Power spectrum estimators

Most of the lecture was based on astro-ph/9708020, by Tegmark, Hamilton, Vogeley and Szalay; this note is a direct summary of that paper. (Please tell Joanne jcohn@astron.berkeley.edu of any mistakes, typos or more important, thanks!).

1. Introduction

We've been hearing all week about how to calculate theoretical predictions for theories. The big underlying question between theories and measurements is "how well do theory and data agree?" The simplest thing to check in both is the power spectrum. So then the questions become

1. What is the power spectrum predicted by the theory?
2. What is the power spectrum of the data?
3. How well do they agree (i.e. what are the errors)?
4. what is the best theory (i.e. parameters) for the data, and how well does it fit (errors again)?

The field has developed significantly in the last 5-10 years, with some of the "traditional" methods dating back only 10 years! The data (and theoretical predictions) are getting good enough that precision is becoming crucial, in addition the data sets are getting large enough that the speed of a method is an issue. The main new aspect in the "new" methods is that they allow one to deal with systematics more effectively, and often hybrids of many methods are used at once.

2. What is the power spectrum from the data?

This sounds like a very simple question, but it serves to illustrate the problem. The theorist can define the power spectrum immediately: taking the continuous density field $\rho(\mathbf{x})$, we have that fluctuations obey

$$\delta(\mathbf{x}) = \frac{\rho(\mathbf{x}) - \bar{\rho}}{\bar{\rho}} \tag{1}$$

with a correlation function

$$\langle \delta(\mathbf{x})\delta^*(\mathbf{x}+\mathbf{r})\rangle = \xi(\mathbf{r}) = \frac{V}{(2\pi)^3}\int d^3k P(k) e^{-i\mathbf{k}\cdot\mathbf{r}} \tag{2}$$

and the power spectrum $P(k)$ obeys

$$P(k) = \langle |\delta_k|^2 \rangle \tag{3}$$

where the brackets mean ensemble average (which is done in practice as a volume average), and δ_k is the Fourier transform of $\delta(\mathbf{x})$. Note that I will not be too careful about Fourier transform conventions in these notes, the paper does do them in detail in an appendix. The correlation function can also be interpreted as excess probability over random

$$dP = \bar{\rho}^2[1 + \xi(\mathbf{r})]dV_1 dV_2 \tag{4}$$

of finding an object in volume 1 and another in volume 2 simultaneously.

The observer on the other hand has N galaxies, and thus a density field given by

$$n(\mathbf{r}) = \sum_\alpha \delta^D(\mathbf{r} - \mathbf{r}_\alpha) \tag{5}$$

with galaxies at positions $\mathbf{r}_\alpha$. This isn't the same as above for several reasons.

- Instead of a continuous $\delta(\mathbf{r})$ we have a discrete sampling given by $n(\mathbf{r})$.

  Treating this sampling as a Poisson process, we can say that

  $$\text{Prob}(n(\mathbf{r}) = n) = e^{-\mu}\frac{mu^n}{n!} \tag{6}$$

  where

  $$\mu(\mathbf{r}) = \bar{n}(\mathbf{r})(1 + \delta(\mathbf{r})) \tag{7}$$

  and the fluctuations ensemble average, $\langle \delta_{\mathbf{k}} \delta_{\mathbf{k}'} \rangle = (2\pi)^3 \delta^D(\mathbf{k} - \mathbf{k}')P(\mathbf{k})$. As we saw the day before, this leads to shot noise in the measurements.

- We don't know $\bar{\rho}$. The analogue we have is $\bar{n}(\mathbf{r})$. This is the expected number of galaxies, in the particular survey, in a volume dV, in the absence of clustering in the survey. This includes properties of the survey (i.e. the selection function, thus it depends upon position) but also includes an estimate of the average number of galaxies you'd see, which you don't know ahead of time. You need to figure this quantity out somehow or make sure things don't depend too strongly on it (so that errors don't mess up the estimates too much). More later.

- Not only is the density field sampled discretely, it is also not sampled in all directions.

  There is only a finite volume of the density sampled, and the survey can be volume limited, flux limited or something else. In addition, there might be biases, distortions, etc., and other systematics, which will affect how well ρ is sampled.

However, from the theorist, we know we want some quantity that is quadratic in the density, so that we want something that Tegmark et al write abstractly as

$$q_i = \int \int d^3r \, d^3r' E_i(\mathbf{r}, \mathbf{r}') \frac{n(\mathbf{r})}{\bar{n}(\mathbf{r})} \frac{n(\mathbf{r}')}{n\bar{(}\mathbf{r}')} \tag{8}$$

Most of the different methods are just ways of picking out $E_i(\mathbf{r}, \mathbf{r}')$ (aside from the "brute force method"). Essentially you want to give the power spectrum in a bunch of bands, with as much info as possible from the data, but without any biases/systematics that you can get rid of. In addition you want the errors to be calculable and uncorrelated, and you want the calculation to be doable (on a computer or otherwise).

"Keeping as much info as possible from the data" can be made more precise. Fisher matrix techniques (outlined in the paper) can be used to quantify the "information" in a data set with respect to a parameter or parameters of interest, by saying how big the error bars are. If you "clean" the data, and remove noise, or compress it for some other reason, you can check how much information is lost by computing the Fisher matrices before and afterwards. It should be noted that Fisher matrices rely upon you knowing the probability for the data given a particular theory, often these have to be assumed to be gaussian (not always reasonable), or calculated for example using mock catalogues.

Looking at the expression for the estimators, q_i again, Tegmark and co. suggested a way of describing many of them in one language, writing

$$E_i(\mathbf{r},\mathbf{r}') = \psi_i(\mathbf{r})\psi_i^*(\mathbf{r}') \tag{9}$$

where $\psi_i(\mathbf{r})$ differs from estimator to estimator. In more detail,

$$q_i = |x_i^2| \tag{10}$$

where

$$x_i = \int d^3r \left[\frac{n(\mathbf{r})}{\bar{n}(\mathbf{r})} - 1\right] \psi_i(\mathbf{r}) \tag{11}$$

Ideally the ψ_i is chosen to obey

$$\int d^3r \psi_i(\mathbf{r}) = 0 \tag{12}$$

so that in practice the "-1" doesn't matter in the definition of x_i. This is called the integral constraint and will be discussed more later. One thing to note is that this constraint is not automatically met by all the examples below.

Given a $\psi_i(\mathbf{r})$ you are basically giving an estimator. Some examples are

1. "counts in cells"

   Take

   $$\psi_i(\mathbf{r}) = \begin{cases} \bar{n}(\mathbf{r}) & \text{in region } i \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

   so that

   $$x_i = \int_{V_i} d^3r[n(\mathbf{r}) - \bar{n}(\mathbf{r})] \tag{14}$$

   i.e. the excess density in region i, which is basically just the counts in cells.

2. “Fourier transforms”

   Take

   $$\psi_i(\mathbf{r}) \propto \phi(\mathbf{r}) e^{i\mathbf{k}_i\cdot\mathbf{r}} \tag{15}$$

   that is, weighted Fourier coefficients of some sort. Some choices are

   $$\begin{array}{ll} \phi(\mathbf{r}) = \left\{ \begin{array}{ll} 1 & \text{in survey} \\ 0 & \text{otherwise} \end{array} \right. & \text{Fourier transform in survey} \\ \bar{n}(\mathbf{r}) & APM \\ \frac{\bar{n}(\mathbf{r})}{1+\bar{n}(\mathbf{r})P} & FKP \\ efn\ of\ \nabla^2 - \frac{\gamma}{\bar{n}(\mathbf{r})} & narrow\ in\ k\ space \end{array} \tag{16}$$

   which have all been used in the past. The top one, a straight Fourier transform, is the simplest, the second was used in APM, the third is what is called the FKP weighting (after the Feldman, Kaiser and Peacock ’93 paper which described it) and the fourth has the advantage of giving narrow windows in k space. Note that they don’t necessarily obey the integral constraint.

3. “Spherical Harmonics”

   Take

   $$\psi_i(\mathbf{r}) \propto Y_{\ell m} j_\ell(k_i r) \; . \tag{17}$$

4. Take $\psi_i(\mathbf{r})$ to be the eigenmode of a useful operator.

In the above, the top case is the correlation function, which is discussed in another set of notes. Versions of the second are called “the traditional method” by Tegmark et al, and an example of the fourth method is the Karhunen-Loeve transform (KL transform).

The paper discusses in detail the “traditional method”, the KL transform, the brute force likelihood method and the quadratic method. In comparison to the above cases, the quadratic method can be written as

$$q_i = “\mathbf{x} C_{,i}^{-1} \mathbf{x}'' \tag{18}$$

with $C_{,i}$ to be defined later on, instead of $q_i = |x_i|^2$.

As many of the cases do involve the quadratic $q_i = |x_i|^2$, let’s consider what this quantity is. Observationally, you just plug in your data points, with your assumption for $\bar{n}(\mathbf{r})$, and get out some estimator q_i. Theoretically, you can also calculate what this should be. Consider more generally $\langle x_i x_j^* \rangle$. Theory gives a definite prediction for this if you have chosen your ψ_i. That is,

$$\begin{array}{l} \langle \mathbf{x} \rangle = 0 \\ \langle x_i x_j * \rangle = \mathbf{C}_{ij} = \mathbf{N}_{ij} + \mathbf{S}_{ij} \end{array} \tag{19}$$

where the matrices $\mathbf{N}, \mathbf{S}$ are called the noise and signal matrices respectively, and their combination is usually called $\mathbf{C}$. Theoretically you can calculate these matrices given the theory in the following way:

$$\begin{aligned}
\langle x_i x_j * \rangle &= \langle \textstyle\int d^3r \left[\frac{n(\mathbf{r})}{\bar{n}(\mathbf{r})} - 1\right] \psi_i(\mathbf{r}) \int d^3r' \left[\frac{n(\mathbf{r}')}{\bar{n}(\mathbf{r}')} - 1\right] \psi_j^*(\mathbf{r}')\rangle \\
&= \textstyle\int d^3r\, d^3r' \langle \frac{n(\mathbf{r})}{\bar{n}(\mathbf{r})} \frac{n(\mathbf{r}')}{\bar{n}(\mathbf{r}')} \rangle \psi_i(\mathbf{r}) \psi_j^*(\mathbf{r}') \\
&= \textstyle\int d^3r \frac{\psi_i(\mathbf{r})\psi_j^*(\mathbf{r})}{\bar{n}(\mathbf{r})} \langle 1 \rangle + \int d^3r\, d^3r' \langle (1+\delta(\mathbf{r}))(1+\delta(\mathbf{r}'\rangle)) \psi_i(\mathbf{r}) \psi_j^*(\mathbf{r}') \\
&= \textstyle\int d^3r \frac{\psi_i(\mathbf{r})\psi_j^*(\mathbf{r})}{\bar{n}(\mathbf{r})} + \int \frac{d^3k}{(2\pi)^3} \hat{\psi}_i(\mathbf{k}) \hat{\psi}_j^*(\mathbf{k}) P(k) \\
&= \mathbf{N}_{ij} + \mathbf{S}_{ij}
\end{aligned} \tag{20}$$

To get from the first to the second line, we used the integral constraint, $\int d^3r \psi_i(\mathbf{r}) = 0$. Going from the second to the third line used that the Poisson average of the discrete number density obeys

$$\langle n(\mathbf{r}) n(\mathbf{r}')\rangle = \mu(\mathbf{r})\mu(\mathbf{r}') + \delta^D(\mathbf{r} - \mathbf{r}')\mu(\mathbf{r}) \tag{21}$$

where

$$\mu(\mathbf{r}) = \bar{n}(\mathbf{r})(1 + \delta(\mathbf{r})) \tag{22}$$

and δ^D refers to the Dirac delta function. The ensemble average still needs to be done on the third line (hence the brackets in the expression), which are calculated to go to the fourth line. Again, if you have a theory, you can calculate these matrices for a given power spectrum and $\bar{n}(\mathbf{r})$.

To see how this corresponds to the power spectrum more directly, consider the last line, with the definitions of the shot noise and signal, $\mathbf{N}_{ij}$ and $\mathbf{S}_{ij}$ respectively. Say that the $\hat{\psi}_i(k)$ are very narrow in k space, centered at k_i. Then, for a narrow enough window, the power spectrum will vary very slowly, allowing it to be taken out of the integral

$$\mathbf{S}_{ij} = \int \frac{d^3k}{(2\pi)^3} \hat{\psi}_i(\mathbf{k}) \hat{\psi}_j^*(\mathbf{k}) P(k) \sim \delta_{ij} P(k_i) \int \frac{d^3k}{(2\pi)^3} \hat{\psi}_i(\mathbf{k}) \hat{\psi}_i^*(\mathbf{k}) \tag{23}$$

i.e. the signal will be the power spectrum weighted by $\int \frac{d^3k}{(2\pi)^3} \hat{\psi}_i(\mathbf{k}) \hat{\psi}_i^*(\mathbf{k})$. For this reason people define the window function to be

$$W_i(k) = \hat{\psi}_i(\mathbf{k}) \hat{\psi}_i^*(\mathbf{k}) \tag{24}$$

and require

$$\int \frac{d^3k}{(2\pi)^3} W_i(k) = 1 \; . \tag{25}$$

The off-diagonal terms in $\mathbf{C}_{ij}$ correspond to how much interdependence there is between the chosen modes (i.e. covariance).

The authors then covered 4 methods: two "old" (before 1997 that is?!) ones, the traditional and brute force methods, and two "new" methods, KL (and generalizations), and the quadratic method. We'll do each of these in turn.

3. "Traditional Method"

The traditional method takes

$$\psi_i(\mathbf{r}) = \phi(\mathbf{r})e^{i\mathbf{k}_i\cdot\mathbf{r}} \tag{26}$$

i.e. a weighted Fourier transform of the data, with the power in the ith mode being $q_i = x_i x_i^*$. Calculating the expectation value of q_i theoretically, we get

$$\begin{aligned}\langle q_i\rangle &= \langle x_i x_i^*\rangle \\ &= \hat{\psi}_i(0)^2 + \int d^3r\frac{\psi_i(\mathbf{r})\psi_i^*(\mathbf{r})}{\bar{n}(\mathbf{r})} + \int \frac{d^3k}{(2\pi)^3}|\hat{\psi}_i(\mathbf{k})|^2 P(k)\end{aligned} \tag{27}$$

just as we saw earlier, except for the first term. The first term comes from the fact that

$$\int \psi_i(\mathbf{r})d^3r = \int \hat{\psi}_i(k)e^{i\mathbf{k}\cdot\mathbf{r}}d^3r\frac{d^3k}{(2\pi)^3} = \hat{\psi}_i(0) \neq 0 \tag{28}$$

that is, generally one won't satisfy the integral constraint. In this form it is a bit easier to see the problems that arise. First of all, if one misguesses $\bar{n}(\mathbf{r})$ and tries to subtract it out, say getting it off by a constant a, then one will think one has gotten rid of this offset but in fact will be left with an offset $\propto (1-a)\hat{\psi}_i(0)^2$. The second problem is that no matter what the estimate from the data is of the power spectrum, by definition the data will have fluctuations going to zero on the scale of the survey, that is for large k. For the largest scale of the survey, that of the survey itself, there will only be one sample and thus the fluctuations will be zero (fluctuations from the average) on that scale. That is, the data will tend to imply that $P(k) \to 0$ as $k \to 0$ no matter what the true $P(k)$ does. In order to get rid of this problem, taking $\hat{\psi}_i(0) = 0$ will mean that this part of the power spectrum will not enter into the estimates of the power or the errors, as we cannot obtain it observationally.

One can mock up a subtracted power spectrum estimator for these,

$$\tilde{q}_i = \left[\frac{|x_i|^2 - b_i}{A_i}\right] \tag{29}$$

where the explicit expressions for b_i, A_i are in the paper and are derived from $\phi(r), n(\mathbf{r})$ in the theory. In practice this subtraction can be very time consuming and produces the main downside of this method.

A popular choice for the Fourier coefficients is given by

$$\phi(\mathbf{r}) = \frac{\bar{n}(\mathbf{r})}{1+\bar{n}(\mathbf{r})P} \tag{30}$$

called the FKP method. Here P is the averaged power, and this weighting is derived in the paper by Tegmark et al. Its benefit can be found by considering $\langle x_i x_j^*\rangle$. Take the modes and divide them up into different k_i values, $i = 1, N$ the number of galaxies. If we consider L, the smallest dimension of the survey (i.e. the length of the shortest side) then if we look at momenta k and momentum bins k_i satisfying

$$k \gg |k_i - k_j| \gg L^{-1} \tag{31}$$

then we find that

$$\langle x_i x_j^* \rangle \sim 0 \tag{32}$$

that is, the overlap between power spectrum bins is close to zero. Note that large k means small scales. Also,

$$\langle (x_i x_j^*)^2 \rangle \tag{33}$$

is minimized. If the probability distribution for the x_i is gaussian, this second quantity is directly proportional to the errors in the estimate, and thus minimizing it is a good thing. (In practice, if shot noise doesn't dominate and gaussianity holds, you can show that the error in the power goes as one over the square root of uncorrelated volumes in the sample). So this estimate is good at short (compared to the size of the sample) distances, for k_i bins that are separated "enough." In practice, you have what the weights are and just plug them into your data to get your estimate of power in these bands. The main drawback is the integral constraint.

4. Brute Force Method

The Tegmark et al paper mentions this method next, although it is not a power spectrum estimator at all–it does the whole thing, comparing data to theory, in one step. It does what you'd like, gives the most likely parameters, however it does it within the context of strong assumptions, and in addition can be painfully slow to implement. That is, it gives you the "most likely" parameters in the case where the probability distribution for $\mathbf{x}$ is a multivariate gaussian. As $\mathbf{x}$ is the density field with some weighting, one might expect this to be true at very large (linear) scales.

The starting point is to say that the probability is gaussian, so that you can write that the probablity for the data to have value $\mathbf{x}$ given the theory (with parameter vector Θ) is

$$f(\mathbf{x}|\Theta) = \frac{1}{\sqrt{C}} e^{-\frac{1}{2}\mathbf{x}^\dagger C^{-1}\mathbf{x}} \tag{34}$$

where $C = C(\Theta) = N + S$ from before, and $\sqrt{C}$ refers to the square root of the determinant of C. What we want instead is the most probable theory (Θ) given the data $\mathbf{x}$. We can use Bayes theorem for this: that the probability of the data given the theory times the prior for the theory is the probability of the theory given the data times the prior for the data, or separating out the probability of the theory given the data,

$$\mathcal{L}(\Theta|\mathbf{x}) = \frac{f(\mathbf{x}|\Theta) f(\Theta)}{f(\mathbf{x})} \tag{35}$$

If we assume very broad priors for $\Theta, \mathbf{x}$ in addition to the probability for the data given the theory being gaussian, then we have the probability (or likelihood) for any theory given the data, where we just plug in the numbers for the data in the above expression. A broad prior for $\Theta, \mathbf{x}$ means we just treat them as constant.

If the above is the probability for any theory parameters given the data, then the most likely theory is the one with the most probability, i.e. the one that maximizes the likelihood $\mathcal{L}$. So one maximizes this likelihood numerically and ends up with the best set of parameters Θ. If the probabilities are not gaussian then this is not very appropriate and doesn't give something useful–unlike the other methods which give a power spectrum estimator but have extra nice other properties if the underlying theory is gaussian. Here, you really want the theory to be gaussian or else it is just plain wrong. So this is good for data on very large (linear) scales. The other point is that it is very slow to implement, so it is good to compress the data ahead of time.

5. KL transforms

The Karhunen-Loeve method is a data analysis/reduction technique that many people swear by. Recall that we had

$$\langle x_i x_j^* \rangle = \mathbf{C}_{ij} = \mathbf{N}_{ij} + \mathbf{S}_{ij} \tag{36}$$

Consider instead another basis

$$\mathbf{y} = \tilde{\mathbf{B}}\mathbf{x} \tag{37}$$

where

$$\begin{array}{c} \tilde{\mathbf{B}}\mathbf{N}\tilde{\mathbf{B}}^T = \mathbf{1} \\ \tilde{\mathbf{B}}\mathbf{S}\tilde{\mathbf{B}}^T = \mathbf{\Lambda} \end{array} \tag{38}$$

with $\mathbf{1}$ the identity matrix and $\mathbf{\Lambda}$ a diagonal matrix with eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_N\}$. The new variables $\{y_i\}$ obey

$$\langle y_i y_j^* \rangle = \delta_{ij}(1 + \lambda_j) \tag{39}$$

If the x_i were defined using weights ψ_i then it follows that the y_i will have weights $\psi_i' = \sum_j \tilde{\mathbf{B}}_{ij}\psi_j$.

The advantages of finding this basis are the following:

1. The $\{y_i\}$ are orthogonal (statistically independent) variables if the $\{x_i\}$ are gaussian, and independent gaussian random variables to boot.

2. The λ_i have a sort of interpretation of being signal to noise,

$$\lambda_i = y_i^2 - 1 \sim \text{“}S/N''\tag{40}$$

so that the signal can be considered as (signal)(“signal/noise”) or $q_i \sim \frac{y_i^2 - 1}{\bar{n}}$, which is very similar in form to the revised (subtracted) estimator for FKP, equation 29.

3. Large λ_i means that a mode contributes a lot of signal to noise, so that the modes are sorted according to their information (i.e. signal to noise). Usually many modes will have $\lambda_i \ll 1$ and can be thrown out without much loss in constraining power. This is the basic idea behind data compression for KL and a way that it is often used in conjunction with other methods.

4. We have modes sorted according to the largest signal $\mathbf{S}$. However, we could sort the modes according to the largest dependence on something else, for example, $\partial\mathbf{C}/\partial\Theta_i = \mathbf{C}_{,i}$ where Θ_i is a parameter in the theory. Often when this is done, one chooses the basis vectors y_i to be orthonormal,

$$\langle y_i y_j^* \rangle = \delta_{ij} \tag{41}$$

by diagonalizing the covariance matrix $\mathbf{C}$ with some matrix $\mathbf{B}$ which obeys

$$\begin{aligned} \mathbf{BCB}^T &= \mathbf{1} \\ \mathbf{BC}_{,i}\mathbf{B}^T &= \mathbf{\Lambda} \; . \end{aligned} \tag{42}$$

In this case the different modes are sorted by their dependence on $\mathbf{C}_{,i}$. In practice one could consider modes which have the most “signal” with respect to many different parameters and then take the set of modes which includes as many of these as possible (e.g. if you want to get at several parameters at once).

Another version of this is to choose a power spectrum that has features one wants (ie is monotonic in k for regions of interest) and to sort according to this “power spectrum”, which results in windows in k space, ie the power spectrum for different values of k.

5. The KL method is “asymptotically pixelization independent.” One is diagonalizing a large matrix to implement the method, and so often one doesn't take as many pixels as there are galaxies, but a smaller number of pixels, corresponding e.g. to larger regions in the sky. With this smaller number of pixels, one can pick out the modes with the most “signal” (or parameter dependence). If one then refines the pixelization, taking smaller regions in the sky, the formerly best modes remain strong “signal” modes, and one just adds more modes with signal dependence.

6. Troubleshooting. If you take a power spectrum from your theory, calculate the covariance matrix $\mathbf{C}$, and use it to calculate your y_i variables, and then calculate from the data what $\langle y_i y_j^* \rangle$ is, you should find that it is roughly diagonal with variance one. If you get anything strongly different (how likely this is can be estimated from gaussian statistics if they are applicable, ie if one is on large scales), then either you have something which isn't gaussian, or you've put in the wrong power spectrum.

7. Linear Filtering. There's been reference above to “throwing out” modes which one doesn't want. To keep the large λ modes (“signal”) one defines a new set of modes

$$x_i' = \sum_{j=1}^{N} (\mathbf{CB}^\dagger)_{ij} w_j y_j \tag{43}$$

where the weights w_j tell you which modes you are keeping. Some obvious choices are

$$\begin{array}{ll} w_j = 1 & x'_i = x_i \text{ same as original} \\ w_j = 1 \; iff \; \lambda_i > \lambda_c & \text{“optimal subspace''} \\ w_j = \frac{\lambda_j}{\lambda_j + 1} & \text{“Wiener filtering''} \end{array} \tag{44}$$

Wiener fitering interpolates between large and small values of λ more smoothly than just throwing out some and keeping others (for the “optimal subspace”.) Even when one uses another method, KL compression can help make the problem more tractable.

8. Last but not least, it appears that if you did choose the wrong power spectrum in your initial guess (and used it to calculate $\mathbf{C}$ and thus the variables y_i), that you will not bias your power spectrum estimate. (You will get variables that don't have variance 1, though.)

6. Quadratic Estimator

The fourth method they cover is the quadratic estimator method. This is a minimum variance power estimator. One breaks up momentum space into several bins, N' of them (where $N' < N$, the number of galaxies). That is, one chooses k_i where

$$0 < k_1 < k_2 < k_3 < \cdots < k_{N'} < k_{N'+1} = \infty \tag{45}$$

and takes the power to be constant in each “band”, that is there are N' constants p_i, the power corresponding to the band with $k_i - k_{i-1}$. Again, from one's guess for the theory one has a matrix $\mathbf{C} = \mathbf{N} + \mathbf{S}$. Consider

$$\frac{\partial \mathbf{C}}{\partial p_i} = \mathbf{C}_{,i} = \int_{k_i > |k| > k_{i-1}} \hat{\psi}_a(k) \hat{\psi}_b^*(k) \frac{d^3k}{(2\pi)^3} \tag{46}$$

as one can see by substitution. (I'm not sure whether the $\hat{\psi}$ have to obey anything in particular, and I'm not sure how N' is chosen however.)

One can use this matrix to define $i = 1, N'$ different power estimators in different bands:

$$q_i = \frac{1}{2} \mathbf{x}^\dagger \mathbf{C}^{-1\dagger} \mathbf{C}_{,i} \mathbf{C}^{-1} \mathbf{x} \tag{47}$$

which is compressing the data (only N' estimators, not N of them) and although quadratic in the x_i has the central weight that makes it slightly different from the other examples.

If the probability distribution for x_i is gaussian, then the mean and covariance of q_i, are given in terms of the Fisher matrix. So $\langle \mathbf{q} \rangle = \mathbf{F}\mathbf{p}$, that is $\mathbf{q}$ is proportional to the power. And the covariance, which gives the error for a gaussian theory, $\langle \mathbf{q}\mathbf{q}^{\mathbf{t}} \rangle - \langle \mathbf{q} \rangle \langle \mathbf{q}^{\mathbf{t}} \rangle$ is the Fisher matrix, and thus, as the Fisher matrix gives the smallest errors one can get, this estimator has the smallest errors one can get. In practice, one rescales $\mathbf{q}$ by the square root of the Fisher matrix to get something for $\mathbf{p}$ that is less correlated and less noisy.

7. Relations between the methods

These power spectrum estimators are all linked together. For instance, the quadratic and KL methods are linked by the identity

$$2q_i = \sum \lambda'_j y_j^2 \tag{48}$$

where the q_i is the quadratic estimator, the y_j are the KL vectors, defined with a $\mathbf{B}$ obeying $\mathbf{B}\mathbf{C}_{,\mathbf{i}}\mathbf{B}^T = \Lambda'$, with Λ' a diagonal matrix with eigenvalues λ'_j. These give the same estimates after one subtracts off shot noise and properly normalizes.

The FKP and quadratic estimators are the same at small scales (they show this in detail), which makes sense, they are both minimum variance weights when one makes a gaussianity assumption.

The quadratic and brute force methods are related because one can solve for the power in the quadratic method, and then reiterate, using for input values of p_i the results from the previous iteration. Eventually one gets the maximum likelihood solution (they point out that this is equivalent to using the Newton Raphson method). One caution is that if you have a noisy power spectrum, then the error bars you estimate this way will be wrong, and they suggest that instead of interating that you parameterized $P(k)$ with a fitting function and add parameters until χ^2 per dof drops below one. More on this in the paper.

8. Systematic errors

One place that the "new" methods (especially KL) excel is in the treatment of systematic errors. This can be illustrated by our ignorance of $\bar{n}(r)$. Recall that in the FKP method, the estimators q_i had an extra piece,

$$\langle q_i \rangle = \hat{\psi}_i(0)^2 + \int d^3r \frac{\psi_i(\mathbf{r})\psi_i^*(\mathbf{r})}{\bar{n}(\mathbf{r})} + \int \frac{d^3k}{(2\pi)^3} |\hat{\psi}_i(\mathbf{k})|^2 P(k) \tag{49}$$

and we argued that we want $\hat{\psi}(0) = 0$ for two reasons, so that a misestimate of the integral of $\bar{n}(r)$ wouldn't result in an additive error to all the estimates, and so that $P(k)$ wouldn't be forced to go to zero at large scales (small k) even though the data will have no fluctuations on the largest scale.

Rephrase this as: you have a systematic effect that you don't know (in this case $\bar{n}(\mathbf{r})$) and so you take out the mode that depends upon it (in this case $\hat{\psi}(0)$). Stated this way, the generalization becomes: if you have a systematic in some mode or modes, make the results independent of that mode.

Examples they give include incorrectly modeled extinction, which would result in power in a purely angular mode. So you can take out angular modes or more pointedly modes corresponding to dust templates. Another is the misestimate of radial selection function. Here you can take out the purely radial modes. The KL method is the only one which keeps phase information, i.e.

which can tell you if a vector **r** is in the radial or angular direction (and same in Fourier space), so many of these things have to be done in the context of the KL method. Generally you use some projection operator $\mathbf{\Pi}$ on your vector **x**, many different ones are discussed in the text and why certain features are convenient.

There is also an appendix on redshift distortions, and Tegmark and Hamilton (in another paper) show how to get a real space power spectrum from a redshift space power spectrum.

9. Comparisons

The paper ends with a chart comparing the methods. Briefly, on small scales, the traditional is best, on large KL, quadratic and brute force are all good. Part of what makes a method "fine" is that the errors are uncorrelated. For redshift space distortions and getting rid of systematics, the KL method is the best. For the largest scales, you want to use the quadratic method, as it is the fastest (there is a way to make it go linearly in N, the number of galaxies). So in practice you probably use a hybrid, FKP on the smallest scales, KL to compress the data, and quadratic on the largest scales. Something like this was done in SDSS.