Methodology for CMB Angular Power Spectrum Extraction

Evan Biederstedt
December 18, 2014

1 Introduction

The observations from astrophysical surveys often result in signals and images containing contributions from several components or sources. This is particularly troublesome for cosmic microwave background (CMB) experiments, where the CMB signal is mixed with various astrophysical emissions (both galactic and extragalactic) and noise. As a result, much effort has been concentrated in the development of techniques to isolate each emission from all the other components present in data, including unwanted components (e.g. detector noise). The extraction of the CMB signal from the set of multifrequency observations includes getting the best possible map of the CMB with the least possible foreground contamination, but also achieving the best possible estimate of the CMB angular power spectrum, C_l. This short exposition will concentrate on the latter. The usual approach to this problem is to calculate unbiased estimators of the angular power spectrum C_l from the values of sky maps, with many techniques attempting to exploit the sparsity of the CMB power spectrum in different dictionaries (e.g. wavelets, needlets, etc.). We detail a new data-based method for extracting the true underlying CMB angular power spectrum, with the aim of developing similar techniques to extract other emissions in the maps developed by the Planck collaboration in the future.

Algorithms for statistical analysis has been very well-developed for the linear domain. However, data in the "real world" require nonlinear methods to detect and model the kind of dependencies that allow successful prediction of properties of interest. By utilizing the strong theoretical and empirical priors of the CMB power spectrum, our goal is to model the power spectrum C_l with a Gaussian process, a type of fully probabilistic Bayesian model to do nonparametric regression.

Furthermore, recall that the sky maps produced by CMB surveys like WMAP and Planck are just one realization of the underlying true angular power spectrum. Due to the limit imposed by cosmic variance, it is impossible to recover "the perfect" accurate CMB power spectrum. Due to this fundamental uncertainty in estimating the underlying C_l, this problem seems poised for Bayesian methods.

2 CMB Data Analysis

Much previous work in CMB data analysis has been focused on model testing and parameter estimation, e.g. assuming a theoretical model and then deciding if the best fit parameter values describe the data well. For cosmological purposes, the four main steps to the data analysis is (1) clean the timestream of raw data to produce time-ordered data (TOD), (2) use the TOD to create maps, (3) estimate the power spectrum from these maps, and then (4) derive cosmological parameters.

The Planck mission measures the entire sky at nine frequency bands covering $30 - 857$ GHz, with an unprecedented angular resolution ($5'$ arcmin) and sensitivity ($\Delta T/T \sim 2 \times 10^{-6}$). This will allow us to measure the CMB angular power spectrum to the highest accuracy yet, as well as exploring the properties of astrophysical phenomena inside and outside our galaxy.

Cosmological theories really only predict statistical properties of the universe. As a result, we often rely upon the definition of a power spectrum $P(k)$ to describe how much some field varies on different scales. Let's begin by considering the power spectrum for the distribution of galaxies, characterized by the inhomogeneities of the field density labeled $\delta(\vec{x})$, we can write the power spectrum $P(k)$ in the form

$$\langle \vec{\delta}(\vec{k})\vec{\delta}(\vec{k'})\rangle = (2\pi)^3 P(k)\delta^3(\vec{k} - \vec{k}^{'}) \tag{1}$$

where $\delta(\vec{k})$ denotes the Fourier transform of $\delta(\vec{x})$. The power spectrum then is the variance in the distribution, e.g. for a very smooth distribution, $P(k)$ is small, and if there are many underdense/overdense regions in our field, the power spectrum $P(k)$ will be very large. Note the dimensions of the power spectrum is $(\text{length})^3$, i.e. $(k)^{-3}$. Therefore, the dimensionless value $k^3P(k)/2\pi^2$ indicates the "clumpiness" on scale k.

CMB surveys however result in a temperature measured via two angular coordinates on the celestial, a two-dimensional field. So, one usually does not Fourier transform the CMB temperature, but expands the two-dimensional field on the surface of a sphere via spherical harmonics as a function of multipole moment l, not wave number k. Moreover, there is an angular variation in the temperature of the sky when observing the CMB. That is, our maps of the CMB do not show absolute temperature, but rather anisotropies, differences between measurements taken in different directions. By calculating the CMB angular power spectrum, you get the the power at a given angular scale—depending on the angular resolution you use, you will see a smaller/bigger contrast. We usually plot the angular power spectrum as how much temperature varies from point to point on the sky $l(l+1)C_l/2\pi$ (in units microKelvin squared, $(\mu K)^2$) versus the angular frequency, multipole l. Here the value of the multipole l is given by

$$l \simeq \frac{\pi}{\theta} \tag{2}$$

where θ is the angle a certain CMB survey feature subtends in the sky.

The distribution of matter and curvature of ΛCDM seems to be consistent with small fluctuations produced by a statistically isotropic Gaussian random process. In addition, much theoretical work also motivates the conclusion that initial fluctuations of the CMB form a Gaussian random field, namely inflation. Therefore, in the case of the Gaussian-distributed fluctuations, the angular power spectrum C_l contains virtually all the statistical information. The angular power spectrum C_l is the variance of the spherical harmonic coefficients a_{lm}, corrected for beam smearing. It represents the fluctuation power at some given angular scale, i.e. the amplitude of CMB fluctuations at said given angular extension on the sky.

For the spherical harmonic coefficients, we cannot make predictions about any particular a_{lm}, we can only make predictions about the distribution from which they are drawn. Recall the mean is $\langle a_{lm} \rangle = 0$ and the variance is written as $\langle a_{lm} a^*_{l'm'} \rangle = \delta_{ll'} \delta_{mm'} C_l$. So, for a given l, each a_{lm} has the same variance.

By utilizing angular transfer function $T(k, l)$, we can calculate the primordial power spectrum $P(k)$ from our calculated angular power spectrum C_l via

$$C_l = 4\pi \int \frac{dk}{k} T^2(k, l) P(k) \tag{3}$$

Note that this spectrum $P(k)$ depends on the cosmological parameters via scale k and is used to determine said cosmological parameters.

3 The Angular Power Spectrum

We begin by focusing on how the CMB angular power spectrum C_l is usually computed:

The most common pixelization scheme for CMB data analysis is based on HEALPix, the "Hierarchical, Equal Area, isoLatitutde Pixelization" scheme. Using HEALPix, pixels of equal area are organized in a hierarchical system with respect to resolution, with pixel centers lying on the rings of constant latitude. These properties are essential for algorithmic operations such as nearest-neighbor searches, Fast Fourier Transforms (FFTs) on our maps and fast numerical integration on the sphere.

We can represent the fluctuations over the sky in terms of spherical harmonics,

$$\frac{\Delta T}{T}(\theta, \phi) = \sum a_{lm} Y_{lm}(\theta, \phi) \tag{4}$$

i.e we can separate out the contributions of different angular scales by doing multipole expansion. For the purpose of data analysis, we often include W_l, the

window function of a user-defined FWHM (full-width, half-maximum),

$$\frac{\Delta T}{T}(\theta,\phi) = \sum a_{lm} W_l Y_{lm}(\theta,\phi) \tag{5}$$

The beam window function is defined as $W_l = \exp\big[-l(l+1)/(2l^2_{\rm beam}\big]$ for a Gaussian beam with $l_{\rm beam} = \sqrt{8\ln 2}(\theta_{\rm beam})^{-1}$.

Note that for spherical harmonics, the functions are orthonormal on the sphere as usual,

$$\int d\Omega\ Y_{lm}(\theta,\phi) Y^*_{l'm'}(\theta,\phi) = \delta_{ll'}\delta_{mm'} \tag{6}$$

.

Pixelating $\Delta T/T(\theta,\phi)$ corresponds to sampling it at $N_{\rm pix}$ locations (θ_p,ϕ_p), for $p \in [0, N_{\rm pix}-1]$. The sample function values at $\Delta T/T_p$ are then used to estimate a_{lm},

$$\hat{a}_{lm} = \frac{4\pi}{N_{\rm pix}} \sum_{p=0}^{N_{\rm pix}-1} Y^*_{lm}(\theta_p,\phi_p)\frac{\Delta T}{T}(\theta_p,\phi_p) \tag{7}$$

Using these estimated a_{lm} coefficients, we then calculate estimators $\hat{C}_l$,

$$\hat{C}_l = \frac{1}{2l+1}\sum_{m=-l}^{l} a_{lm}a^*_{lm} = \frac{1}{2l+1}\sum_{m}^{l} \|\hat{a}_{lm}\|^2 \tag{8}$$

at each l value. In summary, we use the spherical harmonic domain of the pixelized maps, estimate the values of the spherical harmonic coefficients a_{lm}, and use these values to calculate the unbiased estimator $\hat{C}_l$ for all values of multipole l, and then plot the entire angular power spectrum, temperature vs. multipole.

4 The Angular Correlation Function

Though we usually view CMB temperature fluctuations through spherical harmonics, we can also use a two-point correlation function to express the correlation between temperature perturbations at two points, i and j separated by some angle θ_{ij}. We express the relationship between the angular power spectrum and the angular correlation function as follows:

The two-point correlation function

$$C\big(\theta_{ij}\big) = \langle \frac{\Delta T}{T}(\hat{n}_i)\frac{\Delta T}{T}(\hat{n}_j)\rangle \tag{9}$$

can be viewed as in Fourier space angle on the sky is the CMB angular power spectrum. Note that $\langle \dots \rangle$ represents the average performed over all points at $\hat{n}_i$ and $\hat{n}_j$ separated by the angle $\theta_{ij} = |\hat{n}_i - \hat{n}_j|$ over the sphere.

Then, the angular power spectrum is related to the angular two-point correlation function by the following derivation:

$$\begin{aligned} C(\theta_{ij}) &= \langle \frac{\Delta T}{T}(\hat{n}_i) \frac{\Delta T}{T}(\hat{n}_j) \rangle \\ &= \sum_{lm} \sum_{l'm'} a_{lm} a^*_{l'm'} Y_{lm}(\hat{n}_i) Y_{l'm'}(\hat{n}_j) = \frac{1}{4\pi} \sum_{l} (2l+1) C_l P_l(\cos\theta_{ij}) \end{aligned}$$

where P_l denotes the Legendre polynomial of order l, $\cos\theta_{ij} = \hat{n}_i \cdot \hat{n}_j$ and θ_{ij} is the angle between pixels at position $\hat{n}_i$ and $\hat{n}_j$.

Note we relied upon the "close relation" of spherical harmonic bases Y_{lm}, i.e. by summing over the m corresponding to the same multipole number l we find

$$\sum_m \|Y_{lm}(\theta, \phi)\|^2 = \frac{2l+1}{4\pi} \tag{10}$$

5 The Likelihood Function

The first step of probabilistic inference is to write down the likelihood function. Let's say we wish to estimate the angular power spectrum C_l by explicitly maximizing the likelihood function. Working in the domain of pixels, the temperature values of observed pixels should be described as random variables following a Gaussian distribution. If the temperature fluctuations are Gaussian, then the likelihood function Pr(data | theory) is expressed as a multivariate Gaussian

$$\mathcal{L} = \Pr(\mathbf{d}|C_l) = \frac{1}{(2\pi)^{N_{pix}/2} (\det \mathbf{C})^{1/2}} \exp\Big(-\frac{1}{2} \mathbf{d}^T \mathbf{C}^{-1} \mathbf{d} \Big) \tag{11}$$

where our data vector $\mathbf{d}$ is the vector of noisy temperature map measurements $\mathbf{T}_i$ at a set of pixel locations $\hat{n}_i$, and our covariance matrix $\mathbf{C}$ is an $n \times n$ positive, symmetric matrix. The evaluation of the matrix inversion $\mathbf{C}^{-1}$ and the determinant $\det|\mathbf{C}|$ requires a computational complexity of $\mathcal{O}(n^3)$, which for large values n is computationally intractable. (In our case, we are taking n to be the number of pixels on a sky map, N_{pix}). We often expect the covariance matrix to be in the form $\mathbf{C} = \sigma^2 \mathbb{I} + \mathbf{K}$ where (as we explain below) we can express K_{ij} in terms of a kernel matrix, $K_{ij} = k(x_i, x_j)$.

Let's take the covariance matrix to be the superposition of the signal part $\mathbf{S}$ as a function of the power spectrum C_l^{theory} and the instrumental noise term, i.e. $\mathbf{C} = \mathbf{S} + \mathbf{N}$.

Our pixel-space noise covariance matrix will be written as

$$C_{ij} = N_{ij} + S_{ij} = N_{ij} + \frac{1}{4\pi}\sum_l (2l+1) C_l^{theory} P_l(\hat{n}_i \cdot \hat{n}_j) \tag{12}$$

where the noise matrix is nearly diagonal and written as

$$N_{ij} = \sigma_i^2 \delta_{ij} \tag{13}$$

where σ_i^2 is the rms noise in pixel i, [7], [8]. Similarly, the signal matrix S_{ij} is simply the angular correlation function between two pixels, and is composed of off-diagonal components generated by the cosmological model. As we show below, this covariance function S_{ij} relating the similarity between pixels at position $\hat{n}_i$ and $\hat{n}_j$ will be an example of a kernel (in fact, a Mercer kernel). Indeed, our ability to "substitute" S_{ij} as a kernel in place of K_{ij} is an example of the "kernel trick".

Assuming noise equals zero, we could write the likelihood for temperature fluctuations in full sky coverage as

$$P(\mathbf{T}|C_l^{theory}) \propto \frac{\exp\big(-(\mathbf{T}S^{-1}\mathbf{T})/2\big)}{\sqrt{\det S}} \tag{14}$$

where $\mathbf{T}$ is the temperature map and $S_{ij} = \Sigma_l (2l+1) C_l^{theory} P_l(\hat{n}_i \cdot \hat{n}_j)/4\pi$ with Legendre polynomials P_l and the pixel position on the map $\hat{n}_i$.

Covariance matrices of the form above are often found in Gaussian process models, which we will introduce below.

6 Inference Properties of Gaussians

Except sums of independent and identically distributed (i.i.d.) random variables, nothing in the real world is actually Gaussian. However, it can be argued that inasmuch as linear maps provide the natural language for algebra, Gaussians provide a distinct language for inference. Consider the multivariate Gaussian distribution: for mean vector μ and covariance matrix Σ, the joint probability density is written as

$$\mathcal{N}(\mathbf{x}; \mu, \Sigma) = \frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu)\right] \tag{15}$$

where $\mathbf{x}, \mu \in \mathbb{R}^N$ and $\Sigma \in \mathbb{R}^{N\times N}$. The covariance matrix Σ is positive semidefinite, i.e. it is Hermitian for all eigenvalues greater than zero, and $\mathbf{v}^T\Sigma\mathbf{v}$ is greater than zero for any N-dimensional vector $\mathbf{v}$.

The properties of this distribution are varied and flexible enough to do probabilistic inference:

Property 1: Gaussians are closed under multiplication,

$$\begin{aligned} \mathcal{N}(\mathbf{x}; a, A)\mathcal{N}(\mathbf{x}; b, B) &= \mathcal{N}(\mathbf{x}; c, C)\mathcal{N}(a; b, A + B) \\ \text{where } C \equiv (A^{-1} + B^{-1}) &\text{ and } c \equiv C(A^{-1}a + B^{-1}b). \end{aligned}$$

i.e. the product of two Gaussians results in another un-normalized Gaussian.

Property 2: Gaussians are also closed under linear maps, i.e. linear maps of Gaussians are Gaussians

$$\begin{aligned} p(z) &= \mathcal{N}(z; \mu, \Sigma) \\ \implies p(\mathbf{A}z) &= \mathcal{N}(\mathbf{A}z, \mathbf{A}\mu, \mathbf{A}\Sigma\mathbf{A}^T) \end{aligned}$$

for some vector $\mathbf{A}$.

Property 3: Gaussians are closed under marginalization as well. For jointly Gaussian vectors $\mathbf{x}$ and $\mathbf{y}$, we use the notation

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix} \right)$$

The marginal distribution of $\mathbf{x}$ is therefore written as

$$\int \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix} \right) \mathrm{d}y = \mathcal{N}(x; \mu_x, \Sigma_{xx})$$

From this we can conclude that every finite-dimensional Gaussian is marginal of infinitely many more Gaussians.

Property 4: Gaussians are also closed under conditioning.

$$\mathbf{x}|\mathbf{y} \sim \mathcal{N}\big(\mathbf{x}; \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(\mathbf{y} - \mu_y),\ \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}\big) \tag{16}$$

Let's re-examine the last two properties, 3 and 4. On closer examination, the third property reveals itself as a form of the "Sum Rule"

$$\int p(x, y)\mathrm{d}y = \int p(y|x)p(x)\mathrm{d}y = p(x) \tag{17}$$

which provides a definition of marginal probability. That is, marginal probabilities are sums of join probabilities integrating out all random variables except the target random variables. Furthermore, the fourth property depends on a form of the "Product Rule"

$$p(x, y) = p(x|y)p(y) \tag{18}$$

That is, joint probabilities are the products of conditional and marginal probabilities. The conditional form of "Property 4" above is written as

$$p(x|y) = \frac{p(x,y)}{p(y)} = \mathcal{N}\big(\mathbf{x}; \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(\mathbf{y} - \mu_y),\ \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}\big) \tag{19}$$

.

Therefore, we have shown that Gaussians are in fact closed under all the rules of probability. And as Baye's Rule is simply a consequence of the Sum and Product Rules, we can do Bayesian inference by focusing on processes which are Gaussian.

7 Kernels

Often times in probabilistic inference, it is difficult to know how to represent certain kinds of inputs as fixed-size feature vectors, $x_i \in \mathbb{R}^D$. There are several approaches to this difficulty in the machine learning literature, but one of the most fruitful approaches is to measure the similarity between objects without somehow preprocessing these objects as fixed-sized feature vectors. We use a "kernel function" $\kappa(x, x')$ as the similarity measure between pairs of data points. The kernel is a dot product in a (usually high dimensional) feature space. In this space, our estimation methods are usually linear, but as long as we can formulate everything in terms of kernel evaluations, we never explicitly have to compute in the high-dimensional feature space.

For inputs x and x' in some input space $\mathcal{X}$, the kernel function $\kappa(x, x') \in \mathbb{R}$ can express these arguments as an inner product in another space $\mathcal{V}$. (We normally require this function to be both non-negative $\kappa(x, x') \geq 0$ and symmetric $\kappa(x, x') = \kappa(x', x)$.) For a particular inference problem, we simply construct some "feature space" mapping $\phi : \mathcal{X} \to \mathcal{V}$ such that the kernel function is expressed as

$$\kappa(x, x') = \langle \phi(x), \phi(x') \rangle \tag{20}$$

given $\langle \cdot\ ,\ \cdot \rangle$ is an inner product in feature space $\mathcal{V}$.

For a given kernel κ with inputs $x_1, \cdots, x_n \in \mathcal{X}$, the $n \times n$ matrix

$$\mathbf{K} \equiv \big(\kappa(x_i, x_j)\big)_{ij} = \begin{pmatrix} \kappa(x_1, x_1) & \cdots & \kappa(x_1, x_N) \\ \vdots & \ddots & \vdots \\ \kappa(x_N, x_1) & \cdots & \kappa(x_N, x_N) \end{pmatrix} \tag{21}$$

is referred to as a Gram matrix. If K_{ij} is a real $n \times n$ symmetric matrix of the form

$$\sum_{i,j} c_i c_j K_{ij} \geq 0 \tag{22}$$

for all $c_i \in \mathbb{R}$, then the matrix K_{ij} is positive semidefinite.

If a Gram matrix is positive semidefinite, we refer to this as a Mercer kernel. The significance of a Mercer kernel is a result of Mercer's theorem. This states that if the kernel is Mercer, then there exists a function ϕ mapping $x \in \mathcal{X}$ to (potentially) some infinite dimensional space $\mathbb{R}^D$ such that

$$\kappa(x, x') = \phi(x)^T \phi(x') \tag{23}$$

where the mapping ϕ depends on the eigenfunctions of the kernel κ. To be more explicit, consider a positive semidefinite Gram matrix $\mathbf{K}$. We can write the eigenvector decomposition of this matrix as

$$\mathbf{K} = \mathbf{U}^T \Lambda \mathbf{U} \tag{24}$$

such that Λ is a diagonal matrix with eigenvalues $\lambda_i > 0$. If we focus on one element of the matrix $\mathbf{K}$,

$$k_{ij} = (\Lambda^{\frac{1}{2}} \mathbf{U}_{:,i})^T (\Lambda^{\frac{1}{2}} \mathbf{U}_{:,j}) \tag{25}$$

we can define the mapping as $\phi(x_i) = \Lambda^{\frac{1}{2}} \mathbf{U}_{:,i}$, which gives us

$$k_{ij} = \phi(x_i)^T \phi(x_j) \tag{26}$$

.

The importance of this property becomes apparent when we specifically construct new kernel functions to describe our priors; new Mercer kernels can be constructed from simpler Mercer kernels via a set of standard operations (e.g. the linear combination of two Mercer kernels is in fact a Mercer kernel.) As many methods require the covariance matrix to be positive semidefinite, proving your kernel is satisfies the Mercer theorem shows it is well-defined. (For a full account of constructing kernels, see [6].)

By deciding on a specific kernel, we are also choosing both the measure of similarity and the representation for our data. This is quite powerful, as it turns out many algorithms defined in terms of inner products in input space become tractable if we project inputs into this implicit higher-dimensional feature space and carry out computations there.

Furthermore, by constructing an appropriate kernel, we are choosing the covariance function for correlated observations. That is, if we have some knowledge about the relationship of our inputs of our problem (or more specifically, how our observations at different points relate to each other), we can encode this prior knowledge through our kernel. Indeed, one can think of our choice of kernel as the prior knowledge we have about a problem and its solution.

In the case of Gaussian processes, this insight is very strong, as the kernel we construct establishes how likely different functions are considered a priori. That is, we are determining the prior over an entire set of functions. In a nutshell, this is what a Gaussian process does: it defines prior distributions on functions, allowing one to perform Bayesian nonparametric regression. (By the term "nonparametric", we mean that the complexity of the solution increases with the amount of data inputs.)

8 Gaussian Process

As a conceptual crutch, think of a Gaussian process as a continuous stochastic process which directly defines a prior probability distribution over functions, i.e. an uncountably infinite set of functions. Once encountering some finite set of data, this is converted into a posterior over functions. The more data we see, the stronger this posterior becomes.

Consider a random set of variables which are indexed by some continuous variable, i.e. a function $f(x)$. Some finite subset of inputs $\mathbf{X} = \{x_1, x_2, \cdots, x_N\}$ can be expressed as a subset of random function variables $\mathbf{f} = \{f_1, f_2, \cdots, f_N\}$. In a Gaussian process, any such finite set $p(f(x_1), \cdots, f(x_N))$ is jointly Gaussian, i.e.

$$p(\mathbf{f}|\mathbf{X}) = \mathcal{N}(\mu, \mathbf{K}) \tag{27}$$

with some mean μ and the covariance matrix given by a positive semidefinte kernel function $\Sigma_{ij} = \kappa(x_i, x_j)$, i.e. a Mercer kernel. Recall via the kernel that if inputs x_i and x_j possess some similarity, the output of the function at those inputs should be similar as well, defining the correlations between different points in the process. This choice of covariance function is vital to inference, as it specifically determines the the type of sample functions drawn from the Gaussian process prior. (That is, the sample functions' shape, smoothness, amplitude, lengthscales, and so on, specified by values of the hyperparameters.) Here, kernels may be conceptualized as infinitely large positive semidefinite matrices. Refer to [5] and [4] for further details.

Predictions for regression

When using Gaussian processes for regression problems, we need to take into account of noise on the observed target values, $y = f(x) + \epsilon$, where we take $\epsilon \sim \mathcal{N}(0, \sigma_y^2)$ is independent Gaussian white noise of known variance σ_y^2.

For the sake of notational simplicity, let's begin by considering a zero mean Gaussian process prior on the function variables, $p(\mathbf{f}) = \mathcal{N}(0, \mathbf{K})$. The likelihood of the noise is written as $p(\mathbf{y}|\mathbf{f}) = \mathcal{N}(\mathbf{f}, \sigma^2\mathbf{I})$, where $\mathbf{I}$ is the identity matrix. By integrating over the unobserved function variables $\mathbf{f}$, we get the marginal likelihood

$$p(\mathbf{y}) = \int \mathrm{d}\mathbf{f} p(\mathbf{y}|\mathbf{f}) p(\mathbf{f}) = \mathcal{N}(\mathbf{0}, \mathbf{K} + \sigma_y^2\mathbf{I}) \tag{28}$$

The covariance of the observed noise is then

$$\text{cov}[y_i, y_j] = \kappa(x_i, x_j) + \sigma_y^2 \delta_{ij} \tag{29}$$

(Compare this term to the form of the covariance matrix in Section Five, $\mathbf{C} = \sigma^2 \mathbb{I} + \mathbf{K}$.)

Denoting our prediction outputs as $\mathbf{f}_*$ with a "test set" $\mathbf{X}_*$, the joint distribution of the Gaussian process has the form

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{K}_y & \mathbf{K}_* \\ \mathbf{K}_*^T & \mathbf{K}_{**} \end{bmatrix} \right)$$

again, using mean equaled to zero for notation. The posterior predictive density is then

$$\begin{aligned} p(\mathbf{f}_*|\mathbf{X}_*, \mathbf{X}, \mathbf{y}) &= \mathcal{N}(\mathbf{f}_*|\mu_*, \Sigma_*) \\ \mu_* &= \mathbf{K}_*^T \mathbf{K}_y^{-1} \mathbf{y} \\ \Sigma_* &= \mathbf{K}_{**} - \mathbf{K}_*^T \mathbf{K}_y^{-1} \mathbf{K}_*. \end{aligned}$$

For a single test input, we write this as

$$p(\mathbf{f}_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}(\mathbf{f}_*|\mathbf{k}_*^T \mathbf{K}_y^{-1} \mathbf{y},\ k_{**} - \mathbf{k}_*^T \mathbf{K}_y^{-1} \mathbf{k}_*) \tag{30}$$

where $\mathbf{k}_* = [\kappa(x_*, x_1), \cdots, \kappa(x_*, x_N)]$ and $k_{**} = \kappa(x_*, x*)$.

So, we've established that the quality of our predictions for a Gaussian process model depends on the construction of our covariance function. Recall we mentioned that the hyperparamters of the kernel control global properties like lengthscale and amplitude. Detemining the values is key to proper inference. Gaussian processes also allow us to compute hyperparamters from the training data directly; however, it is usually far too complex and/or computationally intractable to place a prior over the hyperparameters themselves and compute the posterior. Hence, we use the marginal likelihood as a cost function. So, in order to estimate these covariance hyperparameters, we simply maximize the marginal likelihood

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{f}, \mathbf{X}) p(\mathbf{f}|\mathbf{X}) \mathrm{d}\mathbf{f} \tag{31}$$

We've noted previously $p(\mathbf{f}|\mathbf{X}) = \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K})$ and we write the likelihood as $p(y|\mathbf{f}) = \prod_i \mathcal{N}(y_i|f_i, \sigma_y^2)$. The marginal likelihood is therefore written as

$$\log p(\mathbf{y}|\mathbf{X}) = \log \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K}_y) = -\frac{1}{2}\mathbf{y}\mathbf{K}_y^{-1} - \frac{1}{2}\log|\mathbf{K}_y| - \frac{N}{2}\log(2\pi) \tag{32}$$

To understand this expression, note the first term is the "data fit" term, the second term expresses the model complexity, and the third term is a constant. Varying the

values of the hyperparameters affects the tradeoff between the first two terms, e.g. if your fit is very good, the model may be too complex and vice versa.

9 Employing a Gaussian Process to CMB Sky Maps

Consider again our likelihood in Section 5, equation (11). We introduced the "signal part" of the covariance matrix S_{ij} to be the angular correlation function between two pixels i and j

$$S_{ij} = \frac{1}{4\pi}\sum_l (2l+1) C_l^{theory} P_l(\hat{n}_i \cdot \hat{n}_j) \tag{33}$$

such that the inputs are the locations of said pixels, at $\hat{n}_i$ and $\hat{n}_j$. We proceeded to claim that this function could serve as a kernel (and hence be used in our Gaussian Process model). Recall the requirements for a kernel $k(x, x')$: it should be symmetric $k(x, x') = k(x', x)$, it should be a Mercer kernel (positive semidefinite symmetric), and it should provide the suitable similarity measure between x and x' for our intended method. S_{ij} takes the inputs of two pixel positions $\hat{n}_i$ and $\hat{n}_j$, and then computes the angle between the pixels, $\cos\theta_{ij} = \hat{n}_i \cdot \hat{n}_j$. Well, the cosine function is obviously a similarity measure, and it is trivial to see that it is symmetric. Does the cosine satisfy the Mercer matrix condition? Consider a simple dot product kernel, i.e. $k(x, x') = x \cdot x'$. This is perhaps the simplest similarity measure, as we define a mapping Φ from the input vector $x \in \mathcal{X}$ into our feature space $\mathcal{H}$ such that $\Phi : \mathcal{X} \to \mathcal{H}$. This allows us to define the similarity measure for the dot product in feature space $\mathcal{H}$

$$k(x, x') = \langle \mathbf{x}, \mathbf{x}' \rangle = \langle \Phi(x), \Phi(x') \rangle \tag{34}$$

which is positive semidefinite $k(x, x') > 0$ for all $x, x' \in \mathcal{X}$. In terms of the dot product, we can write the cosine kernel as

$$k(x, x') = \frac{x \cdot x'}{|x||x'|} \tag{35}$$

Note that this is simple a scaling of the original dot product. Therefore, we can conclude that the cosine does indeed satisfy the requirements of a kernel.

As previously mentioned, when constructing kernels, there are several techniques to build new kernels from two individual valid kernels, e.g. a positive constant multiplied by a valid kernel is another valid kernel, the linear combination of two valid kernels is itself a valid kernel, and the product of two valid kernels is itself a valid kernel. Our covariance matrix S_{ij} also depends on Legendre polynomials. Via Schoelkopf and Smola (2002), a kernel defined on a sphere is positive semidefinite if and only if its expansion into Legendre polynomials P_l only contains nonnegative coefficients, i.e.

$$k(\xi) = \sum_{l=0}^{\inf} b_l P_l(\xi) \text{ with } b_l \geq 0 \tag{36}$$

As the sum of two valid kernels is also a valid kernel, we conclude that our angular correlation function kernel taking in inputs $k(\hat{n}_i, \hat{n}_j)$ is also a valid kernel.

Having a valid kernel, we can therefore try to extract the true underlying CMB angular power spectrum by fitting a Gaussian Process on all-sky maps. Given the substantial foreground noise, our first step is to use the "cleaned" CMB maps produced by WMAP and/or Planck, and employ a Gaussian Process to calculate the angular power spectrum.

The Planck collaboration has produced a set of "cleaned" temperature and polarization maps of the CMB using several different methods. Commander (based on parametric model fitting in pixel space), NILC (needlet/wavelet-based internal linear combination methods), SMICA (linearly constructed via spectral fitting and filtering), and SEVEM (produced via fitting templates in pixel space). We will begin using cut SEVEM maps with the galactic plane and point sources masked. Using the formalism introduced in Section 3 with the estimators produced by the HEALPix package, we can calculate the values of C_l for each multipole l and plot the results (the usual procedure). We can then use this "estimated" power spectrum as a comparison to check our results.

10 Summary and Future Work

We have detailed a new method to extract the CMB angular power spectrum from all-sky maps produced by CMB surveys based on Gaussian Processes. The next step is to execute similar techniques to isolate astrophysical components in the sky maps produced by Planck data, as well as separate out point sources. Planck is well-positioned to offer full-sky surveys of an extraordinarily wide frequency range of spectral regions which are normally difficult to investigate via ground-based and other more limited surveys. We shall categorize the astrophysical emissions attainable via Planck as diffuse galactic emission, extragalactic emission, and emission from the solar system. All such emissions are dependent on different multipoles l and frequency ν.

Diffuse galactic emissions stem from the interstellar medium (ISM) of the milky way. The ISM itself is composed of hot ionized regions, partly ionised regions of intercloud media, and cold diffuse clouds of molecular and atomic gas. The major contributions include free-free radiation, synchrotron emission, thermal dust, and possibly emissions in the microwave range of spinning and/or magnetic dust particles. Free-free emission results from free electron interacting with ions in an ionised medium, while synchrotron emission is due to charged energetic particles (originating

from supernovae shocks) spiraling in a magnetic field. This often generates emission at high galactic latitudes and is therefore less concentrated in the galactic plane than free-free emission or dust. Interstellar dust refers to silicate and carbon-based grains of particulates ranging in size of a few nanometers to a few micrometers. Due to the different sizes and materials, as well as different sources of energy, galactic dust should be present at a large range of temperatures (from around 5 K to over 30 K) and emissivities. Our current understanding of dust is mostly based on extinction observations in the near-IR to UV domain, as well as emission in the radio to IR domains. Dust becomes the dominant "foreground" source above the 70 GHz range, while synchrotron and free-free emission dominate at approximately 40 GHz and below.

Extragalactic emission is mostly due to clusters of galaxies at large scales and the large background of resolved and unresolved radio and IR galaxies. Any large enough body generating hot ionised gas can produce Sunyaev-Zeldovich (SZ) effects. The SZ effect is due to the inverse Compton scatting of CMB photons on free electrons in an ionised medium. The thermal SZ effect results from photon scattering in high-temperature ionised electron gas, and the kinetic SZ effect results from Thompson scattering on electrons with a global radial bulk motion in respect to the CMB. Observations of these effects grant us an opportunity to study the physics of gas condensation in cluster-size potential wells at very large scales, allowing cosmological investigations of structure formation. Extragalactic "point sources" denote emissions which are not resolved by CMB surveys, from objects such as radio galaxies, IR galaxies, and quasars.

Finally, emissions from the solar system include "local" sources such as planets, satellites, asteroids, as well as the diffuse emission of the zodiacal light. There is also a matter of subtracting the contamination cause by the monopole and dipole contributions to the all-sky maps.

References

[1] E. L. Wright, paper presented at IAS CMB Data Analysis Workshop, 1996, astro-ph/9612006

[2] M. Tegmark, Astrophys. J. 480:L87-L90, 1997, astro-ph/9611130v2

[3] P. Paykari, J. L. Starck, M. J. Fadili, 2012, arXiv:1202.4908

[4] S. Ambikasaran, D. Foreman-Mackey, L. Greengard, D. W. Hogg, M. O'Neil, 2014, arXiv:1403.6015v1

[5] C. E. Rasmussen and C. Williams, "Gaussian Processes for Machine Learning", MIT Press, 2006.

[6] B. Schoelkopf and A. J. Smola, "Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond", MIT Press, 2002.

[7] S. P. Oh, D. N. Spergel, and G. Hinshaw, ApJ, 510, 551, 1999, astro-ph/9805339v2

[8] G. Hinshaw, D. N. Spergel et al., ApJ. Suppl. 148:135, 2003, astro-ph/0302217

[9] P. Natoli, G. de Gasperis, C. Gheller, N. Vittorio, Astron. Astrophys. 372:346, 2001, astro-ph/0101252v2

EB would like to sincerely thank Professor David Hogg for providing the idea for this project, as well as an immeasurable degree of patience throughout the process so far.

Appendix 1: Map-making methodology

We begin with raw data, which must be calibrated, cleaned of "glitches" data (e.g. cosmic ray hits), and checked for systematic errors (i.e. some non-random signal found in the time-ordered data TOD not caused by the sky). So, we are left with TOD, a list of the positions and temperatures of all the pixels observed, in chronological order, in each channel. (For WMAP, Planck, the TOD consists of pairs of pixel positions and the associated temperature difference.)

We then compress the TOD with some map-making algorithm to create sky maps. The motivation behind representing the TOD in maps is to compress the data into something computationally manageable without losing any information. WMAP maps have $\sim 10^7$ pixels, Planck maps have around $\sim 10^8$ pixels. Different bands of maps have different resolutions with beams not spatially invariant or isotropic.

Here assuming a symmetric beam profile, let's review the map-making algorithm used by WMAP (using notation from [9]):

We have an $n-$dimensional vector $\mathbf{d}$ of the TOD $d_1, \ldots, d_n$, N_d sky observations with a given scanning strategy and at some given sampling rate (e.g. three points per FWHM). We use this to estimate the map $\mathbf{m}$, a vector $m_1, \ldots, m_n$ representing N_p temperature values, associated with sky pixels of dimension ∼FWHM/3 Our map is naturally pixelized and represented as a function of temperature, so x_i denotes the temperature at pixel i.

By linearity, we write the TOD vector $\mathbf{y}$ as data = signal + noise, i.e.

$$\mathbf{d}_t = \mathbf{s}_t + \mathbf{n}_t = \mathbf{P}_{tp}\mathbf{m}_p + \mathbf{n}_t \tag{37}$$

where $\mathbf{s}$ is the source, temporally constant but spatially varying, $\mathbf{P}$ is a $N_d \times N_p$ "pointing matrix" and $\mathbf{n}$ is noise, temporal.

The pointing matrix $\mathbf{P}_{tp}$ gives the weight of each pixel p in time sample t. The nonzero row values correspond to pixels being observed, at the time denoted by the row; the nonzero column entries correspond to all the times a given pixel has been observed. The matrix itself is normally quite sparse. There are usually: one nonzero entry for a total power temperature observation, two nonzero entries for a differencing temperature observation, and three nonzeros for a total power polarization observation.

Applying $\mathbf{P}$ on a map "unrolls" the map on a TOD via a given scanning strategy, and applying $\mathbf{P}^T$ on the TOD "sums" the values into a map.

How do we estimate $\mathbf{m}$? Algorithms are based on Generalized Least Squares, i.e. a multivariate Gaussian distribution is used to describe the statistical properties of detector noise (noise properties are also assumed to be piece-wise stationary), and therefore we assign a Gaussian noise likelihood function for the data time stream given the true map as

$$\mathcal{L}(d_t|m_p) = \frac{1}{|2\pi N|^{N_p/2}}\left[-\frac{1}{2}(\mathbf{d}^T - \mathbf{m}^T\mathbf{P}^T)\mathbf{N}^{-1}(\mathbf{d} - \mathbf{P}\mathbf{m}) + \mathrm{Tr}(\ln(\mathbf{N})\right]. \tag{38}$$

That is, we minimize χ^2

$$\chi^2 = \mathbf{n}^T\mathbf{A}\mathbf{n} = (\mathbf{d}^T - \mathbf{m}^T\mathbf{P}^T)\mathbf{A}(\mathbf{d} - \mathbf{P}\mathbf{m}) \tag{39}$$

such that some nonsingular, symmetric matrix $\mathbf{A}$ is the noise inverse covariance matrix, i.e. $\mathbf{A}^{-1} = \mathbf{N} \equiv \langle \mathbf{n}\mathbf{n}^T \rangle$. If we take the noise to have a zero mean $\langle \mathbf{n} \rangle = 0$, then the noise covariance matrix is $\mathbf{N} \equiv \langle \mathbf{n}\mathbf{n}^T \rangle$.

We get an estimator $\widetilde{\mathbf{m}}$ by deriving with respect to $\mathbf{m}$

$$\widetilde{\mathbf{m}} = (\mathbf{P}^T\mathbf{A}\mathbf{P})^{-1}\mathbf{P}^T\mathbf{A}\mathbf{d}. \tag{40}$$

Provided $\langle \mathbf{n} \rangle = 0$, then $\langle \widetilde{\mathbf{m}} \rangle = \mathbf{m}$.

The map covariance matrix is then written as

$$\Sigma^{-1} = \langle (\mathbf{m} - \widetilde{\mathbf{m}})(\mathbf{m}^T - \widetilde{\mathbf{m}}^T) \rangle = (\mathbf{P}^T\mathbf{A}\mathbf{P})^{-1}\mathbf{P}^T\mathbf{A}\langle \mathbf{n}\mathbf{n}^T \rangle \mathbf{A}\mathbf{P}(\mathbf{P}^T\mathbf{A}\mathbf{P})^{-1} \tag{41}$$

We find $\mathbf{A}$ that minimizes the variance of $\widetilde{\mathbf{m}}$ in order to have a so-called loise noise estimator, which is exactly $\mathbf{A}^{-1} = \mathbf{N} \equiv \langle \mathbf{n}\mathbf{n}^T \rangle$. In this case, $\widetilde{\mathbf{m}}$ of the minimum variance (linear and unbiased) estimator.

We then write the GLS solution as

$$\widetilde{\mathbf{m}} = \Sigma^{-1}\mathbf{P}^{-1}\mathbf{N}^{-1}\mathbf{d} \tag{42}$$

such that

$$\Sigma = \mathbf{P}^T\mathbf{N}^{-1}\mathbf{P} \tag{43}$$

If the detector noise distribution is in fact Gaussian, the $\widetilde{\mathbf{m}}$ is indeed the maximum likelihood estimator.

We write the map-making equations as

$$N_{pp'}^{-1} = P_{tp}^T N_{tt'}^{-1} P_{t'p'} \tag{44}$$

$$z_p = P_{tp}^T N_{tt'}^{-1} d_{t'} \tag{45}$$

$$d_p = N_{pp'} z_{p'} \tag{46}$$

resulting a number of sky maps at different frequencies. Sky maps from different channels at the same frequency are written such that each frequency map is represented as the weighted average of all the maps of the different channels at that frequency. For purely cosmological purposes, such maps retain all cosmological information from the TOD. One can therefore measure parameters just as accurately from the maps as from the TOD. (Please refer to [1] and [2] for more background.)

Appendix 2: Spherical Harmonics Convention

Recall Laplace's equation, a second-order partial differential equation written as

$$\nabla^2\phi = 0 \tag{47}$$

where ϕ denotes a scalar function. In spherical coordinates, we write this expression as

$$\nabla^2 f = \frac{1}{\rho^2}\frac{\partial}{\partial\rho}\left(\rho^2\frac{\partial f}{\partial\rho}\right) + \frac{1}{\rho^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial f}{\partial\theta}\right) + \frac{1}{\rho^2\sin^2\theta}\frac{\partial^2 f}{\partial\phi^2} = 0 \tag{48}$$

Spherical harmonics are simply the angular set of solutions to this equation such that the set Y_{lm} forms an orthogonal basis. We define Y_{lm} as

$$Y_{lm}(\theta,\phi) = \lambda_{lm}(\cos\theta)e^{im\phi} \tag{49}$$

$$\lambda_{lm}(x) = \sqrt{\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}} P_{lm}(x), \quad \text{for } m \geq 0$$
$$\lambda_{lm} = (-1)^m \lambda_{l|m|} \quad \text{for } m < 0$$
$$\lambda_{lm} = 0, \text{ for } |m| > l$$

Using $x \equiv \cos\theta$, the associated Legendre Polynomials P_{lm} solve the differential equation

$$(1-x^2)\frac{d^2}{dx^2}P_{lm} - 2x\frac{d}{dx}P_{lm} + \left(l(l+1) - \frac{m^2}{1-x^2}\right)P_{lm} = 0 \tag{50}$$

Note that P_{lm} is related to the ordinary Legendre Polynomials P_l such that

$$P_{lm} = (-1)^m (1-x^2)^{m/2} \frac{d^m}{dx^m} P_l(x) \tag{51}$$

such that $P_l(x)$ are given by the Rodrigues formula

$$P_l(x) = \frac{1}{2^l l!}\frac{d^l}{dx^l}(x^2-1)^l \tag{52}$$