Astronomy 8824: Problem Set 5
Due Tuesday, November 14

Forecasting

You will need to refer to results in Statistics Notes 4.

1. Fisher matrix forecast, linear fit

Suppose you have 20 (x, y) data points generated from a linear relation $y = \theta_1 + \theta_2 x$ with x uniformly distributed in the range $5 < x < 20$ and independent Gaussian errors on y with standard deviation $\sigma = 1$.

What is the Fisher matrix

$$F_{ij} = -\left\langle \frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j} \right\rangle \ ?$$

What is the inverse Fisher matrix F_{ij}^{-1}?

If both the intercept θ_1 and slope θ_2 are to be estimated by fitting the data, what are the expected errors on θ_1 and θ_2?

If the slope θ_2 is known and only the intercept must be estimated from the data, what is its expected error?

How does the expected slope error change if $N = 6$ instead of $N = 20$? How does it change if x runs from 5 to 15 instead of 5 to 20?

Hint: The variance of a uniform distribution is $\langle x^2 \rangle - \langle x \rangle^2 = (x_{\rm max} - x_{\rm min})^2/12$.

2. MCMC for parameters of a linear fit

From the course web page, download the data files
`line.n20.s12.dat`
`line.n20.s17.dat`
`line.n20.s0.dat`
`line.n6.s0.dat`.
These files have (x, y, σ) data points, with $\sigma = 1$ in all cases and x evenly spaced in the range $5 - 20$. They were generated by the program `linedata.py`, which you should look at to check that you understand what it is doing. For the two files labeled '`s0.dat`' the points have been forced to lie exactly on the prescribed line.

(a) For the first two files, compute the best-fit slope and intercept using the formulas we have discussed in class (or given in Numerical Recipes).

(b) Also download the program `line_mcmc.py`, which reads a data file in this format and generates an MCMC for the intercept and slope (θ_1, θ_2) of a linear fit. You can either use this code or refer to it and write your own. Note that the probability is proportional to $\exp(-\chi^2/2)$, and you don't have to compute the constant of proportionality because you only need ratios of probabilities for your MCMC.

Generate an MCMC chain for each of the first 3 files (with $N = 20$). Plot the distribution of (θ_1, θ_2) from your MCMC chain. Compare the marginal distributions of θ_1 and θ_2 to Gaussian distributions with the errors you predicted from Part 1. My plotting routine for this problem is available on the web page as `sm.linemc`.

Compare the most probable elements in your MCMC chains to your results from (a).

(c) Plot instead the distribution of $(\theta_1 + 12.5\theta_2, \theta_2)$. Comment on the result. Relate your interpretation of this result to the Fisher matrix (think particularly about the moments that enter there).

(d) Repeat part (b) for the `line.n6.s0.dat` file.

3. A third parameter

Suppose that with the same ($N = 20$) data you allow a third parameter with a quadratic term, $y = \theta_1 + \theta_2 x + \theta_3 x^2$. For the fiducial model being assumed for the forecast, you adopt $\theta_3 = 0$, but you allow it to be free in the fit.

What is the Fisher matrix in this case? (You can do the matrix inversion numerically.) What are the forecast errors on θ_1, θ_2, and θ_3?

Modify your MCMC code to create a chain for this 3-parameter model. Apply it to the files `line.n20.s0.dat` and `line.n6.s0.dat` and plot the results, with particular attention to θ_3 vs. θ_2. Compare the errors from MCMC to your Fisher matrix forecast.

4. Correlated errors

Download the code `linepluscov.py`, which generates a distribution of points with correlated errors. Run it for 20 points in the range $x = 5 - 20$ with a slope $\theta_2 = 0.5$ and intercept $\theta_1 = 2$ for the random number seeds 12 and 17 used previously for the diagonal case. For this problem we are changing the slope from $\theta_2 = 2$ to $\theta_2 = 0.5$ while keeping $\sigma = 1$. This shrinks the vertical scale relative to the error bar, making the effect of correlations easier to see.

What covariance matrices are being used for the five sets of data points (A,B,C,D,E) that the code produces? (Look at the code to figure out what it's doing, and insert `print cov1` statements if needed.)

Plot the two realizations of $N = 20$ points for each of the five cases, attaching error bars and including the $y = 0.5x + 2$ line for reference. You can use my code `sm.linepluscov` for this purpose if you wish.

Comment on how correlated errors affect the distribution of points. Is the impact similar to what you expect?

OPTIONAL

5. Parameter errors with correlated data errors

Compute the predicted errors on θ_1 and θ_2 for each of the cases from Part 4. You'll now need to compute the Fisher matrix using the expression with the full covariance matrix (Stats Notes 4, page 3) and invert it numerically.

How do correlated data errors affect the expected parameter errors? Does this behavior make sense?

For cases D and E, check your Fisher matrix forecast against MCMC error estimation, as you did in Part 2.