INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, ALLAHABAD

C2 Finite Time Assessment (FTA), April 19, 2022. 11:00AM – 2 PM.

Computational Astrophysics

B.Tech: (IT & ECE) Elective VI – Semester *QP: PC*

Full Marks - PART A: {5} + PART B: {13} + PART C: {11} + PART D: {11} + = 40 (Scaled to 10)

Time - 1hr, Max Time - 3hrs.

CALCULATORS/COMPUTER ALLOWED

Answers should be brief and to the point.

Unnecessary Extra writing will attract negative marks.

This FTA will be of **10 marks** out of **30**. The remaining **20 marks** will be for the online assignments that have been submitted. This ratio of 10:20 may be varied in favour of the performance by the students so that the performance of the students could be maximised.

On the Top margin of each paper Students should write their

***i) Question No, ii) Roll No, iii) Name and iv) Signature*.**

These pages should be scanned and uploaded. Please install Adobe Scan to take the pictures of your answer pages for uploading it (Preferably in PDF). Do not Scan at high resolution so that the file size is large and it becomes difficult to upload it from your end. ***Please UPLOAD each PART (A,B, C, D) Separately as 4 PDF files.***

Do not share your login and password of your IITA e-mail. Any Malpractice of uploading through a single IP no, Uploading someone else's answer Sheet IS A CRIME. THE STUDENT will automatically fail the course.

These are difficult times. Your sincerity towards learning and ethical practice is expected from all of you.

PART-A: 5 Marks.
PART-B: 13 Marks.
PART-C: 11 Marks.
PART-D: 11 Marks.
TOTAL: 40 Marks *will be scaled to 10 Marks*

PART-A Photometric Extraction of Star Flux

1. The CCD data from an Optical Telescope is first corrected of systematic errors such as BIAS, DARK, FLAT correction. This data contains counts from the Celestial Objects (Stars: point sources, Galaxies, Nebulas: extended sources) and a background counts which is denoted as Sky. What is the cause of this Sky counts?
2. In the Techniques of extraction of Star fluxes, discussed in the class, what is assumed about the Sky flux counts? Enumerate at least 2 Assumptions.
3. How is the sky value computed?
4. How are the pixels containing the Star flux extracted from the images?
5. What is the formula of MODE used in this analysis and how is it different than the standard statistical formulation?

[Marks: $5\times1.0 = 5.0$].

Answer should be at least 75% correct to obtain 1 otherwise the marking will be 0.

1. The cause of this Sky count is
   a) scattering from bright objects such as the Moon, artificial lighting of the earth scattered on the sky (Light Pollution).
   b) Unresolved celestial objects in the vicinity.

2. The assumed about the Sky flux counts:
   i) The Sky value of the entire image is the same.
   ii) Most of the pixels in the image is the sky.

3. The sky value is computed by taking the MOD of all the pixels of the image. The sky value is improved by taking 2 / 3σ rejections using standard deviation σ. This is iterated multiple times until the variation in the sky value from one iteration to another does not significantly change.

4. The pixels containing the Star flux extracted from the images by assuming all the positive rejected pixels for the sky value are the star pixels.

5. The Formula for Mode used:
   We referred to the paper by E. Bertin and S. Arnouts (1996) "***SExtractor: Software for source extraction***" A&A Supplement Series **Vol. 117**, page 393-404.

   As used in this paper we are going to estimate the mode with:

   $$MODE = 2.5 \times Median - 1.5 \times Mean \qquad (1)$$

   This expression is different from the usual approximation

   $$MODE = 3.0 \times Median - 2.0 \times Mean \qquad (2)$$

   (e.g. Kendall & Stuart 1977), but was found, from the simulations, to be more accurate with our clipped distributions.

PART-B: Radio Astronomy and Aperture Synthesis

1. Consider the Very Long Baseline Array (VLBA) with 10 antennas, 25 m dia, each, operating at an observing wavelength of 1.3 cm, and longest separation of 5000 miles between the antennas. When making the images with this telescope, what will be your choice of image pixel size. **[3 marks]**

2.(i) Consider the Giant Metrewave Radio Telescope (GMRT) observing at 1.4 GHz with a total bandwidth of 2 MHz divided into 512 channels and "integration" (or averaging) time for each data point to be 2 sec. If the array consists of 30 antennas, each 45 m dia., and recording interferometric data with two polarization products, estimate the data rate. [*Hint: Do they know that FITS will normally store 2x16 bits for each complex numbers? Else, this has to be mentioned.*] **[3 marks]**

(ii) For the above observing set-up, if one wish to use the data to make an image of dimension $\theta_{pb} \times \theta_{pb}$ (where θ_{pb} is FWHM of the Primary Beam of the telescope, what will be optimum pixel size and pixel number for the image? Assume the maximum "baseline" to be 30 km. **[4 marks]**

2. Briefly explain the need of "deconvolution" while making image from radio interferometric data. **[3 marks]**

[Marks: $3 + 3 + 4 + 3 = 13.0$ **].**

Answer should be at least 75% correct to obtain full marks ½ the MARKS for right concept, otherwise the marking will be 0.

Q1. is simply computing the resolution from the diffraction limit (1.22*lambda/B_max,

$$Diffraction\ Limit = 1.22 \times \frac{\lambda}{B_{MAX}}$$

$$\lambda = 1.3cm = 1.3 \times 10^{-2}m; \quad B_{MAX} = 5000 miles = 5000 \times 1609.34 = 8.0467 \times 10^{6}\mathrm{m}$$

$$Diffraction\ Limit = 1.22 \times 1.3 \times 10^{-2}/8.0467 \times 10^{6}$$

$$= 1.971 \times 10^{-9} rad = 1.129 \times 10^{-7} deg = 4.065 \times 10^{-4} Arcsec$$

where B_max is the maximum baseline), and then introduce the required sampling of the "synthesized beam" by at least 3x3 pixel (one can do more oversampling, say 5x5 pixel, but more than that will be unnecessarily expensive from computation point of view.

Q2-(i), this is estimating data rate after averaging, so sampling is not an important factor at that stage. The telescope will produce one complex number (2x16 bits) per polarization produce (hence x2 for two pol.), per baseline (hence x 435 for baselines), per channel (hence x 512 for channels), per 2 sec integration time (hence x3600/2 for data rate per hr), so about 3.2 GB/hr.

Q2-(ii), Primary beam FWHM is given by 1.22*lambda/D, where D is dish dia.; this gives the FoV. Pixel size is ~ (1/3)* (1.22*lambda/B_max). Pixel number then will be (theta_pb / pixel size)^2.

Q3 - the measured visibility (after flagging and calibration) V_o = SxTxWxV_T where V_T is the true visibility function, S,T,W are sampling, tapering and weighting function. The sky image I(l,m) is FT of V_T; so FT of V_o will be FT[SxTxW] convolving FT[V_T]. That is, the FT of observed, calibrated visibility will produce I(l,m) convolved with "dirty beam" B_d, where B_d is FT[SxTxW], a very ill-behaved point spread function. Deconvolution is necessary to take out the effect of this dirty beam and produce a clean image.

PART-C Skewed Data Reduction

1. How to measure skewness of a variable in a dataset?
2. Why Box cox transformation does not work for negative values
3. Which one is best- Log transformation, Box cox, square root, and cube transformation in handling skewed data- Justify your answer?
4. Right skewed or left skewed data- How to handle
5. Differentiate between Extra tree regressor and Random Forest regressor in the context of skewed data handling
6. How Pearson's Coefficient is useful for handling data skewness
7. Explain few applications of spark in astronomical big data analysis
8. Why skewed data is not preferred in data modelling
9. What do you understand by symmetric dataset?
10. Does logarithmic transformation compress the magnitude of data? - If yes justify your answer
11. Explain the usage of Kurtosis in data distribution.

[Marks: $11 \times 1.0 = 11.0$].

Answer should be at least 75% correct to obtain 1 otherwise the marking will be 0.

1.The formula given in most textbooks is **Skew = 3 * (Mean – Median) / Standard Deviation**.

2.Box-Cox transformations are designed for non-negative responses, but can be applied to data that have occassional zero or negative values by adding a constant α to the response before applying the power transformation.

3.Depends on dataset. Need to chk all on particular dataset.

4.log transformation: transform skewed distribution to a normal distribution. ...

1. Remove outliers.
2. Normalize (min-max)
3. Cube root: when values are too large. ...
4. Square root: applied only to positive values.
5. Reciprocal.
6. Square: apply on left skew.

5.Extra tree regressor performs better.

6.Pearson's coefficient of skewness (second method) is calculated by **multiplying the difference between the mean and median, multiplied by three**. The result is divided by the standard deviation

7.Astrospark, Astro image processing

8.Degrades the performance of model. Not able to predict where data is skewed.

9.Gaussian distribution. No skewness.

10.Despite the common belief that the log transformation can decrease the variability of data and make data conform more closely to the normal distribution, this is usually not the case. Moreover, the results of standard statistical tests performed on log-transformed data are often not relevant for the original, non-transformed data.

11.Kurtosis is **a measure of the combined weight of a distribution's tails relative to the center of the distribution**. When a set of approximately normal data is graphed via a histogram, it shows a bell peak and most data within three standard deviations (plus or minus) of the mean.

PART-D: HDF5 Data Format

1. Name the python package which is used as an interface for HDF5?
2. When we should use HDF5 file system.
3. Name the command to display all the groups of HDF5.
4. What is the difference between HDF5 group and HDF5 dataset?

6. Name any three compression filters supported by h5py.
7. How will you print file attributes in HDF5?
8. How will you list all variables of a HDF5 file.
9. Why HDF5 works well for dynamic data.
10. Create an HDF5 file named groups. hdf5 and under this file create a group IIITA and under this group create 2 different datasets BTech and MTech.
11. How will you read and write in HDF5 using h5py.

[Marks: $11 \times 1.0 = 11.0$].

Answer should be at least 75% correct to obtain 1 otherwise the marking will be 0.

1. h5py package.
2. When file size is large or the source of the information is heterogenous.
3. visit(get_all).
4. Datasets: Array-like collections of data.
   Groups: Folder-like containers that hold datasets and other groups.
5. GZIP, LZF, and SZIP.
6.
```
for k in hdfid.attrs.keys():
    print('{} => {}'.format(k, hdfid.attrs[k]))
```
7.
```
for item in hdfid:
    print(item)
```
8. Because it uses B-trees to index table objects.
9.
```
with h5py.File('groups.hdf5', 'w') as f:

    g = f.create_group('IIITA')
   d = g.create_dataset('Btech', data)
   dd = g.create_dataset('Mtech', data)
```
10. f = h5py.File(filename, modeType)
    where
    mode Type: 'w', 'r+', 'r', or 'a'