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

For details and questions, contact Mika Juvela regarding exercises 1–9, Antti Penttilä with ex. 10–12, and Karri Muinonen with ex. 13–15. Return your answers before 18.4. to any of the lecturers.

Ex. 1
Generate a random sample of 100 Poisson distributed random numbers. Calculate the four first moments of the sample distribution.

Ex. 2
Take the data set `outliers.txt` and calculate an estimate for the expectation value. Compare the results from the normal mean and some trimmed mean and Winsorized mean.

Ex. 3
Fit a line to a two-dimensional data set (x, y) of your choice. Compare the results of weighted fits (you need data with error estimates for y) and unweighted fits.

Ex. 4
As a continuation of the previous exercise, try some robust fitting technique (e.g. MAD).

Ex. 5
The observations consist of x values $0.0, 1.0, 2.0$ and the corresponding y values of $0.0, 2.0, 1.0$. Calculate the linear correlation coefficient. Is the correlation statistically significant?

Ex. 6
Carry out a least squares fit of the data set `nonlinear.txt`. Assume the function is $f(x) = a + bx + c\sin(x)$ and determine the parameter values a, b, and c and their uncertainties.

Ex. 7
Take a multivariate data set with four or more variables. Use two methods to visualise the data. Present figures and list some pros and cons regarding the use of these methods for these data.

Ex. 8
Find a routine that takes advantage of the error estimates in both variables (e.g., total least squares) and give an example where the difference to the normal weighted least squares becomes significant. If you cannot find software for this, describe what total least squares does and how the results are expected to differ in some sample cases.

Ex. 9
Calculate with Monte Carlo simulation the uncertainties of the value of the Planck function when the temperature is $20 \pm 2K$ (all other parameters are assumed to be free of errors). How does the situation change when the temperature is $10 \pm 2K$.

Background for Ex. 10–11

The likelihood function

In statistical modeling the measurements m_i that are realizations of a random variable M obey some distribution P. The density of P is a function of the measurements and the parameter(s) x. In Bayesian approach also the parameter is a random variable X and has a prior distribution P_{pr}. The posterior density is then given by

$$D(x|M=m) \propto D_{pr}(x)\, D(m|X=x)$$

Usually we have n measurements instead of just one measurement, so the measurement distribution is a joint distribution of all the M_i's. If all the M_i's are identically and independently distributed (*i.i.d.*) the joint distribution is the product of n distribution P:

$$D(m|X=x) = D(m_1|X=x) \times \ldots \times D(m_n|X=x)$$

The likelihood function L is simply the joint distribution $D(m|X=x)$ but without the normalizing constant (that can be any function of measurements $c := c(m)$), so $D(m|X=x) = c(m)L(x;m) \propto L(x;m)$. Because the likelihood function can be somewhat simpler than D the inference is often done with it:

$$D(x|M=m) \propto D_{pr}(x)\, L(x;m)$$

For example, in the case of n i.i.d. exponential-distributed measurements the likelihood function is

$$L(\lambda;m) = \prod_{i=1}^{n} \lambda \exp(-m_i\lambda) = \lambda^n \exp(-\lambda \sum_{i=1}^{n} m_i) = \lambda^n \exp(-\lambda n \overline{m})$$

Ex. 10
Derive the likelihood function for n i.i.d. Poisson-distributed measurements.

Ex. 11
Derive the likelihood function for n i.i.d. Gaussian-distributed measurements with unknown μ but known σ^2. Try to simplify the likelihood function as much as possible by gathering all terms dependent only on the measurements to constant $c(m)$ and leaving them out of the $L(\mu;m)$.

Background for Ex. 12

The concept of Conjugate prior

A Conjugate prior distribution $D_{pr}(x)$ for measurement (error) distribution $D(m|x)$ or equivalently likelihood function $L(x;m)$ is such that the posterior distribution $D(x|m)$ is of the same distribution family as the prior. For example if measurements m follow exponential distribution and the prior for parameter X is Gamma distribution, then the posterior distribution is also a Gamma distribution.

Ex. 12
The example above: show that when i.i.d. measurements m_i follow exponential distribution and the prior for λ is Gamma with hyperparameters $(a, a/b)$ the posterior is also Gamma. (The parameters for prior distribution are often called hyperparameters.) Use the likelihood function from the first example of exponential measurements. Note that you only need to show that the posterior is proportional to something which is essentially a Gamma distribution, maybe without the normalizing

constant. Normalizing constant can also be a function of hyperparameters. What are the posterior parameters of the resulting Gamma distribution?

Ex. 13

Consider the ill-posed inverse problem of ”fitting” $y = kx + b$ to a single data point y_1 measured at $x = x_1$. Devise and implement a Markov-Chain Monte-Carlo method for the unknowns k and b and discuss the results.

Ex. 14

Assume that the k random variables $X = (x_1, \dots, x_k)^T$ obey a k-dimensional Gaussian probability density function with the means $\mu = (\mu_1, \dots, \mu_k)^T$ and $k \times k$ covariance matrix Σ_X.
a) Derive the means and covariance matrices for the marginal probability density functions.
b) Derive the covariance matrix Σ_F for the n-vector $F = DX$, where D is the $n \times k$ coefficient matrix. This is sometimes called the law of error propagation.

Ex. 15

Assume that x and y are Gaussian random variables with the corresponding means of μ_x and μ_y and standard deviations of σ_x and σ_y, as well as the covariance Σ_{xy}. Based on linearized propagation of errors, what would be the mean (μ_z) and standard deviation (σ_z) of the random variable $z = x/y$? Analyze the linearized propagation of errors in this case. If the random variables x, y, and z represent physical quantities, what is wrong? Can you find a cure for the problem? Can you identify a light-scattering-related question where the present problem can be important?

Different distributions:

$$\begin{aligned}
\text{Exponential } D(X = \lambda|m_i) &= \lambda \exp(-m_i\lambda) \\
\text{Poisson } D(X = \mu|m_i) &= \frac{\mu^{m_i}}{m_i!} \exp(-\mu) \\
\text{Gaussian } D(X = \mu|m_i) &= \frac{1}{\sqrt{2\sigma^2}} \exp(-\frac{(m_i - \mu)^2}{2\sigma^2}) \\
\text{Gamma } D(X = (\alpha, \beta)|y) &= \beta^{-\alpha}\, y^{\alpha-1} \exp(-y\,\beta) \,/\, \Gamma(\alpha) \\
\text{Bernoulli } D(X = \pi|m_i) &= \pi^{m_i}\, (1 - \pi)^{1-m_i} \\
\text{Binomial } D(X = \pi|k = \sum_{i=1}^{n} m_i) &= \frac{n!}{k!(n-k)!}\, \pi^k\, (1 - \pi)^{n-k} \\
\text{Beta } D(X = (\alpha, \beta)|\pi) &= \pi^{\alpha-1}\, (1 - \pi)^{\beta-1} \,/\, B(\alpha, \beta)
\end{aligned}$$