Lab Manual: Data Science in Radio Astronomy I

Introduction

In this lab, you will explore advanced techniques in radio astronomy data analysis, focusing on spectral data. Using a Jupyter notebook as your primary tool, you will analyze real 21cm hydrogen line data from the Allen Telescope Array (ATA), an essential component in studying the structure and dynamics of the Milky Way galaxy.

This lab is divided into two parts, each designed to build your skills in data analysis and connect them to real-world applications in astronomy and the search for extraterrestrial intelligence (SETI). In Part 1, you will experiment with simulated signals to investigate the importance of spectral resolution, a critical factor in detecting faint, narrowband signals often targeted in SETI searches. In Part 2, you will analyze 21cm hydrogen line spectral data from the Allen Telescope Array, fitting Gaussian profiles to explore the velocities and distribution of hydrogen gas.

The provided Jupyter notebook serves as both a guide and a workspace, enabling you to modify and extend code while making predictions, interpreting results, and drawing connections to broader astronomical questions.

Part 1: Introduction to Spectral Data in SETI

Background: Spectral Resolution and Narrowband Signals

In radio astronomy and SETI (Search for Extraterrestrial Intelligence), spectral resolution plays a vital role in detecting and interpreting signals. Spectral resolution refers to the ability to distinguish closely spaced frequencies in the frequency domain. A higher resolution allows us to detect narrowband signals, which are critical for identifying faint sources and distinguishing them from background noise. However, frequency resolution must also be balanced with computational cost, and other factors such as time resolution.

Why Does SETI Focus on Narrowband Signals? Narrowband signals—those concentrated at a single frequency or over a very narrow frequency range—are often considered the most likely type of artificial signal. Natural astrophysical processes, such as thermal radiation from stars or synchrotron radiation from pulsars, typically produce broadband signals that span a wide range of frequencies. Narrowband signals, in contrast, are far more likely to arise from technology, as they require energy to be focused in a precise frequency range.

SETI searches target narrowband signals for several reasons:

1. **Energy Efficiency:** Concentrating energy into a narrow frequency range is far more efficient than broadcasting across a wide spectrum. This makes narrowband signals an appealing choice for advanced civilizations trying to communicate or signal their presence.

2. **Ease of Detection:** Narrowband signals stand out against the natural background noise of the universe, making them easier to detect and identify using modern radio telescopes.

3. **Universality:** Narrowband signals can be transmitted at frequencies that are universally significant, such as the 21cm hydrogen line. Hydrogen is the most abundant element in the universe, and its spectral line is a natural "benchmark" frequency that intelligent beings might use to indicate their presence.

In this part of the lab, you will explore the impact of spectral resolution on our ability to detect narrowband signals. Using simulated data, you will examine how different levels of spectral resolution affect the visibility of a narrowband signal embedded in noise. This exercise highlights the importance of selecting the appropriate spectral resolution in SETI searches, as well as the broader implications for detecting and interpreting weak signals in the frequency domain.

Steps

1. **Signal Generation and Initial Observations:**

   (a) Open the provided Jupyter Notebook file `DataScienceLab.ipynb` and locate the first section titled **Part 1: Exploring Spectral Resolution**. Ensure all required libraries (such as `numpy`, `matplotlib`) are installed and the notebook is properly loaded.

   (b) Run the cell that generates a noisy time-domain signal with a narrowband component and performs FFT analysis with low and high spectral resolutions.

   (c) Observe the two plots produced by the notebook: one showing the Power Spectral Density (PSD) at low spectral resolution (FFT size = 64) and the other at high spectral resolution (FFT size = 2048).

   (d) Record your observations about the differences in the appearance of the signal in the low-resolution and high-resolution plots. Note how the narrowband signal becomes more distinguishable at higher resolutions.

2. **Hypothesis and Predictions:**

   (a) In your lab notebook, describe why you think narrowband signals are easier to detect with higher spectral resolution. Connect your reasoning to the nature of the SETI searches for narrowband signals, considering their artificial origins and their distinct appearance in the frequency domain.

   (b) Predict how the PSD will change if the FFT size is further increased or decreased beyond the provided values. Record your predictions.

3. **Experimentation with Spectral Resolution:**

   (a) Modify the FFT size (`N_low_res` and `N_high_res`) in the Jupyter Notebook to explore additional spectral resolutions. Suggested values include FFT sizes of 32, 128, 512, and 4096.

   (b) Rerun the relevant cells in the notebook to regenerate the PSD plots for the new resolutions.

   (c) Record your observations in your lab notebook for each FFT size. Note the clarity of the narrowband signal and the noise floor at each resolution.

4. **Apply the Welsh Method for Noise Reduction**

(a) Run the cells to compute the power spectral density using Welsh's method.

(b) Experiment with different segment lengths (must be a power of 2) and record the effect this has on frequency resolution and signal-to-noise ratio (SNR).

5. **Analysis and Interpretation:**

(a) Compare the PSD plots at different resolutions and analyze the trade-offs of using higher versus lower spectral resolutions.

(b) Write a brief reflection in your lab notebook about how spectral resolution impacts the detection of weak, narrowband signals, and why this is critical for SETI research.

(c) Summarize your findings in a few sentences, emphasizing the role of spectral resolution in enhancing the visibility of signals of interest against a noisy background.

Part 2: Curve Fitting and Hydrogen Line Analysis

Objective

Use Gaussian fitting techniques to analyze complex hydrogen line profiles and extract meaningful parameters such as amplitude, mean frequency, and line width.

Background

In radio astronomy, fitting spectral profiles with Gaussian models is a standard method for analyzing emission lines. By modeling these lines, astronomers can deduce physical properties of the observed gas, such as temperature, density, and motion.

Steps

1. **Load your Data**

(a) In the `filename` line, point to the path where you've saved your assigned data. This can be done by locating the file on your computer and selecting "Copy as Path". Note, you may need to replace the backslashes (\) with double backslashes (\\) in your file path to read in the data.

2. **Focus on the Region of Interest:**

(a) Identify the region of interest (ROI) around the hydrogen line frequency in the loaded data. Use the provided code to mask the data within this range.

3. **Define the Multi-Gaussian Model**

(a) Examine the provided function for modeling multiple Gaussians with a linear background. Discuss how this model captures the complexity of the spectral profile.

4. **Define the Initial Guesses for the Gaussian Function**

(a) Complete the code to input your initial guesses for the Gaussian functions in a for loop, using the format `initial_guess.extend([amplitude, center + i * spacing, standard_deviation])`

5. **Perform Curve Fitting in the Frequency Domain**

   (a) Use the `curve_fit` method to fit the multi-Gaussian model to the ROI data.

   (b) Record the fitted parameters for each Gaussian (amplitude, mean, sigma).

   (c) Analyze how well the model fits the observed data.

6. **Convert Frequency to Doppler Velocity**

   (a) Complete the `calculate_velocity` function to convert frequency to Doppler velocity using the Doppler equation

$$v = c \cdot \frac{f_{rest} - f_{observed}}{f_{rest}} \tag{1}$$

7. **Plot Data in the Velocity Domain**

   (a) Using the `calculate_velocity` function you just wrote, convert frequency to velocity and make a new plot (Normalized Amplitude vs. Velocity).

8. **Perform Curve Fitting in the Velocity Domain**

   (a) Use the same method as before to define your initial guesses for the Gaussian fit:

`initial_guess.extend([amplitude, center + i * spacing, standard_deviation])`.

9. **Perform Curve Fitting in the Velocity Domain**

   (a) Run the code block to perform the Gaussian curve fit to the velocity data.

10. **Calculate Temperature from the Curve Fit**

    (a) Using the kinetic temperature equation, complete the code to find the temperature of the hydrogen gas from the curve fit parameters.

$$T = \frac{m_H \cdot (\sqrt{2} \cdot \sigma_v)}{2 \cdot k_B} \tag{2}$$

11. **Compare Observations Across Groups:**

    (a) Each group will analyze a unique dataset. Compare the fitted parameters and spectral features across datasets.

12. Discuss how variations in the data reflect differences in hydrogen gas distribution and motion.

13. **Reflect on Applications:**

    (a) Discuss how the ability to detect and model hydrogen line emission relates to SETI and the search for narrowband signals from extraterrestrial intelligence.

Reflection and Discussion

14. Reflect on the challenges of spectral analysis and the importance of careful parameter selection (e.g., FFT size, fitting constraints).

15. Discuss the broader implications of hydrogen line studies for understanding galactic dynamics and detecting extraterrestrial signals.

16. Consider how advancements in data science tools might improve future analyses in radio astronomy and SETI.