Separating the Cosmic Microwave Background from Galactic Foregrounds Using Generalized Morphological Component Analysis

Francesca Gear

Advisor: Dr. Colin Bischoff, University of Cincinnati Department of Physics

University of Cincinnati Capstone Final Report

Paper Submitted April 27th, 2021

Abstract

This work probes the efficacy of Generalized Morphological Component Analysis (GMCA), a novel sparsity-based Bayesian component separation method, on CMB-Stage 4 (CMB-S4) simulated data, and creates a GMCA module for large scale data analysis. The work was done in Terminal, using Python in Jupyter notebooks and NERSC servers from the Department of Energy. Source map and mixing matrix optimizations were performed to minimize chi squared to separate the nine frequency maps into foreground and CMB components. Results show strong correlation between the CMB, dust, and synchrotron component maps given as a result of the code, and the original simulation input maps. The maps created in Python were compared visually to the simulation input maps, and also quantitatively using correlation calculations. We find that our GMCA module accurately and efficiently produces maps that show significant correlation between the GMCA component maps, and the simulation input maps.

1 Introduction

The Cosmic Microwave Background (CMB) is the relic radiation of polarized light from the explosion that created the Universe called the Big Bang. The Big Bang is the explosion that took the Universe from a hot, dense singularity to the expanding, cooling Universe we see today. 380,000 years after the Big Bang, sometimes called the surface of last scattering, the Universe became transparent enough for light to move through, and the CMB was emitted. [6] The CMB has cooled to a chilly 2.725 kelvin, or about 3 degrees above absolute zero, and is one of the most perfect examples of a blackbody known to science (a blackbody is an object that absorbs all radiation falling on it). Since the Universe is expanding and cooling presently, we know the Universe used to be smaller and hotter, which means that the CMB signal we see today is significantly redshifted (from UV and visible light to microwaves, more than any galaxy or quasar ever seen).The CMB is almost uniform, but has slight temperature fluctuations which can be pictured as slight hills and valleys. Inflation creates these scalar perturbations, which form the large sclae galactic structure we see today. The sky signal that we see with our telescopes has 3 main components: the CMB itself, dust, and synchrotron (which we will henceforth refer to as *foregrounds*). These foregrounds must be removed from the sky signal, as they contaminate the CMB data, and create inaccurate power spectra which are what we aim to study. The power spectra of the CMB is used to search for evidence of primordial gravitational waves and learn about inflation, and is a very hot topic in cosmology today.

1.1 The Big Bang

The theory of the Big Bang is the leading explanation of how our universe began, and states that the universe began with a small, dense singularity, and *expanded* over 14 billion years into the universe we see today, evident in Figure 1.

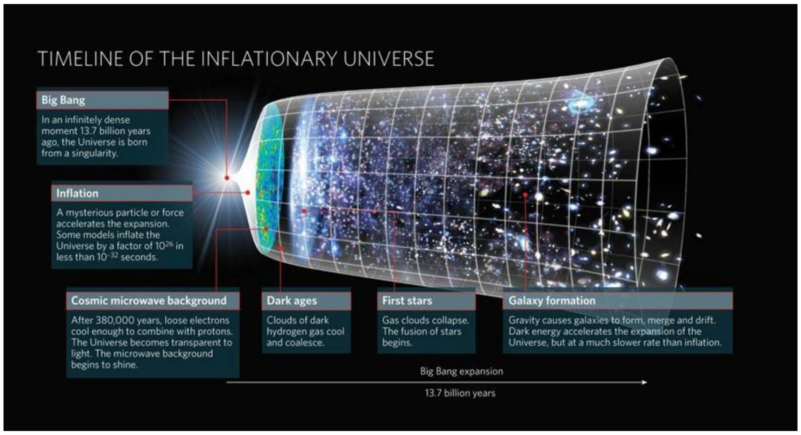

TIMELINE OF THE INFLATIONARY UNIVERSE
Big Bang
In an infinitely dense moment 13.7 billion years ago, the Universe is born from a singularity.
Inflation
A mysterious particle or force accelerates the expansion. Some models inflate the Universe by a factor of 10^{26} in less than 10^{-32} seconds.
Cosmic microwave background
After 380,000 years, loose electrons cool enough to combine with protons. The Universe becomes transparent to light. The microwave background begins to shine.
Dark ages
Clouds of dark hydrogen gas cool and coalesce.
First stars
Gas clouds collapse. The fusion of stars begins.
Galaxy formation
Gravity causes galaxies to form, merge and drift. Dark energy accelerates the expansion of the Universe, but at a much slower rate than inflation.
Big Bang expansion
13.7 billion years

Figure 1. Expansion of the Universe

Since our telescopes and other astronomical instruments aren't able to see back to the very beginning of our universe (the Universe was opaque to light before recombination), we learn most of what we know about the Big Bang using mathematical models. However, the *echo* of the Big Bang can see seen directly in one phenomenon, something we call the Cosmic Microwave Background or CMB. In the first second of the universes life, the temperature was around 10 billion degrees Fahrenheit. The environment was full of a variety of fundamental particles, such as neutrons, electrons, and protons. These fundamental particles either decayed or combined as the universe cooled. Even if you could travel back in time and be present at this exact moment, you wouldn't be able to see what was happening. This is because the universe at that stage was not transparent enough for light to move through. The free electrons would have caused the photons of light to scatter, similar to how sunlight scatters off water droplets in clouds. [7] After 380,000 years, the free electrons combined with atomic nuclei to create neutral atoms, and light could finally shine through. This primordial light is often referred to as the 'afterglow' of the Big Bang, but in astrophysics, we call it the CMB. [6] The CMB was officially discovered (by accident) in 1964 by two Bell Telephone Laboratory workers: Arno Penzias was a physicist, and Robert Woodrow Wilson was a radio-astronomer. They first believed that the anomaly in their equipment was due to debris in their detector, but with the help of a team from Princeton University (led by Robert Dicke), they quickly realized the anomaly was the CMB! They estimated the CMB had a temperature of 3.5 Kelvin, and won the Nobel Prize for their work in 1978. [7]

There have been several missions which have set out to probe the CMB. NASA's Cosmic Background Explorer (COBE) satellite, created maps of the sky in the 1990's. There have been other missions which followed in the footseps of COBE, such as NASA's Wilkinson Microwave Anisotropy Probe (WMAP), and the European Space Agency's Planck satellite. [4] The observations from the ESA Planck satellite first came in 2013, mapped the CMB in unprecedented detail, and revealed that the universe was actually older than we had estimated; 13.82 billion years old versus the previous estimate of 13.7 billion years. The observations of the sky signal from these three satellites is shown in Figure 2. The images from each satellite look slightly different, which is due to the varying angular resolution of these crafts.

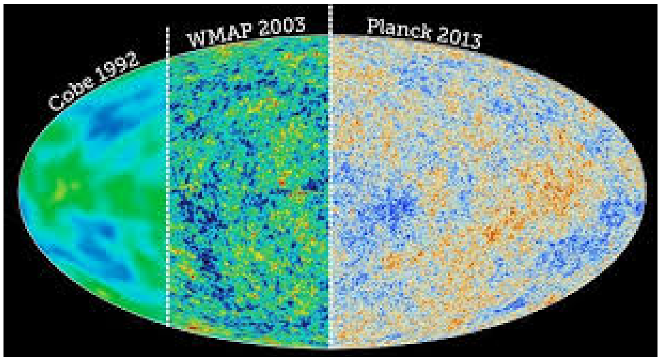


Figure 2. All-Sky signal as seen by COBE, WMAP, and Planck - Works to highlight the increasing angular resolution that comes with advances in technology

Although the data from these missions helped answer many astrophysical questions, it also raised some new ones. One prediction of the Big Bang Theory is that the CMB should be mostly uniform, no matter where you look. However, the data from Planck showed that the Southern Hemisphere of the map appeared slightly warmer than the Northern Hemisphere. The CMB also gives astronomers insight into the composition of the early universe, and helped give rise to the ideas of dark matter and dark energy. We find that only 5% of the universe is made of ordinary matter, such as planets, galaxies, and stars. [6]

Astronomers now have a pretty good idea of what happened in the early universe, but are constantly seeking proof of the rapid inflation that occurred. The theory postulates that in the first second of its life, the universe ballooned in size faster than the speed of light (which doesn't violate Einstein's speed limit, since the speed limit applies to objects inside the universe, not to the expansion of the universe itself). In 2014, using a telescope in the South Pole of Antarctica called Background Imaging of Cosmic Extragalactic Polarization (BICEP2), astronomers found evidence of what seemed to be primordial gravitational waves in the CMB, using particularly the B modes, but it was later determined to be excess signal from galactic dust (which is why foreground separation is so important). The B modes are where astronomers typically search for evidence of primordial gravitational waves, which are another prediction of Inflation Theory but are hard to detect directly. Gravitational waves have previously been confirmed using the movements and collisions of black holes a few tens of masses larger than our sun using the Laser Interferometer Gravitational-Wave Observatory (LIGO) in 2016. As the sensitivity of LIGO increases, we should see black hole gravitational wave events much more frequently. [6]

Since the universe is still expanding, it also expands faster as it grows bigger. Depending on what dark energy turns out to be and a few other unknown questions about the Universe, it is one prediction that eventually, no other galaxies will be visible from earth, or any vantage point inside the galaxy. We can see the distant galaxies moving away from us, but they are moving faster with time. This means that if you wait long enough, distant galaxies will be receding at the speed of light, and their light will never have enough time to reach us. There are also scientists who pose theories of a *multiverse* - the idea that our universe is just one of many, with other universes existing side by side, like bubbles. People who believe these kinds of theories think that in the first major part of inflation, different parts of space-time grew at different rates, which could have carved out different sections (universes) with potentially different laws of physics. [6]

1.2 Cosmic Microwave Background (CMB)

The CMB shows us a snapshot of what the Universe looked like in its extremely early days. It can show us how matter was distributed throughout the Universe, which is very useful information to cosmologists today. The CMB is polarized, which describes the orientation of the light waves perpendicular to the wave vector, whereas unpolarized light has no particular orientation. The CMB is linearly polarized due to Thompson scattering of photons off free electrons in the surface of last scattering. In the CMB, we have Q polarization and U polarization, also called Stoke's parameters. The CMB is made of E and B modes, where E mode polarization is parallel and perpendicular to the direction of the wave vector, and B mode polarization at 45-degree angles to the wave vector, relative to a choice of coordinates. This can seen in Figure 3.

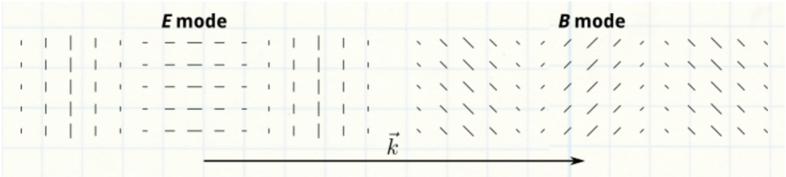


Figure 3. Polarization across a horizontal wave vector, $\vec{k}$ where it is clear that E mode polarization is parallel or perpendicular to $\vec{k}$, and B mode polarization is rotated 45° with respect to $\vec{k}$. [1]

Studying the polarization of the CMB can give cosmologists drastic insight into Inflation and how the Universe began. The most important thing to note about CMB polarization is that it can be described as a combination of E and B Modes. We can learn about Inflation by studying particularly the B modes of the CMB, because curved signatures in the CMB are evidence for these inflationary gravitational waves. This *curled signature* can only be caused by gravitational waves produced by Inflation, so if they are seen in the CMB signal, we find proof of their existence. A prediction of Inflation is that a background of gravitational waves would be produced. These gravitational waves are too faint to be directly detected today, but they can be observed through this signature on the CMB B modes. Since the universe is constantly expanding, the CMB signal we see has been redshifted. During the big bang the CMB was emitted as visible and UV light, but has been redshifted significantly, so that we now observe it in the microwave portion of the electromagnetic spectrum.

1.3 Synchrotron Emission

One of the main foregrounds which affects the CMB data is bright at low frequencies, and is called Synchrotron. Synchrotron radiation is emitted by relativistic electrons (traveling near light-speed) that travel on spiraling paths in the galaxy's magnetic field. [1]

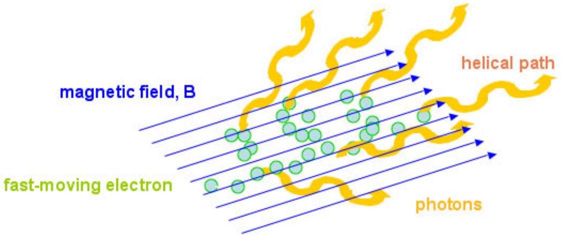


Figure 4. Synchrotron Emission: Energetic charged particles that are spiraling in galactic magnetic fields, - very bright at low frequencies

1.4 Galactic Dust

Galactic dust is the name given to microscopic bits of matter that are floating in interstellar space. Light from surrounding stars heats up these particles, and causes the dust to glow in the microwave portion of the electromagnetic spectrum. The average temperature of the dust hovers around 20 kelvin, meaning its blackbody spectrum peaks at frequencies higher than the CMB, which is why the dust is brightest at high frequencies. [1]

Figure 5. Small particles of mainly silicate and carbonaceous grains floating in the interstellar medium, which are very bright at high frequencies

1.5 Motivation

The goal of our project is to have a workable module of code that will be able to run a GMCA analysis on sets of frequency input maps, which will accurately separate the CMB from galactic foregrounds present in the maps.

2 GMCA Methods

2.1 GMCA

So, now we are going to get into the actual process I used to separate these components, which is called Generalized Morphological Component Analysis, or GMCA. GMCA is a sparsity-based Bayesian component separation method, which means that it relies on the fact that our signal is described by a small number of independent components. By small number, we mean the number of independent signals is smaller than the number of frequency maps. We have 9 frequency maps, so this technique only works as long as we can describe the signal with some number of components that is less than 9, and we happen to be using 3. [3] GMCA is an example of a semi blind source separation method. A completely blind source separation method makes no assumptions about anything before analyzing data. GMCA is partially blind, but we still make the assumption that there are 3 sky components, the CMB, dust, and synchrotron, and we assume that the signal in these maps can be described by a small number of independent components. The main goal of GMCA is to estimate the source components and their emission laws from our data. In our case. The source components are the maps of the CMB, dust, and synchrotron, and the emission laws are encoded in the mixing matrix. Remember that we said the mixing matrix tells us how each component shows up at different frequencies, this is what we mean by the emission laws of the components. A helpful visualization tool for the emission laws of the components is shown below, in Figure 6. We can see in this image that the best frequency for observing clean CMB data is around 90 GHz.

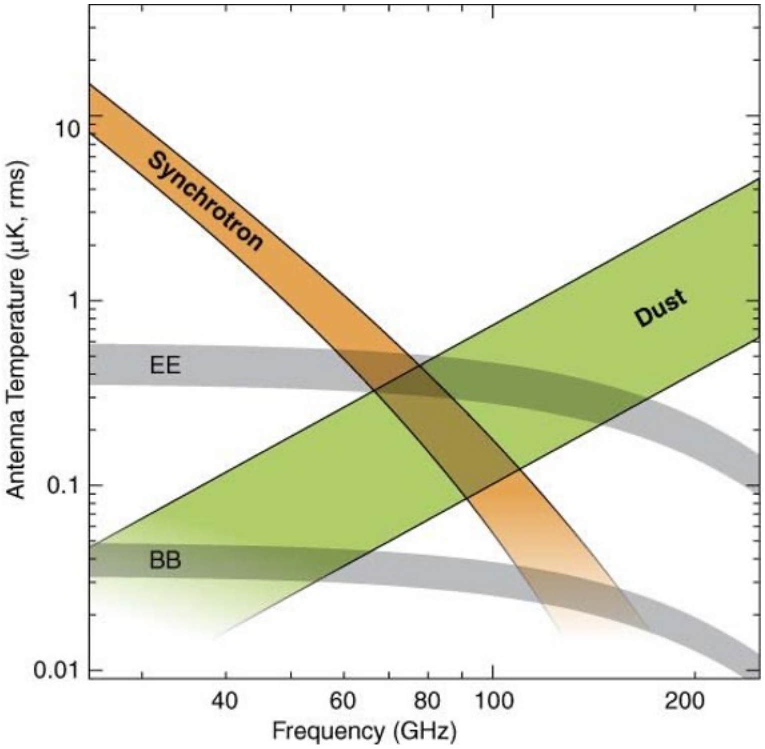


Figure 6. Emission Laws of sky components - The mixing matrix tells us how each component shows up at different frequencies - Highlights the synchrotron being bright at low frequencies, and dust being bright at high frequencies.

GMCA is a component separation method that relies heavily on blind source separation or BSS. BSS is an effective mathematical method to analyze data that is modeled as a combination of components, though blind algorithms will never be as statistically powerful as algorithms that include assumptions about the data. Each observation is modeled as the linear combination of the same sources, plus some noise term. The goal of BSS is to estimate both the mixing matrix, and the source components. This can yield problems, which can be solved with the addition of priors, either about the source components or the mixing matrix. Based on the concept of morphological diversity, the GMCA algorithm allows for the separation of sources which are sparse, and has been shown to perform very well when processing noisy data. There are several applications of GMCA in the data analysis field, including for hyper-spectral analysis, an imaging technique which processes information from frequencies all across the electromagnetic spectrum. In this case, the number of observations can be extremely large, and GMCA is able to further exploit sparsity of the sources, as well as sparsity of the mixing matrix. The GMCA algorithm assumes that all the data follows a linear mixture that is described by a mixing matrix.

For the purpose of CMB estimation from microwave data, GMCA has improved to perform the separation of sources allowing deep cosmological research. This model has proven to be especially efficient when separating sky components from CMB data gathered from the Planck satellite. [5]
In GMCA, the observed sky signal is modeled as a linear mixture of statistically significant and independent components. Then the observations made with a detector are then a noisy linear mixture of n independent sources. This linear mixture model is more conveniently written

$$\mathbf{X} = \mathbf{A}\mathbf{S} + \mathbf{N} \tag{1}$$

where $\mathbf{X}$ is the $m \times t$ data matrix (where the rows are the observed data maps), $\mathbf{A}$ is the $m \times n$ mixing matrix, $\mathbf{S}$ is the $n \times t$ source maps matrix (where the rows are the different sources cmb, synchrotron, and dust, and $\mathbf{N}$ is the $m \times t$ noise matrix. Generally in GMCA, both the sources $\mathbf{S}$ and their emission laws (mixing matrix) $\mathbf{A}$ are partially or completely unknown. GMCA then aims to estimate $\mathbf{S}$ and $\mathbf{A}$ from $\mathbf{X}$. This is typical Blind Source Separation (BSS). [2]

Next we will outline the details of Bayesian Statistics, and how they are used in GMCA.GMCA is a sparsity-based Bayesian component separation method, meaning it relies on bayesian statistics to analyze different components of the data. Bayesian statistics is a system that describes uncertainties using the mathematical language of probabilities. Typical bayesian statistical methods start with existing *prior* beliefs, and update them using data to give *posterior* beliefs. [9] In our research, we added a bayesian prior to our component map optimization, which replaced our chi squared function with a posterior function, which took an additional input, the scale parameter mu (μ), and calculated a total posterior instead of a total chi squared. We added this prior because the results of our original map optimization for the dust map looked odd. Minimizing the chi squared is equivalent to minimizing the log likelihood of a function. In bayesian statistics, we have a likelihood which is the comparison of the model and our data, but we can introduce a posterior from Bayes' theorem, which is essentially the likelihood multiplied by a prior, where the prior is just some prior knowledge about the problem at hand (could be data from another experiment, etc), but in our case, we have a sparsity-based prior. This means that we are putting in our prediction of what the solution should look like as our prior. So if the posterior is the likelihood multiplied by the prior, then instead of maximizing the likelihood, we decide to maximize the posterior. Since we talked about log likelihoods previously, we can also discuss a log posterior. We are trying to maximize a function, which is the same as maximizing the log of that function, and logs are generally easier to work with since our calculations may span many orders of magnitude. This means that the log posterior is the log likelihood multiplied by the log of the prior.

We add these priors to data analysis methods for one main reason. We don't want our optimization functions chasing after noise fluctuations in our data. Our prior says that we still have a likelihood for our data, which is still trying to find an agreement between the model and the data, but the prior is, with all other things being equal, pushing the numbers in the model to have the smallest value possible. Finding the minimal maps that work as source maps helped with the dust map issue we were facing, because theoretically, the dust and CMB maps could cancel each other out and give an inaccurate solution. The added prior will penalize solutions like that, which could have been the reason our dust map appeared to be so similar to the CMB map before we added the prior. Adding the prior fixed the issue with our dust map, and showed much less correlation with the CMB map (which is more reasonable). [2]

2.2 GMCA Module

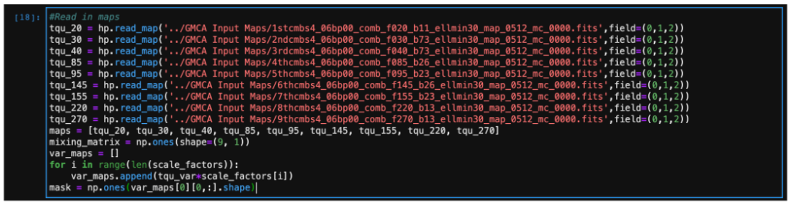
When performing GMCA for my WISE project, I wrote code in a Jupyter Notebook, and had to run each analysis on a new frequency map from scratch. This makes for very slow data analysis, and it would be almost impossible to analyze many frequency maps in a reasonable amount of time. There is a solution though: we can create a 'GMCA module' where I wrote the main components of the GMCA code, and with a little bit of set up and variable defining, calling the GMCA module will run the whole analysis on any and as many maps as we desire. This means that instead of having 200+ cells where I import all the necessary maps, define the mixing matrix and inverse noise variance, and then do a GMCA analysis by hand, I can just import the GMCA module, wait 15 minutes for the analysis to run, and then check my component maps that the code creates. It is a much more efficient way to do large scale data analysis, and creating this module was the main part of this project.

3 Implementation of GMCA

3.1 Python

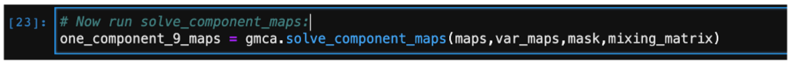
The first optimization we perform when using GMCA is to optimize our source components. We wanted a code that would look like a kind of *black box* to someone looking at the code without knowing exactly what was done. They should see that if they give my code a model, a set of data, a mixing matrix, and inverse noise variances per pixel, the function will spit out a chi squared. To do this, we wrote a code that has those 4 parameters: 1. the model, which is an array of the coefficients for the CMB, dust, and synchrotron maps – originally pixel space coefficients. 2. the data, which is an array of the Q and U maps at the 9 different observing frequencies. 3. the mixing matrix. and 4. and the inverse noise variance which tells us how much noise is in each pixel of our maps. These maps are NSIDE = 512 healpix maps, which essentially means they are maps of a very small piece of the sky. The code will then try various source component maps and find the one which minimizes our chi squared value. The inverse noise variance is calculated based on the amount of noise present in each pixel. The key part of this step is that the mixing matrix is a static input, the code for which was pre-written and sent to me by my advisor, and the source maps are allowed to vary. This is the parameter which changes in the next step.

Now we are going to talk about the second step of GMCA, which is optimizing the mixing matrix. As I mentioned before, this step is a little different from the last one. The process is generally the same, but instead of taking the mixing matrix as an input and varying the source maps to minimize chi squared, we instead use the source component maps we received from the result of the optimization in the last step as the static input, and allow the mixing matrix to vary to minimize chi squared. The end goal of GMCA is to make this an iterative process, where the result of the previous optimization is used as the input for the next optimization. This allows for very accurate data analysis. To create the gmca.py module, I wrote the backbone of the gmca code that I wrote for an earlier research project in a module in my computer's terminal. This allows me to simply import the module when doing data analysis in a Jupyter notebook after setting up some simple variables. For example, Figure 7 shows setting up these variables, and the Figure 8 shows the code where I call the gmca module to run on these defined variables. This is less than 10 lines of code, compared to over 300 in my original gmca analysis. This means I can perform GMCA on larger data sets of cmb-s4 data, or any data set of the correct size.



```
[18]: #Read in maps
      tqu_20 = hp.read_map('../GMCA Input Maps/1stcmbs4_06bp00_comb_f020_b11_ellmin30_map_0512_mc_0000.fits',field=(0,1,2))
      tqu_30 = hp.read_map('../GMCA Input Maps/2ndcmbs4_06bp00_comb_f030_b73_ellmin30_map_0512_mc_0000.fits',field=(0,1,2))
      tqu_40 = hp.read_map('../GMCA Input Maps/3rdcmbs4_06bp00_comb_f040_b73_ellmin30_map_0512_mc_0000.fits',field=(0,1,2))
      tqu_85 = hp.read_map('../GMCA Input Maps/4thcmbs4_06bp00_comb_f085_b26_ellmin30_map_0512_mc_0000.fits',field=(0,1,2))
      tqu_95 = hp.read_map('../GMCA Input Maps/5thcmbs4_06bp00_comb_f095_b23_ellmin30_map_0512_mc_0000.fits',field=(0,1,2))
      tqu_145 = hp.read_map('../GMCA Input Maps/6thcmbs4_06bp00_comb_f145_b26_ellmin30_map_0512_mc_0000.fits',field=(0,1,2))
      tqu_155 = hp.read_map('../GMCA Input Maps/7thcmbs4_06bp00_comb_f155_b23_ellmin30_map_0512_mc_0000.fits',field=(0,1,2))
      tqu_220 = hp.read_map('../GMCA Input Maps/8thcmbs4_06bp00_comb_f220_b13_ellmin30_map_0512_mc_0000.fits',field=(0,1,2))
      tqu_270 = hp.read_map('../GMCA Input Maps/9thcmbs4_06bp00_comb_f270_b13_ellmin30_map_0512_mc_0000.fits',field=(0,1,2))
      maps = [tqu_20, tqu_30, tqu_40, tqu_85, tqu_95, tqu_145, tqu_155, tqu_220, tqu_270]
      mixing_matrix = np.ones(shape=(9, 1))
      var_maps = []
      for i in range(len(scale_factors)):
          var_maps.append(tqu_var*scale_factors[i])
      mask = np.ones(var_maps[0][0,:].shape)
```

Figure 7. Defining variables in order to call GMCA.py



```
[23]: # Now run solve_component_maps:
      one_component_9_maps = gmca.solve_component_maps(maps,var_maps,mask,mixing_matrix)
```

Figure 8. Calling GMCA.py module to run on the variables set up in Figure 7.

3.2 Simulation Input Maps

The best way to separate these foregrounds from the cmb data is with multi-frequency observations. Since we know that synchrotron is bright at low frequencies and dust is bright at high frequencies, we can take measurements of the sky at low, medium, and high frequencies and use the variation in the data to separate the components. For example, I have 9 sets of observing frequencies, starting at 20 GHz which will be synchrotron dominated, going up to 95GHz which will be the best place to observe the CMB (because it is too high for synchrotron to dominate and too low for dust to dominate), and then up to 270GHz which will be dust dominated. The frequency input maps used for these simulations were 20GHz, 30GHz, 40 GHz, 85GHz, 95GHz, 145 GHz, 155 GHz, 220 GHz, and 270 GHz.

As I mentioned, I am using simulated maps of a “relatively quiet” area of the sky called the Southern Hole. These maps contain CMB, synchrotron, dust, and simulated instrument noise. These component maps are combined with a mixing matrix, which tells us how each component shows up at each frequency. These simulations are for the new state of the art CMB detector, which has reduced instrument noise when compared to current instruments. A normal question here is, why are we analyzing these simulations in the first place? The answer to that is to essentially practice separating the components and see how accurately we can do so. Therefore, when the telescopes are actually built, the component separation framework is already in place. CMB-S4 will also be designed to cross critical thresholds in testing inflation, determining the number and masses of neutrinos, constraining possible new light relic particles, providing precise constraints on the nature of dark energy, and testing general relativity on large scales.

4 Results

4.1 Output of GMCA Module

Here, you can see the results of running my gmca.py module on the just the CMB component of one frequency map (95 GHz, which as discussed above, is the cleanest frequency for observing the CMB as the dust doesn’t take a big effect yet, and the synchrotron doesn’t heavily effect frequencies this high - sort of a perfect medium), as showing the result of the full analysis would mean having over 50 images to compare. On the left, we have the results of GMCA: We have the CMB temperature map on top, the Q polarization map in the middle, and the U polarization map on the bottom. On the right, we have the CMB T, Q, and U, input map. We can see that these images are extremely similar, meaning our analysis was successful. Now getting GMCA to work wasn’t the goal of my project, the goal was to create a GMCA module that anyone can use to run on any data sets, but this comparison shows us we have done just that. Slight differences in the two maps can be accounted for, since there was added instrument noise in these simulated maps.

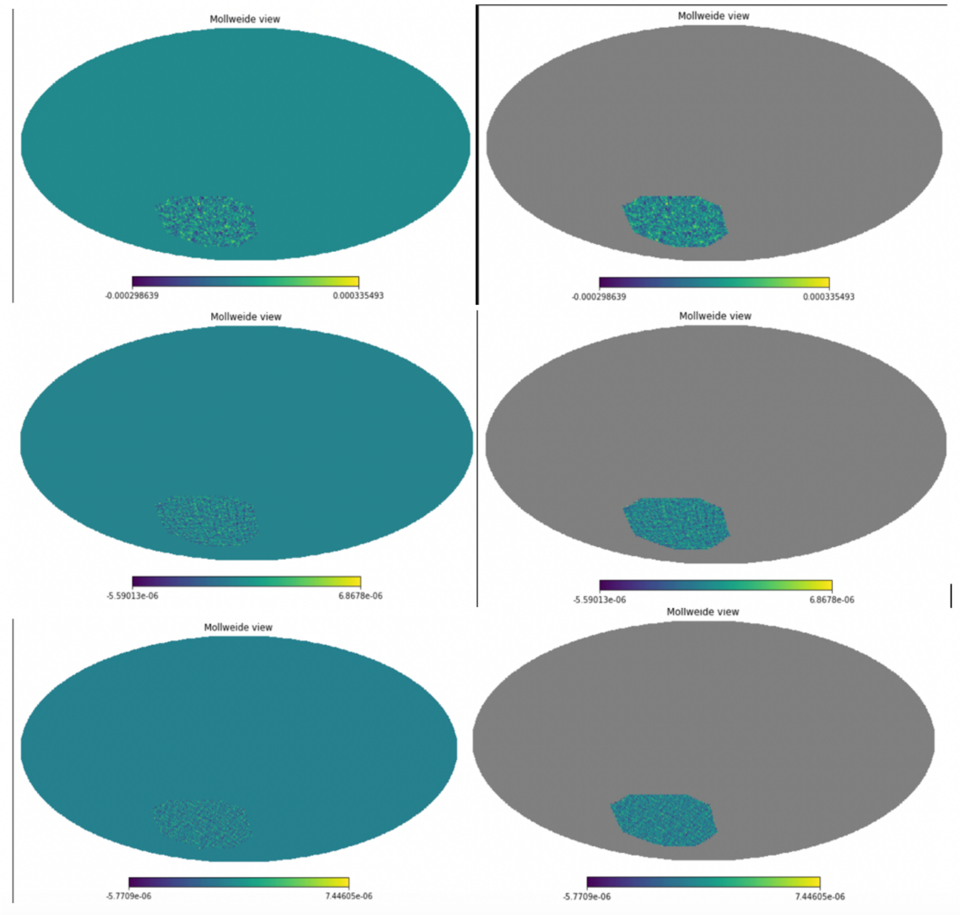


Figure 9. Results of GMCA analysis - Left: CMB temperature map on top, the Q polarization map in the middle, and the U polarization map on the bottom. Right: CMB T Q and U input map.

Visual correlation is evident in all the maps, which is a good sign. This means the models that the code came up with matches fairly well to the original frequency maps whose makeup we know. Although they look good, they are not perfect matches, and there are a couple explanations for this. We expect there to be remnant CMB signal in each of the dust and synchrotron maps, evident from the correlation functions calculated later in this section. Another issue we expect to contaminate our data is the simulated instrument noise.

We also performed a quantitative analysis on our maps, by doing a correlation calculation on our recreated maps, as compared to the original simulated input maps. The correlation coefficient of a set of data is a statistical measure of the strength of the relationship between two variables. The value of a correlation coefficient ranges from -1 to 1, with -1 showing a perfect negative correlation, +1 showing perfect positive correlation, and a value of 0 shows no linear relationship between the two variables. Shown in Figures 10, 11, and 12 are correlations calculated for the CMB, dust, and synchrotron maps respectively, compared to the original input maps. We see that for the CMB E modes, it stays right around 1 which is good. The B modes aren't as high, but we know why this is. The E modes are very bright, and thus have a high signal to noise ratio and are easy to pick out. The B modes are faint and have a lower signal to noise ratio, and don't stand out as much. The correlations for dust and synchrotron are also shown below. There isn't much difference in the correlation of the E and B modes of the synchrotron and dust maps, because the CMB has a much larger asymmetry between the E and B modes than dust or synchrotron. For the CMB, we have a big E mode map which we can reconstruct very accurately, but the B modes have so much more noise that they are harder to construct. We can see that there is good correlation between the E modes, which is due to them being very bright, meaning there is lots of signal present. There is less correlation between the B modes, because they are very faint, and therefore have a higher signal-to-noise ratio. When it comes to the foregrounds, (dust and synchrotron), there is only a mild asymmetry between the E and B modes. This means that the GMCA reconstruction works equally well for the E and B modes, so the graphs show similar plots for E and B modes.

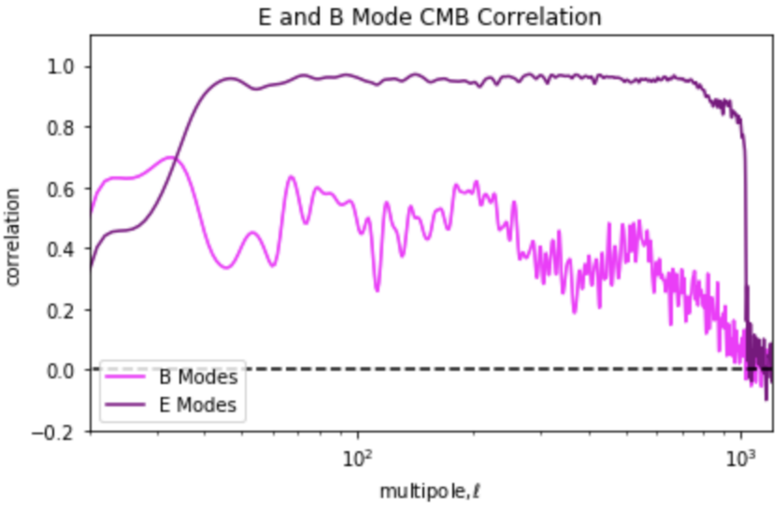


Figure 10. Correlation coefficient plot for CMB

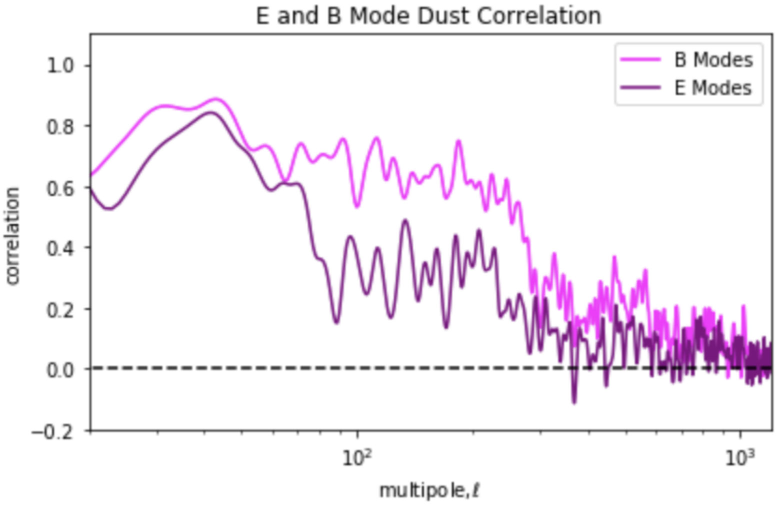


Figure 11. Correlation coefficient plot for dust

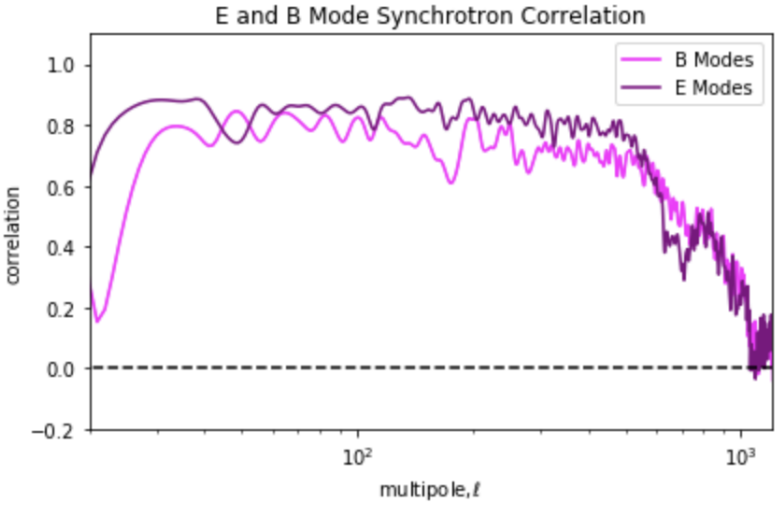


Figure 12. Correlation coefficient plot for Synchrotron

4.2 Challenges and Next Steps

As with any data analysis method, we faced a few challenges while using GMCA. One of the biggest issues came when we were optimizing the source maps with the mixing matrix as the input. The problem was the sheer number of parameters in our functions. The number of parameters was so high because it is determined by the number of maps we have which is 3, multiplied by the number of pixels in each map. So, when we used 22,000 pixels in our map, we ended up having over 68,000 parameters. This used an absurd amount of RAM on my laptop, and never had enough local memory to run successfully. This also created a run time issue: even when the code did finish and said it was unsuccessful, it took over 48 hours sometimes. We combated these issues by using a supercomputer at UC Berkeley called NERSC, which is run by the Department of Energy. This allowed my optimizations to run successfully, and the longest one took just over 2 hours. Another challenge we faced was confidence about our overall goodness of fit. We expect our chi squared value to be similar to the number of degrees of freedom, which for us was around 480,000. Our smallest chi squared value was 18 million. This could be due to how we characterized the inverse noise variance, which would largely skew our data if it is even partially incorrect. The fact that visually, our component maps look good, but we have a poor chi squared value could be due to the fact that there is a relatively low signal to noise ratio in the maps, but these noise terms blow up in the calculation of chi squared. Lastly, the modular form of GMCA.py made it difficult to troubleshoot errors in my code. The line where I call the GMCA.py module would fail, but it was very difficult to go through and determine why.

In terms of future work that could be done in this area, we could further analyze our results, and try to determine what other uncertainties may be present. A phenomenon known as E/B leakage, which refers to the artificial B-mode signal coming from the leakage of E-mode signal when part of the sky is unavailable or excluded, could be occurring, and skewing our data. [8] A colleague of mine designed a machine learning algorithm to detect and correct for E/B mode leakage, an algorithm which could be run on my maps to look for potential E/B mode leakage. Correction of such leakage is one of the preconditions for detecting primordial gravitational waves via the CMB B-mode signal, so it is a very important thing to bear in mind. More generally, there is a new component separation technique called Hierarchical GMCA which was first published in March of this year. HGMCA lets us drop the assumption that the sky has 3 main components, and instead lets the data tell us how many components there are. [10] This allows for the analysis to be fully blind, which makes the results more accurate (less assumptions being made about the data makes for more accurate final results). New component separation techniques will continue to be developed and perfected, in preparation for when CMB-S4 comes online and we have real data to perform our analysis on, instead of simulated signal.

Our correlation calculations showed very good results for the CMB E modes, correlations less than 1 for the CMB B modes, which is expected, due to the higher signal to noise ratio for the B modes. The dust and synchrotron correlations were also less than 1, but we still think our GMCA iteration was successful. Component separation is extremely difficult, and even harder to get perfect results. We were very happy with the strong results for the recreations.

In short, we wanted to create a module that could clean large sets of CMB data from foreground contaminants using a method called Generalized Morphological Component Analysis, with the goal of probing the B mode polarization power spectra to learn about primordial gravitational waves and inflation. Our results show that we were successful, and now have a module which can run GMCA on a variety of data sets. The future in this field is very exciting!! New techniques for component separation will continue to be developed, which will ultimately aid in the search for primordial gravitational waves from the B mode polarization of the CMB.

References

[**1**] *Bischoff, C. (2016, April 05). Galactic Foregrounds. Retrieved November 19, 2020, from https://www.cfa.harvard.edu/ cbischoff/foregrounds/*

[**2**] *Bobin, J., Moudden, Y., Starck, J., Fadili, J., amp; Aghanim, N. (2007, December 04). SZ and CMB reconstruction using Generalized Morphological Component Analysis. Retrieved November 19, 2020, from https://arxiv.org/abs/0712.0588*

[**3**] *Bobin, J., Starck, J., Sureau, F., amp; Basak, S. (2012, June 08). Sparse component separation for accurate CMB map estimation. Retrieved November 19, 2020, from https://arxiv.org/abs/1206.1773*

[**4**] *Bobin, J., Sureau, F., Starck, J., Rassat, A., amp; Paykari, P. (2014, January 23). Joint Planck and WMAP CMB Map Reconstruction. Retrieved November 19, 2020, from https://arxiv.org/abs/1401.6016*

[**5**] *GMCA. (n.d.). Retrieved November 19, 2020, from http://www.cosmostat.org/statistical-methods/gmca*

[**6**] *Howell, E. (2017, November 07). What Is the Big Bang Theory? Retrieved November 19, 2020, from https://www.space.com/25126-big-bang-theory.html*

[**7**] *Hu, W. (n.d.). Cosmic Expansion. Retrieved November 19, 2020, from http://background.uchicago.edu/ whu/beginners/expansion.html*

[**8**] *Liu, H., Creswell, J., Von Hausegger, S., amp; Naselsky, P. (2019, July 19). Methods for pixel domain correction of EB leakage. Retrieved November 19, 2020, from https://arxiv.org/abs/1811.04691*

[**9**] *Spiegelhalter, D., amp; Rice, K. (n.d.). Bayesian statistics. Retrieved November 19, 2020, from http://www.scholarpedia.org/article/Bayesianstatistics*

[**10**] *Wagner-Carena, S., Hopkins, M., Rivero, A., amp; Dvorkin, C. (2020, April 26). A Novel CMB Component Separation Method: Hierarchical Gener-*

alized Morphological Component Analysis. Retrieved November 19, 2020, from https://arxiv.org/abs/1910.08077