7.7 Markov-Chain Monte Carlo

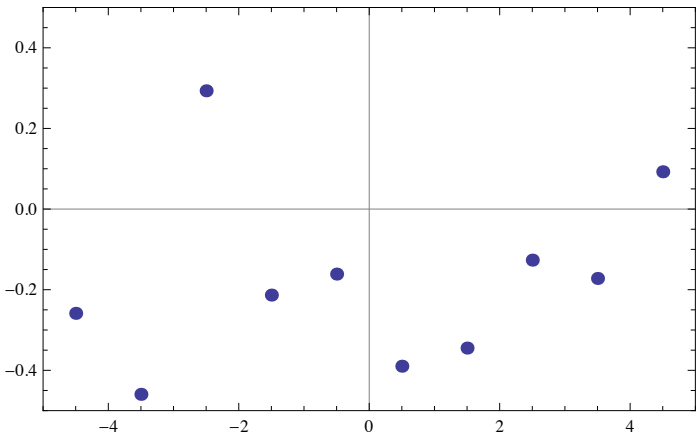

Figure 1: Simulated data

Extend the previous exercise to compute the actual posterior probabilities for three possible polynomials from a constant through to a quadratic. In this case the evidence will have to be calculated: use a MCMC method to compute the evidence by a thermodynamic integration.

Solution

The data are in the file for this example, and are plotted in Figure 1.

We will model these data on the assumption that they are normally distributed around the fitted model (which is a polynomial $P(a, x)$). The as are the polynomial coefficients and x is the independent variable. The likelihood function, for the data values Y_i at a series of abscissae X_i, is as usual

$$\ln \mathcal{L}(a) = -\frac{1}{2\sigma^2} \sum_i (Y_i - P(a, X_i))^2 + \text{constants}$$

Attributing a prior is a difficult problem for our vaguely-defined problem involving polynomials of unknown order. One possibility is the following.

For a problem like this one that is linear in the parameters a, the Hessian matrix

$$\frac{\partial^2 \ln \mathcal{L}}{\partial a_i \partial a_j}$$

depends only on the X_i, not the unknown as or the noise-affected Y_i. The inverse of the Hessian is a covariance matrix whose diagonal elements give the variance in the estimates of the coefficients a. For example, the inverse Hessian for the quadratic case (3 coefficients), for the current dataset, is

$$\begin{pmatrix} 0.00915625 & 0 & -0.000625 \\ 0 & 0.000484848 & 0 \\ -0.000625 & 0 & 0.0000757576 \end{pmatrix}.$$

The prior we will take on the coefficients will be Gaussian, with the standard deviation on each coefficient (denoted ξ) being twice the square root of the corresponding diagonal elements, above:

$$\{0.191377, 0.0440386, 0.0174078\}.$$

The same arguments apply to other orders of polynomial to get priors on the coefficients.

In general then the log of the posterior probability (likelihood $\times$ prior) is

$$\ln \mathcal{P} = -\frac{1}{2\sigma^2}\sum_i (Y_i - P(a, X_i))^2 - \frac{1}{2}\sum_j \frac{a_j^2}{2\xi_j^2} + \text{constants}.$$

The maximum of the posterior gives the estimate of the polynomial coefficients. We consider polynomials of order zero to 3. If a_1 is the constant term, a_2 the coefficient of the linear term, and so on, the maximum posterior estimates are

$$\begin{aligned} a_1 &= -0.139229 \\ a_1, a_2 &= -0.139229, 0.0118769 \\ a_1, a_2, a_3 &= -0.172144, 0.0118769, 0.00204946 \\ a_1, a_2, a_3, a_4 &= -0.172144, -0.0484979, 0.00204946, 0.00444675. \end{aligned}$$

The fits to the data for the various cases are in Figure 2.

To calculate the posterior probability of each of these models, we need to integrate the posterior function $\mathcal{P}$ over the parameters a, and we do this by building a MCMC chain on the posteriors $\mathcal{P}$. In fact, to do a thermodynamic integration, we need to build a series of chains built on the modified function $\mathcal{L}^\gamma \times$ prior where γ is the "inverse temperature".

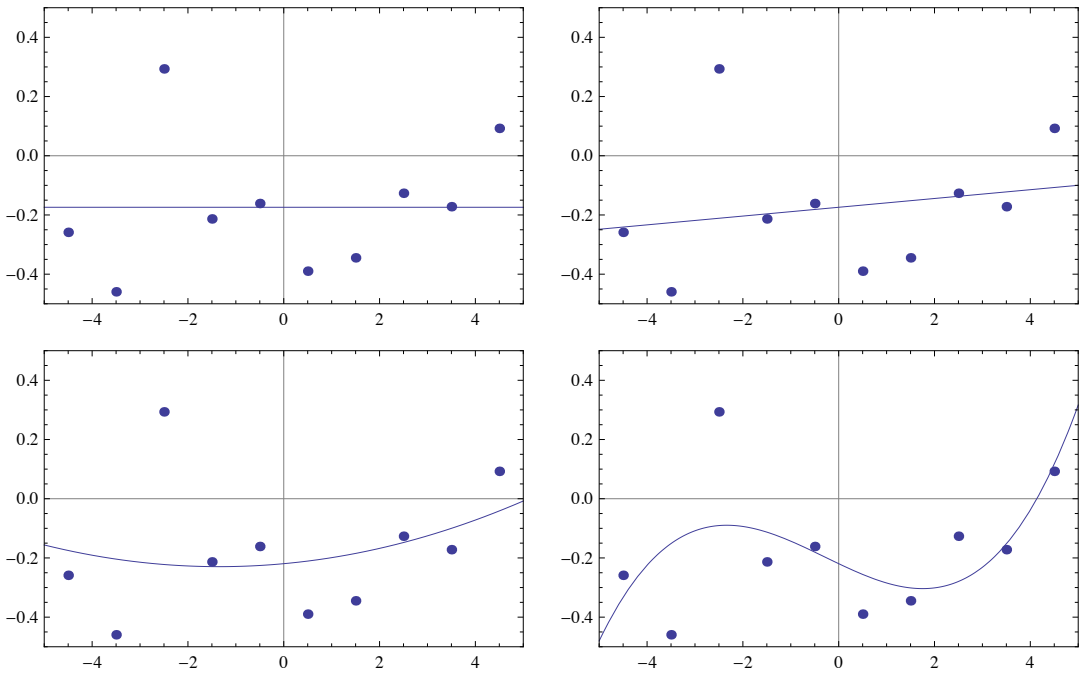

Figure 2: Fits to data

As noted in the previous example, because $\mathcal{P}$ is Gaussian we can do the integrations analytically by Laplace's method. However as an exercise we do this by MCMC and have a useful check from the analytical result.

We will assume incidentally that each of the models is a priori equally probable.

Constructing the chains and doing the integrations follows closely the example in the text, with a few changes.

We start creating the chains with the $\gamma = 0$ case. The starting point for the chain is the maximum posterior estimate for the as, perturbed randomly in each co-ordinate by the standard deviation in each coefficient, as given by the diagonal elements of the inverse Hessian (as above). The chains were 400000 repetitions long, and the final value of each chain was used as the starting point for the next value of γ. Chains were thinned to every 100th value so the chains used for the integrations were 4000 long.

The detail of the proposal distribution needs a little attention for this exercise. Recall that the proposal effects a jump in each coefficient, cycling through the set of coefficients (three for a quadratic fit and so on). Now the posterior is much narrower in some coefficients than others, so it makes sense to use the diagonals of the inverse Hessian (above) to define the size of the jumps in each co-ordinate. This greatly increases the number of accepted jumps.

Another improvement results from noticing that the inverse Hessian is not diagonal, so that jumps along co-ordinate axes are not along the axes of symmetry of the posterior distribution. Working in the principal axes of the inverse Hessian is the solution: the direction of the jumps is fixed by the eigenvectors of the inverse Hessian, and the size of the jumps is fixed by the associated eigenvalues. This is particularly useful when there are correlations between the estimates of the coefficients, so that the posterior distribution is elliptical and not aligned with the coefficient co-ordinate axes.

two quality checks have to be made on any string of random numbers generated by MCMC. One is that successive numbers are more or less independent; this is not vital for the numerical integrations we are doing, but it is a waste of computation to use numbers that are not independent. Checks of the power spectra show "white" spectra at $\gamma = 1$, although by $\gamma = 0$ the spectra are noticeably pink. This reflects the greater width of the target function as γ gets smaller – ideally one would adjust the proposal distribution to make it wider for smaller γ.

Of more importance is the requirement for "burn-in" to have occurred, as otherwise the target distribution is not properly sampled by the numbers in the chain. A simple check is to look at the standard deviation of the numbers

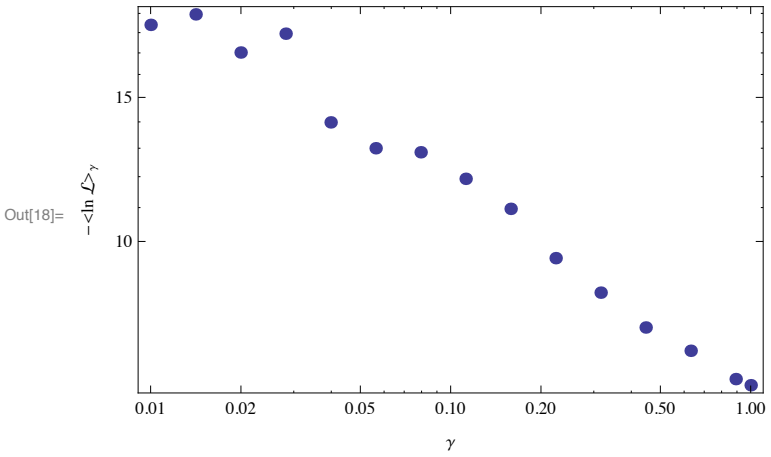

Figure 3: The expectation $\langle \ln \mathcal{L} \rangle_\gamma$ as a function of γ.

in the chain (say, 100 contiguous numbers) as a function of position in the chain. If this standard deviation is not evolving then either the chain is stuck (in which case the standard deviations will make no sense) or else burn-in has occurred. For the chains in this example burn-in seems to occur quickly, which is probably a result of having a well-tuned proposal distribution.

The values of γ to use need some attention. For our example (where the posterior is actually Gaussian) it can be shown that the trend of the key quantity (defined in the text) $\langle \ln \mathcal{L} \rangle_\gamma$ is $\sim 1/\gamma$, with $\langle \ln \mathcal{L} \rangle_0$ being finite. What this means is that much of the change in $\langle \ln \mathcal{L} \rangle_\gamma$ happens at small γ. To capture this, our calculations were made from $\gamma = 0.01$ with steps in γ of a factor of $\sqrt{2}$. Values for $\gamma = 0$ and $\gamma = 1$ were also computed.

Figure 3 gives an example for the quadratic (three coefficients) case.

For this example, a numerical integration of the data in the plot gives a value for the integral of the posterior, denoted $\mathcal{E}_3$, of 0.000162. This is the evidence for a quadratic polynomial model. A analytic integration of the posterior (using the Laplace approximation, exact in this case) shows that this is correct to within 3%.

Here are all the results.

$$\begin{aligned} \mathcal{E}_1 &= 0.000685 \\ \mathcal{E}_2 &= 0.000361 \\ \mathcal{E}_3 &= 0.000162 \end{aligned}$$

$$\mathcal{E}_4 = 0.000089.$$

It is worth trying to calculate the evidence for the cubic case by a brute-force numerical integration: it is surprisingly difficult to get the right answer, reflecting the "curse of dimensionality."

From these we can calculate immediately the probability α of the constant, linear, quadratic and cubic models:

$$\begin{aligned}\alpha_1 &= 0.526\\ \alpha_2 &= 0.277\\ \alpha_3 &= 0.127\\ \alpha_4 &= 0.070.\end{aligned}$$

The constant model is the most probable.

It's interesting to look at a classical model-fitting approach with χ^2. We can get the minimum χ^2 from each model by maximizing the likelihood (NB no prior here!). From this we get the standard p-value, the probability of the residuals from the fit arising by chance - given the model. The degree of the χ^2 distribution we need to compute this is the number of degrees of freedom, the number of data points - the number of parameters.

We get

$$\begin{aligned}\chi^2_1 &= 11.44\\ \chi^2_2 &= 10.99\\ \chi^2_3 &= 10.59\\ \chi^2_4 &= 7.56.\end{aligned}$$

The p-values take some account of the "over-fitting" allowed by having many parameters, via the degrees of freedom in the calculation.

$$\begin{aligned}p_1 &= 0.24\\ p_2 &= 0.20\\ p_3 &= 0.16\\ p_4 &= 0.27.\end{aligned}$$

If we were stuck in a classical world, we might pick the cubic model: the deviations of the data from the model are "more probable" than for other models. The pitfalls of this approach should be apparent by this stage of the book. For instance, the ps do not sum to unity (nor should they).