THE ASTROPHYSICAL JOURNAL, 503:492–501, 1998 August 20
© 1998. The American Astronomical Society. All rights reserved. Printed in U.S.A.

COSMOGRAPHY AND POWER SPECTRUM ESTIMATION: A UNIFIED APPROACH

UROŠ SELJAK
Harvard-Smithsonian Center for Astrophysics, Cambridge, MA 02138
Received 1997 October 24; accepted 1998 March 30

ABSTRACT

We present a unified approach to the problems of reconstruction of large-scale structure distribution in the universe and determination of the underlying power spectrum. These have often been treated as two separate problems, and different analysis techniques have been developed for both. We show that there exists a simple relation between the optimal solutions to the two problems, allowing us to solve for both within the same formalism. This allows one to apply computational techniques developed for one method to the other, which often leads to a significant reduction in the computational time. It also provides a self-consistent treatment of linear reconstruction by first computing the power spectrum from the data themselves.

Subject headings: cosmology: theory — large-scale structure of universe — methods: analytical

1. INTRODUCTION

The issue of optimal power spectrum reconstruction from a set of noisy and sparsely sampled data has recently received a lot of attention in the field of large-scale structure (LSS) and cosmic microwave background (CMB) anisotropies (Bond, Knox, & Jaffe 1997; Górski 1994; Hamilton 1997a; Tegmark 1997; Tegmark et al. 1997a; Vogeley & Szalay 1996). Most of this work has been inspired by the existing and forthcoming large data sets (*COBE*, MAP, Planck, SDSS, 2DF, etc.). The quality and amount of information in these surveys is or will be so large that it will allow one to make high-precision tests of cosmology. The precision to which one can test cosmological models is limited to the accuracy with which one can extract useful information from the data, and so it is clearly important to develop statistical techniques that will allow one to minimize the loss of information. This can be formalized with the use of information theory, which allows one to define the requirements for an optimal estimator (Tegmark, Taylor, & Heavens 1997b). The central object in this discussion is the Fisher information matrix, defined as the ensemble average of the matrix of second derivatives of the minus log–likelihood function with respect to the parameters we wish to estimate. Its inverse provides a minimum bound for the covariance matrix of the parameters, known in the literature as the Cramér-Rao inequality (e.g., Kendall & Stuart 1969). The natural question to ask is which methods allow us to reach this theoretical limit. Maximum likelihood is obviously a preferred method in this sense, since it allows one to directly explore the likelihood function, finding its maximum and second derivatives around it. Unfortunately, to compute the likelihood function one has to invert the data matrix for each set of parameters, which becomes computationally prohibitive for large data sets. While linear compression methods, such as the Karhunen-Loève method (Bond 1995; Bunn & Sugiyama 1995; Tegmark et al. 1997b; Vogeley & Szalay 1996), allow us to reduce the size of the system if the data are oversampled, the compression is not sufficient for future surveys with millions of data points. The problem is present in all linear methods, because they keep the information on individual measurements, while what one wants to extract from a stochastic Gaussian process (thought to be a good model for matter distribution in the universe on large scales) is a quadratic quantity in the data. Under the assumption of isotropy, this can be averaged over all the pairs contributing to the same angular separation or mode amplitude, which then contains all the information about the cosmological model. One should therefore be able to compress the data to a correlation function or power spectrum estimator without losing any information in the case of Gaussian random processes. Recently, a quadratic estimator for the power spectrum has been proposed that allows one to reach the limit given by Fisher matrix directly (Hamilton 1997a). The explicit form of the estimator is particularly simple in its pixelized form if one assumes Gaussian distribution of the data (Tegmark 1997), and it has been shown that it gives results equivalent to the likelihood analysis (Bond et al. 1997). While the Gaussian assumption breaks down because of nonlinear evolution on small scales, it should be a good approximation for the CMB data and for galaxy and weak lensing surveys on large scales. Even in the nonlinear regime the Gaussian estimator remains a useful approximation, being much simpler to compute than the general expression. The main complication in the nonlinear regime is that the covariance matrix depends on the fourth moment, which has to be computed either using Monte Carlo N-body simulations or perturbation theory. Although both the correlation function and power spectrum have been used in the past to measure the statistical correlations between the data, there are several advantages to the power spectrum analysis. For example, a power spectrum can exhibit physical processes, such as acoustic oscillations in CMB, more clearly than the correlation function. More importantly, the observations of nonscalar quantities, such as velocity flows, weak lensing, or polarization of CMB, can be more easily related to the underlying theory in signal eigenmode space than in real space, where the correlation functions are correlated (Kaiser 1992; Seljak 1997a).

Computing the power spectrum from a survey is not the only interesting question that one can address with survey data. Often one wants to reconstruct the true underlying distribution, given the noisy and sparsely sampled set of observations. Such a reconstructed image can be used to search for individual objects, such as cold and hot spots in CMB maps or clusters, superclusters, filaments, and walls,

to count the number density of such objects as a function of their mass or luminosity or to look at their individual morphologies and substructure. Using a reconstructed map one can study the topology of the universe or compare different observations of the same region. A simple and powerful method of reconstruction is Wiener filtering (WF), which minimizes the variance in the class of linear estimators (Rybicki & Press 1992). Just like the minimum-variance estimator, WF does not require the data to be Gaussian distributed, but can be shown to be optimal in this limit, since it coincides with the maximum posterior probability estimator. As such, it is clearly the method of choice in the CMB and LSS reconstruction, where the deviations from Gaussianity are either small or negligible and has successfully been applied in these cases (Bunn, Hoffman, & Silk 1995; Tegmark 1997; Zaroubi et al. 1995 and references therein). The main shortcoming of WF is that it requires as input the power spectrum of the underlying field, and this is often not known in advance. In the absence of any external information it has to be obtained from the data themselves. We will show in this paper how to optimally extract the power spectrum from the WF map, thus providing a self-consistent treatment of WF.

The two analysis techniques, power spectrum estimation and image reconstruction, have so far been disconnected in the literature. Naive power spectrum estimation from the WF maps leads to biased results (a well-known consequence of WF to return zero in the regimes of low signal-to-noise ratio) and was not used for this reason. Optimal power spectrum estimation techniques are generally thought to require very specific analysis techniques, too different from image reconstruction to be able to address both problems simultaneously. It would, however, be useful if a single analysis method produced both power spectrum estimation and image reconstruction. It would be even more useful if such analysis would not increase the computational time for either of the methods. This would then be the preferred method even if only one application were the primary objective of the analysis. In this paper we show that there is a simple connection between the two methods, which gives the two desired properties above. This provides a unified approach to the two analysis techniques and allows one to obtain both the image reconstruction and minimum-variance power spectrum estimator within the same analysis. Moreover, this connection allows one to translate various calculational methods from one application to the other. For example, transformation to Fourier space is commonly employed in WF analysis and leads to a significant reduction of computational time for some problems. The same transformation can also be used for the minimum-variance power spectrum estimator. Conversely, approximation techniques developed for the minimum-variance power spectrum estimator can also be applied to WF analysis and can reduce its computational time significantly, allowing large maps to be reconstructed.

In this paper we concentrate on deriving a set of tools for image reconstruction and power spectrum estimation, paying special attention to the connections between the two. The main goal is to present a self-contained treatment of the two problems, so some of the ideas presented will not be fundamentally new. Given the many possible applications of these methods to the existing and forthcoming LSS and CMB surveys, it seems important to present them with a unified treatment. Among the possible applications are galaxy and weak lensing surveys, peculiar velocity flows, Lyα forest from quasar spectra, and CMB temperature and polarization maps. All these have a common feature of mapping the universe on large scales, where the techniques explored here work best. Some of these applications will be presented elsewhere. The outline of this paper is as follows: in § 2 we review the minimum-variance estimator. In § 3 we first review the WF reconstruction procedure and then show its connection to the minimum-variance estimator. In § 4 we discuss the reconstruction and estimation in signal eigenmode basis, where the calculations often become less computationally intensive. Section 5 applies the general expressions from previous sections to the case of plane wave expansion, which is of particular importance to the LSS surveys and where translational invariance and fast Fourier transform (FFT) techniques can be utilized to reduce the computational time. This is followed by the general discussion in § 6. Treatment of external parameters and other constraints is left for the Appendix.

2. MINIMUM-VARIANCE POWER SPECTRUM ESTIMATOR

In this section we review the formalism related to the minimum-variance estimator. Although this section has no major new results in itself, it is included here for completeness, since it establishes the notation used in the rest of the paper. For further details see Bond et al. (1997), Hamilton (1997a), and Tegmark (1997).

Let us assume that we measure the quantities $d(\boldsymbol{r}_i)$ at N positions $\boldsymbol{r}_i$, either in the sky or in space. We arrange these into a vector $\boldsymbol{d} = \{d(\boldsymbol{r}_i)\}(i = 1, \ldots, N)$. The measured quantities $\boldsymbol{d}$ can either be scalars, such as temperature anisotropy or density perturbation; components of a vector (e.g., radial peculiar velocity); or components of a tensor, such as polarization of microwave background or galaxy ellipticity in the weak lensing analysis. Each measurement consists of a signal and noise contributions, $\boldsymbol{d} = \boldsymbol{R}\boldsymbol{s} + \boldsymbol{R}_b\boldsymbol{S}_b + \boldsymbol{n}$. Here $\boldsymbol{R}$ is an $N \times M$ response matrix and $\boldsymbol{s} = \{s_\lambda\}(\lambda = 1, \ldots, M)$ are the underlying field coefficients that we wish to estimate. A true underlying field always has an infinite number of coefficients, but for computational reasons only a finite number of these can be estimated. For this reason we group all the modes that we do not estimate into $\boldsymbol{s}_b$, with $\boldsymbol{R}_b$ as the corresponding response matrix. The underlying field is parameterized with covariance matrices $\boldsymbol{S} = \langle \boldsymbol{s}\boldsymbol{s}^\dagger \rangle$ and $\boldsymbol{S}_b = \langle \boldsymbol{s}_b{}^\dagger \boldsymbol{s}_b \rangle$. The noise vector $\boldsymbol{n} = \{n_i\}(i = 1, \ldots, N)$ is parameterized with the noise covariance matrix $\boldsymbol{N} = \langle \boldsymbol{n}\boldsymbol{n}^\dagger \rangle$. This noise matrix is often known a priori, uncorrelated with the signal and diagonal in real space (the former is not exactly true if the noise is coupled to the signal, as is the case for shot noise in galaxy clustering; the latter is not true if $1/f$ noise is important, as is the case in some, but not all, CMB experiments). When we take the response functions to be the mode expansion coefficients, then the signal covariance matrices $\boldsymbol{S}$ and $\boldsymbol{S}_b$ are also diagonal and only depend on the amplitude of the mode. The correlation matrix of the data is given by

$$\langle \boldsymbol{d}\boldsymbol{d}^\dagger \rangle \equiv \boldsymbol{C} = \boldsymbol{R}\boldsymbol{S}\boldsymbol{R}^\dagger + \boldsymbol{C}_b + \boldsymbol{N} = \sum_l \Theta_l \boldsymbol{Q}_l + \boldsymbol{N} . \quad (1)$$

Here $\boldsymbol{R}\boldsymbol{S}\boldsymbol{R}^\dagger = \sum_{l=1}^{l_{\max}} \Theta_l \boldsymbol{Q}_l$ was simplified by summing over all the M_l modes $\boldsymbol{s}_l = \{s_{m_l}\}(m_l = 1, \ldots, M_l)$ whose mode amplitude contributes to the portion of the power spectrum parameterized with Θ_l, so that $\langle s_{m_l} s^*_{m_l} \rangle = \Theta_l$. The aliasing term $\boldsymbol{C}_b$ was similarly rewritten from individual modes to a

sum (or integral) over the power spectrum, $\boldsymbol{C}_b = \boldsymbol{R}_b \boldsymbol{S}_b \boldsymbol{R}_b^\dagger = \sum_{l=l_{\rm max}+1}^{\infty} \Theta_l \boldsymbol{Q}_l$. We can introduce a projection matrix $\boldsymbol{\Pi}_l$, which consists of ones along the diagonal corresponding to the M_l modes and zeros otherwise, in terms of which $\boldsymbol{Q}_l = \boldsymbol{\Pi}_l \boldsymbol{R} \boldsymbol{R}^\dagger \boldsymbol{\Pi}_l$. For example, in the spherical harmonic decomposition of temperature anisotropies in the sky we can decompose them as $\Delta(\boldsymbol{n}_i) = \sum a_{lm} Y_{lm}(\boldsymbol{n}_i)$, where $Y_{lm}(\boldsymbol{n}_i)$ are the spherical harmonics in the direction $\boldsymbol{n}_i$. If we choose to model the field with all the modes up to $l_{\rm max}$ then the underlying field is $\boldsymbol{s} = \{s_\lambda\} = \{a_{lm}\}(\lambda = l^2 + l + m + 1,\ l = 1,\ \ldots,\ l_{\rm max},\ m = -l,\ \ldots,\ l)$ and $\boldsymbol{s}_b = \{a_{lm}\}(l = l_{\rm max}+1,\ \ldots,\ \infty)$, while the response matrices are $R_{\lambda,i} = Y_{lm}(\boldsymbol{n}_i)(l = 1,\ \ldots,\ l_{\rm max})$ and $R_{b(\lambda,i)} = Y_{lm}(\boldsymbol{n}_i)(l = l_{\rm max} + 1, \ldots, \infty)$. The correlation matrix has contributions from all l, $\langle a_{lm} a^*_{l'm'} \rangle = C_l \delta_{ll'} \delta_{mm'}$, and we can identify $\Theta_l = C_l$. The projection matrix $\boldsymbol{\Pi}_{l,\lambda,\lambda'}$ has ones for $\lambda = \lambda' = l^2 + l + m + 1, m = -l, \ldots, l$ and zeros otherwise. Finally $Q_{l,ij} = \sum_m Y_{lm}(\boldsymbol{n}_i) Y^*_{lm}(\boldsymbol{n}_j)$ and using the addition theorem for spherical harmonics we find $Q_{l,ij} = [(2l + 1)/4\pi] P_l(\boldsymbol{n}_i \cdot \boldsymbol{n}_j)$, where P_l is the Legendre polynomial. The above notation is, however, more general and allows to parameterize the power spectrum with fewer parameters. If the survey size D is such that $\Delta l \sim 1/D \gg 1$, or if we have reasons to believe that the underlying power spectrum is smooth, then we may group together neighboring mode amplitudes in the power spectrum and parameterize them with a single parameter Θ_l, where now l simply counts the parameters we are trying to estimate and $l_{\rm max}$ is their total number. The corresponding $\boldsymbol{Q}_l$ is still a sum over squares of all individual modes that contribute to the lth parameter Θ_l. For surveys with different geometry a different expansion (e.g., in plane waves) may be more appropriate, but it can nevertheless be written in the above form (replacing some of the sums with integrals). Of course, other parameterizations of the power spectrum are also possible, for example, its amplitude and slope, but for our purposes parameterization in terms of its value at a given mode amplitude will be the most useful form, since it can be directly related to the underlying field $\boldsymbol{s}$ that we wish to estimate. Other parameterizations with fewer parameters could, however, be used for the power spectrum $\boldsymbol{S}_b$ that characterizes the modes $\boldsymbol{s}_b$ that we do not estimate, especially when we do not have sufficient information on them, but we nevertheless wish to estimate their influence on the parameters we are estimating.

Given a set of measurements $\boldsymbol{d}$ we wish to find the most probable set of parameters $\boldsymbol{\Theta}$. For Gaussian theories all the information from the data is contained in the likelihood function,

$$L(\boldsymbol{d}\,|\,\boldsymbol{\Theta}) = (2\pi)^{-N/2} \det(\boldsymbol{C})^{-1/2} \exp\left(-\tfrac{1}{2}\boldsymbol{d}^\dagger \boldsymbol{C}^{-1} \boldsymbol{d}\right), \quad (2)$$

where $\boldsymbol{C}$ implicitly depends on the parameters $\boldsymbol{\Theta}$. To find the most probable set of parameters one needs to find the maximum of the likelihood function above.[1] By Taylor expanding the log of the likelihood function $L(\boldsymbol{d}\,|\,\boldsymbol{\Theta})$ around the model parameters $\hat{\boldsymbol{\Theta}}$ that maximize it, we obtain the matrix of second derivatives, called the curvature matrix. Because the likelihood function is nonlinear as a function of parameters, one has to sample it in a number of parameters to find the location of the maximum and the second derivatives around it. Rather than computing the curvature matrix, which is computationally intensive, one often computes its ensemble average, called the Fisher matrix:

$$F_{ll'} = -\left\langle \frac{\partial^2 \ln L(\boldsymbol{\Theta}; \boldsymbol{d})}{\partial \Theta_l \partial \Theta_{l'}} \right\rangle_{\hat{\boldsymbol{\Theta}}}. \quad (3)$$

The angle brackets denote ensemble averaging. When the likelihood function can be approximated as a Gaussian around the maximum and the maximum likelihood estimator is close to the true value, then one can interpret the inverse of the Fisher matrix $\boldsymbol{F}^{-1}$ as an estimate of the covariance matrix of the parameters $\hat{\boldsymbol{\Theta}}$:

$$\langle \hat{\boldsymbol{\Theta}} \hat{\boldsymbol{\Theta}}^\dagger \rangle - \langle \hat{\boldsymbol{\Theta}} \rangle \langle \hat{\boldsymbol{\Theta}} \rangle^\dagger = \boldsymbol{F}^{-1}. \quad (4)$$

Note that the Gaussian assumption of the likelihood function does not rely on the Gaussianity of the data; the likelihood function will be Gaussian around the maximum provided that sufficient independent modes contribute to each Θ_l by the central limit theorem. Cramér-Rao inequality (e.g., Kendall & Stuart 1969) states that the best an unbiased estimator can do is reach the limit given by the Fisher matrix, i.e., $\sigma(\hat{\Theta}_l) \geq (\boldsymbol{F}^{-1})_{ll}^{1/2}$. Clearly, a brute force maximum likelihood search will achieve this limit, but it is not the fastest method, and the computational cost in many cases becomes prohibitively expensive.

The computational difficulties mentioned above lead a number of authors to investigate a quadratic estimator that allows one to find the maximum of the likelihood more rapidly (Hamilton 1997a; Bond et al. 1997; Tegmark 1997). For the case where the data are Gaussian distributed with zero mean the estimator is

$$\hat{\Theta}_l = \tfrac{1}{2} \sum_{l'} F^{-1}_{ll'} (\boldsymbol{d}^\dagger \boldsymbol{C}^{-1} \boldsymbol{Q}_{l'} \boldsymbol{C}^{-1} \boldsymbol{d} - b_{l'}). \quad (5)$$

Here b_l is the noise term, which can be obtained by computing the ensemble average of $\boldsymbol{d}^\dagger \boldsymbol{C}^{-1} \boldsymbol{Q}_l \boldsymbol{C}^{-1} \boldsymbol{d}$, assuming $\Theta_l = 0$ for $l \leq l_{\rm max}$. This gives

$$b_l = \mathrm{tr}\left[(\boldsymbol{N} + \sum_{l_{\rm max}+1}^{\infty} \Theta_l \boldsymbol{Q}_l) \boldsymbol{C}^{-1} \boldsymbol{Q}_l \boldsymbol{C}^{-1}\right]. \quad (6)$$

The term $\sum_{l_{\rm max}+1}^{\infty} \Theta_l \boldsymbol{Q}_l$ has been left out in previous work because it was implicitly assumed that all the relevant modes are being estimated. In practice, this is not the case when the data are sparsely sampled and have low noise. In this case the mixing between the modes aliases power from small scales to large scales and this term may dominate over the usual noise term. One could argue, of course, that if the data still have some excess of power above the noise then one should try to estimate this power rather than treat it as an additional noise term. There are several instances when this may not be practical. For example, when the sampling is very sparse (e.g., in supernova measurements of velocity field), then there are simply too many small-scale modes to include them all in the parameterization. Another possibility is that we have already measured the power on small scales accurately by some other means, so we do not need to include those modes for estimation. In both cases one can include the effect of small-scale modes as an additional noise term, which has to be included in equation (6) for the estimator to be unbiased. In any case one should at least check whether the aliasing is important by computing its contribution, assuming a reasonable (or already measured)

[1] All the discussion here is also valid from a Bayesian point of view, provided that the likelihood is interpreted as an a posteriori probability. In the case of a uniform prior in the parameters the above expressions remain the same, otherwise the appropriate prior has to be added (e.g., Bond et al. 1997).

power spectrum for the unestimated small-scale modes. The Fisher matrix $\boldsymbol{F}$ is given by

$$F_{ll'} = \tfrac{1}{2} \operatorname{tr} (\boldsymbol{Q}_l \boldsymbol{C}^{-1} \boldsymbol{Q}_{l'} \boldsymbol{C}^{-1}) . \quad (7)$$

This estimator has the nice properties of being unbiased, $\langle \hat{\Theta} \rangle = \Theta$, and has the minimum variance as defined in equation (4) (Tegmark 1997).

The estimator remains unbiased when the Gaussian assumption is dropped (since its mean depends only on the second moment of the distribution), and so one can continue to use it in the nonlinear regime. In this case the estimator as written above is no longer a minimum-variance estimator, although its computational simplicity compared with the general case (Hamilton 1997a) may outweigh the loss of information in practical applications. The more important difficulty in this case is that the covariance matrix for the estimators is no longer given by the inverse of $\boldsymbol{F}$ as defined in equation (7), because it depends on the four-point function of the distribution, which does not vanish for the non-Gaussian process. It has to be computed explicitly, for example, by using Monte Carlo methods on N-body simulations or by perturbation theory. Provided this is properly computed then the estimator above remains a useful working model for the power spectrum estimation even in the non-Gaussian regime.

Although we wrote the optimal estimator in equation (5) in terms of $\boldsymbol{F}^{-1}$, inverting the Fisher matrix may not be stable when the matrix is nearly singular. This happens if the binning of the power spectrum is too narrow for the survey geometry, i.e., $\Delta l \ll 1/D$, where D is the size of the survey. Two simple solutions around this problem follow:

1. Parameterize the power spectrum with fewer parameters, essentially coarse graining the spectrum. This way the Fisher matrix is not singular and can be inverted. One has to be careful to include the effects of a changing power spectrum across the band. Each mode of $\boldsymbol{R}$ contributing to $\boldsymbol{Q}_l = \boldsymbol{\Pi}_l \boldsymbol{R}\boldsymbol{R}^{\dagger} \boldsymbol{\Pi}_l$ gives an estimate of $\boldsymbol{S}$ at that mode amplitude, and if this is changing across the band then it is better to estimate a quantity that is flat across the band. To do this we divide each mode component in $\boldsymbol{d}^{\dagger} \boldsymbol{C}^{-1} \boldsymbol{Q}_l \boldsymbol{C}^{-1} \boldsymbol{d}$ with the prior power spectrum (one that we believe describes well the change of power spectrum across the band) evaluated at the amplitude of that mode. In the end we put this term back by multiplying the power spectrum estimate with the prior power spectrum evaluated at the mode amplitude to which we want to assign the estimator. This procedure will guarantee an unbiased result (see Bond et al. 1997 for a similar treatment) and can be quite important when the power spectrum is steep and binning coarse.

2. Do not attempt to invert $\boldsymbol{F}$ but instead quote filtered estimators $N_l \boldsymbol{F} \hat{\boldsymbol{\Theta}}$, where $N_l = (\sum_{l'} F_{ll'})^{-1}$. The filtering function $N_l \boldsymbol{F}$ is a bell-shaped function in l' for a given l, and its width around l tells us the spectral resolution at that l. The covariance matrix for this estimator is given by $N_l N_{l'} F_{ll'}$ (Tegmark 1997; Bond et al. 1997). More complicated methods that make the power spectrum estimators uncorrelated have also been proposed (Hamilton 1997b; Tegmark 1997; Bond et al. 1997). They still require one to report the full filtering matrix and are useful mainly if one wishes to further compress the estimators, for example, for the purpose of plotting them on a graph.

Once a power spectrum estimate has been obtained, one wants to test different cosmogonies or extract cosmological parameters from it. The interesting question is whether the optimal way to perform this step is using the data or the quadratic estimators. A comparison between the two methods should help to elucidate the similarities and differences between the two. In the linear case we use the actual data in the analysis and try to maximize the likelihood function. This should be done using the full probability distribution of the data, but, in practice, for large data sets one can only perform the analysis under the assumption that the data are Gaussian distributed. One then computes the likelihood function using the correlation matrix of the model one is testing and compares it with the maximum likelihood value. In the quadratic case one first compresses the data to the power spectrum estimators, which are quadratic in the data, and performs the likelihood analysis using the covariance matrix of these estimators. Here again one should use the covariance matrix obtained from the model one is testing and, for consistency, do the power spectrum analysis using that model as well. While the actual values of power spectrum estimators do not depend sensitively on the assumed model, the covariance matrix does if one is limited by the sampling variance, because in this limit the error on the estimate is dominated not by the noise but by the assumed power spectrum. It should be emphasized that in both cases to test a model one has to use the covariance matrix from that model, not from the best-fit model. In general, one therefore cannot attach model-independent error bars to the power spectrum estimator for the purpose of model testing. If we assume that the covariance matrix has been calculated using the model that one is testing, then the only difference between the linear and quadratic approach lies in the Gaussian assumption for the model entering the likelihood function, either the actual data or quadratic power spectrum estimators. The answer then depends on the particular application; if we have reasons to believe the data are Gaussian distributed (as should be the case, e.g., for CMB anisotropies), then linear analysis is clearly correct. For quadratic analysis the distribution for a given power spectrum estimator Θ_l is roughly $\chi^2(M_l)$, where M_l is the effective number of independent modes contributing to Θ_l. Only if M_l is large can this distribution be approximated as a Gaussian by central limit theorem. On the other hand, if the distribution of individual data points is not Gaussian the linear analysis is not correct (or becomes much more complicated if a non-Gaussian likelihood function is used), while the distribution of a quadratic estimator can still be Gaussian in the same large M_l limit. Of course in this case the covariance matrix for the estimators will not be given by equation (7), which was derived under the Gaussian assumption for the data, but provided that this matrix can be obtained by some other means, the Gaussian distribution for the quadratic estimators is correct. The answer therefore depends on the nature of the data and on the size of the survey. In general, for estimations on large scales that are sampling variance limited and likely to be Gaussian, a quadratic estimator will be the least reliable, but since these modes also have the largest errors attached, the overall error may not be very important, at least for finding the most probable model. The effects of the Gaussian assumption are somewhat more important for setting the confidence limits, where the tails of distribution are important (Bond et al. 1997).

The main advantage of the quadratic estimator is that it compresses the data to a small number of values and their

covariance matrix, and they are then easier to manipulate than the original data set. The compression here can be truly enormous, from millions of measurements to hundreds of power spectrum estimates in the case of future surveys. As we argued above, for the purpose of parameter extraction the power spectrum covariance matrix cannot be assigned in a model-independent way, and in principle one would have to repeat the quadratic estimator for each model one is testing. For practical purposes this is not necessary, and one can approximately attach meaningful error bars to the power spectrum estimators. Since we assumed that the power spectrum is known, whereas in reality this is something we wish to estimate, one is tempted to iterate the power spectrum estimation until the convergence is reached. This, however, will not be the best solution, because if the estimators have a large sampling variance then some will be low and some high, and those that are low will also have a smaller attached sampling error, since it was based on the assumption that the true power spectrum is the measured one. This will therefore bias any subsequent estimation of parameters. Two ways to resolve this problem are either to group the parameters so that the sampling variance is reduced or to find a smooth curve that goes through the power spectrum estimators and use that as the assumed power spectrum in the next step of iteration. The criterion for the smoothness of the curve should be based on the requirement that $\chi^2 = (\boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}})^\dagger \boldsymbol{F}^{-1}(\boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}}) \approx l_{\max}$, where $\boldsymbol{\Theta}$ is the parameterization of the smooth power spectrum. In other words, the assumed power spectrum should have sufficient detail to be reasonably close to the measured values, yet it should not attempt to match those values more accurately than permitted by the attached errors. Since the covariance matrix in χ^2 depends on the assumed power spectrum in the previous step, it may be necessary to repeat the iteration two more times, to a total of three estimations.

3. WIENER FILTER RECONSTRUCTION

The goal of image reconstruction deals with the following problem: given incompletely sampled and noisy data $\boldsymbol{d}$ we want to reconstruct the true underlying field $\boldsymbol{s}$ or true image $\boldsymbol{Rs}$ so that the reconstructed field is in some sense the closest to the real field. One way to approach this problem is to require the estimated field $\hat{\boldsymbol{s}}$ to be a linear function of the data, $\hat{\boldsymbol{s}} = \boldsymbol{\Phi d}$, where $\boldsymbol{\Phi}$ is an $N \times M$ matrix. Then one can minimize the variance of the residual

$$\langle (\boldsymbol{s} - \hat{\boldsymbol{s}})(\boldsymbol{s} - \hat{\boldsymbol{s}})^\dagger \rangle \tag{8}$$

with respect to $\boldsymbol{\Phi}$. This way we find the WF estimator of the underlying field,

$$\hat{\boldsymbol{s}} = \langle \boldsymbol{s}\boldsymbol{d}^\dagger \rangle \langle \boldsymbol{d}\boldsymbol{d}^\dagger \rangle^{-1} = \boldsymbol{S}\boldsymbol{R}^\dagger \boldsymbol{C}^{-1}\boldsymbol{d} \, . \tag{9}$$

The variance of residuals is given by

$$\langle \boldsymbol{r}\boldsymbol{r}^\dagger \rangle = \langle (\hat{\boldsymbol{s}} - \boldsymbol{s})(\hat{\boldsymbol{s}} - \boldsymbol{s})^\dagger \rangle = \boldsymbol{S} - \boldsymbol{S}\boldsymbol{R}^\dagger \boldsymbol{C}^{-1}\boldsymbol{R}\boldsymbol{S} \, . \tag{10}$$

The reconstruction here is written in terms of mode expansion coefficients. To make a real space map we transform them back to the real space, giving the reconstructed image $\hat{\boldsymbol{d}} = \boldsymbol{R}\hat{\boldsymbol{s}}$.

For Gaussian random fields Wiener filtering coincides with the maximum probability estimator (Zaroubi et al. 1995), and so it is optimal. To show this one can write the joint probability of signal $\boldsymbol{s}$ and noise $\boldsymbol{n}$ as a product of individual probabilities, under the assumption that they are uncorrelated:

$$P(\boldsymbol{s}, \boldsymbol{n}) = P_s(\boldsymbol{s})P_n(\boldsymbol{n}) = (2\pi)^{-(M+N)/2} \det (\boldsymbol{S})^{-1/2} \times \det (\boldsymbol{N})^{-1/2} \exp \left[-\tfrac{1}{2}(\boldsymbol{s}^\dagger \boldsymbol{S}^{-1}\boldsymbol{s} + \boldsymbol{n}^\dagger \boldsymbol{N}^{-1}\boldsymbol{n}) \right] . \tag{11}$$

For simplicity we have assumed that the noise from aliasing of unestimated modes is included in the noise matrix $\boldsymbol{N}$. We will derive this term explicitly in equation (18). The conditional probability for $\boldsymbol{s}$ given the data is $P(\boldsymbol{s} \mid \boldsymbol{d}) \propto P_s(\boldsymbol{s})P_n(\boldsymbol{d} - \boldsymbol{Rs})$, which after completing the square gives

$$P(\boldsymbol{s} \mid \boldsymbol{d}) \propto \exp \left[-\tfrac{1}{2}(\boldsymbol{s} - \hat{\boldsymbol{s}})^\dagger (\boldsymbol{S}^{-1} + \boldsymbol{R}^\dagger \boldsymbol{N}^{-1}\boldsymbol{R})(\boldsymbol{s} - \hat{\boldsymbol{s}}) \right] , \tag{12}$$

where $\hat{\boldsymbol{s}}$ is given in equation (9). The covariance matrix of residuals $\boldsymbol{S}^{-1} + \boldsymbol{R}^\dagger \boldsymbol{N}^{-1}\boldsymbol{R}$ is identical to the one in equation (10), as can be obtained by matrix manipulation using the matrix expansion $(\boldsymbol{I} + \boldsymbol{X})^{-1} = (\boldsymbol{I} - \boldsymbol{X} + \boldsymbol{X}\boldsymbol{X} - \cdots)$. The posterior probability for $\boldsymbol{s}$ is therefore maximized by the WF solution $\hat{\boldsymbol{s}}$.

In the above example we wanted to find the most probable field $\hat{\boldsymbol{s}}$ given the data $\boldsymbol{d}$. On the other hand, in the previous section we wanted to maximize the probability or likelihood function of the data as a function of the parameters Θ_l independent of the underlying field $\hat{\boldsymbol{s}}$. In this case one wants to marginalize over the parameters $\hat{\boldsymbol{s}}$ (Rybicki & Press 1992):

$$\begin{aligned} P(\boldsymbol{d}) &= \int P(\boldsymbol{s}, \boldsymbol{d} - \boldsymbol{Rs}) d^M \boldsymbol{s} \, , \\ &= (2\pi)^{-(M+N)/2} \det (\boldsymbol{S})^{-1/2} \det (\boldsymbol{N})^{-1/2} \\ &\quad \times \exp \left(-\tfrac{1}{2}\boldsymbol{d}^\dagger \boldsymbol{C}^{-1}\boldsymbol{d} \right) \\ &\quad \times \int \exp \left[-\tfrac{1}{2}(\boldsymbol{s} - \hat{\boldsymbol{s}})^\dagger (\boldsymbol{S}^{-1} \right. \\ &\quad \left. + \boldsymbol{R}^\dagger \boldsymbol{N}^{-1}\boldsymbol{R})(\boldsymbol{s} - \hat{\boldsymbol{s}}) \right] d^M \boldsymbol{s} \, , \\ &= (2\pi)^{-N/2} \det (\boldsymbol{C})^{-1/2} \exp \left(-\tfrac{1}{2}\boldsymbol{d}^\dagger \boldsymbol{C}^{-1}\boldsymbol{d} \right) . \end{aligned} \tag{13}$$

We see that to marginalize over a set of parameters $\boldsymbol{s}$ one has first to find their maximum probability value $\hat{\boldsymbol{s}}$, after which the integration over the parameters can be trivially performed. In this example the final result does not depend on the underlying field, and we could have directly written the probability distribution (or the likelihood function) for the data given the theory, as we did in the previous section (the derivation here merely provides a more rigorous justification for the statement in footnote 1). In other cases (e.g., when the data depend nonlinearly on the field) going through this intermediate step is necessary.

WF only uses information on the mean and variance of statistical distribution, which completely characterizes Gaussian random fields and so is optimal for such applications. On the other hand, by using only this information it may be less than optimal when applied to a strongly non-Gaussian field. A typical example is image reconstruction from incompletely sampled data in Fourier space, when the image has significant point sources. In this case nonlinear methods, such as the maximum entropy method, are the methods of choice, since they allow one to better extract the smooth component underneath the ripples produced by point sources (Narayan & Nityananda 1986). But in the applications to LSS or CMB with small or no deviations from Gaussianity, mean and variance contain all the information on the statistical properties of the field and WF is the optimal method. It should be emphasized that even in

non-Gaussian situations WF still minimizes the variance as defined in equation (8) among the class of linear estimators. For example, for cluster reconstruction from weak lensing there exist several methods that are linear in the data (Squires & Kaiser 1996), but since WF explicitly minimizes the variance in equation (8) it is in this sense optimal in this class of reconstruction methods (Seljak 1997b). However, typically in such applications minimizing the variance may not be the only way to define the best image reconstruction, and in general there is no optimal reconstruction for all purposes.

To make a connection between WF and the minimum-variance estimator we rewrite equation (5) as

$$(\boldsymbol{F}\hat{\boldsymbol{\Theta}})_l = \tfrac{1}{2}[(\boldsymbol{\Pi}_l \boldsymbol{R}^\dagger \boldsymbol{C}^{-1}\boldsymbol{d})^\dagger (\boldsymbol{\Pi}_l \boldsymbol{R}^\dagger \boldsymbol{C}^{-1}\boldsymbol{d}) - b_l] \, . \quad (14)$$

By comparing equations (9) and (14) one finds that the minimum-variance power spectrum estimator can be simply expressed in terms of a WF reconstructed field,

$$(\boldsymbol{F}\hat{\boldsymbol{\Theta}})_l = \tfrac{1}{2}(\boldsymbol{\Pi}_l \boldsymbol{S}^{-1}\hat{\boldsymbol{s}}^\dagger \hat{\boldsymbol{s}}\boldsymbol{S}^{-1}\boldsymbol{\Pi}_l - b_l) \, , \quad (15)$$

where

$$b_l = \mathrm{tr}\, [\boldsymbol{\Pi}_l \boldsymbol{R}^\dagger \boldsymbol{C}^{-1}(\boldsymbol{N} + \boldsymbol{C}_b)\boldsymbol{C}^{-1}\boldsymbol{R}\boldsymbol{\Pi}_l] \quad (16)$$

and

$$F_{ll'} = \tfrac{1}{2}\mathrm{tr}\, (\boldsymbol{C}^{-1}\boldsymbol{Q}_l\boldsymbol{C}^{-1}\boldsymbol{Q}_{l'}) = \tfrac{1}{2}\sum_{m_l}\sum_{m_{l'}} |\, \boldsymbol{\Pi}_l \boldsymbol{R}^\dagger \boldsymbol{C}^{-1}\boldsymbol{R}\boldsymbol{\Pi}_{l'}\,|^2_{m_l m_{l'}} \, . \quad (17)$$

These expressions provide additional insight into the minimum-variance power spectrum estimator and its relation to the WF. To compute the modes both methods first weight the data by multiplying them with the inverse covariance matrix $\boldsymbol{s} = \boldsymbol{R}^\dagger \boldsymbol{C}^{-1}\boldsymbol{d}$. This is not surprising, since it is just a generalization of the usual inverse variance weighting of the data, where now a given measurement can be downweighted either because it has a large measurement error or because it is strongly correlated with other measurements (in which case it does not provide additional information). The minimum-variance power spectrum estimator then simply averages over all the modes contributing to the lth parameter in the spectrum, while WF filters those that are below the noise. To properly normalize the power spectrum estimates and to provide their covariance matrix we also need to subtract the noise bias b_l and compute the Fisher matrix $\boldsymbol{F}$. This estimator therefore improves on the naive power spectrum estimation obtained by the simple average $M_l^{-1}(\boldsymbol{\Pi}_l \hat{\boldsymbol{s}}^\dagger \hat{\boldsymbol{s}}\boldsymbol{\Pi}_l)$, which leads to biased results in the low signal-to-noise regimes (e.g., Bunn et al. 1996). It is also the best possible estimator, as discussed in the previous section. Clearly it should be used whenever WF on the data is performed and one wants to obtain a power spectrum estimate as well. Since WF requires the power spectrum as input in the absence of any external information one can use the actual power spectrum as measured from the data themselves for WF reconstruction (or a smoothed version of it as discussed in § 2). This gives a self-consistency to the WF reconstruction in the sense that it does not have to rely on any ad hoc assumptions when there is no external information. Since WF is basically multiplying the modes with signal/(signal + noise) one can see that the modes with no statistically significant excess of power are being filtered out and replaced with zero. Only the modes where the power spectrum does show an excess of power above the noise will be kept in the reconstruction, thus providing the most conservative reconstruction that is consistent with the data. For Gaussian random fields this is in fact the optimal reconstruction.

The most expensive operation in computing WF is inverting $\boldsymbol{C}$, and once this inversion is obtained it is straightforward to compute b_l and $\boldsymbol{F}$, so computing the power spectrum from WF estimators is not significantly more expensive than computing WF itself. This inversion is $O(N^3)$ operations, if one uses Cholesky decomposition.[2] This is not computationally feasible for systems larger than $N \sim 10^4$, and so it is useful to reduce the size of the system when possible. We discuss this in the next section.

4. SIGNAL EIGENMODE REPRESENTATION

As mentioned in the previous section the computational cost of performing both WF and minimum-variance power spectrum estimation scales as $O(N^3)$. Clearly, reducing the size of the matrix is the first simplification one should attempt to try. One way to achieve this is to note that the data are often oversampled; in the case of *COBE*, for example, there are about 4000 pixels after the galactic cut, but only about 900 modes are required to fit these data, as *COBE* has no information about the modes above $l \sim 30$. Similarly, if the signal-to-noise ratio for each measurement is small (as is the case in weak lensing, with intrinsic ellipticity of individual galaxies being much larger than the signal we are after), then the number of data points will in general be much larger than the number of modes that can be extracted. This suggests that we should transform the data to the signal eigenmode basis first and perform all the operations there. For computational reasons we will do this by dividing the data first with the inverse of the noise matrix, so that our data in the signal eigenmode basis are $\tilde{\boldsymbol{d}} = \boldsymbol{R}^\dagger \boldsymbol{N}^{-1}\boldsymbol{d}$[3]. Using this vector as a new data set we can derive analogous expressions to equations (9), (16), and (17),

$$\hat{\boldsymbol{s}} = [\boldsymbol{S}^{-1} + \boldsymbol{S}^{-1}(\boldsymbol{R}^\dagger \boldsymbol{N}^{-1}\boldsymbol{R})^{-1}\boldsymbol{R}^\dagger \boldsymbol{N}^{-1}\,\boldsymbol{C}_b\,\boldsymbol{N}^{-1}\boldsymbol{R} + \boldsymbol{R}^\dagger \boldsymbol{N}^{-1}\boldsymbol{R}]^{-1}\tilde{\boldsymbol{d}} \equiv \boldsymbol{D}^{-1}\tilde{\boldsymbol{d}} \, , \quad (18)$$

$$b_l = \mathrm{tr}\, [\boldsymbol{\Pi}_l \boldsymbol{S}^{-1}\boldsymbol{D}^{-1}\boldsymbol{R}^\dagger(\boldsymbol{N}^{-1}\boldsymbol{C}_b\boldsymbol{N}^{-1} + \boldsymbol{N}^{-1})\boldsymbol{R}\boldsymbol{D}^{-1}\boldsymbol{S}^{-1}\boldsymbol{\Pi}_l] \, , \quad (19)$$

and

$$F_{ll'} = \tfrac{1}{2}\sum_{m_l}\sum_{m_{l'}} |\, \boldsymbol{\Pi}_l \boldsymbol{S}^{-1}\boldsymbol{D}^{-1}\boldsymbol{R}^\dagger \boldsymbol{N}^{-1}\boldsymbol{R}\boldsymbol{\Pi}_{l'}\,|^2_{m_l m_{l'}} \, , \quad (20)$$

whereas equation (15) remains unchanged.

The role of the correlation matrix $\boldsymbol{C}$ in real space has now been replaced by $\boldsymbol{D}$, which has dimensions $M \times M$. If M is significantly smaller than N then a substantial savings in computational time can be achieved by having to invert a smaller matrix. This matrix has three contributions: signal, aliasing, and noise terms. By using this expression we also

[2] Iterative procedures such as multigrid methods coupled with Jacobi or Gauss-Seidel iteration can, in principle, provide significant speedups for very large systems. Such methods could be used for WF but not for power spectrum estimation, because to compute $\boldsymbol{b}$ or $\boldsymbol{F}$ one has to solve $\boldsymbol{C}^{-1}\boldsymbol{R}$, meaning the iteration has to be repeated M times and one is better off computing $\boldsymbol{C}^{-1}$ explicitly. A possible solution is to compute $\boldsymbol{C}^{-1}\boldsymbol{d}$ using an iterative method and repeat this for a set of Monte Carlo realizations of the data and noise. By averaging over these one can compute $\boldsymbol{F}$ and $\boldsymbol{b}$ directly.

[3] This is computationally advantageous only if the noise matrix is diagonal. If this is not the case then one can perform the transformation directly on the data, $\tilde{\boldsymbol{d}} = \boldsymbol{R}^\dagger \boldsymbol{d}$. This leads to a similar set of expressions to those given in thetext and will not be explicitly presented here.

need to compute $\boldsymbol{R}^{\dagger}\boldsymbol{N}^{-1}\boldsymbol{R}$, which in principle requires M^2N multiplications, although if one can take advantage of plane wave expansion and FFTs (or their analogs for different basis functions) then this can be significantly sped up (see § 5). To compute the aliasing term this matrix needs to be inverted, which is $O(M^3)$, but computing the complete aliasing term in fact results in $O(MN^2 + M^2N)$, which becomes the most expensive operation. Fortunately, this term can often be neglected, in which case computing $\boldsymbol{D}$ simplifies it further (it also makes it symmetric and hence it can be inverted with Cholesky rather than LU decomposition). Matrix $\boldsymbol{D}^{-1}$ also gives the variance of residuals as shown in equation (12) (where now the generalized noise term has been explicitly divided into aliasing and noise contributions). The same manipulations also reduce the computational cost to compute $\boldsymbol{b}$ and $\boldsymbol{F}$ as well. All the expensive operations scale with M only, and there is no $O(N^3)$ left in the problem.

The advantage of this transformation has been noted for WF applications (e.g., Bunn et al. 1996), but obviously it is equally useful for the minimum-variance power spectrum estimation. The computational advantage of signal eigenmode representation is particularly important for weak lensing or CMB polarization analysis, if only one of the two independent modes is excited (as is the case for weak lensing and often also for CMB polarization). Then by transforming to signal eigenmodes the size of the system is immediately reduced by a factor of 2 (and hence the computational cost by a factor of 8), even if we use the same number of eigenmodes as the number of data points (Seljak 1997b). A further advantage in using the expressions above is that no pixelization of the data is necessary. This is because the response matrix $\boldsymbol{R}$ can be evaluated at the exact positions of the measurements and so the projection of the data to a given mode involves no approximations. This would be particularly important if one wants to determine the power spectrum on small scales, where pixelization can introduce significant errors in the method. The only potential disadvantage of working on this basis is that the matrix $\boldsymbol{D}$ can be singular, even though it is regularized with the addition of $\boldsymbol{S}^{-1}$. This happens, for example, when the data are sparse or there are large unobserved regions in the survey. This is not so much a problem for WF, since one can use singular value decomposition (SVD) (Press et al. 1992) to decompose the matrix and zero all the small eigenvalues, which would destabilize the inversion. One cannot, however, use this to estimate the power spectrum, because it leads to biased results.

5. PLANE WAVE EXPANSION

We now apply the formulae derived in previous sections to the case where the eigenmodes are plane waves. This is the case of significant importance, since it can be applied to the analysis of galaxy survey data (in one, two, and three dimensions), weak lensing, Lyα forest, and small-scale CMB anisotropies. We will assume that the noise is uncorrelated between the measurements, as is usually the case here. There are two simplifications with respect to the general case: the first is if noise is diagonal, which allows one to compute $\boldsymbol{R}^{\dagger}\boldsymbol{N}^{-1}\boldsymbol{R}$ with $O(MN)$ operations instead of $O(M^2N)$. The second is the use of FFT, which further reduces the operational count of this operation to $O(N \ln N)$. We also discuss an approximation scheme that can be developed for the solution, allowing very large systems to be solved. Although an analogous approximation can also be developed for the general case, the two advantages of plane wave expansion make it particularly suitable to apply it in this case.

Given the data $\boldsymbol{d}$ we first interpolate them to a grid with N_g points in each of N_d dimensions. The size of the grid D in each dimension has to be such that it covers all the data available,[4] while the grid spacing should be at least one-half of the smallest scale we hope to resolve with the data. The reason for this choice will become apparent later. We will assume here, for simplicity, that the interpolation is simply an assignment to the nearest grid point. This is not essential because as we will show all the operations scale as $O(N \ln N)$, where $N = N_g^{N_d}$ is the number of pixels and N_d is the dimension of the survey, so one can easily increase the number of pixels with only (almost) linear increase in computational time. If there is more than one data point assigned to a given pixel then we average over all the points by inverse noise weighting. If all the data have equal noise variance σ^2, then this amounts to simple averaging of the data, $d_{\boldsymbol{r}} = N_{\boldsymbol{r}}^{-1} \sum_{k=1}^{N_{\boldsymbol{r}}} d_k$, where the sum is over all $N_{\boldsymbol{r}}$ data points that contribute to the $\boldsymbol{r}$th grid point. The variance at that point is given by $N_{\boldsymbol{r}}^{-1}\sigma^2$, where σ^2 is the noise variance for each measurement. The data are now represented with the values $d_{\boldsymbol{r}}$ and the corresponding noise variance $\sigma_{\boldsymbol{r}}^2 = N_{\boldsymbol{r}}^{-1}\sigma^2$. We can define two new vectors, inverse noise weighted data $h_{\boldsymbol{r}} = d_{\boldsymbol{r}}/\sigma_{\boldsymbol{r}}^2$ and inverse noise $\sigma_{\boldsymbol{r}}^{-2}$. Note that if a certain grid point $\boldsymbol{r}$ has no galaxies contributing to it then the corresponding values $d_{\boldsymbol{r}}$, $h_{\boldsymbol{r}}$, and $\sigma_{\boldsymbol{r}}^{-2}$ are set to zero so that all the points on the grid have a well-defined value.

The response matrix $\boldsymbol{R}$ for the plane wave expansion is $R_{\boldsymbol{k},\boldsymbol{r}} = e^{i\boldsymbol{k}\cdot\boldsymbol{dr}}$, where $\boldsymbol{k}$ and $\boldsymbol{r}$ are the Fourier mode and real space position in the box, respectively, and can be parameterized with N_d-dimensional indices $\boldsymbol{l}$ and $\boldsymbol{x}$. The Fourier mode $\boldsymbol{k}$ is obtained from its index as $\boldsymbol{k} = 2\pi\boldsymbol{l}/D$, while the real space position is related to the index as $\boldsymbol{r} = \boldsymbol{x}D/N_g^{N_d}$, where D is the size of the grid. While each component of $\boldsymbol{x}$ runs between zero and $N_g - 1$, components of $\boldsymbol{l}$ run between $-N_g/2$ and $N_g/2$, with the two extreme values being equal (corresponding to the Nyquist frequency). If the data are nonscalar quantities then the above expression has to be generalized by adding a function $\chi(\boldsymbol{k})$ to the response matrix $\boldsymbol{R}$. For example, if the data are two shear components in the case of weak lensing, then $\chi(\boldsymbol{k}) = (\cos 2\phi_k, \sin 2\phi_k)$ for the two shear components, where ϕ_k is the phase of the Fourier mode $\boldsymbol{k}$. We will ignore this factor in the following since it can easily be added to the final expressions. Projecting the data to Fourier space as in the previous section gives

$$\boldsymbol{R}^{\dagger}\boldsymbol{N}^{-1}\boldsymbol{d} = \sum_i e^{i\boldsymbol{k}\cdot\boldsymbol{dr}_i}\frac{d(\boldsymbol{r}_i)}{\sigma^2(\boldsymbol{r}_i)} \approx \sum_{\boldsymbol{r}} e^{i\boldsymbol{k}\cdot\boldsymbol{dr}}\frac{d_{\boldsymbol{r}}}{\sigma_{\boldsymbol{r}}^2} = \sum_{\boldsymbol{r}} e^{i\boldsymbol{k}\cdot\boldsymbol{r}} h_{\boldsymbol{r}} = \tilde{h}_{\boldsymbol{k}}\ , \tag{21}$$

where we denoted with $\tilde{h}_{\boldsymbol{k}}$ the Fourier transform of $h_{\boldsymbol{r}}$, which can be computed using FFT. Note that the only approximation in the above expression was to replace the real positions of measured data with their interpolated position in the grid, and, as mentioned above, this can always be made more accurate by increasing the dimension of the

[4] Since we are sampling the field in a finite volume one can only determine the power spectrum convolved with the window of the data and one cannot resolve it better than $\Delta k \sim 2\pi/D$. If the power spectrum is smooth over Δk then this additional aliasing from nearby modes can be ignored.

grid. Since computing the Fourier transform with FFT involves $O(N \ln N)$ operations, the computational time is small even for very large ($\sim 10^8$) matrices. Of course, one can always perform the direct summation over all the individual data points, in which case the operation count will be $O(MN)$ with no approximation because of pixelization and may be more advantageous when M is small.

Next we need to compute the matrix $\boldsymbol{D}$ (eq. [18]). $\boldsymbol{S}^{-1}$ is diagonal in Fourier space, while noise transformation can be written as

$$(\boldsymbol{R}^{\dagger}\boldsymbol{N}^{-1}\boldsymbol{R})_{\boldsymbol{k}_1,\boldsymbol{k}_2} = \sum_i \frac{e^{i\boldsymbol{k}_1\cdot\delta\boldsymbol{r}_i}e^{-i\boldsymbol{k}_2\cdot\delta\boldsymbol{r}_i}}{\sigma^2(\boldsymbol{r}_i)} \approx \sum_{\boldsymbol{r}} e^{i(\boldsymbol{k}_1-\boldsymbol{k}_2)\cdot\boldsymbol{r}}\sigma_{\boldsymbol{k}}^{-2} = (\tilde{\sigma}^{-2})_{\boldsymbol{k}_1-\boldsymbol{k}_2}\,, \quad (22)$$

which is the Fourier transform of $\sigma_{\boldsymbol{r}}^{-2}$ evaluated at $\boldsymbol{k}_1 - \boldsymbol{k}_2$. This can again be obtained from FFT of the same grid as before, provided the initial grid was oversampled by a factor of 2 in each dimension, so that we are only interested in the modes that are less than one-half of the Nyquist frequency of the original grid. This part of the calculation is therefore also $O(N \ln N)$ as advocated in the previous section, instead of being $O(M^2N)$ [without pixelization to a grid and FFT it is $O(MN)$]. Finally, when required, we also need to compute the aliasing term in equation (18). If only the small-scale aliasing is important it can be obtained by first computing

$$(\boldsymbol{R}^{\dagger}\boldsymbol{N}^{-1}\boldsymbol{R}_b)_{\boldsymbol{k}_1,\boldsymbol{k}_2} = (\tilde{\sigma}^{-2})_{\boldsymbol{k}_1-\boldsymbol{k}_2}\,, \quad (23)$$

where $\boldsymbol{k}_2$ now corresponds to one of the modes we do not estimate. We can then subsequently perform the matrix multiplications in equation (18) to obtain the aliasing term. Of course, this requires that the dimension of the grid be sufficiently large to encompass all of the modes that alias power into the modes that we wish to estimate, so one should always check to see if the results change by changing the dimension of the grid. When this becomes too large one is better off computing the correlation matrix due to unwanted modes in real space analytically and then project it to Fourier space. Because the correlation matrix is not diagonal one cannot perform this with a single FFT but rather with N of these, and the operation count becomes $O(N^2 \ln N)$. Once this is done then everything has been projected to Fourier space and all the subsequent operations scale with M instead of N. In particular, inverting $\boldsymbol{D}$ and computing $\boldsymbol{b}$ and $\boldsymbol{F}$ are all $O(M^3)$ operations.

For very large M performing these matrix manipulations becomes computationally too expensive, so it is worth exploring approximations to the above expressions, which would allow one to compute them more rapidly. When the data are Poisson sampled one can simplify the analysis by noting that to weight the data one can use the mean number of measurements instead of the actual value, i.e., $\sigma_{\boldsymbol{r}}^2 = N_{\boldsymbol{r}}^{-1}\sigma^2$ is being replaced with $\sigma_{\boldsymbol{r}}^2 = \bar{N}_{\boldsymbol{r}}^{-1}\sigma^2$ and $d_{\boldsymbol{r}} = N_{\boldsymbol{r}}^{-1}\sum_{k=1}^{N_{\boldsymbol{r}}} d_k$ with $d_{\boldsymbol{r}} = \bar{N}_{\boldsymbol{r}}^{-1}\sum_{k=1}^{N_{\boldsymbol{r}}} d_k$, where $\bar{N}_{\boldsymbol{r}}$ is the mean density of measurements. We continue to keep the subscript $\boldsymbol{r}$ attached to this quantity, since it may vary across the survey (for example, because of varying selection function as a function of distance in galaxy surveys). If we assume a white noise power spectrum, $\Theta_l = \text{const}$ for all l, then one can solve the system exactly without any inversion. The reason is that the covariance matrix $\boldsymbol{C}$ becomes diagonal in real space, $C_{\boldsymbol{r}\boldsymbol{r}'} = (\sigma^2 + \sigma_s^2)\delta_{\boldsymbol{r}\boldsymbol{r}'}$, where $\sigma_s^2 = \Theta_l\bar{n}(\boldsymbol{r})$ is the theoretical variance, with $\bar{n}(\boldsymbol{r})$ being the mean density of measurements at location $\boldsymbol{r}$. This matrix can be inverted trivially in real space and $h_{\boldsymbol{r}} = (\boldsymbol{C}^{-1}\boldsymbol{d})_{\boldsymbol{r}} = d_{\boldsymbol{r}}/(\sigma^2 + \sigma_s^2)$ is a simple inverse weighting of the data, where the weight consists now of both noise and signal variance. If the survey is compact, then one can locally approximate the power spectrum as white noise and make $\boldsymbol{C}$ diagonal but vary theoretical variance σ_s^2 using the actual power spectrum Θ_l. This corresponds to the Feldman, Kaiser, & Peacock (1994) weighting, which is exact for the power spectrum estimation in the so-called classical limit, where the wavelength of interest is much shorter than the size of the survey (Hamilton 1997a; Tegmark et al. 1997a). The WF estimator becomes

$$\hat{s}_{\boldsymbol{k}} = \Theta_l\tilde{h}_{\boldsymbol{k}}\,, \quad (24)$$

with the variance of residuals

$$\langle r_{\boldsymbol{k}}\, r_{\boldsymbol{k}'}^{*}\rangle = \Theta_l[\delta_{\boldsymbol{k}\boldsymbol{k}'} - \Theta_l\tilde{\boldsymbol{C}}^{-1}(\boldsymbol{k}-\boldsymbol{k}')]\,. \quad (25)$$

The bias and Fisher matrix are given by

$$b_l = M_l\sum_i \frac{\sigma_{\boldsymbol{r}_i}^2}{(\sigma^2+\sigma_s^2)^2}$$

$$F_{ll'} = \sum_{\boldsymbol{k}}\sum_{\boldsymbol{k}'} |\,\tilde{\boldsymbol{C}}^{-1}(\boldsymbol{k}-\boldsymbol{k}')\,|^2\,. \quad (26)$$

Here $\boldsymbol{k}$ and $\boldsymbol{k}'$ are the wavevectors corresponding to parameters Θ_l and $\Theta_{l'}$, respectively, and $\tilde{\boldsymbol{C}}^{-1}$ is a Fourier transform of $(\sigma^2 + \sigma_s^2)^{-1}$. We see that a complete solution to the problem can be obtained without performing any matrix inversion and all the operations are $O(N \ln N)$, a huge advantage over the general case. Because of this we implicitly assumed that all the relevant modes are being estimated and ignored the aliasing term. If $\bar{n}$ is not varying with $\boldsymbol{r}$ then the weighting of the data is uniform and this gives a simple criterion for the signal-to-noise ratio of the modes given by $\bar{n}\Theta_l/\sigma$. One cannot take advantage of FFT to transform $\boldsymbol{C}^{-1}$ directly in this case because of k dependence in real space. Instead one can do a couple of FFTs varying σ_s^2 with k (for example, one FFT for each l using Θ_l to compute σ_s^2) and then combining them into a single matrix so that for a given pair $\boldsymbol{k}$, $\boldsymbol{k}'$ in $\tilde{\boldsymbol{C}}^{-1}$ one chooses Θ_l, which corresponds to $\boldsymbol{k}$ (Hamilton 1997a). The computational cost is still approximately $O(N \ln N)$, and the error is negligible provided that the Fisher matrix is nearly diagonal, which is the assumption that goes into the classical limit. This procedure has been applied to the power spectrum estimation in the case of galaxy surveys (Feldman et al. 1994), but one can apply it to other data sets and to the case of WF reconstruction as well, whenever the survey geometry is compact and the data sampling relatively uniform. One can further improve the reconstruction by performing the full analysis for the largest modes, where the classical approximation is the least reliable, and use the classical approximation on smaller scales.

6. DISCUSSION

In this paper we explore the close connection between optimal methods for image reconstruction and the minimum-variance power spectrum estimator. We show that for the case of Gaussian random fields both can be obtained within the same formalism, providing a unifying link between the two. This is because the first step, inverse variance weighting of the data, is common to both methods as one would expect: if a given measurement has a large

noise or is strongly correlated with the other measurements then it does not add new information and it should be downweighted by the inverse of the covariance matrix. Once this operation is performed then the power spectrum estimator simply averages over all the modes contributing to a given spectral bin, while WF filters the mode according to the signal-to-noise ratio. If the requirement of Gaussianity is dropped, then the two methods remain useful approximations. WF cannot claim to be optimal for all applications, although by definition it still minimizes the variance among all the linear functions of the data and therefore continues to be useful in many instances. Similarly, while there exists a minimum-variance power spectrum estimator (Hamilton 1997a), its expression is generally too complicated to compute and the Gaussian approximation remains useful. One has to be careful in this case to provide proper error estimates, which in general should be done either using second-order perturbation theory or using N-body Monte Carlo simulations. Alternatively, one can use a quick (and dirty) Monte Carlo estimate of error bars by bootstrapping the WF estimators s_l contributing to a given power spectrum bin l. This means drawing a set of WF estimates from the original estimated set with a replacement, so that some of the values will be drawn several times, while others will not be drawn at all. One then repeats the power spectrum estimation procedure for each drawn set, and the scatter in the estimators gives an estimate of the error. To the extent that the scatter in the power spectrum estimates is indicative of the error on their average this will give reasonable results. However, for the bootstrap method to be formally applicable one needs the data to be independent and identically distributed, and this is not necessarily the case because of noise, sparse sampling, and nonlinear evolution. Moreover, bootstrapping will underestimate the error where there are not enough independent realizations of a given mode amplitude. This is mainly a problem on large scales, where a Gaussian error estimate often suffices. Clearly this procedure needs to be carefully tested with N-body simulations before it can be recommended, but it should at least supersede the error estimates based on the variance in subsamples of the data (e.g., Lin et al. 1996).

There is another advantage in relating the two procedures that we attempted to stress in the paper. It is the possibility of translating various computational schemes devised for one method to the other. As an example, WF in the signal eigenmode space is often faster than in the space of real data, and while this has been explored in the past for WF analysis (Bunn et al. 1996) it can be applied in the same way also to the minimum-variance power spectrum estimator. Conversely, approximations such as the classical limit (Feldman et al. 1994) used for power spectrum analysis can be applied to WF as well. In general there is a distinctive advantage of having a unified approach to the two problems, since finding a fast solution to one problem will also lead to a fast solution to the other problem. Finally, being able to compute the power spectrum estimate also answers the question as to what power spectrum to use in WF in the absence of any prior information: one simply uses the minimum-variance power spectrum measured from the data themselves or one that is consistent with it (e.g., a smoothed version of this power spectrum). We hope this will provide a unified frame to many seemingly unrelated statistical approaches and will add a further incentive to those analyzing the data to use these powerful statistical methods on existing and future surveys.

I would like to thank Lam Hui and the participants at the Aspen Workshop on Large Scale Structure for useful conversations. I would also like to acknowledge the hospitality of the Aspen Center for Physics, where part of this work was completed.

APPENDIX A

EXTERNAL CONSTRAINTS

The minimum-variance power spectrum estimator and WF in the form we presented in the text have to be modified when the data are contaminated or do not contain information for some modes, when additional constraints are placed on the data, or when there are additional external parameters not included in the model. For example, in CMB data, the monopole and dipole are contaminated by mean temperature and by our motion with respect to the CMB frame, respectively. Because of incomplete coverage these modes contaminate higher modes as well and have to be removed from the data set. Similarly, in LSS galaxy surveys the mean density is unknown and obtained from the measured data themselves, so the $\boldsymbol{k} = 0$ mode mixes and contaminates the higher $\boldsymbol{k}$-modes as well if the modes are coupled. Other constraints that are not in the form of one of the modes can be placed on the data as well. For example, one may first want to remove a quadratic or cubic trend from the Lyα forest to account for the variable QSO continuum before reconstructing the density field or its power spectrum. Although in some instances we want to remove these components from the data without knowing their actual values, in other cases we may be interested also in their best reconstructed values. We also need to specify what do we mean by removing a certain component. One possibility is to project the data to a subspace orthogonal to these modes, so that all the unreliable information in these modes will be destroyed (Tegmark 1997). From a Bayesian point of view one wants instead to marginalize over the modes we are not interested in by integrating over their probability distribution. Although seemingly different, all these cases lead to the same solution, which can easily be incorporated into the formalism in the text. To show this let us first discuss the case of simultaneous reconstruction and external parameter determination. Following Rybicki & Press (1992) we can write the data vector in the presence of external parameters as $\boldsymbol{d} = \boldsymbol{Rs} + \boldsymbol{Lq} + \boldsymbol{n}$ (again we have implicitly put the aliasing modes into the noise term), where $\boldsymbol{L}$ is the $N \times M_q$ matrix with known coefficients and M_q is the number of external parameters. For example, when determining the mean $\boldsymbol{L} = (1, 1, \ldots, 1)$. If we assume a uniform prior for $\boldsymbol{q}$ [$P_q(\boldsymbol{q}) \propto$ const] then

$$P(\boldsymbol{q}, \boldsymbol{s}, | \boldsymbol{d}) \propto P_s(\boldsymbol{s}) P_n[\boldsymbol{d} - (\boldsymbol{Rs} + \boldsymbol{Lq})] \, . \tag{A1}$$

We can now simultaneously extremize the posterior probability for $\boldsymbol{q}$ and $\boldsymbol{s}$. This gives

$$\hat{\boldsymbol{q}} = (\boldsymbol{L}^\dagger \boldsymbol{C}^{-1} \boldsymbol{L})^{-1} \boldsymbol{L}^\dagger \boldsymbol{C}^{-1} \boldsymbol{d} ,$$
$$\hat{\boldsymbol{s}} = \boldsymbol{S}\boldsymbol{R}^\dagger \boldsymbol{C}^{-1} (\boldsymbol{d} - \boldsymbol{L}\hat{\boldsymbol{q}}) . \tag{A2}$$

If we wish to know the probability distribution of the data independent of external parameters and the underlying field we marginalize over these parameters by integrating $P_s(\boldsymbol{s}) P_n[\boldsymbol{d} - (\boldsymbol{R}\boldsymbol{s} + \boldsymbol{L}\boldsymbol{q})]$ over $d^{M_q}\boldsymbol{q} d^M \boldsymbol{s}$ as we did in equation (13). This gives (Rybicki & Press 1992; Zaroubi et al. 1995)

$$P(\boldsymbol{d}) \propto \exp \left\{ -\tfrac{1}{2} \boldsymbol{d}^\dagger [\boldsymbol{C}^{-1} - \boldsymbol{C}^{-1} \boldsymbol{L} (\boldsymbol{L}^\dagger \boldsymbol{C}^{-1} \boldsymbol{L})^{-1} \boldsymbol{L}^\dagger \boldsymbol{C}^{-1}] \boldsymbol{d} \right\} . \tag{A3}$$

Although we presented here the Bayesian derivation, the maximum likelihood method (in the frequentist sense) leads to exactly the same likelihood function for the data (Rybicki & Press 1992). To compute the power spectrum in the presence of external parameters we therefore need to replace $\boldsymbol{C}^{-1}$ in previous sections with $\boldsymbol{C}^{-1} - \boldsymbol{C}^{-1} \boldsymbol{L} (\boldsymbol{L}^\dagger \boldsymbol{C}^{-1} \boldsymbol{L})^{-1} \boldsymbol{L}^\dagger \boldsymbol{C}^{-1}$ and then proceed as before. Note that

$$[\boldsymbol{C}^{-1} - \boldsymbol{C}^{-1} \boldsymbol{L} (\boldsymbol{L}^\dagger \boldsymbol{C}^{-1} \boldsymbol{L})^{-1} \boldsymbol{L}^\dagger \boldsymbol{C}^{-1}] \boldsymbol{d} = \boldsymbol{C}^{-1} (\boldsymbol{d} - \boldsymbol{L}\hat{\boldsymbol{q}}) \equiv \boldsymbol{C}^{-1} \boldsymbol{\Pi} \boldsymbol{d} , \tag{A4}$$

which simply removes the contribution of best-estimated external parameters $\hat{\boldsymbol{q}}$ from the data. We defined the projection operator $\boldsymbol{\Pi}$, which has the property that $\Pi \boldsymbol{L} = 0$, and so we see that this is also equivalent to projecting the data to the subspace orthogonal to the unwanted modes or external parameters (Tegmark et al. 1997a). Subtracting the best-estimated external parameter, orthogonalizing the data, and marginalizing the data all lead to the same result, as advertised above.

Using the Woodbury formula one can further simplify the new correlation matrix (Rybicki & Press 1992):

$$\boldsymbol{C}^{-1} - \boldsymbol{C}^{-1} \boldsymbol{L} (\boldsymbol{L}^\dagger \boldsymbol{C}^{-1} \boldsymbol{L})^{-1} \boldsymbol{L}^\dagger \boldsymbol{C}^{-1} = \lim_{\sigma^2 \to \infty} (\boldsymbol{C} + \sigma^2 \boldsymbol{L}\boldsymbol{L}^\dagger)^{-1} , \tag{A5}$$

so instead of computing the expression on the left-hand side above one can add to $\boldsymbol{C}$ a term proportional to $\boldsymbol{L}\boldsymbol{L}^\dagger$ with a large variance (see also Bond et al. 1997). This is not surprising, since we assumed a uniform prior for $\boldsymbol{q}$ above, which is equivalent to having a large variance in the correlation matrix for this parameter. This is often computationally the simplest approach, especially when we are not really interested in the external parameters themselves. The discussion here is another example of the close connection between WF and the minimum-variance power spectrum estimator. Both give the same result when dealing with external parameters or other linear constraints and again a single solution solves both problems.

REFERENCES

Bond, J. R. 1995, Phys. Rev. Lett., 74, 4369
Bond, J. R., Knox, L., & Jaffe, A. H. 1997, astro-ph/9708203
Bunn, E. F., Hoffman, Y., & Silk, J. 1996, ApJ, 464, 1
Bunn, E. F., & Sugiyama, N. 1995, ApJ, 446, 49
Feldman, H. A., Kaiser, N., & Peacock, J. A. 1994, ApJ, 426, 23
Górski, K. M. 1994, ApJ, 430, L85
Hamilton, A. J. S. 1997a, MNRAS, 289, 285
Kaiser, N. 1992, ApJ, 388, 272
Kendall, M. G., & Stuart, A. 1969, The Advanced Theory of Statistics, Vol. 2 (London: Griffin)
Lin, H., et al. 1996, ApJ, 471, 617
Narayan, R., & Nityananda, R. 1986, ARA&A, 24, 127
Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. 1992, Numerical Recipes in Fortran (2d ed.; New York: Cambridge Univ. Press)
Rybicki, G., & Press, W. H. 1992, ApJ, 398, 169
Seljak, U. 1997a, ApJ, 482, 6
———. 1997b, ApJ, in press
Squires, G., & Kaiser, N. 1996, ApJ, 473, 65
Tegmark, M. 1997, Phys. Rev. D, 55, 5895
Tegmark, M., Hamilton, A., Strauss, M., Vogeley, M., & Szalay, A. 1997a, astro-ph/9708020
Tegmark, M., Taylor, A., & Heavens, A. 1997b, ApJ, 480, 22
Vogeley, M. S., & Szalay, A. S. 1996, ApJ, 465, 34
Zaroubi, S., Hoffman, Y., Fisher, K. B., & Lahav, O. 1995, ApJ, 449, 446