Parameter estimation

Daniel Mortlock (mortlock@ic.ac.uk)
Last modified: September 12, 2013

1 Introduction

Parameter estimation is one of the most commonly encountered tasks in statistical analysis and one of simplest conceptually. Given some measured data, $\boldsymbol{d} = \{d_1, d_2, \ldots, d_N\} = \{d_i\}$, what constraints can be placed on the values of the parameters, $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_{N_p}\} = \{\theta_j\}$, of a model which is assumed to have generated the data? The full answer to this question is the posterior probability distribution in the parameters' values, $\Pr(\boldsymbol{\theta}|\boldsymbol{d})$, conditional both on the data and the overall framework or paradigm being considered (although this is left implicit, as every probability here is conditioned on this same information). In this context, Bayes's theorem is most usefully written in the same form as in Eq. (**??**), giving

$$\Pr(\boldsymbol{\theta}|\boldsymbol{d}) = \frac{\Pr(\boldsymbol{\theta})\Pr(\boldsymbol{d}|\boldsymbol{\theta})}{\int \Pr(\boldsymbol{\theta'})\Pr(\boldsymbol{d}|\boldsymbol{\theta'})\,\mathrm{d}\boldsymbol{\theta'}}, \tag{1}$$

where the only role of the denominator here is to ensure that the posterior is unit-normalised.

Question: *What information could violate Eq. (1)?*

1.0.1 Marginalisation

It is often the case that it is only some subset of the parameters are of real interest, the others being nuisance parameters included only out of necessity. A common example might be a model in which the data are (assumed to be) subject to normally-distributed measurement noise of unknown mean; the mean of the distribution is the parameter of interest, but the unknown variance must be included as a parameter as well.

Question: *What would be the (erroneous) result of just using some fixed best-fit estimate of the variance?*

In such cases the full Bayesian result is the joint posterior distribution in both the parameter(s) of interest and the nuisance parameter(s), but this can be made less cumbersome by marginalising (*i.e.*, integrating out) the nuisance parameters. If, in an $N_p = 4$ parameter model the parameters θ_1 and θ_2 are of interest but θ_3 and θ_4 are nuisance parameters, then the posterior distribution of interest is

$$\Pr(\theta_1, \theta_2|\boldsymbol{d}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Pr(\theta_1, \theta_2, \theta'_3, \theta'_4|\boldsymbol{d})\,\mathrm{d}\theta'_3\,\mathrm{d}\theta'_4. \tag{2}$$

In general any sub-set of the parameters can be marginalised over this way – it is even possible, if not very interesting, to marginalise over *all* the parameters – but the result should just be unity in every case.

1.0.2 Improper priors

The parameter estimation formalism described above is complete and rigorous if the prior is a correct, unit-normalised distribution in the parameters $\boldsymbol{\theta}$ that satisfies $\Pr(\boldsymbol{\theta}) \geq 0$ for all $\boldsymbol{\theta}$ and $\int \Pr(\boldsymbol{\theta'})\,\mathrm{d}\boldsymbol{\theta'} = 1$. In some cases, however, it is possible to obtain the correct parameter posterior with an improper prior which does not satisfy the normalisation constraint above.

The simplest example is if the prior in a one-parameter model is taken to be an infinitely-broad uniform distribution. This cannot be normalised and so the prior can only be specified as a constant, which for simplicity is taken to be unity: $\Pr'(\theta) = 1$, where the prime is used to denote the fact that the distribution is not correctly normalised. If the integral of the likelihood over the full range of possible values of the model parameter, $\int_{-\infty}^{\infty} \Pr(\boldsymbol{d}|\theta')\,\mathrm{d}\theta'$, is finite then a correctly normalised posterior distribution is

$$\Pr(\theta|\boldsymbol{d}) = \frac{\Pr'(\theta)\Pr(\boldsymbol{d}|\theta)}{\int_{-\infty}^{\infty}\Pr'(\theta)\Pr(\boldsymbol{d}|\theta')\,\mathrm{d}\theta'} = \frac{\Pr(\boldsymbol{d}|\theta)}{\int_{-\infty}^{\infty}\Pr(\boldsymbol{d}|\theta')\,\mathrm{d}\theta'}, \tag{3}$$

where the second expression is for the case $\Pr'(\theta) = 1$ discussed above. As reasonable as this final result is, the process leading up to is flawed, and improper priors should not be used without very careful thought. A more correct way of obtaining the above result would be to adopt a uniform prior

$$\Pr(\theta) = \mathrm{U}(\theta; a, b) = \frac{\Theta(x-a)\,\Theta(b-x)}{b-a}, \tag{4}$$

and then take the limits $a \to -\infty$ and $b \to \infty$ *after* calculating the posterior. In the case that the above integral is finite the result would match that obtained by using the improper prior; but if the likellihood was, *e.g.*, constant for all values of θ greater than some value, then the limit might not exist.

1.0.3 Unnormalised posteriors

An implication of the above results is that the normalisation of the posterior is generally unimportant when doing parameter estimation. It is only the dependence of the posterior on the parameters that matters; multiplying it by a constant factor would not change any subsequent inferences. Alternatively, the prior could be calculated with any arbitary (positive) choice of normalisation and then multiplied through by its inverse to obtain a correctly normalised prior. Indeed, this is what the denominator in Eq. (1) effecitvely does. This normalisation procedure is generally simple numerically – the potentially hard task is finding the peaks in the posterior – but often impossible analytically.

For that reason it is often convenient to operate with an unnormalised (but normalisable) posterior distribution, denoted $\Pr'(\boldsymbol{\theta}|\boldsymbol{d})$. This is not uniquely defined – it can be any function of the form $c\Pr(\boldsymbol{\theta}|\boldsymbol{d})$, provided only that $c > 0$. A number of the following examples become far easier to understand by using the obvious unnormalised posterior

$$\Pr'(\boldsymbol{\theta}|\boldsymbol{d}) = \Pr(\boldsymbol{\theta})\Pr(\boldsymbol{d}|\boldsymbol{\theta}), \tag{5}$$

or even using an improper prior as well, to give

$$\Pr'(\boldsymbol{\theta}|\boldsymbol{d}) = \Pr'(\boldsymbol{\theta})\Pr(\boldsymbol{d}|\boldsymbol{\theta}). \tag{6}$$

As with the use of improper priors, care must be taken that none of the $\boldsymbol{\theta}$-dependence is lost when working with the unnormalised posterior. Conversely, any method which obtains the correct $\boldsymbol{\theta}$-depdendence of the posterior is acceptable[1].

[1] It is even more important to remember that short-cuts involving improper priors and unnormalised posteriors

1.1 Incomplete representations of the posterior distribution

The posterior probability distribution $\Pr(\boldsymbol{\theta}|\boldsymbol{d})$ is the full result of any parameter inference problem – any reduced form of this represents a loss of information (although in the case of marginalisation the information lost is possibly unimportant). If the posterior distribution is analytic then it may be possible to represent this full distribution with a compact algebraic formula, but outside of textbook problems this is a very rare (if happy) state of affairs.

More common is that the information in the posterior must be condensed into a few numbers (*e.g.*, an estimate and uncertainty) or a simple plot (*e.g.*, showing a univariate marginal distribution in one parameter or the bivariate marginal distribution in two parameters). Indeed, any non-analytic posterior distribution will inevitably be represented by a finite set of numbers that cannot encode an arbitrary density in even one variable. Some reasonable options for representing a posterior distribution are described here.

1.1.1 A posteriori point estimates

The most extreme case of compressing the posterior distribution is where the $\Pr(\boldsymbol{\theta}|\boldsymbol{d})$ is to be represented by a single number. This is very much against the underlying principles of Bayesian inference – why go to all the trouble of self-consistently treating uncertainty, only to then disregard the uncertainty in one's conclusions? – but sometimes necessary. The generic motivation for committing this apparent crime is that a decision must be made: a bet placed or a design choice finalised or a route chosen. In all such examples there is an implied loss in making a bad choice, and there is also an entire field, decision theory, that provides logical guidance about how to proceed. In the absence of any further information (*i.e.*, a loss function or similar) there is no single algorithm for best mapping from $\Pr(\boldsymbol{\theta}|\boldsymbol{d})$ to an estimate $\hat{\boldsymbol{\theta}}$, although there are a few standard and "sensible" choices. (Any time there is a choice or option appears it is an indicator that the rules of Bayesian inference are not being rigorously followed.)

Maximum a posteriori probability estimate

Perhaps the most obvious is the maximum a posteriori probability (MAP) estimate, in which $\hat{\boldsymbol{\theta}}$ is taken to be the value of $\boldsymbol{\theta}$ for which $\Pr(\boldsymbol{\theta}|\boldsymbol{d})$ is highest. Mathematically, the MAP estimate can be written as

$$\hat{\boldsymbol{\theta}}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\theta}} \left[\Pr(\boldsymbol{\theta}|\boldsymbol{d})\right]. \tag{7}$$

One disadvantage of the MAP estimate is that it is not uniqely defined if the posterior has multiple modes or a flat peak. Moreover, if the posterior has multiple peaks then it is quite possible that $hat\boldsymbol{\theta}_{\mathrm{MAP}}$ is arbitrarily far away from most of the posterior mass. If $\hat{\boldsymbol{\theta}}_{\mathrm{MAP}}$ is defined then the MAP estimate can seem a reasonably intuitive choice: it is the best bet option given all the available information, in the sense that $\hat{\boldsymbol{\theta}}_{\mathrm{MAP}}$ is the most probable draw from $\Pr(\boldsymbol{\theta}|\boldsymbol{d})$.

Question: *What is the MAP estimate of θ given the (admittedly unrealistic) posterior density $\Pr(\theta|\boldsymbol{d}) = 0.01\,\delta_{\mathrm{D}}(\theta - 0.34) + 0.99\,\mathrm{N}(\theta; 14.55, 0.31^2)$? Is this sensible? Is this useful?*

If the posterior distribution is uniform (and possibly even if it is not) then $\hat{\boldsymbol{\theta}}_{\mathrm{MAP}} = \hat{\boldsymbol{\theta}}_{\mathrm{ML}}$, the maximum likelihood estimate of $\boldsymbol{\theta}$ that might have been obtained if the prior information was

are only ever valid for parameter estimation (and sometimes not even then). The generally more difficult problem of model comparison requires considerably more care in this regard.

ignored. However, while the maximum likelihood estimate is invariant under reparameterisations, the MAP estimate is not, in general, because of the need to consistently transform the prior distribution.

Question: *What is the MAP estimate of* θ *if* $\Pr(\theta|\boldsymbol{d}) = \mathrm{N}(\theta; 1.1, 0.3)$*? Assuming a uniform prior in* θ*, how would this result change if the parameter was instead* $\psi = \sinh(\theta)$*?*

Posterior mean

Another common option is the posterior mean, defined by

$$\hat{\boldsymbol{\theta}}_{\mathrm{mean}} = \int \Pr(\boldsymbol{\theta}'|\boldsymbol{d})\, \boldsymbol{\theta}' \mathrm{d}\boldsymbol{\theta}'. \tag{8}$$

This is a more global quantity than $\hat{\boldsymbol{\theta}}_{\mathrm{MAP}}$, as it incorporates information from the entire posterior distribution, but that does not necessarily make it more useful. There is no guarantee that $\hat{\boldsymbol{\theta}}_{\mathrm{mean}}$ exists – if the tails of $\Pr(\boldsymbol{\theta}'|\boldsymbol{d})$ are too heavy then the above integral might not be defined. Another potential problem is that the posterior probability that $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{\mathrm{mean}}$, can be arbitrarily small, and there is no reason to prevent the absurd situation that $\Pr(\hat{\boldsymbol{\theta}}_{\mathrm{mean}}|\boldsymbol{d}) = 0$ (although it is only absurd because of the choice, not supported by the rules of Bayesian inference, to regard the posterior mean as significant).

Question: *What is the posterior mean of* θ *given the posterior density* $\Pr(\theta|\boldsymbol{d}) = 1/\{\pi[1 + (\theta - 43.69)^2]\}$*?*

Question: *What is the posterior mean of* θ *given the (admittedly unrealistic) posterior density* $\Pr(\theta|\boldsymbol{d}) = 0.3\,\mathrm{N}(\theta; -452.1, 1.2^2) + 0.7\,\mathrm{N}(\theta; 6996.4, 3.0^2)$*? Is this sensible? Is this useful?*

1.1.2 Credible regions

The natural extension of characterising a posterior distribution $\Pr(\boldsymbol{\theta}|\boldsymbol{d})$ by a single value, $\hat{\boldsymbol{\theta}}$, is to complement this with some measure of the uncertainty in $\boldsymbol{\theta}$. The standard way to do this is to define credible regions (or credible intervals in the case of a single parameter) that enclose a specified fraction, f (with $0 \leq f \leq 1$), of the posterior probability. There is, again, ambiguity in this choice, as there are, in general, infinitely many regions of parameter space that enclose a fraction f of the posterior probability. It is also common practice to consider multiple, nested credible regions containing different fractions of the posterior probability (*e.g.*, 1-, 2- and 3-"sigma" regions enclosing 68.2%, 95.4% and 99.7% of the posterior, implying $f = 0.682$, $f = 0.954$ and $f = 0.997$, respectively).

In the univariate case of a posterior distribution in a parameter θ (that may be the result of marginalisation), any interval between θ_{min} and θ_{max} that satisfies

$$f = \Pr(\theta_{\mathrm{min}} \leq \theta \leq \theta_{\mathrm{max}}) = \int_{\theta_{\mathrm{min}}}^{\theta_{\mathrm{max}}} \Pr(\theta'|\boldsymbol{d})\, \mathrm{d}\theta' \tag{9}$$

is valid. For a given choice it is reasonable to simply report the limits. If the posterior is unimodal and the limits have been chosen to be reasonably central, to use the limits together with an estimate $\hat{\theta}$ to report uncertainties as $\theta = \hat{\theta}^{+(\theta_{\mathrm{max}} - \hat{\theta})}_{-(\hat{\theta} - \theta_{\mathrm{min}})}$, although great care should be taken to specify the nature of both the estimate and the choice of interval used.

In a multivariate problem the same freedom is present.

Highest posterior density regions

The most obvious – and generally "best" – choice is the highest posterior density (HPD) region. Except in the case of some extreme distributions (see below), the HPD is uniquely defined by the value of the probability density, p, for which

$$f = \int \Pr(\boldsymbol{\theta}'|\boldsymbol{d})\,\Theta[\Pr(\boldsymbol{\theta}'|\boldsymbol{d}) - p]\,\mathrm{d}\boldsymbol{\theta}'. \tag{10}$$

All parameter values such that $\Pr(\boldsymbol{\theta}|\boldsymbol{d}) \geq p$ are inside the HPD credible region. The boundary (or, in the case of one dimension, limits) is defined by the points such that $\Pr(\boldsymbol{\theta}|\boldsymbol{d}) = p$; hence the models on the boundary are, appealingly, all equally probable. For this reason, bivariate parameter contstraints are often illustrated by showing HPD contours enclosing various fractions of the probability: these contours give information about both the probability density and the integrated probability. The fact that every point inside the HPD region is more probable than every point outside leads to another appealing property of the HPD region: it is smaller than any other credible region containing the same fraction of the posterior probability.

Question: *For what univariate distributions* $\Pr(\theta|\boldsymbol{d})$ *are the HPD regions undefined for general values of* f*? For what univariate distributions* $\Pr(\theta|\boldsymbol{d})$ *are the HPD regions undefined for a particular value of* f $(= 0.5$*, say)?*

Upper and lower bounds

If the aim is to place an upper bound on the value of a single parameter θ then it is reasonable to choose $\theta_{\mathrm{min}} = -\infty$ and then find θ_{max} such that $\int_{-\infty}^{\theta_{\mathrm{max}}} \Pr(\theta'|\boldsymbol{d})\,\mathrm{d}\theta' = f$ where, typically, $f \gtrsim 0.9$. In this case it is reasonably to state that the data (in combination with whatever prior information was used) give a $100f\%$ upper bound of $\theta = \theta_{\mathrm{max}}$.

(A similar result holds for a lower bound.)

2 Grid-based methods

2.1 Introduction

The simplest numerical method of approximating a posterior distribution $\Pr(\boldsymbol{\theta}|\boldsymbol{d})$ is to base all calculations on an array of values of the (not necessarily normalised) density evaluated on a regular grid. Integrals over the distribution are approximated by a simple sum, and it is fairly straightforward to calculate the other standard derived quantities.

The advanatages of grid-based methods are:

- accuracy/speed, as the regular spacing of the evaluations means that there is minimal redundancy – the errors in the estimates of standard derived quantities decrease reciprocally with the number of samples;
- repeatability, as (assuming only that the grid parameters are held fixed and the posterior density can be evaluated directly) the results will be identical if re-calculated;
- simplicity, as there is nothing sophisticated about the algorithm of setting up a regular grid;
- no normalisation requirement, as it is possible to work with a grid of unnormalised posterior values.

Another utility of these grid-based methods is they are often used to post-process samples drawn from the posterior using the the algorithms described in Section 3.

The disadvantages – or, really, limitations – of grid-based methods are:

- some external knowledge of the range of parameter values for which the posterior is significant is required (and while the prior range might be sufficient, often it is not);
- potential "systematic" errros if there is the posterior is periodic or varies on a scale comparable to the separation between grid points;
- such methods are only practical for problems of low dimensionality (anything beyond 3-4 parameters becomes computationally unfeasible);
- it can be unwieldy to write code to handle a grid of an arbitrary number of dimensions.

The algorithms required to generate an array of posterior samples and then to process them into useful outputs are very simple, and certainly much simpler than their pure mathematical expression, especially within the context of multiple dimensions. They are much more naturally described in terms of algorithms. Here the methods are presented for the case of a bivariate posterior distribution $\Pr(\boldsymbol{\theta}|\boldsymbol{d}) = \Pr(x, y|\boldsymbol{d})$ where $N_p = 2$ and $\boldsymbol{\theta} = (x, y)$; the generalisation to higher (or lower) dimensional problems is conceptually straightforward.

2.2 Grid generation

Access to (*i.e.*, the ability to evaluate) the distribution $\Pr(x, y|\boldsymbol{d})$ is not a sufficient starting point for grid-based methods. Some information is required to decide what range of parameter values are to be considered: it must be possible to determine values for $x_{\min}$, $x_{\max}$, $y_{\min}$ and $y_{\max}$ such that $1 - \Pr(x_{\min} \leq x \leq x_{\max}, y_{\min} \leq y \leq y_{\max}|\boldsymbol{d}) \ll 1$ (*i.e.*, that the region bounded by $x_{\min}$, $x_{\max}$, $y_{\min}$ and $y_{\max}$ contains almost all of the probability). However it is not sufficient to simply make this range arbitarily large, as then the majority – if not all – of the grid points would fall in regions of low probability. There are general algorithms for doing this (*e.g.*, the sampling methods described in Section 3.3), but to have to resort to them would largely defeat the purpose of a grid-based approach. Hence it is assumed here that reasonable values for $x_{\min}$, $x_{\max}$, $y_{\min}$ and $y_{\max}$ are known from some external information, although it cannot be over-emphasised that if such limits cannot easily be obtained then grid-based methods can be rendeed immediately useless.

The next decision to be made is the resolution of the grid, defined by the number of columns, N_{c}, and the number of rows, N_{r}. This choice is subject to the usual trade-off of accuracy *vs.* speed. A bare minimum in each dimension is ~ 10; anything much over $\sim 10^2$ is usually unnecessary for the posterior distributions usually encountered in real-world problems. As with the range, this is subject to some degree of trial and error.

The grid hence covers the range $x_{\min} \leq x \leq x_{\max}$ and $y_{\min} \leq y \leq y_{\max}$ with a $N_{\mathrm{c}} \times N_{\mathrm{r}}$ array of cells of area $\Delta x \times \Delta y$, where $\Delta x = (x_{\max} - x_{\min})/2$ and $\Delta y = (y_{\max} - y_{\min})/2$. From this point follow this prescriptive algorithm:

1. For each combination of column, $c(\in \{1, 2, \ldots, N_{\mathrm{c}}\})$ and row, $r(\in \{1, 2, \ldots, N_{\mathrm{r}}\})$ calculate[2]

$$(x_c, y_r) = \left[x_{\min} + \frac{c - 1/2}{N_{\mathrm{c}}}(x_{\max} - x_{\min}), y_{\min} + \frac{r - 1/2}{N_{\mathrm{r}}}(y_{\max} - y_{\min})\right],$$

   where the points are chosen to lie in the middle of each grid cell.

2. For each element in the array calculate the unnormalised posterior,

$$p'_{c,r} = \Pr(x, y) \Pr(\boldsymbol{d}|x, y).$$

3. Normalise the posterior samples numerically by calculating

$$p_{c,r} = \frac{p'_{c,r}}{\sum_{c=1}^{N_{\mathrm{c}}} \sum_{r=1}^{N_{\mathrm{r}}} p'_{c,r}}.$$

   (Even though this step is sometimes unnecesary, it is inexpensive numerically and simplifies the subsequent analysis.)

A piece-wise constant approximation to the posterior is now provided by

$$\Pr(x, y|\boldsymbol{d}) \simeq \frac{1}{(x_{\max} - x_{\min})(y_{\max} - y_{\min})} \sum_{c=1}^{N_{\mathrm{c}}} \sum_{r=1}^{N_{\mathrm{r}}} \tag{11}$$

[2]It is sufficient to just calcuate $\{x_c\}$ and $\{y_r\}$ sepaarately, but for the memory and processing requirements are so minor that it is much easier to have access to the full array of (x_c, y_r) values.

$$\Theta[x-(x_c-\Delta x/2)]\,\Theta[(x_c+\Delta x/2)-x]\,\Theta[y-(y_c-\Delta y/2)]\,\Theta[(y_c+\Delta y/2)-y]\,p_{c,r},$$

which is zero outside the area covered by the grid. More complicated interpolation schemes could also be used to go from $\{p_{c,r}\}$ to a distribution defined for all x and y, but the main point is that the continuous function $\Pr(x_c, y_r|\boldsymbol{d})$ is now encoded (albeit approximately) in the finite set of numbers $\{p_{c,r}\}$.

2.3 Post-processing

A variety of useful quantities can be estimated by performing simple numerical (*e.g.*, sums, sorting, *etc.*) operations on $\{p_{c,r}\}$. Some representative examples are given here.

2.3.1 Marginal distributions

The marginal distributions $\Pr(x|\boldsymbol{d})$ and $\Pr(y|\boldsymbol{d})$ are approximated by

$$\Pr(x|\boldsymbol{d}) \simeq \frac{1}{x_{\max}-x_{\min}} \sum_{c=1}^{N_{\rm c}} \Theta[x-(x_c-\Delta x/2)]\,\Theta[(x_c+\Delta x/2)-x] \sum_{r=1}^{N_{\rm r}} p_{c,r} \tag{12}$$

and

$$\Pr(y|\boldsymbol{d}) \simeq \frac{1}{y_{\max}-y_{\min}} \sum_{r=1}^{N_{\rm r}} \Theta[y-(y_r-\Delta y/2)]\,\Theta[(y_r+\Delta y/2)-y] \sum_{c=1}^{N_{\rm c}} p_{c,r}, \tag{13}$$

respectively.

2.3.2 HPD credible regions

To find the HPD credible region containing a fraction f from a grid of posterior density evaluations, $\{p_{c,r}\}$, requires several distinct steps:

1. Sort (or index) the $\{p_{c,r}\}$ from the highest value to the lowest.
2. Accumulate the total fraction of the probability using the sorted values of p, stopping when the sum first exceeds f.
3. Identify this (or the previous, or an interpolated) value of $p(f)$ as that which encloses the fraction f.

A contour drawn at a level $p(f)$ through the array $\{p_{c,r}\}$ encloses the HPD credible region containing a fraction f of the posterior probability.

An analogous operation in one dimension defines the extent of the (marginalised) HPD credible interval.

2.3.3 MAP estimate

An approximation to the MAP estimate of (x, y) is the canonical grid point (x_c, y_r) corresponding to the highest value of $p_{c,r}$. This could be refined using a more sophisticated interpolation algorithm.

2.3.4 Posterior mean and covariance

If the distribution $\Pr(x, y|\boldsymbol{d})$ is known (how?) to be fairly simple (*e.g.*, unimodal and reasonably normal) it might be reasonable to charcaterise it by the posterior mean, $(\theta_{1,\text{mean}}, \theta_{2,\text{mean}}) = (x_{\text{mean}}, y_{\text{mean}})$ and the covariance matrix $\mathsf{C} = [(C_{1,1}, C_{1,2}), (C_{2,1}, C_{2,2})] = [(C_{x,x}, C_{x,y}), (C_{x,y}, C_{y,y})]$. The natural estimates for these quantities are

$$\hat{\theta}_{p,\text{mean}} = \sum_{c=1}^{N_\text{c}} \sum_{r=1}^{N_\text{r}} p_{c,r} \theta_{p,c,r}, \tag{14}$$

and

$$\hat{C}_{p,p'} = \sum_{c=1}^{N_\text{c}} \sum_{r=1}^{N_\text{r}} p_{c,r} (\theta_{p,c,r} - \hat{\theta}_{p,\text{mean}}), (\theta_{p',c,r} - \hat{\theta}_{p',\text{mean}}), \tag{15}$$

where $p \in \{1, 2\}$ and $p' \in \{1, 2\}$.

3 Sampling methods

3.1 Introduction

A large set of samples generated from a distribution can used to estimate (to arbitrary precision, in the limit of an infinite number of draws) any aspect of the generating distribution, and algorithms exist to generate samples from (essentially) arbitrary distributions. These facts combine to explain the preponderance of sampling methods for practical parameter estimation problems. The best algorithms can, with minimal tuning, automatically find and explore the peaks of a probability distribution, allowing full (if approximate) Bayesian parameter inference.

While there is a bewildering number of simulation (*i.e.*, sample generating) algorithms in use, their basic output is the same: a potentially correlated set of samples $\{\boldsymbol{\theta}_s\} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_{N_\mathrm{s}}\}$ with associated weights $\{W_s\} = \{W_1, W_2, \ldots, W_{N_\mathrm{s}}\}$, drawn from the target posterior distribution $\Pr(\boldsymbol{\theta}|\boldsymbol{d})$. Given these samples and weights, the distribution can be approximated as

$$\Pr(\boldsymbol{\theta}|\boldsymbol{d}) \simeq \frac{\sum_{s=1}^{N_\mathrm{s}} W_s \, \delta_\mathrm{D}(\boldsymbol{\theta} - \boldsymbol{\theta}_s)}{\sum_{s=1}^{N_\mathrm{s}} W_s}, \tag{16}$$

where the denominator can be ignored if the weights have been scaled to add to unity. While most distributions of interest are much smoother than the above sum of delta functions, any derived quantities of interest are inevitably obtained by some sort of averaging, either over a small region (in the case of binning) or globally (in the case of an integral quantity). It is really the values of these derived quantities that can be estimated to high accuracy by sampling the posterior distribution.

3.2 Post-processing

A list of samples $\{\boldsymbol{\theta}_s\}$ is almost never interpreted or presented directly, as the human mind cannot make sense of huge lists of numbers. The closest thing to any direct presentation of the samples is a scatter plot, although even this is problematic if the samples have non-uniform weights, as there is no simple way to reflect this graphicly. Rather, the uitlity of $\{\boldsymbol{\theta}_s\}$ is to calculate various derived quantities; algorithms for doing so are presented here.

3.2.1 Posterior mean and covariance

The natural estimate of the posterior mean of the p'th parameter is

$$\hat{\theta}_{p,\mathrm{mean}} = \frac{\sum_{s=1}^{N_\mathrm{s}} W_s \, \theta_{p,s}}{\sum_{s=1}^{N_\mathrm{s}} W_s}. \tag{17}$$

The corresponding covariance matrix then has elements given by

$$\hat{C}_{p,p'} = \frac{\sum_{s=1}^{N_\mathrm{s}} W_s \, (\theta_{p,s} - \hat{\theta}_{p,\mathrm{mean}})(\theta_{p',s} - \hat{\theta}_{p',\mathrm{mean}})}{\sum_{s=1}^{N_\mathrm{s}} W_s}. \tag{18}$$

3.2.2 HPD credible regions

As the boundary of the HPD credible region depends in part on the probability density, it is necessary to bin the samples from the posterior in some way, which in turn implies that a range for each of the parameters must be chosen. A simple, if not completely robust, option is to use the posterior mean and variance calculated as described above to determine a (hopefully) inclusive range in each parameter. For the p'th parameter a reasonable range might be given by $\theta_{p,\mathrm{min}} = \hat{\theta}_{p,\mathrm{mean}} - 5\sigma_p$ and $\theta_{p,\mathrm{max}} = \hat{\theta}_{p,\mathrm{mean}} + 5\sigma_p$, where $\sigma_p = \hat{C}_{p,p}^{1/2}$. The weighted posterior sample can then be binned into a grid and the methods described in Section 2 followed to obtain HPD credible regions.

Other methods (*e.g.*, using Delaunay tessellation) of going from the posterior samples to an estimate of the posterior density field are available, but while greater precision is possible, the price in terms of algorithmic complexity is high.

3.3 Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC) methods are, currently, the generic workhouse used to perform Bayesian parameter estimation. They are conceptually simple and in principle can provide samples from any (posterior) distribution . . . at least in the limit of an infinite number of density evaluations. In practice MCMC methods are not, of course, a pancea, and great care must be taken to check the results obtained.

3.3.1 Markov chains

A Markov chain is defined as an *ordered* sequence of points $\{\boldsymbol{\theta}_s\} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_{N_\mathrm{s}}\}$, in which the position of the s'th point, $\boldsymbol{\theta}_s$, depends explicitly only on $\boldsymbol{\theta}_{s-1}$. The s'th sample hence is drawn from a distribution of the from $\Pr(\boldsymbol{\theta}_s|\boldsymbol{\theta}_{s-1}, \boldsymbol{d})$, which is independent of $\boldsymbol{\theta}_{s-2}, \boldsymbol{\theta}_{s-3}, \ldots$ (and, in principle, also independent of $\boldsymbol{\theta}_{s-1}$, although in this case the ordering is arbitrary and so the sample is not really a chain; none of the MCMC processes considered here have this independence). As such, Markov chains cannot be the optimal way to explore an unknown distribution, as the information obtained from almost all previous steps is disregarded – but *some* information is retained. However, the fact that that the dependence is simple means that it is much easier to understand the stochastic properties of a Markov chain than that of a more complicated sampling algorithm in which all previous samples are utilised.

An human analogy of a Markov process is an amnesiac who can only remember the previous day: their current state/location is linked directly only to yesterday's, although the integrated state (of, *e.g.*, wealth) does propagate information from previous days. Another analogy is that of a person exploring the London Underground: their possible next journeys are determined only by their current stop (even if the current stop is the net result of many trips).

Algorithm

The MCMC algorithms considered here for obtaining the next element of the chain, $\boldsymbol{\theta}_s$, given the previous point, $\boldsymbol{\theta}_{s-1}$, all proceed according to:

1. Draw a trial point from a (as yet unspecified) proposal distribution $\Pr(\boldsymbol{\theta}_\mathrm{trial}|\boldsymbol{\theta}_{s-1}, \boldsymbol{d})$, the form of which can be chosen to increase the efficiency of the algorithm.

2. Accept the trial point with probability $\Pr(\text{accept}|\boldsymbol{\theta}_{\text{trial}}, \boldsymbol{\theta}_{s-1}, \boldsymbol{d})$, which for some algorithms is always unity (*i.e.*, the trial point is always accepted).
3. If the trial point is accepted then set $\boldsymbol{\theta}_s = \boldsymbol{\theta}_{\text{trial}}$; otherwise set $\boldsymbol{\theta}_s = \boldsymbol{\theta}_{s-1}$.

The two-step proposal and acceptance process could be combined into a single distribution by marginalising over the trial positions, giving

$$\begin{aligned}\Pr(\boldsymbol{\theta}_s|\boldsymbol{\theta}_{s-1}, \boldsymbol{d}) &= \int \{\Pr(\text{accept}|\boldsymbol{\theta}'_{\text{trial}}, \boldsymbol{\theta}_{s-1}, \boldsymbol{d}) \Pr(\boldsymbol{\theta}'_{\text{trial}}|\boldsymbol{\theta}_{s-1}, \boldsymbol{d}) \, \delta_{\text{D}}(\boldsymbol{\theta}_s - \boldsymbol{\theta}'_{\text{trial}}) \\ &\quad + [1 - \Pr(\text{accept}|\boldsymbol{\theta}'_{\text{trial}}, \boldsymbol{\theta}_{s-1}, \boldsymbol{d})] \, \delta_{\text{D}}(\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s-1})\} \, \text{d}\boldsymbol{\theta}'_{\text{trial}}, \end{aligned} \tag{19}$$

although it more intuitive to think about the two-step process. This process is repeated a large number of times (at least $\sim 10^3$ and as many as $\sim 10^7$ in some applications), building up a large set of samples.

As stated, however, there is no guarantee that this algorithm will sample from the desired distribution – at no point has the posterior $\Pr(\boldsymbol{\theta}|\boldsymbol{d})$ appeared in the formalism.

Exhaustiveness

The chain must be capable of exploring the entire parameter space, which remove certain forms of $\Pr(\boldsymbol{\theta}_s|\boldsymbol{\theta}_{s-1}, \boldsymbol{d})$ from consideration. If, *e.g.*, this distribution is non-zero only if $|\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s-1}|$ is less than some finite value then the chain would not be able to sample a posterior distribution with two widely-separated regions of non-zero probability. Also, if $\Pr(\boldsymbol{\theta}_s|\boldsymbol{\theta}_{s-1}, \boldsymbol{d})$ was periodic in structure then samples could never be generated in certain regions.

These considerations are critical mathematically, but rarely important in practice.

Stationary distribution

The main requirement for MCMC to be useful is that the samples are, at least in the limit of high s, drawn from $\Pr(\boldsymbol{\theta}|\boldsymbol{d})$. The distribution of the s'th element is given by working through the integral in Eq. (19), which reveals that

$$\begin{aligned}\Pr(\boldsymbol{\theta}_s|\boldsymbol{\theta}_{s-1}, \boldsymbol{d}) &= \Pr(\boldsymbol{\theta}_s|\boldsymbol{d}) \\ &\quad + [\textbf{more terms here}] \end{aligned} \tag{20}$$

Hence an MCMC algorithm will sample the desired distribution if, for all parameter combinations $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$,

$$\begin{aligned}&\Pr(\boldsymbol{\theta}_1|\boldsymbol{d}) \Pr(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \boldsymbol{d}, \text{trial}) \Pr(\text{accept}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{d}) \\ &= \Pr(\boldsymbol{\theta}_2|\boldsymbol{d}) \Pr(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \boldsymbol{d}, \text{trial}) \Pr(\text{accept}|\boldsymbol{\theta}_2, \boldsymbol{\theta}_1, \boldsymbol{d}), \end{aligned} \tag{21}$$

which can be rearranged to give

$$\frac{\Pr(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \boldsymbol{d}, \text{trial}) \Pr(\text{accept}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{d})}{\Pr(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \boldsymbol{d}, \text{trial}) \Pr(\text{accept}|\boldsymbol{\theta}_2, \boldsymbol{\theta}_1, \boldsymbol{d})} = \frac{\Pr(\boldsymbol{\theta}_2|\boldsymbol{d})}{\Pr(\boldsymbol{\theta}_1|\boldsymbol{d})}. \tag{22}$$

This condition is known as "detailed balance".

Burn-in

If the approximate location of the peak of the posterior is known then it is perfectly legitimate – and, indeed, preferable – to start sampling from that location in parameter space [*i.e.*, if $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{d})$ is known to be close to the peak of $\Pr(\boldsymbol{\theta}|\boldsymbol{d})$ then the first sample should be set to $\boldsymbol{\theta}_1 = \hat{\boldsymbol{\theta}}$ so that the chain is immediately sampling from the target distribution]. If the posterior is known (how?) to be unimodal then repeat chains can be started from the same position.

Unfortunately, it is general not known where the posterior is appreciable, but one of the key attributes of most MCMC algorithms is that they can usually be run with an arbitrary starting point. The resultant chain will then tend to propagate towards the peak of the posterior, a process known as "burn-in". The only problem is that these first points can be regions of parameter space with arbitrarily small probability that would most likely not have been sampled in a chain of plausible length.

The crude, if effective, solution to this problem is simply to remove these first elements from the chain, which can be done after sampling is completed. No general rigorous algorithms for doing this exist, but heuristic options abound. A simpe guide is that any sample which has an appreciable probability, relative to the peak probability sampled would have been a plausible first sample were it possible to draw directly from the distribution. Under the assumption that the chain has eventually reached the region(s) of high probability the samples that are after it has first reached this region are acceptable.

Convergence tests

MCMC algorithms could be run indefinitely – there is no universal stopping criterion. In most cases the errors on any estimates (*e.g.*, of means, covariances, *etc.*) decreases with the number of samples as ${N_\mathrm{s}}^{1/2}$, although the scaling of this convergence to the "truth" depends on the degree of correlation in the chain (and the nature of the target distribution).

One key idea is that multiple independent chains should converge on the same stationary distribution, and so any quantities derived from different chains should be consistent with being drawn from the same distribution. If multiple chains exhibit significant differences then it is ikely that the target distribution has not been sampled well; but passing the test is *not* a guarantee that the sampling has been effective. (Although this now has the feeling of a null test, a non-Bayesian concept that is somewhat out of place here.)

Gelman, A. and Rubin, D. B. (1992) devised a heuristic convergence test based on this principle that just uses the empirical means and variances of the marginal distributions of independent chains (that have been pruned of pre-burn-in elements). The test can be applied to each parameter, so without loss of generality a single-parameter model is considered.

1. Consider the N_c chains of (equal) length N_s with elements $\{\theta_{c,s}\}$ (where $c \in \{1, 2, \ldots, N_\mathrm{c}\}$ and $s \in \{1, 2, \ldots, N_\mathrm{s}\}$).

2. Calculate the mean of each chain as

$$\bar{\theta}_c = \frac{1}{N_\mathrm{s}} \sum_{s=1}^{N_\mathrm{s}} \theta_{c,s}.$$

3. Calculate the (empirical) variance of each chain as

$$\sigma_c^2 = \frac{1}{N_\mathrm{s} - 1} \sum_{s=1}^{N_\mathrm{s}} (\theta_{c,s} - \bar{\theta}_c)^2.$$

4. Calculate the (empirical) mean of all the chains (*i.e.*, the best estimate for the posterior mean of the distribution):

$$\bar{\theta} = \frac{1}{N_\mathrm{c}} \sum_{c=1}^{N_\mathrm{c}} \frac{1}{N_\mathrm{s}} \sum_{s=1}^{N_\mathrm{s}} \theta_{c,s} = \frac{1}{N_\mathrm{c}} \sum_{c=1}^{N_\mathrm{c}} \bar{\theta}_c.$$

5. Calculate the average of the individual chains' variances:

$$\sigma^2_\mathrm{chain} = \frac{1}{N_\mathrm{c}} \sum_{i=1}^{N_\mathrm{c}} \sigma_c^2.$$

6. Calculate the (empirical) variance of the chains' means:

$$\sigma^2_\mathrm{mean} = \frac{1}{N_\mathrm{c}} \sum_{c=1}^{N_\mathrm{c}} (\bar{\theta}_c - \bar{\theta})^2.$$

7. Calculate the ratio

$$\hat{R} = \frac{\frac{N_\mathrm{c}-1}{N_\mathrm{c}} \sigma^2_\mathrm{chain} + \frac{1}{N_\mathrm{c}} \sigma^2_\mathrm{mean}}{\sigma^2_\mathrm{chain}}.$$

8. If the chains are well mixed and have all sampled the target distribution then $\sigma^2_\mathrm{chain} \simeq \sigma^2_\mathrm{mean}$ and $\hat{R} \simeq 1$.

9. If the chains have sampled different parts of the target distribution then their individual variances will be less than the variance between the estimates of the chains and $\hat{R} > 1$.

10. The common heuristic approach is to regard the chains as converged if $\hat{R} \lesssim 1.2$.

Thinning

The unmodified outputs of all MCMC algorithms are a set of correlated samples of the posterior distribution and, moreover, the nature of the correlations are unknown a priori. One way of side-stepping this issue to "thin" the chain, using only a fraction of the samples. Information is lost in this process, but the gain in obtaining a (almost) uncorrelated set of samples is considerable, not least because of the decreased storage requirements.

The degree of thinning required to obtain a reasonably uncorrelated sample can be judged from an auto-correlation plot of a chain.

3.3.2 The Metropolis-Hastings algorithm

The Metropolis-Hastings (MH) algorithm is probably the most commonly used variety of MCMC, as it is simple, intuitive and (with some refinements) effective on a wide variety of problems. MH sampling algorithms are effectively a guided random walk through parameter space that preferentially sample from the high-probability regions while exploring the full range of possibilities.

Algorithm

The MH algorithm involves a two-step process for generating each step in the chain. Given a previous point $\boldsymbol{\theta}_{s-1}$, the next point is obtained by:

1. Draw a trial point from the proposal distribution $\Pr(\boldsymbol{\theta}_{\rm trial}|\boldsymbol{\theta}_{s-1})$, the form of which can be chosen to increase the efficiency of the algorithm.

2. Accept the trial point with probability

$$\Pr(\text{accept}|\boldsymbol{\theta}_{\rm trial},\boldsymbol{\theta}_{s-1}) = \min\left[\frac{\Pr(\boldsymbol{\theta}_{\rm trial}|\boldsymbol{d})}{\Pr(\boldsymbol{\theta}_{s-1}|\boldsymbol{d})},1\right] = \min\left\{e^{\ln[\Pr(\boldsymbol{\theta}_{\rm trial}|\boldsymbol{d})]-\ln[\Pr(\boldsymbol{\theta}_{s-1}|\boldsymbol{d})]},1\right\},$$

   which is unity if the trial point is more probable than the previous point, but is only zero if $\Pr(\boldsymbol{\theta}_{\rm trial}|\boldsymbol{d}) = 0$.

3. If the trial point is accepted then set $\boldsymbol{\theta}_s = \boldsymbol{\theta}_{\rm trial}$; otherwise set $\boldsymbol{\theta}_s = \boldsymbol{\theta}_{s-1}$.

This last step is critical to the algorithm: a MH chain must inevitably include sequences in which $\boldsymbol{\theta}_s = \boldsymbol{\theta}_{s+1} = \boldsymbol{\theta}_{s+2} = \ldots$; if it does not then the chain is almost certainly not sampling the target distribution correctly (as described below).

Choice of proposal distribution

The one aspect of the MH algorithm that can be tuned is the form of the proposal distribution. While most combinations of proposal and target distributions will result in full sampling in the limit of infinite samples, there is a clear motivation for the Markov chain to explore the target distribution as efficiently as possible. That the proposal distribution can be tuned to do this can be seen easily by considering two extreme cases.

If the proposal distribution is very concentrated relative to the scales on which the posterior varies then $\Pr(\boldsymbol{\theta}_{\rm prop}|\boldsymbol{d}) \simeq \Pr(\boldsymbol{\theta}|\boldsymbol{d})$ for all proposed points $\boldsymbol{\theta}_{\rm prop}$. The acceptance probability is then $\Pr(\text{accept}|\boldsymbol{\theta}_{\rm prop},\boldsymbol{\theta}) \simeq 1$, and almost all jumps are accepted but the distribution is explored very slowly. (An analogy would be someone trying to explore a mountain range by taking steps of a millimetre.)

The other extreme is that the proposal distribution is much broader than region of parameter space for which the posterior is appreciable. If $\boldsymbol{\theta}_{s-1}$ is in a region of high posterior density then the odds are that $\boldsymbol{\theta}_{\rm prop}$ will be outside the high posterior region and so $\Pr(\text{accept}|\boldsymbol{\theta}_{\rm prop},\boldsymbol{\theta}) \simeq 0$, and almost all jumps will be rejected. The resultant chain would just contain copies of the current position until a jump was finally accepted. The final chain would then contain very few independent points and the exploration of the target distribution would be very inefficient. (An analogy would be trying to explore a mountain range by taking steps of a light year.)

Given that proposal distributions can be too narrow or too broad, the implication is that there is a range of intermediate values that are well suited to the target distribution. The most efficient sampling occurs if the acceptance ratio is in the range from $\sim 20\%$ to $\sim 40\%$, although these figures are distribution-dependent. More to the point, these values are only concerned with the efficiency of the sampling scheme – even an acceptance ratio of a $\sim 5\%$ or $\sim 90\%$ will produce a useful set of samples run for long enough. (And a large set of highly-correlated samples can be made less unwieldy by thinning, as described below.)

The most common proposal distribution is a multi-variate normal, so that

$$\begin{aligned} \Pr(\boldsymbol{\theta}_{\rm trial}|\boldsymbol{\theta}_{s-1}) &= {\rm N}(\boldsymbol{\theta}_{\rm trial};\boldsymbol{\theta}_{s-1},\mathsf{C}) \\ &= \frac{1}{(2\pi|\mathsf{C}|)^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{\theta}_{\rm trial}-\boldsymbol{\theta}_{s-1})^{\rm T}\mathsf{C}^{-1}(\boldsymbol{\theta}_{\rm trial}-\boldsymbol{\theta}_{s-1})\right], \end{aligned} \tag{23}$$

where C is the covariance matrix. A sample from this distribution can be obtained by following steps:

1. Generate a N_p-dimensional vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_{N_p})$ with components drawn from $x_p \sim {\rm N}(0, 1^2)$.
2. Calculate $\boldsymbol{\theta}_{\rm trial} = \boldsymbol{\theta}_{s-1} + \mathsf{L}\boldsymbol{x}$, where L is any matrix such that $\mathsf{L}\mathsf{L}^{\rm T} = \mathsf{C}$, an obvious option being a Cholesky decomposition of C.

In the absence of any information about $\Pr(\boldsymbol{\theta}|\boldsymbol{d})$ there is no reason to prefer correlations between parameters, in which case the covariance matrix is diagonal, so that $C_{p,p'} = \delta_{p,p'}C_{p,p}$) and the trial point can be generated component by component by using $\theta_{{\rm trial},p} \sim {\rm N}(\theta_{{\rm trial},p}; \theta_{s-1,p}, C_{p,p})$, for $p = 1, 2, \ldots, N_p$.

One way of learning about the posterior distribution is by from the initial samples drawn; if the "guessed" proposal distribution is inefficient then the sampling will not be very effective, but will still provide some information about the range of the high-probability region in each variable (and about the correlations). A reasonable choice is then to use the empircal covariance matrix of the chain, calculated using Eq. (18) for the proposal distribution.

Even if the parameters are not strongly correlated, some information is required to set sensible scales for the "width" of the proposal in each of the coordinates. Particularly in the context of astronomical problems, in which some quantities have extremely high or low values, a proposal of ${\rm N}(0, 1^2)$ could be quite useless.

Stationary distribution

Consider a simple two-state problem in which only two parameter values, $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are possible . . .

3.3.3 Gibbs sampling

A very different form of MCMC is provided by Gibbs sampling, in which only a single parameter is explored in any one step. This is useful if (all) the conditional distributions of the target distribution are known and can be sampled from easily. As such, Gibbs sampling is particularly

useful for sampling distributions of high dimensionality in which most of the paramters are not directly linked, as is often the case for heirarchical models.

Algorithm

If all the conditional posterior distributions $\Pr(\theta_i|\theta_1, \theta_2, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_{N_p-1}, \theta_{N_p}, \boldsymbol{d})$ are known and can be sampled from efficiently (*e.g.*, because they are standard distributions), then Gibbs sampling is a very powerful method of exploring the full posterior distribution $\Pr(\boldsymbol{\theta}|\boldsymbol{d})$.

Given an initial position $\boldsymbol{\theta}_0 = (\theta_{1,1}, \theta_{2,1}, \ldots, \theta_{N_p,1})$, standard Gibbs sampling involves drawing from each conditional distribution successively, according to

$$\begin{aligned}
\theta_{1,2} &\sim \Pr(\theta_2|\theta_{2,1}, \theta_{3,1}, \ldots, \theta_{N_p-1,1}, \theta_{N_p,1}, \boldsymbol{d}),\\
\theta_{2,2} &\sim \Pr(\theta_2|\theta_{1,2}, \theta_{3,1}, \ldots, \theta_{N_p-1,1}, \theta_{N_p,1}, \boldsymbol{d}),\\
&\vdots\\
\theta_{N_p-1,2} &\sim \Pr(\theta_{N_p}|\theta_{1,2}, \theta_{2,2}, \ldots, \theta_{N_p-1,1}, \boldsymbol{d}),\\
\theta_{N_p,2} &\sim \Pr(\theta_{N_p}|\theta_{1,2}, \theta_{2,2}, \ldots, \theta_{N_p-1,2}, \theta_{N_p-1,2}, \boldsymbol{d}),
\end{aligned}$$

where, in each case, the latest value for all the parameters (other than that being sampled) is used. The sequence then repeats, with a new value of θ_1 obtained that is conditional on the updated values of all the other parameters.

Hence this sequence of draws yields N_p new samples

$$\begin{aligned}
\boldsymbol{\theta}_1 &= (\theta_{1,2}, \theta_{2,1}, \ldots, \theta_{N_p-1,1}, \theta_{N_p,1}),\\
\boldsymbol{\theta}_2 &= (\theta_{1,2}, \theta_{2,2}, \ldots, \theta_{N_p-1,1}, \theta_{N_p,1}),\\
&\vdots\\
\boldsymbol{\theta}_{N_p-1} &= (\theta_{1,2}, \theta_{2,1}, \ldots, \theta_{N_p-1,2}, \theta_{N_p,1}),\\
\boldsymbol{\theta}_{N_p} &= (\theta_{1,2}, \theta_{2,2}, \ldots, \theta_{N_p-1,2}, \theta_{N_p,2}),
\end{aligned}$$

where some care must be taken with the indexing. Clearly successive samples are correlated, and one option is to only retain the N_p'th sample; however $\boldsymbol{\theta}_{s+N_p}$ is *not* independent of $\boldsymbol{\theta}_s$, despite the fact that every parameter has been explored.

Markov properties

Gibbs sampling is clearly a Markov process, as $\boldsymbol{\theta}_s$ depends explicitly only on $\boldsymbol{\theta}_{s-1}$. In constrast with MH sampling, Gibbs sampling produces a chain with out repeated elements (except in the extreme case that one of the conditional distributions is a delta function).

Exhaustiveness

There are certain classes of distribution that Gibbs sampling will never fully explore, most obviously those with separate peaks that are not aligned in any of the sampling variables. Whereas the MH algorithm will, with a proposal distribution of broad support, eventually sample from an arbitrary distribution, Gibbs sampling has no equivalent of the proposal distribution that can be adjusted.

Stationarity

The stationary distribution of can be most easily seen in the simple case of a two-dimensional model for which $\boldsymbol{\theta} = (x, y)$. Assuming the s'th sample to be a "move" in the x-direction, it is distributed according to

$$x_s \sim \Pr(x_s, |y_{s-1}, \boldsymbol{d}) \tag{24}$$

and

$$y_s \sim \delta_{\rm D}(y_s - y_{s-1}). \tag{25}$$

But if (x_{s-1}, y_{s-1}) is itself already a draw from $\Pr(x, y|\boldsymbol{d})$ then the same will hold for the new sample. Hence Gibbs sampling will produce (correlated) draws from the target distribution.

Utility

A common situation in which this is the case is a heriarchical model, which can have a large number of parameters (*e.g.*, hundreds or more, making MH sampling hopelessly ineffecient), most of which are not directly linked to each other. In that case it is usually much easier to make draws from the simple(r) conditional distributions than from the joint distribution.

Question: *Consider N_p samples drawn from a normal distribution, $\theta_i \sim {\rm N}(\theta_i|\mu, \sigma^2)$ of unknown mean μ and unknown variance Σ^2 for which only noisy measurements $d_i \sim {\rm N}(d_i|\theta_i, \sigma^2)$ of known variance σ^2 are available. What is the sampling distribution of the data, $\Pr(\boldsymbol{d}|\boldsymbol{\theta}, \mu, \Sigma)$, expressed in terms of the above distributions? What is the joint posterior distribution $\Pr(\boldsymbol{\theta}, \mu, \Sigma|\boldsymbol{d})$? What is the form of the conditional distributions $\Pr(\theta_i|\theta_1, \theta_2, \ldots \theta_{i-1}, \theta_{i+1}, \ldots, \theta_{N_p}, \mu, \Sigma, \boldsymbol{d})$, $\Pr(\mu|\theta_1, \theta_2, \ldots, \theta_{N_p}, \Sigma, \boldsymbol{d})$ and $\Pr(\Sigma|\theta_1, \theta_2, \ldots, \theta_{N_p}, \mu, \boldsymbol{d})$? Describe a Gibbs sampling scheme to generate samples from the full posterior distribution.*

3.3.4 Metropolis-within-Gibbs sampling

It is generally possible to combine Metropolis(-Hastings) and Gibbs sampling if neither algorithm is itself suitable for the entire sampling problem.

[MORE HERE]

3.4 Nested sampling

Nested sampling Skilling (2004) is a relatively new algorithm that has become popular in astronomy and cosmology, in part due to the availability of he MultiNest code Feroz *et al.* (2009). Unlike MCMC methods the nested sampling produces independent samples; however they are not uniformly weighted, which is a something of a disadvantage (or at least an inconvenience).

[MORE HERE]

References

Feroz, F., Hobson, M. P., and Bridges, M. (2009). MULTINEST: an efficient and robust Bayesian inference tool for cosmology and particle physics. *Monthly Notices of the Royal Astronomical Society*, **398**, 1601–1614.

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

Skilling, J. (2004). Nested sampling. In *AIP Conference Proceedings of the 24th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, volume 735 of *Lecture Notes in Physics, Berlin Springer Verlag*, pages 395–405.