Spectral Estimation from Irregular Arrays

F. P. Bretherton
J. C. McWilliams
ABSTRACT

The problem of spectral estimation from data sampled at irregularly distributed measurement locations is considered from the point of view of an optimal fit of a spectral model when the amount of data available is large but finite. Several methods are shown to be asymptotically equivalent to the optimal one (as the amount of data becomes large), but several others, previously advocated, are shown to be either biased estimations or to have significantly larger estimation uncertainties than the optimal one. This problem is then reconsidered in a Bayesian formulation. The estimation problem is shown to be a combination of the prior prejudices of the investigator and the ability of the data to yield further information beyond the initial prejudices. In particular, the choice of the spectral model is best made in the context of this combination. The problem of an optimal array design is also considered: given a set of prior prejudices and a spectral model to be tested, what is the distribution of measurement points which yields the maximum amount of spectral information per measurement? An extensive set of optimal arrays is presented, with particular focus on red spectra (i.e., ones monotonically decreasing with wavenumber). Because the optimal estimation procedure can be computationally cumbersome for large arrays, an efficient approximation is developed for red spectral regimes. Finally, a comparison of our estimation procedure with other methods is made, in which it is shown that an analysis of bias, aliasing, and estimation uncertainty can be made in a unified framework for a large class of methods.
TABLE OF CONTENTS

Abstract
Table of Contents
List of Operators and Symbols

1. Introduction 1

2. Fitting Model Covariances 5
   2.1 Definitions 5
   2.2 Spectral Models 8
   2.3 Maximum Likelihood Estimate 9
   2.4 Generalized Least Squares 10
   2.5 An Iterative Approximation 14
   2.6 Asymptotic Efficiency 17
   2.7 An Example 20
   2.8 Consistency Tests 25

3. A Bayesian Formulation 30
   3.1 Bayesian Inference 30
   3.2 A priori Probabilities 30
   3.3 Examples 34
   3.4 Conditional Probabilities 36
   3.5 A posteriori Probabilities 36
   3.6 The Choice of Features 40
   3.7 Resolution and Aliasing in Short and Long Arrays 54
   3.8 Equally Spaced Observations 71
4. Optimal Array Design  
   4.1 Definitions  
   4.2 Remarks  
   4.3 A Standard Example  
   4.4 Realizations  
   4.5 Array Points  
   4.6 Spectral Shape  
   4.7 Features  
   4.8 Quasi-uniformly Spaced Arrays  

5. Higher Dimensions  
   5.1 Spectral Estimation  
   5.2 Isotropic Array Design  
   5.3 Anisotropic Array Design  

6. Computational Aspects of Large Arrays  
   6.1 Cholesky Factorization of the Covariance Matrix  
   6.2 Near Neighbor Array Truncation  
   6.3 Scale Dependent Array Redefinition  

7. Comparison with Other Methods  
   7.1 A priori Estimates  
   7.2 Data Adaptive Estimates  
   7.3 Model Fitting  
   7.4 A Unified Analysis of Methods  

8. Disclaimers
LIST OF OPERATORS AND SYMBOLS

The operators and symbols of this paper are either listed below or are only used within the section of the paper within which they are defined.

A. Operators

<table>
<thead>
<tr>
<th>Operators</th>
<th>Name</th>
<th>Where Introduced</th>
</tr>
</thead>
<tbody>
<tr>
<td>( )</td>
<td>ensemble mean</td>
<td>§2.1</td>
</tr>
<tr>
<td>( )^M</td>
<td>mean over available realizations</td>
<td>§2.1</td>
</tr>
<tr>
<td>( )^*</td>
<td>complex conjugate</td>
<td>§2.1</td>
</tr>
<tr>
<td>( )</td>
<td>a priori mean</td>
<td>(2.7)</td>
</tr>
<tr>
<td>( )^T</td>
<td>estimated value</td>
<td>§2.3</td>
</tr>
<tr>
<td>( )^-1</td>
<td>matrix transpose</td>
<td>§2.3</td>
</tr>
<tr>
<td>det( )</td>
<td>matrix determinant</td>
<td>§2.3</td>
</tr>
<tr>
<td>tr( )</td>
<td>matrix trace</td>
<td>§2.3</td>
</tr>
<tr>
<td>( )'</td>
<td>departure from ensemble mean</td>
<td>(2.15)</td>
</tr>
<tr>
<td>( )^ω</td>
<td>estimate obtained from weights</td>
<td>§2.4</td>
</tr>
<tr>
<td>( )^-ω</td>
<td>estimate obtained optimal weights</td>
<td>§2.4</td>
</tr>
<tr>
<td>( )^-(-)</td>
<td>estimate obtained optimal weights</td>
<td>§2.4</td>
</tr>
<tr>
<td>( )^-n</td>
<td>estimate from n^{th} iteration</td>
<td>§2.5</td>
</tr>
<tr>
<td>( )^ML</td>
<td>maximum likelihood estimate</td>
<td>§2.3</td>
</tr>
<tr>
<td>( )</td>
<td>a vector</td>
<td>§5.1</td>
</tr>
</tbody>
</table>
### B. Symbols

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Name</th>
<th>Where Introduced</th>
</tr>
</thead>
<tbody>
<tr>
<td>(C^P)</td>
<td>feature covariance function</td>
<td>§2.4</td>
</tr>
<tr>
<td>(C(\xi))</td>
<td>covariance function</td>
<td>(2.2)</td>
</tr>
<tr>
<td>(\Delta I)</td>
<td>information gain</td>
<td>(4.3)</td>
</tr>
<tr>
<td>(\delta(k))</td>
<td>Dirac delta function</td>
<td>(3.49)</td>
</tr>
<tr>
<td>(\delta_{pq})</td>
<td>Kronecker delta function</td>
<td>(2.25)</td>
</tr>
<tr>
<td>(\Delta k)</td>
<td>spectral resolution scale</td>
<td>§3.2</td>
</tr>
<tr>
<td>(\Delta s)</td>
<td>transformed spectral resolution scale</td>
<td>§3.3</td>
</tr>
<tr>
<td>(E(k))</td>
<td>biased spectral estimate</td>
<td>§7.1</td>
</tr>
<tr>
<td>(E'(k))</td>
<td>spectral density</td>
<td>(2.2)</td>
</tr>
<tr>
<td>(\varepsilon)</td>
<td>feature amplitude parameter</td>
<td>(3.10)</td>
</tr>
<tr>
<td>(F^P)</td>
<td>principal component eigenvector</td>
<td>(3.24)</td>
</tr>
<tr>
<td>(F^P(k))</td>
<td>spectral feature</td>
<td>(2.7)</td>
</tr>
<tr>
<td>(G^P)</td>
<td>orthonormal features constructed from (G^P)</td>
<td>(3.31)</td>
</tr>
<tr>
<td>(G^P_k)</td>
<td>a priori spectral feature</td>
<td>(3.5)</td>
</tr>
<tr>
<td>(H_{kl})</td>
<td>aliasing matrix</td>
<td>(3.18)</td>
</tr>
<tr>
<td>(H^P_{kl})</td>
<td>fundamental matrix</td>
<td>(7.14)</td>
</tr>
<tr>
<td>(I)</td>
<td>information</td>
<td>(4.2)</td>
</tr>
<tr>
<td>(k)</td>
<td>wavenumber</td>
<td>(2.2)</td>
</tr>
<tr>
<td>(K)</td>
<td>wavenumber vector magnitude</td>
<td>§5.1</td>
</tr>
<tr>
<td>(k_{u}, k_u)</td>
<td>limits of wavenumber range for spectral block features</td>
<td>(2.37)</td>
</tr>
<tr>
<td>(k_m)</td>
<td>mid-point wavenumber for a spectral block</td>
<td>(2.42)</td>
</tr>
<tr>
<td>(k_{\text{min}}, k_{\text{max}})</td>
<td>limits of wavenumber range spanned by features</td>
<td>(2.38)</td>
</tr>
<tr>
<td>Symbol</td>
<td>Name</td>
<td>Where Introduced</td>
</tr>
<tr>
<td>--------</td>
<td>----------------------------------------------------------------------</td>
<td>------------------</td>
</tr>
<tr>
<td>( \lambda_p )</td>
<td>spectral feature coefficient for ( F^p ) ( ) or ( G^p ) ( ) or ( F^p )</td>
<td>(2.7) or §3.6 or (3.24)</td>
</tr>
<tr>
<td>( M )</td>
<td>number of realizations</td>
<td>§2.1</td>
</tr>
<tr>
<td>( M )</td>
<td>effectively measurable vector subspace</td>
<td>§3.6</td>
</tr>
<tr>
<td>( \mu_j )</td>
<td>feature coefficient for ( f^j ) ( ) or ( G^j )</td>
<td>§2.7 or (3.6)</td>
</tr>
<tr>
<td>( N )</td>
<td>number of data points</td>
<td>§2.1</td>
</tr>
<tr>
<td>( N )</td>
<td>effectively aliased vector subspace</td>
<td>§3.6</td>
</tr>
<tr>
<td>( \nu_k )</td>
<td>relative spectral deviation</td>
<td>(2.51)</td>
</tr>
<tr>
<td>( P )</td>
<td>number of spectral features</td>
<td>(2.7)</td>
</tr>
<tr>
<td>( p_0 )</td>
<td>resolution integer</td>
<td>(3.10)</td>
</tr>
<tr>
<td>( \phi(x) )</td>
<td>random process</td>
<td>§2.1</td>
</tr>
<tr>
<td>( r_c )</td>
<td>critical eigenvalue</td>
<td>§3.6</td>
</tr>
<tr>
<td>( \rho(A</td>
<td>B) )</td>
<td>probability density for ( A ) given ( B )</td>
</tr>
<tr>
<td>( r_p )</td>
<td>principal component eigenvalue</td>
<td>(3.24)</td>
</tr>
<tr>
<td>( R_{pq} )</td>
<td>information matrix</td>
<td>(2.19)</td>
</tr>
<tr>
<td>( R_{st} )</td>
<td>transformation of ( H )</td>
<td>(7.26)</td>
</tr>
<tr>
<td>( s(k) )</td>
<td>attention function</td>
<td>(3.9)</td>
</tr>
<tr>
<td>( T_{kl} )</td>
<td>a priori covariance matrix</td>
<td>(3.2)</td>
</tr>
<tr>
<td>( W(A) )</td>
<td>functional to be minimized</td>
<td>(2.11)</td>
</tr>
<tr>
<td>( w(k) )</td>
<td>attention weighting function</td>
<td>(3.9)</td>
</tr>
<tr>
<td>( \hat{w}(k,\xi) )</td>
<td>window function</td>
<td>(7.3)</td>
</tr>
<tr>
<td>( x )</td>
<td>spatial coordinate</td>
<td>§2.1</td>
</tr>
<tr>
<td>( \xi )</td>
<td>lag distance</td>
<td>§2.1</td>
</tr>
<tr>
<td>( \Xi(\xi) )</td>
<td>lag space transformation</td>
<td>(4.12)</td>
</tr>
</tbody>
</table>
1. Introduction

Estimation of the power spectrum of a random process is of fundamental importance in many fields of science. For a band-limited continuous process, the standard procedure is to sample at equal intervals, Fourier transform, and analyze the distribution among frequency bands of sums of squares of Fourier amplitudes. When applicable, this method is computationally efficient and yields reliable results of which the limitations are well understood (Blackmon and Tukey, 1959; Cooley and Tukey, 1965). Frequently, however, particularly in geophysical disciplines, data are only available from samples at irregular intervals with large gaps. The basic theory of spectral estimation is then much less well developed and a plethora of different methods have been proposed to obtain an estimate, but few guidelines exist about the errors, sensitivity, resolution, and aliasing to be expected. These methods include simple interpolation followed by treatment as a continuous process (as, for example, has been done in most estimations of the atmospheric kinetic energy spectrum, e.g., Saltzman and Fleisher, 1962), consideration as an inverse problem (Backus and Gilbert, 1970), least squares estimates of individual Fourier components (Zetler et al., 1968), grouping data points by pairs of given separation and fitting a covariance function (Julian and Kline, 1974), maximum likelihood fits of model spectra (Jones, 1977), and many others.

This paper attempts to lay out a theoretical framework by which the merits of each method can be judged and the more general issues discussed. Attention is concentrated on the statistical uncertainties inherent in all spectral estimates from a given but arbitrary array
of sampling points, under circumstances when the amount of data available is finite but moderately large. The analysis is based upon the assumption that a reasonable first guess is available for the spectrum to be estimated, and a method is proposed by which refinements of, or corrections to, that first guess can reasonably be inferred from the data. Key concepts are the following ones:

(i) the objective fitting to the data of model covariance functions (or, equivalently, spectra) which are functions of a finite number of adjustable parameters,

(ii) approximate estimates of the minimum possible statistical uncertainties in the fitted parameters (i.e., "optimal" estimates), leading to definition of a class of "asymptotically efficient" fitting procedures,

(iii) specification of tests to determine whether the data are inconsistent with the complete family of model spectra under consideration,

(iv) a Bayesian formulation of the estimation problem that shows how the choice of the family of model spectra reflects a compromise between the preconceptions and interests of the investigator on the one hand and the ability of the observational array to yield information about such interests on the other hand,

(v) identification of the intrinsic alias of an array (i.e., those spectral features about which no information is available from the observations, no matter how many realizations are obtained) and the broader effective alias, which is dependent upon the amount of data available.
(vi) an algorithm for optimal array design which allows an investigator to determine the distribution of a given number of observation points that will yield the maximum information about a specified set of potential refinements of a first guess spectrum,

(vii) some approximations to the ideal spectral algorithm that reduce the amount of computing required to implement the recommended spectral estimation procedures, and

(viii) a common formulation within which most spectral estimation procedures can be compared and judged relative to an optimal procedure.

No apology is made for redefining the problem of "calculating the spectrum" in terms of estimating the best fit of an objective model. The authors believe that some form of model fitting, implicit or explicit, will be found to underlie every approach to calculating or interpreting spectra, and explicit formulation allows one to identify the criteria actually being used (see Section 7). The theory is illustrated most extensively for random processes in one dimension (space), but it is easily extended to two or more dimensions (see Section 5). Indeed, the concepts can even, in principle, be adapted to nonstationary processes. A serious restriction, however, is the assumption that the underlying process is Gaussian, specifically that the mean values of fourth order products of data values are related in a certain manner to the mean second order products. A major future task is to examine how robust the conclusions are if this assumption is replaced by alternate ones. Finally, it should be noted that, as in all inference
involving random processes, successful estimation requires a judicious blend between the assumptions which are brought to the analysis, the questions which are asked of it, and the data which can be adduced and organized to answer these questions. There is no unique solution, and usually more important than the final estimate itself is an understanding of the evidence and the limitations which are associated with it.
2. **Fitting model covariances**

2.1 Definitions

The approach will be illustrated for a one-dimensional, band-limited, homogeneous, Gaussian, random process \( \phi(x) \), with zero mean,

\[
\overline{\phi}(x) = 0, \tag{2.1}
\]

and non-negative power spectral density \( \overline{E}(k) \).\(^1\) The overbar (\(-\)) denotes the ensemble mean over a very large number of realizations, \( \phi(x) \), and is equivalent to averaging a single realization over a very large range of \( x \). "Gaussian" means that every finite subset \( \{\phi(x_r)\} \), \( r = 1, \ldots, N \), is joint normally distributed. The spectrum, or equivalently the covariance function

\[
\overline{C}(\xi) = \overline{\phi(x)\phi(x + \xi)} = \int_{0}^{\infty} \overline{E}(k) \cos k\xi \, dk \tag{2.2}
\]

then completely defines the process.

The available data are the values

\[
\phi_r = \phi(x_r) \quad r = 1, \ldots, N \tag{2.3}
\]

at a given array of data points \( \{x_r\} \), for a sample of \( M \) independent realizations of \( \phi(x) \). Mean values over this sample will be denoted by

\(1\)These restrictions are, to some extent, unnecessary though convenient.

The dimensionality is easily increased (see Section 5), the estimation and removal of the mean is a simpler and more accurately solvable problem than the present one (e.g., Bretherton, Davis, and Fandry, 1976), and inhomogeneous or non-Gaussian processes admit spectral estimation algorithms which are analogous to, though more complicated than, the ones given below.
The total number of data values $M \cdot N$ is finite but moderately large. For conceptual clarity, we will always consider the formal limit $M \to \infty$; that is, we take many repeated samples from the same array, so that

$$\text{Limit}_{M \to \infty} \frac{M}{(\quad)} = \left( \quad \right). \quad (2.4)$$

However, in practice, similar results come from considering only a single realization ($M = 1$), but extending the array over a large range of $x$ in a qualitatively repetitive manner ($N \to \infty$). The second alternative is difficult to treat concisely, though heuristic arguments indicate it should be equivalent to the first limiting process.

The sample mean covariance products

$$\overline{C}_{rs}^M = \frac{\phi_r \phi_s^M}{(\quad)} , \quad r,s = 1, \ldots N, \quad (2.5)$$

over each pair of array points are raw estimates of the true covariance matrix

$$\overline{C}_{rs} = \int_0^\infty \overline{E}(k) \cos k (x_r - x_s) \, dk. \quad (2.6)$$

If the number of realizations $M$ were infinitely large, the only problem would be the inversion of the Fourier transform (2.6). In practice, there are random sampling errors in $\overline{C}_{rs}^M$, and $\overline{C}$ must be estimated in a manner consistent with, but not uniquely determined by, $\overline{C}_{rs}^M$.

An immediate extension of this situation is to a process $\phi(x,t)$ sampled continuously in time at discrete observation points $x_r$, requiring estimation of the frequency-wavenumber spectrum. Given long
simultaneous records at each array point, Fourier transforming yields complex components \( a_\omega(x) \) for each frequency \( \omega \). The estimate of the wavenumber spectrum corresponding to some frequency \( \omega_0 \) then follows by replacing the sample mean covariances \( \overline{C}_{rs}^M \) by the cospectral estimates \( a_\omega(x_r) a_\omega^*(x_s)^M \), averaged over \( M \) degrees of freedom provided by \( M/2 \) discrete frequencies within a narrow band around \( \omega_0 \).
2.2 Spectral models

It will be supposed that, based upon previous experience, theoretical insight, or plain prejudice, the investigator brings to the analysis some idea of the spectrum to be expected, specifically a "first guess," \( \overline{E}(k) \). Attention is then focused on a family of possible spectra

\[
E(k;\lambda) = \overline{E}(k) \left( 1 + \sum_{p=1}^{\infty} \lambda_p F^p(k) \right),
\]

(2.7)

where each "feature" \( F^p(k) \) defines a question being asked of the analysis, and \( \lambda_p \) is an associated coefficient. The set \( \lambda \equiv \{ \lambda_p \} \) then defines a specific correction to the first guess and a corresponding positive definite covariance matrix

\[
C_{rs}(\lambda) = \int_0^\infty E(k;\lambda) \cos k|x_r - x_s| \, dk.
\]

(2.8)

Criteria for choosing suitable \( F^p(k) \) are discussed in Section 3. Meanwhile, it will be assumed that the family (2.7) is broad enough to include the actual spectrum \( \overline{E}(k) \); that is, for some \( \lambda \),

\[
\overline{E}(k) = E(k;\lambda).
\]

(2.9)

The use of (2.7) will also be confined to circumstances for which the linear sum \( \sum_p \lambda_p F^p(k) \) is fairly small compared to unity for all \( k \)--thus avoiding the obvious inconsistency of negative \( E(k) \). This will be so if \( \overline{E}(k) \) is a reasonably good approximation to \( \overline{E}(k) \), and if the amount and distribution of the data is adequate to determine \( \lambda \) to moderately high precision. If preferred, a wider class of nonlinear but non-negative spectral models \( E(k;\lambda) \) could be defined and fitted \textit{ab initio}. 
The arguments in this paper would then concentrate on refinements of this process once a crude fit has been obtained (so that a linearization in \( \lambda \) of the nonlinear model is possible) and on the intrinsic sensitivities and uncertainties in such refinements.

2.3 Maximum likelihood estimate

A variety of techniques may be used to find a set of values \( \hat{\lambda} \) of the coefficients \( \{\lambda_p\} \) that make the spectrum \( E(k;\hat{\lambda}) \) consistent with the data. These will be discussed in some detail before passing in Section 3 to the equally important question of how to select the family of model spectra under consideration. It will be found that some are essentially equivalent, but others are significantly inferior. One of the simplest is the maximum likelihood estimator proposed by Jones (1977).

To derive this, note that for a Gaussian random process with covariance function \( C(\xi;\lambda) \), the joint probability distribution \( \rho(\phi_1,\ldots,\phi_N) \) \( d\phi_1,\ldots,d\phi_N \) for the data values \( \{\phi_r\} \) at the array points \( \{x_r\} \) is:

\[
\rho(\phi|\lambda) d\phi = (2\pi)^{-1/2} N \left( \text{det } C \right)^{-1/2} e^{-1/2 \phi^T C^{-1} \phi} d\phi, \tag{2.10}
\]

where \( \phi^T \) is the transpose of the column vector \( \phi = \{\phi_r\} \), \( C \) is the matrix \( C_{rs}(\lambda) \), and \( \text{det } C \) and \( C^{-1} \) are respectively the determinant and inverse of \( C \). Although this expression can be conceived of as a function of \( \phi \), with unlikely data sets \( \{\phi_r\} \) being distinguished by large values of the exponent, it is also for large \( N \cdot M \) a sensitive function of \( \lambda \). Thus, for any given \( \{\phi_r\} \), we are led to the most likely estimate \( \lambda^{ML} \) as that which maximizes \( \rho(\phi|\lambda) \); i.e., for which the exponent in (2.10),

\[
W(\lambda) = 1/2 \left[ \phi^T C^{-1}(\lambda) \phi + \log \text{det } C(\lambda) \right], \tag{2.11}
\]
is a minimum. Remembering that

\[
\frac{\partial}{\partial \lambda} C^{-1} = -C^{-1} \frac{\partial C}{\partial \lambda} C^{-1}
\]

\[
\frac{\partial}{\partial \lambda} \log \det C = \text{tr} \left| C^{-1} \frac{\partial C}{\partial \lambda} \right|
\]

where \( \text{tr} \) denotes the trace of the matrix, this condition is

\[
\frac{\partial W}{\partial \lambda} = \text{tr} \left\{ \frac{\partial C}{\partial \lambda} \left( (C^{-1}_\phi)(C^{-1}_\phi)^T - C^{-1} \right) \right\} = 0 . \tag{2.12}
\]

Because of the matrix inverse \( C^{-1} \), equations (2.12) are inextricably nonlinear (even though for the model (2.7) \( \frac{\partial C}{\partial \lambda} \) is independent of \( \lambda \)), and their solution requires substantial computation. Jones has developed minimization techniques for \( W(\lambda) \) which are effective for a few hundred data points.

2.4 Generalized least squares

A different fitting technique is generalized least squares; i.e., choosing the value of \( \lambda \) that minimizes

\[
W(\lambda) = 1/2 \sum_{u,s} \left( \overline{C}^{-M}_{rs} - C_{rs}(\lambda) \right) \omega_{rs;uv} \left( \overline{C}^{-M}_{uv} - C_{uv}(\lambda) \right), \tag{2.13}
\]

where the matrix \( \overline{C}^{-M} \) describes the raw covariances (2.5), and the weights \( \omega_{rs;uv} \) are such that (2.13) is positive definite. This procedure involves solving the set of linear equations

\[
\sum_{u,s} \overline{C}^{P}_{rs} \omega_{rs;uv} \left( \overline{C}^{-M}_{uv} - \sum_{q} \lambda_q \overline{C}^{q}_{uv} \right) = 0 , \ p = 1, \ldots, P, \tag{2.14}
\]

to give the estimate \( \hat{\lambda}^{\omega} \). Here the matrices \( \overline{C}^{P} \) are the contributions to the covariance associated with the individual features \( F^P \) and are
independent of \( \lambda \) (i.e., from (2.2) and (2.3),

\[
C_{rs}^{P} = \int_{0}^{\infty} \hat{E}(k) F^{P}(k) \cos k|x_r - x_s|dk.
\]

Such an estimate \( \hat{\lambda}^{\omega} \) is a random variable, depending upon the
particular realizations sampled, with an estimation error

\[
\lambda^{\omega'} = \hat{\lambda}^{\omega}(\phi) - \bar{\lambda}.
\]  \hspace{1cm} (2.15)

For any given \( \omega_{rs;uv} \), equations (2.14) yield a unique \( \hat{\lambda}^{\omega} \), and
according to equation (2.9) the model family is consistent. It follows
immediately by commuting the ensemble mean with linear operations that
the estimate is unbiased; i.e.,

\[
\lambda^{\omega'} = 0.
\]  \hspace{1cm} (2.16)

However, the central result of this section is that for differing
\( \omega \), the least mean square estimation errors \( \lambda^{\omega^2} \) occur for the optimal
weights

\[
\omega_{rs;uv} = \frac{M}{2} \overline{(C)}^{-1}_{ru} \overline{(C)}^{-1}_{sv}.
\]  \hspace{1cm} (2.17)

The errors are then described by the \( n \times n \) covariance matrix

\[
\lambda_{pq} \lambda_{pq}' = \overline{(R)}^{-1}_{pq},
\]  \hspace{1cm} (2.18)

where

\[
\overline{R}_{pq} = \frac{M}{2} \sum_{u,v} c_{rs}^{P} \overline{(C)}^{-1}_{ru} \overline{(C)}^{-1}_{sv} c_{uv}^{q}
\]  \hspace{1cm} (2.19)

is the information matrix for the fitting process. Indeed, \( \hat{\lambda}^{\omega} \) for optimal
weights (which we shall denote by \( \hat{\lambda}^{(-)} \)) yields the minimum error covariances.
not only among the class of generalized least squares fit, but also among
the class of all unbiased estimators that depend linearly on the raw co-
variances (or rather their differences from first guess covariances) in the
form
\[ \lambda_p = \sum_{r,s} d_{p,rs} (C_{rs}^M - C_{rs}), \]  
(2.20)
where the \( d_{p,rs} \) are arbitrary coefficients.

To prove this result, the starting point is a calculation of the
sampling error covariances themselves. Since \( \phi \) is gaussian, all fourth
order cumulants vanish; i.e.,
\[ \phi r s u v = \phi r s u v + \phi r u s v + \phi r u v s + \phi r v s u . \]  
(2.21)
Thus, the mean square error in the sample covariances is
\[ U_{rs}; uv = \left( \frac{C_{rs}^M - C_{rs}}{C_{uv}^M - C_{uv}} \right) \left( C_{rs}^M - C_{rs} \right) \left( C_{uv}^M - C_{uv} \right) \]
\[ = \frac{1}{M} \frac{C_{ru} C_{sv} + C_{rv} C_{us}}{C_{ru} C_{su} + C_{rv} C_{sv}} . \]  
(2.22)
If we accept the convention that \( U \) will only be multiplied by expressions
which are symmetric within the index pairs \( (r, s) \) and \( (u, v) \), this matrix
may be inverted to give
\[ (U^{-1})_{rs}; uv = \frac{M}{4} \left[ \begin{array}{cccc}
-1 & -1 & -1 & -1 \\
(C_{ru})^{-1} & (C_{sv})^{-1} & (C_{rv})^{-1} & (C_{su})^{-1}
\end{array} \right], \]  
(2.23)
which is precisely equivalent to the optimal weights (2.17). Thus
\[ \bar{R}_{p}^{\text{pq}} = \sum_{r,s} C_{rs}^{P} (U^{-1})_{rs}; uv C_{uv}^{Q} . \]  
(2.24)
Now suppose that \( \lambda \) is given by expression (2.20), with the restrictions
that \( d_{p,rs} \) is symmetric in \( (r, s) \) and that the \( \lambda_p \) are unbiased. Then
\[ \bar{\lambda}_p = \frac{\lambda_p}{\gamma_p} = \sum_{r,s} d_{p,rs} (c_{rs}^M - c_{rs}^\gamma) \]

\[ = \sum_{r,s} d_{p,rs} (c_{rs}^\gamma - c_{rs}^\gamma) \]

\[ = \sum_{q \in \mathbb{P}} \alpha_q \left( \sum_{r,s} d_{p,rs} c_{rs}^q \right), \]

hence

\[ \sum_{r,s} d_{p,rs} c_{rs}^q = \delta_{pq} \quad p,q = 1, \ldots, \mathbb{P}. \] \hspace{1cm} (2.25)

Given any linear functional of the spectral parameters,

\[ \alpha = \sum_{p} \alpha_p \lambda_p, \] \hspace{1cm} (2.26)

a combination of equations (2.22) through (2.25) shows that the mean square estimation errors in \( \alpha \) will be

\[ \alpha^{-2} = \sum_{p,q} \alpha_p \alpha_q \left[ \sum_{r,s} d_{p,rs} d_{q,uv} U_{rs;uv} \right] \]

\[ = \sum_{p,q} \alpha_p \alpha_q \left[ \delta_{pq} \sum_{r,s} \left( d_{p,rs} - \sum_{p',r',s'} R^{-1}_{pp'} c_{p',r's'} U^{-1}_{r's';rs} U_{rs;uv} \right) \right. \]

\[ \left. \left( d_{q,uv} - \sum_{u',v'} U^{-1}_{uv;u'v'} c_{u'v'}^- U_{u'v';q} \right) \right] \]

which shows that \( \alpha^{-2} \) is indeed a minimum when

\[ d_{p,rs} = \sum_{q} \delta_{pq} \frac{1}{R} \left[ \frac{M}{2} \sum_{u,v} c_{uv} U_{ur}^{-1} U_{vs}^{-1} \right], \] \hspace{1cm} (2.27)

and the minimum is

\[ \sum_{p,q} \delta_{pq} \alpha_p \alpha_q. \] \hspace{1cm} (2.27) is identical to (2.14)

with the optimal weights (2.17).
This approach is, of course, standard in estimation theory, dating back to the work of Gauss in the 19th century. Before proceeding further, a few properties of the information matrix (2.19) should be noted. Since \( \bar{R} \) is proportional to the number of realizations \( M \), the parameter uncertainty \( \lambda' \) is \( O(1/M^{1/2}) \) as \( M \to \infty \). \( \bar{R} \) is intrinsically positive definite and dimensionless, though the magnitudes of its elements do depend on the normalizations chosen for the features \( \mathbf{P}^D \). If the array \( \{x_r\} \) can be divided into widely separated segments, so that correlations \( \bar{C}_{rs} \) between all points \( x_r, x_s \) in different segments are negligible, then the associated matrix elements \( \bar{C}_{rs}^{-1} \) are similarly separated, and the total information matrix \( \bar{R}_{pq} \) is the sum of contribution for each segment separately. Thus, it is to be expected that, if instead of many realizations \( M \) for a given array, there is only one realization but the array is extended \( M \) times in a qualitatively similar manner to contain \( MN \) data points, then \( \bar{R} \) and the optimal estimate \( \hat{\lambda}(-) \) will not be substantially changed. Finally, computation of \( \bar{R} \) and the optimal estimate \( \hat{\lambda}(-) \) requires prior knowledge of the true ensemble mean \( \bar{\lambda} \), which is the objective of the estimation. Thus, in practice the \( \hat{\lambda}(-) \) are inaccessible and approximations to the ideal must be used instead.

2.5 An iterative sequence

A good first approximation to the optimal estimate \( \hat{\lambda}(-) \) comes from using the first guess \( \bar{C} \) instead of \( \bar{C} \) in equation (2.17); that is, we minimize the expression

\[
\hat{W}(\lambda) = \frac{1}{2} \sum_{r,s,u,v} \left( C_{rs} - C_{rs}(\lambda) \right) \frac{M}{2} (C)_{ru}^{-1} (C)_{sv}^{-1} \left( C_{uv} - C_{uv}(\lambda) \right).
\] (2.28)
This yields an estimate \( \hat{\lambda}^{(1)}(\phi) \) which is unbiased but has random errors that are somewhat greater than the optimal. However, an improvement comes from using \( C^{(1)} = C(\hat{\lambda}^{(1)}) \) in (2.17); that is, we minimize

\[
W^{(1)}(\lambda) = \frac{1}{2} \sum_{r,s,u,v} \left( \frac{C_{rs} - C_{rs}(\lambda)}{M} \right) \frac{M}{2} \left( C^{(1)} \right)^{-1}_{ru} \left( C^{(1)} \right)^{-1}_{sv} \left( \frac{C_{uv} - C_{uv}(\lambda)}{M} \right).
\]

(2.29)

Note that the covariance features \( C^{P}_{rs} \) entering the definition of \( C_{rs}(\lambda) \) are unchanged [see (2.7) and (2.8)]; only the weights \( \omega_{rs;uv} \) are adjusted in the iteration. It will be shown shortly that, with probability nearly unity, the sequence \( \hat{\lambda}^{(1)}, \hat{\lambda}^{(2)}, \ldots \) converges rapidly to a limit \( \hat{\lambda}^{(\infty)}(\phi) \). The resulting estimator is non-linear, but does satisfy exactly the equations

\[
\text{tr} \left[ C^{P} (C^{(\infty)})^{-1} \left( \frac{C^{M}}{n+1} - C^{(\infty)} \right) (C^{(\infty)})^{-1} \right] = 0 \quad p = 1, \ldots, P. \tag{2.30}
\]

Comparison with equations (2.12) shows that \( \hat{\lambda}^{(\infty)} \) is identical to the maximum likelihood estimate \( \hat{\lambda}^{ML} \) for the data sample of \( M \) realizations.

The proof of convergence makes use of the assumption that \( M \) is large. Critical is the largest eigenvalue of the matrix

\[
J_{pq} \equiv \frac{\partial \hat{\lambda}^{(n+1)}}{\partial \hat{\lambda}(n)} q_p,
\]

(2.31)
evaluated for \( \hat{\lambda} \) close to \( \hat{\lambda}^{(\infty)} \). If this has magnitude less than unity, the deviations \( \hat{\lambda}^{(n)} - \hat{\lambda}^{(\infty)} \) decrease with each iteration, whereas if its magnitude is greater than unity, the sequence diverges, or at least oscillates finitely. The matrix \( J \) is a random variable dependent upon the data sample \( \phi \) under consideration. Its essential properties are
that its ensemble mean and variance are both $O(1/M)$ and, like all constructs obtained by averaging over a large sample of independent realizations, it is approximately joint normally distributed (n.b., the central limit theorem applies). Thus, typical elements of $J$ are $O(1/M^{1/2})$ and, as $M \to \infty$, the probability of the largest eigenvalue not being less than unity is exponentially small. Indeed, typically the convergence ratio, $\hat{\lambda}^{(n+1)}/\hat{\lambda}^{(n)}$, is $O(1/M^{1/2})$ and the sequence converges rapidly.

To see this, note that at each stage of the iteration

$$
\hat{\lambda}_p^{(n+1)} - \bar{\lambda}_p = \sum_{r,s} d^{(n)}_{p,rs} \left( C_{rs} - \bar{C}_{rs} \right),
$$

(2.32)

where $d^{(n)}$ depends on $\hat{\lambda}^{(n)}$, but always satisfies the constraints (2.25). The derivatives $\partial d^{(n)}/\partial \hat{\lambda}^{(n)}$ are well defined, and $d^{(0)}$ is, of course, determined solely by the first guess $\hat{\lambda}^{(0)} = 0$. Now $C_{rs} - \bar{C}_{rs}$ is $O(1/M^{1/2})$, so equations (2.32) can be regarded as generating an expansion of $\hat{\lambda}^{(n+1)} - \bar{\lambda}$ in powers of the random variable $C_{rs} - \bar{C}_{rs}$, which is valid asymptotically as $M \to \infty$. A similar expansion

$$
J = J_0 + J_1 \left( C_{rs}^M - \bar{C}_{rs} \right) + J_2 \left( C_{rs}^M - \bar{C}_{rs} \right) \left( C_{rs}^M - \bar{C}_{rs} \right) + \ldots
$$

exists for the matrix $J$. The key result is that the leading term $J_0$ of this expansion vanishes. Since the ensemble average of $C_{rs}^M - \bar{C}_{rs}$ is zero,

$$
\bar{J} \sim J_2 \left( C_{rs}^M - \bar{C}_{rs} \right)^2 = O(1/M),
$$

and

$$
\bar{J}^2 \sim J_1^2 \left( C_{rs}^M - \bar{C}_{rs} \right)^2 = O(1/M).
$$
The vanishing of $J_0$ follows from the remark that if the weights $w$ are held constant during the ensemble averaging, a generalized least squares fit is always unbiased, whatever the actual values of the weights; that is, the constraints (2.25) are valid independently of $\lambda$. Thus,

$$J_{pq} = \frac{3}{2\lambda(n)} \left[ \sum_{r,s} d_{p,rs}^{(n)} \left( \overline{C}_{rs}^M - \overline{C}_{rs} + \overline{C}_{rs} - \overline{C}_{rs} \right) \right]$$

$$= \sum_{r,s} \frac{3}{2\lambda(n)} d_{p,rs}^{(n)} \left( \overline{C}_{rs}^M - \overline{C}_{rs} \right)$$

$$+ \frac{3}{2\lambda(n)} \left[ \sum_{p} \overline{\lambda}_p \left( \sum_{r,s} d_{p,rs}^{(n)} C_{rs}^p \right) \right].$$

The first term is directly proportional to $\overline{C}_{rs}^M - \overline{C}_{rs}$, and the coefficient is non-singular in this difference. The second term vanishes by (2.25), which establishes the result that $J_0 = 0$.

2.6 Asymptotic efficiency

To evaluate different fitting procedures, a helpful concept is that of asymptotic efficiency, introduced by Dzhaparidze and Yaglom (1974). An estimator $\hat{\lambda}(\phi)$, based upon a data sample of $M$ independent realizations, is asymptotically efficient if the estimation error, $\lambda' = \hat{\lambda}(\phi) - \lambda$, which is a random variable of typical magnitude $O(1/M^{1/2})$, is such that

$$M^{1/2} \overline{\lambda'} \rightarrow 0$$

and

$$M \lambda' T \rightarrow \frac{1}{\overline{R}}$$

as $M \rightarrow \infty$ (where the superscript "T" denotes matrix transpose).

Condition (2.33) states that for sufficiently large $M$, any systematic bias in the estimator is much smaller than the random sampling error.

Condition (2.34) states that the covariance matrix for this vector
sampling error is negligibly larger than the optimal defined by equation (2.18). Of course, the minimum M for which these statements hold with useful accuracy will vary from case to case. However, a reasonable interpretation is that, under the circumstances in which we are interested, all asymptotically efficient estimators are effectively equivalent, the differences between them being negligible compared to the uncertainties in each.

Clearly the optimal linear estimator $\lambda^{(1)}$ defined in Section 2.4 is asymptotically efficient since both conditions are satisfied exactly. On the other hand, the generalized least squared estimator $\lambda^{(1)}$ based upon first guess weights in (2.28) satisfies the first condition but not the second. Although the sampling error covariance $\lambda^{(1)}_p, \lambda^{(1)}_q$ is $O(1/M)$, the constant of proportionality is, in general, larger than $MR^{-1}$. However, since $\lambda^{(1)} - \lambda$ is $O(1/M^{1/2})$, $\lambda^{(1)}$ differs from $\lambda$ only by $O(1/M^{1/2})$. Thus, the error covariances for the next iteration $\lambda^{(2)}$ differ from $\lambda$ only by a correction of relative magnitude $O(1/M^{1/2})$, and do satisfy the second condition. $\lambda^{(2)}(\phi)$ is not quite unbiased, but the ensemble average $\lambda^{(2)}$, is quadratic in $\lambda^{M} - \lambda$ and $\lambda^{(2)}$, is thus $O(1/M)$, satisfying the first condition. Thus, although the leading approximation $\lambda^{(1)}(\phi)$ of the iterative series is not asymptotically efficient, the second and all subsequent approximations are, including the maximum likelihood estimator, $\lambda^{ML} = \lambda^{(\infty)}$.

In practice, if the first guess $\hat{E}(k)$ is indeed a good approximation to the ensemble mean, the leading term $\lambda^{(1)}$ is often a quite adequate representation of the optimal estimate. In any case, little is gained by repeating the iteration beyond $\lambda^{(2)}$, or by invoking the full precision of the maximum likelihood estimator.
It should not be assumed from this discussion that all fitting procedures are equally valid. A comprehensive analysis would be very lengthy, but some further examples should be mentioned. A common approach to spectral estimation for data series with gaps is to interpolate smoothly between the observation points, resample at equal intervals, discrete Fourier transform, and fit a smooth curve to the raw spectral estimates that result. It should be noted that, if the interpolated curve is, on the whole, smoother than the original $\phi(x)$, these spectral estimates will be severely biased, with too little power at high wavenumbers and the resulting fit can be significantly misleading.

There is, however, a variant of the interpolation technique, requiring much more computation, that is asymptotically efficient. Indeed, it yields results precisely equivalent to the least squares estimators discussed in Section 2.5. To obtain the leading approximation, $\hat{\lambda}^{(1)}$, the value of $\phi$ between data points $x_r$ is interpolated according to the least mean square error estimate (Gandin, 1965; Bretherton, Davis and Fandry, 1976) using as a covariance function the first guess $C(\xi)$. The resulting continuous function of $x$ is then Fourier transformed and, averaging the square of the amplitude of the Fourier components over $M$ realizations, one obtains a biased estimate $E^M(k)$ of the true spectrum. Then, instead of fitting equation (2.7) to $E^M$ directly, biased estimates $\hat{E}^{(\gamma)}(k)$ and $\hat{E}^{(p)}(k)$ are computed for an infinite number of realizations of each of $P+1$ random process with corresponding spectra $\hat{E}(k)$ and $\hat{E}(k)P(k)$, respectively. A set of parameters $\lambda^*_q$ is then chosen to make
\[ \int_{F_{\text{p}}}^{P_{\text{q}}} \left[ \mathcal{N}_{\text{p}}(k) - E^{(\nu)}(q) - \sum_{q} \lambda^{*}_{q} \hat{E}(q)(k) \right] \, dk = 0 \quad (2.35) \]

for each of \( p = 1, \ldots, P \). It can readily be shown that the coefficients \( \lambda^{*}_{q} \) obtained by solving (2.35) are identical to the \( \hat{\lambda}^{(1)}_{q} \) obtained by minimizing \( \hat{W}(\lambda) \) in equation (2.28). This procedure may be iterated using the covariance function \( C^{(1)}(\xi) \) to yield \( \hat{\lambda}^{(2)} \). Note that the use of comparison estimates to correct for interpolation bias (at great computational expense) is essential for a satisfactory result.

In yet another technique, the first step is to group together all pairs of data points \((x_{r}, x_{s})\) for which the separation \( \xi = |x_{r} - x_{s}| \) lies in a suitable range \((\xi_{p-1}, \xi_{p})\). The raw covariances \( \hat{\phi}_{r-s}^{M} \) are then immediately averaged over the group to give a more stable raw estimate \( \hat{C}^{P} \) of the covariance function at some intermediate point, \( \xi^{*}_{p} \). This procedure is computationally simple and, except for some uncertainty in the choice of \( \xi^{*}_{p} \), yields an unbiased estimate for \( \overline{C}(\xi^{*}_{p}) \). Unfortunately, the sampling errors \( \hat{C}^{P} - \overline{C}(\xi^{*}_{p}) \) are correlated for different \( p \) and, in general, are significantly larger than an optimal procedure would allow. Even if further fitting of a model \( C(\xi, \lambda) \) to the \( \hat{C}^{P} \) is accomplished in an optimal unbiased manner, condition (2.33) cannot be satisfied and the final estimate is not asymptotically efficient. Further studies are appropriate to determine how large in practice is the loss of information associated with this grouping.

2.7 An example

As an illustration of the preceding methods, consider the irregular 20-point array \((N = 20)\) in Figure 1. We wish to estimate departures from a red, power law spectrum over more than two orders of magnitude in wave-number, with uniform attention to equal intervals in \( \log k \). The array
covers a range in x of 28.8 non-dimensional units with a minimum spacing of 0.16. Imagine that data are available at these points from 100 independent realizations (M = 100) of \( \phi(x) \), making a total of 2000 data points. One such random realization is also shown in Fig. 1, where the true spectrum is assumed to be the same as the first guess spectrum; namely,

\[
E(k) = \frac{1}{\nu + k^2},
\]

where \( \nu = .002 \). In the spectral model (2.7), there are 8 features (\( P = 8 \)), each of which selects a wavenumber band of width one octave (Fig. 2):

\[
P_p^p(k) = \begin{cases} 
.25 & k_{lp} \leq k \leq k_{up} \\
0 & \text{otherwise}
\end{cases}
\]

(2.37)

where the wavenumber band intervals are defined by

\[
\begin{align*}
k_{lp} &= k_{\text{min}} + \left( \frac{2^{p-1} - 1}{2^p - 1} \right) \left( k_{\text{max}} - k_{\text{min}} \right) \\
k_{up} &= k_{\text{min}} + \left( \frac{2^p - 1}{2^p - 1} \right) \left( k_{\text{max}} - k_{\text{min}} \right)
\end{align*}
\]

(2.38)

and the bandwidths by

\[
k_{up} - k_{lp} = 2^{p-1} \left( \frac{k_{\text{max}} - k_{\text{min}}}{2^p - 1} \right).
\]

In Fig. 2, the values \( k_{\text{min}} = .125 \) and \( k_{\text{max}} = 25.625 \) were chosen. The bandwidths for these features therefore vary by a factor of \( 2^7 = 128 \), and the ratio of wavenumbers spanned is approximately \( 2^8 = 256 \).

The information matrix \( \tilde{R} \) from (2.19) for these features is listed in Table 1(a). Note that the largest elements lie along the principal diagonal, but there are sizable off-diagonal elements as well. These off-diagonal terms imply a fundamental confusion or aliasing between
the spectral estimates in different wavenumber blocks. In this example, aliasing is not very serious. The precise measure of this, from (2.18), is the inverse matrix $\bar{R}^{-1}$, which is listed in Table 1(b). Since $\bar{R}^{-1}$ is approximately diagonal, the estimation errors are roughly independent, and their average magnitudes are proportional to the square root of the diagonal elements (see (2.18)).

Artificial data with a spectrum identical to that of the first guess (2.36) were generated in the computer by using a pseudo-random number generator. Figure 2 shows the spectra estimated by the first iteration of the least mean square error method $\lambda^{(1)}$ and by the maximum likelihood method $\lambda^{ML} = \lambda^{(\infty)}$. In the iteration from one to the other, the relative change in the $\lambda^{(n)}$ was on the order of 10% for each $n$, which is of the expected order of $M^{-1/2}$. In this logarithmic plot, the deviations in each wavenumber band from $E(k)$ (which for $k >> n^{1/2} = .04$ is a straight line with slope $-2$) are directly proportional to the coefficients $\lambda_p$. Note that the recovery of the true spectrum (2.36) from the estimation procedure is quite successful, and that the differences between $E(k; \lambda^{(1)})$ and $E(k; \lambda^{(\infty)})$ are small.

Given the scatter in the deviations from $\bar{E}(k)$, our interest often focuses on estimating any systematic trends. This requires fitting a smoother final curve to the estimated spectrum in Fig. 2; for example, we might fit the straight line

$$\log E(k) = m_1 + m_2 \log k$$  \hspace{1cm} (2.39)

for unknown coefficients $m_1$ and $m_2$. If the estimates in each wavenumber band were independent and of equal uncertainty (i.e., $\bar{R}^{-1}_{pq} = \delta_{pq}$), a simple least squares fit would suffice; indeed, a
very satisfying result could be estimated by eye in such a case. Because of correlations and non-uniformities in the errors, however, a more elaborate procedure is required.

The minimum uncertainty in the final outcome arises if the coefficients are estimated from a generalized least squares fit to the \( \hat{\lambda}_q \), using as weights the information matrix \( \bar{R}_{pq} \). We refer to this process as second stage fitting, which, if the features of the second stage lie in the space of the original features \( F^P(k) \), is precisely equivalent to initially fitting the second stage features as described in Sections 2.4–2.5. Consider a set of features \( f^j(k) \), \( j = 1, \ldots, J \), which are a linear combination of (2.37),

\[
f^j(k) = \sum_p A_{jp} F^P(k) .
\] (2.40)

We shall fit a model of the type (2.7) based upon the \( f^j \). To approximate the function (2.39), we define

\[
A_{jp} = \alpha_j \left[ \log \frac{k_{mp}}{k_{min}} \right]^{j-1}
\] (2.41)

and set \( J = 2 \). \( \alpha_j \) is a normalization such that \( \max_{(p)} A_{jp} = 1 \), and \( k_{mp} \) is a mid-point wavenumber in the blocks defined by (2.38):

\[
k_{mp} = k_{min} + \left( \frac{2^{p-1/2} - 1}{2^P - 1} \right) (k_{max} - k_{min}) .
\] (2.42)

If the \( \mu_j \) are now defined as the coefficients of the \( f^j \), then they are optimally estimated by minimizing

\[
W(\mu) \equiv \frac{1}{2} \sum_{p,q=1}^P \left[ \hat{\lambda}_p - \sum_{j=1}^J \mu_j A_{jp} \right] \bar{R}_{pq} \left[ \hat{\lambda}_q - \sum_{j=1}^J \mu_j' A_{jq} \right] .
\] (2.43)
The result is shown in Fig. 3, for $\mu_1$ and $\mu_2$ values of -.033 and 0.22 respectively (which were obtained from (2.43) and the $\Lambda^{(1)}$ of Fig. 2).

We can interpret the second coefficient as a best final estimate of the spectral slope as follows: for $\sum_j \mu_j \xi_j^j \ll 1$ and $\nu \ll k^2$,

$$\log E(k; \mu) = \log \tilde{E}(k) + \log \left(1 + \sum_j \mu_j \xi_j^j\right)$$

$$\approx .25 \left[ \mu_1 - \mu_2 \frac{\log k_{\text{min}}}{\log(k_{\text{max}}/k_{\text{min}})} \right]$$

$$- \left[ 2 - \mu_2 \frac{.25}{\log(k_{\text{max}}/k_{\text{min}})} \right] \log k$$

(2.44)

Thus, the true slope of -2 is estimated to be -1.976 from 100 data realizations.

The exact problem solved, (2.43), is not precisely equivalent to fitting the smooth model (2.39). The block features (2.37) are somewhat artificial in that, as a final result, a discontinuous spectrum like those shown in Figs. 2-3 would probably be unacceptable to most investigators. The second stage fitting of (2.43) is the best approximation to (2.39) which can be optimally fit consistent with the original spectral model (2.36)-(2.37). However, there is clearly little loss of precision in interpreting Fig. 3 by the smooth representation (2.44).

The principal advantage of the intermediate display (Fig. 2), involving more features than are finally contemplated, is to permit subjective judgment as well as objective tests of whether the data are indeed consistent with the assumed final model family.

The uncertainties in $\mu_1$ and $\mu_2$ are described by a 2 x 2 covariance matrix. It is the inverse of the second-stage information matrix,

$$R_{ij} = \sum_{p,q} A_{jp} \overline{R_{pq}} A_{jq}$$

(2.45)
which has as elements the coefficients of $\mu_1^2$, $\mu_1\mu_2$, $\mu_2\mu_1$, and $\mu_2^2$ in equation (2.43). The probability distribution for $\mu_1$ (regardless of the value of $\mu_2$) is normal, centered on $\mu_1 = -0.033$ with standard deviation $(.13)^{1/2} = 0.36$ (see Table 2b). The difference from the first guess $\mu_1 = 0$ is clearly not significant, which is consistent with the source of the data being analyzed (i.e., random realizations with the spectrum (2.36)). We can reach a similar conclusion about $\mu_2$.

2.8 Consistency tests

It was assumed in (2.9) that the model family being fitted to the data is consistent, that the true ensemble mean $\overline{E}(k)$ is a member of the family for some $\lambda$. There can be no way of demonstrating from the data that this is true; we can only make tests that demonstrate it is probably false. Such tests will now be described. For example, the final fit in (2.43) contained only two features, a correction to the magnitude of the first guess spectrum as a whole, and a correction to the power law. The coefficient of a third feature of the type (2.40)-(2.42), which is approximately

$$f^3(k) \sim \left[ \log \left( k/k_{\min} \right) \right]^2,$$

was assumed to be zero. Are the data inconsistent with such an assumption?

In constructing a test to answer such a question, it is important to distinguish the new aspects described by a feature $f^3(k)$ from those already permitted by $f^1$ and $f^2$. Any finite set of features $\{f^j(k)\}$ defines a $J$-dimensional subspace over the whole space of possible spectral variations, for which the set is a basis. The information
matrix $R$ generates a norm $\mu^T R \mu$ over the range of alternatives $\mu$ within this space, which is a measure of the ability of the observational array to narrow down the probable $\mu$'s. Given any sequence of features $f^1(k), f^2(k), \ldots$, there is a unique new sequence $g^1(k), g^2(k), \ldots$ that for each $J$ spans the same subspace, yet is orthonormal with respect to the norm. Set

$$g^i(k) = \sum_{j=1}^{J} \alpha_{ij} f^j(k)$$

(2.47)

and

$$v_i = \sum_{i=1}^{J} \beta_{ij} \mu_j \ ,$$

(2.48)

where $\alpha_{ij}$ is a lower triangular matrix, the inverse of the lower triangular factor $\beta_{ij}$ in the Cholesky decomposition of $R_{ij}$ (see Section 6). Thus,

$$\alpha_{ij} = \beta_{ij} = 0 \quad \text{if } j > i$$

and

$$\sum_{i',j'=1}^{J} \alpha_{i'j'} \alpha_{ij} = \delta_{ij} \ ,$$

so that for all $\mu$ and $J$

$$\sum_{j=1}^{J} \mu_j f^j(k) = \sum_{i=1}^{J} v_i g^i(k)$$

(2.49)

and

$$\mu^T R \mu = v^T v \ .$$

(2.50)

Equation (2.49) shows that the set $\{g^i\}$ does indeed provide a basis for each $J$-dimensional subspace generated by the $\{f^j\}$. Furthermore, according to (2.50), the estimation errors in the coefficients $v$ are uncorrelated with unit variance, so the feature $g^i(k)$ describes the new, independent possibilities for spectral variation introduced by $f^i(k)$, beyond those already permitted by $f^1, \ldots, f^{j-1}$. 
For the example in hand, the first three orthonormal features are shown in Fig. 4. The coefficients $\hat{V}_1$ estimated from the data $(E(k; \chi^{(1)})$ in Fig. 2) are .74, .41, and -.59, respectively. These should be compared with an estimation uncertainty (a standard deviation) of ±1, showing that none of these corrections are significantly different from zero. Thus, the restriction to the first two features in Section 2.7 was quite consistent with the observations.

It is also of interest to notice how the total estimation uncertainty $\varepsilon$ varies with the number of features retained in the final fit. We define a relative spectral deviation by

$$\nu(k) \equiv (E(k; \mu) - \bar{E}(k))/\bar{E}(k)$$

and $\varepsilon$ by

$$\varepsilon \equiv \frac{1}{k_{\text{max}} - k_{\text{min}}} \int_{k_{\text{min}}}^{k_{\text{max}}} \frac{1}{\sqrt{2}} \nu^2(k) \, dk.$$

Because of (2.50), we can calculate $\overline{\nu^2}$ as either

$$\overline{\nu^2}(k) = f^T \hat{R} f$$

or

$$\overline{\nu^2}(k) = g^T g$$

For our example $\varepsilon$ assumes the values .0012, .0030, .0049 for $J = 1, 2, 3$ respectively. By definition, $\varepsilon$ must monotonically increase with $J$. The fact that it increases nearly linearly indicates that all three parameters can be estimated with comparable skill. Alternatively, we can conclude that the experiment described in Section 2.7 can successfully estimate the model coefficients in the first three terms of an expansion in powers of log $k$. 
This test shows whether a quadratic expression rather than a linear one is really necessary in Fig. 3. There might be some other regard in which the adequacy of the spectral model (2.7) is in question. Such a contingency is readily tested in a similar manner—specify the feature $f(k)$ of interest, isolate the associated independent aspects $g(k)$, and compare the estimated coefficient with a normal distribution of zero mean and unit variance. However, one large caveat is in order. The basic assumption that the process $\phi(x)$ is Gaussian is, in practice, often not very good. The only place it has really been used is in equation (2.21), relating the covariance of the sampling errors of the raw covariances to the covariance matrix itself. If the fourth order cumulants do not vanish, the generalized least squares estimate is still unbiased, but the uncertainties in the estimate will not be correctly given by $\mathbf{R}$ and, in practice, they will often be enhanced. This means that the results of significance tests based on the Gaussian assumption should be treated with care, with due regard to any available information about the size of fourth order moments relative to second order ones.

Finally, it is worth noting that, in principle, the same approach can be used to test whether $\phi(x)$ is indeed homogeneous in space. If the likely mode of inhomogeneity can be specified, e.g., all variances increasing linearly with $x$, a corresponding feature can be added to the model covariance matrix and an estimated coefficient determined. If zero lies outside the likely range of uncertainty of this estimate, the data indicate that $\phi(x)$ is probably not homogeneous. A sequence of such tests can cover the most important issues, though it must
always be remembered that the probability of a spurious non-null result from at least one of a group of such tests is greater than that from any individual one, so an indefinitely long list will always produce some apparently significant inconsistencies.
3. Bayesian Formulation

3.1 Bayesian inference

Following the previous discussion of techniques for model fitting, this section addresses the more general questions of criteria for the choice of particular model families, the intrinsic limitations of observational arrays, and some fundamental philosophy of the inferential process. It is helpful to formulate the problem in Bayesian terms, not because any investigator precisely follows this approach, but because it does illuminate the nature of the analysis and the subjective decisions inherent in it. Bayesian inference requires a statement of 'a priori' probabilities for the various hypotheses about the system under consideration, and known conditional probabilities for the outcome of certain observations given each one of these hypotheses. Examination of the actual outcome of such observations then permits a deduction of 'a posteriori' probabilities over the hypotheses. These conclusions reflect influences both of the information contained in this data and the assumptions or experience implied in the 'a priori' statement.

3.2 A priori probabilities

The most difficult step in this process is often to find a satisfactory formulation of the hypotheses. A quantitative statement necessarily requires great restrictions upon the range of possibilities, yet it must not be so arbitrary as to be probably inconsistent with the evidence. In the spirit of refining knowledge of a statistical process
about which much has already been determined, it will be supposed that
the process under consideration is indeed stationary, band-limited, and
Gaussian with zero mean, and has a spectrum \( E(k) \) which is close to the
first guess \( \hat{E}(k) \). The most general specification of the relative difference

\[
\nu(k) = \frac{E(k) - \hat{E}(k)}{E(k)}
\]

requires consideration of functions \( \nu(k) \) that are vectors in an infinite
dimensional space. However, since we will be concerned only with \( \nu(k) \)
that are in some sense smooth, we will avoid unnecessary technicalities
by considering only the values at a finite set of wavenumbers within
the band limit which are integer multiples of some constant \( \Delta k \). The
resolution \( \Delta k \) is chosen so small that this discrete representation is
not a practical restriction on the variations in \( E(k) \). The following
is a convenient notation: symbols \( k \) and \( l \) used as arguments, as in
\( E(k) \), will continue to denote the dimensional wavenumber. Used as
suffices, as in \( \nu_k \), they will denote the integers \( k/\Delta k, l/\Delta k \). Also,

\[
E_k = E(k)\Delta k
\]

will be the total variance associated with the bandwidth \( \Delta k \) centered
about the dimensional wavenumber \( k \).

A complete set of a priori probabilities now comes from defining
a probability distribution \( \rho(\nu) \) over the vector \( \nu \). Attention will be
concentrated on the statement

\[
\rho(\nu)d\nu = (2\pi)^{-Q/2}(\det T)^{-1/2} \rho^{-1/2} \frac{\nu^T T^{-1} \nu}{\nu_1 \cdots \nu_Q},
\]

(3.2)
where \( T \) is a positive definite \( Q \times Q \) matrix with inverse \( T^{-1} \) and determinant \( \det T \). Equation (3.2) states that the a priori most probable spectrum is \( \tilde{E}(k) \) (because the probability density peaks at \( \nu = 0 \)), but the relative deviations at wavenumber \( k \), the \( \nu_k \) from (3.1), are joint normally distributed, having covariance
\[
\begin{pmatrix} \nu_k \\ \nu_l \end{pmatrix} = T_{kl} \quad .
\]

(3.3)

We will hereafter use the tilde \( \sim \) to denote averages over the prior distribution (3.2). Consistent with our concentration upon small refinements to the first guess, we suppose that
\[
\overline{(\nu_k^2)} = T_{kk} << 1 \quad .
\]

(3.4)

This condition ensures that the probability of negative \( E(k) \) is indeed negligible. Now any positive definite \( T_{kl} \) can always be factored as
\[
T_{kl} = \sum_p G^p_k G^p_l \quad ,
\]

(3.5)

and the easiest way to describe it is to specify a set of functions \( G^p(k) \), \( p = 1,2,... \) which are the independent modes of variation of \( \nu(k) \). This specification is equivalent to provisionally selecting a model family
\[
E(k; \mu) = \hat{E}(k) \left\{ 1 + \sum_p \mu_p G^p(k) \right\} \quad ,
\]

(3.6)

where the \( \mu_p \) are a priori independently normally distributed, with zero mean and unit variance. Then
\[ v_k = \sum_p \mu_p \mathbf{G}_p^k \quad \text{(3.7)} \]

generates a set of \( v \)'s with probability distribution (3.2). Although formally unlimited, the sequence of features satisfies

\[ T_{kk} = \sum_p (G_p^k)^2 \ll 1 \quad \text{(3.8)} \]

and hence it is necessary that \( G_p^k \to 0 \) rapidly as \( p \to \infty \). If the \( G_p^k \) are smooth functions of \( k \), equation (3.7) is always well defined (i.e., the rank of the matrix \( G_p^k \) is equivalent to that of the matrix formed by adjoining the extra column \( v_k \) to the right of \( G_p^k \)--see Franklin, 1968, p. 481), though its inverse expressing \( \mu \) in terms of \( v \) may not be. Thus, for arbitrary \( v \), the exponent \( v^T \mathbf{T}^{-1} v \) will normally be very large, implying that such a spectrum is unlikely, unless \( v(k) \) is relatively lacking in the details which are expressed only in \( G_p^k \) for large \( p \). The convergence condition (3.8) requires, in effect, that possible deviations \( v(k) \) be ranked in priority, those whose spectral shape is deemed likely or interesting by the investigator appearing as a \( G_p^k \) with moderately large amplitude, those shapes deemed unlikely being accorded a very small amplitude a priori.
3.3 Examples

As an example, suppose that

\[ s(k) = \int_{0}^{k} w(k') dk' \]  

(3.9)
is a mapping of the complete wavenumber domain into the internal 0 ≤ s ≤ 1. s(k) hereafter will be called the attention function. The positive weight function w(k) measures the relative importance attached by the investigator to high resolution at wavenumber k. A suitable set of a priori features would be

\[ G_k^p = \varepsilon \sqrt{\frac{4}{\pi p_0^2}} e^{-\frac{1}{2} \left( \frac{p-1}{p_0} \right)^2} \cos (\pi(p-1)s(k)). \]  

(3.10)

Here, \( p_0 \) is a large integer and \( \varepsilon \) is a small parameter. The wavenumber resolution associated with each feature increases systematically with \( p \), but the reduction in amplitude when \( p \gg p_0 \) implies that fine structure over wavenumber bands narrower than \( 1/(\pi p_0 w(k)) \) is relatively unlikely. Approximating the sum over \( p \) in (3.5) by an integral,

\[ T_{kl} \approx \varepsilon^2 e^{-\left( \frac{\pi p_0}{2} w(k)(k-1) \right)^2}, \]  

(3.11)

for \( k, l \) large compared to \( \Delta k \). Thus, deviations \( \psi(k) \) separated in wavenumber by more than \( 2/(\pi p_0 w(k)) \) are essentially independent. So \( p_0 \) controls the resolution permitted in \( \psi_k \), weighted in wavenumber according to \( w(k) \). On the other hand, \( \varepsilon \) controls the local standard deviation, uniform relative to \( \psi(k) \) across the whole wavenumber range. Thus, this formulation of the a priori possibilities is an approximation
to the intuitive statement "complete freedom, with as much resolution as possible consistent with reasonable accuracy."

A different set of features with many of the same properties is

\[ G^p(k) = \varepsilon, \text{ if } \frac{P-1}{p_0} < s(k) \leq \frac{P}{p_0} \text{ and } p \leq p_0, \]  
\[ = 0, \text{ otherwise,} \]  

(3.12)

\[ w(k) = \begin{cases} 
D^{-1} \left[ 1 + \frac{k - k_{\text{min}}}{D} \right]^{-1} \left[ \ln \left( 1 + \frac{k_{\text{max}} - k_{\text{min}}}{D} \right) \right]^{-1} & \text{for } k_{\text{min}} \leq k \leq k_{\text{max}} \\
0, & \text{otherwise,} 
\end{cases} \]  
\[ \text{(3.13)} \]

where \( D = (k_{\text{max}} - k_{\text{min}})/(2^P - 1) \) is the size of the first wavenumber block in (2.37)-(2.38). Equation (3.12) is especially convenient for computation, though it leads to the somewhat artificial result of discontinuous spectra, as shown in Fig. 2. For both specifications (3.10) and (3.12), the near diagonality of \( T_{kl} \) implies that, for distributions \( \nu_k = \nu(k) \) that vary only slightly over a resolution bandwidth (as defined in (3.12)),

\[ \sum_{k,l} \nu_k T_{kl}^{-1} \nu_l \approx \frac{1}{\varepsilon^2} p_0 \int w(k) \nu^2(k)dk \]
\[ \approx \frac{1}{\varepsilon^2} \sum_s (\nu(k_s))^2 \]  
\[ \text{(3.14)} \]

where the sum in (3.14) is for \( s = 1/2p_0, 3/2p_0, \ldots, 1-1/2p_0 \). These expressions are not, however, valid for \( \nu(k) \) that have substantial
structure near the resolution limit. For such \( v(k) \), the right-hand side of (3.14) is a serious underestimate. Indeed, for the second specification (3.12), any \( v(k) \) that is not essentially uniform across a block has zero probability of occurrence, with \( v^T T^{-1} v \) being infinite as the formal expression of this fact through (3.2).

3.4 Conditional probabilities

For the Gaussian process \( \phi \) defined by a particular value \( v \), the conditional probability of observations at \( \{ x_r \} \) yielding the values \( \{ \phi_r \} \) in a single realization follows at once from the remark that the \( \{ \phi_r \} \) are joint normally distributed with zero mean. This probability is

\[
\rho(\phi|v) \, d\phi = (2\pi)^{-N/2} \, (\det C(v))^{-1/2} \, e^{-1/2 \phi^T C^{-1}(v) \phi} \, d\phi_1 \cdots d\phi_N ,
\]

where \( C(v) \) is the \( N \times N \) covariance matrix \( \sum_{r,s} \phi_r \phi_s = C_{rs}(v) \). Accordingly, the probability of occurrence of a compound data set \( \phi \) arising from a sequence of \( M \) independent realizations is

\[
\rho(\phi|v) \, d\phi = (2\pi)^{-MN/2} \, e^{-1/2 \, M \{ \text{tr}(C^{-1} C^M) \} + \log \det C} \, d\phi_1 \cdots d\phi_{MN} .
\]

This result has already been used in the discussion of the maximum likelihood estimator (equation (2.10)).

3.5 A posteriori probabilities

Bayes' theorem asserts that, for a given set of a priori probabilities \( \rho(v) \) and a given data set \( \phi \), then the a posteriori probability density for \( v \) is
\[ \rho(\mathcal{V} | \phi) = \frac{\rho(\mathcal{V} | \phi) \rho(\mathcal{V})}{\rho(\phi)} \]
\[ = a e^{-1/2 M \left\{ \text{tr}(C^{-1} \overline{C} M) + \log \det C \right\}} e^{-1/2 \mathbf{v}^T T^{-1} \mathbf{v}} \]

where \( a \) is a normalization constant such that the integral over all \( \lambda \) is unity (see, for example, Jeffreys, 1967). Equation (3.17) gives the probability that the data \( \phi \) come from a process with statistics given by any particular value of \( \mathcal{V} \). Hence we can determine from (3.17) not only an estimate of the central tendency \( \hat{\mathcal{V}} \), but also measures of the uncertainty in that estimate.

A full discussion requires some further simplification of equation (3.15), in which the first exponent is replaced approximately by a positive quadratic function of \( \mathcal{V} \), centered around a minimum that depends on \( \phi \), but with curvature effectively uniform over the range of most interest. Then this curvature may be calculated at \( \mathcal{V} = 0 \) (i.e., by using the first guess spectrum \( \hat{\mathcal{V}}(k) \)). Such an approximation to the first exponent was given in equation (2.28). In the representation defined by equation (3.1), the exponent in (3.15) (and the function minimized in Section 2.5 becomes

\[ \hat{W} (\mathcal{V}) = \frac{1}{2} M (\mathcal{V} - \hat{\mathcal{V}}(1))^T H (\mathcal{V} - \hat{\mathcal{V}}(1)), \]

where

\[ H_{kl} = \sum_{r,s,u,v} \frac{1}{2} \hat{E}_k \cos k(x_r - x_s) (\hat{\mathcal{C}}^{-1})_{ru} (\hat{\mathcal{C}}^{-1})_{sv} \cos 1(x_u - x_v) \hat{E}_1 \]

and

\[ \sum_{1}^{n} H_{kl} \hat{V}_{(1)} = \sum_{r,s,u,v} \frac{1}{2} \hat{E}_k \cos k(x_r - x_s) (\hat{\mathcal{C}}^{-1})_{ru} (\hat{\mathcal{C}}^{-1})_{sv} \overline{C}^{-M} \]

(3.20)
Inserting (3.18) into equation (3.17) we obtain

\[ \rho(\nu|\phi) = \alpha \, e^{-1/2 \, M(\nu - \nu^*)^T \, H(\nu - \nu^*) \, e^{-1/2} \, \nu^T \nu} \]

\[ = \alpha_* e^{-1/2(\nu - \nu^*)^T \, (MH + T^{-1})(\nu - \nu^*)} \]

(3.21)

(3.22)

where

\[ \nu^* = (MH + T^{-1})^{-1} \, MH \nu^{(1)} \]

(3.23)

and \( \alpha_* \) is an alternative normalization serving the same purpose as \( \alpha \).

Equations (3.21)-(3.23) are central to the remaining discussion of this section. We see that the likely values of \( \nu \) are clustered around \( \nu^* \), constrained by two distinct effects. The first exponent in equation (3.21) is directly due to the data and would lead to \( \nu^{(1)} \), the unbiased linear estimator that closely approximates the maximum likelihood estimate defined by equation (2.12). The second exponent in equation (3.21) is independent of observations and is entirely due to the prior preference for \( \nu = 0 \). If the measurements dominate \( (MH \gg T^{-1}) \), \( \nu^* \sim \nu^{(1)} \), and if the information provided by the data is slight, then \( \nu^* \) is heavily biased towards zero. The total uncertainty surrounding \( \nu^* \) is determined by \( MH + T^{-1} \) and reflects the more stringent of the two constraints.

Together they restrict the important values of \( \nu \) to be small, so replacing \( \frac{C}{C} \) by \( \frac{C}{C} \) and ignoring cubic and higher powers of \( \nu \) in (3.15) is justified in retrospect.

The dimensionless matrix \( H \) in (3.19) expresses the fundamental properties of the observational array, defining the resolving power and aliasing inherent in spectral estimates involving small refinements to \( \tilde{E}(k) \). It is identical to the information matrix \( R \) for the features (3.1), and depends on the choice of features only through the grain size \( \Delta k \),
being otherwise a continuous function of the suffices $k,l$. The elements $H_{kl}$ are positive, and the matrix as a whole is non-negative. However, since $\gamma^{-1}$ is finite and positive definite, it follows from (3.19) that the quadratic form $\nu^T H \nu$ vanishes if and only if

$$\sum_k \nu_k E_k \cos k (x_r - x_s) = 0 \text{ for all } r,s ;$$

(3.24)

i.e., the null space of $H$ is the set of spectral features for which the feature covariance vanishes between all pairs of observation points. This null space is the intrinsic alias of the array, within which the observations can yield no discriminations however large the number of realizations $M$. Within this subspace the final estimate $\nu^*$ is determined entirely by the a priori probabilities. In practice, particularly for large arrays, sampling uncertainties preclude discrimination in the larger vector-subspace for which $\nu^T H \nu$ is merely small (compared to $M^{-1} \nu^T T^{-1} \nu$), defining the effective alias. The latter depends on the choice of $T$, and hence on the preferences of the investigator; the former does not. Note that the estimate $\hat{\nu}^{(1)}$ defined by equation (3.20) is only used in the form $H \nu^{(1)}$ in (3.23). Thus the nonexistence of the inverse $H^{-1}$ causes no problems.
3.6 The choice of features

Equations (3.21)-(3.23) provide a rational basis for discussion of the choice of families of spectral models for fitting to particular data sets. If the number of parameters, $P$, in such a family is too large, or features are selected for which the array has little skill, the final estimate will have uncertainties that are intolerable. If $P$ is too small, or inappropriate features were selected, the family may be inadequate to encompass the true spectrum $\bar{E}(k)$, and will be demonstrably inconsistent with the data. It should be recognized that, even in the framework of Bayesian inference, there is no unique or optimal choice of model. The choice is a compromise between the initial preferences or prejudices of the investigator and the ability of the observational array to supply information on those preferences. A careful analysis, preferably prior to any data actually being taken, will normally result in modification of either the preferences of the investigator (T) or the array design (H), or both. This process is analogous to the setting of budgets for a research organization. The perceived needs usually exceed the resources available. Priorities must be determined, resulting usually in refinement downwards of the plans and perceived requirements, and sometimes in renewed efforts to obtain more resources, iterating until a match results. Only after the judgments which result in a compromise are complete would the real task (the analysis of the data set itself) actually begin. There is nothing fundamentally unsound about modifying the statement of prior probabilities, provided there is no feedback from the particular sample of observations to be used in the analysis.
The authors have identified three different approaches to this compromise, reflecting different emphases on the desiderata of precision, resolution, minimum aliasing, and stability of the estimate. High precision means that the relative uncertainty $\nu'(k)^2$ in the final estimate should be small, measured either by a global upper bound or by some weighted average over wavenumber. For an unbiased estimator, precision implies accuracy. High resolution means that the analysis is able to distinguish effectively between slightly different wavenumbers; it is indicated locally by the minimum $\Delta k$ at which the correlation coefficient between uncertainties $\nu'(k + \Delta k)$ and $\nu'(k)$ falls below some threshold (0.5, say). In general terms, aliasing means inability to distinguish power between widely separated wavenumbers. A precise definition of the intrinsic alias of an irregular array was given in Section 3.5. Characterization of what leads to effective aliasing (also defined in Section 3.5) is difficult and will be discussed later (Section 3.7). Stability of the estimate means that when minor details of the a priori probabilities are changed, the major features of the estimate $\hat{v}(k)$ should be relatively unaffected.

The approaches to compromise will be illustrated using the example of Section 2.7, for which they yield rather similar results. The starting point will be a priori probabilities defined by equations (3.5) and (3.10) or (3.12), with the resolution parameter $p_o$ initially substantially larger than will eventually be needed, so that equation (3.14) is an adequate approximation. The key input from the investigator is a tentative attention function, with the selection of the parameters $\varepsilon^2$ and $p_o$ being delayed until later in the analysis. Note
that an appropriate attention function is often suggested by the
abscissa of the anticipated graphical display for the final spectrum.

The first method, that of principal components, concentrates upon
the eigenvalues $r_p$ and eigenvectors $F^p(k)$ of the matrix equation

$$MHF = rT^{-1}F.$$  \hspace{1cm} (3.25)

The eigenvalues are stationary values of the ratio

$$r = \frac{\nu^T MH\nu}{\nu^T T^{-1} \nu}$$  \hspace{1cm} (3.26)

between the constraints imposed by observation and those imposed a priori.

When normalized according to

$$F^p (MH + T^{-1}) F^q = \delta_{pq},$$  \hspace{1cm} (3.27)

the spectral model,

$$E(k;\lambda) = \sum_k \left[ 1 + \sum_p \lambda_p F^p \right],$$

has the property that each coefficient $\lambda_p$ is independent (both a priori
and a posteriori) and is determined after the measurements to within a
sampling error of standard deviation unity. If the corresponding eigen-
value is large ($r_p >> 1$), the value of $\lambda_p$ is determined by the data, and
the most probable a posteriori estimate,

$$\hat{\lambda}^*_p = \frac{r_p}{1 + r_p} \hat{\lambda}^ {ML}_p,$$

is close to the unbiased maximum likelihood estimate $\hat{\lambda}^ {ML}_p$. If, on the
other hand, $r_p << 1$, $\hat{\lambda}^*_p$ is only a small fraction of $\hat{\lambda}^ {ML}_p$ and is strongly
biased towards the a priori estimate of zero. Thus, given an array
matrix $H$ and prior covariances $T$, the eigensolutions $F_k^p$ are the natural features for the problem.

The solution of the system (3.25) involves finding the eigenvalues of the $P \times P$ information matrix (2.19), here evaluated as

$$ R_{pq} = M \sum_{k,l} G_{pq}^k H_{kl} G_{ql}^l, \quad (3.28) $$

where $G_{pq}^k$ are the factors of the prior covariance matrix $T$ (see equation (3.5)). These eigenvalues are well defined and the eigensolutions of (3.27) are linear combinations of the $G_{pq}^k$, even if the inverse $T^{-1}$ does not exist. Thus, the prior covariances indeed delimit the range of possibilities from which the observations select further.

The authors have found from many examples that the principal components are usually well ordered in the sense that, when arranged in order of decreasing $r_p$, successive eigenfunctions are either peaked in new wavenumber bands or add increasing resolution in bands already covered. After the investigator chooses a critical value $r_c$ greater than unity (in a manner to be described later), the vector subspace $M$ spanned by those eigenfunctions $r_p \geq r_c$ defines a class of "effectively measurable" features. The dimension $P$ of this subspace is the best single measure of the resolving power of the analysis. The orthogonal complement subspace $N$, spanned by those eigenfunctions for which $r_p < r_c$, is the effective alias (and includes the intrinsic alias for which $r_p = 0$). Any vector $\nu(k)$ can be uniquely decomposed into the sum of a part in $M$ and a part in $N$, each of which can be determined independently, both a priori and a posteriori. Once $r_c$ and hence the boundaries of $M$ have been determined, it is a useful practice to
truncate the set of features under consideration to some basis set over \( M \). This is formally equivalent to redefining \( T \) so that it is very improbable that any vector in the effective alias should be an amplitude other than zero, thus minimizing the appearance in the final estimate of details that are not firmly implied by the observations and keeping the estimator unbiased (n.b., equation (3.23)). The residual uncertainty in the spectral estimate is then

\[
\int_0^1 v'(k)^2 \, ds = \frac{\varepsilon^2}{p_0} \sum_{l=1}^{P} \frac{1}{1 + r_p} ,
\]

(3.29)

which, for suitable choices of \( \varepsilon, p_0, \) and \( r_c \), will be as small as required. Equation (3.29) can be derived by differencing (3.25) and (3.27), applying the approximations implicit in (3.14), and noting that \( v_k = \sum_p \lambda_p e_p^k \) and that \( \lambda_p^k \lambda_q^p = \delta_{pq} \).

In the second approach to the negotiation process, attention is concentrated upon an ordered infinite sequence of raw features \( G_k^p \) of known shape but arbitrary amplitude (because any amplitude factor will cancel in the treatment below). For example, these features might be

\[
G_k^p = \alpha_p \frac{\varepsilon}{\sqrt{p_0}} \cos \{\pi(p-1)s(k)\} , \quad p = 1, 2, 3, \ldots ,
\]

(3.30)

where each successive feature introduces additional resolution and 

\[
\alpha_p = \sqrt{2} - \delta_{p1}(\sqrt{2} - 1). \quad \text{The coefficient in (3.30) is introduced only for easy comparison with other approaches (e.g., see (3.31) below).}
\]

From this set of features is constructed an orthonormal set \( \{G_k^p\} \), each member being independent under the observations with coefficient \( \lambda_p \) determined to within \( \lambda_p^2 = 1 \). Thus, for each \( P \), \( G_k^p \) is a linear combination of \( \{G_k^q\}, \quad q = 1, \ldots, P, \) and
Thus, $G^P$ is that part of $G^P$ which can be determined from the observations independently of its predecessors, scaled so that the sampling uncertainty in its coefficient is unity. For each $P$ the total residual relative variance,

$$\bar{v}_k^2 = \sum_{p=1}^{P} (G^P_k)^2,$$

is computed for each $k$, and the process is terminated in $P$ when (3.32) first exceeds an acceptable limit. For comparison with the first approach, note that, for the raw features defined in equation (3.30),

$$\int_0^1 \frac{1}{\sqrt{v_k^2}} ds = \frac{\varepsilon^2}{p_o} \text{tr} \{R^{-1}\},$$

where $R^{-1}$ is the inverse of the $P \times P$ information matrix (3.28).

Equation (3.33) follows directly from an explicit integration of (3.32) and the normalization (3.31).

The normalization adopted in this second (and the following third) approach is relative to the observations alone (i.e., one based upon the probability distribution defined by the matrix $MH$ rather than $MH + T^{-1}$). This is because the prior probability distribution is in fact adjusted in these analyses so as to exclude all features which lie in the effective alias and thus assure stability in the sense defined above. Furthermore, by the criteria of asymptotic efficiency (Section 2.6), once the set of features is restricted to be finite (as in the second and third approaches), the estimation procedures based upon the two normalizations are equivalent.
This second approach emphasizes above all else the resolution of
the spectral estimate, which is distributed over wavenumber precisely
according to the preconceived weighting function \( w(k) = ds/dk \) specified
by the investigator. Aliasing is considered only indirectly, in that,
if some linear combination of the raw features lies in the intrinsic
alias of \( H \), the inverse \( R^{-1} \) will not exist, and the variance (3.33)
will be infinitely large, so that the sequence (3.30) will be terminated
before the inclusion of a raw feature that permits description of any
\( \nu(k) \) about which the observational array yields no information. A
similar remark applies to the effective alias. The truncation procedure
is equivalent to redefining the a priori probabilities to be

\[
T_{kl} = \frac{1}{\gamma} \sum_{p=1}^{P} G_k^P G_l^P
\]

(3.34)

instead of (3.5). Note that in this second approach the a priori
probabilities have not previously been stated in obtaining equations
(3.30)-(3.33). When one defines them by equation (3.34), one obtains
principal component eigenfunctions \( \Phi_k^P = (1 + \gamma)^{-1/2} G_k^P \) with eigenvalues
\( r_p = \gamma^{-1} \) for \( p \leq P \) and zero \( r_p \) for \( p > P \). Obviously, \( \gamma \ll 1 \).

In the third approach no attempt is made explicitly to derive
independent features. Rather the attention function \( s(k) \) is adjusted
so that each diagonal element \( R_{pp} \) of the information matrix associated
with block features of width \( \Delta s \) and amplitude \( \xi \) (i.e., equation (3.12))
has the same value \( \Lambda \). If it turns out that the resulting matrix \( R_{pq} \)
is approximately diagonal, then the associated spectral block amplitudes
are indeed approximately independently estimated, and the spectral
parameters \( \mu_p \) are determined to within standard deviation \( \Lambda^{-1/2} \).
Since, for features spanning all wavenumbers at which $E(k) \neq 0$,
$$\sum_{p,q} R_{pq} = \varepsilon^2 MN/2, \Lambda = \varepsilon^2 MN/2P.$$ The preceding properties allow a clear interpretation of the results, essentially comparable to the usual graphical presentation of spectral estimates from equally-spaced arrays, except that the abscissa is not $k$ but $s(k)$, a function which is determined in the analysis.

If, on the other hand, the $H$ matrix implies major aliasing involving widely separated wavenumber bands, such a procedure will leave substantial off-diagonal terms in $R$ and the interpretation is more complex. However, if the aliasing is only moderate, further adjustments in $s(k)$ aimed at minimizing

$$\text{tr}(R^{-1}) \approx \sum_{p=1}^{P} \frac{1}{R_{pp}} \left\{ 1 + \sum_{q \neq p} \frac{R^2_{pq}}{R_{pp} R_{qq}} + \ldots \right\}, \quad (3.35)$$

will select the distribution of resolution with wavenumber that yields the minimum weighted average variance for $P$ features (n.b., for block features (3.12), the average variance formula (3.33) holds generally because of (2.18)).

This adjustment process is best done approximately, by initially calculating $R_{pq}$ for block features of width $\Delta$s far smaller than ultimately required, and then combining together adjacent blocks in appropriate ways. If two blocks are combined, with indices $p_1$ and $p_2$, the resulting matrix elements are $R_{p_1 q} + R_{p_2 q}$ and $R_{p_1 p} + R_{p_2 p}$ for the off-diagonal terms in column $q$ or row $p$ respectively and $R_{p_1 p_1} + 2R_{p_1 p_2} + R_{p_2 p_2}$ for the diagonal term. Thus the corresponding columns and rows of the original $R$ are simply added, in a manner which conserves "R-substance" for an array while changing the effective resolution of the analysis.
The ratio $R_{pq}^2/R_{pp}R_{qq}$ which appears in (3.35) is a direct measure of the degree to which features $p$ and $q$ can be determined independently. It is necessarily less than or equal to unity, approaching unity when the ability to discriminate between them is much less than the ability to measure almost any combination of them. Since "R-substance" is intrinsically positive for block features, sizable matrix elements far off-diagonal cannot be eliminated, though ones near-diagonal will disappear through block combinations. The presence of the latter is a consequence of the resolution available from the observational array actually being less than that sought. The former have not been a significant problem in the examples studied so far, once suitable attention has been paid to obvious aliases such as seeking to estimate $E(k)$ for $k$ much larger than $\pi/\Delta x$, where $\Delta x$ is the minimum array spacing. In the absence of off-diagonal terms, the sum of the diagonal elements $\text{tr}(R)$ is fixed, and the sum of their inverses $\text{tr}(R^{-1})$ has a minimum when all elements are equal.

These three approaches have been applied to the example of Section 2.7. The initial attention function is

$$s = \begin{cases} 
0 & k \leq .125 \\
.18 \ln(10k - .25) & .125 \leq k \leq 25.625 \\
1 & k \geq 25.625 
\end{cases} \quad (3.36)$$

which is consistent with (2.38) and (3.13). The logarithmic nature of $s$ is suggested by the investigator's interest in departures of the spectrum from a power law, and the wavenumber endpoints are approximately related to the maximum and minimum spacings in the array in
Fig. 1 (this relationship is made explicit in Section 4). The a priori features are defined by

\[ G_k^p = 0.25 \quad (p = 1, \ldots, 8) \quad , \tag{3.37} \]

in each of 8 non-overlapping blocks of width \( \Delta s = 1/8 \), with \( G^p \) vanishing outside of the \( p \)th block.

In the first approach, the eigenfunctions (3.25) are linear combinations of the \( G^p \)'s:

\[ \phi_k^p = \sum_q G_k^p A_{qp} \quad , \tag{3.38} \]

where

\[ \sum_q R_{pq}^t A_{qp} = r_p A_{pp} \quad , \tag{3.39} \]

and

\[ \sum_p A_{qp} A_{qp} = \delta_{pp} \frac{1 + r_p}{1 + r_p} \quad . \tag{3.40} \]

The eight eigensolutions \( A_{qp} \) are plotted in Fig. 5 against the index \( q \). Fig. 5 can also be interpreted as plots of \( F_k^p \) against \( \log k \), where the abscissa \( q \) represents \( \log k \), and the \( A_{qp} \) values should be connected as in a bar graph, rather than a line graph (and further multiplied by \( \epsilon = 0.25 \)). Thus, the \( F_k^p \) are discontinuous, as are the \( G_k^p \), with no fine structure within each block of width one octave in \( k \). The largest eigenvalues (smallest \( p \)) are associated with relatively simple features. The number of zero crossings in \( F_k^p \) vs. \( k \) increases with the index \( p \).

Imagine now that an investigator sought to estimate the amplitude of a feature \( g_k \) with finer wavenumber resolution than can be represented in the basis set of block features \( G_k^p \) \( (p = 1, \ldots, P) \) from which the prior
probability matrix $T_{kl}$ has been constructed in (3.5). For example, $g_k$ might be alternately positive and negative across one of the wave-number blocks in (3.12). In this case, $g$ is orthogonal to all of the other basis features,

$$\sum_k G^p_k g_k = 0$$

for all $p$, and the ratio

$$\frac{T^{-1} g}{T g}$$

is infinite, which is an expression of the a priori impossibility of the spectral feature represented by $g$. Such a feature thus lies in the intrinsic alias, hence in the effective alias for all values of $r_c$.

Inspection of Fig. 5 shows that the first four eigenfunctions provide reasonable coverage over blocks 3-8, with resolution in this range further enhanced by eigenfunctions 5 and 6. However, blocks 1 and 2 are poorly covered, and the ability to discriminate between them comes only from the eighth eigenfunction, which is strongly peaked in block 1. Since the local variance in the final spectral estimate is given by

$$\bar{\nu}_k^2 = \sum_{p=1}^P \left( \frac{P}{k} \right)^2 = (.25)^2 \sum_{p=1}^P (A_{q_kp})^2$$

(3.41)

where $q_k$ is the label for the block within which $k$ lies, the inclusion of the last eigen function implies a heavy penalty in terms of precision. Indeed, since $r = 0.7 < 1$, the coefficient of this feature in the spectral estimate is determined more by the a priori assumptions than
the data. The measurable subspace includes eigenfunctions 1-7 (note
the list of eigenvalues in the caption of Fig. 5), and eigenfunction 8
lies within the effective alias. The attention function might perhaps
be modified to exclude block 1; however, this modification is accom-
plished automatically, because none of the basis functions for the
measurable subspace has significant amplitude in block 1. The wave-
number averaged variance is given by equation (3.29) with, here,
\( \varepsilon = .25 \) and \( p_0 = 8 \):
\[
\sqrt{\int_0^1 ds \, v'^2} = 0.10 ,
\]
(3.42)
which shows that the final spectral estimated based upon 100 realizations
should be accurate to ±10%. Inclusion of the eighth eigenfunction only
somewhat increases this figure to 0.12, because the variance in the first
block is limited by the a priori assumptions (n.b., the unit added to
\( r_p \) in (3.29)). The rejection of this eighth eigenfunction is therefore
primarily to retain stability of the estimate in the following sense.
Because \( \sigma_p ^2 \propto \varepsilon \), \( r_{pq} \propto \varepsilon ^2 \) and \( r_p \propto \varepsilon ^2 \). When \( r_p \gg 1 \), the direct contri-
bution from the a priori parameter \( \varepsilon \) thus cancels from the estimate of
average variance. When \( r_p \ll 1 \), however, the a priori influence domi-
nates, and both the estimate \( \hat{\nu} \) and its variance \( \nu'^2 \) will be strongly
dependent upon the value of \( \varepsilon \) selected. Thus, in this example, the
estimate for block 1 is not stable.

An analysis analogous to that above can be performed over an
extended wavenumber range. We define \( T_{kl} \) by equation (3.5), \( C^p_k \) by
equation (3.12) (except with \( p_0 = 12 \)), and the weighting function \( w(k) \)
is non-zero over a broader interval. In equation (3.13), we formally
replace $k_{\text{max}}$ by a value extended by four octaves from that given following (2.38) (i.e., $k_{\text{max}} = 409.625$) while retaining the previous values of $k_{\text{min}}$ and $D$ (i.e., 0.125 and 0.1, respectively). There is now a larger set of eigenvalues of equation (3.24), and the eigensolutions extend over a broader wavenumber range. However, the larger $r_p$ are little altered (the largest is increased by only 8%), and the corresponding eigensolutions have a similar shape over the previous wavenumber range (see Fig. 6 for the first principal component matrix column $A_{ql}$; $q = 1, \ldots, p_0$). The $r_p$ values for $p = 9-12$ are very much smaller than those for smaller $p$ (e.g., $r_4 = 8.4$, $r_5 = 4.2$, $r_9 = 10^{-2}$, and $r_{12} = 10^{-5}$), and their associated $F^p_k$ or $A_{pq}$ have their maximum amplitudes for $k$ or $q$ values in the extended wavenumber range. This is an expression of very slight estimation skill for wavenumber bands significantly higher than the resolution scale of the array (n.b., from Fig. 1, we can estimate the resolution scale as $\pi/\Delta x_{\text{min}} = 19.2$, which is in the block corresponding to $p = 8$ [see Fig. 2]). Another measure of the reduced estimation skill in the extended range is provided by the normalized information matrix based upon (3.28),

$$R'_{pq} = \frac{R_{pq}}{\sqrt{R_{pp} R_{qq}}}.$$  \hspace{1cm} (3.35)

Equation (2.18) states that estimation errors in the block coefficients $\lambda^p$ have covariances $R^{-1}_{pq}$; thus, the proximity to unity of off-diagonal elements in $R'_{pq}$ indicates a degree of intrinsic indistinguishability (i.e., aliasing) of spectral amplitudes in wavenumber bands $p$ and $q$. The matrix $R'_{pq}$ for the $p_0 = 12$ extended range analysis is listed in
Table 3. Notice that for $p,q < 8$, the aliasing is rather small, indicating that the array of Fig. 1 is reasonably successful at distinguishing these wavenumber bands (except perhaps the first and second), while for $p$ or $q \geq 9$, the aliasing is nearly total. In the face of such small $r_p$ and large aliasing, a prudent investigator would, for this particular array, revise his prior statement of the attention function to reduce the wavenumber range of interest.

In the second approach we adopt the form (3.30) for the a priori sequence of features. The first four orthonormal features derived from these are shown for the example of Section 2.7 in Fig. 7a, and the residual variance $\nu_k^2$ when the series is truncated at $P = 4$ and $P = 6$ is shown in Fig. 7b. A small amplitude for a normalized feature means that it is well measured, and the relative ineffectiveness of the array at wavenumbers between .1 and .2 is reflected in the local peak in $\nu_k^2$. In this approach the compromise consists of truncating at the desired level of precision, then modifying the attention function so that $ds/dk = w$ is reduced in the offending region, and finally re-computing. Reducing $ds/dk$ locally to zero would be equivalent to eliminating block 1 (and practically the same as eliminating the eighth eigenfunction) in the previous treatment. Multiplying $ds/dk$ by a factor proportional to the inverse of the variance in Fig. 7b should provide a better fit, the goal being to have $\nu_k^2$ uniform across the interval under consideration, with $P$ as large as possible for an acceptable upper bound on $\nu_k^2$.

The third approach leads to the wavenumber block widths shown in Fig. 8, with the corresponding information matrix listed in Table 4. Note that, in this case, a substantial measure of independence between
blocks is in fact achieved (i.e., $R_{pq}$ is roughly diagonal), which encourages a graphical presentation of the best fit as in Fig. 8, where the lines $E(k;\lambda)=\hat{E}(k)\sqrt{\frac{1}{k}}$ are also plotted for each block. Since increasing a block width increases $R_{pp}$, and tends to decrease the local estimation variance $(R_{pp}^{-1})$, there is the usual trade-off between precision and resolution long recognized in the conventional analysis of equally-spaced arrays. The additional refinement of precisely minimizing (3.35) does not seem to be warranted here—the correlation coefficients between neighboring blocks are small enough to allow the straightforward graphical interpretation.

3.7 Resolution and aliasing in short and long arrays

When comparing these three approaches, the relationships between a number of parameters must be kept clearly in mind. These are the standard deviation $\varepsilon$ and resolution number $p_o$ assumed for the original a priori probabilities, the standard deviation $\varepsilon_p = p_o^{-1/2} \{\text{tr}(R+I)^{-1}\}^{1/2}$ and cut-off number $P$ for the a posteriori probabilities, the number of realizations $M$ and data points $N$, and the critical value $r_c$ determining the boundary of the measurable subspace $M$. Adjustments of the attention function $s(k)$ to minimize the effects of aliasing may occur in all three approaches. It is worth distinguishing two limiting cases. The first is the short array ($N$ small) in which the resolution is determined by the individual locations of the observation points. A large number of realizations is necessary to achieve high precision. The second case is that of the long array ($N$ large), where the resolution is set by the investigator low enough that many widely separated segments,
each having $N_o$ data points, are contributing approximately independently and additively to the information matrix, and it is the product $MN/N_o$ which governs precision. $M$ may even be unity. It will be assumed that there is no long range order; that is, the properties of the array can be described statistically, with separate segments of size $N_o$ being homogeneous and uncorrelated.

For the short array, the number of non-zero eigenvalues is determined by the array geometry, and is at most $1/2 N(N+1)$. As the a priori resolution $p_o$ is increased beyond a certain limit, the non-trivial eigenfunctions become completely measurable and insensitive to $p_o$, whereas the new eigenfunctions introduced are all in the intrinsic alias with $r_p = 0$. For each non-trivial eigenfunction, $r_p$ is proportional to $\varepsilon^2/p_o M$, hence everything is insensitive to $p_o$ if $\varepsilon^2/p_o$ is kept constant. However, to have stability of the final estimate to changes in $\varepsilon^2/p_o$, and to eliminate partial biases from that estimate, the value of $r_c$ determining the boundary of the measurable subspace should always be chosen substantially greater than unity, enabling replacement of $\text{tr}(R + I)^{-1}$ by $\text{tr}(R^{-1})$, and $1 + r_p$ by $r_p$. Then, writing

$$r_p = \frac{\varepsilon^2}{p_o MN} r_p^*$$

the average relative a posteriori variance from (3.29) is

$$\int ds \sqrt{\bar{\nu}_s^2} \approx \frac{\varepsilon^2}{p_o} \sum I \frac{1}{r_p} = \frac{P}{MN} \left[ \frac{1}{P} \sum I \frac{1}{r_p^*} \right],$$

(3.44)

where $P$ is the smallest integer for which $(\varepsilon^2/p_o MN) r_p^* > r_c$. Thus the choice of
\[ r_c = \frac{e^2}{P^0} \frac{MN}{P} r^* \] (3.45)

for given \( P \), provided only that this is large compared to unity, will ensure the complete independence of the final estimate from the parameters in the a priori probabilities, though not from the attention function \( s(k) \). It is, of course, supposed here that the array geometry, and hence the value of \( N \), is kept fixed. The expression \( \frac{1}{P} \sum_1^P \frac{1}{r^*} \) expresses the average value of \( 1/r^* \) in the measurable subspace. Since \( r^* \) decreases to zero as \( P \) increases, attempts to increase the resolution beyond a certain point by increasing the cut-off parameter \( P \) encounter a rapid loss of precision; the tradeoff is more severe than the familiar one of variance proportional to resolution.

For the long array, on the other hand, increasing \( P \) introduces new eigenfunctions that have eigenvalues within the range of interest. Indeed, it appears that, as \( N \) and \( p_0 \) become infinite, the information matrix \( R \) has a combination of a discrete eigenvalue spectrum and a continuous one. The continuous spectrum includes

\[ r(\sigma) = \frac{e^2}{P^0} M \frac{X}{2\pi} \int d\sigma \frac{dk}{d\sigma} (f(k))^2 \] (3.46)

with associated eigenfunctions

\[ e_0(s) = \delta(s - \sigma) + \frac{h(s)}{s - \sigma} \] (3.47)

\( X \) is the length of the array, \( h(s) \) is a continuous function of \( s \), and

\[ f(k) = \lim_{X \to \infty} \left\{ \frac{\pi}{X} E(k) \sum_{R,S} \cos k (x_R - x_S) C_{RS}^{-1} \right\} \] (3.48)

For finite \( p_0 \), the detailed form of the eigenfunctions (3.47) is sensitive
to the nature of the cut-off in the a priori features as \( p \) exceeds \( p_o \),
but the eigenvalues are characterized by a peak in \( s \) of width \( \mathcal{O}(1/p_o) \),
and the spacing of the eigenvalues is \( \Delta s = \mathcal{O}(1/p_o) \). Thus, the whole
range of \( s \) is covered with a degree of resolution proportional to \( p_o \),
but individual eigenfunctions cannot be traced continuously as \( p_o \)
varies. The possibility of infinitely fine spectral detail need cause
no concern since, in the double limiting process \( N, p_o \to \infty \), the array
has already been assumed to be much larger than required for any given
resolution. Since \( X \) in (3.46) is proportional to \( N \), the formulae
(3.43)-(3.45) are still appropriate, but now \( N \) is large and \( P \) should
be regarded as proportional to \( p_o \). If \( P/p_o \) is held constant as
\( p_o \to \infty \), then the measurable subspace will include functions with
resolution \( \Delta s \approx 1/p_o \) over the complete range of \( s \) for which
\[ \int \frac{2^2(k)}{2} \frac{dk}{ds} \frac{X}{N} \]
is greater than a critical value roughly independent of the other
parameters, and the a posteriori variance (3.44), for a given \( P/MN \), is
insensitive to the a priori parameters.

This rough analysis shows that for long arrays the critical
control in the first approach is the resolution parameter \( p_o \) initially
chosen by the investigator. The cut-off \( P \) should be a major fraction
of \( p_o \) to avoid excluding from the measurable subspace substantial
portions of the range of \( s \), and, if this choice results in \( \nu_k^2 \) being
unacceptably large, the value of \( p_o \) rather than \( P/p_o \) should be reduced.
A somewhat different choice is implied in the second and third ap-
proaches, because the a priori parameters do not enter the specifi-
cation of the orthonormal set \( G^P \) at all, and the only control is the
cut-off \( P \). In all cases, though, the critical control is a resolution
or cut-off parameter, whether a priori or a posteriori. This is, of course, the usual situation with uniformly sampled data.

Before deriving the result (3.46) we will consider qualitatively the circumstances that give rise to aliasing in short and long arrays, and how its appearance depends on the choice of attention function s(k). The objective is to help the investigator recognize general tendencies. Quantitative answers can be found only by exact computation.

According to Section 3.6, any spectral feature \( v(k) \) can be written as the sum of two parts, a measurable part in subspace \( M \), for which the ratio \( r \) in (3.26) is larger than a critical value \( r_c \), which in practice is set substantially greater than unity, and an unmeasurable part in subspace \( N \), for which \( r \) is less than \( r_c \). Two spectral estimates differing only in the features in \( N \) cannot usefully be distinguished by the observational array, though one may be preferred a priori. It is recommended that, once the effective alias \( N \) has been identified, the a priori probabilities should be adjusted so that the amplitude of every feature in \( N \) is exactly zero, thus restricting the family of spectra under consideration to those in \( M \). This keeps the spectral estimator unbiased and avoids giving the impression of information being available when in fact there is none.

Now the ratio \( r \) may be small either because \( v^T(MH)v \) is small (i.e., the array is yielding little information) or because \( v^T T^{-1} v \) is large (i.e., the feature \( v \) is a priori very improbable). We distinguish three cases. When both the first and second statements are true, the feature \( v(k) \) has justly been excluded from consideration as irrelevant. When the first estimate is true but not the second, the
amplitude of the feature in the final estimate $v^*$ of equation (3.23) is determined essentially by the a priori probabilities and is quite uncertain. The class of such features will be known as the intermediate alias, and they are what is affected by the restriction of the permitted spectra to those in $M$. Elimination of this intermediate alias is what is meant by "with minimum aliasing" in Section 3.6. When the second statement is true but not the first, the amplitude in $v^*$ is already very small, essentially because of the a priori assumptions, and there is little effect from a restriction to $M$. For the long arrays referred to in equations (3.46)-(3.48), features just beyond the limit of resolution are in this category. Depending on the amount of observational information actually available, and the confidence level demanded by the investigator, features in this class should be considered for testing according to the procedures in Section 2.8 as to whether the a priori probabilities are indeed consistent with the data.

To understand qualitatively when $v^T H v$ may be small, consider the matrix

$$K(\xi, \eta) = \frac{1}{2} \sum_{U^r, Y^s} \delta(\xi - |x_r - x_s|) C_{ru}^{-1} C_{sv}^{-1} \delta(\eta - |x_u - x_v|), \quad (3.49)$$

valid for all $\xi, \eta \geq 0$. Clearly $K$ is the Fourier transform of $H(k, l)/E(k)E(l)$ and is positive semidefinite. Then, for any feature $v(k)$, for which the covariance function is

$$C^a(\xi) = \int_0^\infty E(k) v(k) \cos k\xi \, dk \quad , \quad (3.50)$$

we have
\[
\nu^T H \nu = \int_0^\infty d\xi \int_0^\infty d\eta \ K(\xi,\eta) \ \text{C}^a(\xi) \ \text{C}^a(\eta) .
\] (3.51)

Since \( \text{C}^a(\xi) \) is smooth on scale \( 1/k_0 \), where \( k_0 \) is the bandwidth of \( E(k) \), the delta functions in (3.49) may be computed on this same resolution. Before the relative magnitude of \( \nu^T (\text{MR}) \nu \) can be estimated, it is important that \( \nu \) be normalized. In order to consider separately the effects of changing attention function and resolution, we will not use the measure provided by the a priori probabilities, but will suppose instead that

\[
\int_0^\infty \{ \text{C}^a(\xi) \}^2 d\xi = 1 .
\] (3.52)

This is arbitrary, but sufficient for our purposes.

For a short array, the values of \( \xi \) and \( \eta \) covered by non-zero \( \text{C}^a \) may range over the whole length of the array. Equation (3.49) shows that \( K(\xi,\eta) \) vanishes for all \((\xi,\eta)\) except those values equal to one of the lags \( |x_r - x_s| \) between pairs of array points \((r,s)\). Since \( \text{C}^{-1}_{rs} \) is a strictly positive definite matrix, \( \nu^T H \nu \) will be exactly zero if and only if \( \text{C}^a(|x_r - x_s|) = 0 \) for all pairs \((r,s)\), which is equivalent to equation (3.24). This condition defines the intrinsic alias for the array, the class of features which lie in \( N \) regardless of the value of \( M \) or the statement of a priori probabilities. Since \( M \) must be large for the statistics to be meaningful for a short array, \( M(\nu^T H \nu) \) will be small only if \( \text{C}^a(\xi) \) is small when \( \xi = |x_r - x_s| \) for every \( r,s \); that is, it should be small compared to its typical values for \( \xi \) in between those lags. Whether such a \( \text{C}^a(\xi) \) is permitted a priori remains to be seen.
For a long array, there are many distinct lags $|x_i - x_s|$ in any given interval of $\xi$ or $\eta$, lags which arise in widely separated segments of the array and contribute additively to $K$. Thus it is only the average properties of $K(\xi, \eta)$ over many shorter segments that are relevant, and in some respects the result is best regarded as a continuum. We will also suppose that contributions from segments further apart than a certain distance $L_o$ are independent. $L_o$ measures the extent of the long-range order in the array. However, if there is a characteristic short-range structure (such as the absence of observation points closer together than $\Delta x_{\min}$) which almost all segments have in common, it will be reflected in the average. In particular, working outwards from the origin, there is always a peak in $K$ at $\xi = \eta = 0$ (corresponding to the lag from the self-identical pairs $r = s$), surrounded by a region where $K$ vanishes, extending out to $\xi, \eta = \Delta x_{\min}$. Whether there is a further peak at $\xi = \Delta x_{\min}, \eta = 0$ or $\Delta x_{\min}$, depends on whether $\Delta x_{\min}$ is the time greatest lower bound of adjacent lags in almost every segment, or whether for $\xi \geq \Delta x_{\min}$ there is an effectively continuous lag distribution. For a strictly uniformly spaced array, the lag distribution is a row of delta functions. For a quasi-uniformly spaced array with a local spacing which varies continuously and slowly along its length, the lag distribution is peaked at $\xi = 0$, but is continuous for $\xi > \Delta x_{\min}$. For all $\xi > L_o$, the limiting length of long range order, the lag distribution is certainly continuous.

Qualitatively, the behavior of $K(\xi, \eta)$ for small or moderate $\xi, \eta$ is dominated by the lag distribution and is fairly insensitive to the spectral shape $E(k)$, provided the latter is positive everywhere. If
K is written as a sum \( \sum \frac{f(\xi)}{\alpha} f(\eta) \), for example by Cholesky factorization (see Section 6), then according to equation (3.51) the only way that \( v^T H v \) can be small is for \( \int_{0}^{\infty} \frac{f(\xi)}{\alpha} C^\alpha(\xi) d\xi \) to vanish or be small for almost every \( \alpha \). Numerical experimentation shows that, for each \( \alpha \), \( f(\xi) \) has the same structure as the lag distribution, though with coefficients that depend on \( \alpha \) and \( \xi \), and as \( \alpha \) varies the \( f(\xi) \) sample the whole space \( \xi \geq 0 \). Thus, the information from the observational array will be small if and only if \( C^\alpha(\xi) \) largely avoids all the regions of \( \xi \) in which there is a substantial lag density. In particular, this implies at least that

\[
C^\alpha(0) \approx 0 \quad \text{and} \quad C^\alpha(\xi) \approx 0 \quad \text{for} \quad \xi > L_0 \quad ,
\]

and if there is a well defined minimum spacing \( \Delta x_{\min} \)

\[
C^\alpha(\Delta x_{\min}) \approx 0 \quad .
\]

Characterizing the intermediate alias involves determining which, if any, of the functions \( C^\alpha(\xi) \) not excluded by the a priori assumptions can satisfy these constraints.

The class of circumstances for which \( v^T T^{-1} v \) is large is independent of the array geometry but does depend upon \( s(k) \) and the cut-off parameter \( p_0 \). To characterize those \( C^\alpha(\xi) \) which are permitted as not a priori improbable, we expand the corresponding \( v(k) \) as

\[
v(k) = \sum_{p} a_p \cos \{ \pi(p-1)s(k) \} \quad \text{(3.53)}
\]

and determine the coefficients \( a_p \). Since according to the formulation (3.10)
\[ \nu_{-1}^{T_{-1}} = \sum_{p=1}^{\infty} a_p^2 \exp \left( \frac{p-1}{p_0} \right)^2 , \]

any significant power in \( a_p \) for \( p > p_0 \) will distinguish a highly improbable feature. In the alternative formulation (3.12), the reduction in amplitude of high resolution features is different in detail but has the same general character, namely that functions \( v(k) \) varying on a finer scale than \( \Delta k = \pi p_0 \) \( ds/dk \) are highly unlikely, and the same criterion is appropriate.

To estimate \( a_p \), suppose that \( \cos\{\pi(p-1)s(k)\}, p = 1,2,\ldots \) is indeed a complete set of functions over the range \( 0 \leq k \leq k_0 \), where \( k_0 \) is the band limit of \( E(k) \). This may require slight modification of \( s(k) \), so that everywhere in the range the weight function \( ds/dk \) is non-zero, though possibly very small. Then, using the usual orthogonality conditions

\[ a_p = \int_{0}^{1} v(k(s)) \cos \{\pi(p-1)s\} \, ds \]

\[ = \int_{0}^{\infty} g_p(\xi) \cdot C^a(\xi) \, d\xi , \]

where

\[ g_p(\xi) = \int_{0}^{k_0} \frac{1}{E(k)} \frac{ds}{dk} \cos \{\pi(p-1)s(k)\} \cos{\{k\xi\}} \, dk \ . \]

The function \( g_p(\xi) \) can, in general, only be computed numerically, but since our interest is in \( p \) large, its qualitative characteristics can readily be estimated by the method of stationary phase. This yields

\[ g_p(\xi) \sim \frac{1}{2\pi\sqrt{(p-1)}} \frac{1}{\sqrt{\frac{d^2s}{dk^2}}} \left( \frac{1}{E(k)} \cos \left( \pi(p-1)s(k) - k\xi \pm \frac{\pi}{4} \right) \right) + O \left( \frac{1}{p^{3/2}} \right) , \]

(3.55)

where \( k(\xi) \) is defined by
\[ \frac{ds}{dk} (k) = \frac{\xi}{\pi(p-1)} , \quad (3.56) \]

and the \( \pm \) is according to whether \( d^2s/dk^2 \) is negative or positive.

Equation (3.55) shows that, for large \( p \), \( g_p(\xi) \) is highly oscillatory, with amplitude slowly varying according to \( k(\xi) \). When differentiating the phase of the cosine with respect to \( \xi \), all terms in \( dk/d\xi \) vanish, so the locally dominant wavenumber is precisely \( k \) as determined from (3.56). The expression (3.55) is the leading term of an asymptotic expansion for large \( p \), with fixed \( \xi \), valid in the range where

\[ k\xi = \pi(p-1)k \frac{ds}{dk} \gg 1 \]

and

\[ \xi < \pi(p-1) \left( \frac{ds}{dk} \right)_{\text{max}} . \]

For larger \( \xi \) there is no stationary phase solution and the integral decreases exponentially as \( p \to \infty \). As usual for a stationary phase approximation, the inclusion of higher order corrections proportional to \( p^{-3/2} \), \( p^{-5/2} \), ..., changes the amplitude of the expression (3.55) but not its oscillatory behavior. It is not strictly valid to extend this solution in to \( \xi = 0 \), but a smooth continuation completes the neighboring oscillation with half wavelength approximately appropriate to \( k\xi = \pi/2 \), or of wavenumber defined by

\[ k \frac{ds}{dk} = \frac{1}{2(p-1)} \quad , \quad (3.57) \]

As \( \xi \) increases from zero, for the first few oscillations \( k(\xi) \) varies more rapidly with \( \xi \) than a strict application of stationary phase would permit, but \( g_p(\xi) \) is approximately sinusoidal with wavenumber given by (3.57).
In Figure 20(a) is sketched a typical attention function $s(k)$, in the form of $k \frac{ds}{dk}$ versus log $k$. The plateau of constant $k \frac{ds}{dk}$ is the primary wavenumber range of interest, in which $s(k)$ is proportional to log $k$. The slopes at either end portray a much reduced weight function outside this primary range. For any given $p$, the ordinate $1/(2(p-1))$ selects two wavenumbers on these slopes, one small, one large, which describe the oscillations near the origin of two additive stationary phase approximations to $g_p(\xi)$. Of most interest is the high wavenumber solution, sketched in Figure 20(b). As $\xi$ increases, the wavenumber decreases slowly as

$$k \frac{ds}{dk} = \frac{k\xi}{\pi(p-1)}$$

moves up the slope towards the primary range. The significant conclusion is that the first few oscillations of $g_p(\xi)$ are approximately sinusoidal, with a dominant wavenumber that for moderate $p$ is always close to the upper end of the primary range of attention in the region where $k \frac{ds}{dk}$ is decreasing rapidly as $k$ increases. For all larger $k$ up to the band limit $k_o$, there is always a solution of (3.57) for some larger $p$, even though it may be very much greater than $p_o$. For the range of $\xi$ corresponding to these first few oscillations, the low wavenumber component of $g_p(\xi)$ always is effectively constant.

We are now in a position to characterize those features $C^a(\xi)$ which are not a priori unlikely. When considering the intermediate alias, we may restrict attention to $C^a(\xi)$ which are very small for all $\xi$ greater than a few multiples of $\Delta x_{\text{min}}$, and certainly small for all $\xi > L_o$. This range will also be a small number of wavelengths.
of the wavenumber $k_*$ defined by (3.57) with $p = p_0$. Then the integral (3.54) becomes effectively the Fourier cosine transform, for wavenumber $k$ corresponding to $p$ by (3.57). This is not exactly true, since $f_p(\xi)$ is not monochromatic, but it is clear that for a feature to be permitted it is necessary that $C^a(\xi)$, regarded as an even function of $\xi$, be smooth on all wavenumbers greater than $k_*$, in the sense that such Fourier components vanish or be extremely small. In addition, to avoid contributions to $a_p$ from the low wavenumber component of $g_p(\xi)$, it is necessary that $\int_0^{\infty} C^a(\xi) d\xi \approx 0$. Since for the effective alias $C^a(0) \approx 0$,

$$C^a(\Delta x_{\min}) \approx 0,$$

and, as an even function of $\xi$, $d/d\xi C^a(0) = 0$, there is no way all these conditions can be met if $k_*$ is only slightly larger than $\pi/\Delta x_{\min}$. On the other hand, if $k_* \approx 2\pi/\Delta x_{\min}$, some suitable $C^a(\xi)$ can almost certainly be found. Thus the intermediate alias will be significant if the primary range of $s(k)$ extends much beyond the wavenumber $\pi/\Delta x_{\min}$, but is assuredly eliminated if $k ds/dk$ is decreasing rapidly for all wavenumbers greater than $\pi/\Delta x_{\min}$.

Quite different considerations apply when we consider the limits of resolution within the primary range of interest. Consider

$$C^a(\xi) \approx e^{1/2} \frac{\xi^2/\xi_o^2}{\xi_o} \xi \sin k_o \xi,$$

which corresponds to

$$v(k) \approx \frac{1}{E(k)} \frac{\partial}{\partial k_o} e^{-1/2(k-k_o)^2 \xi_o^2},$$

and distinguishes two neighboring wavenumber bands centered on $k_o$ but separated by $2/\xi_o$. Multiplying by $g_p(\xi)$ and integrating over $\xi$, the only significant contributions to $a_p$ come from where $C^a(\xi)$ and $g_p(\xi)$
oscillate synchronously; that is, where $k(\xi) = k_0$ and
\[
\xi = \pi(p-1) \left. \frac{ds}{dk} \right|_0.
\]
Thus $a_\perp$ has magnitude proportional to $\exp -1/2 \left( \pi(p-1) \left. \frac{ds}{dk} \right|_0 \xi \right)^2$, and the feature is permitted if this decreases as $p \to \infty$ faster than $\exp -1/2 \left( \frac{p}{p_0} \right)^2$; that is, if
\[
\xi_0 < \pi p_0 \left. \frac{ds}{dk} \right|_0.
\]
Thus the maximum wavenumber resolution allowed by the a priori probabilities is $2/\pi p_0 \left. \frac{ds}{dk} \right|_0$. Now if $\xi_0$ exceeds $L_0$, the extent of long-range order, there is certainly substantial overlap between $C^a(\xi)$ and $K(\xi, \eta)$, so according to equation (3.51) there is considerable information from the observational array, regardless of whether the feature is permitted or not. For a sufficiently short array, on the other hand, the factor $\xi$ in $C^a(\xi)$ ensures that the effective overlap is small, and there is little information. Thus for a short array the feature is in the intermediate alias if (3.58) is satisfied, otherwise it is excluded as irrelevant. For a long array, on the other hand, it is measurable if (3.58) is satisfied, and is potentially inconsistent with the data otherwise. Note that, with suitable a priori assumptions and large enough $N$, a long array always permits indefinitely fine resolution, even for wavenumbers such that $k > \pi/\Delta k_{\min}$, whose low resolution features are aliased. This somewhat paradoxical result is a consequence of the different ranges of the associated covariance functions.

To conclude this section, we present briefly a derivation of equation (3.46). This follows from the behavior of $K(\xi, \eta)$ when $\xi, \eta$ are both large. Now in equation (3.49) indices $(r,s)$ and $(u,v)$ can be exchanged within pairs simultaneously without affecting anything
so for $\xi, \eta \neq 0$ we may assume without loss of generality that $x_r > x_s$ and

$$
x_r = \theta + 1/2 \zeta + 1/2 \xi$$
$$x_s = \theta + 1/2 \zeta - 1/2 \xi$$
$$x_u = \theta - 1/2 \zeta \pm 1/2 \eta$$
$$x_v = \theta - 1/2 \zeta \mp 1/2 \eta$$

for some discrete $\theta$ and $\zeta$, where the upper sign holds if $x_u > x_r$.

Since $C_{ru}^{-1}$ and $C_{sv}^{-1}$ effectively vanish when

$$|x_r - x_u| = |\zeta + 1/2 (\xi \mp \eta)|$$

and

$$|x_s - x_v| = |\zeta - 1/2 (\xi \mp \eta)|$$

respectively are greater than some distance $L$, contributions to $K$ are very small unless both $|\xi|$ and $1/2 |\xi \mp \eta|$ are less than $L$. For $\xi, \eta > L$ the second choice of sign is ruled out. Furthermore, if

$$1/2 (\xi + \eta) > L + L_o,$$

the segments $(x_r, x_u)$ and $(x_s, x_v)$ are separated by a distance at least $x_u - x_s = 1/2 (\xi + \eta) - \zeta > L_o$, and are effectively independent. Keeping $\xi$ and $\eta$ fixed, $\theta$ ranges over the whole length $X$ of the array. As $X \to \infty$, the total (3.49) approaches $X$ times an average value, which because of the independence of segments $(r,u)$ and $(s,v)$, can be written as the product of the average from each segment separately. Thus

$$K(\xi, \eta) \approx \frac{1}{2} X \int_{-L}^{L} f(1/2 (\xi - \eta) - \zeta) \hat{f}(1/2 (\xi \mp \eta) + \zeta) d\zeta,$$

where

$$\hat{f}(\zeta) = \lim_{X \to \infty} \frac{1}{X} \sum_{r,u} \delta(\zeta - (x_r - x_u)) C_{ru}^{-1}.$$
The preceding equation is the Fourier transform of equation (3.48). The approximation to \( K(\xi, \eta) \) above is valid for all \( \xi, \eta \) greater than \( L + L_0 \), so that, Fourier transforming and multiplying by \( E(k) E(l) \),

\[
H(k, l) \approx \frac{1}{2\pi} X f(k)^2 \delta(k - 1) + H^{(1)}(k, l) ,
\]

where \( H^{(1)} \) is smooth on scales \( 1/(L + L_0) \). For block features of with \( \Delta s = 1/p_0 \) and amplitude \( \varepsilon \), the corresponding information matrix has the structure

\[
R_{st} = \frac{\varepsilon^2 \Delta X}{p_0} \frac{1}{2\pi} \int ds \int dt \ f(k)^2 \delta_{st} + R^{(1)}_{st} ,
\]

where \( R^{(1)}_{st} \) is a smooth function of \( s \) and \( t \), proportional to \( \varepsilon^2 \Delta X/p_0^2 \). For large \( p_0 \), the first, diagonal term in \( R_{st} \) dominates, leading to the eigenvalue (3.46) and eigenfunction \( \delta_{s} \) for each integer \( \sigma \). As \( p_0 \to \infty \), this eigenfunction becomes the first part of equation (3.47). Substituting the full form (3.47) into the expression for \( R_{st} \) above, we obtain an integral equation for \( h(s) \) which is in principle soluble for all \( \sigma \), provided \( r(\sigma) \) has the value (3.46).

For completeness, we include some remarkable results involving the function \( f(k) \) defined in (3.48), even though they have no obviously important application. Some limitations on the structure of the R-matrix for long arrays are implied by the remark that, viewed from the "natural" bandwidth \( \Delta k = 2\pi/X \) (so that \( E_k = 2\pi/X E(k) \)), the column total for (3.19) becomes

\[
\sum_l H_{kl} = f(k) , \quad (3.59)
\]
where \( f(k) \) is defined by (3.48). However, according to the
of the preceding paragraph, the power in the central delta function is

\[
H_{kk} = (\Delta k)^2 H(k,k) = f(k)^2 .
\]  

Since the off-diagonal terms \( H_{kl} \) are all positive, it follows that

\[
f(k) < 1 ,
\]  

with the diagonal term dominating the sum of the off-diagonal elements
if and only if \( f(k) \equiv 1 \). Furthermore, since \( \sum_{k,l} H_{k,l} = 1/2 N \), we also
have

\[
\int_0^\infty f(k) \, dk \approx \sum_k \Delta k \, f(k) = \frac{\pi}{\Delta x} ,
\]  

where \( \Delta x \) is the average spacing over the length of the array. The
condition \( f(k) \equiv 1 \) will be seen in Section 4 to be satisfied only for
a continuously sampled array or a very red spectrum, and by (3.62) can
only be satisfied for a restricted range of wavenumbers.

However, the limitation of diagonal dominance to \( f(k) \equiv 1 \) is not
in practice as restrictive as might at first sight appear, because
the diagonal band has been defined as having infinitesimal width \( 2\pi/X \),
and the significant off-diagonal terms often tend to cluster around the
diagonal. Lumping into block features of finite bandwidth \( dk/ds \Delta s \),
therefore, seems usually to yield a matrix \( R_{st} \) in which the diagonal
terms are substantially larger than those immediately above and below
over a wide range of wavenumbers, with more distant terms negligible.

The extent of and rationale for this behavior is, of course, of great
concern, particularly in the third approach, but so far the authors have
been unable to construct a convincing argument to derive it generally.
3.8 Equally spaced observations

This discussion would be incomplete without a brief review of the simplest case of all—N equally spaced observations of a process \( \phi(x) \) that is periodic in \( x \) with period \( L = N \Delta x \) (e.g., \( x \) could measure distance around a latitude circle on the earth). The spectrum \( E_k \) is then discrete, quantized by \( \Delta k = 2\pi/L \). The covariance matrix is

\[
C_{rs} = E_0 + \sum_{k=1}^{\infty} E_k \cos \frac{2\pi}{N} k(r-s),
\]

(3.63)

where, by our previous convention, \( E_0 = 1/2 \hat{E}(0) \Delta k \). This may also be written as a finite sum

\[
C_{rs} = \frac{1}{2} E^*_0 + \sum_{k=1}^{N/2-1} E^*_k \cos \frac{2\pi}{N} k(r-s) + \frac{1}{2} E^*_{N/2} (-1)^{r-s},
\]

(3.64)

where

\[
E^*_k = E_k + E_{N+k} + E_{2N+k} + \ldots
\]

\[
+ E_{N-k} + E_{2N-k} + \ldots
\]

for \( 1 \leq k \leq N/2 - 1 \) and

\[
E^*_0 = 2(E_0 + E_N + E_{2N} + \ldots)
\]

\[
E^*_{N/2} = 2(E_{N/2} + E_{3N/2} + E_{5N/2} + \ldots).
\]

(3.65)

\( E^*_k \) is the aliased spectrum obtained by summing over the members of an alias class formed by successive reflections about the Nyquist wave-number \( N \Delta k/2 \). This situation is special in that \( C \) can be inverted explicitly for all \( E_k \):
The basic matrix (3.19) then has the elements

\[
(C^{-1})_{rs} = 4/N^2 \left\{ \frac{1}{2E_0^*} + \sum_{k=1}^{N/2-1} \frac{1}{E_k^*} \cos \frac{2\pi}{N} k(r-s) + \frac{(-1)^{r-s}}{2E_{N/2}^*} \right\}
\]

(3.66)

This matrix has the structure shown in Fig. 10, with extensive off-diagonal terms where \((k, 1)\) are members of the same alias class. A natural choice of features \(F_p^k\) for \(p = 0, 1, \ldots, N/2\) is to set \(F_p^k = 1\) if \(k\) belongs to the alias class headed by wavenumber \(p\), with \(F_p^k = 0\) otherwise. Such features make no attempt to distinguish the magnitude \(E_k^*/E_k\) of relative spectral variations at different wavenumbers within the same alias class, but they do lead to an estimate \(\hat{E}_p^k\) of the total variance associated with each distinct class. The information matrix is particularly simple; namely,

\[
R_{pq} = M_{pq}^0 \quad p = 1, \ldots, \frac{1}{2}N - 1
\]

\[
= \frac{1}{2} M_{pq}^0 \quad p = 0 \text{ or } \frac{1}{2}N
\]

(3.68)

For large \(M\), these values are what is to be expected from a discrete Fourier transform of the original data, giving raw spectral estimates.
\( \xi_{\Delta}/\xi_{\Delta} \) which are independent and distributed as \( \chi^2 \)-square with 2 \( M \) degrees of freedom, except when \( p = 0 \) or \( 1/2 \) \( N \), in which case there are only \( M \) degrees of freedom in each estimate. In addition, it is readily shown that if the set of features \( \{ \xi_{\Delta k} \} \) is extended in any way, expanding the subspace of possibilities to include any discrimination between aliased wavenumbers, the new information matrix is always singular, and the new feature \( \xi_{\Delta k}^{N/2 + 1} \), orthonormalized on the measurement capability, has infinite amplitude. Thus there is no information whatsoever available from the observations to distinguish \( \xi_{\Delta k}/\xi_{\Delta k} \) for different \( k \) in the same alias class. This is, of course, also apparent from examination of the way Fourier components contribute to the sample values \( \phi(x_\tau) \) for equally spaced \( x_\tau \). Finally, because of the independence and (nearly) equal uncertainties in the \( \xi_{\Delta k}^* \), the traditional exchange of resolution for precision by averaging adjacent band estimates can be seen to be an optimal estimation process.
4. Optimal array design

4.1 Definitions

The negotiation process described in Section 3—between prior prejudices, the choice of spectral features, and final spectral uncertainty—can to some extent be circumvented if the investigator can specify the array from which the data are taken. In this section the previous results are used to address a different problem: given an a priori statement of the expected spectrum $\tilde{E}(k)$, the (equally probable) features $F^P(k)$ to be tested, and the expected number of realizations $M$, what is the configuration of $\{x_r\}$, $r = 1, \ldots, N$, which will yield the maximum new information about the spectrum? Obviously for a stationary process only the relative positions, $\xi_r = x_{r+1} - x_r$, $r = 1, \ldots, N-1$, matter; therefore, we shall concentrate upon the optimal configuration of adjacent lags $\{\xi_r\}$.

We assume a spectral model of the form (2.7), with prior probabilities $\tilde{\lambda}_p = 0; \overline{\lambda}_p \lambda_q = \delta_{pq}$. (4.1)

For any particular array $\{x_r\}$, the a priori and a posteriori probability distributions for the $\lambda_p$ are given by (3.2) and (3.22), respectively, with $M R = R$ (i.e., we are not restricting ourselves to the features (3.1)) and $T = I$.* According to standard theory (Middleton, 1960), the gain in information $\Delta I$ between these two probability distributions is equal to the change in

$$I \equiv - \int_{-\infty}^{\infty} \rho \log \rho \; d\lambda$$

(4.2)

*We have assumed a $T_{pq} = \delta_{pq}$, which is not consistent with our assumption of small spectral refinements (e.g., equation (3.4)). This inconsistency is unimportant, however. In what follows, for a consistently small $T' = \epsilon^2 T$ ($\epsilon < 1$), we can redefine the number of realizations to be $M' = \epsilon^2 M$, and the optimal array design problem for $(T', M')$ is formally equivalent to that for $(T, M)$ except for an additive constant in the information gain $\Delta I$. 

Thus,

\[ \Delta \mathcal{I} = \frac{1}{2} \log \det (R + I) \]  \tag{4.3} 

\[ = \sum_{q=1}^{P} \frac{1}{2} \log (1 + r_q) \]  \tag{4.4}

where the \( r_q \) are the eigenvalues of the information matrix \( R \). The largest contributions to \( \Delta \mathcal{I} \) come from those principal components for which \( r_q \gg 1 \), which are determined more by observation than initial assumption (see Section 3.6). They are also those for which the initial feature amplitude is irrelevant to the array design. If the principal component has a small eigenvalue, the data add little new information \( (\log (1 + r_p) \approx 0) \).

The optimization problem is to choose \( \{x_r\} \) such that \( \Delta \mathcal{I} \) in (4.3) is maximized. This will be achieved numerically, working through a number of examples so that qualitative trends can be identified. In Section 4.8 a partial theory is presented, seeming valid in the case of a very red first guess spectrum. At a stationary value of \( I \), for each \( x_r \),

\[ \frac{\partial \Delta \mathcal{I}}{\partial x_r} = \frac{1}{2} \text{tr} \left\{ (R + I)^{-1} \frac{\partial R}{\partial x_r} \right\} \]

\[ = \sum_{p,q=1}^{P} \sum_{s,t,u=1}^{N} \left[ \frac{\partial C_{rs}}{\partial x_r} \frac{\partial C_{tu}}{\partial x_r} \frac{\partial C_{pq}}{\partial x_r} \right] \left( \frac{\partial C_{uv}}{\partial x_r} \frac{\partial C_{vw}}{\partial x_r} \frac{\partial C_{wt}}{\partial x_r} \right) \]

\[ = 0. \]  \tag{4.5}

For moderate values of \( N \) and \( P \), values of \( \Delta \mathcal{I} \) and \( \partial \Delta \mathcal{I}/\partial x_r \) can be readily computed numerically for any initial \( \{x_r\} \). Efficient iterative routines (e.g., Fletcher and Powell (1963)) then modify \( \{x_r\} \) until a local maximum of \( \Delta \mathcal{I} \) is attained. There exists no certain technique for finding the global maximum.
4.2 Remarks

The problem of optimal array design is dependent upon a number of variables. They are $N$, $M$, $\tilde{E}(k)$, and $P^P(k)$. For convenience, we shall restrict our attention to the $P^P(k)$ in equation (3.12), for a fixed amplitude $\varepsilon = .25$; thus, the variables governing the features are $w(k)$ and $\Delta s = 1/P$. Below we shall present a small set of examples which indicate the dependence of the optimal arrays on each of these variables.

Firstly, however, we can establish a general feature of the array optimization. Consider a spectrum which is band limited (i.e., non-zero only for $k \leq k_o$) and features which span the entire wavenumber band (e.g., $k_{\min} = 0$ and $k_{\max} = k_o$ in (3.13)). Then

$$\sum_p c^p_{rs} = \varepsilon c_{rs}$$

(4.6)

and

$$\sum_{p,q} R_{pq} = \varepsilon^2 \frac{NM}{2} .$$

(4.7)

Since $R_{pq} \geq 0$ for all $p$ and $q$,

$$\sum_{p,q} R_{pq} \geq \text{tr} |R| = \sum_p r_p .$$

(4.8)

When the number of realizations, $M$, is very large (such that $r_p \gg 1$ for all $p$),

$$\Delta I \approx \frac{1}{2} \log \det R = \frac{1}{2} \log \prod_p r_p .$$

(4.9)

For any set of positive numbers, the geometric mean is bounded by the arithmetic mean; thus,

$$\Delta I \leq \frac{P}{2} \log \left( \sum_p r_p \right) \leq \frac{P}{2} \log \frac{\varepsilon^2 NM}{2} .$$

(4.10)
Equation (4.10) provides a bound upon the possible information gain.* Moreover, it indicates that this bound is achieved (i.e., the inequalities in (4.10) become equalities) only when $R_{pq}$ becomes proportional to the identity matrix. This optimal $R$ structure is generally unattainable in any particular situation, as shown below, but it does represent an optimizing tendency for the $R$ matrix in array design.

Before considering more practical examples, we shall design an optimal least array. This is defined by the following circumstances: the spectrum is band-limited white noise (non-zero only for $k < k_0$), the single feature ($P = 1$) is a wavenumber block function but with a generally different band limit ($k_{\text{min}} = 0$ and $k_{\text{max}} = \hat{k}k_0$), and the array consists of only two points ($N = 2$) and thus a single lag distance $\xi$. The information matrix is only a single number in this case; namely,

$$R = (\epsilon \hat{k})^2 \frac{(1 - \hat{S})^2 + (\hat{S} - S)^2}{(1 - S^2)^2} \quad \text{for } \hat{k} < 1$$

$$= \epsilon^2 \quad \text{for } \hat{k} \geq 1 \quad (4.11)$$

where $\epsilon$ is the feature amplitude (e.g., (3.12)) and

$$S = \sin(k_0 \xi)/(k_0 \xi)$$

$$\hat{S} = \sin(\hat{k} \hat{k} \xi)/(\hat{k} \hat{k} \xi) \ .$$

Equation (4.11) is a function of $\hat{k}$ and $\xi$; the optimal least array design selects $\hat{\xi} = \xi(\hat{k})$ such that $R$, hence $\Delta T$ in (4.3), is a maximum. Because $R$ is only a number here, the number of realizations, $M$, is irrelevant to the array design (though not to the subsequent data analysis).

*The authors are grateful to Mr. C. Bretherton for this proof.
For the particular value \( \hat{k} = 0.5 \), \( R(\xi) \) is plotted in Fig. 11a. This curve illustrates two general characteristics of optimal array design: there exist multiple local maxima of \( \Delta I \), and the maxima in \( \Delta I \) are broad rather than sharp (i.e., moderate changes in \( \xi \) from the optimum value yield reasonably small changes in \( \Delta I \)). As a matter of practice, these characteristics encourage approximate rather than precise solutions to array optimization. In Fig. 11a, the optimal separation for \( \hat{k} = 0.5 \) is \( \hat{\xi} \approx 1.3 \pi/k_0 \). A Nyquist lag \( \xi_N \) can be defined as \( \pi/k_N \), where \( k_N \) is the Nyquist wavenumber (see Section 3.7) and is associated with the highest wavenumber resolved. Thus, \( \hat{\xi} \) lies between the Nyquist lags associated with the spectral and feature cut-offs. Fig. 11b shows that this is generally true of \( \hat{\xi} \) for \( \hat{k} \leq 0.85 \). For small \( \hat{k} \), \( \hat{\xi} \) is closer to \( \pi/k_0 \) than \( \pi/\hat{k}k_0 \) (i.e., the spectral cut-off controls the optimal spacing). For \( \hat{k} \to 1^- \), \( \hat{\xi} \to 1.2\pi/k_0 \), which is not bounded by the two Nyquist lags; however, \( R(\xi) \) becomes very flat in this limit (with \( \partial R/\partial \xi = 0 \) at \( \hat{k} = 1 \)) and the optimal lag value loses meaning. For \( \hat{k} \geq 1 \), the problem is ill-posed; that is, any array has equivalent information.

4.3 A standard example

Throughout Sections 2-3, various analyses were presented with simulated data taken from the 20-point array shown in Fig. 1. This array is, in fact, a nearly optimal one for the circumstances listed in Section 2.7 (i.e., \( M = 100 \), \( \bar{E}(k) \) from (2.36), and \( F_P(k) \) from (2.37)). The array must be described as only nearly optimal, both because the authors have no means of proving that it yields a global maximum of \( \Delta I \) and because another
nearly optimal solution has been found (see Table 5). The two nearly
optimal arrays are quite similar. Both have their least spacings in
the middle, with increases in $\xi_T$ toward the array edges. One is sym-
metric in $\xi_T$ about the central lag, the other is not. Their information
gain $\Delta T$ and their least spacings differ by less than 1%, while their
extents differ by only 2%. Note that the information matrix for the
first array (Table 1) is far from the identity matrix structure identi-
fied in Section 4.2 as the (in this case unattainable) ideal for an
optimal array.

A Nyquist wavenumber for these arrays, $k_N = \pi/(\text{least } \Delta x)$, has the
value $k_N = 19$. From (2.38) we note that $k_N$ lies well within the final
feature's wavenumber interval (n.b., $[k_{\lambda 0}, k_{u 0}] = [12.8, 25.6]$). Thus,
the least $\Delta x$ in these arrays has been selected to resolve the highest
wavenumber feature by a criterion qualitatively similar to that identi-
fied in the least array design of Section 4.2. From Table 5 one can see
that the arrays have several lags corresponding to $k_N$. The greatest $\Delta x
in these arrays can be identified with the largest half cycle which fits
within the array, $k_{\delta} = \pi/(\text{greatest } \Delta x)$. From Table 5, $k_{\delta} = .11$ for the
two arrays. This wavenumber lies barely just below the first feature's
interval (n.b., $[k_{\lambda 1}, k_{u 1}] = [.125, .225]$); that array 1 has very few
lags corresponding to wavenumbers within this first interval is reflected
in the relative smallness (i.e., smallness of estimation skill) of $R_{11}$
in Table 1(a). For reasons discussed below, it appears to be a general
property of optimal design that the requisite least lags are established
with highest priority (i.e., with greatest contribution to $\Delta T$), and the
larger lags required by the low-wavenumber features are established
secondarily. For example, when \( N \) is reduced the range of the adjacent lags \( \{ \xi_r \} \) is little altered, but \( k_g \) is of necessity much smaller (see Section 4.5).

A more precise lag analysis can be made as follows. Analogous to \( s(k) \) in (3.9), we can define a lag space mapping \( \Xi(\xi) \) which transforms \([\xi_{\text{min}}, \xi_{\text{max}}]\) into \([0,1]\). For the features (2.37), defined by the weighting function \( w(k) \) in (3.13), it is appropriate to define

\[
\Xi(\xi) = \frac{\log \left[ 1 + \frac{\xi - \xi_{\text{min}}}{D} \right]}{\log \left[ 1 + \frac{\xi_{\text{max}} - \xi_{\text{min}}}{D} \right]},
\]

(4.12)

where \( D = (\xi_{\text{max}} - \xi_{\text{min}})/(2^P - 1) \). The unit interval in \( \Xi \) is thus split into \( P \) equal sub-intervals. The distribution of the number of lags within these sub-intervals will provide our principal means of describing optimal arrays.

Figure 12 shows the lag distribution for the two arrays of Table 5 and the mapping parameters

\[
\begin{align*}
\xi_{\text{min}} &= \pi/k_{u8} = 0.12 \\
\xi_{\text{max}} &= \pi/k_{\lambda 1} = 25.13
\end{align*}
\]

(4.13)

\( P = 8 \).

Thus, the mapping (4.12) has its range defined by the wavenumbers defining the features (2.37)-(2.38), and its partitioning is equivalent to the number of features. Two types of lag distributions are shown in Figure 11, adjacent (e.g., the \( \{ \xi_r \} \) of Table 5) and independent. The latter is the total number of array point separations, \( x_i - x_j \), excluding in each \( \Xi \) sub-interval one member of every lag pair which is excessively overlapping. An excessive overlap is defined to be an intersection in
physical space of the intervals \([x_j, x_i]\) which exceeds half of a mid-point value for a particular \(\Xi\) sub-interval. For (4.12)-(4.13), the mid-point lags are

\[
\xi_{\text{mp}} = \xi_{\text{min}} + D(2^D - 1/2 - 1).
\] (4.14)

Obviously, it is the totality of lags, not just the adjacent ones, which matters in spectral estimation, and lags which are excessively overlapping yield essentially redundant information about covariances on that scale. The independent lag distribution in Fig. 12 has the property of being nearly uniform in \(\Xi\) for the first 5 or 6 sub-intervals and then tailing off as \(\Xi \to 1\). Thus, an array designed to address nearly equal attention in \(\log k\) has independent lags nearly uniformly distributed in \(\log \xi\). The two different arrays in Table 5 have qualitatively similar lag distributions.

4.4 Realizations

The number of realizations which an investigator expects to obtain from an array influences the optimal array choice. In this section, we consider the optimal arrays which differ from the standard array of Section 4.3 only in their value of \(M\). The lag distributions for various \(M\) are shown in Fig. 13. As \(M\) decreases from the standard value of 100, shorter lags become favored over longer ones, and the distribution of independent lags becomes more a decreasing rather than uniform function of \(\Xi\). An interpretation of this tendency is that, as \(M\) decreases, the a priori probabilities play a more important role in the estimation and the optimal array achieves its greatest gain in information by concentrating on effectively fewer features.
(the ones associated with high wavenumbers). As $M$ increases from 100 to infinity, the lag distribution is little altered, indicating that the standard array lies within the asymptotic regime for large $M$.

4.5 Array points

In this section, we consider changes in the nearly optimal arrays as the number of array points, $N$, is varied. The resulting lag distributions are shown in Fig. 14 for $N$ between 10 and 30. To a remarkable degree the distributions are self-similar, differing from each other only in amplitude, both for adjacent and total independent lags. At least for red spectra, therefore, it is unimportant how an investigator apportions his total number of measurements $N\cdot M$ between realizations and array points once he has enough of each (for this example, $N \geq 15$ and $M \geq 20$ seem to be enough).

4.6 Spectral shape

The standard array (Section 4.3) was designed for the case of a red spectrum (i.e., $3E/3k < 0$). We selected this with geophysical prejudice since such spectra characterize many processes in the oceans and atmosphere. However, the general problem of spectral estimation embraces many types of spectra. In this section, the influence upon optimal arrays of different spectral shapes is examined, while $N$, $M$, and $\{P^p(k)\}$ are kept as in Section 4.3. Note from the definitions of $R$ and $\Delta I$ (i.e., equations (2.19) and (4.3)) that it is only the shape and not the magnitude of $E(k)$ which is relevant here.
Fig. 15 shows the optimal array lag distributions for various \( E(k) \). They include a progression from very red to white (i.e., \( k^{-4}, k^{-2}, k^{-1}, \) and \( k^{-0} \)), all with a common cut-off wavenumber \( k_0 = 40 \) (compare \( k_0 \) with the range covered by the spectral features in (2.37) and Fig. 15), as well as two spectra of mixed type, white-red and blue-red (i.e., \( 1 + k^{-2}, k^2 + k^{-2} \)).

The two red spectra in Fig. 15a have very similar lag distributions, as similar as the two standard case solutions in Fig. 12. The white spectral array in Fig. 15b, however, has a completely different distribution. It has nearly uniform spacing with \( \xi_r \approx .10 \) for all \( r \). This spacing is intermediate between the feature and spectral cut-off lags (i.e., \( \pi/k_u8 = .12 \) and \( \pi/k_o = .077 \)). This property is shared with the least array in Section 4.2 and Fig. 11b (n.b., \( \hat{k} = k_u8/k_o = .6 \) here). Of special interest is the spectrum \( \alpha k^{-1} \) in Fig. 15b. It can be interpreted as the transition between red and white spectra in that two different array solutions have been found, with nearly equal information gain \( \Delta I \), yet with distinctly different lag distributions. One distribution has approximately equal spacing (as is optimal for a band-limited white spectrum), and the other has a lag distribution very similar to those for the red spectra in Fig. 15a.

The mixed type spectra are shown in Fig. 15c. The wavenumber at which the transition between types occurs is \( k_T = 1.625 \), which is also the value of \( k_u4 = k_5 \) in (2.38). Thus, the highest four features lie in the red spectral regime, while the lowest four are in the white or blue regime. Note that \( \xi_T = \pi/k_T = 1.9 \) and \( \Sigma_T = \Sigma(\xi_T) = 0.55 \). By analogy to the least array of Section 4.2, we expect no adjacent lags are required for \( \Sigma \geq .55 \) for the low wavenumber, white regime. (Since \( \hat{k} \ll 1 \) here we
might even expect an absence of adjacent lags near $E_T$—see Fig. 11b). Fig. 15c confirms that there is indeed a cut-off to the adjacent lag distribution below $E_T$. For a blue spectrum at low wavenumbers, the cut-off in adjacent lags is even more pronounced.

We conclude, therefore, that the type of array which is optimal can vary considerably with the assumed spectrum when the features focus nearly uniform attention upon equal increments in log $k$. For a red spectrum, the distribution of independent lags is nearly uniform in log $q$ within a range defined by $\pi$ times the inverses of the wavenumbers associated with the highest and lowest features; at the largest lags in this range the distribution tails off (e.g., Fig. 12). For a white spectrum, the optimal array has nearly uniform adjacent spacing with separation near $\pi$ times the inverse of the highest wavenumber in the white regime. For blue spectral regimes, the optimal arrays have their largest adjacent spacings on scales smaller than $\pi$ times the inverse of the highest wavenumber in the regime (n.b., no realizable spectrum can be blue for all $k$).

4.7 Features

As the final variable in optimal array design, we consider the influence of the features selected. One aspect of the features is the degree of spectral resolution sought. As a variant of the standard array, consider a reduction of the total number of features $P$ from 8 to 4, while the wavenumber range $[k_{\text{min}}, k_{\text{max}}]$ and minimum block width remain fixed. This requires that adjacent wavenumber blocks have widths differing by a factor of four rather than two. The weighting function $w(k)$ remains
as in (3.13), but the size of $\Delta s$ in (3.12) doubles. The lag distributions for this and the standard optimal arrays are shown in Fig. 16. The two arrays have reasonably similar distributions, although the coarser features array has relatively fewer lags at intermediate $E$ and more at large $E$. The coarseness of the features does allow the optimal $R$ matrix (Table 6) to come somewhat closer to the ideal identity matrix structure (i.e., the $R_{pp}$ differ by ratios which can be greater than 4 in Table 1 but are always less than 3 in Table 6).

Alternatively, one can retain the feature resolution of the standard array ($P = 8$) and change the wavenumber weighting function (3.13) to one which is uniform in $k$:

$$w(k) = \begin{cases} \frac{1}{k_{\text{max}} - k_{\text{min}}} , & k_{\text{min}} \leq k \leq k_{\text{max}} \\ 0 , & \text{otherwise} \end{cases} \quad (4.15)$$

For such features, the optimal arrays all have nearly uniform spacings, with values near $\pi/k_{\text{max}}$ or $\pi/k_o$, for each of the spectra shown in Fig. 15. The case which comes closest to equal spacing is that with the features (4.15) plus the band-limited white spectrum of Fig. 15b; there the $\{\xi_r\}$ all lie within the range $[.095, .099]$. In general, the features chosen by the investigator strongly influence the optimal array design. The degree of resolution sought is less important than the weighting function $w(k)$, however.

Finally, we remark that for any band-limited spectrum (i.e., $E = 0$ for $k > k_o$), block features covering the interval between $k_{\text{max}} = k_o$ and $k_{\text{min}} = 0$, weighting function as in (4.15), and $P = N/2 + 1$, then a uniform array spacing, with $\{\xi_r\} \equiv \{2\pi/k_o\}$, is optimal. Thus, the traditional analysis described in Section 3.8 is in fact an optimal match between analysis and array design.
4.8 Quasi-uniformly spaced arrays

The preceding results may also be interpreted against an approximate calculation, applicable when the first guess spectrum is very red and the number of array points large, so that the points can be locally fairly uniformly spaced while the spacings may vary by a large factor over the length of the array. Many of the optimal array designs computed according to equation (4.5) appear to have this property. An approximate estimate of the information matrix $R$ can be obtained by replacing the array by independent additive segments, each precisely uniform. For such a segment of $n$ points at spacing $\Delta x$, the contribution to the $H$ matrix is given by equation (3.67) as

$$H(k,l) = \frac{2}{\Delta k} e(k,\kappa) e(l,\kappa) \{ \delta(k-1) + \delta(k+1-\kappa) + \delta(|k-1|-\kappa) + \ldots \}$$

(4.16)

where

$$\Delta k = \frac{2\pi}{n\Delta x}$$

is the wavenumber resolution based upon the complete length of the segment,

$$H(k,l) = H_{kl}/(\Delta k)^2$$

is the information density matrix in wavenumber space,

$$e(k,\kappa) = \frac{E_k}{\sum_{m=-\infty}^{\infty} E(k + m\kappa)}$$

(4.17)

is the ratio of the first guess spectral density to the fully aliased value obtained by reflections about 0 and the Nyquist wavenumber

$$\frac{1}{2} \kappa = \frac{\pi}{\Delta x}$$.
and the sequence of Dirac delta function in (4.16) describes the fundamental component followed by various aliases of increasing order. Note that

\[ \delta(k - 1) = \frac{1}{\Delta k} \delta_{kl} \cdot \]

The factor of 2 in \( \mathcal{H} \) should be omitted when \( 2k/\kappa \) is an integer, but this is a negligible correction to what follows. If the spectrum is red, the higher order aliases are relatively unimportant, as may be seen from examination of the ratio function \( e(k, \kappa) \) when \( E(k) \propto k^{-m} \) for \( m \geq 2 \). For large \( m \), \( e(k, \kappa) \) is close to unity for \( 0 \leq k < 1/2 \kappa \), and very small for larger \( k \). We will consider the red noise limit \( m \to \infty \), replacing \( e(k, \kappa) \) by a step function in \( k/\kappa \) and ignoring all but the first term in equation (4.16). Contributions from, for example, the second delta function require either \( k \) or \( l \) to be greater than \( 1/2 \kappa \) and are thus small corrections.

Now add contributions from segments with various spacings \( \Delta x \), each of which has \( n(\Delta x) \) array points, to yield the total fundamental matrix

\[
\mathcal{H}^{(o)}(k, l) = \sum_{\Delta x < \pi/k} \frac{1}{\pi} \Delta x \ n(\Delta x) \ \delta(k - 1) \\
= \frac{1}{\pi} \lambda(k) \ \delta(k - 1) \ , \quad (4.18)
\]

where \( \lambda(k) \) is the total length of the array that is covered by points of spacing less than \( \pi/k \). The approximations made imply that the segments so covered behave as if the sampling of \( \phi(x) \) were continuous, whereas if the spacing is greater than \( \pi/k \) they yield no information. The delta function in (4.18) implies perfect discrimination between
wavenumbers, though clearly the resolution must be limited by the minimum of \( l/(nAx) \) for the various segments. In reality the array is slowly varying, and the scale over which the spacing changes significantly sets the resolution limit for this analysis.

Given an attention function \( s(k) \), with block features defined by small equal intervals \( \Delta s \), the fundamental information matrix is

\[
R^{(o)}_{st} = (\Delta s)^2 \frac{dk}{ds} H^{(o)}(k,1) \frac{dl}{dt} = \frac{1}{\pi} \Delta s \frac{dk}{ds} \lambda(s) \delta_{st}. \tag{4.19}
\]

This is a diagonal matrix with eigenvalues

\[
\lambda_s = \frac{\Delta s}{\pi} \frac{dk}{ds} \lambda(k) \tag{4.20}
\]

at the value of \( k \) defined by incrementing \( s \) from 0 to 1 at intervals \( \Delta s \). These eigenvalues will shortly be used to find the optimum choice for \( n(\Delta x) \).

When \( m \) is only moderately large, aliasing is not entirely negligible and \( R_{st} \) is not diagonal. There is a small continuous background, dominated by the second delta function in equation (4.16), which represents the first alias with \( k, l \) both close to \( 1/2 \). Specifically this term makes a contribution

\[
R^{(1)}_{st} = \frac{(\Delta s)^2}{\pi} \frac{dk}{ds} \frac{dl}{dt} \left| e\left(\frac{k}{k+1}\right) \right| = \left| e\left(\frac{1}{k+1}\right) \frac{dl}{dk} \right| = k+1
\]

\[
\approx \frac{(\Delta s)^2}{\pi} \left(\frac{dk}{ds}\right)^2 \frac{dl}{dk} \left| \kappa = 2k \right| \left(\frac{m}{k+1}\right)^2 \text{ as } m \to \infty \tag{4.21}
\]

which is non-zero only in a narrow band about the diagonal where
\( m(k-l/k+l) = O(1) \). Note that this continuum \( R^{(1)} \) implies a lack of resolution between nearby wavenumbers, and is proportional to \( \frac{d\lambda}{dk}\bigg|_{k=2k} \) (i.e., to the number of points with spacing closed to the Nyquist limit \( \Delta x = \pi/k \)), and is associated directly with folding about that limit. However, the effect of this background upon the eigenvalues (4.20) is moderately small, being formally \( O(1/m) \).

We now use the estimate (4.19) to choose \( n(\Delta x) \) so as to maximize the information gain \( \sum_s \log \lambda_s \) from observations of a given number of realizations \( \phi(x) \). \( \Delta s \) will be supposed small, so sums may be replaced by integrals. \( n \) will be regarded as a function of the continuous variable \( k = \pi/\Delta x \), so that \( n(k) \, dk \) is the number of adjacent lags in the range \( (k, k+\Delta k) \), and

\[
\ell(k) = \pi \int_k^\infty \frac{1}{k^T} \, n(k') \, dk'.
\]

(4.22)

This distribution is of course subject to the constraints

\[
\int n(k) \, dk = N,
\]

(4.23)

\[
n(k) \geq 0,
\]

(4.24)

of which the second is non-linear. For any infinitesimal change \( \delta n \), the change in information is

\[
\delta I = \sum_s \frac{1}{\lambda_s} \, \delta \lambda_s
\]

\[
= \int_0^1 \frac{ds}{\lambda_s} \, \frac{dk}{ds} \left\{ \int_k^\infty \frac{1}{k^T} \, \delta n(k') \, dk' \right\}
\]

\[
= \int_0^\infty \frac{1}{k} \, f(k) \, \delta n(k) \, dk
\]

(4.25)
where
\[ f(k) = \int_0^k \frac{dk'}{\lambda_s(k')} \]

is a function that is yet to be determined. The constraint (4.23) requires that
\[ \int \delta n(k) \, dk = 0 \]

but equation (4.25) shows that a change transferring \( n(k) \) from wave-numbers where \( 1/k \, f(k) \) is small to those where it is a maximum will always increase \( I \), unless such a change is prohibited by the constraint (4.24). Thus we are led to the following necessary conditions for a maximum:

either \[ \frac{1}{k} \, f(k) = \text{constant} = \mu \quad (4.26) \]
or \[ \frac{1}{k} \, f(k) < \mu \quad \text{and} \quad n(k) = 0 \quad (4.27) \]

Equation (4.26) leads immediately to
\[ \frac{1}{\lambda_s} \frac{df}{dk} = \mu \]

which implies
\[ \lambda(k) \propto \frac{ds}{dk} \quad (4.28) \]

and, if it held for all values of \( k \), would achieve the upper bound (4.10). However, since \( \lambda(k) \) is a monotonically decreasing function of \( k \), and the weight function \( ds/dk \) certainly is not, equation (4.28) cannot hold everywhere, and there must be at least one interval in which (4.27) is applied and \( \lambda \) is independent of \( k \). The location of these regimes is best determined graphically. In Fig. 17 the solid
curve is the weight function \( ds/dk \), and in the intervals \( P'Q \) and \( Q'R \), \( \lambda(k) \) is proportional to \( ds/dk \) and decreases with \( k \). However, in the intervals \( PP' \) and \( QQ' \), \( n(k) = 0 \) and \( \lambda(k) \) is constant, represented by the horizontal dotted lines \( PP' \), \( QQ' \). The vertical placing of each of these dotted lines is such that the net shaded area between it and the solid curve vanishes. To see that this does indeed describe the solution of conditions (4.26), (4.27), note that

\[
\frac{1}{k} f(k) = \int_0^k \frac{\Delta s \frac{dk'}{\lambda(k')}}{\pi \frac{ds}{\lambda(k')}} = \frac{\pi}{\Delta s} \frac{1}{k} \int_0^k \frac{ds}{\lambda(k')} \, dk',
\]

which is proportional to the average of \( (ds/dk)/\lambda(k) \) over the interval \((0,k)\). This average increases to unity over the range \( PP' \), is equal to unity along \( P'Q \), drops below unity in \( QQ' \), and is unity in \( Q'R \). The profile \( PP'QQ'R \) for \( \lambda(k) \) is uniquely determined and is readily seen indeed to define a maximum for the information gain.

This analysis reproduces qualitatively the previously demonstrated features of optimal array designs. For the weight function \( ds/dk \) proportional to \( 1/k \) in the range \((k_{\min}, k_{\max})\) and zero outside, it predicts a concentration of spacings close to \( \pi/k_{\max} \) as \( \lambda(k) \) jumps from 0 in \( k > k_{\max} \) to \( \eta/k \) for \( k \) just less than \( k_{\max} \). There should then be a constant number of adjacent lags for which the spacing lies in each octave range for which

\[
e^{-1} k_{\min} < k < k_{\max},
\]

with none at greater spacings than \( \pi \eta/k_{\min} \). This is not dissimilar from the curve for adjacent lags \( N = 30 \) in Fig. 14, or for the \( N = 20, m = 4 \) in Fig. 15a, except that the cutoff for large spacings appears
to be around $6\pi/k_{\min}$ rather than $2.7 \pi/k_{\min}$, and the corresponding points are distributed at somewhat shorter lags. The reason for this discrepancy is unclear, as is the precise relationship to the analysis in terms of independent lags. It should be emphasized, however, that the neglect of the alias terms, $H^{(1)}$, etc., in equation (4.25) constitutes a substantial approximation valid only in the limit of very red spectra.
5. Higher Dimensions

5.1 Spectral estimation

The preceding formulation is easily generalized to random processes in higher spatial dimensions, a property which is necessary for broad geophysical usefulness. If $k$ is the vector wavenumber, then the appropriate analog of the spectral model (2.9) is

$$E(k; \lambda) = \tilde{E}(k) \left[ 1 + \sum_{p} \lambda_p^p \tilde{P}(k) \right]; \quad (5.1)$$

that is, the spectrum and features are functions of $k$, while $\lambda_p$, the parameter to be estimated from observations, remains a scalar quantity. The associated covariance function $C(\xi; \lambda)$ is now related to $E$ by a multi-dimensional Fourier transform,

$$C(\xi; \lambda) = \frac{1}{2} \int_{-\infty}^{\infty} \int E(k; \lambda) \cos (k \cdot \xi), \quad (5.2)$$

where $E(k) = E(-k)$. For any array $|x_r|$ and a spectral grain size $\Delta k$, there exists a matrix analogous to (3.19) defined by

$$H_{kl} = \frac{1}{8} \sum_{u,v} E_k \cos k \cdot (x_r - x_u) \tilde{C}^{-1}_{ru} \tilde{C}^{-1}_{sv} \cos l \cdot (x_u - x_v) \tilde{E}_l, \quad (5.2)$$

where $E_k = E(k) \Delta k$ as in Section 3.2. The information matrix $R_{pq}$ can then be defined by

$$R_{pq} = M \sum_{k,l} \tilde{p}^p_k \tilde{q}^q_l H_{kl} \tilde{F}^q_l, \quad (5.3)$$

and the modeling fitting discussion of Sections 2.3-2.6 is directly applicable. So also is the Bayesian formulation of Sections 3.4-3.5.
5.2 Isotropic array design

It is not our intention in this paper to extensively focus on higher dimensional aspects of spectral estimation. Let a few illustrations suffice. Consider first the question of optimal array design in two dimensions for an isotropic random process.

For \( K = |k| \), we define the isotropic spectrum \( E_I(K) \) by

\[
E_I(K) = \pi K E(k), \quad (5.4)
\]

and assume that \( E \) is a function of \( K \) only. Then,

\[
C(ξ) = C(|ξ|) = \int_0^\infty E_I(K) J_0(K|ξ|) \, dK, \quad (5.5)
\]

with an analogous formula for \( C_P \) with \( E_I(K) \) replaced by \( E_I(K) F_P(K) \). The optimal array is one which maximizes the information gain \( ΔI \) defined in (4.3). The technique for finding the optimal array point locations \( \{x_r \} \) is again an iterative one; however, in two dimensions there are \( 2N \) unknowns, the two coordinates for each of the \( N \) array points. We designate these unknowns by \( ξ_α, \, α = 1, \ldots, 2N \), and define them by

\[
\begin{align*}
ξ_{2r-1} &= x_r \\
ξ_{2r} &= y_r
\end{align*}
\quad \text{for } r = 1, \ldots, N \tag{5.6}
\]

The counterpart of (4.5), the information gradient, is

\[
\frac{\partial ΔI}{\partial ξ_α} = \sum_{p, q} \left\{ \sum_{u, v, w} \left[ \begin{array}{ccc}
p \quad A_r(ω), w & C_r(ω), u & C^{-1\alpha}_ν \quad C^{-1}_u \\
A_r(ω), w & C_r(ω), u & C^{-1}_ν \quad C^{-1}_u \\
+ \sum_{s, t} B_{r(ω), w} C_{rs} C^P_{\alpha} C^{-1}_w C^{-1}_t C^q \quad C^q_{uv} \end{array} \right] \right\}.
\]

\[
\cdot (R + I)^{-1}_{pq}, \quad α = 1, \ldots, 2N,
\]

\[
\tag{5.7}
\]
where

\[ r(\alpha) = \begin{cases} 
(\alpha + 1)/2, & \alpha \text{ odd} \\
\alpha/2, & \alpha \text{ even} 
\end{cases} \quad (5.8) \]

and

\[ A_{r(\alpha),w}^p = \begin{cases} 
\frac{x_{r(\alpha)} - x_w}{E_{r(\alpha),w}} \frac{dC_p}{dE} \left( E_{r(\alpha),w} \right), & \alpha \text{ odd} \\
\frac{y_{r(\alpha)} - y_w}{E_{r(\alpha),w}} \frac{dC_p}{dE} \left( E_{r(\alpha),w} \right), & \alpha \text{ even} 
\end{cases} \quad (5.9) \]

\[ B_{r(\alpha),w} = \begin{cases} 
\frac{x_{r(\alpha)} - x_w}{E_{r(\alpha),w}} \frac{dC(E_{r(\alpha),w})}{dE}, & \alpha \text{ odd} \\
\frac{y_{r(\alpha)} - y_w}{E_{r(\alpha),w}} \frac{dC(E_{r(\alpha),w})}{dE}, & \alpha \text{ even} 
\end{cases} \quad (5.10) \]

In these expressions, \( E_{uv} = |x_u - x_v| \), and \( C \) is a function of \( E \) only.

Now consider the problem of two-dimensional array design for isotropic features and a red, isotropic spectrum. We assume (1) \( E_{\perp}(K) \) has the functional form of (2.36) (with \( \nu = .002 \) as before);* (2) \( F^p(K) \) are logarithmic block functions of the form (2.37)-(2.38) (with \( K_{\min} = .125 
\]

*The one-dimensional spectrum \( E_{\perp}(k) \) is similar to, but not identical with, the isotropic spectrum \( E_{\perp}(K) \). Formally, the first is related to \( C(\xi) \) by a one-dimensional Fourier transform, while the second is related to it by the Bessel transform (5.5). For a power law asymptotic regime, however, it is easy to show that

\[ E_{\perp}(K) \propto K^{-m} \iff E_{\perp}(k) \propto k^{-m} \]

for all \( m > 1 \).

---

\[ \frac{x_{r(\alpha)} - x_w}{E_{r(\alpha),w}} \frac{dC_p}{dE} \left( E_{r(\alpha),w} \right), \alpha \text{ odd} \]

\[ \frac{y_{r(\alpha)} - y_w}{E_{r(\alpha),w}} \frac{dC_p}{dE} \left( E_{r(\alpha),w} \right), \alpha \text{ even} \]

\[ \frac{x_{r(\alpha)} - x_w}{E_{r(\alpha),w}} \frac{dC(E_{r(\alpha),w})}{dE}, \alpha \text{ odd} \]

\[ \frac{y_{r(\alpha)} - y_w}{E_{r(\alpha),w}} \frac{dC(E_{r(\alpha),w})}{dE}, \alpha \text{ even} \]
$K_{\text{max}} = 25.625$, and $P = 4$); (3) the number of realizations is large ($M = 100$); and (4) the number of array points is intermediate ($N = 21$).

Three arrays are shown in Figs. 18a,b,c which have local maxima in the information gain $\Delta I$. It is not our intention here to make a thorough study of two-dimensional isotropic array design. However, several qualitative aspects of this problem can be described on the basis of these solutions:

(i) the multiplicity of local maxima in $\Delta I$ is much greater in two dimensions than in one;

(ii) the range of nearest neighbor lag distances is quite similar in one and two dimensions (i.e., $\xi_{\text{min}} = 0.20$ and $\max|x_{i+1} - x_i| = 6.1$ for the $P = 4$, one-dimensional array of Fig. 16, whereas $\xi_{\text{min}} = 0.22-0.24$ and the maximum separation of nearest neighbors is $6.2-6.4$ for the three two-dimensional arrays of Fig. 18a,b,c); *

(iii) the optimal two-dimensional arrays are in fact one-dimensional when $\tilde{E}(k)$ and the $P^p(k)$ are isotropic (n.b., the line array of Fig. 18c is superior to the asterisk and cross configurations of Figs. 18a,b). The most efficient manner of obtaining the range of lags needed (see Section 4.3), independent of direction, is to have all the array points in a line.

*Precise correspondences should not be expected between the one-dimensional and two-dimensional array design problems described here. In the former problem it is $E_{\perp}(k)$ which has the form (2.36) and in the latter it is $E_{\parallel}(K)$.
In summary, the isotropic array design problem in two, and presumably higher, dimensions is unimportantly different from the problem in one dimension. Even the greater multiplicity of solutions is largely irrelevant, since the additional solution types are inferior in their information gain, \( \Delta I \).

5.3 Anisotropic array design

Anisotropy in the spectral model (5.1) is therefore crucial to there being an important distinction between one and two dimensional array design. We shall now consider two simple examples of anisotropy. We assume \( \hat{E}(k) \) is isotropic but consider features which distinguish between different wavenumber vector directions. The types of anisotropy which we shall consider are the following:

\[
F^P(k) = \begin{cases} 
  F^P(K) \cos \theta_k, & p = 1, \ldots, Q \\
  \text{and/or} \\
  F^{P-Q}(K) \sin \theta_k, & p = Q+1, \ldots, P 
\end{cases} 
\]  

(5.11)

where \((K, \theta_k)\) is a polar representation of the vector \( \hat{k} = (k, \ell) \). For comparison with the preceding isotropic problem, we shall put \( Q = 4 \) and identify the \( F^P(K), p = 1, \ldots Q \), with the previous isotropic, wavenumber block features. In the first example below, we put \( P = Q = 4 \) and thus included only the \( \cos \theta_k \) features of (5.11); these are capable of distinguishing any anisotropy between the \( x \) and \( y \) coordinate directions. Contours of these \( F^P(k) \) are shown in Fig. 19a. In the second example below, we put \( P = 2Q = 8 \) and thus include an equal number of \( \sin \theta_k \) features as well. Contours of the second type of feature in (5.11) are shown in Fig. 19b. Together the two types of features can distinguish anisotropies along four different coordinate axes (i.e., \( x = 0, y = 0, x = \pm y \)).
For an anisotropic array design problem, the $\Delta I$ gradients are given correctly by (5.6)-(5.8). However, some of the component matrices in (5.7) differ from their isotropic forms when we use the features of (5.11). The covariance matrix $C_{rs}$ can be evaluated from (5.5) and $B_{rs}$ can be evaluated from (5.10), since $E$ is isotropic, but the anisotropic feature covariances are defined by

$$C^P(\Xi, \theta_{\Xi}) = \left\{ \begin{array}{ll}
- \cos \theta_{\Xi} \int_0^{\infty} E_I(K) F^P(K) J_2(KE) dK , & p = 1, \ldots, Q \\
and/or
- \sin \theta_{\Xi} \int_0^{\infty} E_I(K) F^{P-Q}(K) J_2(KE) dK , & p = Q+1, \ldots, P
\end{array} \right. \quad (5.12)$$

where $(\Xi, \theta_{\Xi})$ is a polar representation of the lag vector $\xi$. The matrices $C^P_{rs}$ in (5.7) are evaluated by (5.12) for $\xi rs \equiv x_r - x_s$. The final ingredient in the anisotropic problem is a replacement for the $A$ matrix in (5.9); namely,

$$A_{r(\alpha), w} = \frac{3}{\partial \xi_{\alpha}} C^P(x_{r(\alpha)} - x_{w})$$

$$= \left\{ \begin{array}{ll}
\cos \theta_{\Xi} \frac{\partial}{\partial \xi_{\alpha}} C^P(\Xi, \theta_{\Xi}) - \frac{\sin \theta_{\Xi}}{\Xi} \frac{\partial}{\partial \theta_{\Xi}} C^P(\Xi, \theta_{\Xi}) , & \alpha \text{ odd} \\
\sin \theta_{\Xi} \frac{\partial}{\partial \xi_{\alpha}} C^P(\Xi, \theta_{\Xi}) + \frac{\cos \theta_{\Xi}}{\Xi} \frac{\partial}{\partial \theta_{\Xi}} C^P(\Xi, \theta_{\Xi}) , & \alpha \text{ even}
\end{array} \right. \quad (5.13)$$

where $\Xi$ and $\theta_{\Xi}$ are evaluated for $\xi_{r(\alpha), w} \equiv x_{r(\alpha)} - x_{w}$.

Locally optimal array solutions (i.e., with $\Delta I$ a local maximum) are shown in Figs. 18d,e,f,g,h,i for configurations analogous to the previous isotropic solutions. The isotropic and anisotropic problems are identical except for the features: they are $F^P(k)$, $p = 1, \ldots, 4$, from (5.11) for Figs. 18a,b,c; $F^P(k)$, $p = 1, \ldots, 4$, from (5.11) for Figs. 18d,e,f; and $F^P(k)$, $p = 1, \ldots, 8$ from (5.11) for Figs. 18g,h,i.
The most dramatic distinction between the three types of solutions is the inversion of the array configurations in their $\Delta I$ ranking. For isotropic features the line is superior, for two axis anisotropy the cross is superior, and for four axis anisotropy the asterisk is superior. Furthermore, the discrepancies in $\Delta I$ increase for increasing anisotropy. For four axis anisotropy (Figs. 18g,h,i), the line is inferior to the asterisk by nearly two units in $\Delta I$. The primary reason for this large discrepancy is that the line is incapable of distinguishing between the two types of features in (5.11). For an orientation along a coordinate axis (the y-axis in Fig. 18i), no information at all is available about the second set of features in (5.11). (The angle $\theta_{\frac{\pi}{2}}$ has the values $\pm \pi/2$, $C^p = 0$ for $Q + 1 \leq p \leq P$ from (5.12), and $R_{pq} = 0$ for $Q + 1 \leq (p$ and/or $q) \leq P$.) The range of anisotropic array, near neighbor lag distances is similar to, but somewhat larger than, that for the isotropic arrays. The smallest lags are in Fig. 18d and are nearly 50% smaller than the smallest isotropic lags; the largest lags also occur in Fig. 18d and are only 6% larger than the largest isotropic lags.
6. Computational Aspects of Large Arrays

If there are $N$ observation points and $P$ spectral features, with $N \gg P$, the direct computation of the information matrix

\[ R_{pq} = \frac{1}{2} M \sum_{u,s}^{N} C_{rs}^{P} C_{ru}^{-1} C_{sv}^{-1} C_{uv}^{q} \]  \hspace{1cm} (6.1)\]

and of the auxiliary expression

\[ Z_{p} = \frac{1}{2} M \sum_{u,s}^{N} C_{rs}^{P} C_{ru}^{-1} C_{sv}^{-1} C_{uv}^{M} \]  \hspace{1cm} (6.2)\]

involves $O(N^3P)$ arithmetic operations (recall from (2.14) and (2.17) that $\lambda = R^{-1}Z$). Because of the factor $N^3$ this can be very expensive when $N$ is large, both in computer time and in intermediate storage. Furthermore, it would seem intuitively to be unnecessary, since in a large array widely separated segments are essentially independent of each other and contribute only additively to $R$ and $Z$.

We shall now discuss in general terms two approaches to reducing this rapid computational increase with $N$; the first is simple to implement but is not as powerful as the second. They illustrate the tradeoffs between many realizations from a small array and a few realizations from a large array containing many observation points. The same approximations are made in the expressions for both $R_{pq}$ and $Z_{p}$, which preserves the unbiased character of $R^{-1}Z$.

6.1 Cholesky factorization of the covariance matrix

The following is a discussion of Cholesky factorization, which plays a key role in most of the computations described in this paper. The Cholesky decomposition of a positive definite symmetric matrix $C$ is into a lower triangular matrix $L$ and its upper triangular transpose $L^T$, such that

\[ C = L L^T \]  \hspace{1cm} (6.3)\]
At each stage $u$ of the decomposition algorithm, ($u = 1, \ldots, N$, where $N$ is the dimension of $C$), $C$ is replaced by the $(N - u + 1) \times (N - u + 1)$ matrix

$$C^{(u)}_{rs} = C_{rs} - \sum_{l=1}^{u-1} L_{rv} L_{sv}; \quad r, s = u, \ldots, N. \quad (6.4)$$

Then the $u^{th}$ column of $L$ is defined by

$$L_{ru} = 0 \quad \text{if} \quad r = 1, \ldots, u - 1$$

$$= \frac{C^{(u)}_{ru}}{\sqrt{C^{(u)}_{uu}}} \quad r = u, \ldots, N. \quad (6.5)$$

Note that if $C$ is banded of width $m$ (i.e., if $C_{rs} = 0$ for all $|r - s| > m$), then $L$ is likewise strictly banded of the same width $m$. Because $L$ is triangular with non-zero diagonal elements, it is readily inverted, and

$$C^{-1} = (L^{-1})^T L^{-1}. \quad (6.6)$$

This is an efficient stable procedure for inverting positive definite matrices.

An important interpretation of this process comes from the remark that, because of (6.3), the random variable $\{\phi_r\}$ can be expressed by

$$\phi_r = \sum_u L_{ru} \varepsilon_u. \quad (6.7)$$

In (6.7), $\phi_r$ is a linear combination of independent random variables $\{\varepsilon_u\}$, each with zero mean and unit variance

$$\varepsilon_u \varepsilon_v = \delta_{uv}. \quad (6.8)$$

Thus (6.7) and (6.8) provide a model of the random process $\phi(x)$,
adequate to describe it completely at the observation points $x_r$. If
the model is extended to arbitrary $x$ by the definition

$$L_{xu} = \sum_{v} C_{xv} L_{uv}^{-1}, \quad (6.9)$$

then the interpolated function

$$\hat{\phi}_x = \sum_{u} L_{xu} c_u \quad (6.10)$$

is the least mean square error linear estimate for $\phi(x)$ given only the
information at the observation points (Gandin, 1965). Note that $L_{xu}$
is a continuous function of $x$, equal to the previous definition (6.5)
when $x = x_r$. Since $L_{uv}^{-1}$ vanishes unless $v \leq u$, $L_{xu}$ is a linear com-
bination of $C_{xu}$ and its $(u - 1)$ predecessors $C_{x1}, C_{x2}, ..., $ only--
indeed precisely that linear combination that vanishes at $x = x_1,$
$x_2, ..., x_{u-1}$. Thus the inclusion in the model of the random variable
$c_u$ leaves the representation at the previous $u - 1$ observation points
unaffected. If, as will be assumed hereafter,

$$x_r < x_{r+1}, \quad r = 1, 2, ..., \quad (6.11)$$

then for a sufficiently smooth covariance function $C(x)$, the function
$L_{xu}$ is small in the region $x < x_u$, peaks in $x_u \leq x \leq x_{u+m}$ and vanishes
if $x > x_{u+m}$. This property will be used later in interpolating large-
scale aspects of $\phi(x)$ to observation points that have been omitted in
the calculation as redundant.
There is also a correspondence between Cholesky decomposition and the orthonormalization of a sequence of vectors in a linear vector space. For any linear functional over \( \phi(x) \)

\[
\langle \alpha, \phi \rangle = \alpha(\phi) = \int \alpha(x)\phi(x) \, dx ,
\]

we define the norm

\[
||\alpha||^2 = (\alpha(\phi))^2 = \int \int C(x - y) \alpha(x) \alpha(y) \, dx \, dy .
\]

(6.12)

The space of the infinite dimensional vector \( \alpha \) is the class of all possible measurements of \( \phi(x) \). A basis set for such measurements is provided by samples at arbitrary points \( x_1, x_2, \ldots \), typified by the Dirac delta function

\[
\alpha(x) = \delta(x - x_r)
\]

so that

\[
\langle \alpha, \phi \rangle = \Sigma_r \alpha_r \phi_r .
\]

(6.13)

The norm is then

\[
||\alpha||^2 = \sum_{r,s} C_{rs} \alpha_r \alpha_s .
\]

(6.14)

(6.15)

Taking now a specific ordered sequence of points \( x_1, x_2, \ldots \), the basis vectors \( \delta(x - x_1), \delta(x - x_2), \delta(x - x_3), \ldots \) span a set of measurement subspaces \( V_1, V_2, V_3, \ldots \) of increasing dimension, each including its predecessor. However, in general, these basis vectors are not orthogonal according to the norm as defined. A classical problem in linear algebra is to derive a basis sequence \( \{e_u, u = 1, 2, \ldots \} \) spanning the
same subspaces, that is orthonormal. The solution comes by subtracting at each stage \( u \) from the vector \( \delta(x - x_u) \) just that part that is parallel to some vector in \( V_{u-1} \) and then normalizing the result. It is easily seen that this yields

\[
e_u(x) = \sum_{v \leq u} \mathbf{L}^{-1}_{uv} \delta(x - x_v),
\]

(6.16)
i.e., the vector with the \( u \)th row of \( \mathbf{L}^{-1} \) as its oblique components.

These interpretations show how the diagonal term

\[
\sigma^2_u = C_{uu} = (\hat{\phi}_u - \bar{\phi}_u)^2
\]

(6.17)
that arises at each stage of the decomposition algorithm describes the new information required to define \( \phi_u \) over and above that already available from the values \( \phi_1, \ldots, \phi_{u-1} \). It is thus inherently positive, but if \( x_u \) is very close to \( x_{u-1} \), \( \sigma_u^2 \) may be much smaller than \( C_{uu} \). In this case, omission of the complete rows and columns involving point \( u \) in the decomposition provides a model (6.9) for \( \phi(x) \) with fewer factors \( \mathbf{L}_{xy} \) than there are observation points, but with a good interpolated value \( \hat{\phi}_x \) at \( x = x_u \) instead of the observation \( \phi_u \). Such a modified procedure yields an approximate factorization

\[
C_{xy} = \sum_{u=1}^{n} L_{xu} L_{yu}, \quad (n < N)
\]

(6.18)
which we denote by

\[
C = \mathbf{L} \bowtie \mathbf{L}^T.
\]

(6.19)
The brackets in the matrix representation (6.19) are a reminder that the number of columns in \( \mathbf{L} \) is less than the number of rows.
Another point worthy of comment is that the $u^{th}$ row of $L^{-1}$ has the structure

$$L_{ur}^{-1} = - \frac{1}{\sigma_u} \beta_{ur} \quad r < u$$

$$= \frac{1}{\sigma_u} \quad r = u$$

$$= 0 \quad r > u$$

(6.20)

where

$$\beta_{ur} = \sum_{s < u} C_{us} C_{sr}^{-1}$$

(6.21)

is the coefficient multiplying $\phi_r$ in the optimal estimate $\hat{\phi}_u$ of $\phi_u$ from its predecessors, and

$$\sigma_u^2 = C_{uu} - \sum_{r,s < u} \beta_{ur} \beta_{us} C_{rs}$$

(6.22)

is the residual variance after that estimate. Note that in (6.21) $C_{sr}^{-1}$ is the inverse of the $(u - 1) \times (u - 1)$ matrix $C_{rs}$, $r,s < u$, and is not the $(s,r)$ element of the inverse of the complete matrix $C$.

Now if $C_{rs}$ is banded of width $m$ and $\sigma_u$ is not small, the only way $L_{ur}^{-1}$ can be significantly non-zero for $|r - u| >> m$ is if there exists an $s$ for which both

$$u - m \leq s \leq u$$

and $C_{sr}^{-1}$ is substantial. However, for such an $s$ inevitably

$|s - u| >> m$. Since, if $L^{-1}$ is banded, so by equation (6.6) is $C^{-1}$, it would appear consistent to assert that:

If $C_{rs}$ is a symmetric matrix, banded of width $m$, and strongly positive definite in the sense that the residuals
the elements \( C_{rs}^{-1} \) become very small for \(|r - s| >> m\).

The authors are unaware of any proof of this assertion, other than the experience provided by many calculations. It is, however, of major practical importance in this theory, providing the basis for both approaches to an approximate calculation of \( R, Z \) when \( N \) is large.

6.2 Near neighbor array truncation

In the first approach, suppose that there is an interval \( \xi_o \) such that both the first guess covariance function \( C(\xi) \) and all the feature covariances \( C^p(\xi) \) vanish when \( \xi > \xi_o \). The value of \( \xi_o \) is, of course, controlled by the finest spectral resolution under consideration.

Then there is certainly some value of \( m \) for which \(|x_{r+m} - x_r| > \xi_o\) for all \( r = 1, \ldots N \), and hence the covariance matrix \( C_{rs} \) and feature matrices \( C^p_{rs} \) are banded of width \( m \). If the array is very long \((N >> m)\) and \( m \) is reasonably small, the calculation proceeds quite efficiently by replacing the approximately banded \( L^{-1} \) in (6.20) by the strictly banded lower triangular matrix \( L^{*-1} \), where

\[
L^{*-1}_{ur} = \begin{cases} 
0 & r < u - m^* \\
- \frac{1}{\sigma_u^*} \beta_r^* & u - m^* \leq r < u \\
\frac{1}{\sigma_u^*} & r = u \\
0 & r > u 
\end{cases}
\]  

(6.23)

Here \( \beta_u^*, \sigma_u^* \) are the values of (6.21), (6.22) when \( r, s \) are restricted to the range

\[
u - m^* \leq r, s < u
\]  

(6.24)
and \( m^* \) is some suitable multiple of \( m \) (say \( m^* = 1.5m \)). Thus the estimate \( \hat{u}_u \) is evaluated only from the \( m^* \) nearest predecessors, rather than from all \( u - 1 \) predecessors. Because of the bandedness of \( C \) and the preceding assertion about \( C^{-1} \), \( \hat{u}_u \) is quite insensitive to the addition of distant data. Even if \( m^* \) is only slightly larger than \( m \), this approximation is usually a very good one. \( R_{pq} \) is obtained by multiplying together the strictly banded matrices

\[
L^{*\perp} C^p(L^{*\perp})^T \quad \text{and} \quad L^{*\perp} C^q(L^{*\perp})^T
\]

and taking the trace \( Z_p \) follows in a similar manner. In this approximation, the only data points contributing to \( R_{pq} \) in equation (1) are those sets \( (r,s,u,v) \) for which

\[
|r - s| \leq m, \quad |u - v| \leq m^* \\
|r - u| \leq m^*, \quad |s - v| \leq m^*
\]

(6.25)

Thus segments in the array separated by more than \( m + m^* \) intermediate points are effectively independent, contributing additively to expressions (6.1) and (6.2). For a very large array, the important property of array design is the joint probability distribution of lags \( |x_r - x_s| \) among sets of points satisfying the conditions (6.25), as sampled by the many such sets as the point \( u \) moves from 1 to \( N \). A crude analysis of array properties was given in Section 4, based upon the probability distribution of individual nonredundant lags rather than upon the joint distribution of \( 1/2 \ (m + m^*) (m + m^* + 1) \) relevant lags defined by (6.25).
The procedure just outlined is straightforward to apply, and when $N >> m$ it can result in very substantial computational savings. It does, however, have disadvantages when the array $\{x_r\}$ and the features $C^P(\xi)$ have a large dynamic range, with some regions of closely spaced data points efficiently sampling the small-scale aspects of $\psi(x)$, and with other regions of quite widely separated data points covering the larger scales. The interval $\xi_o$ tends to be set by the largest scale of interest in the spectrum. It may then contain, in some places, a large number of observation points, requiring the band width $m$ to be correspondingly large. The redundancy in expression (6.1) arises because, when considering small scales, the array may be divided into effectively independent segments, whose contribution to $R$ is dominated by elements of $C^{-1}$ that are close to the diagonal, whereas, for the larger scales, the densely sampled regions provide a more than adequate representation from which many observations could be omitted. The objective of the second approach is to find a representation for $C^{-1}$ that clarifies this scale separation.

6.3 Scale dependent array redefinition

This second approach depends on the Cholesky factorization of individual features

$$C^P_{rs} = \sum_u s^P_{ru} s^P_{su}$$

(6.27)

where the index $u$ labeling the individual factor runs over a relevant subset (to be defined later) of the observation points. Then equation (6.1) shows that

$$R_{pq} = \frac{1}{2} M \sum_{u,v} (C^P_{uv})^2$$

(6.28)
where
\[
G_{pq}^{uv} = \sum_{r,s=1}^{N} C^{-1}_{rs} s^p r^q s^v
\]  
(6.29)

is a key matrix element connecting the distinct factors identified by the pairs \((p,u)\) and \((q,v)\). Supposing for simplicity that the first guess covariance function can be written as the sum of the feature covariances
\[
C(\xi) = \sum_{l=1}^{P} C^P(\xi)
\]  
(6.30)

the factorization (6.27) is equivalent to replacing \(\phi(x)\) by the model
\[
\phi(x) = \sum_{u,p} s^p x^u \varepsilon^p_u
\]  
(6.31)

where each of the \(\varepsilon^p_u\) is independently distributed with zero mean and unit variance. Given a set \(\{\varepsilon^q_v\}\) which determines \(\phi(x)\) at the observation points, the best estimate (least mean square error) for the \(\varepsilon's\) that can be recovered from the \(\{\phi_r\}\) is
\[
\hat{\varepsilon}^p_u = \sum_{q,v} G_{pq}^{uv} \varepsilon^q_v
\]  
(6.32)

and the residual covariance in that estimate is
\[
(\varepsilon^p_u - \hat{\varepsilon}^p_u)(\varepsilon^q_v - \hat{\varepsilon}^q_v) = s^p_{uv} - C^p_{uv}
\]  
(6.33)

Thus the diagonal element \(G_{uu}^{pp}\) describes how well the individual coefficient \(\varepsilon^p_u\) can be measured, being unity for a perfect estimate, and the more general element describes the cross aliasing between the pairs \((p,u)\) and \((q,v)\). The information matrix \(R\) is the sum of squares of these elements.
To evaluate the matrix elements (6.29) efficiently, we need a representation of $C^{-1}_{rs}$ that is approximately banded and matched to the length scales of $S^P_{xu}$ and $S^Q_{xv}$ with redundant sample points eliminated. To do this we group the terms on the right-hand side of (6.31) into a sequence of levels

$$
\phi(x) = \psi_1(x) + \psi_2(x) + \ldots + \psi_m(x)
$$

(6.34)

arranged in order of increasing scale. Each $(p,u)$ is associated with a definite level, $\lambda(p,u)$, and the terms associated with a given level $\lambda$ are consolidated into a single factor for each $u$

$$
\psi_\lambda(x) = \sum_{u} L^\lambda_{xu} c_u
$$

(6.35)

where

$$
\overline{\psi_\lambda(x) \psi_\lambda(y)} = \sum_{u} L^\lambda_{xu} L^\lambda_{yu} = \sum_{\lambda(p,u)=\lambda} S^P_{xu} S^P_{yu}.
$$

(6.36)

The function of $\lambda(p,u)$ will be chosen so that the band width over non-redundant sample points of the consolidated factor $L^\lambda_{xu}$ is constant and equal to $m$, and so that a fairly uniform and controlled fraction of the sample points $u$ at level $\lambda$ is eliminated as redundant at level $(\lambda+1)$. This leads to an efficient representation, with cross-aliasing between levels reduced to a minimum.

At any level $\lambda > 1$, the accumulated contributions of smaller scale

$$
\Theta_{\lambda-1}(x) = \sum_{\lambda'-1}^{\lambda-1} \psi_{\lambda'}(x)
$$

are effectively "noise," with covariance
\[ E^\hat{\ell} = \theta_{\hat{\ell}-1}(x_r) \theta_{\hat{\ell}-1}(x_s) \]
\[ = \sum_{\ell'=1}^{\ell-1} \left\{ \sum_u L_{ru} L_{su}' \right\} . \quad (6.37) \]

Then, in the notation of (6.19):

\[ E^\ell = E^{\hat{\ell}-1} + L^\ell \gg L^{\hat{\ell}T}, \quad (6.38) \]

and, as may be verified by direct substitution,

\[ (E^\ell)^{-1} = (E^{\hat{\ell}-1})^{-1} - (E^{\hat{\ell}-1})^{-1} L^\ell \gg H^\ell < L^{\ell T}(E^{\hat{\ell}-1})^{-1} , \quad (6.39) \]

where

\[ H^\ell = (I + L^{\ell T} (E^{\hat{\ell}-1})^{-1} L^\ell)^{-1} \quad (6.40) \]

If \( n_\ell \) is the number of nonredundant sampling points at level \( \ell \), \( H^\ell \) is an \( n_\ell \times n_\ell \) matrix, whereas \( E^\ell \) is in principle \( N \times N \). Equations (6.39) and (6.40) provide an important recurrence relation for \( (E^\ell)^{-1} \) in terms of \( (E^{\hat{\ell}-1})^{-1} \). This is particularly valuable in the contracted form

\[ (E^\ell)^{-1} L^\ell \gg = (E^{\hat{\ell}-1})^{-1} L^\ell > H^\ell . \quad (6.41) \]

These equations show how at each level the noise \( \theta_{\ell-1}(x) \) mitigates against the effective discrimination of the signal \( \psi_\ell(x) \). A measure of this discrimination is provided by the matrix element

\[ \langle (L^\ell)^T (E^\ell)^{-1} L^\ell \rangle = I - H^\ell . \quad (6.42) \]

When \( L^\ell \gg L^{\ell T} \) is large compared to \( E^{\hat{\ell}-1} \), the signal to noise ratio is good and \( H^\ell \) is small compared to the unit matrix. This second approach is useful only when \( H^\ell \) is small for each \( \ell \), so that the
discrimination of the distinct levels $\psi_\lambda(x)$ is close to being perfect.
This will be so if at each level the matrix

$$U^\lambda,^l,^l+1_> = (L^\lambda)^{-1} L^l+1_> \quad (6.43)$$

is large; we designate its order as $U \gg 1$. In any case, $H^\lambda$ can be estimated from the iterative formula

$$H^l+1 = (I + <U^l+1,^l (I - H^l) U^l,^l+1_>)^{-1}, \quad (6.44)$$

showing that $H = O(1/U^2)$ for large $U$. Note that the scales larger than those of level $\lambda$ are not included in the discrimination matrix (6.42), either as signal or noise.

Before proceeding to evaluate $C^{-1}$ for use in (6.29), we must establish an appropriate hierarchy of sampling points $\{x_u\}$ and explain how the attribution $\lambda(p,u)$ of a level to each factor $S^p_{x,u}$ is arrived at. For each $p$ define a scale $\xi_p$ such that

$$C^p(\xi) = 0 \quad \text{if } |\xi| > \xi_p, \quad (6.45)$$

and suppose the features so ordered that

$$\xi_1 < \xi_2 < \cdots < \xi_p < \cdots < \xi_p.$$ 

One might formally use this separation by features to define the levels, but if the array has wide dynamic range, with very disparate adjacent lags $x_{r+1}-x_r$, this tends to be inefficient, leading to poor discrimination between the $\psi_\lambda(x)$. Instead we include in $\psi_\lambda(x)$ only those parts of $\{\phi_r\}$ that can be described by a model of bandwidth $m$, mixing items from different features where necessary.
For \( l = 1, 2, \ldots, m \), we perform the following two-step iteration. In step 1, for a suitable small parameter \( \delta \), designate as redundant at level \( \ell \) any sample point \( x_r \) included at level \((\ell-1)\) for which

\[
0 < x_r - x_u < \delta \xi_{\text{Pmin}}(u,\ell)
\]

(6.46)

where \( u \) is the last previous point not designated as redundant. For \( \ell = 1 \), \( \text{Pmin} = 1 \) for all \( u \). In step 2, for each \( u = 1, \ldots, n_\ell \) in the nonredundant set let \( \text{Pmax}(u,\ell) \) be defined as the largest \( p \) for which

\[
x_{u+m} - x_u > \xi_p
\]

(6.47)

This will define locally in \( u \) the largest scale to be included in \( \psi_\ell(x) \). Then let \( \text{Pmin}(u,\ell+1) = \text{Pmax}(u,\ell) + 1 \) and return to step 1 until the features are exhausted. For each nonredundant point \( u \) the level \( \ell(p,u) \) is ascribed to the feature \( p \) if

\[
\text{Pmin}(u,\ell) \leq p \leq \text{Pmax}(u,\ell)
\]

The typical economy in sampling points in passing from level \( \ell \) to \( \ell+1 \) by this procedure may be estimated by remarking that in the internal \( \xi_{\text{Pmax}}(u,\ell) \) lie approximately \( m \) sample points, whose average spacing is therefore \( \xi_{\text{Pmax}}(u,\ell)/m \). On the other hand, the minimum nonredundant spacing at level \((\ell+1)\) is \( \delta \xi_{\text{Pmax}}(u,\ell)+1 \). Hence the reduction factor is approximately

\[
m\delta (\xi_{P+1}/\xi_p)
\]

(6.48)

which can be made large by a suitable choice of \( m \) and \( \delta \).

Next, note that for each feature \( p \) there is now a relevant set of nonredundant sample points \( x_u \), though at various levels. The Cholesky factorization (6.27) thus proceeds according to the standard
algorithm (6.4)-(6.5) for all points \((r,s)\) in the relevant set. There-
after (6.9) is used to interpolate \(S^P_{xu}\) to those points that are redundant
at level \(\ell\) but not at \((\ell-1)\). Summing over \(p\) between \(p_{\text{min}}\) \((\ell,u)\) and
\(p_{\text{max}}\) \((\ell,u)\) to obtain \(\psi _{\ell}(x)\psi _{\ell}(y)\) according to (6.36) and then refactoring
and interpolating, we obtain the consolidated factors \(L^\ell_{xu}\) labeled by
each nonredundant point \(u\) at level \(\ell\). This apparently complex procedure
is necessary to ensure that each covariance factor \(S^P_{xu}\) is included in
one and only one consolidated factor \(L^\ell_{xu}\). This procedure involves
substantial manipulation, but the matrices involved are all strictly
banded so the computation is limited by the bandwidth. There is an
approximation (controlled by the magnitude of the parameter \(\delta\)) implied
by writing \(C^P_{rs}\) in the form (6.27) with fewer factors \(u\) than the original
observation points \(r,s\). If \(\delta\) is small, this error is very slight and
the total covariance implied by the consolidated factors closely
approximates the first guess matrix \(C^P_{rs}\).

The procedure now is to evaluate \(U^{1,2,3,...}\), etc., according to
equation (6.43) with \((L^\ell)^{-1}\) replaced by the strictly banded approxi-
mation \((L^\ell)^{-1}\) defined in equation (6.23). Note that \(L^\ell\) here is the
strictly triangular matrix based on the sample points \(x_r\) which are
nonredundant at level \(\ell\), but \(U^{\ell,\ell+1}\) is not triangular, having \(n_\ell\) rows
but only \(n_\ell+1\) columns. Thus \(U^{\ell,\ell+1}\) is banded in the sense that each
\(u\) is associated nontrivially only with those \(r\) that lie between \(u\) and
\(u+m\). Then, starting from \(H^1 = 0\), evaluate \(H^2, H^3, ..., H^m\) according
to (6.44). If each \(H^\ell\) is in fact small, the significant elements
are confined to a band of width \(m\) about the diagonal, controlled by
the elimination of each stage of redundant sampling points.
The reconstruction of $\C^{-1}$ proceeds in the reverse order, concentrating upon the matrix element $\langle (L^\ell)^T \C^{-1} L^\ell \rangle$ which measures how well $\psi_\ell(x)$ can be discriminated from the noise $\theta_\ell(x)$ and from larger scales $\psi_{\ell+1}(x)$. Defining

$$J_{\ell m} = H_{\ell m}$$

and

$$J_{\ell} = H^\ell + H^\ell U_{\ell,\ell+1} > J_{\ell+1} < U_{\ell+1,\ell} H^\ell,$$  \hspace{1cm} (6.49)$$

for $\ell = \ell_m - 1, \ldots, 1$, and making use of equation (6.42), it follows from equation (6.39) that

$$\langle (L^\ell)^T \C^{-1} L^\ell \rangle = (I - H^\ell) - (I - H^\ell) U_{\ell,\ell+1} J_{\ell+1} < U_{\ell+1,\ell} (I - H^\ell),$$  \hspace{1cm} (6.50)$$

and

$$\langle (L^{\ell-1})^T \C^{-1} L^{\ell-1} \rangle = (I - H^{\ell-1}) U_{\ell-1,\ell} > H^\ell (I - U_{\ell,\ell+1} J_{\ell+1} < U_{\ell+1,\ell} (I - H^\ell)), $$  \hspace{1cm} (6.51)$$

These formulae provide important information about the structure of the matrix $\C^{-1}$, particularly when the spectrum is strongly red. Then $U_{\ell,\ell+1}$ is large, of order $U$, and $U_{\ell,\ell+1} H^{\ell+1}$ is small, of order $(1/U)$, so that

$$J_{\ell} \approx H^\ell \approx (U_{\ell+1,\ell} U_{\ell,\ell+1})^{-1}.$$  \hspace{1cm} (6.52)$$

If $\psi_\ell(x)$ were perfectly measurable, the matrix element $\langle (L^\ell)^T \C^{-1} L^\ell \rangle$ would equal the $n_\ell \times n_\ell$ unit matrix $I$. Because of small-scale noise, however, it is reduced slightly and broadened in bandwidth to $I - H^\ell$. However, the last term in (6.50) describes a major reduction. Indeed, because of it (ignoring terms of order $1/U^2$)

$$\langle (L^\ell)^T \C^{-1} L^\ell \rangle \approx I - U_{\ell,\ell+1} (U_{\ell+1,\ell} U_{\ell,\ell+1})^{-1} U_{\ell+1,\ell},$$  \hspace{1cm} (6.53)$$
which implies that any function \( \psi_\lambda(x) \) that can be written in the form \( L_x^\lambda U^\lambda \) is essentially unmeasurable. This result is scarcely surprising for such functions belong to the class \( \psi_{\lambda+1}(x) \) which has been separated off. The cross term \( \langle (L^{-1})^T C^{-1} L^\lambda \rangle \) is seen from equation (6.51) to be dominated by the factor \( U^{\lambda-1,\lambda} \gg H^\lambda \), which is \( O(1/U) \) and moderately small. More distant cross terms such as \( \langle (L^{-2})^T C^{-1} L^\lambda \rangle \) contain additional factors \( U^{\lambda-2,\lambda-1} \) and are even smaller. Thus, when \( U \) is large, the functions \( \psi_\lambda(x) \) can be well measured from this array, and small limited corrections suffice to determine the interference or aliasing from other levels. Note that the error covariances \( E^\lambda \) and \( C = E^\lambda m \) do not appear explicitly in (6.50) and (6.51). The use of matrix elements selects the appropriate sampling points for the level under consideration. Of central importance from a practical point of view is that when \( U \) is large the matrix elements at each level \( \lambda \) are approximately banded, and that cross terms between adjacent levels are small, with cross terms between more widely separated levels negligible.

Now evaluate the feature matrix elements \( G_{pq}^{uv} \) in equation (6.28) according to

\[
G_{pq}^{uv} = \sum_{\lambda, \lambda'} \left( (L^\lambda)^{-1} s^p_u \right)^T (L^\lambda T C^{-1} L^{\lambda'}) \left( (L^{\lambda'})^{-1} s^q_v \right) . \tag{6.54}
\]

The zero order approximation for large \( U \) is particularly simple. Then all elements for which \( \lambda' \neq \lambda \) can be ignored and

\[
R_{pq} \approx \sum_{\lambda} \sum_{u,v} \left| \left( (L^\lambda)^{-1} s^p_u \right)^T (L^\lambda)^{-1} s^q_v \right|^2 , \tag{6.55}
\]
where the summation is over those $u,v$ for which

$$\lambda(p,u) = \lambda(q,v) = \lambda.$$

The effectively complete separation between levels does not necessarily lead to $R$ being diagonal. Cross aliasing between features occurs within each level because of the irregular array spacing. To the next approximation in $1/U$, (6.50) involves banded corrections of order $H = 1/U^2$, and (6.51) can be adequately described by

$$<(L^{-1})^TC^{-1}L^k > \approx (\frac{U^{-1} - 1}{U} L) H^k (I - U^k, U^{k+1})^{-1} <U^k \lambda+1, U^k \lambda>$$

(6.53)

with cross terms separated by more than one level being ignored.

Although expression (6.53) is $O(U^{-1})$, it appears squared in $R_{pq}$ and hence contributes at $O(1/U^2)$.

It will be seem from this discussion that this second approach implies two different types of approximation. The first is a geometrical one associated with modeling a feature covariance by a limited number of factors, interpolating to intermediate sample points that are nonredundant one level down. The grouping of different features into a single level is controlled by the bandwidth $m$, through equation (6.48) and the scale $P_{max}^{(u,\lambda+1)}$ which can just accommodate $m$ thinned out sample points at level ($\lambda+1$). The second type of approximation is the simplification of expressions (6.50), (6.51) and the neglect of most of the matrix element $G_{pq}^{(u,v)}$ that contribute to $R_{uv}^{(u,v)}$.

The crucial point for this to be valid is that the consolidated factors should have $L_{xu}^{\lambda+1} > L_{xu}^\lambda$. This appears to require $C_{xu}^{p+1}(0) > C_{xu}^p(0)$; that is, the first guess spectrum defined by equation (6.33) must be
strongly red. The precise connection with the previous condition is unclear. An obvious question arises when the first guess spectrum is nearly white, yet the feature dynamic range is so great that the first approach is very cumbersome. It would appear that then there is no major economy available. The matrices (6.50), (6.51) are not limited to elements close to the diagonal and there is no sense in which $C^{-1}$ is narrowly banded. The problem lies in the use of some features with a large reach $\xi_p$, without the ability to separate scales inherent in a red spectrum. To achieve such a high degree of spectral resolution, even for one feature, requires a full treatment of the inverse $C^{-1}$ and a large amount of calculation. An obvious remedy is to relax the spectral resolution requirements, permitting a smaller interval $\xi_0$ and band limit $m$. 
7. **Comparison with Other Methods**

According to Davis and Regier (1977), there are three basic types of methods for estimating spectra from a given array: a priori, data adaptive, and model fitting. We now present arguments that, although philosophically these methods may look very different, in the circumstances of most interest there is much in common between them and they can be subjected to quantitative intercomparison.

7.1 A priori estimates

In the first (a priori) approach, the spectral estimate $\hat{E}(k)$ is a linear combination of the pairwise covariances $\overline{C}_{rs}^M$ estimated at pairs $(r,s)$ of observation points from $M$ realizations of the data:

$$\hat{E}(k) = \sum_{r,s} \alpha_{rs}(k) \overline{C}_{rs}^M,$$  \hspace{1cm} (7.1)

where the $\alpha_{rs}$ are chosen independently of the observations. If the number of realizations $M$ is large, then $\hat{E}(k)$ approximates the ensemble mean value $\overline{E}(k)$, which is related to the true spectrum by

$$\overline{E}(k) = \int \overline{w}(k,l) \overline{E}(1) \, dl,$$  \hspace{1cm} (7.2)

where

$$\overline{w}(k,l) = \sum_{r,s} \alpha_{rs}(k) \cos \left( \frac{k}{r-s} \right)$$  \hspace{1cm} (7.3)

is a 'window function' that determines the resolution and aliasing of the estimate. There are various ways of selecting the coefficients $\alpha_{rs}$, either by deriving them from the output of a linear filter acting on the process $\phi(x)$ so that they have the form...
\( \alpha_{rs} = \beta_r e^{\frac{ikx_r}{s} - \frac{ikx_s}{s}} \) (7.4)

(e.g., Thompson, 1971) or by optimizing to make \( \hat{W}(k,l) \) resemble as closely as possible for each \( k \) a delta function with a narrow peak around \( l = k \) and side lobes kept to a minimum (Davis and Regier, 1977).

Another example in this class of spectral estimators is the interpolation method in which \( \phi(x) \) is interpolated between data points then the spectrum of the resulting continuous function is estimated in the regular manner.* A further example is the 'bin method' of Julian and Kline (1974), where a smooth curve is fitted by least squares to \( C_{rs}^M \) regarded as a function of \( \xi = |x_r - x_s| \), the result being Fourier transformed to give \( \hat{E}(k) \).* Two properties all these examples have in common are the following: a definite prescription is given for \( \alpha_{rs}(k) \), and the function \( \hat{E}(k) \) is the outcome of the analysis, with the inversion for the spectrum \( \hat{E}(k) \) from

\[
\hat{E}(k) = \int \hat{W}(k,l) \hat{E}(k) \, dl \quad (7.5)
\]

treated qualitatively or left tacit.

7.2 Data adaptive estimates

In a data adaptive scheme such as the maximum likelihood method of Capon (1969), the estimate is

\[
\hat{E}(k) = \frac{1}{\sum_{rs} e^{\frac{ikx_r}{s} - \frac{ikx_s}{s}}} \left( \sum_{rs} \frac{ikx_r}{s} \Delta_{rs}^M - \frac{ikx_s}{s} \right) \quad (7.6)
\]

*Refer to Section 2.6 for more discussion of these methods.
involving the inverse of the matrix $\widehat{\mathbf{C}}^M$ of raw covariance estimates. An estimate of this type is highly non-linear in the raw covariance estimates, but it can formally be expressed in the equations of Section 7.1, where the coefficients $\alpha_{rs}(k)$ are themselves functions of the data being analyzed. They can again be derived as the output power spectral density of a certain linear filter formed by the coefficients (see (7.4))

$$
\beta_r(k) = \gamma(k) \sum_u e^{ikx_u} \left( \widehat{\mathbf{C}}^M \right)_{ur} e^{-ikx_r},
$$

(7.7)

where $\gamma(k)$ is a normalization factor chosen to make

$$
\sum_r \beta_r(k) = 1.
$$

(7.8)

It is readily shown that, subject to the idealization of replacing $\widehat{\mathbf{C}}^M$ by $\widehat{\mathbf{C}}$ in (7.6), such a choice of coefficients minimizes the ensemble average of the filter output power at each wavenumber when it is applied to $\phi(x)$, subject to the constraint (7.8). The latter ensures that when the filter is applied to a narrow bandwidth signal of definite wavenumber the amplitude transmitted at the wavenumber is unaltered. The physical interpretation of this constraint is a little unclear, but it implies that $\hat{E}(k)$ is always an overestimate of the true spectrum $\overline{E}(k)$, though to some minimal extent consistent with the observational array and the spectrum being observed. An alternative, the data adaptive spectral estimate, has been proposed by Davis and Regier (1977), in which the constraint is applied to the weighted integral of the power of an input signal rather than to a specific Fourier amplitude. Again, however, the procedure for inferring $\hat{E}(k)$ from $\hat{E}(k)$ is treated at best qualitatively.
7.3 Model fitting

In the third type of approach the inference of $\overline{E}(k)$ is dealt with directly by assuming an analytic form for the answer and then choosing the undetermined coefficients so that the implied covariance best fits the data. A good example is the work of Jones (1977) in which all values of the undetermined parameters are assumed to be equally likely a priori, and, given a particular data set, the most likely a posteriori values are computed exactly. This type of approach has been criticized on the grounds that the outcome is largely determined by the initial assumptions about the form of the spectrum and that the true spectrum is almost certainly not of this form. Recall that, in Section 2, it was shown how several different model fitting procedures are essentially equivalent and others are not as accurate.

7.4 A unified analysis of methods

In discussing these approaches, we will assume that the amount of data available, specifically the number of realizations $M$, is large but not infinite; that indeed the investigator has formed a reasonable first guess $\hat{E}(k)$ about the spectrum $\overline{E}(k)$ of the process under investigation, either from previous experience or from a preliminary crude look at the data itself; that the convention has been adopted that $\overline{E}(k)$ is in some sense smooth except insofar as the data implies otherwise; that for the purposes of intercomparison of sampling errors between approaches the process $\phi(x)$ may be assumed Gaussian, or at least with a known relationship between fourth-order cumulants and second-order moments; and finally that considerations of computational
expense are of secondary importance. In a qualitative sense, these assumptions are probably appropriate in most cases of interest. Their quantitative validity depends on the accuracy in detail of the specific approximations made in deriving the standard formulation below. Thus considerable judgment is still required in selecting the most appropriate method.

The formalism developed in Section 3 enables all three approaches to be discussed in a single framework. Central is the approximate calculation of the relative spectrum

\[ \hat{v}(k) = (\hat{E}(k) - \bar{E}(k))/(\hat{E}(k)) \] (7.9)

which is implied for a given sample of data from M realizations of \( \phi(x) \) by inverting equation (7.5) for both the a priori and data adaptive methods and from equation (2.12) for the model fitting method. The inversion process is necessary to remove the systematic biases present in the estimators \( \hat{E}(k) \), and to establish a common variable for intercomparison. The accuracy should be such that \( \hat{v}(k) \) differs from the true correction \( \bar{v}(k) \) by an amount that is small compared to the random fluctuations

\[ v'(k) = \hat{v}(k) - \bar{v}(k) \]

To assist the calculation we also assume that \( \hat{v} \) is small compared to unity, so that only lowest order terms in an expansion in powers of \( v \) need be retained. This requires both the first guess \( \hat{E} \) be a good estimate of the mean \( \bar{E} \), and that the fluctuations \( E' \) about \( \bar{E} \) are fairly tightly constrained by the data (as they must be when M is
large). A key quantity is the matrix $H(k,l)$ defined as the inverse of the covariance matrix $\bar{v}'(k)\bar{v}'(l)$ of the sampling fluctuations. An important technical difficulty is that the inversion of equation (7.5) is an ill-posed problem, so $\bar{v}'(k)\bar{v}'(l)$ is not well defined. However, this shortcoming does not apply to its inverse $H$, which measures the complete information about $E(k)$ provided by the analysis procedure.

Now the functions $v(k)$ are all continuous functions of $k$, but to avoid unnecessary complexity we shall suppose as in Section 3 that the wavenumber domain is sampled discretely and finitely at a spacing $\Delta k$ fine enough to be no practical constraint, and we shall describe $v(k)$ by the finite dimensional vector $v = \{v_k\}$.

We now evaluate the constraints on $\bar{v}$ implied by the existence of a set of observations, bearing in mind that there is a random sampling error present in any particular set of $M$ realizations. For the a priori methods, the coefficients $\alpha_{rs}(k)$ are definitely known. The removal of bias then proceeds without approximation by inversion of equation (7.5); i.e.,

$$\hat{v}_k = \frac{1}{E_k} \sum_1 (W^{-1})_{kl} (E_{1l} - E_{1l}'),$$

(7.10)

where $W^{-1}$ is the inverse of the matrix $W$, insofar as it exists. $\hat{v}$ may be regarded as a random variable, fluctuating about a definite mean $\bar{v}$ with a joint normal probability distribution. Since equation (7.9) is linear, $\bar{v}$ differs from $v$ only by an arbitrary vector $v_0$ in the null-space of $W$, the intrinsic alias of the analysis. From equation (7.3) it is apparent that this null-space includes at least all those spectral variations that have no impact on the covariance matrix $C_{rs}$.
for any pair \((r,s)\) of observation points. Also the \textit{range} of \(\hat{W}\), the
set of all \(E - \hat{E}\) that can be generated by some \(v\), is a vector space \(U\)
with finite dimension at most \(1/2 N(N+1)\), with one dimension for each
of the distinct linearly independent functions \(\alpha_{rs}(k)\). A fluctuation
\(\nu'\) is associated with a unique estimator sampling error in \(U\):

\[
E' = E - \bar{E} = \hat{W}(\nu')
\]

(7.11)

From equation (2.22) the covariance matrix of such errors is

\[
\frac{E'E}{1} = \frac{1}{M} Q_{kl} = \frac{1}{M} \sum_{\hat{r},\hat{s} \in \hat{V}} \alpha_{rs}(k) \left( \hat{C}_{ru} \hat{C}_{sv} + \hat{C}_{rv} \hat{C}_{su} \right) \alpha_{uv}(1)
\]

(7.12)

Here, in accordance with the assumption that \(\hat{V}\) is small, \(\hat{C}\) has been
replaced by \(\hat{C}\). \(Q_{kl}\) is a positive definite matrix with inverse \(Q^{-1}\) in
the space \(U\). Thus, since the fluctuations \(E'\) are joint normally dis-
tributed, so are the \(\nu'\), with probability distribution proportional

to

\[
\exp \left\{ - \frac{M}{2} (\nu - \hat{\nu})^T H (\nu - \hat{\nu}) \right\}
\]

(7.13)

where

\[
H_{kl} = \sum_{\hat{k}',\hat{l}'} \hat{E}_{k'} \hat{W}_{k'k} \left( Q^{-1} \right)_{k'1'} \hat{W}_{1'1} \hat{E}_{1}
\]

(7.14)

and \(\hat{\nu}\) is defined by

\[
\hat{E}_{k} - \hat{E}_{k} = \sum_{1} \hat{W}_{kl} \left( \hat{E}_{1} \hat{\nu}_{1} \right)
\]

(7.15)

Note that all quantities here are well defined, though \(\nu\) and \(\hat{\nu}\) are
unconstrained by (7.13) and (7.15) to the extent of an arbitrary
additive vector \(\nu_0\) from the null-space of \(W\). The fundamental matrix \(H\)
measures the total information available from the estimator \(E(k)\),
and is the quantitative basis for intercomparison studies.
For data adaptive methods the sensitivity to bias and fluctuations can still be estimated, albeit under the assumption that the fluctuations are small. For example, in equation (7.6)

\[ \frac{1}{(C^{-1})^2} = \gamma^{-1} - \gamma^{-1} (C^{-1} - \gamma^{-1}) \gamma^{-1} + O(\varepsilon^2), \]

so

\[ \hat{E}_k = \frac{1}{2} \left( \sum_{r,s} i k x e r^{-1} C_{rs} e^{-i k x s} \right) \]

and

\[ \hat{E}_k - \frac{\gamma}{k} = \frac{\gamma}{k} \sum_{r,s} i k x e r^{-1} (C_{uv} - \gamma^{-1}) \gamma^{-1} e^{-i k x s} + O(\varepsilon^2). \]

Thus to a first approximation,

\[ \overline{E}_k - \overline{\gamma} = \frac{\gamma}{k} \sum_{r,s} \overline{E}_k e^{i k x e r^{-1} e^{-i k x s}}^2 \overline{E}_1 \overline{\gamma}_1 \] (7.16)

and

\[ \overline{E}_k (\hat{E}_k \hat{E}_k) = \left( \frac{\gamma}{k} \right)^2 \sum_{r,s} \overline{E}_k e^{i k x e r^{-1} e^{-i k x s}}^2 \overline{E}_1 \overline{\gamma} \] (7.17)

Equations (7.16) and (7.17) replace (7.2) and (7.12) respectively, with \( \hat{\nu} \) defined by replacing \( \overline{\nu} \), \( \overline{E} \) in (7.6) by \( \hat{\nu} \), \( \hat{E} \). Substituting in equation (7.14), we obtain

\[ H_{kl} = \frac{1}{2} \overline{\gamma} \sum_{r,s} i k x e r^{-1} e^{-i k x s}^2 \overline{E}_1 \] (7.18)

which can be compared with equation (3.19). Note that the spectrum \( \overline{E}_k \) here is two-sided, as opposed to the one-sided spectra used elsewhere in this paper, and the wavenumbers have a range including negative as well as positive values.

For the model fitting methods, the most important consideration is that the model for \( \hat{E}_k \) be defined broadly enough to include quite
general $\hat{v}_k$ in equation (7.9). The form for $H$ then depends on how

the fitting is done. It was shown in Section 2.2 that if generalized

least squares are used the least uncertainty in the result comes from

using the weights defined in (2.17). Again replacing $\bar{C}$ by $\hat{C}$, this

leads to an $H$ matrix identical to (7.18). To the approximation con-

cidered here the same is true using the maximum likelihood fit of

Jones (1977). On the other hand, if the weights (2.17) were replaced

by some other expression such as $\delta_{ru} \delta_{sv}$ corresponding to simple least

squares, the $H$ matrix would be substantially altered from the optimal

one.

So far this discussion has shown how the conditional probability

of the spectrum defined by $\nu(k)$, given a particular set of observations,
can always be put in the form (7.13), where the matrix $H$ depends on the
observation array, on the general form of the spectrum as anticipated
in the first guess, and on the analysis technique used. As shown in
Section 2, there is a definite choice of the coefficients $\alpha_{rs}(k)$ that
maximizes the quadratic form $\nu'^T H \nu'$ for every $\nu'$. To within the
approximations made, this choice is implied by the maximum likelihood
methods of Jones (1977) and Capon (1969), by the optimal least squares
fitting of Section 2.4, and by the interpolation method if the inter-

polation is the least mean square error linear estimator for a process

with covariance function $\hat{C}(\xi)$. However, equation (7.13) implies nothing

about the magnitude of an arbitrary vector $\nu_o$ that can be added to $\nu$

provided only that it doesn't change the covariance matrix $C_{rs}$. This

is a typical ill-posed inverse problem, and further progress requires

additional assumptions about the smoothness of $\nu(k)$. 

In Section 3.2 these additional assumptions took the form of an a priori probability distribution for \( V \), by writing

\[
V_k = \sum_p \mu_p G^p_k, \quad V_k V_1 \equiv T_{kl} = \sum_p G^p_k G^p_l ,
\]  

(7.19)

where the \( \mu_p \) are independent normally distributed random variables with zero mean and unit variance a priori, and the \( G^p(k) \), \( p = 1,2, \ldots \), are smooth functions of \( k \), increasing in complexity and decreasing amplitude as \( p \) increases. Before pursuing this approach further it is worth considering alternative formulations of the inverse problem (e.g., Parker, 1977). These depend on the specification of a norm \( ||v||^2 \) in a separable Hilbert space. This norm is formally equivalent to the exponent \( v^T T^{-1} v \) in the Gaussian a prior probability distribution assumed in Section 3. Parker considers a finite set of observations, exactly accurate, for which the expected values are given linear functionals of \( v \). The analogue here would be the \( 1/2 N(N+1) \) values of

\[
C_{rs}(v) - \hat{C}_{rs} = \int v(k) \hat{v}(k) \cos k(x_r - x_s) \, dk .
\]  

(7.20)

He also supposes that \( v(k) \) can be written as the sum of a finite set of features \( F^p(k) \) and a residual which is orthogonal to the set:

\[
v(k) = \sum_p \lambda_p F^p(k) + v_\lambda(k) ,
\]  

(7.21)

where the inner products are such that

\[
(v_\lambda, F^p) \equiv v_\lambda^T T^{-1} F^p = 0
\]  

(7.22)

for \( p = 1, \ldots P \). Then a particular parameter set \( \{\lambda_p\} \) is accepted as consistent with the observations if \( ||v_\lambda||^2 \) is less than a critical
value. Analysis shows that this criterion is equivalent to comparing the residual

$$C''_{rs} = \frac{C^M_{rs}}{C_{rs}} - \sum_{p=1}^{P} \lambda_p C^P_{rs}$$  \hspace{1cm} (7.23)$$

between the observations and the model predictions with the fluctuations in $C_{rs}$ that would be expected if $\nu(k)$ were indeed a random variable with the stated a priori distribution. Then

$$||\nu_k||^2 = \sum_{\mathbf{u};\mathbf{v}} C''_{rs}(\mathbf{u}^{-1})_{rs;uv} C''_{uv}$$  \hspace{1cm} (7.24)$$

where $U_{rs;uv}$ is the covariance matrix of $C''_{rs}(\nu)$ given the constraints that

$$(\nu,\mathbf{r}^P) = \sum_{\mathbf{q}} (\mathbf{r}^P,\mathbf{r}^Q) \lambda_q, \quad p = 1,..P$$  \hspace{1cm} (7.25)$$

hold for fixed $\{\lambda_p\}$. The norm (7.24) is distributed as chi-square with $1/2 N(N+1)$ degrees of freedom, providing one makes a familiar interpretation of Parker's accept-reject criterion in terms of confidence levels. Note that this argument is concerned solely with the fit of a particular incomplete finite model to the observed covariance matrix, supposing the latter to be exactly known. Residuals are ascribed to inadequacies in the model, not to sampling errors in the data.

When uncertainties in the data are included, Parker's treatment becomes algebraically complex because of the combination of discrete accept-reject criteria with continuous probability distributions. A simpler procedure is simply to add the norms associated with sampling errors and with model uncertainties. This is formally equivalent to
the Bayesian derivation in Section 3, interpreting the norms as exponents in Gaussian probability distributions. The quadratic norms then add because the effects are independent, yielding a combined quadratic norm.

Returning now to the question of encompassing different approaches to spectral estimation, we recognize a certain arbitrariness about the final inversion by minimizing the role of the a priori probabilities. It is, however, necessary to pursue an argument of the type presented in Section 3 in order to obtain a solution to the inverse problem. In this approach, we recommend the definition of an attention function $s(k)$, which makes explicit the investigator's relative emphasis on different wavenumber bands, a spectral resolution scale $\Delta s$, an a priori accuracy parameter $\epsilon$, and a definition of the subspace $M$ of effectively measurable spectral features. The latter can be defined by examining the eigenvalues and eigenfunctions of the matrix

$$R_{st} \equiv M \epsilon^2 \frac{dk}{ds} H_{kl} \frac{dl}{dt}, \quad (7.26)$$

which is simply a renormalized form of the fundamental matrix $H_{k,l}$, transformed into the wavenumber coordinates $s$ and $t$. $R$ bears the same relation to $H$ as the information matrix $R$ bears to the aliasing matrix $H$ as defined in equation (3.28), when the basis features $e_k^p$ are of amplitude $\epsilon$ and vanish outside a wavenumber interval of $0(\Delta s)$. The eigenvectors $F^p(s)$ of $R$ span $M$ when they are limited to $p$ values with the associated eigenvalues such that $\lambda_p > \lambda_c = 0(1)$. Of course, once $M$ is defined, any linear combination of the $F^p(s)$ provides an equally acceptable basis set of measurable spectral features. The size (dimension) of this space, the parameter $P$, is the best single measure of the resolving power of the analysis.
It was shown in Section 2 that there is a class of optimal spectral estimators for which the quadratic form $\nu^T \mathbf{H} \nu$ is greater, whatever the shape of $\nu$, then for estimators not in the class. Thus, for non-optimal estimators the ratio $(\nu^T \mathbf{H} \nu)/(\mu^T \mu)$ is reduced, in particular when $\mu(s)$ is one of the principal components $F^P(s)$ for the optimal estimators. Each one of these components is thus less well measured, and the overall performance is definitely inferior, with the effective alias quite possibly enlarged. A quantitative statement of the reduction in performance requires detailed consideration of the matrix elements $\mu^P \mathbf{R} Q$. Note that the classification of certain a priori and data adaptive methods as optimal is only after the bias has been carefully removed. The information is potentially there, but if it is misinterpreted it can be very misleading.

This discussion shows that within this framework the choice of a suitable finite model for fitting to the data is not arbitrary, but should preferably span just the effectively measurable subspace $M$. If the model space $F$ includes any vector in the orthogonal complement $N$, then for that feature the amplitude is determined mainly by the a priori assumptions rather than by the observations. If there is any vector in $M$ not included in $F$, there will be some values of $\nu$ for which the model is demonstrably inconsistent with the observations. Such a choice of model is easily accomplished by letting the basis set $F^P(k)$, $p = 1,\ldots,P$ be identical to the eigenfunctions $F^P(s)$ for which $r_p > r_c$, though clearly any other linear combination of these functions would do as well. According to the interpretation in Section 3, such a cut-off amounts to a redefinition of the a priori possibilities. The acceptance
or rejection of a particular set \{\lambda_p\} is then based on confidence levels defined by the discrepancies from observations being ascribed solely to sampling errors. Alternatively, following Parker, it could be asserted that in reality we have

$$\mu(s) = \sum_{P} \lambda_p G^p(s) + \mu_\star(s) ,$$

(7.28)

where \mu_\star lies in \mathcal{N} and only the first part has been determined from the observations. The acceptance criterion is essentially unaltered because variations \mu_\star in \mathcal{N} make a negligible difference to the expected covariances \mathcal{C}_{rs}(\nu).

Finally, the accuracy parameter \varepsilon must be chosen so that the space \mathcal{M} does indeed define \mu(s) to within tolerable limits of uncertainty. Probably the most satisfactory measure is the maximum mean square residual uncertainty \(\sqrt{|\mu'(s)|^2}\) for any s. However, a more convenient one is the integral over s as indicated in equation (3.34). Reducing \varepsilon restricts the boundary \mathbb{P} of \mathcal{M}, and hence makes these tolerances more strict. The final decision is up to the investigator.
8. **Disclaimers**

The present paper, for all its length, is at best the beginnings of a general theory of spectral estimation. The theoretical framework presented has been successful in unifying a large number of different approaches to spectral estimation and, the authors believe, quantifies much of what investigators in practice intend. On the other hand, there are some aspects of this paper that do not fit into the previous approaches. For example, fruitful for array design is the concept of the determination of a spectrum in terms of preconceived features as the legitimate goal of an experiment. An important aspect of this theoretical framework is its potential generalizability to quite complicated circumstances, even though most of the paper has adopted the relatively simple assumptions stated in Section 2.1. At various points several possible generalizations were briefly discussed, but further investigations are needed into the robustness of the present conclusions for inhomogeneity or multiple dimensions or non-vanishing fourth order cumulants. Furthermore, since an investigator's prejudices and judgments about scientific importance are such an integral part of spectral estimation, much of what the authors advocate is rooted in their own prejudices and judgment. We are well aware that any truly general theory must be supported by a much broader set of examples than we have considered here.
ACKNOWLEDGMENTS

The authors appreciate the assistance of Ms. Julianna Chow in computation and Ms. Holly Howard in preparing the manuscript.
REFERENCES


Julian, P. R., and A. K. Cline, 1974: The direct estimation of spatial wavenumber spectra of atmospheric variables.


Parker, R., 1977: Linear influence and underparameterized models.


Thompson, R., 1971: Spectral estimation from irregularly spaced data.

FIGURE CAPTIONS

1. A 20-point data array which is approximately optimal (see Section 4.3) for \( M = 100 \), \( \hat{E}(k) \) as in (2.36), and \( F^P \) as in (2.37). Also shown is a random realization of \( \phi(x_i) \), with the data values connected by straight lines (n.b., this is the optimal interpolation for a spectrum proportional to \( k^{-2} \)).

2. Spectral estimates for (2.36)-(2.37) and \( M = 100 \). Shown are \( E(k; \lambda) \) for \( \lambda = \hat{\lambda}^{(1)} \) and \( \lambda = \hat{\lambda}^{(\infty)} = \hat{\lambda}^{\text{ML}} \). Along the abscissa the boundaries for the wavenumber bands (2.38) are marked.

3. A spectral estimate for (2.40)-(2.43). This is a second stage fit of the smooth model (2.39) to the spectral estimate in Figure 2.

4. The first three orthonormal features \( g^J(k) \) from (2.47).

5. Elements of the principal component transformation matrix, \( A_{pq} \), introduced in (3.38), and satisfying the conditions (3.39)-(3.40), for the example of Section 2.7. Note the changes of scale in the ordinates as \( p \) increases. The associated eigenvalues are \( \{ \gamma_p \} = \{ 7.8, 6.2, 5.2, 4.5, 4.0, 3.3, 2.9, 0.7 \} \).

6. The first row of the principal component transformation matrix \( A_{1q} \) for the extended range block feature analysis discussed in the text.

7. Spectral estimation based upon the Fourier basis features (3.30) and the adjusted prior probabilities (3.34). \( E(k) \) and \( w(k) \) are as in (2.36) and (3.13); that is, this is an alternative analysis of the example in Section 2.7. Panel (a) shows the principal components for \( P = 4 \) and \( \gamma^{-1} = 4 \); panel (b) shows the error \( \nu^2(k) \) from (3.32) for \( P = 4 \) and \( P = 6 \).
8. The spectral estimate for the example of Section 2.7, based upon six block features of the type (3.12) (with amplitude 0.25 and with wave number block boundaries \( \{k_p, k_u\} \) shown along the log \( k \) axis). Plotted are the spectral estimates based upon \( \hat{\lambda}(1) \) from (2.28) and the expected uncertainty \( \pm \sqrt{\nu_k} \) from (3.32).

9. Feature covariance functions \( C^P(\xi) \) for the eight logarithmic block features (2.37) and the spectrum (2.36) (with an arbitrary normalization). Note the scale changes between panels (a) and (b).

10. The matrix \( H_{kl} \) from (3.67) for an aliased equally spaced array. The array has ten data points (\( N = 10 \)) and the spectrum is uniform \( (E(k) = 1) \) for \( k \in [0, 3\pi/\Delta x] \) (i.e., it extends to three times the Nyquist wavenumber \( \pi/\Delta x = N(\Delta k/2) \)). At the truncation wavenumber \( E_{3N/2} = 1/2 \ E(3\pi/\Delta x) \) for the same reason a factor of 1/2 is required for \( E(0) \).

11. Aspects of the optimal design of the least array. Panel (a) shows \( R(\xi) \) from (4.11) versus \( \xi \) for the relative feature cut-off wavenumber value \( \hat{k} = 0.5 \). Panel (b) shows the optimal \( \hat{\xi} \) as a function of \( \hat{k} \).

12. The distribution of adjacent and independent lags for the two 20-point arrays of Table 5 and Section 4.3. The points \( \xi_p = (p-1/2)/\Delta p = 1, 3.04, 4.38, 8.82, 17.73, 35.52 \) from (4.14).

13. The distribution of adjacent and independent lags for different values of \( M \), the number of realizations.
14. The distribution of adjacent and independent lags for different values of N, the number of array points. For N = 30, two arrays are shown which are analogous to the two shown in Figure 12 and which differ by only .005 in ΔL.

15. The distribution of adjacent and independent lags (right panels) for various spectral slopes E(k) (left panels).

16. The distribution of adjacent and independent lags for different feature resolution (i.e., different P).

17. A schematic solution to the optimal array design for the long, quasi-uniformly spaced array and a red spectrum. The solution has a cumulative, fractional array length L(k) from (4.22) such that it is proportional to the weight function (segments P'Q and Q'R) or a constant (segments PP' and QQ').

18. Two-dimensional array configurations with local maxima in the information gain ΔI. Columns are for different degrees of anisotropy in the features and rows are for different types of configurations. The circumstances are defined in detail in the text. Stars indicate the best configuration for each array design problem (each column).

19. Contours of the anisotropic block wavenumber features of (5.11):
(a) $F^p \alpha \cos \theta \overline{k}$, (b) $F^p \alpha \sin \theta \overline{k}$. $K'P$ and $K'Q$ define the block boundaries.

20. (a) A typical attention function $s(k)$. The solution to equation (3.57), $k_r(p_o)$, for $p = p_o$ is also indicated;
(b) A typical $g_p(\xi)$, defined following equation (3.54), and approximated by stationary phase in equation (3.55).
Table 1

(a) $\bar{R}$ for (2.36)-(2.37)

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>.9</td>
<td>.6</td>
<td>.1</td>
<td>.1</td>
<td>.0</td>
<td>.0</td>
<td>.0</td>
<td>.0</td>
</tr>
<tr>
<td>.6</td>
<td>3.1</td>
<td>.8</td>
<td>.1</td>
<td>.2</td>
<td>.1</td>
<td>.0</td>
<td>.0</td>
</tr>
<tr>
<td>.1</td>
<td>.8</td>
<td>4.6</td>
<td>.9</td>
<td>.4</td>
<td>.2</td>
<td>.1</td>
<td>.1</td>
</tr>
<tr>
<td>.1</td>
<td>.1</td>
<td>.9</td>
<td>4.8</td>
<td>.9</td>
<td>.4</td>
<td>.2</td>
<td>.1</td>
</tr>
<tr>
<td>.0</td>
<td>.2</td>
<td>.4</td>
<td>.9</td>
<td>4.1</td>
<td>.9</td>
<td>.3</td>
<td>.2</td>
</tr>
<tr>
<td>.0</td>
<td>.1</td>
<td>.2</td>
<td>.4</td>
<td>.9</td>
<td>4.8</td>
<td>.9</td>
<td>.3</td>
</tr>
<tr>
<td>.0</td>
<td>.0</td>
<td>.1</td>
<td>.2</td>
<td>.3</td>
<td>.9</td>
<td>5.9</td>
<td>1.2</td>
</tr>
<tr>
<td>.0</td>
<td>.0</td>
<td>.1</td>
<td>.1</td>
<td>.2</td>
<td>.3</td>
<td>1.2</td>
<td>6.4</td>
</tr>
</tbody>
</table>

(b) $\bar{R}^{-1}$ for (2.36)-(2.37)

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>13.1</td>
<td>-2.8</td>
<td>.3</td>
<td>-.1</td>
<td>.0</td>
<td>.0</td>
<td>.0</td>
<td>.0</td>
</tr>
<tr>
<td>-2.8</td>
<td>4.0</td>
<td>-.6</td>
<td>.1</td>
<td>-.1</td>
<td>.0</td>
<td>.0</td>
<td>.0</td>
</tr>
<tr>
<td>.3</td>
<td>-.6</td>
<td>2.4</td>
<td>-.4</td>
<td>-.1</td>
<td>.0</td>
<td>.0</td>
<td>.0</td>
</tr>
<tr>
<td>-.1</td>
<td>.1</td>
<td>-.4</td>
<td>2.2</td>
<td>-.4</td>
<td>.0</td>
<td>.0</td>
<td>.0</td>
</tr>
<tr>
<td>.0</td>
<td>-.1</td>
<td>-.1</td>
<td>-.4</td>
<td>2.7</td>
<td>-.4</td>
<td>.0</td>
<td>.0</td>
</tr>
<tr>
<td>.0</td>
<td>.0</td>
<td>.0</td>
<td>-.1</td>
<td>-.4</td>
<td>2.2</td>
<td>-.3</td>
<td>.0</td>
</tr>
<tr>
<td>.0</td>
<td>.0</td>
<td>.0</td>
<td>.0</td>
<td>.0</td>
<td>-.3</td>
<td>1.8</td>
<td>-.3</td>
</tr>
<tr>
<td>.0</td>
<td>.0</td>
<td>.0</td>
<td>.0</td>
<td>.0</td>
<td>-.3</td>
<td>1.6</td>
<td></td>
</tr>
</tbody>
</table>
Table 2

(a) $R_{jj}$, from (2.45)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>52</td>
<td>32</td>
</tr>
<tr>
<td>32</td>
<td>23</td>
</tr>
</tbody>
</table>

(b) $R_{jj}^{-1}$, from (2.45)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>.13</td>
<td>-.18</td>
</tr>
<tr>
<td>-.18</td>
<td>.30</td>
</tr>
</tbody>
</table>
Table 3

\[
R'_{pq} \text{ from (3.35)}
\]

\[
\begin{array}{cccccccccccc}
1.00 & .62 & .05 & .03 & .02 & .01 & 0 & 0 & .01 & 0 & .01 & 0 \\
.62 & 1.00 & .31 & .04 & .02 & .01 & 0 & 0 & .01 & 0 & .01 & 0 \\
.05 & .31 & 1.00 & .20 & .04 & .03 & .01 & 0 & .01 & 0 & .02 & 0 \\
.03 & .04 & .20 & 1.00 & .16 & .07 & .03 & 0 & .01 & 0 & .02 & 0 \\
.02 & .02 & .04 & .16 & 1.00 & .15 & .05 & 0 & .12 & 0 & .06 & 0 \\
.01 & .01 & .03 & .07 & .15 & 1.00 & .15 & .08 & .33 & .18 & .29 & .24 \\
0 & 0 & .01 & .01 & .03 & .05 & .15 & 1.00 & .15 & .94 & .42 & .70 & .60 \\
0 & 0 & .01 & .02 & .04 & .08 & .15 & 1.00 & .19 & .95 & .79 & .86 \\
.01 & .01 & .03 & .06 & .12 & .33 & .94 & .19 & 1.00 & .46 & .75 & .65 \\
0 & .01 & .02 & .04 & .08 & .18 & .42 & .95 & .46 & 1.00 & .92 & .97 \\
.01 & .01 & .02 & .06 & .11 & .29 & .70 & .79 & .75 & .92 & 1.00 & .98 \\
.01 & .01 & .02 & .05 & .10 & .24 & .60 & .86 & .65 & .97 & .98 & 1.00 \\
\end{array}
\]
Table 4

$R_{pq}$ for the estimation of Fig. 8

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>5.8</td>
<td>.9</td>
<td>.3</td>
<td>.2</td>
<td>.1</td>
<td>.0</td>
</tr>
<tr>
<td>.9</td>
<td>6.0</td>
<td>1.2</td>
<td>.5</td>
<td>.2</td>
<td>.1</td>
</tr>
<tr>
<td>.3</td>
<td>1.2</td>
<td>5.8</td>
<td>1.1</td>
<td>.4</td>
<td>.2</td>
</tr>
<tr>
<td>.2</td>
<td>.5</td>
<td>1.1</td>
<td>6.2</td>
<td>1.1</td>
<td>.4</td>
</tr>
<tr>
<td>.1</td>
<td>.2</td>
<td>.4</td>
<td>1.1</td>
<td>6.4</td>
<td>1.1</td>
</tr>
<tr>
<td>.0</td>
<td>.1</td>
<td>.2</td>
<td>.4</td>
<td>1.1</td>
<td>6.6</td>
</tr>
</tbody>
</table>
Table 5
Nearly Optimal Arrays for Section 4.3

<table>
<thead>
<tr>
<th>Property</th>
<th>Array 1</th>
<th>Array 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$</td>
<td>x_r - x_{r+1}</td>
<td>5.87, 2.96, 2.57, 2.66, 1.32, .20, .17, .17, .17, .17, .17, .17, .17, .17, .37, .65, .73, 1.46, 3.06, 5.72</td>
</tr>
<tr>
<td>$\Delta I$</td>
<td>4.02</td>
<td>4.05</td>
</tr>
<tr>
<td>greatest $\Delta x$</td>
<td>28.8</td>
<td>29.4</td>
</tr>
<tr>
<td>least $\Delta x$</td>
<td>0.164</td>
<td>0.163</td>
</tr>
</tbody>
</table>
Table 6

$\bar{R}$ for Section 4.7

<table>
<thead>
<tr>
<th></th>
<th>5.6</th>
<th>1.0</th>
<th>.3</th>
<th>.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>11.2</td>
<td>1.8</td>
<td>.5</td>
<td></td>
</tr>
<tr>
<td>.3</td>
<td>1.8</td>
<td>10.7</td>
<td>1.6</td>
<td></td>
</tr>
<tr>
<td>.1</td>
<td>.5</td>
<td>1.6</td>
<td>14.9</td>
<td></td>
</tr>
</tbody>
</table>
$\phi(x)$

array point locations

$\Delta x = 1$

Figure 1
Figure 2

\( E(k; \hat{\lambda}(1)) \)

\( E(k; \hat{\lambda}(\infty)) \)

\( \{k_{\perp}, k_{\parallel}\} \)

Figure 2
$E(k; \mu), J = 2$

Figure 3
Figure 4
Figure 5

- $A_{q1}$
- $A_{q2}$
- $A_{q3}$
- $A_{q4}$
- $A_{qp}$
- $A_{q6}$
- $A_{q5}$
- $A_{q7}$
- $A_{q8}$
Figure 6
Figure 7a
\[ \sum_{q=1}^{P} \left[ G^q(k) \right]^2 \]

Figure 7b
Figure 8
Figure 9
Figure 10

\[ \Delta k = \Delta l = 1 \]
Figure 11a
Figure 11b

\[ \frac{\partial \Delta I}{\partial \xi} \bigg|_{k=1} = 0 \]

Optimal
\[ k_0 \xi / \pi \]
Figure 12
Figure 13
Figure 14

Figure 14 shows the number of lags for adjacent and independent distributions with different sample sizes (N=10, N=20, N=30). The x-axis represents lag values ranging from 0 to 1.0, and the y-axis represents the number of lags ranging from 0 to 14.
\( \hat{E}(k) = \frac{1}{\nu} \frac{1}{1 + (k/\sqrt{\nu})^m} \)

---

**Figure 15**

(a) Plot of \( \hat{E}(k) \) with curves for \( m = 2 \) and \( m = 4 \).

(b) Detailed view of \( \hat{E}(k) \) showing approximate \( k^2 \) behavior.

(c) Further analysis with \( m = 0 \) and \( m = 1 \).
Figure 16

Number of lags vs. $H$

- $P = 4$
- $P = 8$

Independent

Adjacent

$H$

0 0.25 0.5 0.75 1.0
Figure 17

\[ w = \frac{ds}{dk} \]
ISOTROPIC P = 4

Figure 18abc

(a) 
ΔI = 3.11

(b) 
ΔI = 3.17

(c)* 
ΔI = 3.27
Figure 18def
ANISOTROPIC

Figure 18ghi

\[ P = 8 \]

\[ \Delta I = 1.70 \]

\[ \Delta I = 1.20 \]

\[ \Delta I = 0.20 \]

\[ \Delta I = -0.20 \]

\[ X \]

0 1 2 3

Figure 18ghi
Figure 19
Figure 20

(a) 
\[ k \frac{d\xi}{dk} \]

(b) 
\[ g_p(\xi) \]

\[ \delta \xi \approx \frac{2\pi}{k^*} \]

0