provided we are prepared to make multiple passes over the data. Note to the *cognoscenti*: we do not really solve these normal equations as stated, as that would be needlessly inaccurate. Rather one can make use of a row-wise QR decomposition, most often using Givens rotations.

This leads to another general strategy:

> *Consider algorithms needing multiple passes over the data.*

[38] http://en.wikipedia.org/wiki/Accuracy.

These are almost inevitably slower, but can need fewer resources at any one time and so may be feasible. This is how programs from long ago like SAS[39] work, and there is an R package `biglm` which takes this approach for linear models and GLMs.

Another general strategy is to

> *Make use of multiple CPUs by parallelizing the computations.*

Increasingly computers are coming with multiple CPU cores—even basic machines have two CPUs on a single chip and up-market servers have quad-cores and two or more such chips. This trend is bound to continue for quite a while, and computer clusters containing 256 or more[40] CPUs are now fairly common. Programming parallel computations is not easy, and the available hardware is beginning to run far ahead of available software. There are at least two fundamental problems

- Location of data. Many statistical algorithms need repeated use of the same pieces of data and of data created from earlier computations (as we have seen e.g. for GLMs). Moving that data around between multiple computers is a bottleneck. Even where the CPUs are in the same computer or even on the same chip, moving data around can be an issue since modern CPUs get their speed by maintaining two or three levels of local cache memory.
- Synchronization. Somehow calculations need to be arranged so that CPUs are not waiting for other CPUs to finish.

The one relevant area where a lot of work has been done on parallelization is numerical linear algebra—experience is that gains with just two CPUs are often small, but with 4 or 8 it is possible to get a substantial speedup. *However*, it is also possible for the data-passing issues to dominate so that using multiple CPUs is several times *slower* than using just one. It is not easy to anticipate when this might happen, as the authors of the mixed-effect models packages for R such as `lme4` have found.

The most trivial form of parallel computation, using several CPUs for separate simulations, is particularly well suited to the methods of this module. This is sometimes known as 'embarrassingly parallel' programming.

Note that it may become increasingly important to use multiple CPUs, as the conventional folk-wisdom of Moore's Law ('computer power doubles every 18 months') is showing signs of slowing down—Asanovic and ten others (2006, p. 6) suggest a doubling of single-CPU power every five years.

Visualization

Visualization of large datasets is an important topic – see Unwin *et al.* (2006) for the viewpoints of the Augsburg school. One dataset explored in Chapter 11 there is discussed at `http://www.public.iastate.edu/~hofmann/infovis/` (with videos).

Sampling can help here, unless the aim is to spot outliers and other exceptional cases.

[39] still using a design from the 1960s and 70s.
[40] up to 100,000s

References

Aarts, E. and Korst, J. (1989) *Simulated Annealing and Boltzmann Machines*. John Wiley and Sons.

Albert, J. (2007) *Bayesian Computation with R*. New York: Springer.

Asanovic, K. and ten others (2006) The landscape of parallel computing research: A view from Berkeley. Technical Report UCB/EECS-2006-183, EECS Department, University of California, Berkeley.

Asmussen, S. and Glynn, P. W. (2007) *Stochastic Simulation. Algorithms and Analysis*. New York: Springer.

Bowman, A. and Azzalini, A. (1997) *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford: Oxford University Press.

Box, G. E. P., Hunter, W. G. and Hunter, J. S. (1978) *Statistics for Experimenters*. New York: John Wiley and Sons.

Buckland, S. T. (1984) Monte Carlo confidence intervals. *Biometrics* **40**, 811–817.

Casella, G., Lavine, M. and Robert, C. (2001) Explaining the perfect sampler. *American Statstician* **55**, 299–305.

Chen, M.-H., Shao, Q.-M. and Ibrahim, J. G. (2000) *Monte Carlo Methods in Bayesian Computation*. New York: Springer.

Chernick, M. R. (2008) *Bootstrap Methods. A Practioner's Guide*. Second Edition. New York: Wiley.

Chib, S. (1998) Estimation and comparison of multiple change-point models. *J, Econometrics* **86**, 221–241.

Chib, S. and Greenberg, E. (1995) Understanding the Metropolis–Hastings algorithms. *American Statistician* **49**, 327–335.

Collett, D. (1991) *Modelling Binary Data*. London: Chapman & Hall.

Cook, D. and Swayne, D. F. (2007) *Interactive and Dynamic Graphics for Data Analysis*. New York: Springer.

Dagpunar, J. (2007) *Simulation and Monte Carlo. With Applications in Finance and MCMC*. Chichester: Wiley.

Davison, A. C. (2003) *Statistical Models*. Cambridge: Cambridge University Press.

Davison, A. C. and Hinkley, D. V. (1997) *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.

Davison, A. C., Hinkley, D. V. and Young, G. A. (2003) Recent developments in bootstrap methodology. *Statistical Science* **18**, 141–157.

Efron, B. (1982) *The Jackknife, the Bootstrap, and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B. (1983) Estimating the error rate of a prediction rule. improvements on cross-validation. *Journal of the American Statistical Association* **78**, 316–331.

Efron, B. and Tibshirani, R. (1993) *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Efron, B. and Tibshirani, R. (1997) Improvements on cross-validation: The .632 bootstrap method. *Journal of the American Statistical Association* **92**, 548–560.

Evans, M. and Swartz, T. (2000) *Approximating Integrals via Monte Carlo and Deterministic Methods*. Oxford: Oxford University Press.

Gamerman, D. and Lopes, H. F. (2006) *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Second Edition. London: Chapman & Hall/CRC Press.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004) *Bayesian Data Analysis*. Second Edition. Chapman & Hall/CRC Press.

Gelman, A. and Hill, J. (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science* **7**, 457–511.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Geyer, C. (1999) Likelihood inference for spatial point processes. In *Stochastic Geometry. Likelihood and Computation*, eds O. E. Barndorff-Nielsen, W. S. Kendall and M. N. M. van Lieshout, Chapter 3, pp. 79–140. London: Chapman & Hall/CRC.

Geyer, C. J. and Thompson, E. A. (1992) Constrained Markov chain maximum likelihood for dependent data (with discussion). *JRSSB* **54**, 657–699.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996) *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.

Green, P. J. (1995) Reversible junp Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Grenander, U. and Miller, M. (1994) Representations of knowledge in complex systems (with discussion). *Journal of the Royal Statistical Society series B* **56**, 549–603.

Hall, P. (1992) *The Bootstrap and Edgeworth Expansion*. Springer-Verlag.

Hall, P. (2003) A short pre-history of the bootstrap. *Statistical Science* **18**, 158–167.

Harrell, Jr., F. E. (2001) *Regression Modeling Strategies, with Applications to Linear Models, Logistic Regression and Survival Analysis*. New York: Springer-Verlag.

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Jarrett, R. G. (1979) A note on the interval between coal-mining disasters. *Biometrika* **66**, 191–3.

Jerrum, M. (1995) A very simple algorithm for estimating the number of k-colorings of a low-degree graph. *Random Structures and Algorithms* **7**, 157–165.

Jöckel, K.-H. (1986) Finite sample properties and asymptotic efficiency of Monte Carlo tests. *Annals of Statistics* **14**, 336–347.

Kendall, W. S. (2005) Notes on perfect simulation. In *Markov Chain Monte Carlo. Innovations and Applications*, eds W. S. Kendall, F. Liang and J.-S. Wang, pp. 93–146. Singapore: World Scientific.

Kirkpatrick, S., Gelatt, Jr, C. D. and Vecchi, M. P. (1983) Optimization by simulated annealing. *Science* **220**, 671–680.

Kushner, H. J. and Lin, G. G. (2003) *Stochastic Approximation and Recursive Algorithms and Applications*. Second Edition. New York: Springer-Verlag.

Lauritzen, S. and Spiegelhalter, D. J. (1988) Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society series B* **50**, 157–224.

Liu, J. S. (2001) *Monte Carlo Strategies in Scientific Computing*. New York: Springer.

Meng, X. L. and Wong, W. H. (1996) Simulating ratios of normalizing constants vis a simple identity: a theoretical exploration. *Statistica Sinica* **6**, 831–860.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E. (1953) Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1091.

Morgenthaler, S. and Tukey, J. W. eds (1991) *Configural Polysampling. A Route to Practical Robustness*. John Wiley and Sons.

Pincus, M. (1970) A Monte-Carlo method for the approximate solution of certain types of constrained optimization problems. *Operations Research* **18**, 1225–1228.

Politis, D. N., Romano, J. P. and Wolf, M. (1999) *Subsampling*. New York: Springer-Verlag.

Propp, J. and Wilson, D. (1996) Exact sampling with coupled markov chains and applications to statistical mechanics. *Random Structures and Algorithms* **9**, 223–252.

Rao, J. N. K. and Wu, C. F. J. (1988) Resampling inference with complex survey data. *Journal of the American Statistical Association* **83**, 231–241.

Ripley, B. D. (1979) Algorithm AS137. Simulating spatial patterns: dependent samples from a multivariate density. *Applied Statistics* **28**, 109–112.

Ripley, B. D. (1987) *Stochastic Simulation*. New York: John Wiley and Sons.

Ripley, B. D. (1988) *Statistical Inference for Spatial Processes*. Cambridge: Cambridge University Press.

Ripley, B. D. (1996) *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

Ripley, B. D. (2005) How computing has changed statistics. In *Celebrating Statistics: Papers in Honour of Sir David Cox on His 80th Birthday*, eds A. C. Davison, Y. Dodge and N. Wermuth, pp. 197–211. Oxford University Press.

Ripley, B. D. and Kirkland, M. D. (1990) Iterative simulation methods. *Journal of Computational and Applied Mathematics* **31**, 165–172.

Robert, C. P. and Casella, G. (2004) *Monte Carlo Statistical Methods*. Second Edition. New York: Springer.

Roberts, G. O. and Rosenthal, J. S. (1998) Markov-chain Monte Carlo: Some practical implications of theoretical results. *Canadian Journal of Statistics* **26**, 5–31.

Roberts, G. O. and Tweedie, R. L. (2005) *Understanding MCMC*. New York: Springer.

Roeder, K. (1990) Density estimation with confidence sets exemplified by superclusters and voids in galaxies. *Journal of the American Statistical Association* **85**, 617–624.

Shao, J. and Tu, D. (1995) *The Jackknife and the Bootstrap*. New York: Springer.

Simon, J. L. (1997) *Resampling: The New Statistics*. Second Edition. Resampling Stats.

Snijders, T. A. B. (2001) The statistical evaluation of social network dynamics. In *Sociological Methodology – 2001*, eds M. Sobel and M. Becker, pp. 361–395. Boston and London: Basil Blackwell.

Snijders, T. A. B. (2006) Statistical methods for network dynamics. In *Proceedings of the XLIII Scientifc Meeting, Italian Statistical Society*, pp. 281–296. Padova: CLEUP.

Staudte, R. G. and Sheather, S. J. (1990) *Robust Estimation and Testing*. New York: John Wiley and Sons.

Tanner, M. A. (1996) *Tools for Statistical Inference*. Third Edition. Springer-Verlag.

Tanner, M. A. and Wong, W. H. (1987) The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82**, 528–540.

Unwin, A., Theus, M. and Hofmann, H. (2006) *Graphics of Large Datasets. Visualizing a Million*. Springer.

Venables, W. N. and Ripley, B. D. (2002) *Modern Applied Statistics with S*. Fourth Edition. New York: Springer-Verlag.

Yu, B. and Mykland, P. (1998) Loooking at Markov samplers through cusum path plots: A simple diagnostic idea. *Statistics and Computing* **8**, 275–286.

Tuesday Practical

Ex 1 We repeat the calculations for figure 1. We will use 999 random permutations (in fact the figure used all $2^{10} = 1024$ permutations).

```
library(MASS)
shoes
t.test(shoes$A, shoes$B, paired=TRUE)
d <- shoes$A - shoes$B
t.test(d)
t.test(abs(d)) # the most extreme difference
R <- 999
tperm <- numeric(R)
for(i in 1:R) {
    a <- 2*rbinom(20, 1, 0.5) - 1
    tperm[i] <- t.test(a*d)$statistic
}

op <- par(mfrow = c(1, 2))
truehist(tperm, xlab = "diff", xlim=c(-5,5))
x <- seq(-5, 5, 0.1)
lines(x, dt(x,9))
plot(ecdf(tperm), xlim=c(-5,5), do.points=FALSE)
lines(x, pt(x,9), lty=3)
par(op)
```

How can you use these data to perform a Monte Carlo test?

Ex 2 Dataset `cd4` in package `boot` provides 20 'before' and 'after' measurements of CD4 counts on HIV-positive patients.

(a) Read the help page for the background.

(b) Find a 90% confidence interval for the correlation (as in the preliminary material).

(c) Let us start by finding a 90% Monte-Carlo confidence interval based on bivariate normality. As ever, there is more than one way to do compute this! Here's an approach making use of the facilities of package `boot`:

```
library(MASS); library(boot)
cd4.rg <- function(data, mle) mvrnorm(nrow(data), mle$m, mle$v)

cd4.mle <- list(m=mean(cd4), v=var(cd4))
cd4.boot <- boot(cd4, corr, R=999,
                 sim = "parametric", ran.gen = cd4.rg, mle = cd4.mle)
cd4.boot
boot.ci(cd4.boot, type=c("norm", "basic", "perc"), conf=0.9)
boot.ci(cd4.boot,  type=c("norm", "basic", "perc"), conf=0.9,
        h=atanh, hinv=tanh)
```

(d) Now find a bootstrap confidence interval. `corr` is a function in package `boot` to compute weighted correlations.

```
cd4.boot <- boot(cd4, corr, stype="w", R=999)
cd4.boot
boot.ci(cd4.boot, conf=0.9)
boot.ci(cd4.boot, conf=0.9, h=atanh, hinv=tanh)
```

(e) The last part gave a warning about being unable to compute Studentized intervals. We can remedy that by

```
corr.fun <- function(d, w = rep(1, n))
{
    n <- nrow(d)
    w <- w/sum(w)
    m1 <- sum(d[,1]*w); m2 <- sum(d[,2]*w)
    v1 <- sum(d[,1]^2*w) - m1^2; v2 <- sum(d[,2]^2*w) - m2^2
    rho <- (sum(d[,1]*d[,2]*w) - m1*m2)/sqrt(v1 * v2)
    i <- rep(1:n, round(n*w))
    us <- (d[i, 1] - m1)/sqrt(v1)
    xs <- (d[i, 2] - m2)/sqrt(v2)
    L <- us*xs - 0.5*rho*(us^2 + xs^2)
    c(rho, sum(L^2)/n^2)
}
cd4.boot <- boot(cd4, corr.fun, stype="w", R=999)
boot.ci(cd4.boot, type="stud", conf=0.9)
boot.ci(cd4.boot, type="stud", conf=0.9,
        h=atanh, hdot=function(r) 1/(1-r^2), hinv=tanh)
```

but you will need to consult Davison & Hinkley (1997, practical 2.3) to understand the calculations.

(f) We can use the double bootstrap to see how well we have done at getting 90% coverage. This will take a couple of minutes: if you have more time increase `M`, if less decrease `R`.

```
page(nested.corr) # a function in the 'boot' package
cd4.nest <- boot(cd4, nested.corr, R=999, stype="w", t0=corr(cd4), M=249)

op <- par(pty = "s", xaxs = "i", yaxs = "i")
qqplot((1:999)/1000, cd4.nest$t[,2], pch=".", asp = 1,
       xlab="nominal", ylab="estimated")
abline(a = 0, b = 1, col = "grey")
par(op)
```

Now work out what corrections are needed to get a more accurate 90% interval.

Ex 3 This replicates part of figure 2, the version with edge-correction.

(a) First we quickly get a rough idea of the MLE:

```
library(spatial)
towns <- ppinit("towns.dat")
fac <- (69*68)/(2*40*40)
tget <- function(x, R=3.5) fac*pi*(Kfn(x, R, 1)$y)^2
Tget <- function(x, R=3.5) sum(dist(cbind(x$x, x$y)) < R)
t0 <- tget(towns)
R <- 100
cv <- seq(0, 1, 0.2)
res <- numeric(length(cv)) # res[1] = 0
for(i in 2:6)
    res[i] <- mean(replicate(R, tget(Strauss(69, c=cv[i], r=3.5))))
plot(cv, res, type="l")
abline(h=t0, col="grey")
```

(b) This suggests zooming in to $(0.4, 0.5)$. We do will do more runs: computers are fast enough these days.

```
R <- 1000
cv <- seq(0.4, 0.5, len=6)
res <- numeric(length(cv))
sds <- numeric(length(cv))
for(i in seq_along(cv)) {
    z <- replicate(R, tget(Strauss(69, c=cv[i], r=3.5)))
    res[i] <- mean(z)
    sds[i] <- sd(z)/sqrt(R)
}
plot(cv, res)
abline(h=t0, col="grey")
abline(lm(res ~ cv))
arrows(cv, res-1.96*sds, cv, res+1.96*sds,
       angle=90, code=3, length=0.1, xpd=TRUE)
```

(c) Now make use of these results to estimate the MLE of c, and give some indication of the inaccuracy. Choose some more simulations to do to get a more accurate result for the same amount of computation.

(d) How well can we do with polysampling? We do need to estimate the ratio of normalizing constants, and we need the uncorrected counts.

```
c0 <- 0.45
runs <- numeric(R); rs <- numeric(R)
for(i in 1:R) {
    xx <- Strauss(69, c=c0, r=3.5)
    runs[i] <- tget(xx)
    rs[i] <- Tget(xx)
}
cv <- seq(0.4, 0.5, len=6)
res <- numeric(length(cv))
for(i in seq_along(cv))
    res[i] <- mean(runs * (cv[i]/c0)^rs)/mean((cv[i]/c0)^rs)
points(c0, mean(runs), col="blue"); lines(cv, res, col="blue")
```

(e) Now let us try stochastic approximation. The sequence (a_n) was chosen by trial-and-error.

```
R <- 1000
doit <- function(ave=FALSE) {
    res <- numeric(R)
    cv <- runif(1, 0.4, 0.5) # initial guess.
    for(i in 1:R) {
        a <- 0.5*(i+5)^-0.7
        err <- tget(Strauss(69, c=cv, r=3.5))/t0  - 1
        cv <- cv - a*err
        res[i] <- cv
    }
    if(ave) cumsum(res)/(1:R) else res
}
res <- doit()
plot(res, type="l", ylim=c(0.4, 0.5))
for(i in 2:5) lines(doit(), col=i)
```

That was a näive form of Robbins–Munro. Clearly we can get a more accurate estimate by local averaging, and in what is known as Polyak–Ruppert averaging we can get optimum

convergence rates, e.g.

```
res <- doit(TRUE)
plot(res, type="l", ylim=c(0.4, 0.5))
for(i in 2:5) lines(doit(TRUE), col=i)
```

Experiment with other choices of $a_n = An^{-\gamma}$: what is a good choice will depend on whether averaging is done.

Ex 4 We continue the LD50 example from the lectures. Using `MCMCpack` we had

```
ldose <- rep(0:5, 2)
numdead <- c(1, 4, 9, 13, 18, 20, 0, 2, 6, 10, 12, 16)
sex <- factor(rep(c("M", "F"), c(6, 6)))
SF <- cbind(numdead, numalive = 20 - numdead)
resp <- rep(rep(c(1,0), 12), times = t(SF))
budworm <- data.frame(resp, ldose = rep(ldose, each = 20),
                      sex=rep(sex, each = 20))
summary(glm(resp ~ sex*ldose, family = binomial, data = budworm))

library(MCMCpack)  ## loads package 'coda'
fit <- MCMClogit(resp ~ sex*ldose, data = budworm)
summary(fit)
plot(fit)
acfplot(fit) # suggests thining
fit <- MCMClogit(resp ~ sex*ldose, data = budworm, mcmc=1e5, thin = 20)
summary(fit)
HPDinterval(fit)
```

Now we can explore the posterior distribution of LD50: this uses translucent points so that a build-up of colour indicates density.

```
## plot beta vs alpha for females
library(MASS)
contour(kde2d(fit[,1], fit[,3], n=50), xlab="alphaF", ylab="betaF")
points(fit[, c(1,3)], pch=".", col=rgb(0,0,1,0.2))
ld50F <- as.mcmc(2^-fit[,1]/fit[,3])
ld50M <- as.mcmc(2^-(fit[,1]+fit[,2])/(fit[,3] + fit[,4]))
range(ld50M); range(ld50F)
ld50 <- mcmc(cbind(M=ld50M, F=ld50F))
plot(ld50)
acfplot(ld50)
HPDinterval(ld50)
```

There is no hint of negative slopes. You can use `codamenu()` to explore the analysis facilities.

Ex 5 Now we try out BUGS. The first thing we need is a model file `budworm.bug`, which needs to be copied from the notes to the current directory. (Note that `openbugs` writes files in the current directory, so it must be writable.)

```
library(R2WinBUGS)
options(BRugsVerbose = FALSE)  # reduce chatter
budworm.sim <- openbugs(list("numdead", "ldose"),
                        list(alphaM = 0, betaM = 0, alphaF = 0, betaF = 0),
                        c("alphaM", "alphaF", "betaM", "betaF"),
                        model.file = "budworm.bug", DIC = FALSE,
                        n.chains = 1, n.iter = 10000)
budworm.sim
plot(budworm.sim)
res <- mcmc(budworm.sim$sims.matrix)
```

The last line gives results in `coda` format which can be plotted etc as before.

Now experiment with multiple starting points. To do so we set up a function for `inits`.

```
inits <- function()
   list(alphaM = rnorm(1,0,10), betaM = rnorm(1),
        alphaF = rnorm(1,0,10), betaF = rnorm(1))
budworm.sim <- openbugs(list("numdead", "ldose"), inits,
                        c("alphaM", "alphaF", "betaM", "betaF"),
                        model.file = "budworm.bug", DIC = FALSE,
                        n.chains = 5, n.iter = 1000)
budworm.sim
plot(budworm.sim, TRUE)

args <- lapply(1:budworm.sim$n.chains,
               function(i) mcmc(budworm.sim$sims.array[, i, ]))
res <- do.call(mcmc.list, args)
densityplot(res) # and so on
```

Again, we convert the results to `coda` format.

Ex 6 There is code in the lecture notes for two appoaches to the change-point problem for coal-mining disasters. Try them out. The data are[41]

```
D <- c(4,5,4,1,0,4,3,4,0,6,3,3,4,0,2,6,3,3,5,4,5,
       3,1,4,4,1,5,5,3,4,2,5,2,2,3,4,2,1,3,2,1,1,1,1,
       1,3,0,0,1,0,1,1,0,0,3,1,0,3,2,2,0,1,1,1,0,1,0,
       1,0,0,0,2,1,0,0,0,1,1,0,2,3,3,1,1,2,1,1,1,1,2,
       4,2,0,0,0,1,4,0,0,0,1,0,0,0,0,0,1,0,0,1,0,1)
```

Things you might want to do with the results include

```
attach.bugs(coalmining.sim)
op <- par(mfrow=c(2,2))
plot(density(changeyear))
frame()
plot(density(b[,1]), xlab="beta1")
plot(density(b[,2]), xlab="beta2")
detach.bugs()
par(op)
```

[41] This version from Gamerman & Lopes (2006, p. 145) has the correct total, unlike that in Albert (2007).

Thursday Practical

The R code for these exercises has deliberately not been hidden inside a package—the intention is that you work through the R and BUGS code and understand what it is doing (not necessarily in the practical class but as part of the review work on this module).

Ex 7 Consider the SAT-V coaching example, for which the lecture notes give BUGS code. This exercise is based on Gelman *et al.* (2004, Appendix C), which is available on-line.

(a) Run the code in the lecture notes.

(b) Replace the normal prior on the school-improvement means by a t_4 distribution and repeat.

(c) Now consider doing the simulations in R itself, using a Gibbs sampler.

```
J <- 8
y <- c(28, 8, -3, 7, -1, 1, 18, 12)
sigma.y <- c(15, 10, 16, 11, 9, 11, 10, 18)

theta.update <- function() {
    V.theta <- 1/(1/tau^2 + 1/sigma.y^2)
    theta.hat <-(mu/tau^2 + y/sigma.y^2)*V.theta
    rnorm(J, theta.hat, sqrt(V.theta))
}
mu.update <- function() rnorm(1, mean(theta), tau/sqrt(J))
tau.update <- function() sqrt(sum((theta-mu)^2)/rchisq(1, J-1))

n.chains <- 5
n.iter <- 1000
thetai <- paste ("theta[", 1:J, "]", sep="")
sims <- array(, c(n.iter, n.chains, J+2),
              dimnames = list(NULL, NULL, c(thetai, "mu", "tau")))
for(m in 1:n.chains) {
    mu <- rnorm(1, mean(y), sd(y))
    tau <- runif(1, 0, sd(y))
    for(t in 1:n.iter) {
        theta <- theta.update(); mu <- mu.update(); tau <- tau.update()
        sims[t, m, ] <- c(theta, mu, tau)
    }
}
library(R2WinBUGS)
print(monitor(sims), digits=3)
library(coda)
args <- lapply(1:5, function(i) mcmc(sims[,i,]))
z <- do.call(mcmc.list, args)
summary(z)
densityplot(z)
plot(z, ask=T)
```

Explore the output further, and especially the size of the largest θ_i.

(d) We now replace the normal prior for the school means by a t_4 prior, which we do by writing $T = N/\sqrt{V}$ for $\nu V \sim \chi^2_\nu$.

```
nu <- 4
mu.update <- function() rnorm(1, sum(theta/V)/sum(1/V), sqrt(1/sum((1/V))))
tau.update <- function() sqrt(rgamma(1, J*nu/2+1, (nu/2)*sum(1/V)))
V.update <- function() (nu*tau^2 + (theta-mu)^2)/rchisq(J,nu+1)

for(m in 1:n.chains) {
    mu <- rnorm(1, mean(y), sd(y))
    tau <- runif(1, 0, sd(y))
    V <- runif(J, 0, sd(y))^2
    for(t in 1:n.iter) {
        theta <- theta.update(); V <- V.update()
        mu <- mu.update(); tau <- tau.update()
        sims[t,m,] <- c(theta, mu, tau)
    }
}
```

Now analyse as before.

(e) Finally, consider t_ν where ν is also an unknown parameter. This is again a Gibbs sampler, using with a Metropolis step to update $1/\nu$ (Gelman *et al.*, 2004, pp. 292, 454).

```
log.post <- function(theta, V, mu, tau, nu, y, sigma.y)
{
    sum(dnorm(y, theta, sigma.y, log=TRUE)) +
        sum(dnorm(theta, mu, sqrt(V),log=TRUE)) +
            sum(0.5*nu*log(nu/2) + nu*log(tau) -
                 lgamma(nu/2) - (nu/2+1)*log(V) - 0.5*nu*tau^2/V)
}

nu.update <- function(sigma.jump.nu = 1)
{
    nu.inv.star <- rnorm(1, 1/nu, sigma.jump.nu)
    if(nu.inv.star <= 0 | nu.inv.star > 1) {
        # do nothing
    } else {
        nu.star <- 1/nu.inv.star
        log.post.old <- log.post(theta, V, mu, tau, nu, y, sigma.y)
        log.post.star <- log.post(theta, V, mu, tau, nu.star, y, sigma.y)
        r <- exp(log.post.star - log.post.old)
        nu <- ifelse(runif(1) < r, nu.star, nu)
    }
    nu
}

sims <- array(, c(n.iter, n.chains, J+3),
              dimnames = list(NULL, NULL, c(thetai, "mu", "tau", "nu")))
for(m in 1:n.chains) {
    mu <- rnorm(1, mean(y), sd(y))
    tau <- runif(1, 0, sd(y))
    V <- runif(J, 0, sd(y))^2
    nu <- 1/runif(1, 0, 1)
    for(t in 1:n.iter) {
        theta <- theta.update(); V <- V.update()
        mu <- mu.update(); tau <- tau.update(); nu <- nu.update()
        sims[t,m,] <- c(theta, mu, tau, nu)
```

```
    }
}
```

Once again, analyse the results.

Ex 8 Suppose that the eighth school had shown a mean improvement of 120 and not 12 and re-analyse the data.

Ex 9 Consider the Australian AIDS survival example sketched in the lecture notes.

Try the code there, and experiment with tuning `scale`. To set the problem up you will need

```
library(MASS)
make.aidsp <- function() {
  cutoff <- 10043 # 1987-07-01 with origin 1960-01-01
  btime <- pmin(cutoff, Aids2$death) - pmin(cutoff, Aids2$diag)
  atime <- pmax(cutoff, Aids2$death) - pmax(cutoff, Aids2$diag)
  survtime <- btime + 0.5*atime
  status <- as.numeric(Aids2$status)
  data.frame(survtime, status = status - 1, state = Aids2$state,
    T.categ = Aids2$T.categ, age = Aids2$age, sex = Aids2$sex)
}
Aidsp <- make.aidsp()
library(survival)
```

Use the `mcmc` function from package `coda` to convert the simulations to `coda` objects, and explore the diagnostics of the latter package.

Would there be any point in centring the explanatory variables in this problem?