# Non parametric inference with CMB data

A Preprint

**Ceccomancini** Loretucci Melnic Meo Russo

May 31, 2020

## Abstract

The cosmic microwave background (CMB) is the radiation left over from just 380000 years after the Big Bang. The fluctuation of the CMB radiation field is crucial to test Cosmological models and to have an estimate of some characteristic parameters of our Universe such as the total energy density and the dark matter density of the Universe: both can be estimated by studying the peaks of the CMB's *power spectrum*. For this reason a critical statistical question is how accurately can these peaks and the whole power spectrum be estimated: in the following study, we ask this question in a *nonparametric framework*, therefore with a minimal assumptions that are also justified from a physical and statistical point of view. The aim is to estimate the power spectrum and to build a *confidence ball* around it in a function's space that tells us how accurately the behavior of the estimated function, especially around the peaks, can be recovered starting from the WMAP satellite data. For the heteroskedastic nature of data building the confidence ball is not trivial, for this reason in our work we decided to do some approximations to facilitate the computation.

## 1 Physics Background

The Cosmic Microwave Background (CMB from now on) is the radiation left by a phase transition of the Universe which occurred about 380000 years after the Big Bang, it can be thought as a snapshot of the Universe at that age. The CMB was generated by the expansion and cooling of the Universe, in particular when the Universe, cooling down to a temperature of about 3000 K, passed from a plasma state consisting of a coupled photon-baryon fluid to a state where photons flew free through space, this is called the *recombination* period: the radiation measured today is made up of these primordial photons, now cooled to about 2.7 K, these photons are in the microwave range of the electromagnetic spectrum. The CMB is a measurement used to support the Big Bang theory and in cosmology this radiation is converted into a temperature fluctuation field which provides us information about the macro-state of each sector of the Universe 380000 years after the Big Bang.

The CMB temperature fluctuation field is very important to provide critical tests on cosmological models, in particular on the following cosmological parameters:

- $h$, the Hubble constant is the rate of the Universe's expansion;
- $\Lambda$, the cosmological constant, it acts as a negative pressure that accelerate the Universe's expansion;
- $\Omega$ is the Universe energy density, that can be divided in *baryonic density*, $\Omega_b$, *dark matter density*, $\Omega_d$, *radiation density*, $\Omega_r$.

The object of study is the CMB's *power spectrum* derived from the temperature fluctuation field, that is the difference in the local temperature from the average temperature mapped on a sphere as a function of angles:

$$Z(\theta, \phi) = \frac{T(\theta, \phi) - \overline{T}}{\overline{T}} \tag{1}$$

where $T(\theta, \phi)$ is the local temperature, $\overline{T}$ is the average temperature, $0 \leq \theta \leq \pi$ is the latitude and $0 \leq \theta \leq 2\pi$ is the longitude. In this definition $Z$ is a theoretical function defined on a sphere but, looking at the Fig.(1) where we have the WMAP satellite clean data, we can have an idea of how it should looks like.

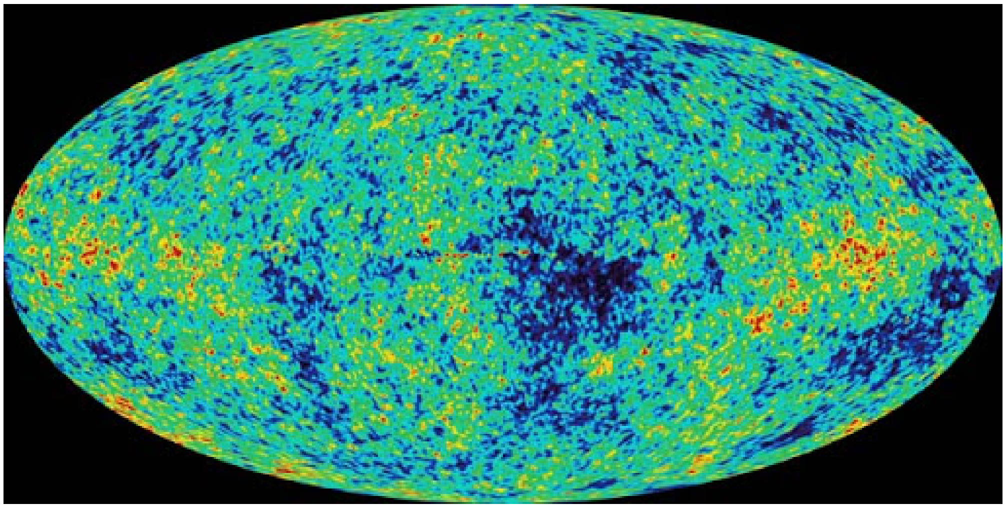

Figure 1: The CMB data collected from the WMAP satellite. The data are already cleaned by the various noise sources, such as the noise given by the presence of the Milky Way radiation.

The *power spectrum* summarizes these fluctuation as functions of angular frequency and it is featured by the presence of some *peaks*: their location and height are of critical importance because are directly related to the parameters described above. From a physical point of view, these peaks are due to areas with higher local gravitational potential that cause overdensities during the perfect fluid phase, these fluctuations in the gravitational field produced acoustic oscillations that are the reason of the peaks of the *power spectrum*. Without going into the details, we describe the physics of the peaks we are interested in:

- the **first peak** is given by the total energy density $\Omega$, because more matter and energy imply more gravitational attraction and a stronger compression, so a greater oscillation amplitude;
- the **third peak** existence provides a proof of *dark matter* existence and its amplitude is useful to estimate the fraction of dark matter in the Universe, in particular this peak could appear just if the Universe already was in a "matter-dominated era" before the *recombination*.

The aim of the work is to estimate the power spectrum and to build a *nonparametric confidence set* for it and to test if, even with minimal assumptions and without parametric assumptions, we can resolve with high accuracy the spectrum's peaks.

## 2 From regression to many normal means

The first step of the analysis is to decompose the function in Eq.(1) and the observed fluctuations into spherical modes keeping a *nonparametric framework*. We know that our data and the temperature fluctuation field live on the sphere surface; since the *spherical harmonics* are an orthonormal base on the sphere, we can use them to expand any function defined on a sphere like the one defined in Eq.(1). The set of spherical harmonics is $\{Y_{l,m}(\theta,\phi)\}$ and each of them is defined as follows:

$$Y_{l,m}(\theta,\phi) = \sqrt{\frac{2l+1}{4\pi}\frac{(l-|m|)!}{(l+|m|)!}}P_l{}^{|m|}(cos\theta)e^{im\theta} \tag{2}$$

where $l$ is the *multipole index* and $P_l{}^m(x)$ are the relative Legendre polynomials. Expanding $Z(\theta,\phi)$ via *spherical harmonics* base we can write the following exact equation:

$$Z(\theta,\phi) = \sum_{l=1}^{\infty}\sum_{m=-l}^{l} a_{l,m}Y_{l,m}(\theta,\phi) \tag{3}$$

so now we have transformed the problem of estimating the unknown function in Eq.(1) into estimating a set of coefficients $\{a_{l,m}\}$.

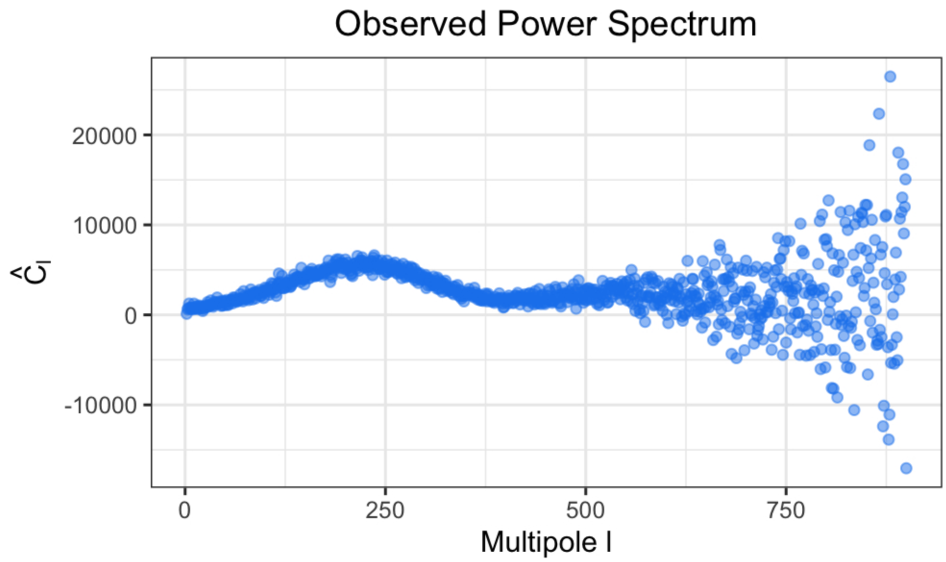


Figure 2: Observed $C_l$ as a function of the multipole $l$.

We can treat them as random variables with zero mean and we are interested in their absolute squared expected value that is defined as *power spectrum*:

$$C_l := E|a_{l,m}|^2 = \frac{1}{2l+1}\sum_{m=-l}^{l}|a_{l,m}|^2 \tag{4}$$

where $l$ is called the *multipole index*. We expect the $C_l$ to be as large as that group of spherical harmonics with same $l$ is able to capture the characteristic fluctuations of the function in Eq.(3).

What we measure for each $l$ is :

$$\hat{C}_l = C_l + \epsilon_l \qquad\qquad l = \{2, 3, ..., 900\} \tag{5}$$

where $\hat{C}_l$ is the *observed power spectrum*, $C_l$ is fixed and unknown and $\epsilon_l$ is a random variable that we assume being Normal, so $\epsilon \sim N(0, \sigma_l^2)$ with known variance.

$C_l$ is a function of $l$ and this is nothing but a regression problem, where we want to estimate an unknown function $f(l) \equiv C_l$ given some noisy observation of it. To simplify the computation we shall normalize the $l$ into the [0,1] segment, trasforming it into $x_i = l/L_{max}$.
Rewriting everything into a more familiar form we get:

$$Y_i = f(x_i) + \sigma_i\epsilon_i \tag{6}$$

where $\epsilon_i \sim N(0, 1)$.
In the [0,1] interval, assuming that our function belongs to the $L_2$ space, or more specifically to a Sobolev space, we can expand it in series using a complete orthonormal set of functions. In our example, we will use the cosine basis: $\phi_0(x) = 1$, $\phi_j(x) = \sqrt{2}\cos(j\pi x)$ for $j = \{1, 2, 3, ...\}$. This basis has some nice properties: if the function is very smooth then the coefficients decay "fast", for this reason we can approximate the real $f$ as its partial sum $f_n$:

$$f(x) \approx f_n(x) = \sum_{j=1}^{n}\theta_j\phi_j(x) \tag{7}$$

where the parameters $\theta_j$ are the Fourier coefficients given by:

$$\theta_j = \int_0^1 f(x)\phi_j(x)dx \tag{8}$$

Since $f$ belongs to a Sobolev space of order $m$ and we expanded $f$ in a cosine series, then the parameters $\theta_j$ belong to a Sobolev ellipsoid: $\Theta(m, c) = \{\theta : \sum_j a_j^2\theta_j^2 \leq c^2\}$, where $a_j^2 \sim (\pi j)^{2m}$ and the $\theta_j$ are smaller as $j$ grows by the definition of Sobolev ellipsoid.

The approximation of Eq.(7) introduces a bias in the Risk, but as shown in [2] (Lemma 8.4) , this term, for a Sobolev ellipsoid $\Theta(m, c)$ goes to zero in the worst case at least as $1/n$.

We can build the following estimator for the Fourier coefficients $\theta_j$:

$$Z_j = \frac{1}{n} \sum_{i}^{n} Y_i \phi_j(x_i) \tag{9}$$

This is nothing but a discrete form of the integral in Eq.(8) where we have put $Y_i$ instead of $f(x)$. We can evaluate the expectation of $Z_j$:

$$\mathbb{E}[Z_j] = \frac{1}{n} \sum_{i}^{n} \mathbb{E}[Y_i] \phi_j(x_i) = \sum_{i}^{n} f(x_i) \phi_j(x_i) \frac{1}{n} \approx \int_0^1 f(x) \phi_j(x) dx = \theta_j \tag{10}$$

wich tells us that we have an unbiased estimator. Instead for the variance we have:

$$\mathbb{V}[Z_j] = \frac{1}{n^2} \sum_{i}^{n} \mathbb{V}[Y_i] \phi_j^2(x_i) = \frac{1}{n^2} \sum_{i}^{n} \sigma_i^2 \phi_j^2(x_i) = \sigma_{j,n}^2 \tag{11}$$

It can be shown that $Z$ is asymptotically normal :

$$Z \approx N_n(\theta, \Sigma_n) \tag{12}$$

At this point we can switch our focus from the regression problem in Eq.(6) to the *many normal means* framework, where we want to estimate the mean vector of a normal random vector given one observation of it, with known $\Sigma_n$ Eq. (12).

The easiest estimator for $\theta$ is given by $\hat{\theta} = Z$, that is the MLE, but we can do better in terms of the MSE risk. Assuming that the vector belongs to some specific space of sequence, for example an $l_2$ ball, it can be proved that the MSE risk in this class of function can be lower. More precisely, if we estimate $\theta$ as $Z$, the risk is $R(Z, \theta) = \sum_j \sigma_{j,n}^2$ (just $\sigma^2$ in case the variances are all the same) but assuming that it belongs to the $l_2$ ball, where $\Theta(c) : \{\theta : \sum \theta_j^2 \leq c^2\}$, it can be shown that the minimax risk, for the case where the variances are all the same, is:

$$\liminf_{n \to \infty} \inf_{\hat{\theta}} \sup_{\theta \in \Theta(c)} R(\hat{\theta}, \theta) = \frac{\sigma^2 c^2}{\sigma^2 + c^2} \tag{13}$$

In particular we restrict our attention to the Sobolev ellipsoid space of parameters where it can be shown that the minimax risk goes asymptotically to zero with a rate of $n^{-2m/(2m+1)}$ (Pinsker's theorem for Sobolev ellipsoids, [2] 7.32 ).
In general, if we make some assumptions on the space in which the parameters belong we can improve our result in estimating the mean vector of Eq. (12). In spaces of functions where the elements have spatially homogeneous smoothness, we can achieve the minimax risk using a particular class of estimators that will be introduced in the next chapter.

## 3 Linear estimators

Now we want to introduce *linear estimators* in our problem to shrink closer to zero all the $Z_j$ that are not so important in the *power spectrum* description and because we will show that, considering *linear estimators*, we are not losing generality, at least in terms of minimax risk.
We will focus our attention on this type of estimators, i.e. the ones in the form: $\hat{\theta}_j = w_j Z_j$, for all $j$. These estimators achieve the minimax risk on the class of parameters of the ellipsoids, i.e. $\Theta(c) = \{\theta : \sum a_j^2 \theta_j^2 \leq c^2\}$, with $a_j \to \infty$, that is just the generalization of the class of Sobolev ellipsoids. In particular, we know from the Pinsker's theorem for ellipsoids [2] 7.36 that the general minimax risk, $R_n = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} R(\hat{\theta}, \theta)$, and the minimax risk evaluted just on the linear estimators space $L$, $R_n^L = \inf_{\hat{\theta} \in L} \sup_{\theta \in \Theta} R(\hat{\theta}, \theta)$, are close to each other as $n \to \infty$: so we are restricting our research on the linear estimators space but, at least asymptotically, the theorem ensures we are not losing generality in terms of minimax risk.
The $w$ is called *modulator*, it's a vector that allows us to build a different linear estimators for each estimator $Z_j$ that we

have. In particular, we use a specific case of monotone modulators that are the *nested subset selection (NSS) modulators* class $M_{NSS}$, where the $w$ vector is of the form: $w = (1, .., 1, 0, ...0)$. With this choice of modulator we can write our power spectrum function as following:

$$\hat{f}(x) = \sum_{j=1}^{J} Z_j \phi_j(x) \tag{14}$$

where J represents the index of the last non zero coefficient of $w$.

Moreover, for this class of estimators we can rewrite the empirical risk as:

$$\hat{R}_J(\hat{\theta}) = \sum_{j \leq J} \sigma_{j,n}^2 + \sum_{j > J} (Z_j^2 - \sigma_{j,n}^2)_+ \tag{15}$$

where the first sum is associated to the one terms in $w$, while the second sum to the zero terms in $w$. We shall notice that Eq. (14) is just the modified SURE. We should find the $\hat{J}$ that minimizes the risk in Eq. (15) and define our estimate of $f$ the estimator in Eq.(14) with that $J$.

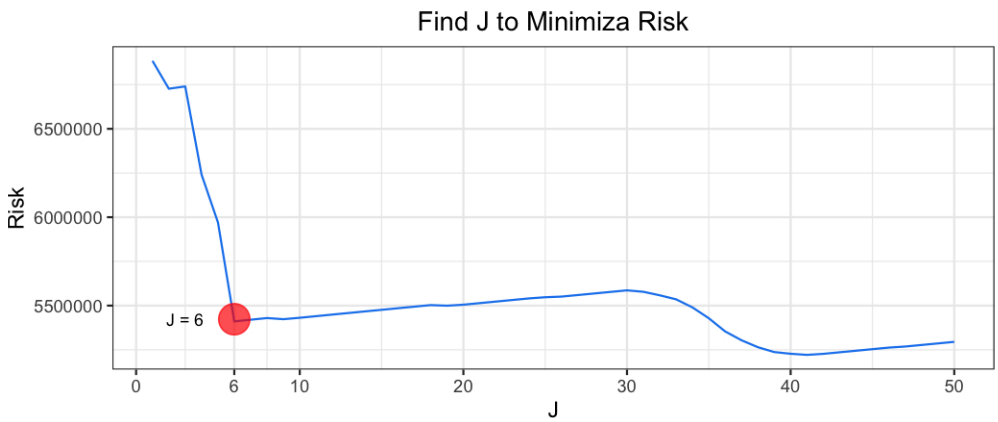


Figure 3: Plot of the NSS risk. It is possible to see the first local minimum at $J = 6$. The point $J = 40$ gives a function that is too "wiggly" and it is a property that we do not expect from the real function.

Beran and Dümbgen proved in 1998 that the value of $J$ that minimizes the empiricial risk Eq. (15) is asymptotically the one that minimizes the real risk: $|R(\hat{J}) - R(J^*)| \to 0$, where $J^*$ is the $J$ that minimizes the real risk.

## 4 Confidence bands

From the estimate $\hat{f}$ of some function $f$ we would like to estimate the confidence bands for it, that are usually of the form: $\hat{f}(x) \pm c\hat{se}(x)$, where $\hat{se}(x)$ is the estimated standard error for $\hat{f}$ and $c$ is some constant greater than zero. In trying to do so we get into the *bias problem*.
When trying to build the confidence bands we get:

$$\frac{\hat{f}(x) - f(x)}{\hat{se}(x)} = \frac{\hat{f}(x) - \overline{f}(x)}{\hat{se}(x)} + \frac{\overline{f}(x) - f(x)}{\hat{se}(x)} = Z(x) + \frac{bias(\hat{f})}{\sqrt{\mathbb{V}(\hat{f})}} \tag{16}$$

where $Z$ converges usually to a standard normal from which we derive the confidence intervals and $\overline{f}$ is the expectation of $\hat{f}$. Since in non parametric inference the estimators are usually endowed with some bias and variance the second term of the right side of Eq. (16) never goes to zero and for this reason the confidence band is never centered around the true function $f$ but is centered around $\overline{f}$.
Since we are estimating the function $f$ "indirectly", the bias problem doesn't show up (asymtotically) because we are going to estimate confidence balls around the vector $\theta$.
Given the parameter $\theta$ we want to estimate a confidence set for it, namely a confidence set $B_n = \{\theta : ||\theta - \hat{\theta}||^2 \leq r_n^2\}$ at a level of confidence $1 - \alpha$ is defined as:

$$\inf_{\theta \in \mathbb{R}^n} P(\theta \in B_n) \geq 1 - \alpha \tag{17}$$

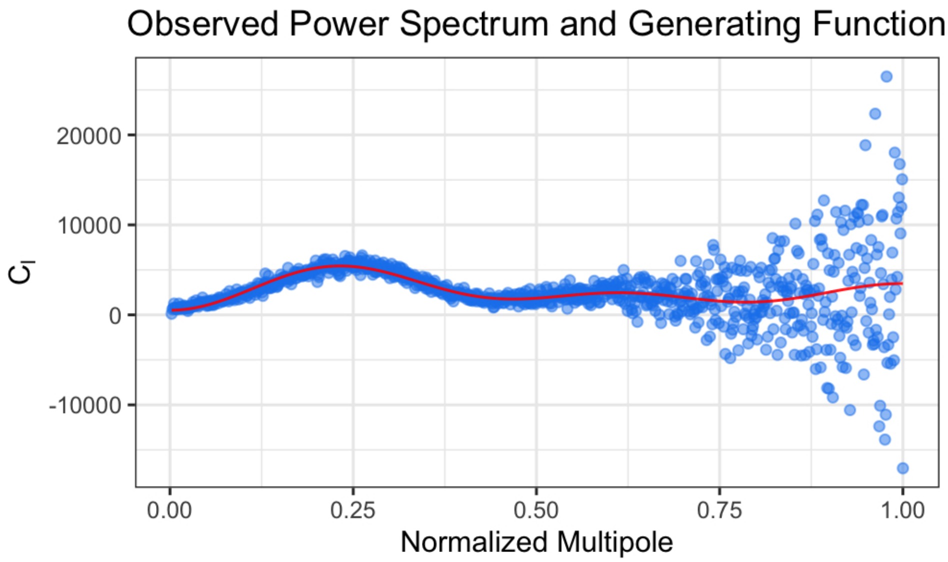


Figure 4: Estimated function $\hat{f}(x)$ with $\hat{J} = 6$ as described in Eq. (14). The first peak is easy to identify whereas the second and third one can not be resolved easily.

In particular there are several ways to compute such sets, we will focus on the Beran-Dümbgen-Stein pivotal method [2] 7.8.
Let $\hat{\theta}$ be a NSS estimator for $\theta$, $L_J = ||\hat{\theta} - \theta||_2^2$ the loss function, $\hat{R}_J$ the Stein's unbiased risk and define as *pivot*:

$$V_J = \sqrt{n}(L_J - \hat{R}_J) \tag{18}$$

it can be shown that:

$$\frac{V_{\hat{J}}}{\hat{\tau}_{\hat{J}}} \approx N(0, 1) \tag{19}$$

where $\hat{\tau}^2$ is the estimator of $\mathbb{V}(V_J)$ and $\hat{J}$ is the index that minimizes the NSS risk as we saw in Eq. (15).
This result can be used to estimate a confidence ball for $\theta$. Let $r_n^2 = \hat{R}_{\hat{J}} + \frac{\hat{\tau} z_\alpha}{\sqrt{n}}$ and define the confidence ball:

$$B_n = \{\theta \in \mathbb{R}_n : ||\hat{\theta} - \theta||_2^2 \leq r_n^2\} \tag{20}$$

it can be shown that:

$$\mathbb{P}(\theta \in B_n) = \mathbb{P}\left(L_J \leq \hat{R}_{\hat{J}} + \frac{\hat{\tau} z_\alpha}{\sqrt{n}}\right) = \mathbb{P}\left(\frac{V_{\hat{J}}}{\hat{\tau}_{\hat{J}}} \leq z_\alpha\right) \to 1 - \alpha \tag{21}$$

In our case the errors we have for the parameters are all different, they are low for low $l$s and high for greater $l$s. We won't take into account this difference in errors and build confidence balls in a computationally simplified way. From the last assumption , defining $\sigma$ as the average of all the $\sigma_i$, we can derive an expression for $\hat{\tau}^2$:

$$\hat{\tau}^2 = \frac{2\sigma^4}{n} \sum_j [(2\hat{w}_j - 1)(1 - c_j)]^2 + 4\sigma^2 \sum_j \left(Z_j^2 - \frac{\sigma^2}{n}\right) [(1 - \hat{w}_j) + (2\hat{w}_j - 1)c_j]^2 \tag{22}$$

where $c_j$ is 0 if $j \leq n - \hat{J}$ and $1/\hat{J}$ otherwise.

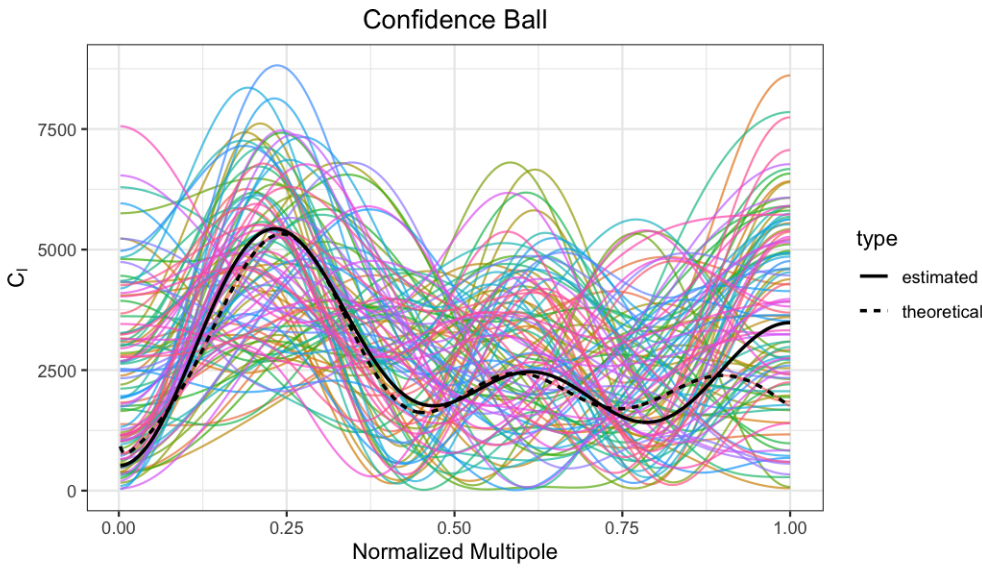


Figure 5: Confront between the estimated function, the theoretical power spectrum for some cosmological parameters generated through the Python CAMB package and some functions that lie in the confidence set. Confronting the estimated and theoretical model we can see that for the first 2 peaks the estimation is good but not for the third one.

## Appendix A : Practical Implementation R Code

Data and `.Rmd` file are available at the GitHub repository [4].

```r
# Import libraries ---------------------------------------------------------

library(latex2exp)
library(ggplot2)
library(tidyverse)
library(gridExtra)
theme_set(theme_bw())

# Data import and manipulation ---------------------------------------------

# Import
data <- read.csv("space.txt", sep="")
str(data)

# Show data:
ggplot( data, aes( x = ell, y = Cl ) ) +
  geom_point(colour = 'dodgerblue2', alpha = 0.5) +
  labs(x = 'Multipole␣l', y = TeX('$\\hat{C_l}$')) +
  theme(plot.title = element_text(hjust = 0.5)) +
  ggtitle('Observed␣Power␣Spectrum')

# Data manipulation
l = data$ell
L.max = max(data$ell)
x = l/L.max
y = data$Cl
se = data$se
n = nrow(data)
```

```
# Nonparametric regression and Risk Optimization ——————————————

# cosine basis:
cos.j = function(j,x){
  1*(j==0) + sqrt(2)*cos(pi*j*x)*(j>0)
}

# Fourier parameters estimation (our Z_j):
z = rep(NA, n)
for(i in 1:n){
  z[i] = 1/n*sum(y*cos.j(i - 1,x))
}

# standard error of
z_se = rep(NA, n)
for (j in 1:n){
  z_se[j] = sqrt( sum( se**2 * cos.j( j - 1 , x)**2 ) / n**2 )
}

# NSS Risk
risk = function(J){
  R = 0
  for (i in 1:n){
    tmp = z[i]**2 - z_se[i]**2
    R = R + (i <= J)*(z_se[i]**2) + (i > J)*(tmp > 0)*(z[i]**2 - z_se[i]**2)
  }
  return(R)
}
risk_vec = Vectorize(FUN = risk, vectorize.args = 'J')

# Risk optimization:
J.seq = seq(1,50)
ggplot() +
  geom_line(aes(x = J.seq, y = risk_vec (J.seq) ), colour = 'dodgerblue2') +
  annotate("point", x = 6, y = 5423000, colour = "red", size = 7, alpha = 0.7) +
  annotate("text", x = 3.5, y=5423000, label = "J = 6") +
  scale_x_continuous(breaks=c(0, 6, 10, 20, 30, 40, 50)) +
  labs(x = 'J', y = 'Risk') +
  theme(plot.title = element_text(hjust = 0.5)) +
  ggtitle('Find J to Minimiza Risk')

# Estimate values:
J.min = 6
sum.cos = 0
for (i in 1:J.min){
  sum.cos = sum.cos + z[i]*cos.j(i-1, x)
}

# Plot the result:
ggplot() +
  geom_point(aes(x = x, y = y), col = 'dodgerblue2', alpha = 0.5) +
  geom_line(aes(x = x, y = sum.cos), col = 'red', size = 0.5, alpha = 0.9) +
  theme(plot.title = element_text(hjust = 0.5)) +
  ggtitle('Observed Power Spectrum and Generating Function') +
  labs(x = 'Normalized Multipole', y = TeX("C_l" ))
```

```
# Confidence ball ————————————————————————————————————————————

mean_se = mean(se)
mean_se_vec = rep(mean_se, n)
J_hat = 6

# Calculate variance of V (tau^2)
first_term = n - J_hat + J_hat * (1-1/J_hat)**2
second_term = sum( z[ (J_hat + 1) : (n - J_hat) ]**2 - 1/n * mean_se_vec[ (J_hat + 1) : (n
  sum(z[ (n - J_hat + 1) : n]**2 - 1/n * mean_se_vec[ (n - J_hat + 1) : n]**2) * (1 - 1/J_
tau_hat.2 = 2 * mean_se**4/n * first_term + 4*mean_se**2 * second_term

# Calculate the radius:
alpha=0.05
radius.2 = risk(J_hat) + sqrt(tau_hat.2) * qnorm(1 - alpha)/sqrt(n)
radius = sqrt(radius.2)

c('radius:', round(radius, 2))

# Import the theoretical functions value:
conco <- read.csv("conco.csv")

# Get only the first 899 values:
conco = conco$X0
conco = conco[2:900]

# Create a dataframe with the values of each function generated:
func_df = tibble(x = x, y = sum.cos, type = 'estimated')
func_df = rbind( func_df, cbind( cbind( x = x, y = conco ), type = 'theoretical'))
func_df[,1:2] = sapply(func_df[, 1:2], as.double)

theta.hat = z[ 1:J_hat ] # we consider the first six, because of J hat
# A matrix containing in each rows the parameters for a function in the ball
B = 100 # Number of functions to estimate
generated_parameters = matrix(data = NA, nrow = B, ncol=length(theta.hat))
count = 1
for (b in 1:B){

  random.parameters = rep(NA, length(theta.hat)) # an array containing the generated param

  # Clean the coeffcients, keep them only under some conditions:
  check = TRUE
  while (check){

    # Generate the parameter by a uniform distribution, trying to remain the ball.
    for (i in 1:length(theta.hat)){
      random.parameters[i] = runif(1, min = theta.hat[i]
- 0.5 * radius, max = theta.hat[i] + 0.5 * radius )
    }

    # Chcke if the parameters are in the ball:
    diff.thetas = sum((theta.hat - random.parameters)**2)
    if (diff.thetas < radius.2){

      # Check if the function is positive in all points:
      sum.cos = 0
      for (i in 1:6){
        sum.cos = sum.cos + random.parameters[i] * cos.j(i - 1, x)
```

```
        }
        if( sum( sum.cos >= 0 ) == n ){

          # Save the parameters:
          generated_parameters[b, ] = random.parameters
          # Add the function values in the df:
          func_df <- add_row(func_df, x = x, y = sum.cos,
                             type = paste('function', as.character(count), sep ='_') )
          check = FALSE
          count = count + 1
        }
      }
    }
}

# Plot the result:

ggplot(func_df, aes(x = x, y = y, colour = type)) +
  geom_line(alpha = 0.5) +
  geom_line(data = func_df[func_df$type == 'estimated' | func_df$type == 'theoretical', ],
linetype = type), size = 1.2, colour = 'black') +
  theme( plot.title = element_text(hjust = 0.5)) +
  guides(colour = FALSE) +
  ggtitle('Confidence Ball') +
  xlab('Normalized Multipole') +
  ylab( TeX("C_l" ) )
```

## References


[1] Christopher R. Genovese, Christopher J. Miller, Robert C. Nichol, Mihir Arjunwadkar and Larry Wasserman. Nonparametric Inference for the Cosmic Microwave Background. In *Statistical Science 2004*, Vol. 19, No. 2, 308–321 DOI 10.1214/088342304000000161.

[2] Larry Wasserman. *All of non parametric statistics.*

[3] Link to the CAMB package for Python.

[4] Link to the GitHub repository with the data and .Rmd code.