11

ANALYSIS

Increasingly, theorists and even busy experimentalists are turning their attention to the fundamental question of how best to analyze a set of data. The main reason for this focus is that the quality and quantity of data have improved dramatically over the past decade. There is every reason to believe that this trend will continue. Anisotropies in the temperature of the CMB have been measured by dozens of experiments already. The satellites MAP and Planck will take these measurements to the next stage, but there is no reason to think this will be the last stage. There are still polarization and very small scale anisotropies to be measured. The power spectrum of matter is probed in a variety of ways; activity here, too, shows no sign of letting up. After the completion of the Sloan Digital Sky Survey and 2DF, the two largest redshift surveys to date, surveyors have begun planning large weak lensing missions and even deeper galaxy surveys. These larger data sets create new challenges in analysis.

A wonderful/disturbing example of these challenges was given recently by Julian Borrill. He used scaling arguments to show that a brute-force algorithm for making a map from the raw data of the Boomerang CMB anisotropy experiment would take 12 years to run on current computers! Already, data sets are far too large for brute-force calculations. And things are rapidly getting more dire. Since typically the number of arithmetic operations scales as the number of pixels cubed, and since MAP and Planck will have of order 10 to 100 times more data than Boomerang, the situation cries out for creative solutions.

Another reason for the recent focus on analysis is one I hope to convey in this chapter: analysis is exciting. The techniques that have been proposed to deal with the complexity of forthcoming data sets are beautiful. The elegance of these techniques is of course enhanced by their importance. But the elegance is there; for this reason alone, it is well worth working through some of these recent advances.

11.1 THE LIKELIHOOD FUNCTION

The basic building block of contemporary analysis is the likelihood function. This is defined as the probability that a given experiment would get the data it did given a theory. This seemingly simple definition is exceedingly powerful. Once we have the likelihood function, with a caveat or two, we can determine the parameters of the theory (best estimate is the place in parameter space where the likelihood function is largest) along with errors (determined by the width of the likelihood function). We start with a simple example and move on to the likelihood function for the CMB and then a galaxy survey.

11.1.1 Simple Example

Suppose you want to weigh somebody. Since you are a scientist, you know that, in addition to the measurement, you should also report an uncertainty. So you set up 100 different scales and record the person's weight on each of these different scales. Given these 100 numbers, what value should you report for the weight and the uncertainty in the weight? We all know the answer to this question, so let's introduce the formalism of the likelihood function in this simple context.

The likelihood function is the probability of getting the hundred numbers given a theory. Our theory will be that each measurement is the sum of a constant signal (the person's weight) w and noise, with the noise drawn from a Gaussian distribution with mean zero and variance σ_w^2. Thus our "theory" has two free parameters, w and σ_w. If only one data point d was taken, the probability of getting d given the theory would be

$$P[d|w,\sigma_w] \equiv \mathcal{L}(d;w,\sigma_w) = \frac{1}{\sqrt{2\pi\sigma_w^2}} \exp\left\{-\frac{(d-w)^2}{2\sigma_w^2}\right\}. \tag{11.1}$$

Here and throughout, $P[x|y]$ denotes the probability of x given y. Equation (11.1) simply restates the assumptions that $d-w$ is equal to noise and that the noise is drawn from a Gaussian distribution with standard deviation σ_w. In the limit that σ_w becomes very small, this function becomes sharply peaked at $d=w$. Since we are making $N_{\rm m}=100$ independent measurements, the likelihood function is the product of all the individual likelihood functions. That is,

$$\mathcal{L} = \frac{1}{(2\pi\sigma_w^2)^{N_{\rm m}/2}} \exp\left\{-\frac{\sum_{i=1}^{N_{\rm m}}(d_i-w)^2}{2\sigma_w^2}\right\}. \tag{11.2}$$

Notice that, although the data are distributed as a Gaussian, the likelihood function is not Gaussian in all the theoretical parameters (it is in w but is not in σ_w).

We are interested in the value of the theoretical parameters w and σ_w. Thus, we don't want $P[\{d_i\}|w,\sigma_w]$, which is the likelihood function we have computed. Rather, we want $P[w,\sigma_w|\{d_i\}]$. To obtain the latter from the former we can use a simple relation from elementary probability theory,

$$P[B\cap A] = P[B|A]P[A]$$

$$= P[A|B]P[B]. \tag{11.3}$$

In this context $A = \{d_i\}$ and $B = \{w, \sigma_w\}$, so the equality between the two lines of Eq. (11.3) means that

$$P[w, \sigma_w|\{d_i\}] = \frac{P[\{d_i\}|w, \sigma_w]\ P[w, \sigma_w]}{P[\{d_i\}]}. \tag{11.4}$$

The denominator can be rewritten by realizing that when we integrate the probability $P[w, \sigma_w|d]$ over all values of the parameters w, σ_w, we must get 1. So the denominator is equal to the integral of the numerator over w, σ_w. As a result, the denominator does not depend on the parameters w and σ_w, so it does not affect the place in parameter space where the likelihood function peaks or the width of the likelihood function. For the most part, then, we are free to ignore it.

To get the probability we want we need the likelihood function — the first term in the numerator — and also the *prior* probability $P[w, \sigma_w]$. If we possess prior information about these quantities, we might use this information here. If we want to be conservative, and assume nothing, we put in a uniform prior for the parameters. Then,

$$P[w, \sigma_w|\{d_i\}] \propto \mathcal{L}, \tag{11.5}$$

the proportionality constant being independent of the parameters and therefore of little interest. Many people find this idea of using prior information disturbing. Indeed, even the conservative choice of a uniform prior is not as innocent as it sounds. If we had taken the parameter to be σ_w^2 instead of σ_w and we had assumed a prior uniform in σ_w^2 (i.e., that equal intervals of σ_w^2 are equally likely), we would get a different answer for the final probability (try it!). Nonetheless, the dependence on the prior is a problem only in cases where the data are not very discriminatory. If the data do have discriminatory power, then the likelihood function $P[\{d_i\}|w, \sigma_w]$ will be sharply peaked and any reasonable prior will not affect the final results.

We can now find best-fit values for our parameters w and σ_w. Simply find the place in parameter space where the likelihood function is largest. In this simple example, we can proceed analytically by differentiating $\mathcal{L}$ with respect to each of the parameters. First consider the derivative with respect to w.

$$\frac{\partial \mathcal{L}}{\partial w} = \frac{\sum_{j=1}^{N_m}(d_j - w)}{\sigma_w^2(2\pi\sigma_w^2)^{N_m/2}} \exp\left\{-\frac{\sum_{i=1}^{N_m}(d_i - w)^2}{2\sigma_w^2}\right\}. \tag{11.6}$$

For this derivative to be zero, we set the prefactor

$$\sum_{i=1}^{N_m}(d_i - w) = 0 \tag{11.7}$$

or equivalently, the likelihood is a maximum when

$$w = \bar{w} = \frac{1}{N_m}\sum_{i=1}^{N_m} d_i, \tag{11.8}$$

the expected answer. Similarly, we can find what the most probable value of σ_w^2 is by computing

$$\frac{\partial \mathcal{L}}{\partial \sigma_w^2} = \mathcal{L}\left[-\frac{N_{\rm m}}{2\sigma_w^2} + \frac{\sum_{i=1}^{N_{\rm m}}(d_i - w)^2}{2\sigma_w^4}\right] \tag{11.9}$$

and setting it equal to zero. Solving for the variance σ_w^2, we find a most probable value of

$$\sigma_w^2 = \frac{\sum_{i=1}^{N_{\rm m}}(d_i - \bar{w})^2}{N_{\rm m}}, \tag{11.10}$$

again the expected result.

We have found the best-fit values of our theoretical parameters. What is the error in these best-fit values? The error is just proportional to the width of the likelihood function. A simple way to approximate the width is to assume that $\mathcal{L}$ is Gaussian in the parameters, or equivalently that $\ln \mathcal{L}$ is quadratic in the parameters. We know in general that the variance of a Gaussian distribution is twice the inverse of the coefficient of the quadratic term, so we can simply identify the variance (the square of the error) by computing this coefficient. Let's work this out explicitly for w:

$$\begin{aligned}\ln \mathcal{L}(w) &= \ln \mathcal{L}(\bar{w}) + \frac{1}{2}\frac{\partial^2 \ln \mathcal{L}}{\partial w^2}\bigg|_{w=\bar{w}}(w - \bar{w})^2 \\ &= \ln \mathcal{L}(\bar{w}) - \frac{N_{\rm m}}{2\sigma_w^2}(w - \bar{w})^2. \end{aligned} \tag{11.11}$$

Thus the width of the likelihood function at its maximum is $\sigma_w/N_{\rm m}^{1/2}$. This is the one-sigma error in our determination of w. This too is familiar: as more measurements are taken, the noise gets beaten down by a factor of 1 over the square root of the number of independent measurements. It is important to reiterate that the uncertainty on our estimate of the weight is *not* equal to σ_w.

Two numbers then sum up all $N_{\rm m}$ measurements: our best guess for the person's weight — in this case $\bar{w}$ given by Eq. (11.8) — and the error on this estimate, here equal to $\sigma_w/N_{\rm m}^{1/2}$. Therefore, we can compress all 100 measurements into just two by rewriting the likelihood function as

$$\mathcal{L} = \frac{1}{\sqrt{2\pi C_N}} \exp\left\{\frac{-(w - \bar{w})^2}{2C_N}\right\} \tag{11.12}$$

where the variance due to noise is now

$$C_N = \frac{\sigma_w^2}{N_{\rm m}}. \tag{11.13}$$

This form of the likelihood has precisely the same maximum and width as does the form with all $N_{\rm m}$ data points. Thus it has compressed all the information in the likelihood function into two numbers, $\bar{w}$ and C_N.

Moving away from the weight metaphor, we can apply the above to the CMB. Instead of a true weight, the signal s is the true CMB temperature at a given point on the sky. The many measurements of the signal correspond to many measurements of the temperature at that point. The signal at that spot is a constant, and the data are the sum of the constant signal plus noise (atmospheric, instrumental, or both). The compression of all the different measurements into one data point with an associated error as in Eq. (11.12) is called map-making. We will take this up in more detail in Section 11.5. Now, though, we must move beyond the likelihood of Eq. (11.12). For we know that no theory predicts a value for the temperature at a particular position on the sky; i.e., no theory predicts s. Rather all theories predict a distribution of temperatures, from which s at a given pixel is drawn. We must now incorporate this distribution into the likelihood function.

11.1.2 CMB Likelihood

Let's convert the notation of the previous subsection to the CMB. The true temperature anisotropy in a given spot on the sky s replaces w, while the data point $\bar{w}$ (really the average of many measurements) becomes the estimated value of this temperature anisotropy, call it Δ.[1] The variance of this estimator C_N, which represents the spread of the measurements, is also given. How can this set of data (Δ, C_N) be compared with theory? The simplest theories, such as inflation, predict that the signal in a given spot on the sky is drawn from a Gaussian distribution. So the probability that the sky temperature falls in a range between s and $s+ds$ is

$$P(s)ds = \frac{1}{\sqrt{2\pi C_S}} \exp\left\{\frac{-s^2}{2C_S}\right\} ds. \tag{11.14}$$

Here C_S is the variance expected due to the signal alone, independent of any noise. This variance is directly related to the C_l's in a manner we will explore in Section 11.2.

In order to get the likelihood function, we have to convolve the probability distribution of Eq. (11.14) with the likelihood function of Eq. (11.12). Schematically

$$P\left[\Delta|C_S\right] = \sum_s P\left[\Delta|s\right] P\left[s|C_S\right]. \tag{11.15}$$

More concretely, the likelihood function is an integral over all possible values of the true anisotropy:

$$\mathcal{L} = \int_{-\infty}^{\infty} \frac{ds}{\sqrt{2\pi C_S}} \exp\left\{\frac{-s^2}{2C_S}\right\} \frac{1}{\sqrt{2\pi C_N}} \exp\left\{\frac{-(\Delta - s)^2}{2C_N}\right\}. \tag{11.16}$$

The argument of the exponential here is quadratic in s so it is straightforward to carry out the integration over s. Let us rewrite the argument of the exponential as

[1] In Section 11.5 we explore ways to go from the raw data, the timestream, to the pixelized map represented by Δ. For now, we assume that this step has already been taken.

$$-\frac{s^2 C}{2C_S C_N} + \Delta s/C_N - \frac{\Delta^2}{2C_N} = -\frac{C}{2C_S C_N}\left[s - \frac{C_S \Delta}{C}\right]^2 + \frac{C_S \Delta^2}{2CC_N} - \frac{\Delta^2}{2C_N} \quad (11.17)$$

where the full covariance matrix is defined as

$$C \equiv C_S + C_N. \quad (11.18)$$

Changing variables in the integral over s to $x \equiv s - C_S\Delta/C$ leads to

$$\mathcal{L} = \frac{1}{\sqrt{2\pi C_N}} \exp\left\{\frac{C_S \Delta^2}{2CC_N} - \frac{\Delta^2}{2C_N}\right\} \int_{-\infty}^{\infty} \frac{dx}{\sqrt{2\pi C_S}} \exp\left\{-\frac{Cx^2}{2C_S C_N}\right\}$$

$$= \sqrt{\frac{1}{2\pi C}} \exp\left\{\frac{-\Delta^2}{2C}\right\}. \quad (11.19)$$

This is our final expression for the likelihood function for a one pixel experiment. This form is exactly what one expects: the measured temperature should be distributed like a Gaussian with a variance given by the sum of the variances due to noise and signal.

We can easily generalize Eq. (11.19) to the more realistic case with a measurement of N_p pixels. Then the likelihood function is

$$\mathcal{L} = \frac{1}{(2\pi)^{N_p/2}(\det C)^{1/2}} \exp\left\{-\frac{1}{2}\Delta C^{-1}\Delta\right\} \quad (11.20)$$

where now Δ is the data vector consisting of all N_p measurements and C is the full covariance matrix. In general, the noise covariance matrix can often be close to diagonal, but the theoretical covariance matrix is not diagonal. Thus, the hard computational part of evaluating the likelihood function is taking the determinant and the inverse of the $N_p \times N_p$ matrix C. The passage from one theoretical parameter (C_S) to a full matrix of parameters creates complications besides the computational. If there was only parameter, observers could quote results in the form of one number. Now that all correlations need to be included, one needs to allow for many different theoretical parameters, in principle all $N_p(N_p + 1)/2$ elements of the (symmetric) covariance matrix C_S. In practice of course this is not done. First of all, the covariance in all theories depends on the angular distance between two points, so elements of the matrix corresponding to two sets of points separated by the same distance are identical. Equivalently, a given theory is associated with a full set of C_l's; as we will see shortly, these can be used to construct all the elements of C_S. The second simplification is that most experiments have not been sensitive to individual C_l's but rather to the average power over a range of l, i.e., in a given *band*. So, analysts typically fit for *bandpowers*, a fitting which requires even fewer parameters to be determined.

As mentioned the matrix $C = C_S + C_N$ is typically not diagonal. However, we can get some nice insight into the likelihood function by considering the special case

when C is diagonal and proportional to the identity matrix (all diagonal elements the same). In that case,

$$\mathcal{L} \propto \frac{1}{(C_S + C_N)^{N_p/2}} \exp\left\{ -\frac{1}{2} \frac{\sum_{i=1}^{N_p} \Delta_i^2}{C_S + C_N} \right\}. \tag{11.21}$$

We can easily find the value of C_S which maximizes the likelihood function in this case. Differentiating with respect to C_S leads to

$$\frac{\partial \mathcal{L}}{\partial C_S} = \mathcal{L} \times \left[\frac{-(N_p/2)}{(C_S + C_N)} + \frac{1}{2} \frac{\sum_{i=1}^{N_p} \Delta_i^2}{(C_S + C_N)^2} \right]. \tag{11.22}$$

If we set this to zero, we find that the likelihood function is maximized at

$$C_S = \frac{1}{N_p} \sum_{i=1}^{N_p} \Delta_i^2 - C_N. \tag{11.23}$$

Thus a useful rule of thumb for estimating the signal in a CMB experiment is to calculate the variance of the data points (the first term on the right in Eq. (11.23)) and compare it with the average noise per pixel (the second term on the right). If the data has larger variance than the noise, the theoretical signal is simply the difference between the two.

We can also calculate the error on this determination of C_S. As we saw in Section 11.1.1, the error is related to the second derivative of the log of the likelihood function:

$$\sigma_{C_S} = \left(\frac{-\partial^2 \ln \mathcal{L}}{\partial C_S^2} \right)^{-1/2}. \tag{11.24}$$

In this case, it is easy to calculate the derivative. Differentiating Eq. (11.22) once more leads to

$$\frac{\partial^2 \ln \mathcal{L}}{\partial {C_S}^2} = \frac{(N_p/2)}{(C_S + C_N)^2} - \frac{\sum_{i=1}^{N_p} \Delta_i^2}{(C_S + C_N)^3}. \tag{11.25}$$

At the peak of the likelihood we can replace the $\sum_{i=1}^{N_p} \Delta_i^2$ by $N_p[C_S + C_N]$, so

$$\sigma_{C_S} = \sqrt{\frac{2}{N_p}} (C_S + C_N). \tag{11.26}$$

Equation (11.26) is a simplified version of a very handy, useful formula which can be used to assess how accurately a given experiment will determine parameters. This simplified version gives the errors on our one theoretical parameter, C_S. The more general formula gives the corresponding errors when the free parameters are the C_l's themselves. In that case,

$$\sigma_{C_l} = \sqrt{\frac{2}{(2l+1) f_s}} (C_l + C_{N,l}). \tag{11.27}$$

The only change moving from Eq. (11.26) to Eq. (11.27) is that the number of pixels – or equivalently the number of independent measurements – has been replaced by

$(2l+1)f_s$, where f_s is the fraction of the sky covered. This makes perfect sense, for in the full sky limit, one can measure at best $2l+1$ a_{lm}'s; that is, one can sample the distribution characterized by C_l only $2l+1$ times. In fact, this is a fundamental limit on the accuracy with which we can measure the C_l's. Even if there is no noise ($C_N = 0$), there remains a fundamental uncertainty in the theoretical parameters (either C_S or C_l) due to the fact that we only have one sky on which to take measurements. This limit, which we have already encountered in Chapter 8, is called *sample variance*, or in the limit of an all-sky survey *cosmic variance*.

11.1.3 Galaxy Surveys

At first, one might think that analysis of galaxy surveys would be completely different from CMB analysis. There *are* a number of differences. The galaxy distribution is fundamentally 3D, while the CMB anisotropies are a function of angular position only. Also, CMB experiments measure a continuous field, the temperature field, a function of position. Galaxy surveys count discrete objects (galaxies). A survey is simply a list of positions of these objects. Another difference is that the CMB temperatures are drawn from a Gaussian distribution, whereas the galaxies supposedly trace the underlying mass distribution, which — at least on small scales — has already "gone nonlinear." Nonlinearities inevitably produce non-Gaussianity, even if the primordial distribution is Gaussian.

Despite these, and other, differences, analysts have in recent years come to realize that many of the same techniques can be applied to data from both the CMB and galaxy surveys. To solidify the CMB–galaxy survey connection, we need to formalize the concept of a pixel. In the case of the CMB, the notion of a pixel is so natural that I didn't even bother to define it above. For galaxy surveys, following the treatment of Tegmark *et al.* (1998), we can define the data in pixel i as

$$\Delta_i \equiv \int d^3x \; \psi_i(\vec{x}) \left[\frac{n(\vec{x}) - \bar{n}(\vec{x})}{\bar{n}(\vec{x})} \right]. \tag{11.28}$$

Here $n(\vec{x})$ is the galaxy density at $\vec{x}$ and $\bar{n}$ is the expected number of galaxies at $\vec{x}$, i.e., the number there would be if the distribution was uniform. The weighting function ψ_i, which determines the pixelization, will be discussed shortly, but first let's understand operationally how to determine n and $\bar{n}$ from a survey. A simple way is to divide the volume into small sub-volumes, each of which is much smaller than the total survey, but large enough to contain many (e.g., greater than 10) galaxies. The density of a given sub-volume is then the number of galaxies in it divided by its volume. For a uniform survey, the average density $\bar{n}$ would just be the total number of galaxies divided by the total volume.[2]

There are two popular choices for ψ_i, choices which determine the pixelization. First is "counts-in-cells," wherein

[2]A caveat: if the pixels as defined by ψ_i are overlapping, more care must be taken in computing $\bar{n}$ for then the same galaxy could appear in more than one pixel.

$$\psi_i^{\rm CIC}(\vec{x}) = \begin{cases} \bar{n}(\vec{x}) & \text{if } \vec{x} \text{ is in the } i\text{th sub-volume} \\ 0 & \text{otherwise} \end{cases} . \tag{11.29}$$

In this case, the sub-volumes themselves are the pixels, and Δ_i is the over(under)-density in the ith sub-volume. Another useful pixelization scheme is a set of Fourier pixels, which emerge from choosing

$$\psi_i^{\rm Fourier}(\vec{x}) = \frac{e^{i\vec{k}_i\cdot\vec{x}}}{V} \begin{cases} 1 & \vec{x} \text{ inside the survey volume} \\ 0 & \vec{x} \text{ outside survey volume} \end{cases} . \tag{11.30}$$

Here V is the volume of the survey. In this case, the pixels are not spatial, but rather live in the Fourier domain. Still, even in this case, Δ_i is the fractional overdensity in the pixel.

No matter which pixelization is chosen, one cannot hope to write down a simple expression for the likelihood function, the probability of getting a set of $\{\Delta_i\}$ given a theory. The theory of galaxy formation is simply too complicated. Indeed, even if one assumes that the galaxy density perfectly traces the mass overdensity, the complications from gravity alone make the likelihood function non-Gaussian. Nonetheless, progress can still be made by noting that the expectation value of Δ_i is zero by construction, with a covariance matrix[3]

$$\langle \Delta_i \Delta_j^* \rangle = (C_S)_{ij} + (C_N)_{ij} \tag{11.31}$$

exactly like the CMB case. We will discuss the signal covariance matrix in detail in Section 11.2. The noise covariance matrix is actually easier than the corresponding CMB matrix, which depends on the atmosphere, pointing, instrumental noise, scan strategy, and other experimental details. In a galaxy survey, even if there was no signal, the expected value of the square of the density, $\langle n^2(\vec{x}) \rangle$, would still differ from $\bar{n}^2$ simply because there are only a finite number of galaxies in a given sub-volume. Thus, even in the absence of any intrinsic clustering of the galaxies $\langle \Delta_i \Delta_j \rangle$ would be nonzero because of *Poisson noise*. You can show in Exercise 3 that the covariance matrix due to Poisson noise is

$$(C_N)_{ij} = \int d^3x \frac{\psi_i(\vec{x})\psi_j^*(\vec{x})}{\bar{n}(\vec{x})}. \tag{11.32}$$

Armed with this noise covariance matrix and the signal covariance matrix we explore next, galaxy survey analysts can use many of the same techniques as the CMB analysts.

11.2 SIGNAL COVARIANCE MATRIX

Until now, we have sidestepped the question of how the expected variance in a given experiment is related to the underlying power spectrum. That is, we have learned that the predictions of a given theory are a set of C_l's and $P(k)$. If we want to relate theory to experiment — and we do! — we need to know how to turn this set of predictions into a covariance matrix C_S.

[3]The angular brackets here denote an average over the distribution from which Δ_i is drawn.

11.2.1 CMB Window Functions

For simplicity, let us first consider the diagonal element of the covariance matrix:

$$C_{S,ii} \equiv \langle s_i s_i \rangle \qquad \text{(no sum over } i\text{)} \tag{11.33}$$

where the average $\langle \ldots \rangle$ is over many realizations of the theoretical distribution and the subscript i labels the pixel. The temperature difference reported in each pixel can be expressed as

$$s_i = \int d\hat{n} \Theta(\hat{n}) B_i(\hat{n}) \tag{11.34}$$

where B_i is the beam pattern at the ith pixel and Θ is the underlying temperature. As an example, the beam pattern from the MSAM experiment is shown in Figure 11.1. It is typical of the patterns produced by many CMB experiments: the difference of the temperature in two (or more) regions of the sky, and in each region the temperature is sampled by a beam which is roughly Gaussian.

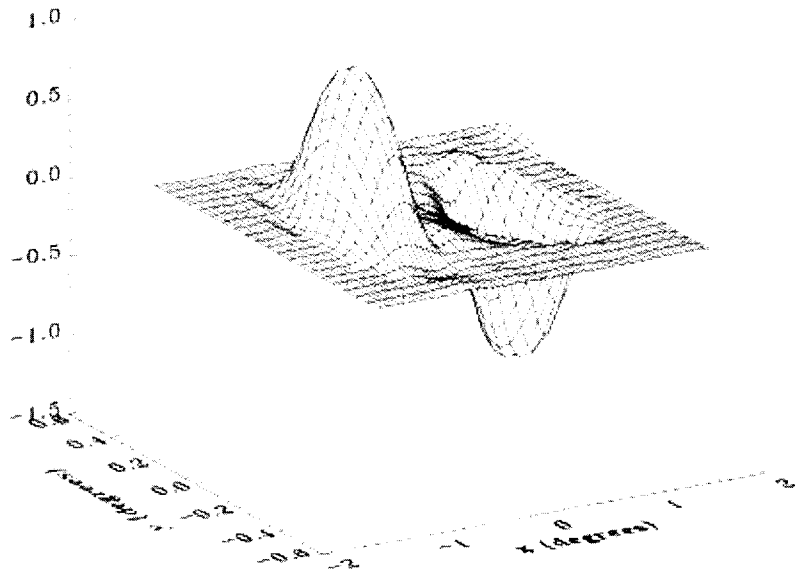

Figure 11.1. The beam pattern for the MSAM experiment (see Wilson *et al.*, 2000, for a summary). The anisotropy reported at a given pixel is roughly the difference between the temperature at ± 0.5 degrees from the center of the pixel. The beamwidth is also on the order of half a degree.

To find C_S we square Eq. (11.34) after expressing the temperature field as an expansion over spherical harmonics as in Eq. (8.60):

$$\frac{C_{S,ii}}{T^2} = \int d\hat{n} \int d\hat{n}' B_i(\hat{n}) B_i(\hat{n}') \sum_{lm} Y_{lm}(\hat{n}) \sum_{l'm'} Y^*_{l'm'}(\hat{n}') \langle a_{lm} a^*_{l'm'} \rangle. \tag{11.35}$$

Using Eq. (8.63), we find that the sums over $l'm'$ collapse to give

$$\frac{C_{S,ii}}{T^2} = \int d\hat{n} \int d\hat{n}' B_i(\hat{n}) B_i(\hat{n}') \sum_l C_l \sum_m Y_{lm}(\hat{n}) Y^*_{lm}(\hat{n}'). \tag{11.36}$$

But $\sum_m Y_{lm}(\hat{n}) Y^*_{lm}(\hat{n}') = (2l+1) P_l(\hat{n} \cdot \hat{n}')/4\pi$,so

$$\frac{C_{S,ii}}{T^2} = \sum_l \frac{2l+1}{4\pi} C_l W_{l,ii} \tag{11.37}$$

where the window function is defined as

$$W_{l,ii} \equiv \int d\hat{n} \int d\hat{n}' B_i(\hat{n}) B_i(\hat{n}') P_l(\hat{n} \cdot \hat{n}'). \tag{11.38}$$

Until now, we have been thinking of $\hat{n}, \hat{n}'$ as three-dimensional unit vectors. If $\hat{n}'$ and $\hat{n}$ are sufficiently close to each other, though, we can use the flat space approximation. The three-dimensional unit vectors can be safely approximated as two-dimensional vectors $\vec{x}, \vec{x}'$ in the transverse directions. The distance between $\vec{x}$ and $\vec{x}'$ (measured in radians) is then equal to the angle between $\hat{n}$ and $\hat{n}'$. In this limit, the argument of the Legendre polynomial in Eq. (11.38) becomes

$$\vec{n} \cdot \vec{n}' = \cos\left(|\vec{x} - \vec{x}'|\right). \tag{11.39}$$

The diagonal window function is therefore

$$W_{l,ii} = \int d^2x \int d^2x' B_i(\vec{x}) B_i(\vec{x}') P_l\Big(\cos\left(|\vec{x} - \vec{x}'|\right)\Big). \tag{11.40}$$

A useful property of Legendre polynomials is that they become equal to the zero-order Bessel function in the limit of large l (the small-angle limit we are working under here). So,

$$\begin{aligned} P_l\Big(\cos\left(|\vec{x} - \vec{x}'|\right)\Big) &\to J_0\left(l|\vec{x} - \vec{x}'|\right) \\ &= \frac{1}{2\pi} \int_0^{2\pi} d\phi e^{-il|\vec{x}-\vec{x}'|\cos\phi}, \end{aligned} \tag{11.41}$$

where the last line is an integral representation of the Bessel function (Eq. (C.21)). We can simplify further by promoting l to a 2D vector with direction chosen so that the angle between $\vec{l}$ and $\vec{x} - \vec{x}'$ is equal to ϕ. Then, the argument of the exponential simplifies to $-i\vec{l} \cdot (\vec{x} - \vec{x}')$. This form is so useful because the $\vec{x}$ integral for example is now

$$\int d^2x B_i(\vec{x}) e^{-i\vec{l}\cdot\vec{x}} \equiv \tilde{B}_i(\vec{l}), \tag{11.42}$$

where $\tilde{B}_i$ is the Fourier transform of the beam pattern. The $\vec{x}'$ integral is the complex conjugate of this, so the window function simplifies to

$$W_{l,ii} = \frac{1}{2\pi} \int_0^{2\pi} d\phi \left|\tilde{B}_i(\vec{l})\right|^2. \tag{11.43}$$

Thus calculating the window function reduces to a two-step process:

- Calculate the 2D Fourier transform of the beam pattern.
- Find the angular average of the square of this transform.

The window function is a function of the experiment only and indeed contains information about the beam size and chopping angle of the experiment. However, it is not the whole story. A complete evaluation of the likelihood function entails calculating all of the elements of the covariance matrix C_S. A given off-diagonal element of the matrix is given by Eqs. (11.37) and (11.38), with one of the indices i changed to j. The matrix is symmetric, so it is characterized by $N_p(N_p+1)/2$ elements where N_p is the number of pixels. Thus, for an N_p-pixel experiment there are really $N_p(N_p+1)/2$ window functions!

11.2.2 Examples of CMB Window Functions

Gaussian Beam. Let us take a break from formalism and calculate a simple (diagonal) window function. Consider a Gaussian beam; this is a good approximation to many CMB experiments. The beam pattern for the ith pixel is

$$B_i(\vec{x}) = \frac{1}{2\pi\sigma^2}\exp\Big(-\frac{(\vec{x}-\vec{x}_i)^2}{2\sigma^2}\Big). \tag{11.44}$$

We may choose $\vec{x}_i$ to be zero for the window function computation. The Fourier transform of the beam is also a Gaussian,

$$\begin{aligned}\tilde{B}_i(\vec{l}) &= \frac{1}{2\pi\sigma^2}\int d^2x e^{-i\vec{l}\cdot\vec{x}}\exp\Big(-\frac{x^2}{2\sigma^2}\Big)\\ &= e^{-l^2\sigma^2/2}.\end{aligned} \tag{11.45}$$

In this simple case, $\tilde{B}$ does not depend on the direction of $\vec{l}$, so there is no need to take the angular average. The window function is then simply the square of the Fourier transform,

$$W_{l,ii} = e^{-l^2\sigma^2}. \tag{11.46}$$

The window function falls off sharply at large l. Large l corresponds to small angular scales. Structure on scales smaller than the beam size is inevitably washed away and undetectable. Figure 11.2 illustrates the series of steps, from beam function to Fourier transform to window function.

There are two subtleties associated with the Gaussian window function. First, one must avoid the temptation to set σ equal to the number which is often quoted in papers, the full width half maximum (FWHM). The latter is twice the value of x for which $B(\vec{x})$ drops to half of its maximum. So $\sigma = \mathrm{FWHM}/(\sqrt{8\ln(2)}) = 0.4245\,\mathrm{FWHM}$. The second subtlety has to do with normalization. It is crucial to determine how observers have normalized their output. The prefactor in Eq. (11.44) ensures that if the temperature field was uniform, the reported temperture would be equal to the underlying one. In this simple case, the choice is obvious; in general factors of 2 can easily be lost.

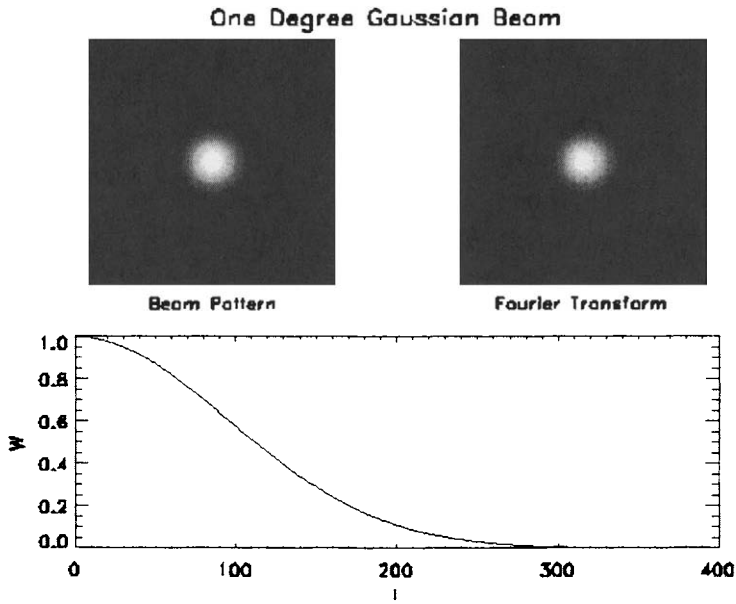

Figure 11.2. A one-degree Gaussian beam is shown in the upper left panel and its Fourier transform, also a Gaussian, in the upper right. The resulting window function is shown in the bottom panel.

Differencing a Gaussian Beam. As another straightforward example, let us consider an experiment which takes the difference between the temperatures at two adjacent points on the sky. For simplicity, let us first assume that the Gaussian beam is infinitely small, so it can be approximated as a Dirac delta function. Then,

$$B(x,y) = \delta(y)\left[\delta(x-x_0) - \delta(x+x_0)\right], \tag{11.47}$$

where the *chopping angle*, or the distance between the plus and minus position, is $2x_0$. The Fourier transform of this is straightforward:

$$\tilde{B}(\vec{l}) = 2i\sin(l_x x_0). \tag{11.48}$$

The window function is the angular average over all $\vec{l}$ directions. Choosing the angle between $\vec{l}$ and the $x-$ axis to be ϕ, we have

$$\begin{aligned} W_l &= \frac{4}{2\pi}\int_0^{2\pi} d\phi\ \sin^2\left[lx_0\cos\phi\right] \\ &= \frac{1}{\pi}\int_0^{2\pi} d\phi\ \left(1-\cos\left[2lx_0\cos\phi\right]\right) \end{aligned}$$

$$= 2\left(1 - P_l\left[\cos(2x_0)\right]\right), \tag{11.49}$$

where the last line follows from Eq. (11.41).

Until now, we have neglected the finite width of the beam. However, this turns out to be very simple to rectify. A realistic beam will be the convolution of the chop described by Eq. (11.47) with a finite beam size:

$$B(x,y) = \frac{1}{2\pi\sigma^2}\int dx'dy' \exp\left\{-\frac{(x-x')^2+(y-y')^2}{2\sigma^2}\right\}$$
$$\times\, \delta(y')\left[\delta(x'-x_0) - \delta(x'+x_0)\right]. \tag{11.50}$$

Recall, though, that the Fourier transform of the convolution of two functions is simply equal to the product of the two Fourier transforms. The angular averaging over this product is unaffected since the Gaussian has no angular dependence. Therefore, the final window function is

$$W_l = e^{-l^2\sigma^2}\left(1 - P_l\left[\cos(2x_0)\right]\right). \tag{11.51}$$

This window function is shown in Figs. 11.3 and 11.4 along with the beam pattern

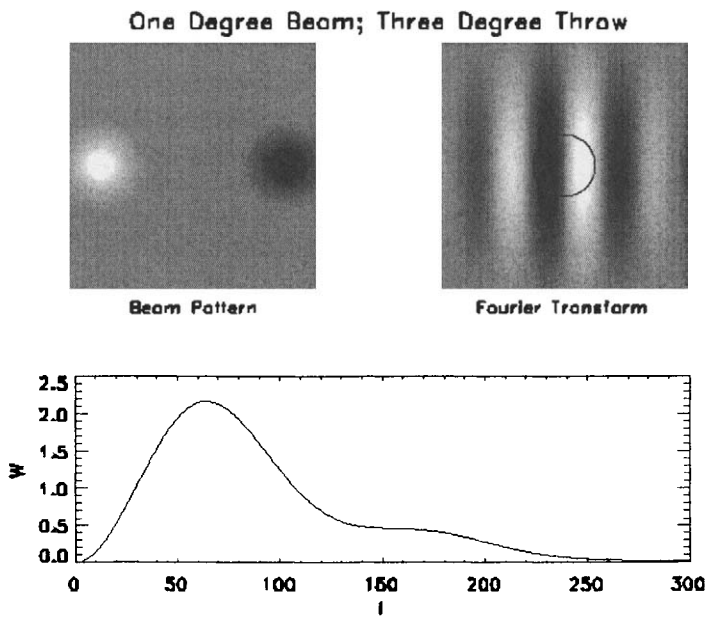

Figure 11.3. Differencing a Gaussian beam. Upper left panel shows the beam pattern and the upper right its Fourier transform. The circle in the upper right corresponds to $l = 50$. The bottom panel shows the window function.

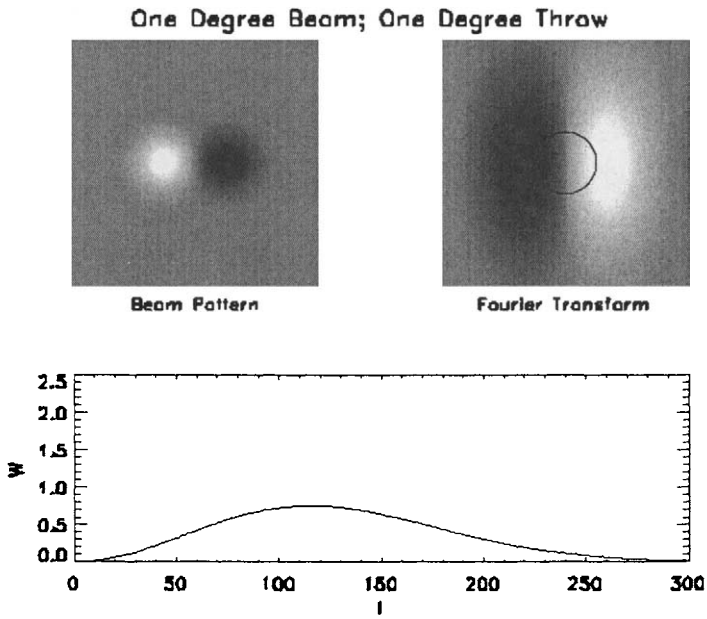

Figure 11.4. Same as Figure 11.3 except the chopping angle is now much smaller. Note the lack of support at large scales due to the reduced chopping angle. This can also be seen in the Fourier transform which vanishes near the center of the circle.

for two different chopping angles. Note that for l much smaller than $1/x_0$, the Fourier transform vanishes. So, unlike the undifferenced beam, there is no support for small l. That is, chopping removes information about structure on large scales. As the chopping angle gets larger, more and more information is obtained about large scales.

11.2.3 Window Functions for Galaxy Surveys

We now consider the signal covariance matrix for galaxy surveys. By setting the term in square brackets in Eq. (11.28) to $\delta(\vec{x})$, we see that the signal covariance matrix is equal to

$$(C_S)_{ij} = \langle \Delta_i \Delta_j \rangle \Big|_{\text{no noise}} = \int d^3x \; d^3x' \; \psi_i(\vec{x}) \psi_j(\vec{x}') \xi(\vec{x} - \vec{x}'). \tag{11.52}$$

The correlation function ξ appears here because it is equal to the expectation value of the product of two overdensities, Eq. (9.7). Since ξ is the Fourier transform of the power spectrum, we see that the signal covariance matrix for a galaxy survey is

$$(C_S)_{ij} = \int d^3x \; d^3x' \int \frac{d^3k}{(2\pi)^3} \frac{d^3k'}{(2\pi)^3} \frac{d^3k''}{(2\pi)^3} \tilde{\psi}_i(\vec{k}) \tilde{\psi}_j^*(\vec{k}') P(k'') e^{i[\vec{k}+\vec{k}'']\cdot\vec{x} - i[\vec{k}'+\vec{k}'']\cdot\vec{x}'}$$

$$= \int \frac{d^3k}{(2\pi)^3} P(k) \tilde{\psi}_i(\vec{k}) \tilde{\psi}_j^*(\vec{k}). \tag{11.53}$$

The second equality follows simply after integrating over $\vec{x}$ and $\vec{x}'$ to get 3D Dirac delta functions and using these to perform the integrals over $\vec{k}'$ and $\vec{k}''$. It is convenient to define the window function as the angular part of this integral

$$W_{ij}(k) \equiv \int \frac{d\Omega_k}{4\pi} \tilde{\psi}_i(\vec{k}) \tilde{\psi}_j^*(\vec{k}) \tag{11.54}$$

so that

$$(C_S)_{ij} = \int_0^\infty \frac{dk}{k} \left[\frac{k^3 P(k)}{2\pi^2}\right] W_{ij}(k). \tag{11.55}$$

Notice that the window function for galaxy surveys has the identical form as that for CMB experiments. In both cases, it is the angular average of the square of the Fourier transform of the weighting function (either B or ψ). Also, you should recognize the quantity in square brackets in Eq. (11.55) as $\Delta^2(k)$, the contribution to the variance per $\ln(k)$. Let's turn to some examples of window functions of galaxy surveys.

Volume-Limited Survey. Consider a survey which observes all galaxies within a radius R from us. If we use Fourier pixelization (Eq. (11.30)), then the Fourier transform of the weighting function is

$$\tilde{\psi}_i(\vec{k}) = \int_{|\vec{x}|<R} \frac{d^3x}{V} e^{-i\vec{k}\cdot\vec{x}} e^{i\vec{k}_i\cdot\vec{x}}. \tag{11.56}$$

We will shortly carry out this integral, square, and then average over all angles to get the diagonal window function of Eq. (11.54). First, though, let's ask what we expect qualitatively. Equation (11.56) is the Fourier transform of the survey volume as a function of $\vec{k} - \vec{k}_i$. The survey volume is a sphere of radius R. In general, when a function is confined to a region $x < R$, the Fourier transform is confined to $k < 1/R$. Here then, $\tilde{\psi}$ will be nonzero only when $|\vec{k} - \vec{k}_i|$ is less than $1/R$. The window function therefore should peak at $k = k_i$ and have a width of order $1/R$.

More quantitatively, the integral in Eq. (11.56) is

$$\begin{aligned}\tilde{\psi}_i(\vec{k}) &= \frac{4\pi}{V|\vec{k} - \vec{k}_i|} \int_0^R dx\ x\ \sin\left(|\vec{k} - \vec{k}_i| x\right) \\ &= \frac{4\pi}{V\left(|\vec{k} - \vec{k}_i|\right)^3} \left[-|\vec{k} - \vec{k}_i| R \cos\left(|\vec{k} - \vec{k}_i| R\right) + \sin\left(|\vec{k} - \vec{k}_i| R\right)\right]\end{aligned} \tag{11.57}$$

The diagonal window function is the angular average of the square of this. Defining $y \equiv |\vec{k} - \vec{k}_i| R$, this angular average is

$$W_{ii}(k) = \frac{(4\pi R^3)^2}{V^2} \int_{-1}^1 \frac{d\mu}{2} \int_0^{2\pi} \frac{d\phi}{2\pi} \frac{(\sin y - y\cos y)^2}{y^6}$$

$$= \frac{8\pi^2 R^6}{V^2} \int_{-1}^{1} \frac{d\mu}{y^6} (\sin y - y\cos y)^2 \tag{11.58}$$

where μ is the cosine of the angle between $\vec{k}$ and $\vec{k}_i$. Integrating over y instead of μ leads to

$$W_{ii}(k) = \frac{9}{2kk_iR^2} \int_{|k-k_i|R}^{(k+k_i)R} \frac{dy}{y} j_1^2(y) \tag{11.59}$$

since the volume of the survey $V = 4\pi R^3/3$. This window function is shown in Figure 11.5 for several different values of k_iR. Notice that modes with wavelength much smaller than the size of the survey $k_iR \gg 1$ do indeed have window functions sharply peaked at $k = k_i$, with a width of order $1/R$. The largest wavelength modes, however, pick up contributions from all scales (e.g., the $k_iR = 3$ curve in Figure 11.5). Not surprisingly, surveys do not do a good job measuring the power on wavelengths comparable to their sizes.

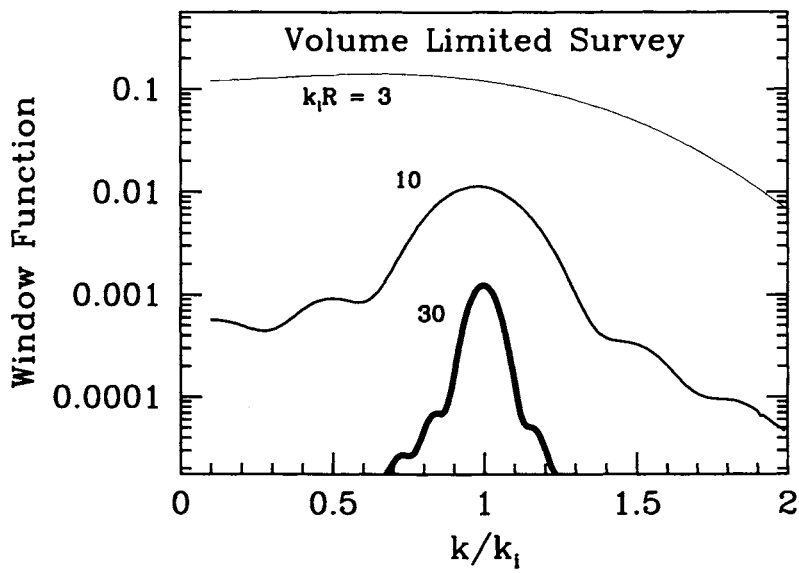

Figure 11.5. The window function for a volume-limited survey. Narrow window functions enable one to determine the power spectrum at the wavenumber of interest more accurately. Modes with wavelengths comparable to the size of the survey (e.g., here $k_iR = 3$) have broad window functions.

The height of the window function is also important, for it determines the amplitude of the signal covariance matrix. When considering modes with wavelengths k_i^{-1} much smaller than R, you will show in Exercise 7 that

$$(C_S)_{ii} \simeq \frac{P(k_i)}{V}. \tag{11.60}$$

It is instructive to compare this to the corresponding element of the noise matrix, as given in Eq. (11.32). For the diagonal elements, $|\psi_i|^2$ in the integrand is $1/V^2$ as long as $\vec{x}$ is in the survey volume. Thus,

$$(C_N)_{ii} = \frac{1}{\bar{n}V}, \tag{11.61}$$

equal to the inverse of the total number of galaxies in the survey. The ratio of the diagonal elements of the signal and noise covariance matrices is therefore

$$\frac{(C_S)_{ii}}{(C_N)_{ii}} \simeq P(k_i)\bar{n}. \tag{11.62}$$

Harking back to Eq. (11.27), we identify this ratio as the ratio of cosmic variance to Poisson noise. A rough estimate is that $\bar{n} \sim 1\,h^3\,\mathrm{Mpc}^{-3}$, so cosmic variance dominates as long as the power spectrum is larger than $1\,h^{-3}\,\mathrm{Mpc}^3$. Looking back to Figure 7.11, we see that on large scales, this is always satisfied. On small scales, eventually the power spectrum does drop beneath $1\,h^{-3}\,\mathrm{Mpc}^3$, the linear power spectrum of standard CDM at $k \sim 10\,h\,\mathrm{Mpc}^{-1}$. On very small scales, therefore, Poisson noise becomes important.

Pencil-Beam Survey. Now consider a survey which is very deep, but also very narrow, with the general shape of a pencil (Figure 11.6). The Fourier transform of the

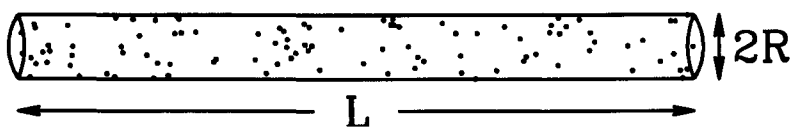

Figure 11.6. A pencil-beam survey with depth L much larger than the typical size of the narrow dimensions, R. The z-axis in the text is taken to be aligned with the long dimension.

weighting function in this case is

$$\tilde{\psi}_i(\vec{k}) = \int \frac{d^3x}{\pi R^2 L} e^{i(\vec{k}_i - \vec{k})\cdot\vec{x}} \Theta(z + L/2)\Theta(L/2 - z)\Theta(R^2 - x^2 - y^2) \tag{11.63}$$

where Θ is a step function, equal to 1 if its argument is positive and zero otherwise. It is easiest to carry out this integral in cylindrical coordinates, wherein $x^2+y^2 = r^2$. If we define $\vec{q} \equiv \vec{k}_i - \vec{k}$, then

$$\tilde{\psi}_i(\vec{k}) = \frac{1}{\pi R^2 L} \int_{-L/2}^{L/2} dz e^{iq_z z} \int_0^R dr r \int_0^{2\pi} d\theta e^{iq_r r\cos\theta}. \tag{11.64}$$

The azimuthal integral can be done using Eq. (C.17), the integral over z using Eq. (C.15), so that

$$\tilde{\psi}_i(\vec{k}) = \frac{2}{R^2} j_0(q_z L/2) \int_0^R dr r J_0(q_r r). \tag{11.65}$$

Finally the integral over r is $RJ_1(q_rR)/q_r$, which you can see by differentiating the integral with respect to q_rR and then using Eq. (C.22). Therefore, the Fourier transform of the weighting function is

$$\tilde{\psi}_i(\vec{k}) = \frac{2}{(q_rR)} j_0(q_zL/2)J_1(q_rR). \tag{11.66}$$

Equation (11.66) indicates that the Fourier transform of the weighting function is anisotropic. Indeed, even before deriving the various flavors of Bessel functions, we should have expected this Fourier transform to be compact along the q_z direction and broad in the transverse plane, i.e., shaped like a disk. This flows from our intuition that the Fourier transform of a function localized within a radius R should be localized within a region $1/R$. Indeed, in the z direction, $j_0(q_zL/2)$ falls off once q_z gets larger than $2/L$. The same holds for $J_1(q_rR)$: it becomes small for $q_r > 1/R$. The ringing associated with these Bessel functions is a manifestation of the fact that the Fourier transform of a top-hat function oscillates for large wavenumbers (e.g., Exercise 6).

To get the window function for a pencil-beam survey, we need to average Eq. (11.66) over all directions $\hat{k}$. This average will differ for different $\vec{k}_i$. Let's choose $\vec{k}_i$ to point in the $\hat{z}$-direction as one concrete example. The averaging will pick up contributions only when $\vec{q} \equiv \vec{k}_i - \vec{k}$ has z-component smaller than L^{-1} and transverse component smaller than R^{-1}. Since the transverse component of $\vec{k}_i$ is zero in this example, many $\vec{k}$ will contribute, as long as their transverse component is smaller than R^{-1}. Therefore, we expect the window function to get contributions from many wavenumbers, not to be sharply peaked around $|\vec{k}_i|$. A similar argument holds for other directions $\hat{k}_i$. Figure 11.7 shows the window function for a pencil-beam survey. As expected, it is broader than that for a symmetric, volume-limited survey: a given scale k_i picks up contributions from smaller scales $k > k_i$.

11.2.4 Summary

We have determined the signal covariance matrix for CMB experiments which measure a filtered version of the temperature in a given set of pixels and for galaxy surveys which measure the overdensity in a given set of pixels. Not surprisingly, the fundamental relation between the covariance matrix and and the underlying power spectrum is very similar in both cases: the connection is provided by a window function determined by the experimental/observational specifications. It is interesting to point out that we have encountered *natural* window functions in Chapters 9 and 10. The angular correlation function of Section 9.1 is simply the signal covariance matrix of measurements of the 2D galaxy distribution. Recall from Eq. (9.13) that this too is an integral of the 3D power spectrum convolved with a window function (we called it a *kernel* back then). The same is true for the peculiar velocity (Eq. (9.29)) and the shear field that can be measured with weak lensing (Eq. (10.32)). In all of those cases, the window function is determined partly by the

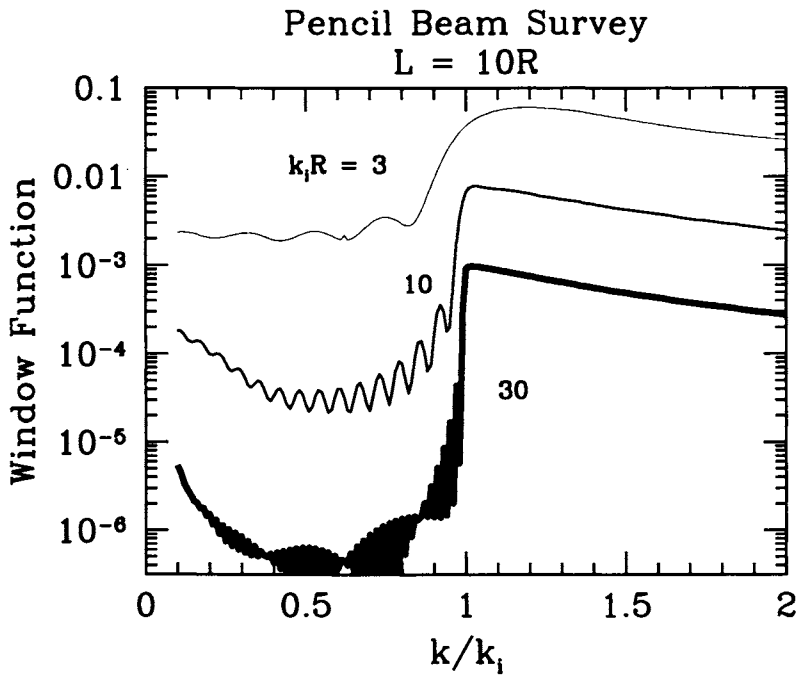

Figure 11.7. The window function for a pencil-beam survey with length 10 times larger than width. These window functions are for $\vec{k}_i$ pointing in the $\hat{z}$-direction, along the long dimension of the survey.

observational strategy (e.g., how deep one goes), but also by the intrinsic nature of the measurement itself. For example, the 2D galaxy distribution intrinsically is a projection of the 3D distribution. So it is seductive to think of CMB anisotropy experiments and 3D galaxy surveys as more pristine measurements of the power spectrum. Analysts are hard to seduce, though. They recognize that, mathematically, the different sets of measurements can all be analyzed in the same fashion. So the likelihood function can be used on virtually all cosmological observations. It is a very powerful tool.

It is also simple. Were it not for the size of modern cosmological data sets, we would be done. At least in the Gaussian case (CMB or even galaxy surveys on large scales), the likelihood function is given by Eq. (11.20). The data points are simply the pixelized temperatures or overdensities, while the covariance matrix is the sum of the noise and signal covariances. The noise covariance matrix is usually estimated from the data itself, while the signal covariance matrix is computed by convolving the theory (i.e., the C_l's or the power spectrum) with the window function. In principle, then, one could compute this likelihood function at many points in parameter space, find its maximum (this constitutes the set of best-fit parameters), and the contour delineating the region in which, say, 95% of the volume

lies. This contour would then be the 95% confidence region of the parameters. Many experiments of the previous decade, especially CMB experiments, have been analyzed in this brute-force fashion. Times are changing, though, and the brute-force approach has already become impractical.

11.3 ESTIMATING THE LIKELIHOOD FUNCTION

To illustrate the need for new techniques of likelihood computation, let us consider a concrete example: the data set from the Boomerang anisotropy experiment (Netterfield *et al.*, 2001). There are 57,000 pixels on the sky covered in this set. The theory and noise covariance matrices are both nondiagonal and both $57,000 \times 57,000$ dimensional matrices. Inverting these beasts with present computers is possible, although slow.[4] If we needed to invert only once, this might be acceptable. But, we need to evaluate the likelihood function at many points in parameter space to find its maximum and the region at which it falls to, say, 5% of its maximum. This would be barely manageable if the parameter space was one-dimensional. A one-dimensional fit, though, would lose most of the information contained in the map. The data are actually sensitive to the power on many different scales. Therefore, the parameter space — the amplitude of the power on these many different scales — is multidimensional, "multi-" here of order 20. The likelihood function should in principle be computed about 10 times in each dimension, for a total of 10^{20} computations. Since each inversion takes several hours, this is not feasible. All of these estimates are for the Boomerang experiment. The MAP satellite will have 10 times as many pixels and be sensitive to a wider range of scales. Planck will be more sensitive still. Thus, we need new techniques, shortcuts, for evaluating the likelihood function and finding its maximum and its width.

11.3.1 Karhunen–Loève Techniques

The first technique was discovered many years ago and reinvented by a number of people over the past few years to deal with both CMB data (Bunn, 1995; Bond, 1995) and with data from galaxy surveys (Vogeley and Szalay, 1996). It is a method for speeding up the computation of the likelihood function. The fundamental idea is simple: any experiment, no matter how good, will have many *modes* which are useless, fundamentally contaminated by noise.[5] If it was obvious which modes were most noisy, then we could greatly simplify the likelihood calculation by not using those modes. If only 10% of the modes carried useful information — and this figure is roughly what is found in many present-day experiments — then the data set would be reduced by a factor of 10. The covariance matrices would now be $(N/10) \times (N/10)$ and the inversion (which scales as N^3) would speed up by a factor of 1000. It is a nice, simple idea. The only problem is finding the modes which are useful.

[4]Note that the problems outlined here assume that we are handed the map of 57,000 pixels. Mapmaking is actually the most difficult computational part of the analysis!

[5]A mode here is defined as a linear combination of the data points.

If the signal and noise covariance matrices were diagonal, then it would be simple to ascertain which modes had high signal-to-noise. The pixels with diagonal elements $C_S > C_N$ would have signal-to-noise greater than 1; the others would be low-signal modes. The problem is to identify the low modes in the more realistic case where the covariance matrices are not diagonal. This is precisely what Karhunen–Loève does. To illustrate the technique, let us first write it down formally and then work out a simple example explicitly.

We assume there are N_p data points, Δ_i. Each data point is presumed to be the sum of both signal s_i and noise n_i. Each of these are assumed to be uncorrelated (the noise knows nothing about the signal and vice versa). Thus the full covariance matrix is

$$\langle \Delta_i \Delta_j \rangle \equiv C_{ij} = C_{S,ij} + C_{N,ij}. \tag{11.67}$$

The Karhunen–Loève method takes advantage of the fact that instead of computing the likelihood function using Δ_i and its covariance matrix C, we could instead use rotated data

$$\Delta'_i \equiv R_{ij}\Delta_j \tag{11.68}$$

where R is a real matrix. The covariance matrix associated with Δ' will be

$$\begin{aligned} C'_{ij} &= \langle (R\Delta)_i (R\Delta)_j \rangle \\ &= R_{ii'} R_{jj'} C_{i'j'} \end{aligned} \tag{11.69}$$

In matrix notation this is simply

$$C' = RCR^T \tag{11.70}$$

where R^T is the transpose of R.

The Karhunen–Loève method consists of three such rotations.

1. R_1 : Diagonalize C_N
2. R_2 : Set $C'_N = I$
3. R_3 : Diagonalize C'_S

The first step is always possible since C_N is a real, symmetric matrix. Once C_N has been diagonalized, it is trivial to perform step 2: simply choose R_2 to be diagonal with elements equal to 1 over the square root of the diagonal elements that emerge from step 1. Finally, step 3 is straightforward since again C_S is a real symmetric matrix. Let's evaluate the new theory and noise covariance matrices. The theory matrix is

$$C'_S = R_3 R_2 R_1 C_S R_1^T R_2 R_3^T. \tag{11.71}$$

Note that, since R_2 is diagonal it is equal to its transpose. The matrix C'_S is a diagonal matrix. Now consider C'_N. After step 2, it was simply the identity matrix. So we need consider only the effect of step 3. In fact, since R_3 is unitary, it has no effect ($R_3 I R_3^T = I$). Thus, C'_N is still equal to the identity matrix. This has

profound implications. It means that the elements of (the diagonal matrix) C'_S are a measure of the signal-to-noise squared of the modes! The data points

$$\Delta'_i = (R_3 R_2 R_1)_{ij} \Delta_j \tag{11.72}$$

then have diagonal covariance matrix

$$\langle \Delta'_i \Delta'_j \rangle = \begin{cases} 1 + C'_{S,ii} & i = j \\ 0 & i \neq j \end{cases} . \tag{11.73}$$

These modes can be ordered according to their signal-to-noise values. Modes with large C'_S can be kept; those with C'_S significantly smaller than 1 can be eliminated from the analysis.

Let us work through a simple example to see how Karhunen–Loève picks out the highest signal-to-noise modes. The example is a simple two-pixel experiment with diagonal noise:

$$C_N = \begin{pmatrix} \sigma_n^2 & 0 \\ 0 & \sigma_n^2 \end{pmatrix} . \tag{11.74}$$

The signal covariance matrix does have correlations between the two pixels so

$$C_S = \sigma_s^2 \begin{pmatrix} 1 & \epsilon \\ \epsilon & 1 \end{pmatrix} \tag{11.75}$$

where σ_s is the expected rms in the pixel and $-1 < \epsilon < 1$ measures how correlated the signal is between the two pixels. Steps 1 and 2 of the Karhunen–Loève method are particularly simple since C_N is diagonal. Thus,

$$R_1 = I \quad ; \quad R_2 = \frac{1}{\sigma_n} I. \tag{11.76}$$

To complete step 3, we need to diagonalize

$$R_2 R_1 C_S R_1^T R_2 = \frac{\sigma_s^2}{\sigma_n^2} \begin{pmatrix} 1 & \epsilon \\ \epsilon & 1 \end{pmatrix} . \tag{11.77}$$

To diagonalize the matrix in Eq. (11.77), we must solve

$$\frac{\sigma_s^2}{\sigma_n^2} \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} 1 & \epsilon \\ \epsilon & 1 \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} = \begin{pmatrix} C'_{S,11} & 0 \\ 0 & C'_{S,22} \end{pmatrix} \tag{11.78}$$

for the rotation angle θ. Carrying out the multiplication on the left side leads to

$$\frac{\sigma_s^2}{\sigma_n^2} \begin{pmatrix} 1 + \epsilon \sin(2\theta) & \epsilon \cos(2\theta) \\ \epsilon \cos(2\theta) & 1 - \epsilon \sin(2\theta) \end{pmatrix} = \begin{pmatrix} C'_{S,11} & 0 \\ 0 & C'_{S,22} \end{pmatrix} . \tag{11.79}$$

Equality in the off-diagonal elements holds if $\theta = \pi/4$, so the new theory covariance matrix is

$$C'_S = \frac{\sigma_s^2}{\sigma_n^2} \begin{pmatrix} 1 + \epsilon & 0 \\ 0 & 1 - \epsilon \end{pmatrix} . \tag{11.80}$$

The rotation matrix which diagonalized the theory matrix is the first one on the left in Eq. (11.78) with $\theta = \pi/4$, so

$$R_3 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}. \tag{11.81}$$

The new modes are $\Delta' = R_3 R_2 \Delta$; explicitly, they are

$$\Delta'_1 = \frac{(\Delta_1 + \Delta_2)}{\sqrt{2}\sigma_n}$$

$$\Delta'_2 = \frac{(-\Delta_1 + \Delta_2)}{\sqrt{2}\sigma_n}. \tag{11.82}$$

The new modes (Eq. (11.82)) and their covariance matrix (Eq. (11.80)) are easy to understand if we consider the special cases of $\epsilon = 0$ and $\epsilon = +1$. If $\epsilon = 0$ the two modes have the same signal-to-noise, σ_s/σ_n. If the expected signal is large, these modes both carry information; if not, the noise swamps the signal. In either case, each mode — the sum and the difference — is equally (un)important. If $\epsilon = +1$ then the theory predicts the same signal in each pixel. In that case, the difference mode (Δ'_2) is worthless, since only noise contributes to it. We see this from the fact that $(C'_S)_{22}$ goes to zero as $\epsilon \to 1$. Its signal-to-noise is zero. The other mode — the sum of the two pixels — has signal-to-noise of $\sqrt{2}\sigma_s/\sigma_n$ since the two measurements beat down the noise by a factor of $\sqrt{2}$. This would of course have emerged from the full 2×2 likelihood analysis. But, using both modes in the analysis is a waste of time, a waste which is detected and obliterated by the Karhunen–Loève method.

Bunn (1995) and Bond (1995) independently analyzed the COBE data by looking at Karhunen–Loève modes. Figure 11.8 shows several such modes: clearly the ones with the highest signal-to-noise are the large-scale modes, indicating that COBE was sensitive to large-angle anisotropy. The smallest signal-to-noise modes are the small-scale modes which COBE did not have the resolution to measure.

Vogeley and Szalay (1996) first applied this technique to galaxy surveys. In this context, the Karhunen–Loève method has another useful feature. Recall from Section 11.2.3 that on large scales, the signal covariance matrix is larger than Poisson noise, while on small scales the reverse is true. When we order the modes, then, large-scale modes will have the largest ratio of signal-to-noise. Thus the Karhunen–Loève basis will preferentially pick out large-scale modes. This is extremely useful because we are often most interested in eliminating small-scale modes — which are afflicted by nonlinearities and bias — from an analysis. The Karhunen–Loève method does this automatically! An example is shown in Figure 11.9. For example, the first eigenmode roughly weights pixels only on the basis of their distance from us. It essentially takes the difference between the number of galaxies close to us and the number at moderate distances. The second mode takes a different component of the dipole, the difference between the number of galaxies on the left and the number on the right. Modes with lower weight take successively higher moments of the galaxy distribution.

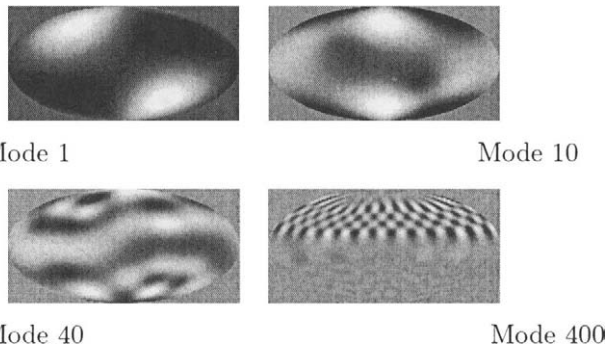

Figure 11.8. Modes of the COBE anisotropy experiment (from Bunn, 1995) ordered according to their signal-to-noise. The mode with the largest signal-to-noise (upper left) is sensitive predominantly to the quadrupole, while the mode with the smallest signal (bottom right) is sensitive to much smaller scale structure.

There are several drawbacks to the Karhunen-Loève method. First, we need to assume a C_S to begin with, in order to identify the modes that are worthless. Although this might seem like a big problem — the modes that are thrown out for one choice of C_S could conceivably be important for another — people who have studied the issue assure us that it is not. They claim that the choice of the important modes is relatively insensitive to the input, assumed C_S. Another drawback is computational. Once the important modes are chosen, C'_S needs to be recalculated at many points in parameter space. In most of parameter space it will not be diagonal at all (it is only diagonal at the special point, the C_S that was chosen as the input spectrum initially). Thus at every point in parameter space, we still need to invert nondiagonal covariance matrices. This drawback is of course partially offset by the fact that — by virtue of the much smaller size of the matrices — the computation is now much faster. Nonetheless in many instances this is not enough to make the full computation managable. We must find still other ways of reducing the computational burden.

Before turning to one such way, let me mention one more use of the Karhunen-Loève method. It is an extremely useful consistency check. This is perhaps best illustrated with an example. The Python experiment (Coble *et al.*, 1999) measured anisotropy over a large (for that time) area, with very low signal-to-noise. In principle, this is a good idea because the large area beats down cosmic variance. In practice, though, it presents challenges because it is difficult to check the consistency of the data. One typically wants to break up the data into several subsets and make sure that each subset sees the same signal. With Python, this was very difficult because noise dominated each subset. One way to check for consistency then is to work with Karhunen-Loève modes. In this basis, each data point d'_i should be drawn from a Gaussian distribution with variance equal to $1 + C'_{S,ii}$. The histogram of $d'_i/(1 + C'_{S,ii})^{1/2}$ should then look Gaussian. Figure 11.10 shows this histogram

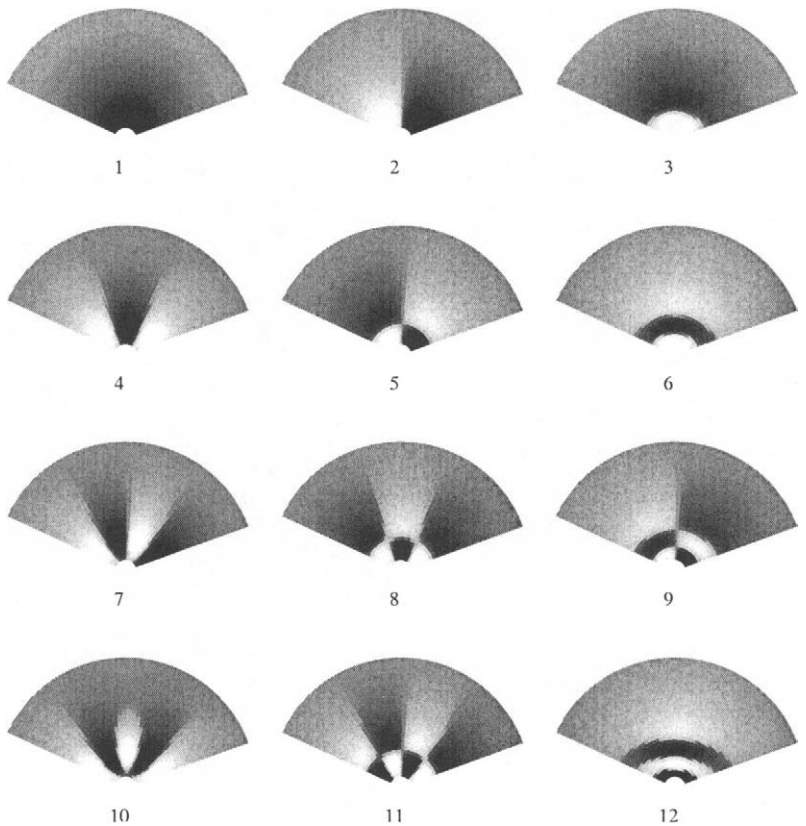

Figure 11.9. The 12 modes with highest signal-to-noise for the CFA2 survey (from Vogeley and Szalay, 1996). Modes pick out the large-scale structure of the galaxy distribution. In each case, we are at the bottom of the slice, and the top region is farthest from us.

using a preliminary noise matrix. There were 650 measurements, but about 70 of these were eliminated for reasons that needn't concern us, so ignore the central spike at $d = 0$. If the remaining 580 data points were distributed as a Gaussian we would expect about two of them (0.3%) to have absolute value greater than 3, and none of them to have absolute value greater than 4. In fact, Figure 11.10 shows that nine modes are more than 4-sigma away from zero. The distribution is definitely not Gaussian with this noise model.

This analysis led the Python team to question the model they had constructed for the noise covariance. (Adjacent points were more correlated than they had allowed for.) Redoing the analysis with a new noise matrix led to the results in Figure 11.11. This technique for testing data quality has been used in a number of other venues, often identifying signs of trouble.

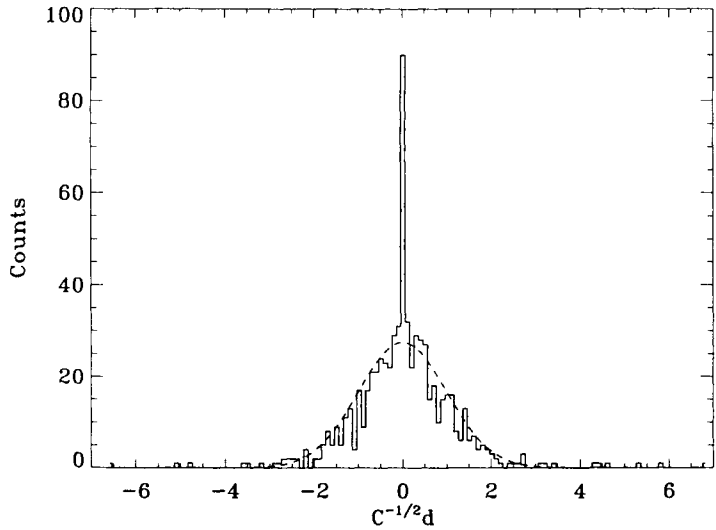

Figure 11.10. Histogram of data from one modulation of the Python CMB anisotropy measurement with a preliminary noise matrix. The data are in Karhunen–Loéve basis in which the covariance matrix is diagonal, so $C^{-1/2}d$ should be distributed as a Gaussian with variance equal to unity. The central spike should be ignored as 70 of the modes have been set to zero. The best fit Gaussian is the solid line. The counts are lower than the best-fit Gaussian in the central region, but above it in the tails.

11.3.2 Optimal Quadratic Estimator

One simple way of speeding up the likelihood calculation is to employ one of many successful *root-finding* algorithms. We are searching for the place where the likelihood function is a maximum, so we want to find

$$\left.\frac{\partial \mathcal{L}}{\partial \lambda}\right|_{\lambda=\bar{\lambda}} = 0 \tag{11.83}$$

where for simplicity I've assumed that the likelihood function depends on only one parameter λ (we'll generalize this shortly) and $\bar{\lambda}$ is its value at the maximum of the likelihood.

An efficient way to find the root is to consider the derivative of the likelihood function evaluated at some trial point $\lambda = \lambda^{(0)}$. Expand this derivative around $\lambda^{(0)}$ in a Taylor expansion:

$$\mathcal{L}_{,\lambda}(\bar{\lambda}) = \mathcal{L}_{,\lambda}(\lambda^{(0)}) + \mathcal{L}_{,\lambda\lambda}(\lambda^{(0)})\left(\bar{\lambda} - \lambda^{(0)}\right) + \ldots \tag{11.84}$$

where I have introduced the notation of writing partial derivatives as subscripted commas:

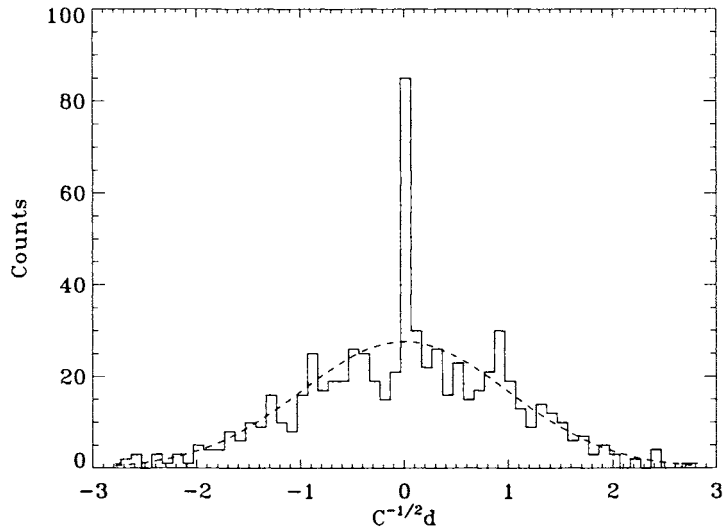

Figure 11.11. Same as Figure 11.10, but this time with an improved noise matrix. Note that there are no data points with absolute value greater than $3C^{1/2}$. The counts are consistent with a Gaussian distribution.

$$\mathcal{L}_{,\lambda} \equiv \frac{\partial \mathcal{L}}{\partial \lambda} \quad ; \quad \mathcal{L}_{,\lambda\lambda} \equiv \frac{\partial^2 \mathcal{L}}{\partial \lambda \partial \lambda}. \tag{11.85}$$

Since $\mathcal{L}$ is maximized at $\bar{\lambda}$, its derivative vanishes there; hence the right-hand side of Eq. (11.84) must also vanish. Setting it to zero leads to a simple expression for $\bar{\lambda}$:

$$\bar{\lambda} \simeq \lambda^{(0)} - \frac{\mathcal{L}_{,\lambda}(\lambda^{(0)})}{\mathcal{L}_{,\lambda\lambda}(\lambda^{(0)})} \tag{11.86}$$

where the $\simeq$ sign acknowledges that we have neglected higher order terms in the Taylor expansion in Eq. (11.84).

The solution in Eq. (11.86) assumes that the likelihood function is a quadratic function of the parameter λ. In fact, it is nothing of the sort: even in the simplest cases the likelihood function is not a quadratic function. For example, far from its maximum, $\mathcal{L}$ typically becomes exponentially small. A much better approximation therefore is that $\mathcal{L}$ is a Gaussian function, so that $\ln(\mathcal{L})$ is quadratic in λ. We can repeat the derivation above since the place where $\mathcal{L}$ is maximized is also the place where $\ln \mathcal{L}$ is a maximum. The estimator for λ is now

$$\hat{\lambda} = \lambda^{(0)} - \frac{(\ln \mathcal{L})_{,\lambda}(\lambda^{(0)})}{(\ln \mathcal{L})_{,\lambda\lambda}(\lambda^{(0)})}. \tag{11.87}$$

Figure 11.12 illustrates the first iteration of this algorithm, called the Newton–Raphson method.

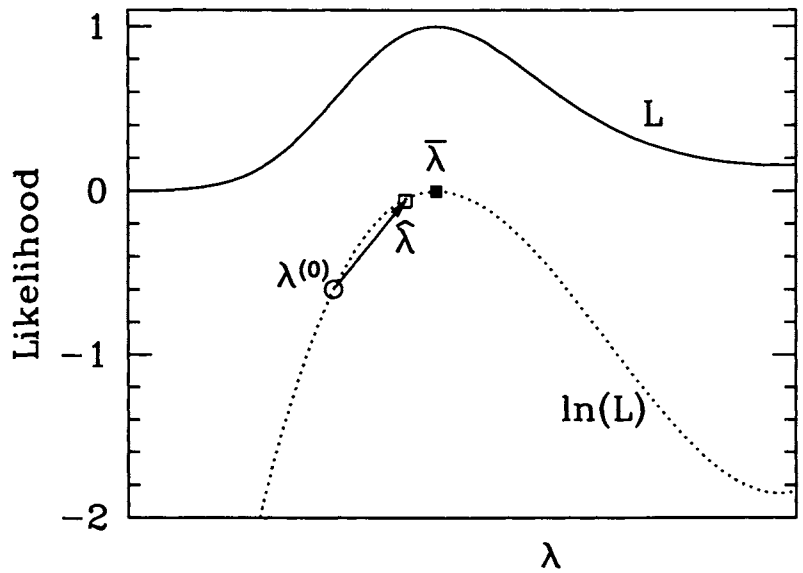

Figure 11.12. A sample likelihood function and its derivatives. The root finding alogrithm starts from the point $\lambda^{(0)}$ (open circle) and moves to $\hat{\lambda}$ (open square) which is quite close to the true maximum of the likelihood, $\bar{\lambda}$ (filled square). The Newton–Raphson technique does this by evaluating the first and second derivatives of $\ln(L)$ at the trial point $\lambda^{(0)}$. The method would not work as well if the derivatives of L were used, because L is not even approximately quadratic away from $\bar{\lambda}$.

To find an estimate for the best-fit value of λ, we need to calculate the derivatives in Eq. (11.87). In the case of the CMB, $\mathcal{L}$ is given explicitly by Eq. (11.20). Thus we need to differentiate the log of Eq. (11.20):

$$(\ln \mathcal{L})_{,\lambda} = \frac{\partial}{\partial \lambda}\left[-\frac{1}{2}\ln(\det C) - \frac{1}{2}\Delta C^{-1}\Delta\right]. \tag{11.88}$$

The covariance matrix here C depends on the theoretical parameter, λ. We can use the identity $\ln\det(C) = \mathrm{Tr}\ln(\mathrm{C})$ and the fact that $C^{-1}_{,\lambda} = -C^{-1}C_{,\lambda}C^{-1}$ to get

$$(\ln \mathcal{L})_{,\lambda} = \frac{1}{2}\Delta C^{-1}C_{,\lambda}C^{-1}\Delta - \frac{1}{2}\mathrm{Tr}[\mathrm{C}^{-1}\mathrm{C}_{,\lambda}]. \tag{11.89}$$

Here, the trace of $C^{-1}C_{,\lambda}$ is the sum of $(C^{-1})_{ij}\partial C_{ji}/\partial\lambda$ over all pixels i and j. Getting the second derivative requires more of the same types of steps. We find

$$(\ln \mathcal{L})_{,\lambda\lambda} = -\Delta C^{-1}C_{,\lambda}C^{-1}C_{,\lambda}C^{-1}\Delta + \frac{1}{2}\mathrm{Tr}[\mathrm{C}^{-1}\mathrm{C}_{,\lambda}\mathrm{C}^{-1}\mathrm{C}_{,\lambda}]$$

$$+ \frac{1}{2}\left(\Delta C^{-1}C_{,\lambda\lambda}C^{-1}\Delta - \mathrm{Tr}[\mathrm{C}^{-1}\mathrm{C}_{,\lambda\lambda}]\right). \tag{11.90}$$

Equation (11.90) gives the second derivative of $\ln \mathcal{L}$ with respect to the parameter λ. By definition, this is minus the *curvature* of the likelihood function:

$$\mathcal{F} \equiv -\frac{\partial^2 \ln \mathcal{L}}{\partial \lambda^2}. \tag{11.91}$$

The curvature is particularly important when evaluated at the maximum of the likelihood function, for there it measures how rapidly the likelihood falls away from the maximum (since the first derivative vanishes). If the curvature is small, then the likelihood changes slowly and the data are not very constraining: the resulting uncertainities on the parameter will be large. Conversely, large curvature translates into small uncertainties.[6]

We could take the ratio of Eqs. (11.89) and (11.90) to get an estimate for the maximum of the likelihood. This is *not* what is usually done. Rather, one typically sets $\Delta\Delta \rightarrow \langle \Delta\Delta \rangle = C$ in the second derivative. Upon doing this, the last line of Eq. (11.90) vanishes, and we are left with

$$\hat{\lambda} = \lambda^{(0)} + F_{\lambda\lambda}^{-1} \frac{\Delta C^{-1} C_{,\lambda} C^{-1} \Delta - \mathrm{Tr}[\mathrm{C}^{-1}\mathrm{C}_{,\lambda}]}{2} \tag{11.92}$$

where F is defined as

$$\begin{aligned} F_{\lambda\lambda} &\equiv \langle \mathcal{F} \rangle \\ &= \frac{1}{2}\mathrm{Tr}[\mathrm{C}_{,\lambda}\mathrm{C}^{-1}\mathrm{C}_{,\lambda}\mathrm{C}^{-1}]. \end{aligned} \tag{11.93}$$

That is, F is the average of the curvature over many realizations of signal and noise, both of which in this case are assumed to be drawn from Gaussian distributions. Here C and its derivatives are evaluated at the input point, $C(\lambda^{(0)})$.

There are several important features about Eq. (11.92), our estimator for the value of λ which maximizes the likelihood function. As the title of this subsection promised, it is a *quadratic* estimator: it is of the general form $A\Delta^2 + B$. The only hard part was determining the coefficients which lead to the best algorithm for finding the root. In the spirit of the Newton–Raphson method, Eq. (11.92) is best used iteratively. One assumes an input spectrum, uses it to determine a new input parameter ($\hat{\lambda}$ in Eq. (11.92)), then uses the new input parameter to find a new best-fit value, and so on until the process converges. In practice, analysts have found that very few iterations are needed until convergence. Nonetheless, we must be somewhat wary of quadratic estimators. It is possible that they will lead us to local maxima in parameter space. Finally, the foregoing discussion assumes there

[6]The correspondence between the curvature matrix and parameter uncertainties is often more quantitative than this. See *Numerical Recipes* (Press *et al.*, 1992), Chapter 15, for a detailed discussion. The bottom line, though, is that, in many cases, the inverse of the curvature matrix is a very good approximation to the error matrix on the parameters. E.g., in the one parameter case we are presently considering, the 1-σ, 68% confidence level on λ is approximately equal to the inverse square root of the curvature matrix.

is only one free parameter, but Eqs. (11.92) and (11.93) are easily generalizable to the more relevant case when many parameters are allowed to vary. If we have many parameters $\lambda_\alpha = \lambda_1, \lambda_2, \ldots$, then the quadratic estimator for each is

$$\hat{\lambda}_\alpha = \lambda_\alpha^{(0)} + F_{\alpha\beta}^{-1} \frac{\Delta C^{-1} C_{,\beta} C^{-1} \Delta - \mathrm{Tr}[\mathrm{C}^{-1}\mathrm{C}_{,\beta}]}{2} \tag{11.94}$$

where the *Fisher matrix* is defined as

$$\begin{aligned} F_{\alpha\beta} &\equiv \langle -\frac{\partial^2(\ln \mathcal{L})}{\partial\lambda_\alpha \partial\lambda_\beta} \rangle \\ &= \frac{1}{2}\mathrm{Tr}[\mathrm{C}_{,\alpha}\mathrm{C}^{-1}\mathrm{C}_{,\beta}\mathrm{C}^{-1}]. \end{aligned} \tag{11.95}$$

Putting aside the details which led to Eq. (11.94), we can appreciate that the result is remarkable. We can now hope to find the values of the parameters which maximize the likelihood function without blindly covering the whole parameter space. A very small number of matrix manipulations suffice to determine these best-fit values. This is a huge advantage, too good to pass up considering the alternative.

We still need to find a way to evaluate the errors on the parameters. Had we evaluated $\mathcal{L}$ everywhere, we could easily identify the region in parameter space ruled out at, say, the 95% level. How do we identify such regions using the quadratic estimator? To answer this question we need to remove ourselves from the derivation above and simply notice that Eq. (11.94) is an estimator for the true best-fit values of the parameters. If we think of it in this way — rather than as the result of a root-finding algorithm — we can study its distribution. Since the distributions for both signal and noise are known (they are assumed Gaussian with covariance matrices C_S and C_N, respectively), we can calculate the estimator's expectation value and variance.

First, let's consider its expectation value:

$$\langle \hat{\lambda}_\alpha \rangle = \lambda_\alpha^{(0)} + F_{\alpha\beta}^{-1} \frac{\langle \Delta C^{-1} C_{,\beta} C^{-1} \Delta \rangle - \mathrm{Tr}[\mathrm{C}^{-1}\mathrm{C}_{,\beta}]}{2}. \tag{11.96}$$

Here the covariance matrix, its derivatives, and the Fisher matrix have all been evaluated at the trial point $\lambda_\alpha = \lambda_\alpha^{(0)}$. The expectation value $\langle \Delta\Delta \rangle$ on the other hand is equal to the *true* covariance matrix, $C(\bar{\lambda}_\alpha)$. We can expand $C(\bar{\lambda}_\alpha)$ about $\lambda_\alpha^{(0)}$:

$$C(\bar{\lambda}_\alpha) \simeq C + C_{,\alpha'} \left(\bar{\lambda}_{\alpha'} - \lambda_{\alpha'}^{(0)} \right). \tag{11.97}$$

Therefore the expectation value of the quadratic estimator is

$$\langle \hat{\lambda}_\alpha \rangle = \lambda_\alpha^{(0)} + \frac{1}{2} F_{\alpha\beta}^{-1} \Big\{ \mathrm{Tr}\Big[\mathrm{C}^{-1}\mathrm{C}_{,\beta}\mathrm{C}^{-1}\mathrm{C}\Big]$$

$$+\mathrm{Tr}\left[\mathrm{C}^{-1}\mathrm{C}_{,\beta}\mathrm{C}^{-1}\mathrm{C}_{,\alpha'}\right]\left(\bar{\lambda}_{\alpha'}-\lambda_{\alpha'}^{(0)}\right)-\mathrm{Tr}[\mathrm{C}^{-1}\mathrm{C}_{,\beta}]\Bigg\}. \tag{11.98}$$

The first term in brackets cancels the third. The remaining trace is twice the Fisher matrix, so upon multiplying by F^{-1}, we are left with

$$\langle\hat{\lambda}_\alpha\rangle=\bar{\lambda}_\alpha. \tag{11.99}$$

So the quadratic estimator we have been considering is *unbiased*: the expectation values of the set of $\hat{\lambda}_\alpha$ are equal to the true parameters $\bar{\lambda}_\alpha$, no matter what set of parameters are assumed at the outset.

We are also interested in the variance of the estimator:

$$\langle(\hat{\lambda}_\alpha-\bar{\lambda}_\alpha)(\hat{\lambda}_\beta-\bar{\lambda}_\beta)\rangle=\left(F^{-1}\right)_{\alpha\beta}. \tag{11.100}$$

This equality, which I will leave as an exercise, holds if we are truly at the maximum of the likelihood function and if the data points really are distributed as a Gaussian. If these conditions hold, then the expected errors on the parameters are equal to the square root of the diagonal elements of F^{-1}. This is a magic limit, for there is a theorem, the *Cramer–Rao inequality*, that states that no method can measure the parameters with errors smaller than this (e.g., Kendall and Stuart, 1969). This makes sense since the errors from a full likelihood computation could not possibly be smaller than the width of the likelihood. This width in turn is determined by the curvature, and the Fisher matrix is simply the ensemble average of the curvature. Equation (11.100) tells us that on average, the quadratic estimator of Eq. (11.94) will reach this optimal limit.

Given any point in parameter space, we can calculate the associated Fisher matrix. Thus a simple way to assign error bars to the parameters determined via the quadratic estimator is to use the Fisher matrix evaluated at that point in parameter space. Bond, Jaffe, and Knox (1998), among others, have shown that this prescription works well: i.e., it agrees with a more complete tracing out of the likelihood contours.

Equation (11.100) is useful for other reasons as well. As is apparent from Eq. (11.95), the Fisher matrix — and hence the expected errors on any set of parameters — can be evaluated *without any data*. It will serve us well in Section 11.4.3 when we set out to determine how well upcoming experiments will be able to determine parameters.

We have derived the quadratic estimator in a way which might lead you to believe that it is restricted to the CMB. Namely, our derivation assumed that the likelihood function is Gaussian, true for the CMB but not for galaxy surveys. Even without the assumption of a Gaussian likelihood function, though, the quadratic estimator of Eq. (11.94) can be applied to galaxy surveys. Like any quadratic estimator, it has a mean and a variance. We have just seen that, for Gaussian distributions, it has the lowest variance possible. On large scales, where the galaxy distribution is Gaussian, therefore, it is extremely relevant. Even on small scales, where nonlinearities add to the variance, it is often competitive with other, more traditional estimators.

11.4 THE FISHER MATRIX: LIMITS AND APPLICATIONS

The Fisher matrix plays a key role in describing the ability of a given experiment to constrain parameters. It is difficult to gain much insight, though, from the definition in Eq. (11.95). Fortunately, in the case of full sky coverage, the Fisher matrix can be computed analytically. This analytic computation can then be extended — via a plausibility argument — to the more realistic case of partial sky coverage. This calculation is presented next for both the CMB and galaxy surveys. The most popular use of the Fisher matrix is as a tool for forecasting. How well do we expect a given experiment (even a hypothetical one) to determine cosmological parameters? The Fisher matrix is ideally suited for this task, and we will see some startling expectations from upcoming experiments.

11.4.1 CMB

The trace in Eq. (11.95) is a sum of the diagonal elements of the matrix $[C_{,\alpha}C^{-1}C_{,\beta}C^{-1}]_{ij}$ where i, j index the pixels used in the map. There are two decisions that need to be made before the Fisher matrix can be computed. First, what pixelization scheme should we use, and second what parameters λ_α are we interested in? For a full-sky CMB experiment, we choose as our parameters the C_l's themselves. That is, we take each individual C_l as a free parameter and ask how well an experiment can determine it. To avoid confusion (both the covariance matrix and the C_l's are C's), let's call each parameter λ_l instead of C_l, at least while working through the algebra. This answers the second question. The best way to deal with the first question — how to pixelize — in the case of the CMB is to use the a_{lm}'s. That is, instead of using the pixelized temperatures $\Theta(\hat{n})$, use

$$a_{lm} = \int d\Omega Y^*_{lm}(\hat{n})\Theta(\hat{n}) \tag{11.101}$$

as the data values. Each *pixel* then is labeled by l and m, so a given row (or column) in the covariance matrix corresponds to a fixed valued of l and m. Explicitly, since we start with the quadrupole,

$$C = \begin{pmatrix} C_{l=2,m=-2;l'=2,m'=-2} & C_{2,-2;2,-1} & \cdots & C_{2,-2;2,2} & C_{2,-2;3,-3} & \cdots \\ C_{2,-1;2,-2} & C_{2,-1;2,-1} & \cdots & C_{2,-1;2,2} & C_{2,-1;3,-3} & \cdots \\ & & & \vdots & & \\ C_{3,-3;2,-2} & C_{3,-3;2,-1} & \cdots & C_{3,-3;2,2} & C_{3,-3;3,-3} & \cdots \\ & & & \vdots & & \end{pmatrix}. \tag{11.102}$$

As usual the covariance matrix is the sum of the signal and noise covariance matrices. From Eq. (8.63), the signal covariance matrix would be $\delta_{ll'}\delta_{mm'}\lambda_l$ (remember that we're using λ_l instead of C_l) if the window function were unity. Let's assume that the experiment measures the anisotropy with a beam size σ. Then the signal covariance matrix must be multiplied by $e^{-l^2\sigma^2}$. The noise covariance matrix is a

little trickier. You will show in Exercise 11 that, in the case of uncorrelated, uniform noise, it is $\delta_{ll'}\delta_{mm'}w^{-1}$. Here w is the *weight* defined as

$$w = \left[(\Delta\Omega)\sigma_n^2\right]^{-1} \tag{11.103}$$

where $\Delta\Omega$ is the size in radians of the real space pixels and σ_n is the noise per pixel. Putting these two together, we have

$$C_{lm;l'm'} = \delta_{ll'}\delta_{mm'}\left[\lambda_l e^{-l^2\sigma^2} + w^{-1}\right]. \tag{11.104}$$

With these simple assumptions, we can take the inverse of the covariance matrix C and also find its derivative with respect to the parameters, the λ_l's. The inverse of the covariance matrix is

$$\left(C^{-1}\right)_{lm;l'm'} = \delta_{ll'}\delta_{mm'}\left[\lambda_l e^{-l^2\sigma^2} + w^{-1}\right]^{-1} \tag{11.105}$$

while the derivative of the covariance matrix with respect to the parameter λ_α is

$$C_{lm;l'm',\alpha} = \delta_{ll'}\delta_{mm'}\delta_{l\alpha}e^{-l^2\sigma^2}. \tag{11.106}$$

We can now construct the Fisher matrix; the only difficult task will be keeping track of indices. Very explicitly,

$$\begin{aligned} F_{\alpha\alpha'} &= \frac{1}{2}C_{lm;l'm',\alpha}C^{-1}_{l'm';l''m''}C_{l''m'';l'''m''',\alpha'}C^{-1}_{l'''m''';lm} \\ &= \frac{1}{2}\left(\delta_{ll'}\delta_{mm'}\delta_{l\alpha}e^{-l^2\sigma^2}\right)\left(\frac{\delta_{l'l''}\delta_{m'm''}}{\lambda_{l'}e^{-l'^2\sigma^2}+w^{-1}}\right)\left(\delta_{l''l'''}\delta_{m''m'''}\delta_{l''\alpha'}e^{-l''^2\sigma^2}\right) \\ &\quad\times\left(\frac{\delta_{l'''l}\delta_{m'''m}}{\lambda_l e^{-l^2\sigma^2}+w^{-1}}\right) \end{aligned} \tag{11.107}$$

with the implicit sum over $ll'l''l'''mm'm''m'''$. Consider first the Kronecker deltas with subscripts m, m', m'', and m'''. There are four of these; summing over all the subscripts besides m contracts these four to

$$\sum_{m'm''m'''}\delta_{mm'}\delta_{m'm''}\delta_{m''m'''}\delta_{m'''m} = \delta_{mm}. \tag{11.108}$$

Then, summing over all m leads to a factor of $2l+1$. The remaining sums over l and its cousins leads to a simple factor of $\delta_{\alpha\alpha'}$. Therefore in the all-sky limit, the Fisher matrix for a CMB experiment is

$$F_{ll'} = \frac{2l+1}{2}\delta_{ll'}e^{-2l^2\sigma^2}\left[C_l e^{-l^2\sigma^2} + w^{-1}\right]^{-2} \tag{11.109}$$

where I have gone back to C_l here since the covariance matrix does not appear.

In an all-sky survey, therefore, the Fisher matrix for the C_l's is diagonal. There are no correlations between adjacent C_l's. The errors on a given C_l expected from an all-sky experiment can be read off from Eq. (11.109). The errors are equal to $\sqrt{F^{-1}}$, so

$$\delta C_l = \sqrt{\frac{2}{2l+1}} \left[C_l + w^{-1} e^{l^2 \sigma^2} \right]. \tag{11.110}$$

As anticipated in Eq. (11.26), there are thus two sources of error: (i) cosmic variance, proportional to the signal itself C_l and (ii) noise — atmospheric or instrumental — as encoded in the weight w and the smoothing determined by the beam width σ. The factor of $2l + 1$ in the denominator also traces back to Eq. (11.26); it is the number of independent samples used to estimate a given C_l.

No experiment will ever cover the entire sky, since the CMB cannot be observed in the plane of our galaxy. Even MAP and Planck, two satellites designed to map CMB anisotropy from space, will therefore cover a fraction of the sky $f_{\rm sky} < 1$. Recalling that the factor of $2l + 1$ in the denominator of Eq. (11.26) counts the number of samples, we could guess that this factor must be multipled by $f_{\rm sky}$. This leads to

$$\delta C_l = \sqrt{\frac{2}{(2l+1) f_{\rm sky}}} \left[C_l + w^{-1} e^{l^2 \sigma^2} \right]. \tag{11.111}$$

This formula enables one to project the errors obtainable for any given experiment. The three characteristics of the experiment which determine the error on C_l are the sky coverage; the weight; and the beam width.

11.4.2 Galaxy Surveys

The analogue of the all-sky CMB experiment is a volume-limited galaxy survey as the volume gets arbitrarily large. This limit applies to all modes with wavelengths k^{-1} much smaller than the typical size of the survey. We have already computed the signal and noise covariance matrices in this limit, Eqs. (11.60) and (11.61), so the covariance matrix for Fourier pixels is

$$C_{\vec{k}_i, \vec{k}_j} = \frac{\delta_{ij}}{V} \left(P(k_i) + \frac{1}{\bar{n}} \right). \tag{11.112}$$

To compute the Fisher matrix, we will need the inverse of this and its derivative with respect to whatever parameters we choose. The inverse is simple,

$$C^{-1}_{\vec{k}_i, \vec{k}_j} = \frac{\delta_{ij} V}{P(k_i) + \frac{1}{\bar{n}}}. \tag{11.113}$$

For our parameters, we will choose the amplitude of the power spectrum in a set of narrow k-bins, each with width Δk. The power in the bin with $k_\alpha < k < k_\alpha + \Delta k$

will be denoted P_α. The derivative of the covariance matrix with respect to P_α is therefore

$$C_{ij,\alpha} \equiv \frac{\partial C_{\vec{k}_i,\vec{k}_j}}{\partial P_\alpha} = \frac{\delta_{ij}}{V} d_{i\alpha} \tag{11.114}$$

with

$$d_{i\alpha} \equiv \begin{cases} 1 & k_\alpha < |\vec{k}_i| < k_\alpha + \Delta k \\ 0 & \text{otherwise} \end{cases} . \tag{11.115}$$

The Fisher matrix is now

$$\begin{aligned} F_{\alpha\alpha'} &= \frac{1}{2} C_{ij,\alpha} C^{-1}_{jj'} C_{j'i',\alpha'} C^{-1}_{i'i} \\ &= \frac{1}{2} \delta_{ij} d_{i\alpha} \frac{\delta_{jj'}}{P(k_j) + \frac{1}{\bar{n}}} \delta_{j'i'} d_{i'\alpha'} \frac{\delta_{i'i}}{P(k_i) + \frac{1}{\bar{n}}} . \end{aligned} \tag{11.116}$$

The sums over j, j, i' are straightforward, leaving

$$F_{\alpha\alpha'} = \frac{1}{2} \frac{\sum_i d_{i\alpha} d_{i\alpha'}}{\left(P_\alpha + \frac{1}{\bar{n}}\right)\left(P_{\alpha'} + \frac{1}{\bar{n}}\right)} . \tag{11.117}$$

As long as the k-bins do not overlap, k_i cannot be in two different bins so the product $d_{i\alpha} d_{i\alpha'}$ requires $\alpha = \alpha'$. The sum then is over all vectors $\vec{k}_i$ in a spherical shell with radius k_α and width Δk. This sum is $4\pi k_\alpha^2 \Delta k V$. Therefore, the Fisher matrix is diagonal, with elements

$$F_{\alpha\alpha'} = \delta_{\alpha\alpha'} \frac{4\pi k_\alpha^2 \Delta k V}{2\left(P_\alpha + \frac{1}{\bar{n}}\right)^2} . \tag{11.118}$$

The error on the power spectrum in this limit is the inverse square root,

$$\delta P_\alpha = \sqrt{\frac{2}{4\pi k_\alpha^2 \Delta k V}} \left(P_\alpha + \frac{1}{\bar{n}}\right) . \tag{11.119}$$

This is identical in form to the errors on the C_l's. The denominator in the prefactor counts the number of modes in a given estimate; the first term is cosmic variance; and the last is the noise, in this case Poisson noise.

11.4.3 Forecasting

One of the great promises of upcoming cosmological experiments is that they will determine many of the presently unknown cosmological parameters. How does one predict the expected uncertainties in cosmological parameters from future experiments? The answer is surprisingly simple. Let's consider a CMB experiment as an example. Start with the following:

- A set of C_l's that are assumed to describe the true universe

- The uncertainty on C_l from a given experiment, δC_l, assumed to be given by Eq. (11.111)
- The set of cosmological parameters, $\{\lambda_\alpha\}$, for which we want to forecast errors

The observed C_l's in this universe will be close to the true C_l's; indeed, if we form

$$\chi^2\left(\{\lambda_\alpha\}\right) = \sum_l \frac{\left(C_l\left(\{\lambda_\alpha\}\right) - C_l^{\rm obs}\right)^2}{(\delta C_l)^2}, \tag{11.120}$$

then we expect this χ^2 to reach a minimum at the point in parameter space where $\lambda_\alpha = \bar{\lambda}_\alpha$, the actual values of the parameters. Of course, we do not now know what those values are, but even without that information, we can ask how quickly $\chi^2\Big(\{\lambda_\alpha\}\Big)$ changes as λ_α moves away from $\bar{\lambda}_\alpha$. If it increases rapidly, then the errors on the parameters will be very small; if the χ^2 changes little, then there will be large errors on the parameters.

To quantify this, we can expand χ^2 about its minimum at $\bar{\lambda}_\alpha$. Let's first do this in the case of one parameter; the generalization to many parameters will be straightforward. In the one-parameter case,

$$\chi^2(\lambda) = \chi^2(\bar{\lambda}) + \mathcal{F}(\lambda - \bar{\lambda})^2. \tag{11.121}$$

The linear term in Eq. (11.121) vanishes since χ^2 is a minimum at $\bar{\lambda}$. The coefficient of the quadratic term is

$$\mathcal{F} = \frac{1}{2}\frac{\partial^2\chi^2}{\partial\lambda^2}\bigg|_{\lambda=\bar{\lambda}}. \tag{11.122}$$

The curvature here, $\mathcal{F}$, measures how rapidly the χ^2 changes away from its minimum. As such, the error on λ is simply $1/\sqrt{\mathcal{F}}$. So all we need to do in order to determine the expected errors on a parameter is calculate $\mathcal{F}$. Note that $\mathcal{F}$ is the curvature of the likelihood function only if the likelihood function is equal to $e^{-\chi^2/2}$, that is, if the errors on the C_l's are Gaussian distributed. In fact, they are not, so $\mathcal{F}$ as given by Eq. (11.122) is not really the curvature, $-\partial^2 \ln\mathcal{L}/\partial\lambda^2$. Nonetheless, the distribution is close enough to Gaussian that the error estimates that arise are expected to be quite accurate.

The second derivative of χ^2 contains two terms:

$$\mathcal{F} = \sum_l \frac{1}{(\delta C_l)^2}\left[\left(\frac{\partial C_l}{\partial\lambda}\right)^2 + (C_l - C_l^{\rm obs})\frac{\partial^2 C_l}{\partial\lambda^2}\right]. \tag{11.123}$$

The second term in the sum over l is traditionally neglected. The idea (as elucidated in *Numerical Recipes*) is that the difference $C_l - C_l^{\rm obs}$ will sometimes be negative, sometimes positive. On average, there will be much cancellation, so the first term will dominate. Thus, the general practice is to take

$$\mathcal{F} \to \sum_l \frac{1}{(\delta C_l)^2}\frac{\partial C_l}{\partial\lambda}\frac{\partial C_l}{\partial\lambda}. \tag{11.124}$$

Equivalently, you can think of the dropping of the second term as taking an average over the whole distribution, thereby replacing the curvature matrix with the Fisher matrix (although again keeping in mind that this is not the true curvature or Fisher matrix since the C_l's are not distributed as Gaussians). The generalization of this to many parameters is simply

$$F_{\alpha\beta} = \sum_l \frac{1}{(\delta C_l)^2} \frac{\partial C_l}{\partial \lambda_\alpha} \frac{\partial C_l}{\partial \lambda_\beta}. \tag{11.125}$$

In order to predict how accurately parameters will be known, then, we simply need to know the experiment's specifications (to determine δC_l) and the derivatives of the C_l's around their (assumed) true values.

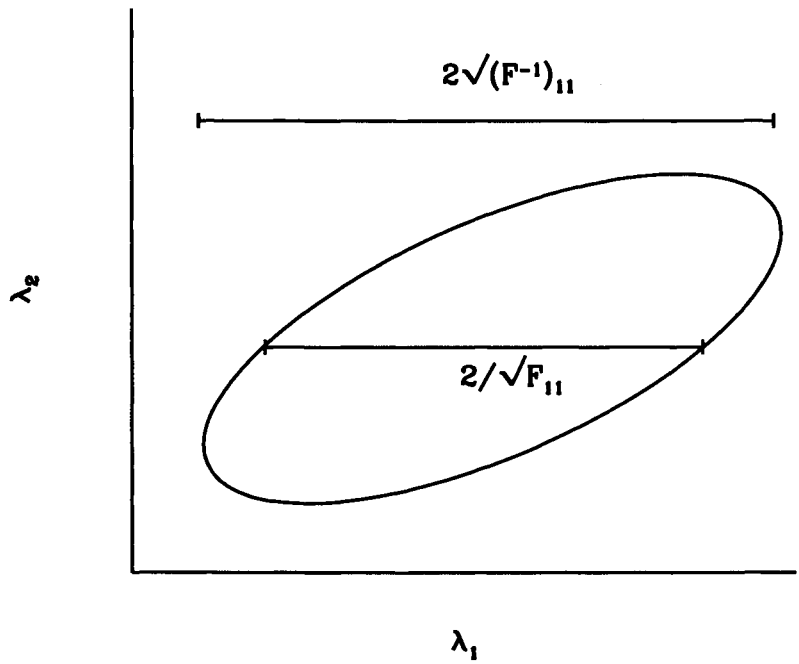

Figure 11.13. Error ellipse in a 2D parameter space. If no prior information is known about λ_2, then the error on λ_1 is $\sqrt{(F^{-1})_{11}}$. If, however, λ_2 is fixed, the error on λ_1, $(F_{11})^{-1/2}$ is smaller.

Assuming a Gaussian distribution, the one-sigma uncertainty on a one parameter fit is $1/\sqrt{F}$. How about if more than one parameter is allowed to vary? Figure 11.13 illustrates the situation in a two-dimensional setting. If the parameter λ_2 is assumed known, then the error on λ_1 is still $1/\sqrt{F_{11}}$. However, if λ_2 is allowed to vary, the error on λ_1 is now $\sqrt{(F^{-1})_{11}}$. It is instructive to prove this explicitly by noting that we are assuming that the joint probability for the two parameters is

$$P(\lambda_1, \lambda_2) \propto \exp\left\{-\frac{1}{2}\lambda_i F_{ij} \lambda_j\right\} \tag{11.126}$$

where I have assumed that the distribution peaks at $\lambda_i = 0$ for simplicity. Allowing λ_2 to vary is equivalent to integrating this probability distribution over all possible values of λ_2. This is referred to as *marginalizing* over λ_2. Then,

$$\begin{aligned} P(\lambda_1) &= \int d\lambda_2 P(\lambda_1, \lambda_2) \\ &\propto \exp\left\{-\frac{\lambda_1^2}{2}\left(\frac{F_{11}F_{22} - F_{12}F_{21}}{F_{22}}\right)\right\} \end{aligned} \tag{11.127}$$

where the second line comes from carrying out the λ_2 integration explicitly. The term in parentheses in the exponential—$[F_{11}F_{22} - F_{12}F_{21}]/F_{22}$—is simply equal to $1/(F^{-1})_{11}$. So the one-sigma error is indeed given by $\sqrt{(F^{-1})_{11}}$.

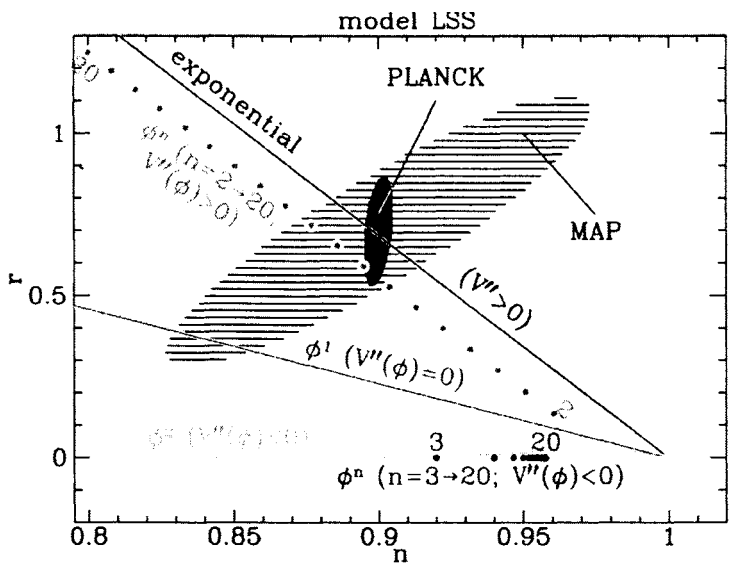

Figure 11.14. Expected 95% uncertainty on the inflationary parameters n and r from MAP and Planck (from Dodelson, Kinney, and Kolb, 1997). (See color Plate 11.14.) Three other parameters (normalization, Ω_B, and h) have been marginalized over. Every inflationary model gives a unique prediction somewhere in this plane; many such predictions are plotted.

Figure 11.14 shows the expected uncertainties from MAP and Planck for two inflationary parameters, those which determine the primordial spectrum. Note that I have fixed the "true" model to be one in which the spectral index $n = 0.9$ and the tensor-to-scalar ratio $r = 0.7$. Different fiducial models often lead to quite different error bars. The ellipse in Figure 11.14 has been drawn after marginalizing over

Table 11.1. Marginalized Errors for ΛCDM for various experiments.

	Map	Planck	Map+SDSS	Planck+SDSS
h	0.22	0.13	0.029	0.022
Ω_m	0.24	0.14	0.036	0.027
Ω_Λ	0.19	0.11	0.042	0.024
$\ln(\Omega_b h^2)$	0.060	0.010	0.050	0.010
m_ν (eV)	0.58	0.26	0.33	0.21
Y_P	0.020	0.013	0.020	0.013
n	0.048	0.008	0.040	0.008
r	0.18	0.012	0.16	0.012
τ	0.022	0.004	0.021	0.004

three other cosmological parameters: normalization, baryon density, and Hubble constant. To do this, one starts with the five-dimensional Fisher matrix, inverts, and then considers only the 2×2 part of the inverse. This 2×2 part defines the ellipses drawn in the figure.

Also plotted in Figure 11.14 are the predictions from a wide variety of inflationary models. The figure argues persuasively that we will indeed be able to distinguish among different inflationary models with upcoming CMB experiments. This is a remarkable finding: we will learn about physics at 10^{15} GeV or higher using CMB data. If more parameters are allowed to vary, then the errors on any one parameter naturally get larger. However, this can be offset by using other observations, most notably those from large-scale structure. Table 11.1 presents the marginalized errors on a number of parameters expected from the MAP, Planck, and Sloan Digital Sky Survey experiments (Eisenstein, Hu, and Tegmark, 1999). Here the errors include the measurement of polarization in MAP and more importantly Planck.

11.5 MAPMAKING AND INVERSION

Until now, we have discussed ways of analyzing a *map*, a collection of pixels with data Δ and noise covariance matrix C_N. How does one make such a map? How do we go from the timestream of data to a pixelized set of spatial Δ's? Most work on this issue has focused on the temperature maps in the CMB, so I'll use this as an example. We will see that mapmaking is essentially an inversion problem. So the mapmaking techniques discussed here are applicable to a broad range of problems in physics and astronomy.

Let's first state the problem. An experiment amounts to a *timestream* of data, d_t. Each number in the timestream corresponds to data taken at a given point on the sky. The data are assumed to be composed of signal plus noise:

$$d_t = P_{ti} s_i + \eta_t. \tag{11.128}$$

Here subscript t denotes a given element in the timestream; i denotes spatial pixel; P is the pointing matrix; s is the underlying temporally constant, but spatially varying signal; and η is the temporal noise. The pointing matrix encodes information about where the receiver is pointing. That is, it associates with each temporal measurement t a given pixel i. It is an $N_t \times N_p$ matrix where N_t is the number of temporal measurements and N_p is the number of spatial pixels. The pointing matrix has a special form: every row has only one nonzero entry equal to 1, corresponding to the pixel on the sky being observed at the time denoted by the row. Each column, however, typically has many nonzero entries corresponding to all the times a given spot has been observed. The noise η is assumed to be Gaussian with a covariance matrix N. There are techniques to determine N directly from the data, but to simplify the discussion, we will assume that N is known.

What is the best way to make a map from this timestream? One forms a χ^2,

$$\chi^2 \equiv \sum_{tt'ij} (d_t - P_{ti}s_i)\, N^{-1}_{tt'}\, (d_{t'} - P_{t'j}s_j)\,, \tag{11.129}$$

and minimizes with respect to the signal s. Indeed, if the noise is Gaussian, then the likelihood function is proportional to $e^{-\chi^2/2}$ and minimizing the χ^2 is equivalent to maximizing $\mathcal{L}$. Taking the derivative with respect to s_i leads to

$$\frac{\partial \chi^2}{\partial s_i} = -2 \sum_{tt'j} P_{ti} N^{-1}_{tt'} \left(d_{t'} - P_{t'j}s_j\right). \tag{11.130}$$

Set the derivative equal to zero:

$$\sum_{tt'j} P_{ti} N^{-1}_{tt'} P_{t'j} s_j = \sum_{tt'j} P_{ti} N^{-1}_{tt'} d_{t'}. \tag{11.131}$$

The terms multiplying s_j on the left are an $N_p \times N_p$ matrix,

$$\left(C_N^{-1}\right)_{ij} \equiv \sum_{tt'} P_{ti} N^{-1}_{tt'} P_{t'j}. \tag{11.132}$$

Multiply both sides by the inverse of this (C_N itself) to find that the χ^2 is minimized when s is equal to

$$\Delta_i = C_{N,ij} P_{ti} N^{-1}_{tt'} d_{t'}. \tag{11.133}$$

In matrix notation,

$$\Delta = C_N P^T N^{-1} d \tag{11.134}$$

where ^{T} denotes transpose. The noise matrix of this map is equal to

$$C_N = \left(P^T N^{-1} P\right)^{-1}, \tag{11.135}$$

a fact which you can verify by taking $\langle \Delta\Delta \rangle$.

A simple limit of Eq. (11.134) emerges when the timestream noise is diagonal and uniform (this is unrealistic). In that case, the elements of C_N become

$$C_{N.ij} \to N \left(\sum_t P_{it}^T P_{tj} \right)^{-1} \tag{11.136}$$

where now N is simply a number, the diagonal element of the timestream noise. Recall that for a given t, P_{ti} is nonzero for only one pixel i. So the product $P_{it}^T P_{tj}$ vanishes unless $i = j$ and the receiver at time t was pointing at pixel i. Thus the sum over t counts the number of times the receiver sampled pixel i; call this number π_i. In this artificial case of uniform, uncorrelated noise, therefore, the noise covariance matrix for the map is diagonal with elements N/π_i. This makes sense: as a given pixel is sampled more times, the standard deviation goes down as $\pi_i^{-1/2}$. The map now becomes

$$\Delta_i \to \frac{1}{\pi_i} \sum_t P_{it}^T d_t. \tag{11.137}$$

That is, one simply averages all the data points corresponding to the given pixel (exactly like Eq. (11.8)).

Figure 11.15 shows a more realistic implementation of Eq. (11.134), from the long-duration balloon flight of Boomerang (Netterfield *et al.*, 2001), launched on December 29, 1998. The map covers a region a region of about 700 square degrees with 7-arcminute pixels. Therefore, the map required $N_p \simeq 50,000$, while the timestream contained of order 2×10^8 data points. Some tricks were needed to avoid direct inversion and multiplication of the large matrices P and N. Nonetheless, the basis for the Boomerang map, indeed for all CMB maps, is Eq. (11.134).

The raw data from which the map is made need not be the timestream. Instead, the raw data could consist of a series of modulations, e.g., the difference between the temperature at two points. Reconstructing a map from a set of modulations sounds like a much different problem than doing the same from the timestream. In fact, it is mathematically identical: the data d is the sum of a signal and noise. The signal can be thought of as a matrix acting on the underlying temperature field. This matrix does not have the exact form as the pointing matrix (i.e., only one nonzero element in each row), but it is a matrix nonetheless, and all the operations in Eqs. (11.134) and (11.135) can be carried out. There is a big advantage in using a map constructed in this fashion as opposed to the modulated data themselves. Ultimately, the main use of the data will be to estimate parameters in a likelihood analysis. As we have seen, one must construct the signal covariance matrix in order to carry out such an analysis. The signal covariance matrix for a map is extremely simple—the window function is simply $P_l(\cos\theta_{ij})$—whereas that for modulated data is quite difficult to obtain (recall Section 11.2).

Equation (11.134) and the corresponding noise matrix in Eq. (11.135) are even more general than this. They apply to any problem in which the data are a sum of some processed signal and noise, i.e., to an extraordinarily wide range of problems in physics and astronomy. Consider just two examples. First, the angular correlation

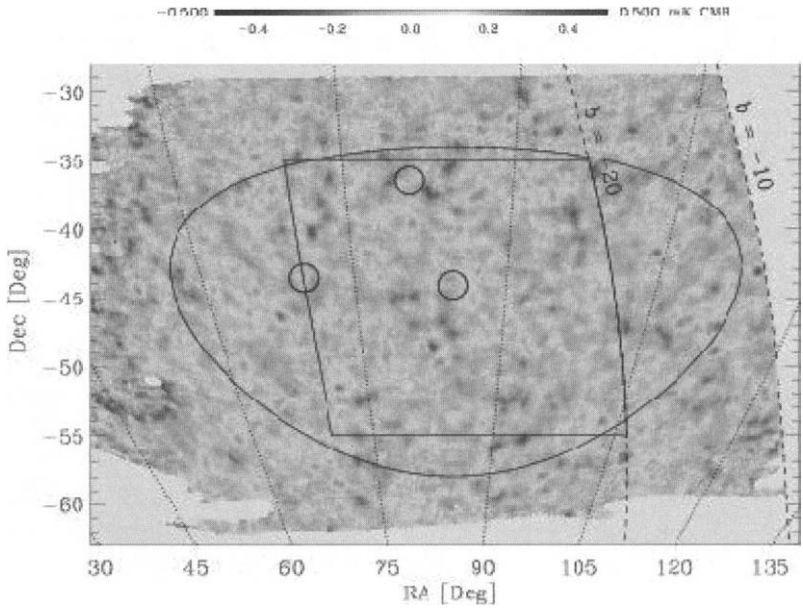

Figure 11.15. A map of the CMB temperature from observations by Boomerang (Netterfield *et al.*, 2001), a long-duration balloon flight at the South Pole. (See also color Plate 11.15.) Hot and cold spots have amplitudes as large as 500μK. Circles shows quasars identified in these radio observations. The large elliptical region delineates data analyzed to obtain bandpowers. The rectangular region is an earlier data set.

function is a sum of the 3D power spectrum processed by a kernel plus noise. One can apply Eq. (11.134) directly to obtain the 3D power spectrum, simply replacing the pointing matrix with the kernel. In the next section, we will see another example, the extraction of the CMB signal from data contaminated by foregrounds.

11.6 SYSTEMATICS

A *systematic* error is one which remains even after averaging over many data samples. Systematic errors are the most worrisome aspect of most cosmological observations. Knowing this, observers typically take many precautions against them, submitting raw data to a wide variety of consistency checks. Many of these tests are the result of common sense and intuition. Nothing much formal can be said about them. Here I want to focus on several ... systematic ways of dealing with such effects.

11.6.1 Foregrounds

One of the biggest obstacles to observing the anisotropies in the CMB are *foregrounds*, other sources of radiation which also emit at microwave frequencies. The

list of foregrounds is long and includes anything in space that might come between us and the radiation left over from the Big Bang. There is dust, synchrotron radiation, and free-free or bremsstrahlung emission, all emanating from our galaxy (but extending to regions of the sky far from the galactic plane). There are also extragalactic sources of radiation, point sources and clusters of galaxies. All of these have the potential to contaminate an experiment searching only for a cosmic signal. The magnitude of this problem is hinted at in the nomenclature. Usually in science, a possible source of systematic error is called a *background*. In CMB physics, we cannot call these things backgrounds: the "B" in CMB precludes that possibility. We must acknowledge that the cosmic signal is coming from farther away than any possible contaminant and we must deal with the real possibility that the cosmic signal will be smaller than some of these contaminants.

The problem of foregrounds has in the past few years been demonstrated to be manageable. There are a number of reasons for this good fortune (it is good fortune: if we were living deeper in the galaxy foreground amplitudes would be considerably larger). First observers have been very successful at finding the coolest portions of the sky and using only these regions. Second, foreground amplitudes have proven to be smaller than the cosmic signal in a fairly wide part of frequency space. Figure 11.16 shows the intensity of several galactic foregrounds and the CMB. The amplitudes of each of these components varies across the sky, but the relative amplitudes shown in Figure 11.16 are fairly typical. At very high frequencies, dust dominates, and at very low frequencies synchrotron and bremsstrahlung become important. But, in the range from 30 to 200 GHz, the CMB anisotropies often have the largest intensities.

The final reason foregrounds can be managed — and the one I want to focus on here — can also be gleaned from Figure 11.16. The spectral shapes of the foregrounds are all different, different from one another and from the blackbody shape of the CMB anisotropy. This raises the possibility that detections at different frequencies can be used to extract the CMB signal from the foregrounds. As analysts, we must find the optimal way to perform this extraction. Given measurements at several different frequencies, what is the best algorithm for finding the cosmic signal? How effective do we expect this extraction to be?

First we need to set up some notation. Instead of intensity (or brightness) B, which has units of ergs cm^{-2} sec^{-1} Hz^{-1} , it is convenient to introduce the brightness temperature or *antenna temperature*, defined as

$$\begin{aligned} T_{\mathrm{ant}} &\equiv \frac{B}{2\nu^2} \\ &= 2\pi\nu f, \end{aligned} \tag{11.138}$$

which has dimensions of kelvins (recall that we are working in units with $k_B = \hbar = c = 1$). The frequency ν is related to the momentum we used in earlier chapters via $p = 2\pi\nu$ and f is the familiar phase space density.

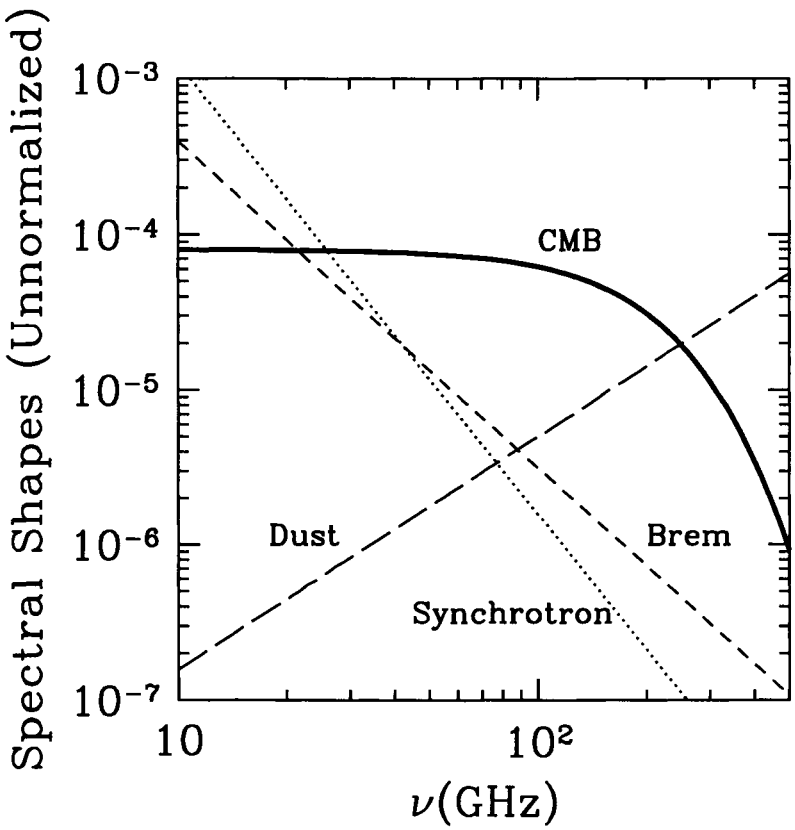

Figure 11.16. Spectral shapes of the dominant galactic foregrounds and the CMB blackbody temperature anisotropy.

For the CMB we know that f is given by Eq. (4.35), so the antenna temperature is

$$\frac{T_{\rm ant}^{\rm cmb}}{T} = \frac{x}{e^x - 1} + \Theta \frac{x^2 e^x}{(e^x - 1)^2} \tag{11.139}$$

where

$$x \equiv \frac{p}{T} = \frac{2\pi\nu}{T}. \tag{11.140}$$

The first term in Eq. (11.139) is the monopole, which does not interest us. The second contains information about the shape of the spectrum of the CMB anisotropies. It is useful then to neglect the first term and write

$$\frac{T_{\rm ant}^{\rm cmb}(\nu)}{T} = \Theta W^{\rm cmb}(2\pi\nu/T) \tag{11.141}$$

where W^α will be the *shape* vector for different components. That for the CMB is

$$W^{\rm cmb} = \frac{x^2 e^x}{(e^x - 1)^2}. \tag{11.142}$$

Every component can be written in this fashion: the product of an amplitude (in this case Θ), which is frequency independent, and the shape vector, which carries no information about the amplitude. For a blackbody distribution like the CMB, the amplitude has a name: Θ is called the *thermodynamic temperature*. Note that the CMB shape vector goes to 1 in the limit of small frequencies, the Rayleigh–Jeans limit. At high frequencies, it falls off exponentially.

I have called W a shape *vector* implying that it has a number of components. These components are the different frequency channels at which a given experiment takes data. We will label these with a subscript c while allowing for many possible foreground components:

$$\vec{W}^\alpha = W_c^\alpha \qquad c = 1, \ldots, N_{\rm ch} \quad ; \quad \alpha = 0, \ldots, N_{\rm foregrounds}. \tag{11.143}$$

Here the CMB component is associated with index $\alpha = 0$. Thus, the data from a given experiment in a given pixel (we focus on only one spatial pixel) on the sky is in the form of a set of antenna temperatures, d_c, at all the different frequency channels. Our model is that this data set is the sum of the contributions from the CMB, foregrounds, and Gaussian noise:

$$d_c = \sum_{\alpha=0}^{N_{\rm foregrounds}} W_c^\alpha \Theta^\alpha + n_c. \tag{11.144}$$

We assume that we know the covariance matrix of the noise N and all the spectral shapes[7] W_c^α. The question is, how do we best determine the CMB anisotropy?

This problem has the exact same form as does the general inversion problem of Section 11.5. We want to obtain estimates of the amplitude Θ^α. We can immediately write down the minimum variance estimator for Θ^α:

$$\Delta^\alpha = (C_N)_{\alpha\beta} W_c^\beta N_{cd}^{-1} d_d \tag{11.145}$$

with covariance matrix

$$\left(C_N^{-1}\right)_{\alpha\beta} = W_c^\alpha (N^{-1})_{cd} W_d^\beta. \tag{11.146}$$

Let's work out a simple example to bring these formulae, which are so ubiquitous, to life.

[7]This assumption is true for the CMB, and very nearly true for synchrotron and bremsstrahlung which have shapes which are fairly constant over the sky. The assumption is worst for dust. A number of groups have explored the consequences of incorrect assumptions about the shapes or allowed for some freedom in the shapes.

Consider an experiment taking measurements at two frequencies, both in the Rayleigh–Jeans regime. The shape vector for the CMB then has two components, both equal to 1:

$$\vec{W}^0 = (1, 1). \tag{11.147}$$

Let's also assume that the noise is uncorrelated from one frequency channel to the next and is uniform with diagonal element, σ_n^2. First let's consider the case of zero foregrounds. In that case, the covariance matrix is just a single number, the inverse of

$$\begin{aligned} WN^{-1}W &\rightarrow \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} 1/\sigma_n^2 & 0 \\ 0 & 1/\sigma_n^2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ &= \frac{2}{\sigma_n^2}. \end{aligned} \tag{11.148}$$

The inverse square root of this is the noise, $\sigma_n/\sqrt{2}$. The two channels reduce the noise by a factor of the square root of 2 (if there were three channels in our example, the factor would be $\sqrt{3}$, etc.). The mimimum variance estimator is given by Eq. (11.145),

$$\begin{aligned} \Delta &\rightarrow \frac{\sigma_n^2}{2}(1,1) \begin{pmatrix} 1/\sigma_n^2 & 0 \\ 0 & 1/\sigma_n^2 \end{pmatrix} \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} \\ &= \frac{d_1 + d_2}{2}. \end{aligned} \tag{11.149}$$

We simply average the two data points.

Now suppose there is one foreground to worry about, say synchrotron emission, with shape vector

$$\vec{W}^1 = (1, 1/2). \tag{11.150}$$

Typically, the intensity of synchrotron emission scales as ν^{-1} (see, e.g., Rybicki and Lightman, 1979), so its antenna temperature falls off as ν^{-3}. Thus a shape vector $(1, 1/2)$ follows from observing at, e.g., $\nu_1 = 20$ GHz and $\nu_2 = 25$ GHz.

Now the covariance matrix C_N is a two by two matrix. Then Eq. (11.146) becomes

$$\begin{aligned} C_N^{-1} = W_c^\alpha (N^{-1})_{cd} W_d^\beta &= \begin{pmatrix} 1 & 1 \\ 1 & 1/2 \end{pmatrix} \begin{pmatrix} 1/\sigma_n^2 & 0 \\ 0 & 1/\sigma_n^2 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 1/2 \end{pmatrix} \\ &= \frac{1}{\sigma_n^2} \begin{pmatrix} 2 & 3/2 \\ 3/2 & 5/4 \end{pmatrix}. \end{aligned} \tag{11.151}$$

Already at this stage, we can glean some important information about the experiment. Recall from the discussion in Section 11.4 that the inverses of the diagonal elements of C_N^{-1} are the unmarginalized variances, the errors if all other parameters are known. In this case, there is one parameter besides the amplitude of the CMB,

the amplitude of the foreground. If we assume that is known, then the error on the CMB temperature will be the inverse square root of the $_{00}$ component of the matrix in Eq. (11.151). This is $\sigma_n/\sqrt{2}$, in agreement with the calculation above. To find the errors if we know nothing about the foreground amplitudes, take the inverse of this to find the covariance matrix,

$$C_N = 4\sigma_n^2 \begin{pmatrix} 5/4 & -3/2 \\ -3/2 & 2 \end{pmatrix}. \tag{11.152}$$

The $_{00}$ component of this gives the marginalized variance, $5\sigma_n^2$. The ratio of the marginalized error to the unmarginalized error is a measure of how severely the unknown foregrounds degrade our ability to determine the CMB temperature. It is called the *foreground degradation factor*, or simply the FDF. In this case, the FDF is equal to $\sqrt{10}$.

In this example, we can now determine the minimum variance estimator for the CMB temperature. Following Eq. (11.145), we write

$$\begin{aligned} \Delta^0 &= 4\sigma_n^2 \, (5/4, -3/2) \begin{pmatrix} 1 & 1 \\ 1 & 1/2 \end{pmatrix} \begin{pmatrix} 1/\sigma_n^2 & 0 \\ 0 & 1/\sigma_n^2 \end{pmatrix} \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} \\ &= -d_1 + 2d_2. \end{aligned} \tag{11.153}$$

This should be no surprise. The best estimator for the CMB temperature is completely insensitive to the amplitude of the foreground component. For, if the foreground really does have shape vector $(1, 1/2)$, the linear combination $-d_1 + 2d_2$ from the foreground is equal to zero.

In real life, one must find the minimum variance estimators at many different spatial pixels. The formula of Eq. (11.145) remains identical in this more general case: one simply adds an index for spatial pixel. It is often most convenient to work with the a_{lm}'s instead of the temperature as a function of angular coordinates. Then, the minimum variance estimator for a_{lm} often weights the different frequency channels differently depending upon how the noise in each channels scales with l. An example is shown in Figure 11.17, based on the five frequencies of the MAP experiment. The figure shows how the best estimator for a_{lm} weighs the five different frequencies in the absence of foregrounds. At low l, all the channels have similar noise, so the best estimator is just the average of the five. The beam size is frequency dependent, however, largest at the lowest frequencies. Therefore, at high l, the lowest channels have no signal. Only the highest frequency channel can be used. Indeed, one sees that the minimum variance estimator gradually drops a channel at a time as l gets larger.

Often prior information on the foregrounds exists, in the form of an estimate of the power spectrum of each foreground component, i.e., its C_l. This prior information can be incorporated into the minimum variance estimate; see, e.g., Exercise 15. Figure 11.18 shows the results of accounting for foregrounds in the MAP experiment. The difference between the weighting scheme in this figure and that in Figure

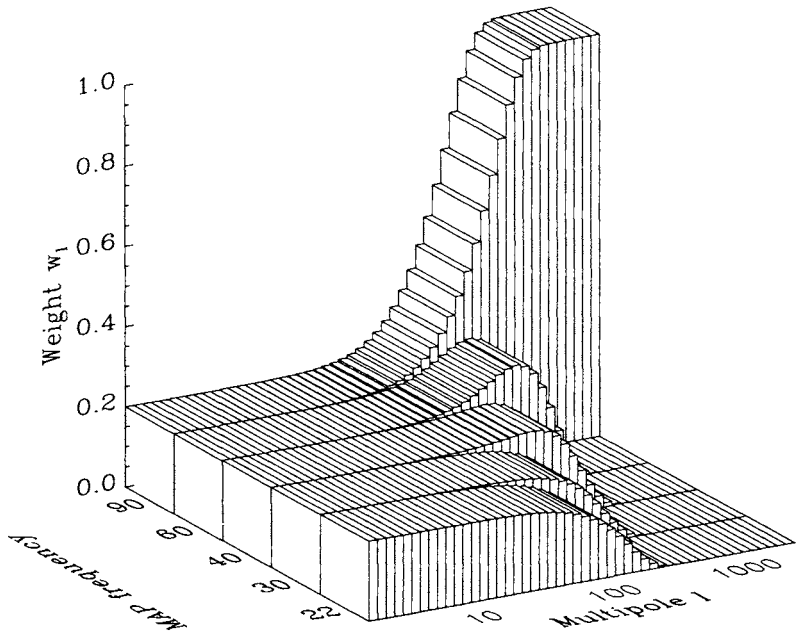

Figure 11.17. The minimum variance linear combination of the frequency channels of MAP in the *absence* of foregrounds (from Tegmark *et al.*, 2000). The noise at low l is identical in all channels, so the minimum variance estimator weights them equally. At high l, the highest frequency channel has the lowest noise, so the best estimator uses only that channel.

11.17 is striking, especially at low l. No longer does one weight all the different channels identically. Rather, a complicated set of weights must be used to project out the foreground contamination.

Note that foregrounds do indeed fit the definition of a systematic effect. If one neglected synchrotron emission in the previous example, the naive estimate of the CMB temperature $(d_1 + d_2)$ would be wrong no matter how small the noise. You might argue that, over the whole sky, the average "wrongness" would cancel out, since there are an equal number of hot and cold spots of foregrounds. The power spectrum, though, the key quantity of interest, *would* be contaminated: it would be the sum of $C_l^{\rm CMB}$ and $C_l^{\rm synch}$, again even if there was no noise.

11.6.2 Mode Subtraction

A common problem in cosmological observations is that a particular mode is contaminated by some external source. An example might be a region of space very close to our galaxy, where foregrounds are particularly important. In a galaxy survey, it might be a dusty region. Dust tends to absorb high-frequency light, and so redden an object's image. This leads to less flux in blue bands, often pushing a galaxy outside the flux limits of a survey. This push can be crucial, because typ-

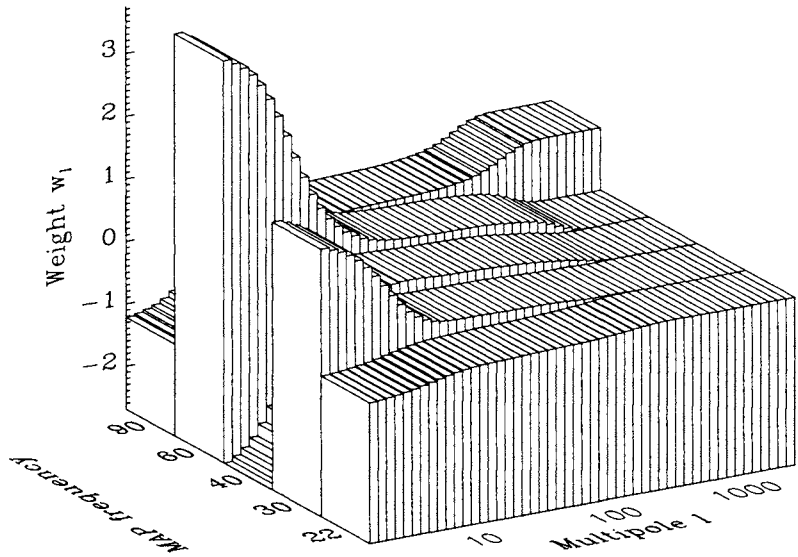

Figure 11.18. The minimum variance linear combination of the frequency channels of MAP in the *presence* of foregrounds (from Tegmark *et al.*, 2000). Compare with Figure 11.17 to see that, especially at low l, foreground contamination dictates a more complicated linear combination of the channels.

ically most of a survey's galaxies lie close to the flux limits. Yet another example is a CMB experiment which has no sensitivity to the average temperature over a given set of pixels because of atmospheric contamination. Another is a galaxy survey in which a given stripe is contaminated because a CCD went bad at the time of observation. And there are many more examples.

One way of dealing with such contamination is to subtract it off. This is commonly done in galaxy surveys by applying the *reddening correction* due to dust from an external dust map (the best one at present is by Schlegel, Finkbeiner, and Davis, 1998).

There is another way of dealing with mode contamination, one which is rapidly growing in importance, as precision cosmology becomes a reality. This technique is based on the twin observations that (i) often there is quite a bit of uncertainty in the amplitude of the contamination and (ii) a given experiment often measures many, many modes. Since the second fact is true, we can often do without the offending mode entirely without losing too much information. Since the first fact is true, we often *should* do without the offending mode, for it may lead us to an incorrect place in parameter space.

How can we eliminate a contaminated mode from an experiment? Let's start with the case where the contaminated mode is a single spatial pixel. In that case, one simple way to ensure that the pixel carries no weight in the likelihood analysis is to add to the covariance matrix a huge number in the diagonal element corre-

sponding to the offending pixel. Then, no matter the value observed in this pixel, the likelihood function will not be affected. This simple idea—adding large noise to a contaminated mode—can be extended to more complicated modes, those not localized to one spatial pixel.

As an example, consider a two-pixel CMB experiment, in which the atmosphere contributes identically to both pixels. Thus, the average temperature of the CMB cannot be determined. The contaminant is assumed to be 100% correlated in the two pixels, so we add to the noise covariance matrix

$$C_{\rm con} = \kappa \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \tag{11.154}$$

where κ is a very large number. Suppose the noise—in the absence of this constraint—is uniform and uncorrelated. Then the new noise covariance matrix is

$$C_N = \sigma_n^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \kappa \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}. \tag{11.155}$$

In the likelihood function, we add the noise covariance matrix to the signal covariance matrix. Again, for simplicity assume that the pixels are far enough away from each other so that the signal is uncorrelated. Then, the likelihood function depends on the full covariance matrix

$$C = \begin{pmatrix} \sigma_n^2 + \sigma_s^2 + \kappa & \kappa \\ \kappa & \sigma_n^2 + \sigma_s^2 + \kappa \end{pmatrix}. \tag{11.156}$$

In particular, the likelihood function depends on the determinant and inverse of this matrix (recall Eq. (11.20)). In the limit that κ is very large, the determinant becomes $2\kappa(\sigma_n^2 + \sigma_s^2)$ and the inverse is

$$C^{-1} \to \frac{1}{2(\sigma_n^2 + \sigma_s^2)} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}. \tag{11.157}$$

Therefore, in this limit, the likelihood function becomes,

$$\mathcal{L} \to \frac{1}{2\pi\sqrt{2\kappa(\sigma_n^2 + \sigma_s^2)}} \exp\left\{ -\frac{(\Delta_1 - \Delta_2)^2}{2(\sigma_n^2 + \sigma_s^2)} \right\}. \tag{11.158}$$

That is, apart from an irrelevant normalization constant, the likelihood function is a Gaussian in $\Delta_1 - \Delta_2$ (the difference between the observed temperatures in the two pixels) with variance equal to $\sigma_n^2 + \sigma_s^2$. Thus adding the constraint matrix of Eq. (11.154) is our way of telling the likelihood function to ignore the average temperature.

In this simple example, we could have written down Eq. (11.158) from the outset, but with more complicated modes, the constraint formalism is extremely powerful. What is the generalization of Eq. (11.154) for an arbitrary contamination? Suppose the contaminated mode is of the form m_i where index i labels the pixels. Thus, in

the average example, m_i would be equal to $(1, 1)$. One adds to the noise covariance matrix the outer product of this vector times a large number κ:

$$(C_{\rm con})_{ij} = \kappa m_i m_j. \tag{11.159}$$

This is precisely what we did above for the average, but this expression is now completely general and allows for elimination of any contaminated mode. These matrices, often called *constraint* matrices, have been used extensively in recent CMB analyses, most notably in the interferometric experiment, DASI.

SUGGESTED READING

I am not a professional statistician and this chapter no doubt glosses over some important concepts in statistics. Nonetheless, I believe this chapter does do justice to the recent flurry of activity in cosmological analysis. Readers interested in more general statistics treatments might consult *The Advanced Theory of Statistics* (Kendall and Stuart). An essential reference is *Numerical Recipes* (Press *et al.*) for all numerical work, and Chapter 15 especially for some of the analysis issues discussed here. A couple of nice early papers on the CMB are Readhead *et al.* (1989) and Bond *et al.* (1991). The former takes the likelihood function further than we did here. For example, it deals with frequentist tests and defines such terms as significance and power, which are extremely important in statistics. The Bond *et al.* paper is a concise introduction to Bayesian analysis of a CMB experiment. Among the ideas explained clearly there are CMB window functions; the likelihood function; dealing with an unknown average; and, to top it off, the idea that CMB experiments will probe the baryon density.

The discussion in Section 11.1.3 is based on a similar treatment in Tegmark *et al.* (1998), which cemented the connection between CMB analyses and galaxy surveys. It also deals with pixelization schemes other than just the two in Section 11.1.3. Window functions for the CMB are discussed in many places. A nice recent treatment is given by Souradeep and Ratra (2001).

Karhunen and Loéve decompositions were introduced to the cosmology community by Bunn (1995) and Bond (1995), who used them to analyze COBE, and by Vogeley and Szalay (1996) in the context of galaxy surveys. The optimal quadratic estimator was introduced by Tegmark (1997) and Bond, Jaffe, and Knox (1998), the latter using the Newton–Raphson motivation I stressed in the text. The former focused on the *minimum variance* aspect, which you can prove in Exercise 10. Earlier, Feldman, Kaiser, and Peacock (1994) had computed an optimal estimator for galaxy surveys which turns out to be the small-scale limit of the optimal quadratic estimator. The Fisher matrix was introduced by Fisher (1935). Knox (1995) computed the Fisher matrix in the all-sky CMB case, Tegmark *et al.* (1998) for galaxy surveys. Jungman *et al.* (1996) used the Fisher matrix (although they didn't call it that) to give the first forecast of parameter determination. There have been many follow-up papers improving and tweaking various parts of the forecast. Some of the improvements are discussed by Eisenstein, Hu, and Tegmark (1999). The curvature matrix and the covariance matrix of errors on parameters are covered in detail in *Numerical Recipes.*

Mapmaking and indeed many of the issues in CMB analysis are reviewed by Bond *et al.* (1999). Foregrounds have been discussed by many authors. The text borrows most heavily from Dodelson (1997) and Tegmark *et al.* (2000). Other works of note include Tegmark and Efstathiou (1996) and Bouchet and Gispert (1999). Mode subtraction along the lines discussed in Section 11.6.2 was introduced by Bond, Jaffe, and Knox (1998). A nice example of its use is in Halverson *et al.* (2002).

EXERCISES

Exercise 1. In the simple example of Section 11.1.1, show that a prior uniform in σ_n^2 gives a final probability distribution in σ_n different from the one in Eq. (11.5).

Exercise 2. In the simple example of Section 11.1.1, we found the error on the signal s. What is the error on the other theoretical parameter of the model σ_n?

Exercise 3. Derive the expression for the covariance matrix due to Poisson sampling, Eq. (11.32).
(a) Divide the survey region into small sub-volumes. Assume that the number of galaxies in a given sub-volume is drawn from a Poisson distribution with mean $\bar{n}$ (assume $\bar{n}$ is constant in all sub-volumes for simplicity),

$$P(n) = \frac{(\bar{n})^n e^{-\bar{n}}}{n!}. \tag{11.160}$$

Determine the expectation values $\langle n \rangle$ and $\langle n^2 \rangle$ for this distribution.
(b) Rewrite Eq. (11.28) as

$$\Delta_i = v \sum_\alpha \psi_i(\vec{x}_\alpha) \left[\frac{n(\vec{x}_\alpha) - \bar{n}}{\bar{n}} \right] \tag{11.161}$$

where α indexes each sub-volume of size v. Using the results of **(a)**, and assuming that there is no intrinsic clustering, determine $\langle \Delta_i \Delta_j \rangle$. Show that it is given by Eq. (11.32). You'll have to change the sums back into integrals.

Exercise 4. Determine the noise covariance matrix in a galaxy survey using counts-in-cells. Assume the cells are spherical with radius R, and find $(C_N)_{ij}$ as a function of the separation between two cell centers; call it r_{ij}. Assume that the survey is volume limited, that is, that $\bar{n}$ is constant everywhere within a volume V.

Exercise 5. Do a full likelihood analysis of the University of California at Santa Barbara's 1990-1 CMB experiment carried out at the South Pole (Gaier *et al.*, 1992).
(a) Determine the window function. The chopping angle was 2.1° and the beam width (FWHM) 1.35°. The anisotropy was measured at nine positions, each separated by 2.1° on the sky. Neglect off-diagonal elements.
(b) Fit a flat band power (i.e., take $\mathcal{C} \equiv l(l+1)C_l/2\pi$ constant) to the following data:

Position	$\delta T(\mu\text{K})$	$\sigma_n(\mu\text{K})$
1	-30.5	25.9
2	-3.2	26.5
3	29.2	26.1
4	-10.8	26.3
5	-8.7	28.8
6	23.1	26.4
7	4.7	26.5
8	-24.7	26.6
9	20.3	25.8

Assume that the noise and the signal are both uncorrelated from pixel to pixel.
(c) The likelihood function you obtain should peak at band power equal to zero. Find the 95% CL upper limit on the band power, defined by the point for which

$$\int_0^{\mathcal{C}_U} d\mathcal{C}\mathcal{L}(\mathcal{C}) = 0.95 \int_0^{\infty} d\mathcal{C}\mathcal{L}(\mathcal{C}). \tag{11.162}$$

This upper limit was reported around the same time as the COBE detection of anisotropies. Compare the two results.

Exercise 6. Find the Fourier transform of the top-hat function $f(x) = \Theta(x + R)\Theta(R - x)$ where Θ is the step function, equal to 1 when its argument is positive and zero otherwise.

Exercise 7. Find the diagonal elements of the covariance matrix C_S for a volume limited survey for modes $k_i \gg R^{-1}$. Show that they are given by Eq. (11.60).

Exercise 8. Compute the off-diagonal window function for the two types of galaxy surveys mentioned in Section 11.2.3, a volume-limited survey and a pencil-beam survey. Take both $\vec{k}_i$ and $\vec{k}_j$ parallel to the z-axis (which in the pencil beam survey is aligned with the long distance L). Plot the window function at $k = k_i$ as a function of k_j. At the point $k_i = k_j$, you should recapture the corresponding points in Figures 11.5 and 11.7.

Exercise 9. Prove Eq. (11.100). Assume there is only one parameter λ and that the likelihood function is Gaussian in the overdensities Δ. Further assume that you have iterated enough times so that the input parameter $\lambda^{(0)}$ is equal to the true value $\bar{\lambda}$.

Exercise 10. Derive the optimal quadratic estimator of Eq. (11.92) by minimizing the variance of a general quadratic estimator subject to the constraint that its expectation value is unbiased. Hint: Use a Lagrange multiplier to enforce the constraint.

Exercise 11. Consider an all-sky CMB experiment with spatial pixels of area $\Delta\Omega$. Assume that the experiment measures the temperature in each pixel with Gaussian noise σ_n. The noise is thus assumed to be uniform (the same everywhere on the

sky) and uncorrelated (from one pixel to the next). Determine the noise covariance matrix for a_{lm}. If the pixel size is cut in half (for the same experiment), each pixel will get less observing time by a factor of 2. The noise will then go up for each pixel by a factor of $\sqrt{2}$. Show that these two changes (smaller pixels; more noise per pixel) leave the noise covariance matrix for a_{lm} unchanged.

Exercise 12. Estimate the expected errors on C_l for the following experiments: (i) COBE, (ii) Boomerang, and (iii) MAP, and (iv) Planck.

Exercise 13. The full-sky limits for the Fisher matrix derived in Section 11.4 can be used to find the optimal quadratic estimator. The results should not surprise you.
(a) Given a set of a_{lm} from a full-sky CMB experiment with uniform weight w, find the optimal quadratic estimator for C_l.
(b) Given a set of pixelized Fourier overdensities from a 3D galaxy survey, find the optimal quadratic estimator for $P(k)$.
(c) From your answers to (a) and (b), discuss qualitatively the effectiveness of the optimal quadratic estimator. When do you expect it to perform differently from the naive quadratic estimator, $\sum_m |a_{lm}|^2/(2l+1)$ for C_l and $\sum_{\vec{k}} |\delta_{\vec{k}}|^2/(4\pi k^2 \Delta k)$ for $P(k)$ (where Δk is the width of the k-bin)?

Exercise 14. Show that the noise covariance matrix of a map using the estimator in Eq. (11.134) is C_N, as given in Eq. (11.135)

Exercise 15. Suppose one had prior information about a foreground, in the form of an assumed power spectrum, C_l^α, where α labels the foreground component.
(a) Find the miminum variance estimator for the temperature and the associated covariance matrix.
(b) Consider the example of Section 11.6.1 with two frequencies in the Rayleigh–Jeans regime and one foreground with shape vector $\vec{W}^1 = (1, 1/2)$. What is the best estimator of the CMB temperature if the foreground has assumed mean equal to zero and variance equal to that of the noise (σ_n^2)? What is the new error on the determination of the CMB temperature? Compare both of these to the case treated in the text when no information about the foreground amplitude was assumed.

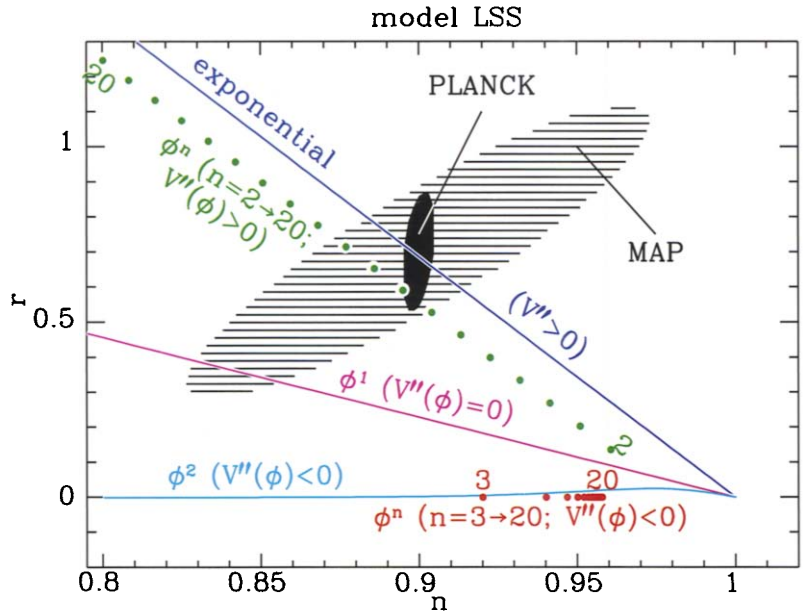

Plate 11.14. Expected 95% uncertainty on the inflationary parameters n and r from MAP and Planck (from Dodelson, Kinney, and Kolb, 1997). Three other parameters (normalization, Ω_B, and h) have been marginalized over. Every inflationary model gives a unique prediction somewhere in this plane; many such predictions are plotted.

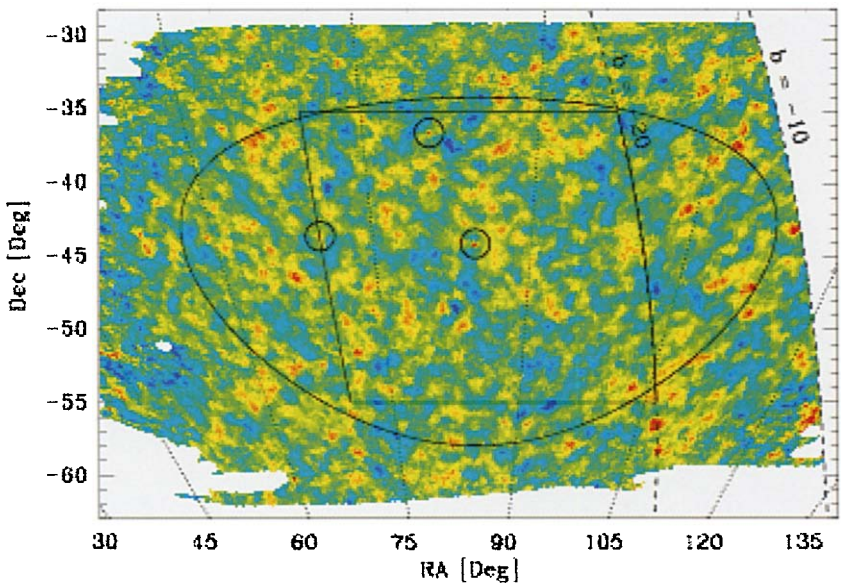

Plate 11.15. A map of the CMB temperature from observations by Boomerang (Netterfield *et al.*, 2001), a long-duration balloon flight at the South Pole. Hot and cold spots have amplitudes as large as 500μK. Circles shows quasars identified in these radio observations. The large elliptical region delineates data analyzed to obtain bandpowers. The rectangular region is an earlier data set.