Draft for Astrostatistics project

Pratyusha Bala

March 2025

1 Introduction

This work is inspired by "Extracting the Global 21-cm signal from Cosmic Dawn and Epoch of Reionization in the presence of Foreground and Ionosphere" by Tripathi et al. (2024) and an effort to find alternative approaches to the methodology presented in their study.

1.1 Main Objective :

Detecting the redshifted H_I 21-cm emission plays a crucial role in exploring the Universe's first billion years. However, given the significantly brighter foreground, detecting it is observationally difficult. The Earth's ionosphere distorts the signal at low frequencies by introducing directional-dependent effects.

Our main aim for this paper is to address the problem of extracting H_I 21-cm signal from the composite all-sky averaged signal, including foreground and ionospheric effects such as refraction, absorption, and thermal emission from the ionosphere's F and D-layers using non-parametric methods like polynomial regression, Gaussian process regression and trend filtering Tibshirani (2014)for the extraction.

1.2 Methodology :

For simulating the signal, we are using the "parametrized" model used in Tripathi et al. (2024). We conduct experiments under various scenarios using a synthetic set of the global 21-cm signals created by altering its parameter space based on the "tanh" parametrized model and the Accelerated Reionization Era Simulations (ARES) algorithm. We also assume a 'perfect' instrument, neglecting instrumental calibration and beam effects. (Tripathi et al., 2024)use Artificial Neural Networks (ANNs). We are exploring the nonparametric methods.

1.3 Origin of the signal

The theory behind the emission: 21-cm signal refers to the electromagnetic radiation with a wavelength of 21 cm (frequency of ~ 1.42 GHz) emitted by neutral hydrogen atoms. A neutral hydrogen atom consists of a proton and an electron, both having intrinsic spins. The state at which both spins are aligned have a slightly higher energy than the the anti-aligned state. When the spins flip from being aligned to anti-aligned, the atom emits energy in the form of the 21-cm radiation.

The theory behind the origin: "Sourced from Wikipedia & Fialkov et al. (2024)"

In the Big Bang model for the formation of the Universe, inflationary cosmology predicts that after a few seconds the Universe expanded rapidly, known as Inflation. As it expanded, it cooled down, allowing fundamental sub-atomic particles to form. It took about 380,000 years before electrons combined with nuclei to form neutral atoms, leading to the release of the Cosmic Microwave Background (CMB) radiation. The next period was called the Dark Ages - a period without any stars or galaxies, filled mostly with neutral hydrogen(shown in the 1).

Gradually, over millions of years, the first stars and galaxies started forming. This epoch is known as the Cosmic Dawn. The first stars emerged producing Ly-α photons that couple the 21-cm spin temperature to the kinetic temperature of the gas, where the gas temperature is colder than the CMB, resulting in the characteristic absorption trough of the H_I 21-cm signal(notice the lowest point around $z \sim 18$ in 1).

As more stars and galaxies formed, the absorption deepens, and continues until the first population of heating sources emerges.The heating causes the neutral gas temperature might to either rise above that of the CMB radiation resulting in an emission 21-cm signal (as shown in Figure 1 at the low-redshift end). This epoch is called the Epoch of Reionization. Finally, the signal vanishes as the neutral hydrogen in the IGM is ionized by galaxies and quasars.

Basically, the signal is found in the Universe where neutral hydrogen is abundant and by tracking this signal over time, we can map the distribution and evolution of neutral hydrogen during the Cosmic Dawn and Epoch of Reionization.

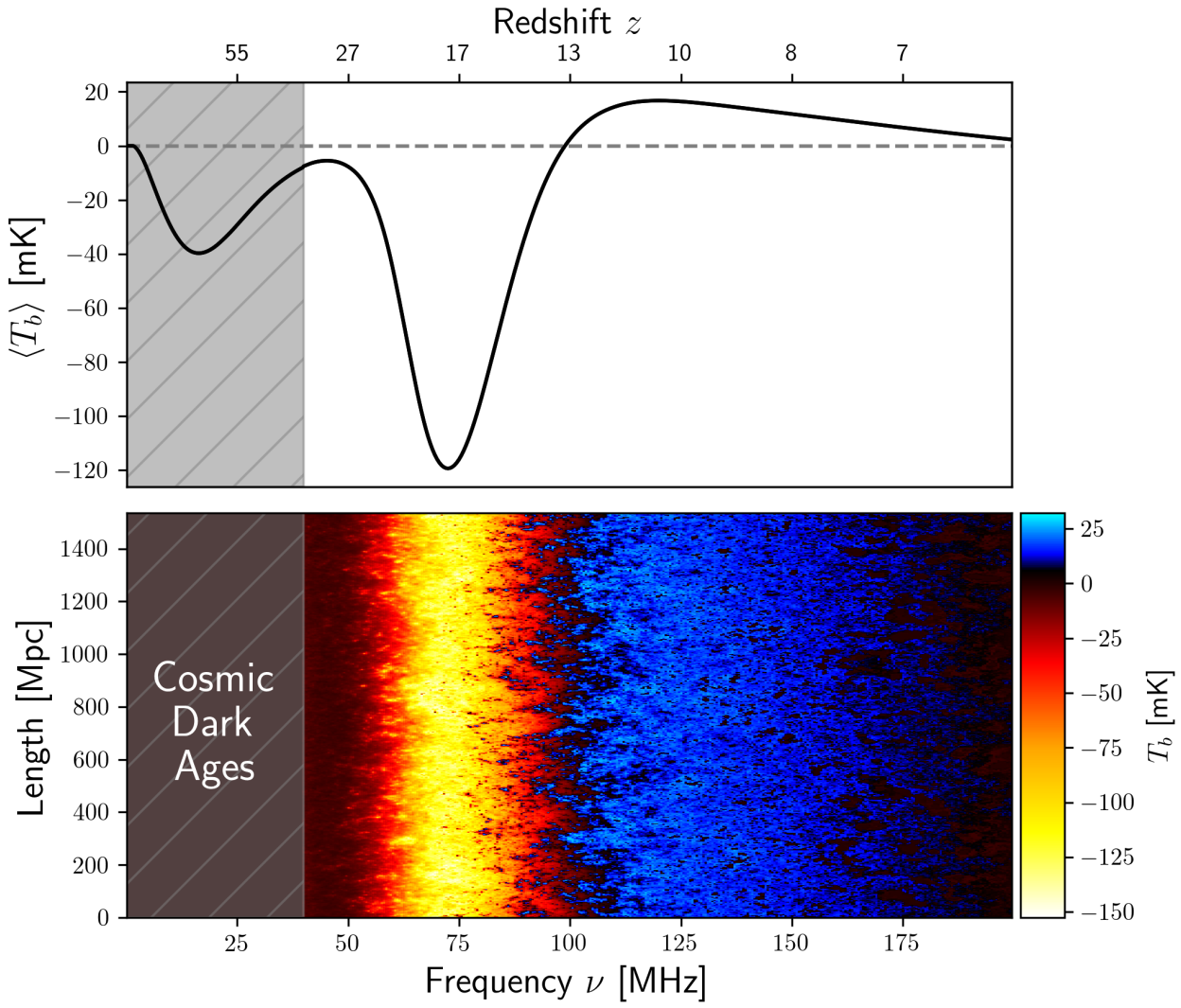

Figure 1: The 21-cm signal across cosmic time by Fialkov et al. (2024).The demonstrated timeline covers (from left to right) the Dark Ages ($z \geq 30$), Cosmic Dawn ($z \sim 10 - 30$), and the entirety of the Epoch of Reionization ($z \sim 10 - 6$, with the process of reionization completed by $z \sim 5$). The sky-averaged (global) signal (top) and a light cone map of spatial fluctuations (bottom) as a function of time (horizontal) and space (vertical) are shown.

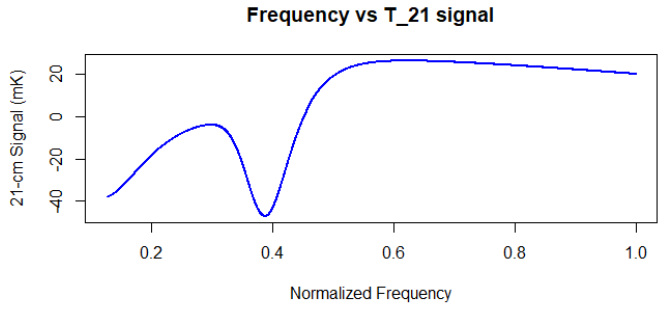

Figure 2: A sample of 21-cm signal generated from ARES algorithm

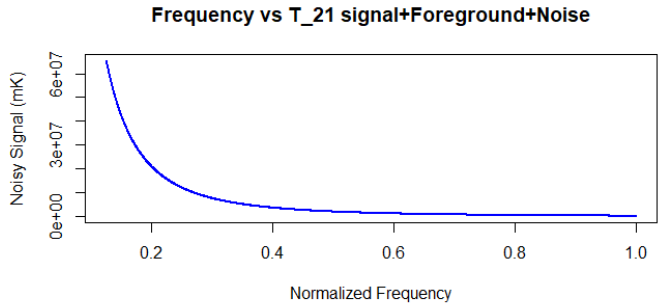

Figure 3: The corresponding Signal with foreground and noise

2 Models :

The quantity we can measure is known as "differential brightness temperature", δT_b. We measure this quantity relative to Cosmic Microwave Background (CMB) followed by :

$$\delta T_b \equiv T_b - T_\gamma \approx 27(1 - x_i)\left(\frac{\Omega_b h^2}{0.023}\right)\left(\frac{0.15}{\Omega_{m,0}h^2}\frac{1+z}{10}\right)^{1/2}\left(1 - \frac{T_\gamma(z)}{T_s}\right)mK$$

where Ω_m and Ω_b signify total matter density and baryon density, respectively, $T_\gamma(z)$ denotes CMB temperature at redshift z, and T_s is spin temperature. We get the right side of the equation to construct the signal, neglecting the peculiar velocity and density fluctuation components in the global signal.

2.1 Parametrized model for simulating the global 21-cm signal

In Tripathi et al. (2024), they use the tanh parametrized model in the ARES algorithm to replicate the global 21-cm signal across the redshift range $6 < z < 40$.

3 Simulation WORKING

3.1 Building of Datasets

Following the same steps as in Tripathi et al. (2024), data sets are created using each parameter value sampled randomly and uniformly from the given parameter range in Tripathi et al. (2024). We have 1000 samples of the actual pure signal (T_{21}), the foreground (T_{fg}), and the thermal noise, on a range of 1024 frequencies, all individually available, which sums up to T_sky_noisy, imitating the noisy signal observed in real life scenarios from physical instruments.

3.2 Gaussian Process Regression

3.2.1 Homoscedastic Noise:

Modeling the observed sky signal $T_{sky}(\nu)$ at frequency ν as a sum of three components, we get,

$$y(\nu) = T_{sky}(\nu) = T_{\text{fg}}(\nu) + T_{21}(\nu) + \varepsilon(\nu)$$

where, $\quad T_{\text{fg}}(\nu) \sim \mathcal{GP}\left(0, K_{\text{fg}}(\nu, \nu')\right), \quad T_{21}(\nu) \sim \mathcal{GP}\left(0, K_{21}(\nu, \nu')\right), \quad \varepsilon(\nu) \sim \mathcal{N}(0, \sigma^2_{\text{noise}})$

The covariance matrix for the full model is given by:

$$K_y = K_{\text{fg}} + \lambda_{21} K_{21} + \sigma^2_{\text{noise}} I$$

where, λ_{21}, the scale factor, is a hyperparameter to control how much of the T_{sky} variance comes from the signal T_{21}.

The posterior means are: $\hat{\mu}_{\text{fg}} = K_{\text{fg}} K_y^{-1} y \; ; \hat{\mu}_{21} = \lambda_{21} K_{21} K_y^{-1} y$

Using Matérn kernels to model both the smoother foreground and comaratively rougher 21-cm signal component. The Matérn kernel is defined as:

$$K_{\text{Matern}}(x, x') = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}|x - x'|}{\ell} \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}|x - x'|}{\ell} \right)$$

where:

- $\nu > 0$ controls the smoothness of the function.
- $\ell > 0$ is the lengthscale.
- K_ν is the modified Bessel function of the second kind.

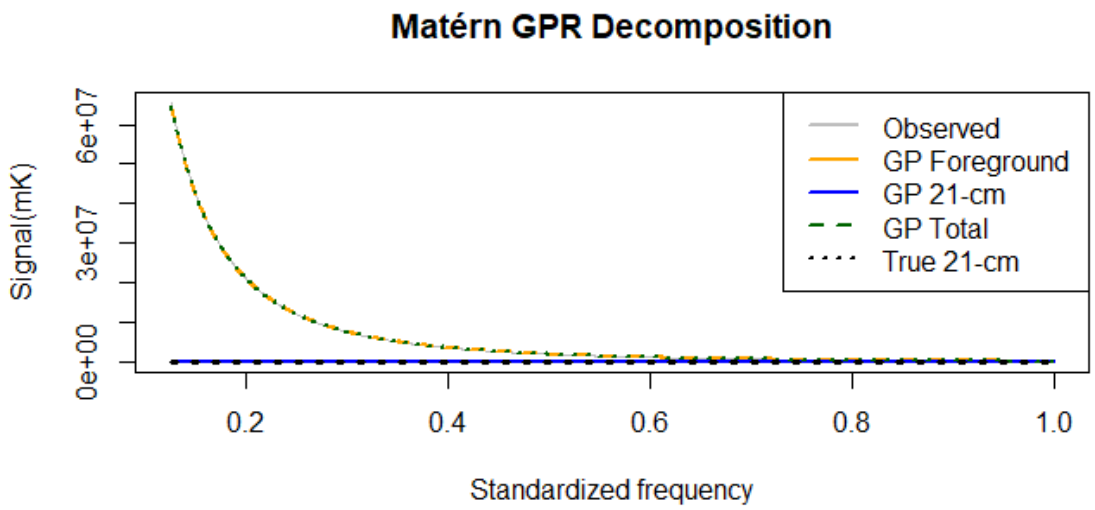

Figure 4: The true 21-cm and the noisy signal vs their homoskedastic GP estimates for a random sample

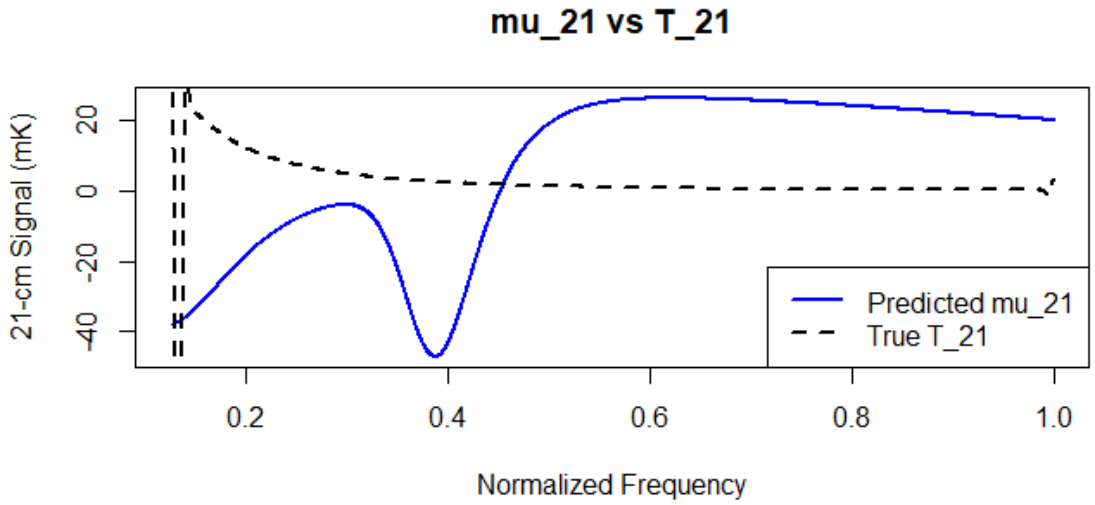

Figure 5: T_{21}(blue) vs. GPR estimate(black dotted).

Optimized the marginal likelihood:

$$\log p(y \mid \theta) = -\frac{1}{2} y^\top K_\theta^{-1} y - \frac{1}{2} \log |K_\theta| - \frac{n}{2} \log(2\pi)$$

with respect to the log-transformed hyperparameters $\theta = (\ell_{fg}, \ell_{21}, \sigma^2_{\text{noise}})$, using `optim`. Used a grid search to find ν by minimizing the RMSE between the GP-estimated 21-cm signal and the actual T_{21}. We used $\nu_{\text{fg}} = 1.5$, $\nu_{21} = 0.5$, $\ell_{\text{fg}} = 0.03$ (smoother) , $\ell_{21} = 0.0024$ (wigglier) and the scaling factor $\lambda_{21} = 10^{-5}$ for a random sample as shown in 2 and 3.The results from this homoscedastic GP is shown in 4 and the extracted 21-cm signal looks like 5 over the actual signal.

3.2.2 Gaussian process with heteroskedastic noise

Assuming, a heteroscedastic noise which depends on input is more proper in this case. Because instrumental noise originates mainly from the receiver temperature, which is dominated by the foreground brightness.

Let us assume the parametric form noise $\propto T_{fg}$.

We write the observed spectrum at frequency ν as

$$y(\nu) = T_{\text{sky}}(\nu) = T_{\text{fg}}(\nu) + T_{21}(\nu) + \varepsilon(\nu), \tag{1}$$

with $\varepsilon(\nu) \,|\, T_{\text{fg}}(\nu) \sim \mathcal{N}\big(0,\ c\, T^2_{\text{fg}}(\nu)\big)$.

Step 1: We first get a smooth estimate of the foreground $\widehat{T}_{\text{fg}}$ via either trendfiltering or heterogenous GP.

Step 2: Dividing the data by the smooth foreground estimate, $\widehat{T}_{\text{fg}}(\nu)$ gives

$$r(\nu) = \frac{y(\nu)}{\widehat{T}_{\text{fg}}(\nu)} - 1 \;=\; \frac{T_{21}(\nu)}{\widehat{T}_{\text{fg}}(\nu)} + \xi(\nu), \qquad \xi(\nu) \sim \mathcal{N}\big(0,\ c\big),$$

which makes this transformer noise *homoskedastic*.

Step 3: We can now model the residual GP with kernel

$$K_r = \lambda_{21}\, K_{21} + c\, I,$$

where K_{21}^{frac} is the covariance of $T_{21}(\nu)/\widehat{T}_{\text{fg}}(\nu)$ and λ_{21} is a scaling factor.

Step 4: Posterior means in the original temperature units are obtained by

$$\widehat{T}_{21}(\nu) = \widehat{T}_{\text{fg}}(\nu)\ K_{21}^{\text{frac}} K_r^{-1} r, \qquad \widehat{T}_{\text{fg}}(\nu) = K_{\text{fg}} K_y^{-1} y.$$

The results are as shown in 6.

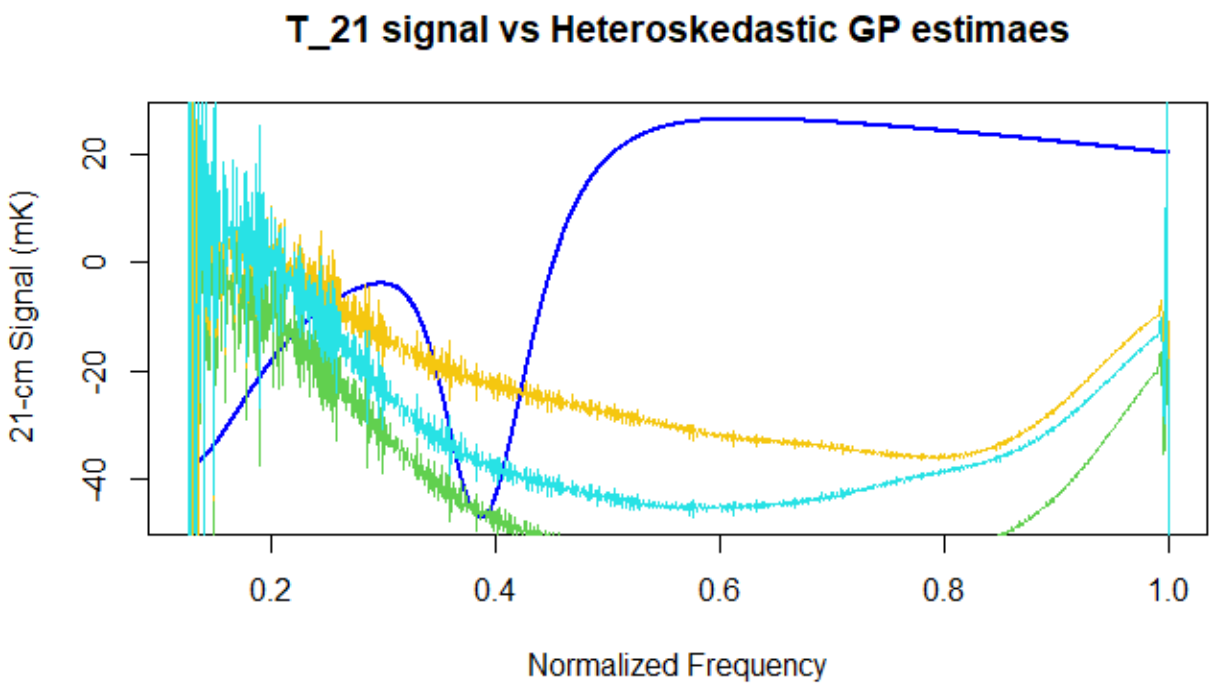

Figure 6: (green) GP assumes the noise GP variance does not inherit the scale for foreground variance. (yellow)GP assumes the noise GP inherits that scale. (sky) GP assumes that noise surface hugs the empirical variances not much tightly, ultimately saying that the noise GP is smoother.

References

Fialkov, A., Gessey-Jones, T., and Dhandha, J. (2024). Cosmic mysteries and the hydrogen 21-cm line: bridging the gap with lunar observations. *Philosophical Transactions of the Royal Society A*, 382(2271):20230068.

Tibshirani, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering.

Tripathi, A., Datta, A., Choudhury, M., and Majumdar, S. (2024). Extracting the global 21-cm signal from cosmic dawn and epoch of reionization in the presence of foreground and ionosphere. *Monthly Notices of the Royal Astronomical Society*, 528(2):1945–1964.