

University of Leicester
Department of Physics and Astronomy
Lecture Notes
Data Analysis Techniques

Dr. R. Willingale

December 3, 2007

Contents

1	Introduction	2
2	Modelling the instrument response	2
3	Systematic errors and noise	5
4	The data analysis problem	8
5	Direct inversion	9
6	Chi-squared fitting and confidence intervals	9
7	Linear, isoplanatic systems, stationary processes and Fourier filtering	14
8	Least squares fitting and the Wiener filter	15

9 Time series analysis	18
10 Fourier analysis of time series	19
11 Maximum likelihood fitting and restoration	25
12 The Bayesian method verses the classical approach	27

1 Introduction

This course is a formal introduction to data analysis rather than just a collection of things you can do with data. Most data you will encounter are produced by some instrument designed for purpose and the basic philosophy I will present is that a very large fraction of data analysis can be described using the same formalism involving an instrument response coupled with systematic errors and noise.

Such a formal approach has advantages. When confronted by a data analysis problem you should try and describe it using the recipe I will provide. Having done this the solution may be immediately apparent, using a well established technique or some method developed in another discipline but equally applicable in the situation presented by your data. Even if the solution is not so obvious the various methods which might be applicable will be clear. It is often the case that new data sets are very similar to old except that some feature of the system is slightly different. Established methods may be applicable if you make some minor adjustment or some acceptable approximation. If all this fails you will have a succinct statement of your problem on which to base a novel solution. You will be well placed to invent your own data analysis techniques.

2 Modelling the instrument response

In general an instrument responds to a function in some input space and produces a data set in some output space. For example the objective of a camera takes the angular distribution of light falling on the aperture and converts it into an intensity distribution over the focal plane. In turn this distribution is converted into coloured dyes over the surface of a plastic film. The input and output *space* are VERY different but if the camera and film are well made and correctly used there is a tight relationship between the original light distribution and the processed result.

The *response* of a large class of instruments can be modelled using an integral equation of the form:

$$g(x', y', z') = \int \int \int f(x, y, z) h(x', y', z', x, y, z) dx dy dz$$

$f(x, y, z)$ is a piecewise continuous function over the input space and $g(x', y', z')$ is a continuous function over some output space. $h(x', y', z', x, y, z)$ is the *response function* and in general depends on both the input and the output space. For simplicity we shall consider a 1-D system:

$$g(x') = \int f(x) h(x', x) dx$$

The extension to higher dimensions is normally straight forward but requires care and attention to details.

Note that if the input is an impulse $f(x) = A\delta(x - a)$ then using the properties of the Dirac delta function:

$$g(x') = A h(x', a)$$

The response function can be thought of as a normalised version of the output we expect from a perfect impulse input. In principle we can measure the response function $h(x', a)$ by stimulating the system with an impulse input at $x = a$.

A *perfect* instrument would have a response:

$$h(x', x) = C\delta(x' - x)$$

where C is just a scaling constant.

A *linear* instrument has a response such that if $f_1 \Rightarrow g_1$ and $f_2 \Rightarrow g_2$ then

$$f_1 + f_2 \Rightarrow g_1 + g_2$$

So, for instance, if we double the input level the output level is also doubled.

An *isoplanatic* system has a response which is only a function of the difference between the input and output coordinates:

$$h(x', x) = h(x' - x)$$

This means that the shape or form of the response is the same for every input position. In this case the system equation becomes a *convolution* of *faltung* integral:

$$g(x') = \int f(x)h(x' - x)dx$$

Although in most cases the underlying response of the instrument is properly modelled by an integral equation involving continuous functions the use of computers and digital techniques naturally leads to a *digitization* of the response. The input and output functions become vectors (or more generally matrices and tensors) and the integration becomes a summation:

$$g_i = \sum_j h_{ij}f_j$$

We must be careful that this representation is a good approximation to the true response by proper choice of the samples in the input and output space. If the function $h(x', x)$ changes rapidly as a function of x or x' then we must use a large number of samples across the change. We must also make sure that all finite (significant) contributions to the integral are included in the range covered by i and j . Now the instrument response has become a *RESPONSE MATRIX*. It is usually only an approximation to reality because of numerical integration errors. If the response is *linear* the elements of h_{ji} are constants and if it is *isoplanatic* then each row (or column) of the matrix is just a circular permutation of the first row (or column).

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 1 & 2 & 3 \\ 3 & 4 & 1 & 2 \\ 2 & 3 & 4 & 1 \end{pmatrix}$$

Because of this the matrix multiplication performs an approximation to a circular convolution. To avoid wrap-around (aliasing) problems you must pad out the rows (or columns) with zeros.

If the instrument involves more than 1 dimension then the input and output matrices (or tensors) can be stacked to form vectors. The response will still be a matrix but will have a complicated banded structure representing the coupling or blurring between the different coordinates. So the simple digitized system equation is still valid but the interpretation of the indices i and j is rather complicated. If the the system is isoplanatic and the response is separable:

$$h(x', y', x, y) = h_x(x' - x)h_y(y' - y)$$

and the response matrix has a so called *Toeplitz form* such that $h_{ij} = h_{mn}$ if $i - j = m - n$. If the system response is isoplanatic but not separable then the stacked response matrix has a *block Toeplitz form*.

3 Systematic errors and noise

Whether we model our instrument using an integral equation or a summation of discrete elements it is always an idealisation. The real world includes something extra in the data that is not included in our prescription. These are, of course, the dreaded errors or instrument *noise* hated by undergraduates. Actually it is these errors or the noise which makes data analysis interesting and a challenge.

Systematic errors are generally things which we could/should have included in the system response but couldn't/didn't. They can be eliminated or at least reduced by calibration.

Noise is something which is inherent in the processes involved in the measurement system but which is not amenable to the instrument response prescription. It can't be included in the response function but must be allowed for in data analysis.

It is conventional to include both systematic errors and noise as an additive term in the response equation:

$$g_i = \sum_j h_{ij} f_j + n_i$$

If each n_i is independent of its neighbours and independent of the input f_j then the noise is said to be *uncorrelated*. In many cases this is a good approximation. Note that this does not mean that n_i is independent of i . The noise can still be a function of position in the output space and uncorrelated.

In many modern instruments the noise arises directly from the inherent statistical nature of the processes involved. This can be thermal noise, for instance in an electrical resistance, or counting statistics when detecting individual particles, electrons, photons, neutrons etc..

The noise can come from the source itself in which case $f(x)$ is essentially a probability density distribution or f_j are probabilities if the samples are essentially integrals over increments in x .

The noise can be instrumental in which case $h(x', x)$ (or h_{ij}) describes the mean behaviour of the system and n_i is something which is added by the measurement process.

In many cases we can think of the output function $g(x')$ as a probability distribution (or g_i as probabilities) and then n_i (or at least a component of n_i) is just characteristic of the output of the instrument.

In any particle counting experiment we must encounter *counting statistics*. g_i will then be the mean number of particles expected in a given sample $x \rightarrow x + \Delta x$ over a given time interval. If the particles are independent (i.e. not bunched or anti-bunched) then the probability of detecting $m = 0, 1, 2, \dots$ events (particles) will be governed by the *Poisson distribution*:

$$p(m) = \frac{\lambda^m \exp(-\lambda)}{m!}$$

where $\lambda = \mu$ is the mean number detected (which is g_i in our description above). The variance of the Poisson distribution is $\sigma^2 = \lambda$ so it will depend on g_i and is NOT independent of the object distribution. The output of a particle counting experiment will also contain spurious or background events which are not related to the desired signal. These must be included in n_i and appear as an additive term in the mean of the Poisson distribution so $g_i = \lambda_i = s_i + b_i$ where s_i is the mean signal count expected in the i th sample and b_i is the mean background count. The variance is then given by the total mean count $\sigma_i^2 = s_i + b_i$.

The Poisson distribution is a limiting case of the *Binomial distribution* when the probability q of detecting an event in a single trial is very small. If we make N trials then:

$$p(m) = \frac{N!}{m!(N-m)!} q^m (1-q)^{N-m}$$

The mean of this Binomial distribution is $\mu = Nq$ and the variance is $\sigma^2 = Nq(1-q)$.

Finally if N is very large and neither q or $1 - q$ is too small then the binomial distribution becomes a continuous probability density, the *normal* or *Gaussian* distribution:

$$p(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$

where $z = \frac{m - Nq}{\sqrt{Nq(1-q)}}$ a standardized continuous variable. It is of course the normal distribution that we usually use to model the additive noise n_i in experiments which don't involve counting events:

$$p(n_i) = \frac{1}{\sigma_i \sqrt{2\pi}} \exp \frac{-(n_i - \mu_i)^2}{2\sigma_i^2}$$

where μ_i and σ_i^2 are the mean and variance of the noise in the i th sample.

In many problems it is necessary to consider a statistical model for the source and noise processes using moments or expectation values to characterise the underlying probability density function. In discrete notation using the sampled functions the first two moments, the mean and covariance are

$$\underline{u}_f = \langle \underline{f} \rangle$$

$$[\phi_f] = \langle (\underline{f} - \underline{u}_f)(\underline{f} - \underline{u}_f)^{\sim} \rangle$$

where $\langle \rangle$ denotes the expectation value and $(\tilde{})$ the transpose. \underline{u}_f represents the mean as a function of position and $[\phi_f]$ describes the correlation between different positions.

A stochastic process is said to be *stationary* if \underline{u}_f is a constant vector and $[\phi_f]$ has a Toeplitz form (or block Toeplitz if we are dealing with a stacked set of multiple dimensions). This means that the correlation function of the process only depends on the difference in position rather than the absolute position.

A matrix closely related to the above moments is the correlation matrix

$$[c_f] = \langle \underline{f} \underline{f}^{\sim} \rangle = [\phi_f] + [\underline{u}_f \underline{u}_f^{\sim}]$$

4 The data analysis problem

In the context of the system equation

$$g(x') = \int f(x)h(x', x)dx + n(x')$$

or

$$g_i = \sum_j h_{ij}f_j + n_i$$

we can state the problem posed in data analysis; given a set of measurements g_i what can we say about the source vector f_j or the source function $f(x)$?

If the noise function $n(x')$ is independent of the unknown source function $f(x)$ then we are trying to solve a Fredholm equation of the first kind. If $n(x')$ is a function of $f(x)$ then the integral equation is said to be a Fredholm equation of the second kind. In either case it is important to remember that in almost all experiments the discrete notation is an approximation to an integral equation.

If the instrument response is a complicated transformation (a Fourier transform say) or if the response is rather poor, then the major problem is to calculate a good estimate of f_j given g_i . This is usually called an *inversion problem*. How easy the inversion is depends on the form of the response matrix and the noise level. Good examples of inversion problems are aperture synthesis in radio astronomy or the correction for the optical blurring in the Hubble telescope. When performing an inversion the central question is how reliable is the result? What features of \hat{f} can we believe and which are due to noise?

If you are lucky (or if you don't make a mistake manufacturing the primary mirror) then the instrument response is good and it is easy to estimate f_j given the data g_i . Then the problem shifts to another plane. What can we deduce about the processes that produce $f(x)$ given the measurements $g_i \Rightarrow f_j$? In this case we can wonder about the physical system beyond the immediate instrument and the problem is a *modelling problem*. Although the response h_{ij} may not present any difficulty the noise n_i will still be a limiting factor. In this case we require to find good estimates of the model parameters and confidence limits on these parameters.

5 Direct inversion

We can attempt to invert the system equation

$$\underline{g} = [h]\underline{f} + \underline{n}$$

by pre-multiplying both sides by the inverse of the response matrix

$$[h]^{-1}\underline{g} = [h]^{-1}[h]\underline{f} + [h]^{-1}\underline{n}$$

If $[h]^{-1}$ exists, $[h]^{-1}[h] = [I]$ the unit matrix and if there is no noise $\underline{n} = 0$ then this procedure will recover the original source vector

$$[h]^{-1}\underline{g} = \underline{f}$$

Unfortunately the inverse $[h]^{-1}$ is usually ill-conditioned (some of the elements of $[h]^{-1}$ are very large because they are formed by division by very small numbers) or doesn't exist at all (the determinant is zero).

When we attempt the inversion the noise term $[h]^{-1}\underline{n}$ blows up and dominates. What is needed is some *pseudo inverse* matrix which is optimized to recover as much of the source vector as possible in the presence of noise. In practice this means suppressing the ill-conditioned elements in the true inverse $[h]^{-1}$ and trying to produce a best estimate of the source vector $\hat{\underline{f}}$.

Apart from the limitations imposed by noise most data sets are very large and the instrument response matrix is therefore rather large although it may be fairly sparse. The accurate inversion of such large matrices is difficult and time consuming and in many cases is not practical.

6 Chi-squared fitting and confidence intervals

Least squares or maximum likelihood fitting can yield model parameter values, c_1, c_2, \dots which are in some way the best estimates of the true values given the data. However on their own these methods can't tell us, firstly, whether or not the model is a good

interpretation (fit) of the data and, secondly, what confidence limits (errors) we should assign to the parameter values found.

The workhorse of methods for testing goodness of fit is the Chi-squared statistic. This can be used to fit a probability distribution to a measured data set, to test the goodness of fit in some fitting procedure, to yield confidence limits for derived parameters and forms the basis of a fitting procedure in its own right.

Chi-squared is based on the hypothesis that the optimum description of a set of data is that which minimises the *weighted* sum of the squares of the differences between the data values and predicted values. The variance of the fit is given by

$$v^2 = \frac{1}{N - k} \sum w_i (g_i - \hat{g}_i)^2$$

where N is the number of data points, k is the number of parameters c_1, c_2, \dots , $\hat{g}_i = y(x, c_1, c_2, \dots)$ some fitting function and w_i is a weighting factor for each data point:

$$w_i = \frac{\frac{1}{\sigma_i^2}}{\frac{1}{N} \sum \frac{1}{\sigma_i^2}}$$

This is a measure of the uncertainty on each data point normalized to the average uncertainty of all the points. Note that if all the data variances are constant $\sigma_i^2 = \sigma^2$ then $w_i = 1$ and v^2 is simply a sum of square differences divided by $N - k$.

Chi-squared is defined as

$$\chi^2 = \sum \frac{1}{\sigma_i^2} (g_i - \hat{g}_i)^2$$

So

$$\frac{\chi^2}{N - k} = \frac{v^2}{\langle \sigma_i^2 \rangle} = \chi_\nu^2$$

where $\nu = N - k$ is the number of degrees of freedom and

$$\langle \sigma_i^2 \rangle = \left[\frac{1}{N} \sum \frac{1}{\sigma_i^2} \right]^{-1}$$

the weighted sum of the variances. χ_ν^2 is called the reduced Chi-squared.

If the fit is good we expect that ν^2 will be close to the underlying variance of the parent process and $\chi_\nu^2 \approx 1$. If χ_ν^2 is too large then the fit variance must be much larger than the expected value and the fit is poor. If χ_ν^2 is too small then you may have under estimated the errors or included too many parameters in the fit.

We can attempt to quantify the *goodness of fit* using the expected probability density distribution of $\chi^2 = z^2$

$$p(z^2, \nu) = \frac{(z^2)^{\frac{\nu-2}{2}} \exp \frac{-z^2}{2}}{2^{\frac{\nu}{2}} \Gamma(\frac{\nu}{2})}$$

The integral of this function between $z^2 = \chi^2$ and $z^2 = \infty$

$$P(\chi^2, \nu) = \int_{\chi^2}^{\infty} p(z^2, \nu) dz^2$$

can be found tabulated in many statistics or data analysis books.

So if you search for parameter values c_1, c_2, \dots which produce the global minimum χ^2 you can estimate the probability that the χ_{min}^2 value found exceeds that value. That is if you performed the experiment an infinite number of times then some fraction of the data sets (*fraction* = $P(\chi^2, \nu)$) is expected to yield a resulting χ^2 larger than the one found from the actual data set.

Unfortunately χ^2 combines two factors, how good the underlying function $y(x, c_1, c_2, \dots)$ describes the data AND the deviations between the best fit function found and the data set. For example you could achieve a $\chi_\nu^2 \approx 1$ if the error estimates were too large and the function was a fairly bad fit, or the error estimates were too small and too many parameters were used to produce the fit.

The formal development and justification for the χ^2 fitting procedure outlined above is rather involved. This is discussed in some detail by Lampton, Margon and Bowyer (1976) [14]. In the definition cited

$$\chi^2 = \sum \frac{1}{\sigma_i^2} (g_i - \hat{g}_i)^2$$

the statistic will only be the *true Chi-squared* if the quantities \hat{g}_i and σ_i^2 are the true population mean and variance values. In any experiment these are only known to nature

and not the experimenter. In the idealized instance that they are the true values each term in the summation will be a sample of a random variable with zero mean and unit variance and in the limit of a very large number of samples ($N \rightarrow \infty$) the statistic will be normally distributed. When the number of samples is small (or finite) the statistic has the sampling probability distribution of χ_{true}^2 with N degrees of freedom $\nu = N$ as quoted above.

What we actually calculate is a χ_{min}^2 using a fitting function to yield \hat{g}_i . It can be shown that this sampling statistic is distributed as χ^2 with $\nu = N - k$ degrees of freedom when using k functional parameters. This minimization provides the *best fit* values for the parameters c_1, c_2, \dots and the χ_{min}^2 value.

Analysis shows that the difference between the χ_{true}^2 and χ_{min}^2 will also have a sampling distribution of χ^2 but with k degrees of freedom

$$\Delta S = \chi_{true}^2 - \chi_{min}^2 = \chi_k^2$$

This difference is independent of the χ_{true}^2 value and the number of data points and only depends on the number of parameters used to generate the χ_{min}^2 .

What we want is the definition of a confidence region or volume in the parameter space so that we can assign *errors* or confidence ranges to the parameters c_1, c_2, \dots . As we vary the parameter values about the minimum so we expect the Chi-squared to increase. If the increase exceeds some limiting value then the probability that the parameters are correct is small ($< 10\%$ say). The limiting value of the increase is χ_{limit}^2 given by

$$P(\chi_{limit}^2, k) = \int_{\chi_{limit}^2}^{\infty} p(z^2, k) dz^2 = 0.90$$

where k is the number of parameters we are varying. So if we fix one of the k parameters at a series of values about the minimum and for each of these fixed values we find the minimum Chi-squared by varying all the remaining parameters we can find the 90% confidence interval for that one parameter. We can do this for each parameter in turn to define the complete confidence region. This procedure ensures the independence of all the fitting parameters. It is not the same as fixing all the parameters at their best fit values and then varying just one parameter at a time. If this is done then $0.90 = P(\chi_{limit}^2, 1)$ (i.e. $k = 1$) and the confidence limits for each parameter are much smaller. In doing this you are not allowing for the interplay between the parameters in the fitting procedure. Actually if there are a large number of parameters it may be legitimate to assume constant values for some of the parameters and not consider them in the error analysis but just fix them at their best fit or some other values. This reduces the number k and the volume

of the confidence region. However you must be consistent in this designation. Once a parameter is assumed constant you can't change your mind later, vary this parameter and start quoting confidence limits for it.

Although the $\Delta\chi^2$ above is independent of χ_{min}^2 the procedure is only sensible if the initial χ_{min}^2 is reasonable. In the above we have implicitly assumed that $\chi_{min}^2 \approx \chi_{true}^2$. There is no point in calculating confidence limits for a model which is poor to start with or if the error estimates are wrong.

It is usually possible to lower the χ_{min}^2 value by introducing more parameters into the fit. So if the fit is poor we can include a new feature into the fitting function to try and model the discrepancy. It is therefore legitimate to ask at what point is the introduction of another parameter no longer justified by the data? This is the purpose of the so called *F-test*.

If we take the ratio of two Chi-squared statistics then the sampling probability distribution is the F distribution

$$p_f(z, \nu_1, \nu_2) = \frac{\chi_1^2/\nu_1}{\chi_2^2/\nu_2}$$

which has a functional form

$$p_f(z, \nu_1, \nu_2) = \frac{\Gamma(\nu_1/2 + \nu_2/2)}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} \frac{z^{1/2(\nu_1-2)}}{(1 + z\nu_1/\nu_2)^{1/2(\nu_1+\nu_2)}}$$

When we construct such a ratio it is the integral of the distribution which is of most interest

$$P(F, \nu_1, \nu_2) = \int_F^\infty p_f(z, \nu_1, \nu_2) dz$$

So to test for the necessity of an additional fitting parameter (or parameters) we take the ratio of the minimum Chi-squared values obtained with and without the new parameters. We then look up this F value in a tabulated form of the integrated probability above using the appropriate ν_1 and ν_2 values. If the ratio is larger (or smaller depending on which way up we take the ratio) than that expected at a given level of confidence then the inclusion of the new parameters is significant at that level. Note that $P(F_{12}, \nu_1, \nu_2)$ is not the same as $P(F_{21}, \nu_2, \nu_1)$ unless ν_1 and ν_2 are rather large so we should test both the ratio and its reciprocal taking care with the order of the ν_1 and ν_2 values.

If the inclusion of some parameters is not justified by a significant improvement in Chi-squared as defined by the F-test then there is no point in deriving confidence limits for those parameters or indeed even including them in the fit. Such parameters should be removed or fixed in the fitting procedure.

7 Linear, isoplanatic systems, stationary processes and Fourier filtering

The linear, isoplanatic system has a system response matrix with the Toeplitz (circulant) form. A circulant matrix is diagonalized by the Discrete Fourier Transform (DFT). Similarly the block Toeplitz form generated by stacking more than 1 dimensions is diagonalized by the multidimensional DFT. The elements of the 1-D DFT matrix are

$$t_{kj} = \exp(i2\pi kj/l)$$

$$t_{kj}^{-1} = \exp(-i2\pi kj/l)/l$$

where $i = \sqrt{-1}$ and l is the number of rows and columns. The properties of the complex exponential function lead to the Fast Fourier Transform algorithm (FFT) and so matrix diagonalisation (and inversion) for such a system can be performed very quickly.

Note that diagonalisation of the linear, isoplanatic system by the DFT can be thought of as the *discrete convolution theorem* the digital analogue of the familiar convolution theorem which holds for the Fourier transform.

Stationary processes have a covariance matrix with a Toeplitz form which is diagonalized by the DFT. By definition, the diagonal elements of the resultant matrix $[\Lambda_f]$ form the *power spectrum* of the stochastic process. (When dealing with continuous functions rather than vectors the power spectrum is the Fourier transform of the autocorrelation function.)

Discrete Fourier filtering is carried out by taking the DFT of the data vector

$$\underline{G} = [t]^{-1} \underline{g}$$

multiplying each element by a filter weighting

$$\underline{G}' = \underline{W} \underline{G}$$

and then taking the inverse transformation

$$\underline{g}' = [\underline{t}] \underline{G}'$$

If we construct a diagonal matrix from the weighting (filter) vector

$$\Lambda_{ii} = W_i$$

we can define a matrix

$$[\underline{w}] = [\underline{t}] [\underline{\Lambda}] [\underline{t}]^{-1}$$

and write the operation in the real domain as a matrix multiplication

$$\underline{g}' = [\underline{w}] \underline{g}$$

A consequence of the above is that if we decide to perform Fourier filtering on our data then we are implicitly treating the underlying processes as stationary and we are performing a convolution in the real domain. A Fourier filter treats all areas of the data plane equally; it performs the same correlation everywhere in the data space so it cannot adapt to changes in the noise or signal correlations or the instrument response as a function of position.

8 Least squares fitting and the Wiener filter

If we consider the inversion problem an estimate of the original source vector $\underline{\hat{f}}$ is required. The error between this estimate and the true vector is:

$$\underline{\varepsilon} = \underline{f} - \underline{\hat{f}}$$

Using the principle of least squares we want to minimize the mean square error

$$\min \langle \tilde{\underline{\epsilon}} \underline{\epsilon} \rangle$$

where $\tilde{\underline{\epsilon}}$ is the transpose of the error vector and $\langle \rangle$ indicates the expectation value of the product.

We want a matrix operator or *filter* $[w]$ such that

$$\hat{\underline{f}} = [w]\underline{g}$$

Using the system equation to substitute for \underline{g} and minimizing the square error yields:

$$[w] = [c_f][\tilde{h}]([h][c_f][\tilde{h}] + [c_n])^{-1}$$

where $[c_f]$ and $[c_n]$ are the correlation matrices of the source and the noise

$$[c_f] = \langle \underline{f} \underline{f}^{\tilde{}} \rangle$$

$$[c_n] = \langle \underline{n} \underline{n}^{\tilde{}} \rangle$$

and the noise is assumed to be independent of the source

$$[c_{nf}] = \langle \underline{n} \underline{f}^{\tilde{}} \rangle = [c_{fn}] = [0]$$

This form of filter $[w]$ was originally applied to time series by Wiener, 1949 [29] and is therefore called a Wiener Filter. It has been used successfully in image processing and other areas of data analysis following the work of Helstrom (1967).

It is only easy to implement the Wiener filter if a diagonalization transformation exists so it has mostly been used as a Fourier filter

$$[t]^{-1} \hat{\underline{f}} = [\Lambda_f][\Lambda_h^*]([\Lambda_h][\Lambda_f][\Lambda_h^*] + [\Lambda_n])^{-1}[t]^{-1} \underline{g}$$

where the diagonal elements of $[\Lambda_h]$ are the DFT of the response function and those of $[\Lambda_f]$ and $[\Lambda_n]$ are the power spectra of the source and noise processes (assumed to be stationary).

The Wiener filter constructed in this way attempts to restore the best estimate of the source vector (function) assuming a particular simple form for the system response, source statistics and noise.

Perhaps a more familiar application of the Principle of least squares is model fitting. Suppose the estimate $\hat{f}(x)$ (or if you prefer $\underline{\hat{f}}$) is modelled by a function $y(x, c_1, c_2, \dots)$ where c_1, c_2, \dots are unknown parameters or coefficients which are often physical attributes of the source under study, temperature, density, etc.. Such a source function would lead to an estimated data function

$$\hat{g}(x') = \int y(x, c_1, c_2, \dots) h(x', x) dx$$

The error function (vector) is then

$$\underline{\varepsilon} = \underline{g} - \underline{\hat{g}}$$

and we can again search for

$$\min \langle \underline{\tilde{\varepsilon}} \underline{\varepsilon} \rangle$$

varying the parameters c_1, c_2, \dots to find the best solution. For example if the instrument response is perfect (a delta function) and $y = c_1 + c_2 x$ we get the familiar linear regression formulae

$$c_1 = \frac{\sum_i g_i \sum_i x_i^2 - \sum_i x_i \sum_i x_i g_i}{N \sum_i x_i^2 - (\sum_i x_i)^2}$$

$$c_2 = \frac{N \sum_i x_i g_i - \sum_i x_i \sum_i g_i}{N \sum_i x_i^2 - (\sum_i x_i)^2}$$

where N is the number of data points.

Similar expressions can be found for the coefficients of higher order polynomials but if y is a complicated function or if the system response is not perfect then we must resort to some form of searching algorithm to find the least squares solution.

9 Time series analysis

Any set of observations of some variable taken at different times is called a *time series*. The times are usually equally spaced and form a continuous sequence but can be unequally spaced and often contain gaps because of viewing constraints or exposure problems. Each sample can be a *snapshot* of the variable or some form of average value taken over a time sample. In most cases the measurement instrument does not blur or mix times so that the response h_{ij} is diagonal with gaps in the data represented by missing elements along the diagonal. In some cases the variable we measure is a summation of different components which suffer different time delays and then off-axis elements in h_{ij} will represent the time delays.

The analysis of time series is the subject of many chapters in books on statistics and signal processing. In general the movements or fluctuations seen are classified into types

- Trends or long term secular movements
- Periodic or cyclical variations
- Irregular or random fluctuations

So the usual problem of time series analysis is trying to extract these features from the data and characterise them rather than deconvolving the instrument response. We can model the time series as a sample of a continuous function $g(t)$ multiplied by a window function to represent gaps or exposure variations.

$$g_i = g(t_i)w(t_i)$$

As with all data analysis, superimposed on top of the underlying behaviour, be it periodic, long term trend or irregular, is noise, so the sampling by the instrument is described by an integral equation of the form

$$g(t_i) = \int_{-\infty}^{+\infty} h(t_i, t)f(t)dt + n(t_i)$$

If this process is effectively instantaneous then $h(t_i, t) = \delta(t_i - t)$ but most instruments will integrate over some small time interval so that $h(t_i, t) = h(t_i - t)$ where $h(t_i - t) = 1$ for $0 < t_i - t < \tau$ and $h(t_i - t) = 0$ otherwise. Providing τ is smaller than the time between samples then each g_i will be independent of its neighbours. Within this model we want to characterise $f(t)$ give g_i .

Trend analysis is best done by linear regression, least squares fitting or Chi-squared fitting as described above. A Chi-squared test of the hypothesis that $f(t)$ is constant is a good way of seeing whether there is evidence for variability of any sort in the data although it can be fooled if there is a very weak periodic signal present as you will see below.

10 Fourier analysis of time series

If the samples are equally spaced at Δt we can look for periodic behaviour using the discrete Fourier transform (DFT) as described under the section on linear, isoplanatic systems.

$$\underline{G} = [t]^{-1} \underline{g}$$

where for a vector of length l the forward transform matrix is

$$t_{kj}^{-1} = \frac{1}{l} \exp(-i2\pi kj/l)$$

This expresses the vector \underline{g} as the linear sum of Fourier components (eigenvectors) with each G_i corresponding to the frequency

$$\nu_i = \frac{i}{l\Delta t}$$

where $i = 0, 1, 2, 3, \dots, l-1$ and l is the total number of samples. G_i is complex but if g_i are real (which is usually the case) then only the first $l/2 + 1$ samples of G_i will be required. The highest frequency in the data is represented by $i = l/2$ with a frequency of $\nu = 1/(2\Delta t)$ the Nyquist frequency (2 samples per period). The remaining samples $i = l/2 + 1 \rightarrow l-1$ are a mirror image of the low frequency samples $G_i = G_{l-i}$ for $i = 0 \rightarrow l/2 - 1$. The real part of G_i represents the cosine (symmetric) components and the imaginary part the sine (antisymmetric) components. (Note the symmetry is about $i = 0$ not $i = l/2$.)

Any periodic component in g_i will appear as a series of peaks in the vector G_i , at the fundamental frequency and the harmonics (integer multiples of the fundamental). The relative amplitudes of these peaks will depend on the shape of the periodic function. The amplitude of the Fourier components will be given by $\sqrt{(G_i G_i^*)}$ and the phase angle by $\tan^{-1}(\text{imag}(G_i)/\text{real}(G_i))$.

There are a number of problems associated with period searching using the DFT. The first is the noise. If the time samples don't overlap the vector G_i is the linear sum of the DFT of f_i and the DFT of the noise vector n_i .

$$G_i = F_i + N_i$$

The power spectrum of the data set (square of the amplitude of Fourier components) is therefore given by

$$G_i G_i^* = (F_i + N_i)(F_i^* + N_i^*) = F_i F_i^* + N_i N_i^* + F_i N_i^* + N_i F_i^*$$

The first two terms on the righthand side are the power spectrum of the signal and noise respectively. If the noise is uncorrelated with the signal then (by definition) the second two cross-terms will be zero. So the data power spectrum is usually a simple sum of the signal and noise power spectra.

By the convolution theorem these power spectra are the Fourier transforms of the auto-correlation functions of the time series of the signal and the noise. If

$$c_f(t') = \int f(t) f(t' + t) dt$$

then

$$C_f(\nu) = \frac{1}{\sqrt{(2\pi)}} \int c_f(t') \exp(-i2\pi\nu t') dt' = \sqrt{(2\pi)} F(\nu) F^*(\nu)$$

Notice that the complex conjugation has been introduced since we are taking an auto-correlation not a convolution, ($t' - t \rightarrow t' + t$). Or in discrete notation

$$\underline{c}_f = [f_c] \underline{f}$$

where $[f_c]$ is a circulant matrix generated from \underline{f} (see earlier section on modelling the instrument response). The circulant for a crosscorrelation is the transpose of that required for a convolution. Then

$$\underline{C}_f = [t]^{-1} \underline{c}_f = l([t]^{-1} \underline{f})([t]^{-1} \underline{f})^* = l \underline{E} \underline{E}^*$$

where l is the number of elements in the vector \underline{f} . Actually multiplication by the circulant form $[f_c]$ performs a discrete circular convolution and the vectors must be padded out with zeros to render the discrete result equivalent to the continuous form. A proof of the above and some discussion of these details is given by Hunt (1971).

If the noise is *white* then there is no correlation between the noise in different data samples and the auto-correlation function of the noise will be a delta function. The power spectrum of the noise will then be a constant since

$$\delta(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-i2\pi\nu t) d\nu$$

In practice the data vector (and hence the noise vector \underline{n}) only contain a finite number of samples so the sample auto-correlation function of the noise vector is approximately a delta function with residual wings. Likewise the noise power spectrum will be flat but will contain residual fluctuations. If the noise is Gaussian with standard deviation σ then the mean power in the DFT spectrum $\underline{N} \underline{N}^*$ is σ^2/l . Any mean offset in the noise will only appear in the N_0 term (sometimes referred to as the DC term by analogy with AC theory). The power samples are distributed as Chi-squared with 2 degrees of freedom (because they are the sum of the squares of the real and imaginary parts of the spectrum, see for example 'An introduction to spectral analysis' J.N.Rayner). If the noise is Poissonian (counting statistics) with a constant mean then in the limit of large l the power samples again are distributed as Chi-squared with 2 degrees of freedom (see D.A.Leahy et al 1982). It should be noted that because the Nyquist term ($i = l/2$) contains only a real component this is distributed as Chi-squared with 1 degree of freedom.

If the signal process $f(t)$ has a stochastic component then this will also add to the mean level per frequency sample but this is seldom white and hence the $\underline{F} \underline{F}^*$ will not be flat. Any periodic component of $f(t)$ will appear as peaks superimposed on the continuum. The significance of such peaks will be determined by the Chi-squared distribution of the continuum fluctuations. In order to normalise these fluctuations the power should be multiplied by $2l/\sigma^2$ so that the mean value of the Chi-squared distribution is 2 as expected.

Analysis of power spectra is complicated by the presence of a window function. Since the continuous function $g(t)$ is multiplied by the window function $w(t)$ in the real domain the spectrum of the signal is convolved by the spectrum of the window function in the Fourier domain. In discrete notation we have

$$\underline{G}' = [W_c] \underline{G}$$

where $[W_c]$ is a circulant matrix constructed from the DFT of the window function \underline{W} . In the trivial case of continuous coverage the circulant will be diagonal because \underline{W} is a delta

function. However if there are missing real domain samples the convolution will produce a mixing of Fourier domain samples and nearby frequency samples will be correlated. The exact form of the correlation depends on the shape of the window function. If the spectral width (frequency resolution) of the window function is W_r , then the square of the coefficient of variation (the fractional variance) of the power is

$$\frac{\text{variance}}{(\text{average power})^2} = \varepsilon^2 \approx \frac{1}{l\Delta t W_r} = \frac{\Delta\nu}{W_r}$$

When there are no gaps $W_r = \Delta\nu = 1/l\Delta t$ so we have $\varepsilon^2 = 1$ and the normalised power samples are distributed as Chi-squared with 2 degrees of freedom. If gaps are introduced then the spectral resolution gets worse (W_r gets larger with off-diagonal elements in $[W_c]$, the precise definition of W_r depends on the exact form of the window function) and the number of degrees of freedom is approximated by

$$k = \frac{2}{\varepsilon^2} \approx \frac{2W_r}{\Delta\nu}$$

(You should recall that the mean of a Chi-squared distribution is k , the number of degrees of freedom, and the variance is $2k$.) To get the correct normalisation for the distribution you must sum the power in the samples across the spectral width W_r and then scale by $2(l - l_g)/\sigma^2$ where l_g is the number of missing samples. The number of frequency samples across W_r is approximated by the ratio of the total number of evenly spaced samples in the time series l divided by the number of exposed samples.

$$l_r \approx \frac{l}{l - l_g}$$

Therefore after summing across the resolution bandwidth and scaling the effective scaling factor is $2l/\sigma^2$ as before. However the gaps have reduced the coefficient of variation ε and the power spectrum has been *smoothed*. As might be expected there is a direct trade between the signal to noise of the power samples and the spectral resolution. This behaviour of the power spectrum is discussed in detail by Blackman and Tukey (1958).

If the gaps in the data are large or frequent then the measured power spectrum \underline{G}' will be distorted. In particular periodic peaks in \underline{G} will be accompanied by side peaks associated with the power spectrum of the window function. Periodic components in the real domain must be found by a matched filtering operation in the Fourier domain looking for the characteristic shape of the power spectrum of the window function in the spectrum.

If the times series is very long then the potential frequency resolution is very high but may be compromised by gaps (or the capacity of the DFT routine). In such a case it

may be better to split the data into a series of shorter observations not so plagued by missing samples. Breaking the data into M smaller series of equal length l/M we can take the power spectrum of each sub-set of data and then add the power spectra samples together to construct the mean power spectrum of the complete data set. If there are very few missing data values in each sub-set the normalised mean power samples will be distributed as Chi-squared with $2M$ degrees of freedom ($\varepsilon^2 = 1/M$) so the signal-to-noise will be improved by order \sqrt{M} and the spectral sampling (resolution) will be reduced to $\Delta\nu = M/l\Delta t$.

If there are a limited number of data samples and many gaps or if the periodic modulation expected has a peculiar pulse shape (not sinusoidal) then the DFT is not well suited to finding a periodic signal. In such cases a better approach is the method of *epoch folding*. Assuming a trial period of duration τ each time sample is converted into a phase by dividing by the period and taking the remainder (modulo function) again divided by the period. If the original time samples have a natural resolution of Δt then to avoid any beating between this sampling and the trial period the times t_i must be artificially perturbed by $\pm\Delta t/2$ using a random number generator. So assuming r_i is a random number taken from a distribution evenly distributed over the interval $-\Delta t/2 \rightarrow +\Delta t/2$, we have phases

$$p_i = \text{modulo}(t_i + r_i, \tau) / \tau$$

The weighted frequency distribution r_j (histogram) of these phase samples is then accumulated by *summing* the data into n_p bins spanning the phase space of $0 \rightarrow 1$ using the phase index for the i th time sample given by

$$j = \text{integer}(p_i n_p) \quad (= 0 \rightarrow n_p - 1)$$

As well as accumulating the data values $g(i)$ using the index j the unweighted frequency distribution or the *occupancy* u_j of each phase bin is also accumulated. This tells us the exposure each phase bin has received including the influence of data gaps. The mean data value for each phase bin can then be calculated

$$\bar{r}_j = \frac{r_j}{u_j}$$

The variance of this mean must be estimated from the errors on the data samples σ_i^2 . If this is constant ($\sigma_i = \sigma$) then

$$\sigma_j^2 = \frac{\sigma^2}{u_j}$$

If there is a modulation in the data at the trial period τ then we expect the phase samples \bar{r}_j to exhibit the pulse shape. If not, then providing there are sufficient data samples so that $u_j \neq 0$ and any secular trends occur over time scales much larger than the trial period, we expect the phase samples to be constant \bar{r} . We can test the hypothesis that the folded data are constant using a statistic

$$S_\tau = \sum_0^{n_p-1} \frac{(\bar{r}_j - \bar{r})^2}{\sigma_j^2}$$

In the absence of periodic or secular variations we expect S_τ to be distributed as χ^2 with $n_p - 1$ degrees of freedom.

So to search for periodic modulation we calculate S_τ for a large number of trial periods and look for values above some specified confidence level. In order to avoid missing significant peaks we must choose periods corresponding to a frequency spacing $1/2T$ where T is the duration of the data set ($T = l\Delta t$) if there are l samples at spacing Δt .

The choice of the number of phase samples n_p depends on what is required of the analysis. If the modulation is roughly sinusoidal and maximum detection sensitivity is needed then $n_p = 2$ will suffice. Each phase sample will then be the summation of approximately half the original data samples but the fractional variance on each will be a factor $l/2$ that of the original data sample. In the other extreme if $n_p = l$ then the variance of each point is on average the same as in the original data. So the signal to noise for detection of a periodic variation by folding can be increased by order $\sqrt{l/2}$ compared with looking for the variation in the raw data. It is precisely this signal to noise gain which makes both epoch folding and the DFT so sensitive in finding periodic variations in long stretches of data.

On the other hand if the pulse shape contains sharp steps or the phase of the pulse is of prime interest then n_p must be chosen commensurate with the mark-space ratio of the pulse and phase accuracy required. Incidentally an easy way to find the phase of the modulation is to calculate the first Fourier component of the folded profile.

$$R_1 = \frac{1}{l} \sum_j \exp(-i2\pi j/l) \bar{r}_j$$

Then the phase of the modulation (relative to $t = 0$) at period τ is

$$\phi_\tau = \tan^{-1}(\text{imag}(R_1)/\text{real}(R_1))$$

11 Maximum likelihood fitting and restoration

The principle of least squares is a simple formulation of a more general method, *maximum likelihood*. Given an estimate or guess for the source function $\hat{f}(x)$ or a model function for the source $y(x, c_1, c_2, \dots)$ it is usually possible to calculate or estimate the probability (or probability density) of observing a particular event or value in the data set.

$$p_i = p(x'_i, \hat{f})$$

or

$$p_i = p(x'_i, y) = p(x'_i, c_1, c_2, \dots)$$

We can then construct a likelihood function from the product of the individual probabilities

$$L(\hat{f}) = \prod_i p_i$$

or

$$L(c_1, c_2, \dots) = \prod_i p_i$$

The maximum likelihood occurs when L is stationary with respect to changes in the estimate function or the fitting parameters.

There are many varied applications of this method but it becomes particularly effective when dealing with event data or data from particle counting experiments.

Suppose the data set (vector) g_i consists of the number of detected events in a series of bins $x'_i \rightarrow x'_i + \Delta x'_i$ where the sampling distance $\Delta x'_i$ need not be constant. The probability in each bin is then given by the Poisson distribution.

$$p_i = \frac{\hat{g}_i^{g_i} \exp(-\hat{g}_i)}{g_i!}$$

The natural logarithm of the likelihood function is then

$$\ln L = \sum g_i \ln \hat{g}_i - \sum \hat{g}_i - \sum \ln g_i!$$

but from the system equation

$$\hat{g}_i = \sum h_{ij} \hat{f}_j$$

so we can substitute for the source estimate and set the partial derivatives with respect to the source elements to zero

$$\frac{\partial \ln L}{\partial \hat{f}_j} = 0$$

The solution is

$$0 = \sum_i \left(\frac{g_i}{\hat{g}_i} - 1 \right) h_{ij}$$

The best fit function (vector) which satisfies this equation can be found by the Richardson-Lucy (RL) algorithm (Richardson 1972, Lucy 1974).

$$\hat{f}_j^{new} = \hat{f}_j \frac{\sum_i \frac{g_i}{\hat{g}_i} h_{ij}}{\sum_i h_{ij}}$$

This scheme has been proven to converge by Shepp and Vardi (1982). It is this RL algorithm that has been used extensively to restore the Hubble Space Telescope images prior to the optical refurbishment.

Maximum likelihood can also be used on event data directly without binning. In this case we assign a probability density to each event using a source function. The expected mean count density for position x' is given by

$$\hat{g}(x') = \int y(x, c_1, c_2, \dots) h(x', x) dx$$

and using the Poisson distribution the probability for $m(x')$ events over the interval $x' \rightarrow x' + \delta x'$ is

$$p(x') = \frac{(\hat{g}(x')\delta x')^{m(x')} \exp(-\hat{g}(x')\delta x')}{m(x)!}$$

If $\delta x'$ is small enough we can ensure $m(x')$ is either 0 or 1. The logarithm of the likelihood function is found by summation over all possible event positions

$$\ln L = \sum m \ln \hat{g}(x') - \sum \hat{g}(x')\delta x' - \sum m \ln \delta x' - \sum \ln m(x)!$$

The last two terms are independent of the parameters c_1, c_2, \dots and are therefore of no consequence. The second term is just the total expected count

$$\sum \hat{g}(x')\delta x' = \hat{C}$$

The first term only contributes when $m_i = 1$ so we can redefine the logarithm of the likelihood function by missing out the uninteresting terms and summing over the detected events

$$\ln L' = \sum_i \ln \hat{g}(x'_i) - \hat{C}$$

We can then use a search algorithm, varying the parameters c_1, c_2, \dots to find the stationary value (maximum) of $\ln L'$. This technique is particularly good if the number of events is rather small and has been used successfully in Gamma and X-ray astronomy for point source searching and fitting. Some finer points about such a procedure are discussed by Cash (1979).

12 The Bayesian method verses the classical approach

In the classical approach to statistical inference parameters and associated confidence limits are estimated directly from the data. It is assumed that if a hypothetical infinite number of identical experiments were carried out then, say, 90 percent of all the confidence intervals calculated from the data would contain the true parameter values. Thus the classical approach is hunting for some true values of the parameters known only to nature

and the confidence level is a statement about a putative population of observations or observers. The only extra information used to arrive at the results is associated with the process of observation.

The problem with the classical approach is how to justify the assigning of asymptotic properties to a particular, single, estimate. The method does not concern itself with the statistics of the parameters only the statistics of the data themselves. You have to make assumptions about the observation process to arrive at results concerning just that process. More importantly you can't turn the above statement around and say that there is a 90 percent chance that the true value of a parameter lies within a particular confidence interval derived from just one data set, i.e. the actual data set.

Conversely Bayesian methods require that a priori knowledge about some parameter, for example a probability distribution, should be combined with information from the data to yield an a posteriori probability distribution describing a property of the source of the parameters. You can then make some probability statement about the parameter values, for example that the probability of a value being within a given interval is such and such. Thus the Bayesian approach turns the problem around and asks what is the probability that possible parameter values give rise to the observed data?

The Bayesian method is based on the use of Bayes' theorem (T. Bayes 1763) which follows directly from the definition of conditional probabilities.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

In the context of data analysis this is more easily expressed in words; the probability of result A given the data B is equal to the probability of getting data B given that the result is A multiplied by the unconditional probability of getting result A divided by the unconditional probability of getting data B . If there is well defined set of possible results A_i then the denominator can be written as $P(B) = \sum P(B|A_i)P(A_i)$. Classically the result of some measurement is the value of a previously unknown but fixed quantity. However the Bayesian use of Bayes' theorem requires that such a quantity is ascribed a probability distribution which describes a priori the likely outcome. If this is done and some data B_1 are achieved then the theorem can be used to assign a conditional probability $P(A|B_1)$ which is an a posteriori probability. If further experiments are performed to ascertain a value for A then the value of $P(A)$, the so called prior, can be replaced by $P(A|B_1)$.

$$P(A|B_2) = \frac{P(B_2|A)P(A|B_1)}{P(B_2)}$$

and so a chain of scientific inference is built up. It is implicit in $P(A|B_2)$ that the result

B_1 is already known and taken into account. The Bayesian method provides a neat way of rationalising scientific discovery although it is debatable whether or not things actually work like this in practice.

The problem with this approach is how to express prior ignorance about the parameters. This is different but no worse than guessing the properties of the observation process required in the classical approach. Both methods require assumptions to be made. Both methods can yield improved results if more data are taken. Both methods lead to the same answer in many cases. What is important is that you are aware of the approach taken and you don't make statements about the results which are inappropriate.

The Bayesian approach is closer to the intuition of astronomers or physicists. They are not just analysing the observation process but trying to infer something about the source or phenomenon under observation. In this sense the Bayesian astronomer wants to go further than the classical statistician and the Bayesian method is an augmentation of purely statistical ideas. References [7], [20] and [12] contain useful discussions about classical verses Bayesian statistics in the context of physics and astronomy.

Conventional Chi-squared fitting lies firmly in the classical camp. All the required properties are assumed of the data or estimated directly from the data. However the results are dependent on the assumptions made about the measurement process and they are not free from prejudice. You can change the result by assuming something different about the data.

The Wiener filter described above is an example of the Bayesian method. The required filter is constructed using an *a priori* estimate of the covariance matrix (function) $[c_f]$ of the input vector \underline{f} which does not come from the data. Deciding on the form of $[c_f]$ is a problem but it does mean that prejudice or prior knowledge about the required distribution can be made to influence the results. For example pictures obtained in medical imaging are very different from astronomical plates. The former contain extended diffuse structures while the latter tend to be full of bright points (stars!) so the power spectrum of medical images tends to roll-off at high spatial frequencies whereas stellar images have a flat extended power spectrum. It is reasonable that the Wiener filter applied in these two cases is different. This topic is discussed in some detail by Hunt (1977) [11].

References

- [1] Andrews H.C. and Hunt B.R., 19***, "Digital image restoration", Prentice Hall
- [2] Bryan R.K. and Skilling J., 1980, Mon.Not.R.astr.Soc.,190
- [3] Cash W., 1979, Ap.J., 228, 939

- [4] Bevington P.R. and Robinson D.K., 19***, “Data Reduction and Error Analysis for the Physical Sciences”, 2nd Edition, McGraw-Hill, Inc.
- [5] Blackman R.B. and Tukey J.W., 1958, “The measurement of power spectra”, Dover publications Inc.
- [6] Brandstätter A., Swift J., Swinney H.L. and Wolf A., 1983, Phys. Rev. Lett. Vol. 51, No. 16, 1442
- [7] Eadie W.T., Drijard D., James F.E., Roos M. and Sadoulet B. “Statistical methods in experimental physics”, 1971, North-Holland, Amsterdam
- [8] Hauser M.G. and Peebles P.J.E., 1973, Ap. J. 185, 757-785
- [9] Helstrom C.W., 1967, J.Opt.Soc.Am., 57, 3, 279
- [10] Hunt B.R., 1971, IEEE Trans. AU-19, 4
- [11] Hunt B.R., 1977, “Bayesian methods in nonlinear digital image restoration”, IEEE Trans. Computers, C-26, 3, 219
- [12] Kraft R.P., Burrows D.N. and Nousek J.A., 1991, Ap. J. 374, 244-355
- [13] Lahav O., Ficher K.B., Hoffman Y., Scharf C.A. and Zaroubi S., 1994, Ap.J. 423, L93-L96
- [14] Lampton M., Margon B. and Bowyer S., 1976, Ap.J., 208, 1177
- [15] Leahy D.A., Darbro W., Elsner R.F., Weisskopf M.C., Sutherland P.G., Kahn S.M. and Grindlay J.E., 1983, Ap. J., 266, 160
- [16] Lichtenberg A.J. and Lieberman M.A., 1983, “Regular and stochastic motion”, Springer-Verlag
- [17] Lucy L.B., 1974, A.J., 79, 745
- [18] Packard N.C., Crutchfield J.P., Farmer J.D. and Shaw R.S., 1980, Phys. Rev. Lett. Vol. 45, No. 9, 712
- [19] Peitgen H.-O. and Saupe D. Eds., 1988, “The science of fractal images”, Springer-Verlag
- [20] Pollock A.M.T., Bignami G.F., Hermsen W., Kanbach G., Lichti G.G., Masnou J.L., Swanenburg B.N. and Wills R.D., 1981, Astron. Astrophys. 94, 116-120
- [21] Press W.H., Teukolsky S.A, Vetterling W.T. and Flannery B.P., 1992, “Numerical Recipes in Fortran - The Art of Scientific Computing”, 2nd Ed., Cambridge University Press
- [22] Rayner J.N., 1971, “An introduction to spectral analysis”, Pion Limited

- [23] Richardson B.H., 1972, *J.Opt.Soc.Am.*, 63, 55
- [24] Roux J.C., Rossi A., Bachelart S. and Vidal C., 1980, *Phys. Lett.* Vol. 77A, No. 6, 391
- [25] Scharf C., Hoffman Y., Lahav O. and Lynden-Bell D., 1992, *Mon.Not.R.astr.Soc.*, 256, 229-237
- [26] Shepp L.A. and Vardi Y., 1979, *IEEE Trans. Medical Imaging*, MI-1, 113
- [27] Spiegel M.R., 1961, "Statistics", Schaum's Outline Series, McGraw-Hill
- [28] Spiegel M.R., 1974, "Fourier Analysis", Schaum's Outline Series, McGraw-Hill
- [29] Wiener N., 1949, "The extrapolation, interpolation and smoothing of stationary time series", p84, John Wiley and Sons, Inc.,
- [30] Wilks S.S., 1938, *Ann. Math. Stat.* 9, 60
- [31] Wolf A., Swift J.B., Swinney H.L. and Vastano J.A., 1985, *Physica* 16D, 285-317
- [32] Wright E.L., Bennett C.L., Górski K., Hinshaw G. and Smoot G.F., 1996, *Ap.J.* 464, L21-L24
- [33] Wright E.L., Smoot G.F., Bennett C.L. and Lubin P.M., 1994a, *Ap.J.*, 436, 443
- [34] Yu J.T. and Peebles P.J.E., 1969, *Ap. J.* 158, 103-113