Data Analysis Project 3

Due Sunday, 6 May, at 5 pm. Turn in the code for all of your work. Note that since this homework serves in place of a final exam, it will be weighted higher than regular homework. As usual, you are encouraged to collaborate, as long as the work that you submit is your own.

Bayesian Analysis of Source Flux in a Low S/N Millimeter Survey

In this final homework you will use your knowledge of Maximum Likelihood (ML) estimators, Baye's theorem, and Monte Carlo methods to perform a Bayesian analysis of source flux in a low signal-to-noise millimeter survey data.

This analysis follows that outlined in Scott et al., 2008, MNRAS, 385, 2225, "AzTEC millimetre survey of the COSMOS field: I. Data reduction and source catalogue." AzTEC is a 1.1 mm 144 pixel bolometer array camera. Survey data for this paper were taken with the camera mounted on the James Clerk Maxwell Telescope (JCMT) on Mauna Kea. The millimeter-wavelength survey data have revealed 50 millimeter sources in the region of an X-ray detected cluster and larger-scale overdensity in the COSMOS field, a field which has been surveyed at multiple wavelengths. The sources are thought to be luminous high-redshift dust obscured galaxies with high star formation rates. You may read the entire Scott et al. paper to familiarize yourself with this survey and analysis techniques. At very least, read Sections 4 & 5 on the source catalog Bayesian analysis. You may also choose to read Wall & Jenkins, Ch. 8 on "Detection and Surveys" to get a more general background of selection effects in surveys and the Schechter function.

Because the survey is low s/n ($< 10\sigma$ for all detected sources), simple estimates of the source fluxes from the map are systematically biased towards higher flux. This is due to the fact that the galaxy luminosity distribution follows a Schechter function, with many more faint sources than bright ones. Therefore, a low s/n source detection is more likely to be a faint source on top of a (randomly) positive noise fluctuation in the map rather than a bright source on top of a (randomly) negative noise fluctuation in the map. This effect is closely related to the more well known Eddington bias.

Your job is to estimate the flux, with 68% confidence interval, for one of the sources in this survey, using Bayesian analysis that uses a non-uniform prior generated from the expected galaxy luminosity function. The paper refers to the flux estimate using this posterior flux distribution (PFD) as "flux de-boosting."

Data may downloaded from the `Homework/Data` folder on Canvas.

This folder contains the files `aztec_sim_data.sav` for IDL users, and `aztec_sim_data.mat` for Matlab users. Python users can read the matlab `.mat` file using the scipy.io.loadmat method. I.e.,

```
>>> import scipy.io
>>> data = scipy.io.loadmat('aztec_sim_data.mat')
```

When you load the file into Python/IDL/Matlab, you will have the following arrays in your workspace (for Python, the `data` variable above is a Python dictionary that has key/value pairs below):

Variable Name	Dimensions	Units	Description
`coadded_signal_map`	1811×1436	Jy/beam	Simulated survey map
`coadded_weight_map`	1811×1436	$(\mathrm{Jy/beam})^{-2}$	Weight map, or estimated inverse noise variance map
`kernel_mean`	1810 x 1436	1/beam	Point Source kernel (PSF) for the instrument/reduction
`phys_ra_map`	1811	deg	Map RA coordinate vector
`phys_dec_map`	1436	deg	Map DEC coordinate vector

These simulated data are courtesy of Jay Austermann and the AzTEC team. They are identical to the data in the Scott et al. paper, except that the map contains a simulated survey map instead of the real survey map, since the analysis of the real data is still in progress and the data has not yet been released to the public.

The `coadded_signal_map` is a map created by coadding timestream data from many hours of scanning the COSMOS field. The map is Wiener (optimal or match) filtered to maximize the source s/n.

The `coadded_weight_map` is a map of the estimated inverse noise variance in each beam, σ_m^{-2}. The noise is Gaussian random in nature.

The `kernel_mean` is a map of the point source kernel, or point source function (PSF) for the telescope, instrument, and data reduction filtering process used to make the coadded_signal_map. It is the effective beam pattern, and therefore resolution, for the map, with a peak value of unity.

1) Download the data, and plot images of the `coadded_signal_map`, `coadded_weight_map`, and `kernel_mean`, in RA/DEC coordinates. You will see that the map is mostly noise, with fluctuations smoothed by the angular size of the point source kernel. In all plots and analysis that follows, use only the low-noise central region of the map, where the weight exceeds $(1.4\ \mathrm{mJy/beam})^{-2}$, i.e., with RMS noise σ_m <1.4 mJy/beam, as is done in the Scott et al. paper.

2) Create a s/n map by multiplying the `coadded_signal_map` by the sqrt of the `coadded_weight_map`. Given the (truncated low-noise) map area and the point source kernel main beam area, what is a statistically significant s/n, i.e., what is the s/n above which level you expect to see less than one such fluctuation due to Gaussian random noise? Plot the s/n map, with a color scale appropriate to show sources with statistically significant s/n.

3) Plot a histogram the fluxes from the low-noise region of the `coadded_signal_map`, and overplot a suitably normalized Gaussian PDF with the mean noise RMS $\langle\sigma_m\rangle$ of the low-noise map region. You should see an excess of non-Gaussian high pixel fluxes due to the presence of sources in the map. Compare your plot to Fig. 5 in the Scott et al. paper.

4) Select a source with a s/n$\approx$ 4 in the low-noise region of the map. Report a simple estimate of the measured source flux and uncertainty, $S_m \pm \sigma_m$ (Jy/beam), using the pixel values from the `coadded_signal_map` and `coadded_weight_map`, respectively.

5) Now, perform a Baysian analysis to estimate the source flux and 68% confidence interval, in order to take into account the systematic bias described in the introduction. To do this, calculate the posterior flux distribution (PFD), using Baye's theorem,

$$p(S_i|S_m,\sigma_m) = \frac{p(S_i)p(S_m,\sigma_m|S_i)}{p(S_m,\sigma_m)}$$

where $p(S_i)$ is the prior distribution of flux densities, $p(S_m,\sigma_m)$ is a normalization constant, and

$$p(S_m,\sigma_m|S_i) = \frac{1}{\sqrt{2\pi\sigma_m^2}}\exp\left(\frac{-(S_m-S_i)^2}{2\sigma_m^2}\right)$$

is the likelihood of observing the data S_m given an intrinsic source flux S_i, assuming Gaussian distributed noise with variance σ_m^2.

The hard part is calculating the prior, $p(S_i)$. To do this, first assume the intrinsic point source flux distribution follows a Schechter luminosity function,

$$\frac{dN}{dS} = N'\left(\frac{S}{S'}\right)^{\alpha+1}\exp(-S/S')$$

where $N(S)$ is the number of sources with flux less than S, and the parameters are $N' = 3200\,\mathrm{deg}^{-2}\ \mathrm{mJy}^{-1}$, $S' = 1.6$ mJy, and $\alpha = -2.0$ (adopted from 850 μm survey counts, with an assumed frequency scaling). To generate the prior, you need to generate a simulated noiseless astronomical sky map by populating an empty map with delta functions equal to source fluxes randomly selected from the assumed Schechter function, and positioned randomly in the map. Then convolve the map with the source kernel, `kernel_mean`. The total number of sources you should inject into the map is a Poisson deviate of the integrated source counts above 0.1mJy for the assumed Schechter function. Once you have the simulated signal-only map, make a histogram of the pixel values, as you did in step 3). This histogram is an extimate of the prior $p(S_i)$.

In more detail, the steps for generating the prior are as follows:

a) Write a procedure/function to calculate the luminosity function $N(S)$ and cumulative distribution function $F(S) = N(S)/N(\infty)$ using the Schechter function above.

b) Follow the procedure described in Wall & Jenkins section 2.6 to create a random flux generator following a Schechter probability distribution.

c) Use your function $N(S)$ to find the expected number of sources in the map. Then, use a Poisson random number generator (your own or a function provided in Python/IDL/Matlab) to randomly generate a number of simulated sources in your map, N_s.

d) For each of the N_s simulated sources, add a delta function to your initially empty map, with a uniformly random position, and a random flux amplitude drawn from the Shechter distribution.

e) Convolve your random delta function map with the point source kernel, `kernel_map`. You may choose to make a small kernel array of say 200×200 pixels around the peak value of the kernel, to make the 2D convolution faster. Use a canned 2D convolution routine, rather than writing your own.

f) Make a histogram of the resulting signal-only simulated map, similar to your histogram in step 3. This histogram is an estimate of your prior $p(S_i)$. You may take the mean histogram of several simulated maps to reduce the noise in the estimation. (In the paper, they generated 10,000 simulated maps to do this.)

6) Finally, plot the prior $p(S_i)$, and overplot the Gaussian likelihood distribution $p(S_m, \sigma_m|S_i)$, and the posterior flux distribution $p(S_i|S_m, \sigma_m)$ for your selected s/n≈ 4 point source from step 4), and compare to the Scott et al. paper, Fig. 6. Estimate the intrinsic flux and 68% confidence interval for the point source flux using the peak value and integrated constant-likelihood 68% probability region of the posterior flux distribution $p(S_i|S_m, \sigma_m)$. Your Bayesian analysis should result in a lower estimate for the flux than your original estimate in step 4). The Scott et al. paper calls this Bayesian flux extimate "flux de-boosting."

7) Briefly discuss how the systematic flux bias that you have characterized has implications for survey detection limits, false detection rates, and completeness.