Small projects for Data-analysis and Inverse Methods in Astronomy, spring 2012.

Select and complete four of these six projects. For details and questions, contact Mika Juvela regarding project 1, Antti Penttilä with projects 2–4, and Karri Muinonen with projects 5–6. Return a short, informal report that shows what you have done, including the result, before 18.4. to any of the lecturers.

Sp. 1
Black body spectrum

Fit modified black body curves to simulated "observational" data. Assume the observations are made at wavelengths 160, 250, 350, and 500 µm and the observed spectrum is the modified black body spectrum with $\beta = 2.0$ and temperatures T in the range 10–20 K.

Step one: the flux densities F observed from a single source are fitted with a modified black body function $F = F_0\, B(T)\, (f/f_0)^{\beta}$ with the flux density F_0, the temperature T, and the spectral index beta as free parameters. Simulate input data with $T = 15K$ and $\beta = 2.0$ and check the dependence of the fitted T and beta values on the noise added to the input spectra.

Step two: the parameters (T, β) obtained for different objects are examined for the presence of a negative correlation between the parameters T and β. Use Monte Carlo simulations to examine the effect of noise on this relation. Take a sample of sources with different T but a fixed value of $\beta = 2.0$. Check how the $(T, beta)$ values derived from synthetic observations are distributed. Can you define a test for the hypothesis that $\beta(T)$ is a decreasing function of temperature?

Sp. 2
Maximum likelihood or maximum a posteriori parameter estimation

The posterior distribution of the parameter under interest has all the (inverted) information we can get from the measurements. However, we often need to make a point-estimate of the parameter, i.e. report a single, most probable value. In classical statistics this is called the maximum likelihood estimate and is found by maximizing the likelihood function with respect to the unknown parameter. In Bayesian approach the same thing is done by founding the maximum value of the posterior distribution for the unknown parameter. In addition to this, an error estimate for the point-estimate is usually useful. This error estimate is called the confidence interval in classical statistics, or Bayesian interval in Bayesian case. In the latter case the interval is found by searching the values x_{lower}, x_{upper} so that

$$\int_{x_{lower}}^{x_{upper}} D(x|m)\, dx \;=\; 1 - \alpha \tag{1}$$

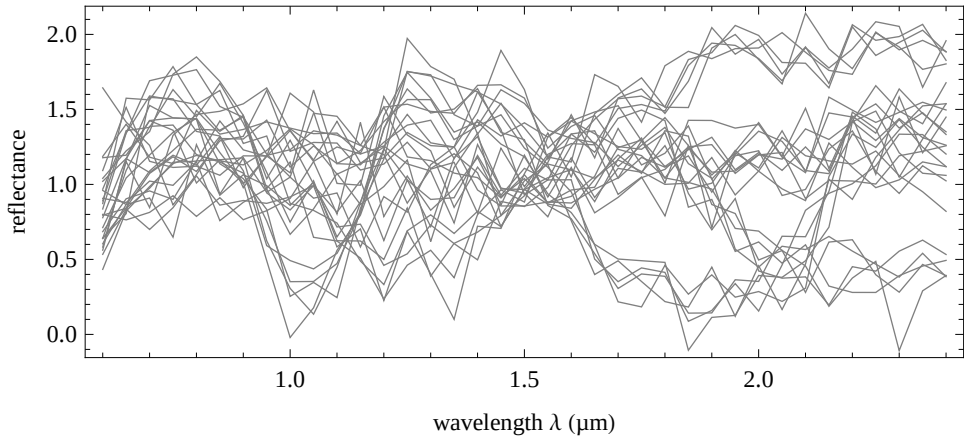
Now α is the confidence (or probability) level for the interval. With probability $1 - \alpha$ the real, unknown value of x is between x_{lower} and x_{upper}. The Eq. 1 does not define the limiting values completely; we still need to define if the interval is *equal tail* or *highest posterior density* (HPD). With equal tail interval both the tails $\int_{-\infty}^{x_{lower}} D(x|m)\, dx$ and $\int_{x_{upper}}^{\infty} D(x|m)\, dx$ have the same probability $\alpha/2$. With HPD the values of the posterior inside (x_{lower}, x_{upper}) must always be greater or equal to the values outside the interval.

a) You have 10 i.i.d. measurements from true/false experiment, let's say $m = (0, 1, 1, 0, 0, 0, 1, 0, 0, 0)$. The distribution for m_i is Bernoulli. Show that the likelihood of m is binomial. Your prior distribution for the probability of success π is Beta. Show that Beta is conjugate prior to binomial.

b) ...let's say that you have reason to believe that the success probability π is 0.5, so the mean of the prior Beta(α, β) should be $\alpha/(\alpha + \beta) = 0.5$. This means that there is actually only one hyperparameter because $\beta = \alpha$. The variance of Beta(α, α), which is $1/(4 + 8\alpha)$, should describe the uncertainty in your prior knowledge. Try two values for prior variance, 0.01 and 0.055. Solve the value for α with these variances. Plot the prior distributions and the posteriors with both the prior variances. If posteriors are without normalization they probably have different scales - do not try to put then in same figure. Find the point-estimates to the success probability π and the 5% equal tail intervals (numerically).

Sp. 3
Asteroid spectra and Principle Component Analysis

Datafile `spectra.dat` has 26 rows — first row has the wavelengths from 0.6 µm to 2.4 µm, and the rest are reflectancies of 25 "asteroids" at those wavelengths (plotted below).



Use principle component analysis to find underlying five groups of asteroids based on their spectra (see e.g. lecture notes from Juvela or Wikipedia on PCA). The first three principle components should be enough to distinct the groups. Plot the observations in the new PCA variable space, e.g. PCA-1 against PCA-2 and PCA-1 against PCA-3.
If you succeed, you should notice that the first 5 rows of data belong to the first group, the second 5 rows to the second group etc. The spectra of those groups correspond to common minerals in asteroids — nickel-iron (obs. 1–5), olivine (obs. 6–10), orthopyroxene (obs. 11–15), plagioclase feldspar (obs. 16–20) and spinel-bearing Allende inclusion (obs. 21–25) (from Binzel et al. (eds.) Asteroids II).

Sp. 4
Gibbs sampling

See lecture slides on Gibbs sampling and do the "Computer task" mentioned there.

Sp. 5

Assume that the observations of two nearby spectral lines are analyzed using a model $f(x)$ (x can denote the wavelength) composed of two Gaussian line profiles and a linear background with altogether eight parameters (or unknowns) a_1, x_1, σ_1, a_2, x_2, σ_2, b, and k:

$$f(x) \;=\; a_1 \exp\left[-\frac{(x-x_1)^2}{2\sigma_1^2}\right] + a_2 \exp\left[-\frac{(x-x_2)^2}{2\sigma_2^2}\right] + b + kx.$$

Given the observational data in the attached file `obs.dat` and assuming Gaussian observational errors with standard deviation of 0.04 (refine this estimate if necessary), solve the statistical inverse problem for the parameters a_1, x_1, σ_1, a_2, x_2, σ_2, b, and k by writing a Markov-Chain Monte-Carlo (MCMC) computer program for the problem. Analyze the result using single-parameter and two-parameter marginal probability densities allowed by the MCMC sampling.

Sp. 6

Consider the monovariate Gaussian probability density function (p.d.f.) $p(x) = n(x;\mu,\sigma)$ with mean μ and standard deviation σ. Let us now assume that $n(x;\mu,\sigma)$ is the a posteriori p.d.f. of a fictitious inverse problem for the unknown x. Let us further assume that a pre-study of the inverse problem gives the following p.d.f. only roughly approximating the actual a posteriori p.d.f.:

$$p(x) \;=\; \begin{cases} \frac{x+\sigma}{8\sigma^2}, & -\sigma \le x \le 3\sigma, \\ 0, & x < -\sigma \;\text{ or }\; x > 3\sigma. \end{cases}$$

Write a computer program that produces a sample of x-values based on applying the MCMC virtual-observation method (see lecture notes) to the monovariate Gaussian p.d.f. by using the approximating p.d.f. above in the generation of proposals. In the virtual-observation method, two values of x are drawn randomly from the approximating p.d.f. and their difference is utilized as the proposal. Utilize the random-walk Metropolis-Hastings algorithm in the MCMC sampling. For $\mu = 0$ and $\sigma = 1$, compare the MCMC and conventional sampling of the Gaussian p.d.f. For example, in the limit of large numbers of samples, do the methods result in converging values for μ and σ? Finally, utilize the approximating p.d.f. directly as the proposal p.d.f. in a random-walk Metropolis-Hastings algorithm and compare the results to those from conventional sampling and sampling with the virtual-observation method.