planets4, 7/8/1996

DETECTION OF EXTRA–SOLAR–SYSTEM PLANETS†

E.T. Jaynes
Arthur Holly Compton Laboratory of Physics
Washington University, St. Louis, MO 63130

Abstract: Stimulated by a proposal of Wm. Hayden Smith and co–workers to detect planets on nearby stars by high resolution imaging, we speculate on the appropriate data analysis method, making use of probability theory to perform the optimal deconvolution of the point–spread function. In this preliminary study, we seek to understand what probability theory has to say about the fundamental problem, by analyzing a simple one–dimensional version. The necessary theoretical principles are developed in the thesis of G. L. Bretthorst (Washington University, May 1987) and in Bretthorst (1988).

Our main message is this: once one is committed to using a computer to analyze the data, the high–resolution imaging problem is completely changed. What one has tried to do in the past by fancy optical and mechanical engineering feats (apodizing, image stabilizing) can be done far better, and at a small fraction of the cost, by the computer.

CONTENTS

One–Dimensional "Baby" Version of the Problem

Our optical system has a point–spread function $(\sin x/x)^2$. A "star" whose image should be a sharp point at position $x = a$ then produces a smeared image proportional to

$$G_1(x) = \frac{\sin^2(x-a)}{(x-a)^2} \tag{1}$$

and a "planet" at $x = b$ gives an offset diffraction pattern

$$G_2(x) = \frac{\sin^2(x-b)}{(x-b)^2}. \tag{2}$$

These are the "model functions" to be built into the computer program. The star and planet have brightnesses A_1, A_2 and so they produce jointly a smeared image

$$f(x) = A_1\, G_1(x) + A_2\, G_2(x) \tag{3}$$

This is observed at the positions $(x_1 \ldots x_N)$, getting a data set $D = (d_1 \ldots d_N)$:

† In *Maximum–Entropy and Bayesian Methods in Science and Engineering*, Vol. 1, G. J. Erickson & C. R. Smith, editors, Kluwer Academic Publishers (1988); pp. 147–160. Some typographical errors in the published version are corrected here.

$$d_i = f(x_i) + e_i, \qquad 1 \leq i \leq N \tag{4}$$

where the e's are white (*i.e.* uncorrelated) Gaussian noise measurement errors. Don't worry about the assumption of whiteness; it will turn out that this assumption can be removed almost trivially at the end, as explained below, so let's keep the problem simple for now.

We shall assume that the noise level is not known in advance, so the computer must estimate it from the data, and use it to determine the accuracy of the other estimates. If the true noise level is known, this extra information can also be given to the computer and it will enable us to improve the accuracy of the estimates, but only slightly.

And for this first look at the problem, it doesn't matter whether the noise e is thought of as measurement noise in the apparatus, atmospheric turbulence noise, or "photon noise" in the phenomenon; the computer is going to estimate the total noise in the data and make proper allowance for it, whatever its source. Later we will see how to teach the computer to distinguish between atmospheric distortion and true noise.

Also, don't worry about the "assumption" of Gaussianity; this will turn out to be not really an assumption at all, but rather the most conservative assignment we could make. If we have any additional information about the noise, which would lead us to assign a nonGaussian probability distribution, this can be built into the computer program and will enable one to get slightly better results than we will obtain below. However, they will not be much better unless that information leads to a wildly nonGaussian noise probability distribution, with a sharp upper bound cutoff.

In short, all the assumptions we are making now are conservative, and removing them will enable us to do still better (at the cost of more computation; but today computing power is plentiful and cheap). The real power of computers to perform sophisticated data analysis is only now beginning to be realized.

The computer's job is now: given the model functions (1), (2) and the data D, tell us whether there is evidence for existence of a planet, and if so, give us the best estimates of the separation $r = b - a$ and the relative brightness of planet and star; and indicate the accuracy of those estimates.

The principles of probability theory, explained in Larry Bretthorst's thesis, determine the data analysis procedure to answer these questions in a way that is optimal (*i.e.,* makes full use of all the relevant information in the data) but is also conservative (*i.e.,* does not mislead us as to the accuracy of its conclusions). The reader is assumed familiar with the general formalism and notation used in the Bretthorst thesis, and we merely apply them to the present problem.

First, the computer is given the model functions $G(x)$, either analytically or measured at the observation points. Then the interaction matrix is

$$g_{jk} = \sum_{i=1}^{N} G_j(x_i)\, G_k(x_i) \tag{5}$$

and this matrix will be calculated by the computer program. For now, approximate it by an integral:

$$g_{12} = \int_{-\infty}^{\infty} G_1(x) G_2(x) dx = 4\pi \left[\frac{2r - \sin 2r}{(2r)^3}\right] \tag{6}$$

It depends only on $r = b - a$, the planetary separation. As $r \to 0$,

$$g_{12} \to g_{11} = \frac{2\pi}{3} \tag{7}$$

[Check: verify $\int (\sin x/x)^4 dx = 2\pi/3$ directly.] Therefore, define

$$g_{12} = \frac{2\pi}{3} u(r) \tag{8}$$

where

$$u(r) = 6 \left[\frac{2r - \sin 2r}{(2r)^3} \right] \tag{9}$$

is the overlap function, normalized to $u(0) = 1$. Note, for later purposes, that as $r \to 0$, $u(r)$ is given asymptotically by

$$u(r) \sim 1 - \frac{1}{5} r^2 + \frac{2}{105} r^4 + \ldots \tag{10}$$

while as $r \to \infty$,

$$u(r) \sim \frac{3}{2} r^{-2} + \ldots \tag{11}$$

The matrix g now becomes

$$g = \frac{2\pi}{3} \begin{bmatrix} 1 & u(r) \\ u(r) & 1 \end{bmatrix}. \tag{12}$$

Since the diagonal elements are equal, a fixed transformation will diagonalize this for any value of r; that is, the orthonormal model functions always have the form

$$H(x) = C \left[G_1(x) \pm G_2(x) \right]$$

so, supplying the proper normalization factors, we have

$$H_1(x; a, b) = \left[\frac{3}{4\pi(1+u)} \right]^{1/2} \left[G_1(x) + G_2(x) \right] \tag{13a}$$

$$H_2(x; a, b) = \left[\frac{3}{4\pi(1-u)} \right]^{1/2} \left[G_1(x) - G_2(x) \right] \tag{13b}$$

which somehow remind one of molecular orbitals. As we see from (1), (2), and (9), they contain the quantities of interest (a, b) as parameters. Once the computer is set to calculate these functions for all (x, a, b), we are ready to analyze any number of data sets (*i.e.,* any number of stars) with it.

The Computation Algorithm

Given a data set $(d_1 \ldots d_N)$, calculate the projections of the data onto the $H(x)$ functions:

$$h_j(a, b) = \sum_i H_j(x_i)\, d_i\,, \qquad j = 1, 2 \tag{14}$$

then a jointly sufficient statistic, which contains all the information the data have to give us about the unknown parameters (a, b) is simply

$$\sum h^2 = 2\overline{h^2}(a, b) = h_1^2 + h_2^2. \tag{15}$$

If the computer is to estimate the noise level from the data and use the student t–distribution, it must also calculate

$$\sum d^2 = N\overline{d^2} = \sum_i d_i^2. \tag{16}$$

Then the joint posterior probability density function for the parameters (a, b), as derived in the Bretthorst thesis, is proportional to

$$p(a, b \mid D, I) \sim \left[\frac{1}{\sum d^2 - \sum h^2}\right]^{(N-m)/2} \tag{17}$$

where m (=2 in the present case) is the number of model functions that we are fitting to the data. But the only parameter of interest is $r = b - a$, so go to the mean and relative coordinates

$$R = (a + b)/2; \qquad r = b - a. \tag{18}$$

The Jacobian of the transformation (18) is one, so the joint posterior probability density for R and r is the same quantity:

$$p(R, r \mid D, I) = p(a, b \mid D, I). \tag{19}$$

The final step is to integrate out the uninteresting parameter R, getting a function of r :

$$p(r) = p(r \mid D, I) = \int dR \; p(R, r \mid D, I) \tag{20}$$

which tells us everything the data have to say about the planetary separation, independently of the brightness and absolute positions of star and planet.

If the data contain evidence for a planet, then $p(r)$ will exhibit a peak at a value of r that represents the "best" estimate of its distance, and the width of the peak will indicate the probable error of that estimate.

The computer program can also give us its best estimate of the relative brightness of the star and planet, and the accuracy of that estimate. The way to do this is explained in the Bretthorst thesis; it is a small further detail available at essentially zero additional computation cost (the estimated brightnesses are just linear combinations of the coefficients h_i already calculated).

If there is no evidence in the data for a planet, then the computer will not be able to fit the data to a model function with $r > 0$, any better than it can to the function with $r = 0$. The posterior distribution $p(r)$ will then peak at $r = 0$, and it will indicate by its width the probable error in detecting the position of a very close planet. That is, if the star does have a planet of detectable brightness, it is extremely unlikely to be further from the star than the width of that distribution.

A major feature of this data analysis procedure is that, thanks to the elimination of nuisance parameters, we can combine the data from many different measurements into one grand final distribution. For example, suppose we have developed a computer program that eliminates atmospheric distortion as a nuisance parameter. Now we take 1000 successive data sets on the same star, in a time so short that the planet has not moved appreciably, perhaps a week. The computer will make allowance for the atmospheric distortion separately in each data set, and the total evidence concerning r will be given by the product of the separate $p(r)$ distributions; we merely add up all the $\log p(r)$ functions from the individual data sets.

The point of this is that, even though any one data set might not have a high enough signal/noise ratio to draw any definite conclusions, the totality of them will. But this evidence could not be extracted by a single analysis of the pooled data, because the atmospheric distortion will vary erratically from one data set to the next. From the pooled data one would be able to consider only a smeared–out average over those erratic variations, with resulting far poorer resolution. The difference between a long time exposure and what an observer can detect at a single instant of good seeing gives only a slight indication of how much this can help.

The computer is not only powerful, but flexible. In the course of a research project, one is almost sure to learn new things about the phenomenon being studied and the capabilities of the apparatus, which change one's views about how the data should be processed. If one is trying to do the processing by optical or mechanical engineering feats, this might involve rebuilding a telescope. If a computer is doing all the data processing, one needs only rewrite a few lines of the program code.

Intuitive Meaning of the Result

One problem we have is that the algorithm and final result (17) are so slick and efficient that, at first glance, it is far from obvious that this is really a sensible data analysis procedure; as it stands, (17) looks unpromising. Put differently, intuition alone would never have been powerful enough to tell us that this is the thing to do. But intuition can be educated; so let's look at the result more closely to see some of the wonderful things that are hiding in (17); first by rewriting it in a form like the usual student t–distribution notation.

We are most interested in the region of the maximum. Let $\hat{a}, \hat{b}$ be a point where the sufficient statistic (15) reaches its absolute maximum (there are generally two such points, because the distribution is symmetric in a and b; so choose the one where $a < b$.) Let $Q(a, b)$ be the departure of the sufficient statistic from that maximum:

$$\sum h^2 = \sum h^2_{max} - Q(a, b) \tag{21}$$

and define the quantity s^2 by

$$\sum d^2 - \sum h^2_{max} = (N - m)s^2. \tag{22}$$

Now the joint posterior probability density (17) is, to within a normalization constant,

$$\left[\frac{1}{1 + \frac{Q(a,b)}{(N-m)s^2}}\right]^{(N-m)/2} \tag{23}$$

which is the form in which we are used to seeing the t–distribution. This is still exact everywhere, only written in different notation; but one sees that we have set it up for a Gaussian approximation. When N becomes large, (23) goes into

$$\exp\left[-\frac{Q(a, b)}{2s^2}\right]. \tag{24}$$

But in the neighborhood of the peak, $Q(a, b)$ can be expanded as a quadratic form:

$$Q(a, b) = Q_{11}(a - \hat{a})^2 + 2Q_{12}(a - \hat{a})(b - \hat{b}) + Q_{22}(b - \hat{b})^2 \tag{25}$$

and so, in spite of first appearances, the exact distribution (17) is very nearly, in the most important region, a bivariate Gaussian; quite accurately so if we have a lot of data. The quantity

$$s^2 = \frac{\sum d^2 - \sum h^2_{max}}{N - m} \tag{26}$$

(where we are still writing m as the number of model functions being fitted to the data, to show the general formula) is just the estimate of the mean–square noise level that probability theory is making.

Let us explain this more fully. The computer finds a "best" model function, by which we mean the one that makes the best least–squares fit to the data, out of the class of possible model functions (*i.e.,* number and range of parameters) that we have specified in setting up the problem. The numerator of (26) is the sum of the squares of the residuals for that "best" model, a measure of how well the best model is able to fit the data.

Now as is clear from (4), anything in the data that the computer cannot fit to that best model, it is obliged to consider as "noise." The denominator of (26) indicates that this total mean–square noise is then ascribed equally to the remaining $(N - m)$ degrees of freedom that are not being fit to the model.

Generally, the integration (20) over R should be performed numerically by the computer. But in the Gaussian approximation (25) we can do it analytically, with the result that the posterior probability density $p(r)$ for r alone is proportional to

$$\exp\left[-\frac{1}{2s^2}\,\frac{Q_{11}Q_{22}-Q_{12}^2}{Q_{11}+2Q_{12}+Q_{22}}(r-\hat{r})^2\right] \tag{27}$$

where $\hat{r} = \hat{b} - \hat{a}$ is the planetary distance at which $\sum h^2$ peaks. Thus the "best" estimate and probable error of that estimate would be given by the (mean) $\pm$ (standard deviation) of (27):

$$(r)_{est} = \hat{r} \pm s\left[\frac{Q_{11}+2Q_{12}+Q_{22}}{Q_{11}Q_{22}-Q_{12}^2}\right]^{1/2}. \tag{28}$$

In this approximation, the accuracy of the distance estimate depends on the estimated RMS noise level s and the coefficients in the expansion (25). Roughly speaking, the larger those coefficients, the sharper the peak in the sufficient statistic, and therefore the more accurate the estimate, as common sense would lead us to expect.

Probability theory never makes gratuitous assumptions about other things that might be in the data. As noted, it simply dumps out everything that it cannot fit to the best model into a bin called "noise," without passing any judgment about whether it is "systematic" or "random."

This is the automatic built–in safety device that prevents probability theory from making over–optimistic claims about the accuracy of its estimates. As is shown in the "Bessel inequality" section of the Bretthorst thesis, anything in the data that cannot be fit to the best model, increases the estimate (26) and broadens the posterior distribution (24), increasing our error estimates for the quantities of interest.

But any further information that we have about other systematic effects that might be in the data, can be given to the computer in the form of more model functions, or more flexible model functions. Then it will be able to fit the data to the new model better, it will perceive by (26) that the noise is smaller than previously estimated, and so it will be justified in claiming smaller estimated errors for the quantities of interest.

Therefore, any information we have about systematic effects that might be in the data, whether or not they are of interest to us in the present problem, and whether we consider them to be part of the "model function" or part of the "noise," should be put into our model. In effect, this tells the computer to be on the lookout for such variations and make allowances for them, so that they do not deteriorate our estimates of the quantities of interest.

We stress the extreme importance of following this policy, which was revealed by the analysis of NMR data in the Bretthorst thesis. Here we are interested in determining the oscillation frequencies present; one would at first suppose that the way to do this is to take the Fourier transform of the data. But an additional systematic effect, exponential decay of the oscillations, is present. If the decay rate is put into the model, then integrated out as a nuisance parameter, one obtains orders of magnitude more accurate frequency estimates than are given by a Fourier transform. The implications of this for the planet problem must not be missed.

In particular, the computer can be taught to distinguish between atmospheric distortion and true noise. A human telescopic observer can to some extent make mental allowance for the atmospheric distortions he is seeing, and concentrate on what the shaky image is telling him about the real object. But a computer can do such a job far better than a human can, if it knows what specific kinds of distortions are to be expected, so that it can make allowance for them quantitatively.

This has been a fast tour through the solution for an oversimplified one–dimensional version of the problem, in order to give an intuitive feel for what is to be done, and what the results will mean.

Generalizations

In the real problem, the solution will need to be fixed up to allow for a dozen picky little details that we have left out above. We do not try to list them all here, because the realities are so complicated that one will not be able to anticipate all these details until a preliminary version of the program is running on real data. But we can indicate the more obvious ones.

The first is, of course, that all this must be restated in two–dimensional form. This is an entirely straightforward computer programming job; much more complicated programs than this have been written and run successfully with the Bretthorst algorithm. So we do not anticipate any problems here.

Perhaps next in importance is to deal with atmospheric turbulence effects. In the past, one has tried to capture instants of good seeing, or to stabilize the image position. But these are things that the computer can do much more easily.

A good example of this process, different in one detail but just the same in principle, is in the Bretthorst thesis where he considers how to eliminate trend distortion from the economic data, so as to detect periodicities if the data have evidence for them. Superposed on the periodic model function of interest, is a trend "nuisance function" $T(x)$ which the computer program estimates and removes.

In the present problem, the different detail is that the nuisance function would not be an additive term in the model function, of the form

$$f(x) + T(x)\,,$$

but it would specify the likely distortions in the independent variable x that might occur:

$$f[x + q(x)].$$

Thus, just as Bretthorst expanded the trend function $T(t)$ in the orthogonal functions and eliminated the coefficients as nuisance parameters, we would now express the distortion function $q(x)$ in some suitable functions (probably a power series in the first tryout), and integrate out the coefficients.

We think that writing and testing a computer program to do this would be great fun because the details are new, and the result would be something never before seen. But some detailed information about the actual kinds of distortion $q(x)$ that occur with real telescopes would enable one to write a smarter program, that does it better or more efficiently.

Also, the computer can be told to deliver, for each data set analyzed, its best estimate of what the atmospheric distortion function $q(x)$ was for that data set. By comparing the planetary distance estimates and estimated distortions for different data sets on the same star, one could learn a great deal about the specific way in which atmospheric distortion affects information loss, that would help in optimizing the procedure to deal with the most distant possible stars.

One of the practical details ignored above is the size of the pixels for which we have data, and whether this is partially under our control. If it is, there is probably an optimal pixel size for these purposes, and some further theoretical analysis is needed to understand it. Needless to say, too coarse a pixel size will lose information by poor resolution. But too fine a pixel may, we suspect, also lose information by wasting some of the available light in the "fences" between the pixels in the detector or by leakage of charge from one pixel to another.

We stress that the only thing that is important is the *amount of information* contained in the data; its exact form and such things as signal/noise ratio for individual pixels do not matter because the computer can always extract the information if it is there.

Presumably, these considerations will not matter for the first tryout of the method; we think that if the near stars have planets – at least, on the Jovian scale – they will be found with the present detectors. But these considerations will become important in the final optimization to study more distant stars.

Also, in the final optimization one will need to take into account correlations in the noise of adjacent pixels. This will enable more sensitive detection and more accurate estimates of planetary distance, because for a given estimated mean–square error the noise vector will be confined by this information into a smaller volume of N–dimensional sample space, and some data vectors that lie just outside that volume will then convey to us significant information about the planet that they could not convey (because they were lost in the noise volume) before we knew of the correlations.

If understanding of the noise mechanism is good enough to determine some respect in which it is known to depart from Gaussian, this will also enable better sensitivity and accuracy for a given mean–square noise level; but it is unlikely to help very much unless there is some known hard constraint on the possible magnitude of the noise.

The computer is readily programmed to take into account these different noise distributions. Fortunately, the change likely to help the most is also the easiest to carry out.

Nonwhite Noise

To take this into account, first get the computer program running which finds the optimal solution discussed above, given the model function matrix

$$G_{ij} = G_j(x_i), \qquad 1 \le i \le N, \quad 1 \le j \le m$$

and data vector $D = (d_1, d_2, \ldots)$ for white (*i.e.*, uncorrelated) noise values e_i. This is the program developed in the Bretthorst thesis.

But now suppose that the inverse correlation matrix of the noise is known to have instead the form M/σ^2 where σ is the RMS magnitude of the noise, known or unknown, and M is an $(N \times N)$ matrix indicating the correlation coefficients. No problem; just use the first program as a subroutine, and write a driver program that feeds it instead the massaged matrix and massaged data

$$G_0 = M^{1/2} G \qquad \text{and} \qquad D_0 = M^{1/2} D, \tag{29}$$

and it will generate the optimal solution for the nonwhite noise.

To prove this, we need only note that the basic sampling distribution for the nonwhite case, from which all else follows, is proportional to $\exp(-Q/2\sigma^2)$, where

$$\begin{aligned} Q &= \Sigma_{ir}[d_i - \Sigma_j G_{ij} A_j]\, M_{ir}[d_r - \Sigma_k G_{rk} A_k] \\ &= [D - GA]^T M [D - GA] \end{aligned} \tag{30}$$

and if M is the unit matrix, this reduces to the uncorrelated case, analyzed in the Bretthorst thesis. So if we had a computer programmed to do the bigger calculation defined by (30), but we gave it a problem in which M is the unit matrix, it would do just the Bretthorst calculation.

Now the crucial point is that M is guaranteed to be symmetric and positive definite, so it can be factored uniquely as

$$M = M^{1/2} M^{1/2} \tag{31}$$

and the two factors can be absorbed into the vectors on either side in (30). Doing this, and using the notation (29), we see that the quadratic form (30) is equal to

$$Q = [D_0 - G_0 A]^T [D_0 - G_0 A] \tag{32}$$

which is of the form used in the Bretthorst calculation. Therefore, if the "bigger" program is fed D, G, and the true matrix M; while the Bretthorst program is fed the massaged D_0, G_0, they will do just the same actual calculation. QED

Warning: Don't Apodize!

Apodizing was a pre–computer way of making the optical system do, crudely and inaccurately, a small part of this computation. But apodizing has serious limitations that a computer does not have. It does suppress the wiggles in the point–spread function; but at a cost that is today not only unacceptable, but unnecessary. We note some of the difficulties with apodization, which can be overcome easily with computers.

Psychologically, apodizing always leaves us not knowing whether something better could have been done, because it is only an intuitive, *ad hoc* device not derived from any theoretical principles or optimality criterion. For this same reason, it leaves us unable to judge the accuracy of our final results.

More serious, apodizing throws away valuable, relevant information, in two respects. The first is simply that any tampering with the pupil amplitude transmittance function throws away photons and sacrifices signal/noise ratio.

To see the second, more fundamental mode of information loss, note that if the pupil transmittance function of the original unapodized optical system is known, a computer can always calculate the apodized diffraction pattern for any apodizing scheme you please; there is no need to alter the optical system for this. But the transformation is not reversible; given an apodized diffraction ring, no computer can recover the original unapodized optical system.

Mathematically, the apodizing operator has no inverse; many different unapodized systems, which would give different information about the object being seen, all generate the same apodized image. Conversely, many different objects, which would be distinguishable by the original unapodized optical systems, all generate the same apodized image and become indistinguishable. That is what we mean by "information loss."

The purpose of a data analysis procedure is to extract as much information as it is possible to get out of the data, pertaining to the question of interest. A philosophy of data analysis that tells us to start by throwing away some of the information in the data, may be of some use as a "quick and dirty" expedient; but fundamentally it is just not a rational way of looking at the problem.

Some years ago, Ronald Bracewell pointed out to me a problem that is very similar both in topic and in theory. Let us recall it for the lesson it teaches us. In the 1950's, the Hanbury Brown–Twiss (HBT) interferometer for measuring stellar diameters was a wonderful, new, almost magical thing and few people understood correctly how it worked.

The Michelson interferometer had failed at rather short mirror separations because the difference in atmospheric turbulence at the two mirror positions washed out the interference fringes. But the HBT method rectified the signals before combining them, and instead of looking at optical interference fringes, looked at correlations in intensity fluctuations in light falling on the two mirrors. This still contained information about the stellar diameter, and in spite of an inherently lower signal/noise ratio for a given amount of light intercepted, the HBT interferometer was able to operate at greater separations, thus measuring smaller angular diameters.

Not surprisingly, some works appeared praising the virtues of the HBT way of looking at the problem, as a great advance in understanding. But Bracewell pointed out that the Michelson interferometer is in principle delivering far more information than is the HBT one; and it does not make sense to suppose that an instrument that yields less information about stellar diameters can be more informative about them.

The problem with the Michelson interferometer was that it was delivering information faster than the technology of the time could handle. The interference fringes were not absent; they were just moving rapidly. If one could record the details of the fringe position and visibility in real time, and analyze the resulting masses of data by computer, the Michelson interferometer would emerge as superior.

The HBT method was, like apodization, a pre–computer way of getting the optical/electronic system to do, crudely, a small part of the computation that an optimal data analysis method would perform. Both of these quick–and–dirty methods were good enough to be usable; but both deliver results far inferior to what a Bayesian computer analysis could give today.

The same lesson was learned again, ten years later, when the Maximum Entropy spectrum analysis method of John Parker Burg was announced. Previously, the Blackman–Tukey (BT) method had removed unwanted "side–lobes" in the Fourier transform of the data by a "lag window" which removed the sharp edges of the measured autocovariance at the end of the data record, shading it smoothly to zero. But the most popular "Hanning window," in removing most of the side–lobes, also threw away half the resolution; BT spectra have lines twice as wide as did the original Schuster periodogram.

In 1967, Burg pointed out that the BT method throws away crucially important information. His maximum entropy method, which conserves all the information in the data, can display spectrum lines orders of magnitude sharper than the BT ones, when the data contain evidence for them. The BT method rapidly became obsolete among those who tried the maximum entropy method, not for philosophical reasons, but because of the superior computer printouts.

But some holdouts remained. John Tukey, in a meeting in Princeton in December 1980, pointed out to the present writer that his windowing method for removing side–lobes is mathematically just a one–dimensional version of apodization – in the belief that this would persuade me of the merits of the BT method! Instead, this made me see clearly where the defects of apodization lie, and how to correct them.

When the application is important enough to justify tying a computer to the optical system, we don't have to put up with such limitations any more. The new rules of conduct are:

(1) Keep your optical system clear and open, gathering the maximum possible amount of light (*i.e.,* information).

(2) Don't worry abut wiggles in the point–spread function; the computer will straighten them out far better than apodizing could ever have done, at a small fraction of the cost.

(3) For the computer to do this job well, it needs only to know the actual point–spread function $G(x)$, whatever it is. So get the best measurement of $G(x)$ that you can, and let the computer worry about it from then on.

(4) What is important to the computer is not the spatial extent of the point–spread function, but its extent in Fourier transform space; over how large a "window" in k–space does the PSF give signals above the noise level, thus delivering relevant information to the computer? Apodizing contracts this window by denying us information in high spatial frequencies, associated with the sharp edge of the pupil function. But this is just the information most crucial for resolving fine detail! In throwing away information, it is throwing away resolution. Apodizing does indeed "remove the foot;" but it does it by shooting yourself in the foot.